PandemIQ Llama: A Domain-Adapted Foundation Model for Enhanced Pandemic Intelligence

Jingmei Yang[1], Mahtab Talaei[1], Britta Lassmann[2], Nahid Bhadelia[2], Ioannis Ch. Paschalidis[1,3,4]

[1]Department of Electrical & Computer Engineering and Division of Systems Engineering, Boston University, Boston, MA
[2]Center for Emerging Infectious Diseases Policy and Research, Boston University, Boston, MA
[3]Department of Biomedical Engineering, Boston University, Boston, MA
[4]Faculty of Computing & Data Sciences, Boston University, Boston, MA
{jmyang, mtalaei, blassman, nbhadeli, yannisp}@bu.edu

Abstract

We introduce *PandemIQ Llama*, a domain-adapted large language model (LLM) designed specifically for pandemic intelligence applications. Building on the pre-trained Llama-3.1-8B model, we conducted continuous training using our curated *Pandemic Corpus*. This dataset was assembled from authoritative public health sources, scientific literature, and specialized knowledge repositories, comprising 508,924 documents totaling 5.8 billion tokens, which is the largest pandemic-domain-specific data cohort for LLM training. Benefiting from our curated large data cohort and through continuous training leveraging extensive computational resources, the developed PandemIQ Llama model can extract critical domain knowledge on pandemics, which is typically underrepresented in general-purpose language models. To evaluate its performance, we conducted a comprehensive comparison of PandemIQ Llama with both prompt-engineered and task-specific fine-tuned models using two tasks: the Biomedical Alert News Question Answering task (1,508 reports with 30 expert-generated questions each) and the Disease Event Type Classification benchmark (4,500 news snippets across eight disease categories). PandemIQ Llama demonstrated substantial improvements over strong baseline models, achieving performance gains ranging from 3.8% to 10.97%. These results suggest that PandemIQ Llama could significantly enhance public health surveillance and analysis capabilities. In addition, our results also suggest that LLMs can perform better with continuous training than fine-tuning on domain-specific tasks. *Social Impact*: Our platform, powered by our model, has been launched and now serves over 100 government and multilateral public health organizations and users across 154 countries. Reports published on our platform are being integrated into the Epidemic Intelligence from Open Sources system run by the World Health Organization. This integration will provide public health decision-makers with a powerful LLM-based tool for pandemic surveillance.

Code — https://github.com/noc-lab/PandemIQ-Llama
Model — Paschalidis-NOC-Lab/PandemIQ-Llama

Introduction

The COVID-19 pandemic revealed critical vulnerabilities in global infectious disease surveillance systems, highlighting the urgent need for enhanced early-warning mechanisms and rapid response capabilities. Recent advances in early warning (MacIntyre et al. 2023) and detection systems (Mollalo et al. 2019; Sundermann et al. 2022), outbreak prediction (Ramchandani, Fan, and Mostafavi 2020; Kim and Ahn 2021), forecasting (Reich et al. 2019; Shastri et al. 2020), and intervention planning (Bastani et al. 2021; Zhang, Guo, and Lv 2024) demonstrate the significant potential of AI-driven surveillance approaches. However, the capacity to identify emerging biological threats remains limited by inadequate monitoring frameworks, diminished prioritization during inter-outbreak periods, and the increasingly complex landscape of novel and resurgent pathogens (Parums 2023; Brownstein et al. 2023). Web-based surveillance platforms have become essential components of global health security infrastructure, yet these systems face substantial challenges in processing the continuous influx of multilingual, unstructured data streams encompassing clinical reports, public health bulletins, news feeds, and social media content in real time. This expanding disparity between information generation rates and analytical processing capacity creates a compelling need for detection systems capable of extracting, interpreting, and contextualizing diverse epidemiological signals with enhanced accuracy and efficiency.

Natural Language Processing (NLP) has emerged as a fundamental component of pandemic surveillance systems, evolving from specialized BERT-based architectures to contemporary large language models. Foundation models such as LLMs demonstrate particular value through their capacity to identify complex patterns within extensive textual datasets. The inherent adaptability of these pre-trained systems across diverse domains and applications renders them exceptionally suitable for health monitoring implementations (Singhal et al. 2023; Xie et al. 2024; Chen et al. 2024). Domain-specific adaptations have yielded substantial advances in biomedical text analysis. BioBERT (Lee et al. 2020) and ClinicalBERT (Alsentzer et al. 2019) enhanced medical text mining through targeted pretraining on biomedical corpora, while COVID-Twitter-BERT (Müller, Salathé, and Kummervold 2023) established the efficacy of social media-adapted models for pandemic intelligence. These BERT-based architectures have markedly improved syndromic signal extraction from social media and traditional sources, resulting in enhanced forecasting accuracy and re-

