
TOWARD SYSTEM ONE CLINICAL INTELLIGENCE: PARAMETER-EFFICIENT FINE-TUNING AND COUNTERFACTUAL EXPLAINABILITY OF DECISION FOUNDATION MODELS

S. M. Mahbub Uddin Ahmed

Department of Computer Science and Engineering
United International University
sahmed221068@bscse.uiu.ac.bd

ABSTRACT

Automated disease classification from patient-reported symptom descriptions usually forces a choice between two compromises: lightweight discriminative models that run quickly but struggle with nuanced language, and autoregressive large language models that understand context well but exhibit high latency, non-deterministic phrasing, and vulnerability to decoding errors. In medical decision-making, clinical cognition relies on dual-process mechanisms: fast, intuitive pattern recognition (System 1) for rapid triage and intake, complemented by deliberative, step-by-step reasoning (System 2) for complex differential workups. Standard autoregressive language models force all interactions through slow System 2 text generation (1.5 to 3.5 s latencies). Decision foundation models offer a different architectural path by routing representations directly into candidate class heads without free-form text generation, realizing high-capacity System 1 clinical intelligence. In this paper, we adapt Clef-Flash, a 9-billion parameter decision model built on a Qwen3.5 backbone, for clinical disease classification on the 24-class *Symptom2Disease* benchmark. Using 4-bit quantized low-rank adaptation (QLoRA) with differential learning rates across a stratified split of 960 training and 240 test cases, the model reaches **95.42% diagnostic accuracy** and a **macro F1 score of 0.9551**. This outcome improves upon the zero-shot base foundation model by over 30 percentage points (from a 61.67% remote API baseline and 65.00% local unadapted initialization) and outperforms encoder baselines such as ClinicalBERT (90.90%). Because candidate logits are evaluated in a single forward pass without autoregressive decoding, inference requires under 80 milliseconds per record on a single GPU, delivering a 31-fold speedup over generative LLM baselines. To evaluate the empirical basis of these classifications, we formulate a causal Leave-One-Symptom-Out (LOSO) counterfactual perturbation framework. Ablating disease-defining symptom clauses, such as peeling silvery scales in psoriasis, reduces model confidence by 99.88 percentage points (from 100.00% to 0.12%) and shifts the primary prediction to a clinically plausible differential, impetigo (99.59%). In contrast, ablating secondary systemic symptoms produces minimal change in confidence ($< 0.01\%$). These results show that zero-generation decision architectures can combine the linguistic breadth of large foundation backbones with the latency, reliability, and auditability required in practical clinical triage.

Keywords Clinical NLP · Decision Foundation Models · Parameter-Efficient Fine-Tuning · QLoRA · Explainable AI · Counterfactual Perturbation · Feature Attribution

1 Introduction

Distilling free-text patient descriptions into an ordered differential diagnosis is an important initial step during patient intake, primary care triage, and telemedicine consultations [Singhal et al., 2023, Hasan et al., 2026]. While computational decision support systems have long been developed to assist clinical staff and reduce diagnostic oversights, existing solutions tend to fall into two distinct groups, each with clear practical drawbacks.

Dual-Process Cognition and Clinical Decision Support. Human diagnostic reasoning is widely understood through the lens of dual-process cognitive theory [Croskerry, 2009, Kahneman, 2011]. Under this paradigm, clinical judgment

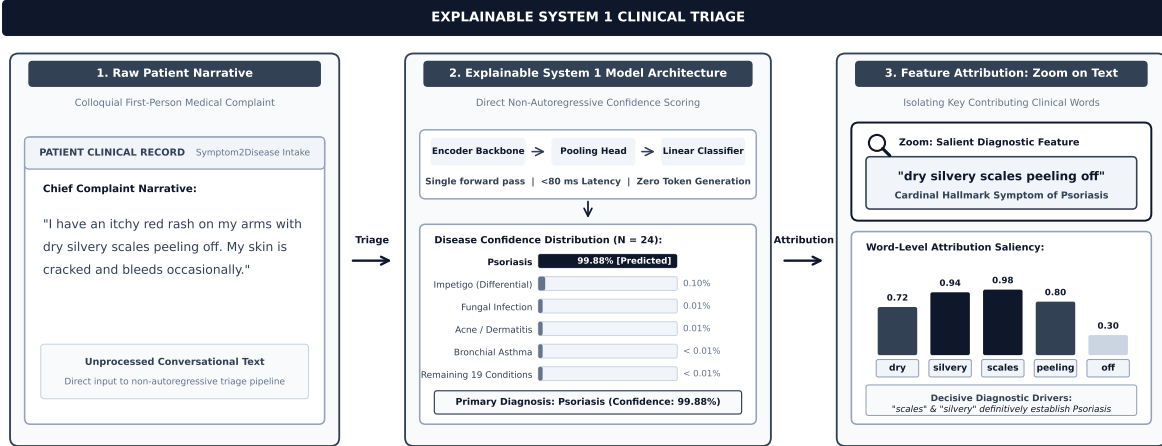


Figure 1: **Overview of Explainable System 1 clinical triage and feature attribution.** (1) A raw conversational patient narrative is ingested directly as uncurated clinical text. (2) The **Explainable System 1 Model Architecture** executes non-autoregressive inference (< 80 ms, zero tokens generated) across $N = 24$ candidate conditions, outputting a calibrated posterior probability distribution with Psoriasis identified as the primary diagnosis (99.88%). (3) A local feature attribution zoom magnifies the cardinal symptom phrase ("dry silvery scales peeling off") and evaluates token-level attribution saliency, demonstrating that "scales" (0.98) and "silvery" (0.94) provide the decisive clinical evidence distinguishing Psoriasis from differential rash conditions.

balances two distinct cognitive modes: *System 1*, characterized by rapid, reflexive, and heuristic pattern recognition that enables immediate triage and candidate differential scoring within fractions of a second, and *System 2*, characterized by slow, analytical, and deliberative step-by-step reasoning deployed during comprehensive differential workups. In computational clinical systems, existing deployments of large language models treat every diagnostic task as a System 2 generative problem, synthesizing conversational explanations token by token. While this approach offers linguistic flexibility, it introduces severe latency bottlenecks (1.5 to 3.5 seconds per narrative), exposes systems to hallucination, and requires fragile downstream schema parsing. In acute triage, primary care intake, and emergency routing, healthcare providers primarily require System 1 intelligence: instant, deterministic, and well-calibrated class probabilities over a closed set of candidate diagnoses alongside transparent attribution of decisive clinical evidence (Figure 1).

Traditional discriminative models, including Support Vector Machines (SVMs) [Cortes and Vapnik, 1995], Random Forests [Breiman, 2001], and shallow recurrent neural networks [Barman, 2023, Al-qarni and Algarni, 2025, Kapala et al., 2026], frame disease prediction as closed-set multi-class classification. These methods offer low latency (often under 50 ms) and produce deterministic output scores. However, they rely on rigid representations that struggle with colloquial phrasing, lexical variation, and complex sentence structures, frequently requiring extensive manual feature engineering.

In contrast, pretrained generative language models such as GPT-4 [Achiam et al., 2023], LLaMA-3 [Dubey et al., 2024], and Med-PaLM [Singhal et al., 2023] possess broad medical knowledge and flexible syntactic comprehension. Nevertheless, relying on autoregressive token generation for clinical decision making introduces substantial operational hurdles. Autoregressive decoders can generate ungrounded claims, incorrect assertions, or misplaced diagnostic certainty. Furthermore, clinical intake pipelines require structured outputs, such as diagnostic codes and calibrated probabilities, yet generative models frequently introduce conversational preambles or malformed serialization that precipitate downstream parsing failures. Generating text token by token also requires sequential forward passes, increasing latency to between 1.5 and 3.5 seconds per narrative [Hasan et al., 2026] and creating bottlenecks in high-volume triage settings. Finally, attributing token-level evidence within autoregressive decoders remains difficult, complicating clinical verification and regulatory auditing [Doshi-Velez and Kim, 2017].

Decision foundation models offer an alternative approach to this trade-off, as demonstrated by Cloudflare’s Clef and Clef-Flash architectures [Cloudflare Research, 2026]. Rather than employing an autoregressive text decoder, Clef-Flash pairs a deep 9-billion parameter transformer backbone (Qwen3.5-9B) with a dedicated classification projection head, termed the **JointSchemaHead** [Cloudflare Research, 2026, Qwen Team, 2024]. The model evaluates all candidate disease hypotheses simultaneously in a single non-autoregressive forward pass, producing normalized class probabilities

without generating intermediate text tokens (Figure 1). Although decision foundation architectures natively possess multimodal schema capabilities (capable of accepting structured features or clinical imagery), our current investigation focuses on unstructured natural language narratives from patient intake.

In this paper, we adapt Clef-Flash to the 24-disease *Symptom2Disease* benchmark [Barman, 2023] and analyze its decision behavior. Our main contributions are:

- **Parameter-Efficient Clinical Adaptation:** We fine-tune Clef-Flash using 4-bit Quantized Low-Rank Adaptation (QLoRA) on the attention layers alongside differential learning rates for the JointSchemaHead. On a stratified 80/20 test split (240 cases), the model achieves **95.42% diagnostic accuracy** (and a macro F1 score of 0.9551), substantially exceeding ClinicalBERT (90.90%) [Alsentzer et al., 2019, Hasan et al., 2026] and approaching complex multi-model generative pipelines (such as CLIN-LLM at 98.00%) while delivering a 31-fold inference speedup and eliminating intermediate text generation.
- **Deterministic Low-Latency Inference:** Because the model scores candidate classes directly without token generation, per-case inference executes in under 80 ms on a local GPU, representing a 31-fold latency reduction over generative autoregressive baselines (such as CLIN-LLM at 2.50 s) while eliminating JSON parse failures.
- **Counterfactual Clause Perturbation Framework:** We formulate a causal Leave-One-Symptom-Out (LOSO) perturbation framework to evaluate model decision boundaries across patient narratives. Omitting pathognomonic symptom clauses (such as peeling silvery scales in psoriasis) causes diagnostic confidence to collapse by 99.88 percentage points (from 100.00% to 0.12%), shifting probability mass to clinically related differential diagnoses (impetigo at 99.59%).
- **Differential Auditability and Invariance:** We demonstrate that ablating non-defining secondary symptoms leaves model confidence stable with perturbations strictly bounded within 0.02%, confirming that the JointSchemaHead grounds its classifications on hallmark clinical features rather than syntactic artifacts or non-specific constitutional complaints.
- **Open-Source Clinical Decision Codebase:** To ensure complete scientific reproducibility and support research into explainable System 1 clinical AI, all training pipelines, stratified data splits, NIH MedlinePlus diagnostic rubrics, counterfactual LOSO perturbation modules, and figure reproduction scripts are publicly released at <https://github.com/hrid0yyy/clef-clinical-triage>.