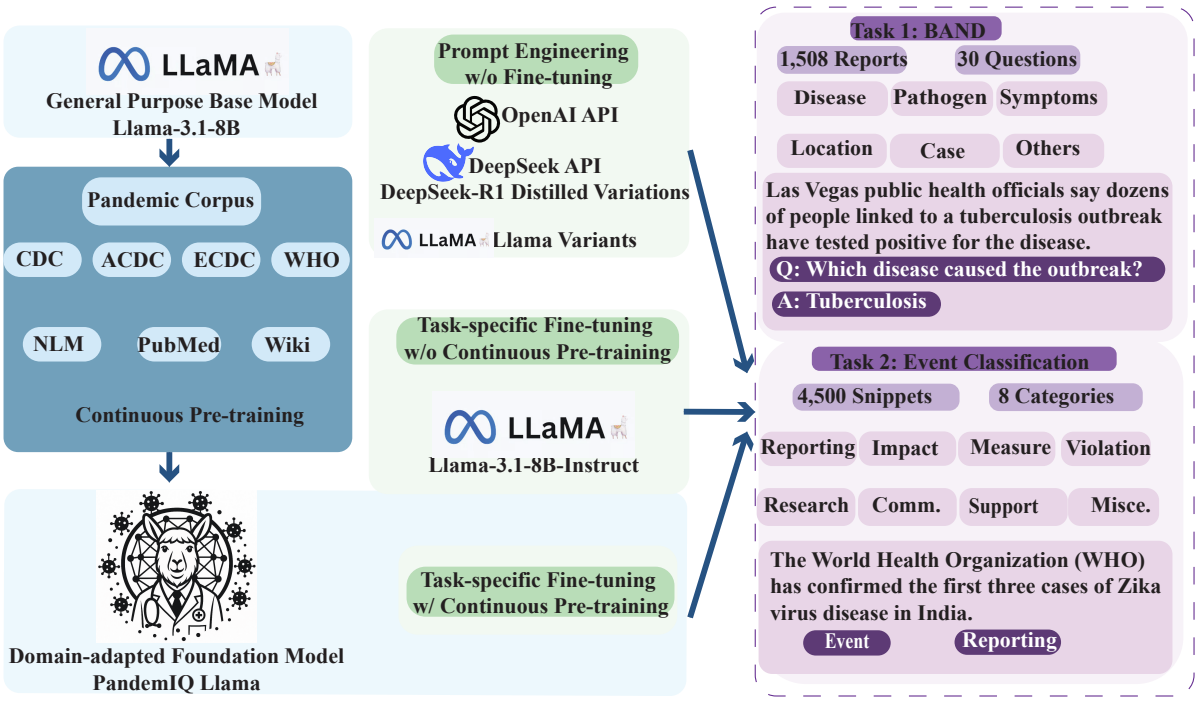

Figure 1: PandemIQ Llama framework. We compared our PandemIQ Llama with three approaches: (1) prompt engineering; (2) task-specific fine-tuning; and (3) fine-tuning of our domain-adapted PandemIQ Llama. *Task* 1*: Biomedical Alert News and Disease (BAND)* (1, 508 alert reports, 30 expert-curated questions covering disease, pathogen, host, location, case counts, etc.) and *Task* 2*: Disease Event Classification* (4, 500 news snippets, eight categories: reporting, impact, measure, violation, research, communication, support, miscellaneous).

duced response latencies (Brownstein et al. 2023; MacIntyre et al. 2023). Contemporary surveillance platforms including BlueDot (BlueDot Inc. 2025) and EPIWATCH employ BERT-based models as foundational prediction engines, demonstrating the practical utility of domain-adapted language models in operational biosurveillance systems (EPIWATCH 2025). Nevertheless, these architectures exhibit inherent constraints, including fixed 512-token context limitations and restricted cross-domain reasoning capabilities, thereby necessitating the development of more sophisticated LLMs for pandemic biosurveillance.

General-purpose LLMs often struggle with domain-specific terminology, lack the deep knowledge required for specialized reasoning, are prone to factual errors, and fail to align with professional standards or regulations (Parums 2023; Brownstein et al. 2023). In the domain of pandemic biosurveillance, the general purpose LLMs demonstrated limited understanding of epidemiological concepts, outbreak dynamics, and biosurveillance terminology. In addition, LLMs trained on *general medical literature* lack specialized expertise for comprehensive pandemic monitoring tasks, particularly in recognizing pathogen-emergence patterns, interpreting complex spatiotemporal outbreak data, and integrating diverse epidemiological signals within appropriate public health frameworks.