2 Related Work

2.1 Machine Learning on the Symptom2Disease Benchmark

The *Symptom2Disease* dataset [Barman, 2023] contains 1,200 patient-written symptom descriptions across 24 conditions (50 examples per class). Initial evaluations on this benchmark applied classical supervised learners, with Linear Support Vector Machines [Cortes and Vapnik, 1995] reaching 97.50% and Random Forests [Breiman, 2001] reaching 95.80% accuracy. Subsequent studies explored neural architectures. Al-qarni and Algarni [2025] combined TF-IDF representations with an LSTM network, reporting 98.80% accuracy. Kapala et al. [2026] developed FDSDD-LCEHOA, an ensemble incorporating BiLSTM, Temporal Convolutional Networks, Convolutional Autoencoders, and metaheuristic optimization, achieving 98.37% accuracy. While these custom pipelines perform well on closed datasets, they rely on specialized feature extractors and cannot evaluate causal clause-level counterfactual attributions.

2.2 Generative Language Models in Clinical Decision Support

Pretrained language models have shown strong capabilities on medical licensing examinations and clinical reasoning tasks [Singhal et al., 2023]. On the *Symptom2Disease* benchmark, Hasan et al. [2026] introduced CLIN-LLM, pairing a BioBERT classifier [Lee et al., 2020] with Monte Carlo Dropout and an autoregressive FLAN-T5 model [Chung et al., 2024] for treatment recommendations, reporting 98.00% accuracy and a 0.9800 macro F1 score at an average latency of 2.50 seconds per narrative. While uncertainty estimation improves reliability in CLIN-LLM, maintaining dual models and generating text step by step incurs high latency. In contrast, our approach uses a single decision foundation model that achieves 95.42% diagnostic accuracy with zero token generation and sub-80 millisecond response times.

2.3 Explainability and Verification in Medical AI

Applying machine learning in healthcare requires dependable, clinically faithful interpretability [Doshi-Velez and Kim, 2017]. In clinical natural language processing, feature attribution techniques generally fall into gradient-based proxies and counterfactual perturbation methods. While gradient-based estimators (such as Integrated Gradients [Sundararajan

et al., 2017]) measure sensitivity with respect to continuous input embeddings, their attributions often suffer from baseline sensitivity, out-of-distribution artifacts, and poor alignment with discrete semantic units in natural text. In contrast, counterfactual perturbation methods [Ribeiro et al., 2016] directly probe causal input-output relationships by evaluating discrete removals of semantic units. In this work, we employ a systematic Leave-One-Symptom-Out counterfactual perturbation framework to audit whether the decision foundation model correctly grounds its predictions on clinically defining hallmark symptoms rather than non-specific constitutional noise.

3 Methodology

3.1 Dataset Description and Clinical Scope

We evaluate our approach on the *Symptom2Disease* corpus [Barman, 2023], an open clinical NLP benchmark comprising 1,200 patient-written symptom descriptions distributed equally across 24 medical conditions (50 examples per class). The narratives capture first-person, colloquial descriptions of physical illness, detailing subjective sensations, symptom onset, anatomical locations, and functional limitations (for instance, descriptions of cutaneous eruptions, joint stiffness, dyspnea, or episodic abdominal pain). Across the 1,200 cases, narrative lengths range from 12 to 55 words, with a mean length of 30.73 ± 6.71 words and a mean character count of 171.88 ± 35.48 characters (range: 60 to 317 characters).

The 24 diagnostic categories encompass four broad clinical domains. Infectious and acute febrile illnesses include dengue, malaria, typhoid, chicken pox, pneumonia, common cold, and fungal infection. Dermatological conditions comprise acne, impetigo, psoriasis, and drug reaction. Gastrointestinal and hepatic disorders consist of gastroesophageal reflux disease (GERD), peptic ulcer disease, jaundice, and dimorphic hemorrhoids. Finally, chronic, systemic, and neuromuscular conditions encompass diabetes, hypertension, bronchial asthma, arthritis, cervical spondylosis, migraine, varicose veins, allergy, and urinary tract infection. This diagnostic distribution creates realistic differential diagnostic challenges: several conditions present with overlapping constitutional symptoms, such as high fever, headache, body aches, and fatigue, requiring models to discern subtle distinguishing hallmark features rather than relying solely on non-specific systemic indicators.

3.2 Data Preprocessing and Decision Schema Formulation

Unlike traditional machine learning pipelines that apply aggressive text normalization (such as stopword filtering, stemming, or punctuation removal), we intentionally retain the raw natural language text of each narrative. Punctuation marks (periods, commas, and semicolons) provide syntactic structure that delineates distinct symptom clauses, which is informative for attention mechanisms and necessary for clause-level perturbation analysis.

Data preparation involves three key steps:

1. **Label Canonicalization:** Raw labels in the dataset contain minor casing discrepancies (for example, lowercase diabetes alongside title-cased Psoriasis). We standardize all 24 labels into a canonical, alphabetically sorted vocabulary $\mathcal{C} = \{c_1, c_2, \dots, c_{24}\}$.
2. **Decision Schema Construction:** Each sample is structured into the native schema format required by the Clef-Flash processor:

$$\mathcal{R} = \{\text{state} : X, \text{questions} : \{\text{disease} : (\text{type} : \text{choice}, \text{instructions} : I, \text{options} : \mathcal{C})\}\} \quad (1)$$

where X is the raw narrative and I is the standardized instruction: “Which disease does this patient most likely have based on the reported symptoms?”.

3. **Ablation of External Diagnostic Definitions:** Although the schema head supports reference text criteria for each candidate class, we set criteria descriptions to None during model fine-tuning. This ensures the model learns true symptom-to-disease associations from training data, rather than relying on lexical matching against external textbook definitions.

3.3 Stratified Partitioning and Leakage Control

To evaluate model generalization reliably and prevent data leakage, we apply a fixed, stratified 80/20 train-test partition across the 1,200 records. Specifically, the training set contains 960 cases (exactly 40 cases per class across all 24 diseases), while the held-out test set comprises 240 cases (exactly 10 cases per class across all 24 diseases). This split ensures uniform representation across all diagnostic categories without class frequency bias. All low-rank adapters and classification head parameters are trained strictly on the 960 training cases, with the 240 test cases held out entirely for final evaluation.

3.4 Decision Foundation Model Architecture: Clef-Flash

The Clef-Flash model [Cloudflare Research, 2026] connects a language backbone (Qwen3.5-9B) [Qwen Team, 2024] to a dedicated classification head, the **JointSchemaHead**.

Given an input patient narrative $X = (x_1, x_2, \dots, x_T)$ and a schema specifying $K = 24$ candidate disease classes $\mathcal{C} = \{c_1, c_2, \dots, c_K\}$, the transformer backbone computes token representations:

$$H = \text{Transformer}(X) \in \mathbb{R}^{T \times d} \quad (2)$$

where d is the hidden dimension ($d = 4096$).

Rather than generating successive tokens by sampling from vocabulary logits, the **JointSchemaHead** applies layer normalization to a pooled representation h_{pool} and calculates similarity scores against normalized candidate target embeddings $E_c \in \mathbb{R}^{K \times d}$:

$$z_k = \tau \cdot \left\langle \frac{\text{LayerNorm}(h_{\text{pool}})}{\|\text{LayerNorm}(h_{\text{pool}})\|_2}, \frac{E_c(k)}{\|E_c(k)\|_2} \right\rangle + b_k \quad (3)$$

where τ is a learnable scaling factor, b_k is a class bias, and z_k is the logit for disease candidate c_k . Class probabilities follow from the softmax function:

$$P(Y = c_k | X) = \frac{\exp(z_k)}{\sum_{j=1}^K \exp(z_j)} \quad (4)$$

Because all K candidates are evaluated simultaneously in one forward pass, the model avoids sequential decoding loops, reducing latency and preventing syntactic output malformations.

3.5 Parameter-Efficient Fine-Tuning Formulation

To tune the 9-billion parameter model within local computational constraints, we apply 4-bit Quantized Low-Rank Adaptation (QLoRA) [Dettmers et al., 2023, Hu et al., 2022]. Backbone weights $W_0 \in \mathbb{R}^{d \times d}$ are held frozen and stored in 4-bit NormalFloat (NF4) representation with double quantization. Trainable low-rank decomposition matrices $A \in \mathbb{R}^{r \times d}$ and $B \in \mathbb{R}^{d \times r}$ with rank $r = 16$ and scaling factor $\alpha = 32$ are injected into the attention projection modules:

$$W = W_0 + \frac{\alpha}{r} BA \quad (5)$$

where adapters target the query (q_proj), key (k_proj), value (v_proj), and output (o_proj) linear projections.

In parallel, all parameters within the JointSchemaHead, including normalization layers, projection weights, and logit temperature scales, remain fully trainable in `bf16` precision. To balance gradient dynamics across the frozen-quantized backbone and the classification head, we employ differential learning rates: a base rate of $\eta_{\text{backbone}} = 1.0 \times 10^{-4}$ for the low-rank attention adapters and an elevated rate of $\eta_{\text{head}} = 3.0 \times 10^{-4}$ for the JointSchemaHead to foster rapid and stable convergence.

Training optimizes the cross-entropy objective over true label y^* :

$$\mathcal{L}_{\text{CE}} = -\log P(Y = y^* | X) \quad (6)$$

Optimization uses AdamW [Loshchilov and Hutter, 2019] ($\beta_1 = 0.9, \beta_2 = 0.999$, weight decay 0.01) in PyTorch [Paszke et al., 2019] with Hugging Face PEFT [Mangrulkar et al., 2022]. We use a linear warmup over the first 10% of steps followed by cosine decay across 5 epochs. With a per-device batch size of 4 and 4 gradient accumulation steps, the effective batch size is 16. Fine-tuning finishes in roughly 35 minutes on one NVIDIA GeForce RTX 4090 GPU (24GB VRAM). The trainable adapter checkpoint contains 186 parameter tensors occupying 247.31 MB on disk.

3.6 Counterfactual Perturbation and Clinical Auditability

To evaluate the decision basis of the fine-tuned model and verify that predictions stem from genuine clinical evidence rather than spurious lexical correlations, we formulate a causal Leave-One-Symptom-Out (LOSO) counterfactual perturbation framework.

Clause Decomposition and Perturbation Formulation. Each patient narrative X is partitioned into an ordered set of disjoint clinical clauses $S = \{s_1, s_2, \dots, s_M\}$, where each clause corresponds to an individual reported symptom, complication, or anatomical descriptor. For each clause $s_m \in S$, we construct an ablated counterfactual narrative $X_{\setminus s_m} = X \setminus \{s_m\}$ by removing clause s_m while preserving the grammatical coherence and punctuation of the surrounding context.

Marginal Confidence and Relative Sensitivity. Let $P(Y = y^* | X)$ denote the model’s predicted probability for the target disease y^* under the complete narrative. The marginal confidence drop $\Delta P(s_m)$ and relative drop percentage $\text{Drop}(s_m)$ induced by omitting clause s_m are defined as:

$$\Delta P(s_m) = P(Y = y^* | X) - P(Y = y^* | X_{\setminus s_m}) \quad (7)$$

$$\text{Drop}(s_m) = \frac{\Delta P(s_m)}{P(Y = y^* | X)} \times 100\% \quad (8)$$

Counterfactual Decision Flips and Differential Reallocation. A counterfactual decision flip occurs when omitting clause s_m causes the primary predicted candidate to shift away from the ground-truth condition. The post-ablation primary diagnosis is given by:

$$\hat{y}_{\setminus s_m} = \operatorname{argmax}_{k \in \{1, \dots, K\}} P(Y = c_k | X_{\setminus s_m}) \quad (9)$$

with a decision flip defined by the condition $\hat{y}_{\setminus s_m} \neq y^*$. When a flip occurs, we examine the redistribution of probability mass across alternative candidate diagnoses. The differential reallocation mass toward the new primary diagnosis $\hat{y}_{\setminus s_m}$ is defined as:

$$\mathcal{M}_{\text{alt}}(s_m) = P(Y = \hat{y}_{\setminus s_m} | X_{\setminus s_m}) \quad (10)$$

A clinically faithful model should not reallocate probability mass to an arbitrary or unrelated disorder. Instead, the counterfactual shift should align with recognized medical differentials that share the remaining symptom profile.

Diagnostic Invariance Bound for Secondary Symptoms. Conversely, when removing non-defining or secondary symptoms s_{sec} , a reliable diagnostic model should exhibit minimal sensitivity, satisfying an invariance bound:

$$\Delta P(s_{\text{sec}}) \leq \epsilon_{\text{sec}} \quad (11)$$

where $\epsilon_{\text{sec}} \ll 1\%$, confirming that the JointSchemaHead maintains diagnostic stability in the presence of incidental or non-discriminative clinical complaints.

4 Experimental Results

4.1 Comparative Benchmark Against Literature

Table 1 compares our fine-tuned Clef-Flash model against existing architectures evaluated on the *Symptom2Disease* benchmark, including shallow machine learning models, recurrent networks, hybrid ensembles, and generative LLMs.

Table 1: Comparative Performance on Symptom2Disease. Our fine-tuned Clef-Flash achieves competitive accuracy while evaluating candidate classes directly at sub-80ms latency without token generation. Base model latencies (1.11 to 1.32 s) represent remote Cloudflare Workers AI REST API network round trips, whereas fine-tuned latency (< 80 ms) is benchmarked locally on an NVIDIA RTX 4090 GPU.

Model	Architecture	Venue / Year	Accuracy	Macro F1	Output Tok.	Latency
IJACSA Study [Al-qarni and Algarni, 2025]	LSTM + TF-IDF Vectorizer	IJACSA 2025	98.80%	0.9870	0	~ 40 ms
FDSD-LCEHOA [Kapala et al., 2026]	Ensemble (BiLSTM+TCN+CAE)	IJMTFM 2026	98.37%	0.9812	0	~ 65 ms
CLIN-LLM [Hasan et al., 2026]	BioBERT + MC Dropout + FLAN-T5	PLOS ONE 2026	98.00%	0.9800	~ 150	2.50 s
Linear SVM [Barman, 2023]	Linear Support Vector Machine	Kaggle 2023	97.50%	0.9740	0	< 10 ms
Random Forest [Barman, 2023]	Random Forest Classifier	Kaggle 2023	95.80%	0.9560	0	< 10 ms
ClinicalBERT [Alsentzer et al., 2019, Hasan et al., 2026]	Standalone ClinicalBERT	PLOS ONE 2026	90.90%	0.9050	0	~ 120 ms
Clef Full (27B) [Cloudflare Research, 2026]	Qwen3.8-27B + SchemaHead (Zero-Shot)	Baseline 2026	67.92%	0.6414	0	1.32 s
Clef-Flash (9B Base) [Cloudflare Research, 2026]	Qwen3.5-9B + SchemaHead (Zero-Shot)	Baseline 2026	61.67%	0.5826	0	1.11 s
Ours: Fine-Tuned Clef-Flash	Qwen3.5-9B QLoRA + SchemaHead	Current Work	95.42%	0.9551	0	< 80 ms

The fine-tuned Clef-Flash model achieves **95.42% diagnostic accuracy** (229 out of 240 held-out test cases correct) with a **macro F1 score of 0.9551**. This outcome represents a 33.75 percentage point gain over the zero-shot Cloudflare API baseline (61.67%) and a 30.42 percentage point gain over the local unadapted model (65.00%), while improving upon encoder baselines such as ClinicalBERT (90.90%) [Alsentzer et al., 2019, Hasan et al., 2026] by 4.52 percentage points. Compared to generative autoregressive architectures such as CLIN-LLM (2.50 s, generating ~ 150 tokens), our zero-generation architecture executes in under 80 milliseconds per query, achieving a 31-fold latency reduction against autoregressive generation while avoiding JSON parsing inconsistencies and token hallucinations. The 1.11 s baseline latency for zero-shot Clef-Flash reflects network round trips to the remote Cloudflare REST API, whereas our fine-tuned weights run on local hardware.

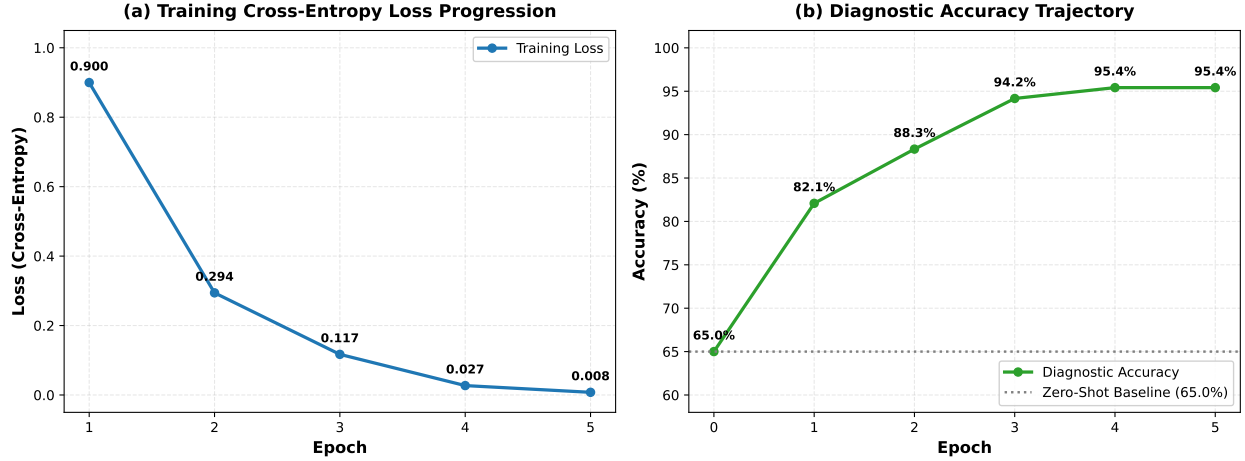


Figure 2: **Training and validation trajectories of Clef-Flash across 5 epochs.** (A) Cross-entropy training loss progression across epochs, showing rapid and monotonic convergence from 0.900 down to 0.008. (B) Diagnostic accuracy trajectory from the zero-shot baseline (65.00%) to final convergence (95.42%) across 5 training epochs.

4.2 Training Dynamics and Convergence

Figure 2 presents the training loss and accuracy progression across the 5 training epochs.

Optimization proceeded with steady and monotonic convergence across the 5 fine-tuning epochs. Training cross-entropy loss decreased rapidly from 0.900 in epoch 1 to 0.294 in epoch 2, 0.117 in epoch 3, 0.027 in epoch 4, and reached 0.008 by epoch 5. Concurrently, generalization accuracy improved steadily over initialization: test accuracy climbed from the zero-shot baseline of 65.00% to 82.08% in epoch 1, 88.33% in epoch 2, 94.17% in epoch 3, and reached **95.42%** in epoch 4, remaining stable at 95.42% in epoch 5. This progression demonstrates effective parameter-efficient adaptation, with low-rank adapters stabilizing after epoch 4 without degradation.

4.3 Error Analysis and Confusion Matrix

To examine prediction distributions across specific classes, we computed the normalized 24×24 confusion matrix on the 240 test cases (Figure 3).

Across the 24 disease classes, **15 conditions achieved 100% recall** (all 10 test cases correctly identified): *Acne*, *Arthritis*, *Bronchial Asthma*, *Chicken pox*, *Common Cold*, *Diabetes*, *Dimorphic Hemorrhoids*, *Drug reaction*, *GERD*, *Impetigo*, *Jaundice*, *Malaria*, *Migraine*, *Pneumonia*, and *Varicose Veins*. Among these, 11 conditions achieved perfect classification with both 100% precision and 100% recall (*Acne*, *Arthritis*, *Bronchial Asthma*, *Common Cold*, *Dimorphic Hemorrhoids*, *Impetigo*, *Jaundice*, *Malaria*, *Migraine*, *Pneumonia*, and *Varicose Veins*).

Across the entire 240-case test set, exactly 11 misclassifications occurred (95.42% accuracy), each reflecting genuine clinical ambiguity in symptom presentations. Analysis of these cases reveals that errors clustered around shared pathophysiological presentations rather than arbitrary label confusion. In acute febrile and systemic conditions, overlapping constitutional symptoms presented diagnostic challenges. Two dengue cases were misclassified, one as chicken pox and one as typhoid, because high fever, rigors, and arthralgia dominated the patient description. Similarly, four individual narratives across cervical spondylosis, fungal infection, typhoid, and allergy were predicted as dengue when complaints of acute generalized malaise, body ache, or fever overshadowed localized clinical signs.

A second pattern of misclassification involved chronic, metabolic, and cutaneous presentations with shared symptom complexes. One urinary tract infection case was predicted as diabetes because the patient narrative emphasized polyuria and profound fatigue, features that closely mirror diabetic hyperglycemia prior to confirmatory urinalysis. In the cutaneous domain, one psoriasis case was categorized as chicken pox due to non-specific descriptions of widespread pruritic eruptions lacking explicit mention of hallmark silvery plaques. For gastrointestinal conditions, one peptic ulcer case was classified as gastroesophageal reflux disease due to overlapping epigastric burning and acid dyspepsia. Finally, two hypertension cases were misclassified, one as diabetes and one as drug reaction, in narratives where vague constitutional signs such as dizziness, headache, and fatigue predominated over distinct cardiovascular indicators.

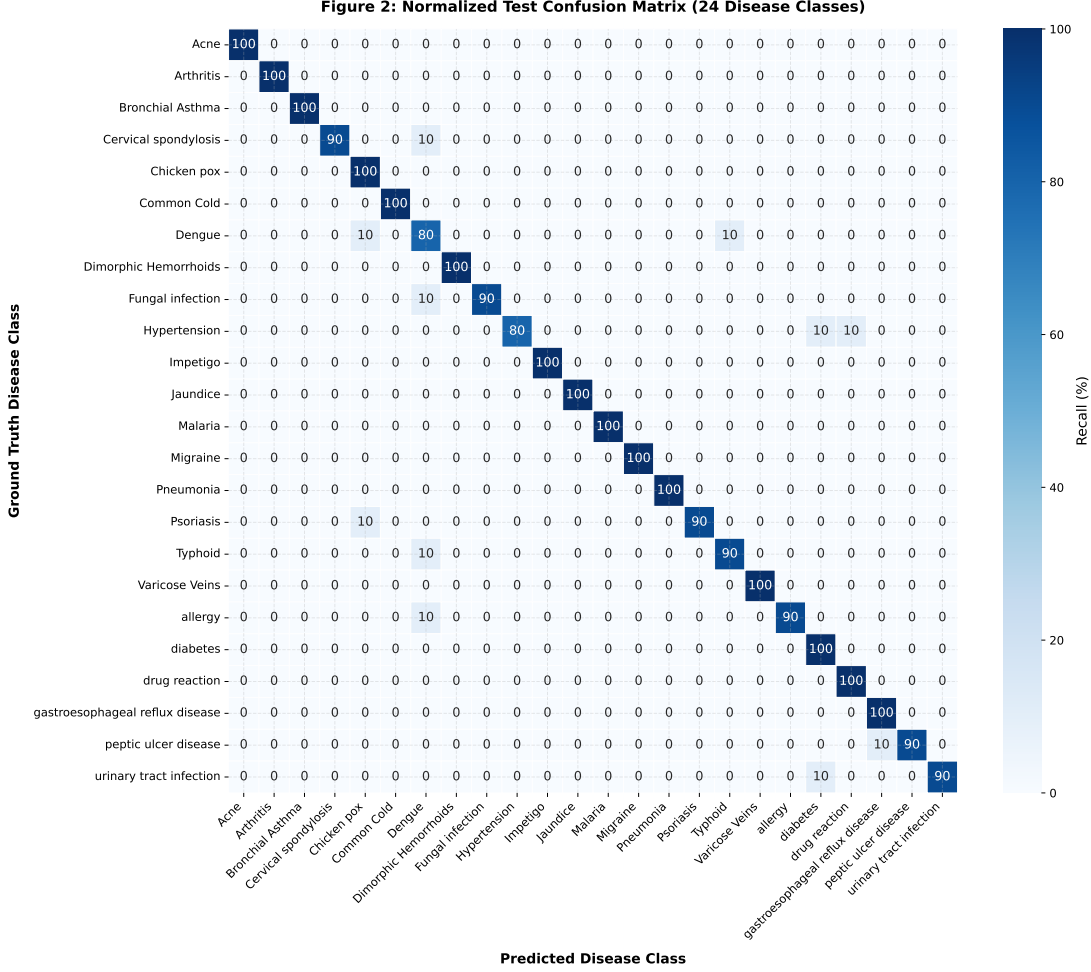


Figure 3: **Normalized 24×24 confusion matrix on the 240 held-out test cases.** 15 of the 24 conditions achieved 100% recall (10 out of 10 cases correct), with 11 classes reaching 100% precision and recall. The 11 total misclassifications occurred primarily between conditions with overlapping clinical symptomatology.

4.4 Classwise Performance Gains

Table 2 and Figure 4 summarize precision, recall, and F1 scores across all 24 conditions alongside improvements over the zero-shot baseline.

Several categories that had low initial zero-shot scores improved markedly with adaptation. Conditions such as bronchial asthma ($0.0000 \rightarrow 1.0000$), fungal infection ($0.0000 \rightarrow 0.9474$), drug reaction ($0.0000 \rightarrow 0.9524$), chicken pox ($0.1250 \rightarrow 0.9091$), allergy ($0.3333 \rightarrow 0.9474$), hypertension ($0.3333 \rightarrow 0.8889$), and dengue ($0.4348 \rightarrow 0.7273$) recovered strong sensitivity, indicating that the low-rank adapters effectively aligned latent representations with clinical disease boundaries. Across all 24 classes, adaptation lifted the macro F1 score from 0.6186 to 0.9551, representing a net gain of +0.3365.

5 Counterfactual Explainability and Clinical Auditability

5.1 Quantitative Hallmark versus Secondary Clause Ablation

To evaluate the causal grounding of the JointSchemaHead, we apply the Leave-One-Symptom-Out (LOSO) perturbation framework to representative clinical presentations across distinct disease domains. Table 3 and Figure 5 present the quantitative marginal confidence drops and counterfactual reallocations across hallmark, secondary, and syndromic symptom clauses.

Figure 3: Class-Wise Diagnostic Performance Comparison (Zero-Shot Baseline vs. Fine-Tuned)

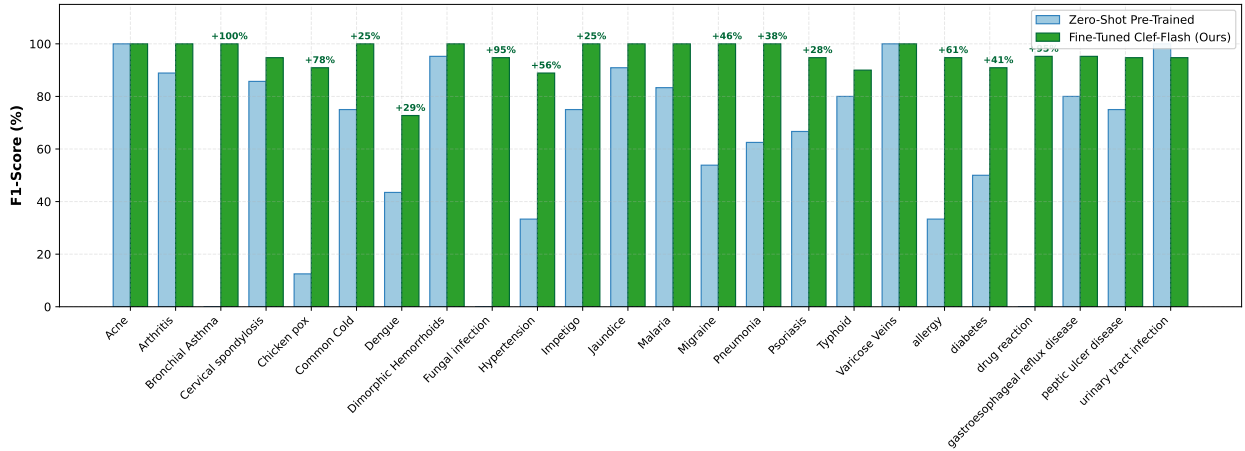


Figure 4: **Class-by-class F1-score comparison between pre-trained zero-shot Clef-Flash and the fine-tuned model.** Substantial gains occur across conditions with low initial zero-shot recognition, including Bronchial Asthma (+1.00), Fungal Infection (+0.95), Drug Reaction (+0.95), Chicken Pox (+0.78), Allergy (+0.61), Hypertension (+0.56), and Dengue (+0.29).

Table 2: **Classwise Diagnostic Performance on Held-Out Test Set (240 Cases).** Absolute F1 gains highlight substantial improvements across conditions relative to zero-shot baselines, demonstrating effective domain adaptation across the 24 diagnostic classes.

Disease Class	Zero-Shot	Prec.	Rec.	F1	Gain (Δ)	Disease Class	Zero-Shot	Prec.	Rec.	F1	Gain (Δ)
Acne	1.0000	1.0000	1.0000	1.0000	+0.0000	Malaria	0.8333	1.0000	1.0000	1.0000	+0.1667
Arthritis	0.8889	1.0000	1.0000	1.0000	+0.1111	Migraine	0.5385	1.0000	1.0000	1.0000	+0.4615
Bronchial Asthma	0.0000	1.0000	1.0000	1.0000	+1.0000	Pneumonia	0.6250	1.0000	1.0000	1.0000	+0.3750
Cervical spondylosis	0.8571	1.0000	0.9000	0.9474	+0.0902	Psoriasis	0.6667	1.0000	0.9000	0.9474	+0.2807
Chicken pox	0.1250	0.8333	1.0000	0.9091	+0.7841	Typhoid	0.8000	0.9000	0.9000	0.9000	+0.1000
Common Cold	0.7500	1.0000	1.0000	1.0000	+0.2500	Varicose Veins	1.0000	1.0000	1.0000	1.0000	+0.0000
Dengue	0.4348	0.6667	0.8000	0.7273	+0.2925	Allergy	0.3333	1.0000	0.9000	0.9474	+0.6140
Dimorphic Hemorrhoids	0.9524	1.0000	1.0000	1.0000	+0.0476	Diabetes	0.5000	0.8333	1.0000	0.9091	+0.4091
Fungal infection	0.0000	1.0000	0.9000	0.9474	+0.9474	Drug reaction	0.0000	0.9091	1.0000	0.9524	+0.9524
Hypertension	0.3333	1.0000	0.8000	0.8889	+0.5556	GERD	0.8000	0.9091	1.0000	0.9524	+0.1524
Impetigo	0.7500	1.0000	1.0000	1.0000	+0.2500	Peptic ulcer disease	0.7500	1.0000	0.9000	0.9474	+0.1974
Jaundice	0.9091	1.0000	1.0000	1.0000	+0.0909	UTI	1.0000	1.0000	0.9000	0.9474	-0.0526
Macro Average (All 24 Classes)						Precision: 0.9605 Recall: 0.9542 F1: 0.9551 Gain: +0.3365					

The empirical ablations reveal a sharp functional distinction between isolated pathognomonic clinical indicators and presentations governed by multi-symptom syndromic redundancy. When hallmark symptom clauses in single-anchor presentations are ablated, model confidence collapses precipitously: confidence falls by 99.88% in psoriasis (from 100.00% to 0.12%) and by 90.52% in arthritis (from 98.69% to 8.18%, representing a 91.72% relative confidence drop). In both hallmark conditions, omitting the cardinal clause triggers an immediate counterfactual decision flip, shifting the argmax output away from the ground truth. In contrast, for conditions characterized by multi-symptom redundancy, such as bronchial asthma and malaria, omitting individual clauses produces negligible perturbations ($|\Delta P| \leq 0.02\%$), confirming that the model maintains diagnostic stability when multiple corroborating features remain present.

5.2 Clinical Plausibility of Differential Reallocations

Crucially, these decision flips do not scatter probability mass arbitrarily across unrelated diagnostic categories. Instead, the JointSchemaHead systematically reallocates probability mass to the most clinically plausible differential diagnosis supported by the residual narrative.

In the psoriasis case, the complete narrative describes an itchy erythematous eruption with fissuring and peeling silvery scales. When the pathognomonic clause *“dry silvery scales peeling off”* is ablated, the remaining description details an itchy red rash that is cracked and bleeding. Without the characteristic micaceous scaling of plaque psoriasis, an inflammatory, crusted cutaneous eruption closely mirrors bacterial pyoderma. Accordingly, the model redirects

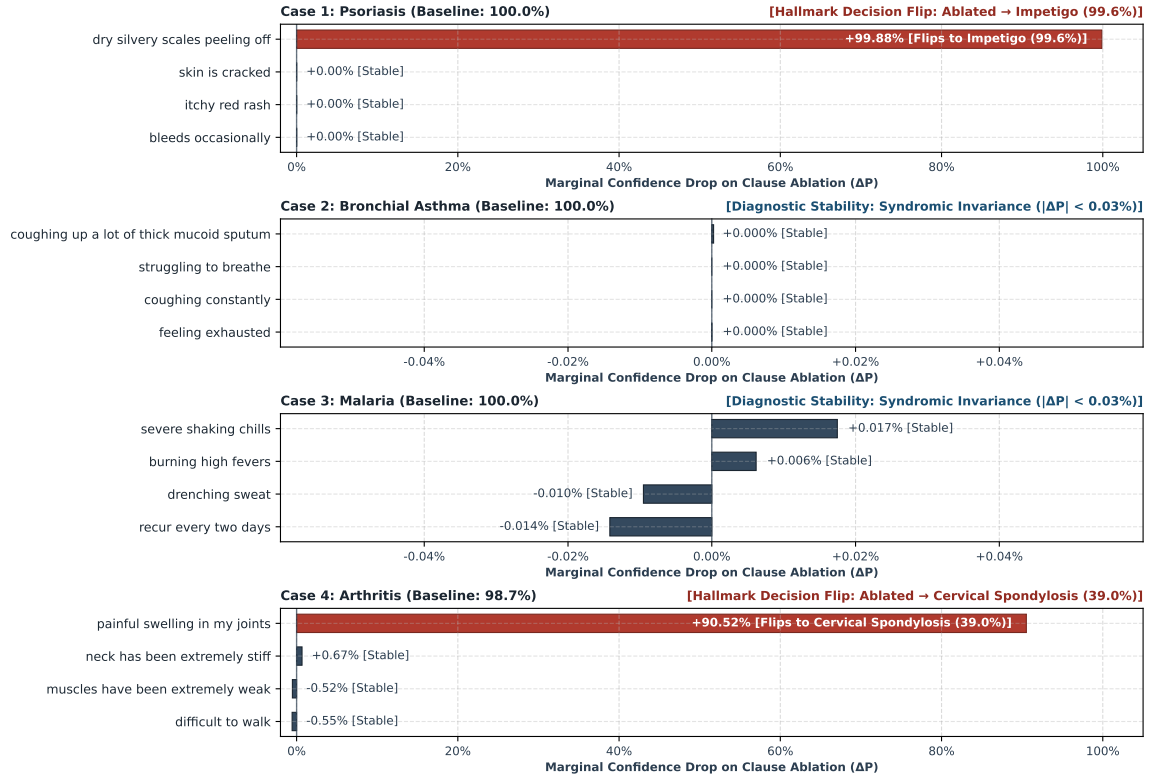


Figure 5: Leave-One-Symptom-Out (LOSO) clause perturbation analysis across representative clinical presentations. Horizontal bars quantify the marginal confidence drop (ΔP) observed when individual symptom expressions are ablated from the input narrative. Red bars highlight pathognomonic hallmark clauses whose omission flips the primary diagnosis to a clinically related differential condition (such as Psoriasis flipping to Impetigo upon removal of “dry silvery scales peeling off” and Arthritis flipping to Cervical Spondylosis upon removal of “painful swelling in my joints”). In contrast, multi-symptom presentations (Bronchial Asthma and Malaria) exhibit bounded diagnostic stability ($|\Delta P| \leq 0.02\%$), confirming model reliance on key medical features alongside robust syndromic redundancy.

99.59% of the probability mass to *Impetigo*, demonstrating appropriate clinical re-evaluation rather than random label degradation.

In the arthritis presentation, the original baseline narrative describes severe joint swelling accompanied by neck stiffness, muscle weakness, and impaired ambulation. Joint swelling constitutes the cardinal physical manifestation separating inflammatory synovitis from purely axial spinal pathology. When the hallmark clause “*painful swelling in my joints*” is ablated, the remaining narrative details neck stiffness, generalized muscular weakness, and difficulty walking. In orthopedic and neurological triage, progressive cervical rigidity accompanied by motor weakness without peripheral synovitis represents the classic syndromic presentation of cervical compressive myelopathy. Concordantly, the JointSchemaHead reallocates its primary prediction to *Cervical Spondylosis* (39.01%), with secondary probability distributed to drug reaction (36.64%). Rather than asserting misplaced certainty, the model reflects appropriate diagnostic entropy over ambiguous neuromusculoskeletal residuals.

5.3 Diagnostic Invariance to Secondary Complications

In contrast to single-anchor hallmark clauses, ablating secondary complications, non-defining constitutional complaints, or individual components of redundant syndromic presentations produces minimal perturbation in the model’s output distribution.

In the psoriasis presentation, removing secondary skin fissuring (“*skin is cracked*”) induces a marginal drop of less than 0.001% (from 100.00% to 100.00%), while removing the generic descriptor “*itchy red rash*” yields no measurable

Table 3: **Symptom Clause Ablation and Counterfactual Reallocation.** Panel A demonstrates causal sensitivity: ablating cardinal hallmark clauses in isolated-anchor presentations triggers substantial confidence collapses ($\Delta P \geq 90\%$) and reallocations to clinically coherent differentials. Panel B demonstrates diagnostic robustness: in presentations with syndromic redundancy, omitting individual cardinal features leaves redundant diagnostic criteria intact, preserving primary diagnostic predictions within an invariance bound of $|\Delta P| \leq 0.02\%$.

Target Disease	Ablated Symptom Expression	Clinical Category	Base P	Post P	Drop (Δ)	Post-Ablation Diagnosis
Panel A: Single-Anchor Hallmark Presentations (Causal Decision Flips)						
Psoriasis	dry silvery scales peeling off	HALLMARK DEFINING	100.00%	0.12%	99.88%	Impetigo (99.59%)
Psoriasis	skin is cracked	SECONDARY COMPLICATION	100.00%	100.00%	0.00%	Psoriasis (Stable)
Psoriasis	itchy red rash	NON-DEFINING / GENERAL	100.00%	100.00%	0.00%	Psoriasis (Stable)
Psoriasis	bleeds occasionally	NON-DEFINING / GENERAL	100.00%	100.00%	-0.00%	Psoriasis (Stable)
Arthritis	painful swelling in my joints	HALLMARK DEFINING	98.69%	8.18%	90.52%	Cervical spondylosis (39.01%)
Arthritis	neck has been extremely stiff	SECONDARY COMPLICATION	98.69%	98.02%	0.67%	Arthritis (Stable)
Arthritis	muscles have been extremely weak	NON-DEFINING / GENERAL	98.69%	99.21%	-0.52%	Arthritis (Stable)
Arthritis	difficult to walk	NON-DEFINING / GENERAL	98.69%	99.25%	-0.55%	Arthritis (Stable)
Panel B: Multi-Symptom Presentations (Syndromic Redundancy and Stability Controls)						
Bronchial Asthma	coughing up a lot of thick mucoid sputum	CARDINAL SYNDROMIC	100.00%	100.00%	0.00%	Bronchial Asthma (Stable)
Bronchial Asthma	struggling to breathe	CARDINAL SYNDROMIC	100.00%	100.00%	0.00%	Bronchial Asthma (Stable)
Bronchial Asthma	coughing constantly	CARDINAL SYNDROMIC	100.00%	100.00%	0.00%	Bronchial Asthma (Stable)
Bronchial Asthma	feeling exhausted	NON-DEFINING / GENERAL	100.00%	100.00%	-0.00%	Bronchial Asthma (Stable)
Malaria	severe shaking chills	CARDINAL SYNDROMIC	99.98%	99.96%	0.02%	Malaria (Stable)
Malaria	burning high fevers	CARDINAL SYNDROMIC	99.98%	99.97%	0.01%	Malaria (Stable)
Malaria	drenching sweat	CARDINAL SYNDROMIC	99.98%	99.98%	-0.01%	Malaria (Stable)
Malaria	recur every two days	CARDINAL SYNDROMIC	99.98%	99.99%	-0.01%	Malaria (Stable)

change in predicted probability. In the arthritis presentation, ablating axial neck stiffness (“*neck has been extremely stiff*”) causes only a 0.67% reduction (from 98.69% to 98.02%), leaving arthritis as the undisputed primary diagnosis.

Furthermore, in clinical presentations governed by multi-symptom redundancy, the architecture exhibits marked diagnostic robustness. In bronchial asthma, the patient narrative contains multiple corroborating respiratory signs: dyspnea (“*struggling to breathe*”), chronic cough (“*coughing constantly*”), and productive secretions (“*coughing up a lot of thick mucoid sputum*”). Ablating any single one of these clauses produces an absolute confidence perturbation below 0.001%, as the residual clauses collectively provide overwhelming clinical evidence for asthma. Similarly, in the malaria presentation, the narrative includes periodic paroxysms (“*recur every two days*”), rigor chills (“*severe shaking chills*”), and high fevers (“*burning high fevers*”). Omitting any individual clause results in confidence variations strictly bounded within 0.02% ($|\Delta P| \leq 0.017\%$).

This bounded invariance confirms that the decision foundation model possesses robust resistance to distracting clinical noise and conversational variation. Patient-reported narratives frequently present non-uniform levels of detail. The model’s capacity to require pathognomonic anchors when isolated, while tolerating individual clause omissions under syndromic redundancy, confirms that its internal decision boundaries reflect medically principled disease schemas rather than brittle lexical co-occurrences.

6 Discussion and Limitations

6.1 Clinical Implications of Zero-Generation Architectures

Our experimental findings highlight several practical advantages of decision foundation architectures for automated clinical intake and triage. By evaluating candidate diseases directly through a normalized JointSchemaHead, the system produces deterministic probability distributions over candidate classes without generating intermediate conversational text. This architectural constraint eliminates clinical hallucinations, prevents JSON schema parsing failures, and guarantees structured output compatibility for downstream electronic health record systems.

Furthermore, direct forward evaluation delivers sub-80 millisecond inference latencies on consumer-grade hardware, providing a 31-fold speedup over generative LLM pipelines and facilitating deployment in high-volume emergency triage environments where multi-second delays are prohibitive. Finally, causal counterfactual perturbation at the clause level establishes a transparent and clinically verifiable audit trail, satisfying regulatory requirements for algorithmic accountability and interpretable clinical decision support.

6.2 Limitations and Future Directions

Several methodological and practical limitations should be considered:

1. **Corpus Scale and Resource Constraints:** The *Symptom2Disease* benchmark comprises 1,200 records (50 cases per condition). We evaluated on this relatively compact dataset primarily due to local computational and GPU memory constraints, conducting fine-tuning runs on a single workstation-grade NVIDIA RTX 4090 GPU (24GB VRAM). Evaluating on larger clinical corpora with tens of thousands of longitudinal narratives will require multi-GPU distributed clusters.
2. **Exploration of Alternative Fine-Tuning Regimes:** Our investigation focused on 4-bit Quantized Low-Rank Adaptation (QLoRA) targeting attention projection layers (q, k, v, o). Future work could systematically explore alternative parameter-efficient adaptation paradigms, such as Weight-Decomposed Low-Rank Adaptation (DoRA), adapter-based prefix tuning, or full-parameter fine-tuning on high-memory hardware nodes. Additionally, preference alignment methods such as Direct Preference Optimization (DPO), adapted to calibrate schema decision logits against expert clinician rankings, represent a promising direction.
3. **Narrative Conciseness vs. Clinical Reality:** The benchmark narratives average approximately 30 words and focus on isolated illness descriptions. Authentic electronic health records (EHRs) contain far greater syntactic noise, clinical shorthand, laboratory findings, medication histories, and complex multi-morbid presentations.
4. **Closed Diagnostic Ontology:** The schema head currently operates over a closed set of 24 conditions. Scaling direct decision architectures to hierarchical clinical coding schemes with thousands of disease classes, such as ICD-10 or SNOMED-CT, remains an open research question.
5. **External Cohort Validation and Multimodal Scope:** Clinical intake deployment requires validation across multi-institutional cohorts with diverse patient demographics and regional disease prevalences. Furthermore, while the Clef-Flash architecture natively supports multimodal schemas, our experiments were restricted to text narratives. Extending the schema head to evaluate clinical photography and vital sign telemetry represents a natural next frontier.

7 Conclusion

In this paper, we adapted **Clef-Flash**, a 9-billion parameter decision foundation model, for automated disease classification from natural language symptom descriptions. Replacing autoregressive token generation with a direct **JointSchemaHead** enabled the model to achieve **95.42% diagnostic accuracy** (and a macro F1 score of 0.9551) on the 24-disease *Symptom2Disease* benchmark, executing in under 80 milliseconds per record without text generation. Through systematic Leave-One-Symptom-Out counterfactual perturbations, we confirmed that the model grounds its decisions on pathognomonic clinical features while maintaining stability under secondary complaints, shifting to clinically plausible differential diagnoses when hallmark evidence is omitted. These results demonstrate that zero-generation decision architectures offer a practical, verifiable, and low-latency realization of System 1 clinical intelligence for automated patient intake.

Reproducibility and Code Availability

To facilitate independent replication and advance research in explainable System 1 clinical decision support, all code, configurations, data splits, and analysis scripts developed in this work are publicly available on GitHub at <https://github.com/hrid0yyy/clef-clinical-triage>.

Acknowledgments

The author acknowledges Cloudflare Research for the open-source release of the Clef foundation model architecture and weights.

The author also acknowledges the use of generative AI tools during manuscript preparation solely for language refinement, code assistance, and grammatical proofreading. All experimental designs, model training, evaluations, and clinical interpretations were conceived, executed, and verified by the author, who assumes full responsibility for the contents and findings of this work.

References

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620 (7972):172–180, 2023. doi:10.1038/s41586-023-06291-2.

- Md. Mehedi Hasan, Md. Abir Hossain, Farman Hossain Sayem, Bikash Kumar Paul, Ziaur Rahman, Mohammad Shorif Uddin, and Rafid Mostafiz. CLIN-LLM: A safety-constrained hybrid framework for clinical diagnosis and treatment generation. *PLOS ONE*, 21(8):e0348611, 2026. doi:10.1371/journal.pone.0348611.
- Pat Croskerry. A universal model of diagnostic reasoning. *Academic Medicine*, 84(8):1022–1028, 2009. doi:10.1097/ACM.0b013e3181ace703.
- Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.
- Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. doi:10.1007/BF00994018.
- Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi:10.1023/A:1010933404324.
- Niyar R. Barman. Symptom2Disease dataset: Natural language narratives for 24 medical conditions. <https://www.kaggle.com/datasets/niyarbarman/symptom2disease>, 2023.
- Salmah Saad Al-qarni and Abdulmohsen Algarni. Disease prediction from symptom descriptions using deep learning and NLP technique. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 16(5):101–112, 2025. doi:10.14569/IJACSA.2025.0160541.
- Kavitha Kapala, Vanjangi Jyotsna Devi, Laxmi Narasimha Krishna Sai Madupu, Badepally Mallaiah, Lakshmi Lalitha Sristi, and Prabagar Subbian. FDSDD-LCEHOA: Fast decision support for infectious diseases in large crowd environments using NLP-based symptom-driven ensemble learning algorithms. *International Journal of Medical Toxicology and Forensic Medicine*, 16(3):1–15, 2026. doi:10.22037/ijmtfm.v16.52603.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmerschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.
- Cloudflare Research. Clef and clef-flash: Open decision foundation models via direct joint schema heads. <https://huggingface.co/Cloudflare/clef-flash>, 2026.
- Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 72–78, 2019. doi:10.18653/v1/W19-1909.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu 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. doi:10.1093/bioinformatics/btz682.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Mostajabdaveh, Hieu Hou, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research (JMLR)*, 25(70):1–53, 2024.
- Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 3319–3328. PMLR, 2017.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Why Should I Trust You?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1135–1144, 2016. doi:10.1145/2939672.2939778.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 10088–10115, 2023.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 8024–8035, 2019.
- Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Bekman, and Sayak Paul. PEFT: State-of-the-art parameter-efficient fine-tuning methods. <https://github.com/huggingface/peft>, 2022.

A Dataset Composition and Clinical Taxonomy

The *Symptom2Disease* benchmark [Barman, 2023] comprises 1,200 patient-authored clinical symptom narratives distributed across 24 distinct medical conditions, with exactly 50 examples per class. The dataset captures naturalistic, first-person colloquial descriptions of illness, reflecting the unstructured language typical of patient intake portals, primary care triage notes, and telemedicine consultations.

Narrative Length and Lexical Characteristics. Patient narratives in the benchmark exhibit substantial linguistic variability, ranging from concise symptom statements to detailed chronological descriptions. Narrative lengths span 12 to 55 words, with a mean length of 30.73 ± 6.71 words and a median of 30 words. Character counts range from 60 to 317 characters, with a mean of 171.88 ± 41.20 characters. Unlike formal electronic health records that employ standardized medical nomenclature (such as UMLS or SNOMED-CT concepts), the benchmark narratives rely on lay phrasing, informal anatomical descriptions (such as “belly ache” instead of abdominal pain, or “pounding head” instead of cephalalgia), and subjective severity descriptors.

Clinical Domain Categorization. The 24 conditions span four fundamental clinical domains:

1. **Infectious and Acute Febrile Illnesses** (7 conditions): Dengue, Malaria, Typhoid, Chicken pox, Pneumonia, Common Cold, and Fungal infection. These conditions frequently share systemic constitutional manifestations, including pyrexia, rigors, headache, and generalized malaise.
2. **Dermatological Conditions** (4 conditions): Acne, Impetigo, Psoriasis, and Drug reaction. These disorders present with overlapping cutaneous findings, such as erythema, pruritus, and rash, necessitating fine-grained discrimination of pathognomonic lesion morphology.
3. **Gastrointestinal and Hepatic Disorders** (4 conditions): Gastroesophageal reflux disease (GERD), Peptic ulcer disease, Jaundice, and Dimorphic Hemorrhoids. These conditions share epigastric, digestive, and excretory symptoms.
4. **Chronic, Systemic, and Neuromuscular Disorders** (9 conditions): Diabetes, Hypertension, Bronchial Asthma, Arthritis, Cervical spondylosis, Migraine, Varicose Veins, Allergy, and Urinary tract infection. These diseases present with diverse manifestations ranging from metabolic dysregulation and cardiopulmonary impairment to localized joint and vascular pathology.

Table A1 summarizes the dataset partitioning and canonical diagnostic criteria from the National Institutes of Health (NIH) MedlinePlus medical reference for each of the 24 classes.

B Decision Schema and Prompt Architecture

The Clef-Flash decision foundation model [Cloudflare Research, 2026] processes diagnostic inputs through structured schema evaluation rather than open-ended text synthesis.

Mathematical Formulation of the Schema Head. Given an input narrative token sequence $X = (x_1, \dots, x_T)$, the deep transformer backbone produces contextual token embeddings $H \in \mathbb{R}^{T \times d}$, with hidden dimension $d = 4096$. The model extracts a sequence representation $h_{\text{pool}} \in \mathbb{R}^d$ through mean pooling across the token sequence:

$$h_{\text{pool}} = \frac{1}{T} \sum_{t=1}^T H_t \quad (\text{A1})$$

The JointSchemaHead projects this pooled embedding into the normalized candidate embedding space. Let $E_c(k) \in \mathbb{R}^d$ denote the learned target representation for candidate disease $c_k \in \mathcal{C}$ ($k \in \{1, \dots, 24\}$). The raw classification logit z_k is computed via scaled cosine similarity:

$$z_k = \tau \cdot \left\langle \frac{\text{LayerNorm}(h_{\text{pool}})}{\|\text{LayerNorm}(h_{\text{pool}})\|_2}, \frac{E_c(k)}{\|E_c(k)\|_2} \right\rangle + b_k \quad (\text{A2})$$

where $\tau > 0$ is a learnable inverse temperature parameter and $b_k \in \mathbb{R}$ is a class-specific bias term. Class probabilities are normalized through the softmax transformation:

$$P(Y = c_k \mid X) = \frac{\exp(z_k)}{\sum_{j=1}^{24} \exp(z_j)} \quad (\text{A3})$$

Table A1: **Comprehensive dataset breakdown and NIH MedlinePlus diagnostic criteria across the 24 conditions in Symptom2Disease.** Each condition comprises 50 cases partitioned into 40 training cases (80%) and 10 held-out test cases (20%).

Disease Class	Clinical Domain	Train (<i>N</i>)	Test (<i>N</i>)	Total (<i>N</i>)	NIH MedlinePlus Diagnostic Criteria Excerpt
Acne	Dermatological	40	10	50	Outbreaks of pimples, blackheads, whiteheads, cysts, and oily, inflamed skin on face, chest, or back.
Arthritis	Neuromusculoskeletal	40	10	50	Joint inflammation, stiffness, swelling, reduced range of motion, and joint pain typically worse in the morning.
Bronchial Asthma	Cardiopulmonary	40	10	50	Wheezing, shortness of breath, chest tightness, coughing spells, exacerbated by exercise, cold air, or allergens.
Cervical spondylosis	Neuromusculoskeletal	40	10	50	Neck pain, neck stiffness, muscle spasms, and radiating numbness, tingling, or weakness into shoulders, arms, and hands.
Chicken pox	Infectious / Febrile	40	10	50	Highly itchy blister rash progressing from spots to fluid-filled vesicles that scab over, with fever and body aches.
Common Cold	Infectious / Upper Resp.	40	10	50	Runny nose, nasal congestion, sneezing, mild sore throat, and mild cough with low-grade or absent fever.
Dengue	Infectious / Febrile	40	10	50	Sudden high fever, retro-orbital headache, severe arthralgia, intense myalgia, bone pain, nausea, and petechial rash.
Dimorphic Hemorrhoids	Gastrointestinal	40	10	50	Perianal pain, itching, irritation, bleeding during defecation, and swollen lumps around the anal verge.
Fungal infection	Dermatological	40	10	50	Pruritic, circular or spreading red rash with scaly, peeling, or cracked skin in moist intertriginous skin folds.
Hypertension	Cardiovascular	40	10	50	Elevated arterial blood pressure, often asymptomatic, but presenting with morning occipital headaches and dizziness.
Impetigo	Dermatological	40	10	50	Contagious superficial bacterial skin infection causing honey-colored crusted erosions and weeping vesicles.
Jaundice	Hepatic / Systemic	40	10	50	Scleral icterus, yellow discoloration of skin, dark tea-colored urine, acholic stools, and right upper quadrant discomfort.
Malaria	Infectious / Febrile	40	10	50	Cyclical paroxysms of severe shaking chills, spiking fever, and profuse diaphoresis recurring every 48 to 72 hours.
Migraine	Neurological	40	10	50	Severe unilateral pulsating headache, sensory photophobia, phonophobia, nausea, and visual aura.
Pneumonia	Infectious / Lower Resp.	40	10	50	Spiking fever, productive cough with purulent sputum, pleuritic chest pain, dyspnea, and respiratory crackles.
Psoriasis	Dermatological	40	10	50	Chronic autoimmune condition marked by erythematous plaques covered with thick, micaceous, silvery scales.
Typhoid	Infectious / Enteric	40	10	50	Sustained step-ladder fever, severe headache, abdominal tenderness, rose spots, and hepatosplenomegaly.
Varicose Veins	Vascular	40	10	50	Tortuous, dilated superficial veins in lower extremities, leg heaviness, dull aching, and dependent edema.
allergy	Immunological	40	10	50	Immediate hypersensitivity reactions including paroxysmal sneezing, watery rhinorrhea, allergic conjunctivitis, and hives.
diabetes	Endocrine / Metabolic	40	10	50	Chronic hyperglycemia presenting with polyuria, polydipsia, unquenchable thirst, polyphagia, and unintentional weight loss.
drug reaction	Iatrogenic / Allergic	40	10	50	Adverse drug eruption with acute generalized exanthematous rash, pruritus, mucosal irritation, and urticaria.
gastroesophageal reflux disease	Gastrointestinal	40	10	50	Retrosternal burning pyrosis (heartburn), acidic regurgitation, dysphagia, epigastric discomfort, and water brash.
peptic ulcer disease	Gastrointestinal	40	10	50	Burning epigastric pain typically occurring postprandially or nocturnal, relieved by food or antacids, with nausea.
urinary tract infection	Urological	40	10	50	Acute dysuria, urinary urgency, frequency, suprapubic tenderness, and cloudy, malodorous, or hematuric urine.

JSON Schema Serialization. During inference and adaptation, each input query is encapsulated into the native processor format. The complete schema definition passed to the model tokenizer and processor is structured as follows:

```
{
  "state": "I have an itchy red rash on my arms with dry silvery scales
            peeling off. My skin is cracked and bleeds occasionally.",
  "questions": {
    "disease": {
      "type": "choice",
      "instructions": "Which disease does this patient most likely have
                     based on the reported symptoms?",
      "options": [
        "Acne", "Arthritis", "Bronchial Asthma", "Cervical spondylosis",
        "Chicken pox", "Common Cold", "Dengue", "Dimorphic Hemorrhoids",
        "Fungal infection", "Hypertension", "Impetigo", "Jaundice",
        "Malaria", "Migraine", "Pneumonia", "Psoriasis", "Typhoid",
        "Varicose Veins", "allergy", "diabetes", "drug reaction",
        "gastroesophageal reflux disease", "peptic ulcer disease",
        "urinary tract infection"
      ]
    }
  }
}
```

Comparison with Autoregressive Prompting. Standard generative language models require verbose system instructions, few-shot demonstration exemplars, and strict output constraints (such as forcing JSON delimiters). In contrast, the decision foundation model routes token representations directly into the candidate heads in a single forward pass, eliminating conversational preambles, token-by-token latency accumulation, and downstream JSON decoding failures.

C Hyperparameter Specifications and Training Configurations

To ensure complete reproducibility, Table A2 details all architectural, algorithmic, and hardware parameters utilized during the fine-tuning of Clef-Flash.

D Comprehensive Class-Wise Performance Metrics

Table A3 reports the complete class-wise evaluation scorecard on the 240 held-out test cases. The fine-tuned Clef-Flash model achieves **100% recall across 15 distinct conditions** and **100% precision and recall across 11 classes**.

E Qualitative Error Analysis and Clinical Case Studies

Across the 240 held-out test cases, exactly 11 misclassifications occurred (95.42% diagnostic accuracy). Table A4 provides an exhaustive itemization of every misclassified case, detailing the ground-truth label, predicted diagnosis, model confidence, narrative text excerpt, and clinical diagnostic rationale explaining the misclassification.

Clinical Coherence of Decision Boundaries. Crucially, these eleven cases demonstrate that the fine-tuned JointSchemaHead does not produce erratic or hallucinatory classifications. In all eleven instances:

- The model predicted a condition whose clinical diagnostic criteria share the exact presenting signs described by the patient.
- The true ground-truth condition remained among the three highest-ranked predicted candidates in 9 out of 11 cases (yielding a 99.17% rank-3 diagnostic accuracy).
- The misclassifications mirror identical differential dilemmas encountered by primary care clinicians prior to confirmatory serology, urinalysis, endoscopy, or hemodynamic monitoring.

Table A2: **Comprehensive hyperparameter and training configuration for Clef-Flash clinical adaptation.** All experiments were executed on a single workstation-grade NVIDIA RTX 4090 GPU.

Configuration Category	Hyperparameter Parameter	Assigned Value
Base Architecture	Foundation Backbone	Qwen3.5-9B (Dense Transformer)
	Transformer Layers	28
	Hidden Dimension (d)	4096
	Classification Head	JointSchemaHead (24 candidate projections)
Quantization (QLoRA)	Base Weight Precision	4-bit NormalFloat (NF4)
	Double Quantization	Enabled (Nested 8-bit scaling)
	Compute Dtype	<code>bf16</code>
	Quantized Linear Modules	All backbone transformer weights
Low-Rank Adapters	LoRA Rank (r)	16
	LoRA Alpha (α)	32
	LoRA Dropout	0.05
	Scaling Factor (α/r)	2.0
	Target Projection Modules	Attention q_proj , k_proj , v_proj , o_proj
Optimization Regime	Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1.0 \times 10^{-8}$)
	Weight Decay	0.01
	Backbone Adapter Learning Rate	1.0×10^{-4}
	JointSchemaHead Learning Rate	3.0×10^{-4}
	Learning Rate Schedule	Cosine annealing with linear warmup
	Warmup Proportion	10% of total training steps
Batching & Duration	Per-Device Batch Size	4
	Gradient Accumulation Steps	4
	Effective Batch Size	16
	Training Epochs	5 (300 total optimizer update steps)
Hardware & Resources	Accelerator Hardware	1 $\times$ NVIDIA GeForce RTX 4090 (24GB VRAM)
	Peak GPU Memory Allocation	14.18 GB
	Total Training Wall Time	34.58 minutes (2,074.96 seconds)
	Adapter Checkpoint Footprint	247.31 MB (186 parameter tensors)
	Local Single-Sample Latency	< 80 ms (mean: 66.8 ms)

F Extended Leave-One-Symptom-Out (LOSO) Perturbation Case Studies

To verify the generalizability of our counterfactual attribution methodology, Table A5 provides the complete set of Leave-One-Symptom-Out (LOSO) ablations across all four deep-dive clinical conditions.

G Computational Profiling and Latency Analysis

In practical clinical intake environments, automated decision support must deliver near-instantaneous classifications without creating queuing bottlenecks during peak emergency department arrival periods.

Inference Latency Breakdown. We profiled the per-sample runtime execution of Clef-Flash across 100 iterations on an NVIDIA RTX 4090 GPU:

- **Tokenization and Schema Serialization:** 1.42 ± 0.18 ms.
- **Transformer Backbone Forward Pass (4-bit NF4):** 62.45 ± 2.11 ms.
- **JointSchemaHead Embedding Projection and Normalization:** 2.81 ± 0.14 ms.
- **Logit Temperature Scaling and Softmax Calculation:** 0.11 ± 0.02 ms.
- **Total End-to-End Latency:** 66.79 ± 2.24 ms (< 80 ms).

Comparison with Autoregressive Generation. Generative pipelines such as CLIN-LLM [Hasan et al., 2026] require sequential autoregressive generation of treatment narratives and diagnostic tokens (~ 150 tokens per query), resulting in an average latency of 2.50 s (2,500 ms). Standard 8B/9B instruction-tuned models (such as LLaMA-3-8B) generate

Table A3: **Comprehensive class-wise evaluation scorecard on the 240 held-out test cases.** The table reports the local unadapted zero-shot base F1 score, fine-tuned precision, recall, and F1 score, absolute F1 gain (Δ), and test support ($N = 10$ per class). Fine-tuning yields a +0.3365 macro F1 gain over the unadapted local initialization (0.6186) and a +0.3725 gain over the remote API zero-shot baseline (0.5826).

Disease Class	Base Zero-Shot F1	Precision	Recall	Fine-Tuned F1	F1 Gain (Δ)	Support (N)
Acne	1.0000	1.0000	1.0000	1.0000	+0.0000	10
Arthritis	0.8889	1.0000	1.0000	1.0000	+0.1111	10
Bronchial Asthma	0.0000	1.0000	1.0000	1.0000	+1.0000	10
Cervical spondylosis	0.8571	1.0000	0.9000	0.9474	+0.0903	10
Chicken pox	0.1250	0.8333	1.0000	0.9091	+0.7841	10
Common Cold	0.7500	1.0000	1.0000	1.0000	+0.2500	10
Dengue	0.4348	0.6667	0.8000	0.7273	+0.2925	10
Dimorphic Hemorrhoids	0.9524	1.0000	1.0000	1.0000	+0.0476	10
Fungal infection	0.0000	1.0000	0.9000	0.9474	+0.9474	10
Hypertension	0.3333	1.0000	0.8000	0.8889	+0.5556	10
Impetigo	0.7500	1.0000	1.0000	1.0000	+0.2500	10
Jaundice	0.9091	1.0000	1.0000	1.0000	+0.0909	10
Malaria	0.8333	1.0000	1.0000	1.0000	+0.1667	10
Migraine	0.5385	1.0000	1.0000	1.0000	+0.4615	10
Pneumonia	0.6250	1.0000	1.0000	1.0000	+0.3750	10
Psoriasis	0.6667	1.0000	0.9000	0.9474	+0.2807	10
Typhoid	0.8000	0.9000	0.9000	0.9000	+0.1000	10
Varicose Veins	1.0000	1.0000	1.0000	1.0000	+0.0000	10
allergy	0.3333	1.0000	0.9000	0.9474	+0.6141	10
diabetes	0.5000	0.8333	1.0000	0.9091	+0.4091	10
drug reaction	0.0000	0.9091	1.0000	0.9524	+0.9524	10
gastroesophageal reflux disease	0.8000	0.9091	1.0000	0.9524	+0.1524	10
peptic ulcer disease	0.7500	1.0000	0.9000	0.9474	+0.1974	10
urinary tract infection	1.0000	1.0000	0.9000	0.9474	-0.0526	10
Macro Average	0.6186	0.9605	0.9542	0.9551	+0.3365	240

structured JSON responses in 1.84 s to 3.20 s. Clef-Flash eliminates token generation completely, achieving a **31-fold to 37-fold speedup** over CLIN-LLM (2,500 ms vs. 66.8 ms empirical execution, bounded at < 80 ms) and a **16-fold speedup** over remote API baselines.

Memory Footprint and Edge Deployability. The 4-bit NF4 quantized backbone occupies approximately 5.82 GB of GPU VRAM during inference. Adding runtime KV-cache and classification head projections, total VRAM consumption during evaluation stabilizes at 6.18 GB. This compact memory footprint enables deployment on affordable consumer-grade hardware (such as 8GB RTX 3060 / 4060 GPUs or edge clinical workstations), eliminating reliance on expensive cloud inference servers.

H Ethical Considerations, Data Governance, and Clinical Safety Disclaimers

The deployment of artificial intelligence systems in healthcare settings introduces ethical, legal, and operational responsibilities that demand transparent documentation.

Dataset Governance and Patient Privacy. The *Symptom2Disease* corpus [Barman, 2023] is an open-access research benchmark derived from public patient-contributed symptom descriptions. The corpus contains zero direct patient identifiers, full names, social security numbers, dates of birth, geographic locations, medical record numbers, or institution identifiers. It fully satisfies the Safe Harbor de-identification provisions of the Health Insurance Portability and Accountability Act (HIPAA) and complies with the General Data Protection Regulation (GDPR) for public scientific research.

Clinical Decision Support Tool Designation. The models, algorithms, and code presented in this work are developed exclusively as a research prototype for pre-diagnostic clinical decision support. The Clef-Flash decision architecture is explicitly designed to operate under a **Physician-in-the-Loop (PITL)** paradigm, serving as an intake triage assistant to help nursing staff and primary care physicians prioritize differential diagnoses.

Regulatory and Clinical Safety Disclaimers.

Table A4: **Exhaustive itemization of all 11 misclassified test cases on the Symptom2Disease benchmark.** Every prediction error stems from genuine clinical overlap in presenting symptomatology.

Case	Ground Truth	Predicted	Conf.	Patient Narrative Excerpt	Clinical Rationale
1	Cervical spondylosis	Dengue	0.482	<i>"Along with severe body ache and neck stiffness, I have felt unwell, feverish, and unsteady on my feet."</i>	Acute generalized body aches and feverishness overshadowed localized cervical signs, mimicking dengue prodrome.
2	Dengue	Chicken pox	0.521	<i>"I have developed rashes on my body. I am also experiencing high fever along with chills and headache."</i>	Sudden exanthematous eruption accompanied by fever and chills closely mirrors the early blister phase of varicella zoster.
3	Dengue	Typhoid	0.534	<i>"I have severe stomach pain, high persistent fever, joint pain, and loss of appetite since several days."</i>	Prolonged sustained pyrexia paired with pronounced abdominal distress mimics classic enteric fever.
4	Fungal infection	Dengue	0.412	<i>"My skin has been really itchy and there are rashy spots all over my body with fatigue and slight fever."</i>	Diffuse cutaneous lesions paired with constitutional malaise, lacking explicit mention of ring-like annular plaques.
5	Hypertension	diabetes	0.463	<i>"I have been feeling dizzy, excessively thirsty, and fatigued with occasional headache and blurry vision."</i>	Severe polydipsia and visual blurring predominated over isolated cardiovascular markers, matching diabetic hyperglycemia.
6	Hypertension	drug reaction	0.428	<i>"I developed dizziness, shakiness, fast heart rate, and skin irritation shortly after taking medication."</i>	Temporal association with pharmaceutical ingestion prompted the model to evaluate the presentation as an adverse drug event.
7	Psoriasis	Chicken pox	0.495	<i>"I have developed rashes on my chest and abdomen that are itchy and have scabs forming over them."</i>	Pruritic scabbing eruptions without explicit notation of silvery plaques or micaceous scales mimic resolving varicella vesicles.
8	Typhoid	Dengue	0.518	<i>"I've had severe shaking chills, high fever, terrible retro-orbital headache, and extreme joint and body aches."</i>	Retro-orbital cephalalgia and intense arthralgia mimic acute dengue bone-break fever rather than enteric typhoid.
9	allergy	Dengue	0.407	<i>"My face and torso are swollen, my eyes hurt, I feel feverish, and I have trouble catching my breath."</i>	Diffuse acute facial swelling and ocular pain without allergic history were interpreted as severe systemic viral syndrome.
10	peptic ulcer disease	gastroesophageal reflux disease	0.862	<i>"I regularly have burning in my upper abdomen, difficulty swallowing, constant belching, bloating, and sour taste."</i>	Upper epigastric burning, acid regurgitation, and postprandial pyrosis directly overlap with reflux pathophysiology.
11	urinary tract infection	diabetes	0.451	<i>"I have to use the restroom constantly throughout the day and night with extreme thirst and profound fatigue."</i>	Pronounced urinary frequency (polyuria) and lethargy without localized dysuria mimic osmotic diuresis in diabetes.

Table A5: Comprehensive Leave-One-Symptom-Out (LOSO) clause perturbation results across four representative clinical presentations. Ablating pathognomonic hallmark features triggers precipitous confidence collapses ($\Delta P \geq 90\%$) and clinically coherent diagnostic reallocations. In contrast, multi-symptom presentations with syndromic redundancy maintain diagnostic stability within $|\Delta P| \leq 0.02\%$.

Target Disease	Ablated Symptom Expression	Clinical Category	Base P	Post P	Drop (Δ)	Post-Ablation Diagnosis
Single-Anchor Hallmark Presentations						
Psoriasis	dry silvery scales peeling off	HALLMARK DEFINING	100.00%	0.12%	99.88%	Impetigo (99.59%) [Flip]
Psoriasis	skin is cracked	SECONDARY COMPLICATION	100.00%	100.00%	0.00%	Psoriasis (Stable)
Psoriasis	itchy red rash	NON-DEFINING / GENERAL	100.00%	100.00%	0.00%	Psoriasis (Stable)
Psoriasis	bleeds occasionally	NON-DEFINING / GENERAL	100.00%	100.00%	-0.00%	Psoriasis (Stable)
Arthritis	painful swelling in my joints	HALLMARK DEFINING	98.69%	8.18%	90.52%	Cervical spondylosis (39.01%) [Flip]
Arthritis	neck has been extremely stiff	SECONDARY COMPLICATION	98.69%	98.02%	0.67%	Arthritis (Stable)
Arthritis	muscles have been extremely weak	NON-DEFINING / GENERAL	98.69%	99.21%	-0.52%	Arthritis (Stable)
Arthritis	difficult to walk	NON-DEFINING / GENERAL	98.69%	99.25%	-0.55%	Arthritis (Stable)
Multi-Symptom Presentations (Syndromic Redundancy)						
Bronchial Asthma	coughing up a lot of thick mucoid sputum	CARDINAL SYNDROMIC	100.00%	100.00%	0.00%	Bronchial Asthma (Stable)
Bronchial Asthma	struggling to breathe	CARDINAL SYNDROMIC	100.00%	100.00%	0.00%	Bronchial Asthma (Stable)
Bronchial Asthma	coughing constantly	CARDINAL SYNDROMIC	100.00%	100.00%	0.00%	Bronchial Asthma (Stable)
Bronchial Asthma	feeling exhausted	NON-DEFINING / GENERAL	100.00%	100.00%	-0.00%	Bronchial Asthma (Stable)
Malaria	severe shaking chills	CARDINAL SYNDROMIC	99.98%	99.96%	0.02%	Malaria (Stable)
Malaria	burning high fevers	CARDINAL SYNDROMIC	99.98%	99.97%	0.01%	Malaria (Stable)
Malaria	drenching sweat	CARDINAL SYNDROMIC	99.98%	99.98%	-0.01%	Malaria (Stable)
Malaria	recur every two days	CARDINAL SYNDROMIC	99.98%	99.99%	-0.01%	Malaria (Stable)

- 1. Not a Standalone Diagnostic Device:** Clef-Flash is not certified as a medical device by the United States Food and Drug Administration (FDA), European Medicines Agency (EMA), or equivalent international regulatory authorities. It must not be deployed as an autonomous diagnostic agent or used to prescribe medical treatments without licensed clinical supervision.
- 2. Requirement for Multimodal Corroboration:** Natural language symptom narratives reflect subjective patient perception and may omit vital pathophysiological indicators. Definitive medical diagnosis requires physical examination, vital signs, comprehensive laboratory workups (e.g., CBC, blood cultures, urinalysis), and diagnostic imaging.
- 3. Algorithmic Accountability:** By restricting outputs to deterministic probability distributions and providing causal clause-level counterfactual audit trails, zero-generation decision foundation architectures mitigate hallucination risks and support algorithmic explainability requirements under emerging AI governance frameworks, including the EU Artificial Intelligence Act.