To address this gap, we developed *PandemIQ Llama*, a domain-adapted foundation model for biosurveillance applications, made possible by training on our large cohort of domain specific data and extensive computational infrastructure. To the best of our knowledge, PandemIQ Llama is the first LLM adapted for pandemic intelligence with validated performance against different benchmark frameworks and pandemic related downstream tasks. The overall design of this framework is illustrated in Fig. 1. We perform continuous pre-training on Llama-3.1-8B with our curated *Pandemic Corpus*, a collection of 5.7 billion tokens from biosurveillance sources, enabling the model to acquire specialized epidemiological knowledge while maintaining general language capabilities. We evaluate PandemIQ Llama against two paradigms of approaches: (1) prompt engineering of general-purpose LLMs without domain adaptation (e.g., (Yang et al. 2024)), and (2) task-specific supervised fine-tuning of general LLMs without domain adaptation in pre-training, and across two challenging tasks: (1) *Biomedical Alert News Question Answering* (1, 508 reports $\times$ 30 expert-curated questions) and (2) *Disease Event Type Classification* (4, 500 news snippets across 8 categories).

Our work offers four key contributions: (1) *Data*: we curated the largest pandemic cohort supervised from *14 domain experts* from Center from Disease Control and Prevention (CDC), WHO, and academic leaders to cover intensive and extensive domain questions; (2) *PandemIQ Llama model*: the first pandemic specific pre-trained LLM is developed to empower a large community of decision mak-

ers and stakeholders with AI tools for informed decision making; (3) *Social impact*: both our data and model will be made available to benefit researchers and policy makers of the pandemic field. In addition, this model will be integrated to EIOS from WHO for broader global impact; (4) *Establishing validation for pandemic biosurveillance*: a rigorous and extensive benchmarking framework for assessing LLM capabilities in biosurveillance, comparing prompt-based and fine-tuning approaches across multiple models to systematically quantify performance is established.

Method

Pandemic Corpus

Effective domain adaptation needs high-quality training data. We collected data from public domain and developed the *Pandemic Corpus*, which is a collection of $508,924$ documents comprising 5.7 billion tokens focused on 16 high-priority pathogens given their pandemic potential and public health significance. Corpus collection was guided by a panel of 14 infectious disease experts. There are seven authoritative sources identified by domain experts:

- Public health authorities, including WHO, CDC, European Centre for Disease Prevention and Control (ECDC) and Africa CDC, provide authoritative surveillance reports, case definitions, response protocols, clinical guidelines, and real-time situation reports during active outbreaks.
- Scientific literature repositories (PubMed Central (PMC) and National Library of Medicine (NLM)) contribute peer-reviewed research with epidemiological models, clinical characterizations, transmission analyses, and technical terminology.
- Wikipedia offers disease outbreak overviews, historical context, and explanations that connect specialized terminology with general understanding.

For each source, we developed tailored scraping and preprocessing pipelines adapted to the documents encountered. This ensured the preservation of critical epidemiological content while removing extraneous elements. The resulting corpus includes documents in 31 languages, with representation from regions where outbreaks frequently occur. This multilingual coverage helps capture early warning signals, as local reporting often appears before international coverage. The corpus also covers a balanced range of priority pathogens, including both common threats and more rare diseases. Documents span historical to recent reports (cutoff: 04/15/2024), providing up-to-date knowledge.

Source	Number of Docs	Number of Tokens (M)
PMC	469,701	5316
NLM	3,865	153
CDC	5,962	135
WHO	4,383	79
ECDC	2,321	43
Wikipedia	22,028	12
Africa CDC	664	1.5

Table 1: Composition of the Pandemic Corpus.

PandemIQ Llama

To develop a model with both general language capabilities and specialized epidemiological knowledge, we adopted a continuous pre-training approach. This method allows for domain adaptation of an existing foundation model while preserving its general language capabilities. Leveraging our Pandemic Corpus, we performed continuous pre-training on the Llama-3.1-8B model (Touvron et al. 2023) to train PandemIQ Llama. Specifically, our implementation leveraged a high-performance computing cluster with five nodes containing 40 NVIDIA A100 (80GB) GPUs. We conducted our pre-training utilizing the HuggingFace Transformers and Accelerate libraries with Fully Sharded Data Parallel (FSDP) in mixed precision training (Zhao et al. 2023). The model underwent 3 epochs of training on the Pandemic Corpus with a global batch size of 240 (6 samples per device), using the AdamW optimizer with a learning rate of 2×10^{-5} and cosine decay schedule with 3% warmup ratio. The entire continuous pre-training process required approximately $1,920$ GPU hours.

Validation Benchmarks

After developing PandemIQ Llama, we evaluate its utility using two complementary tasks: (1) Question answering to test the model's ability to extract epidemiological information (pathogen identification, case counts, affected populations, and spatiotemporal outbreak details) and perform inferential reasoning about outbreak patterns. (2) Multi-label classification to assesses the model's capacity to categorize outbreak reports, with the challenge that reports may belong to multiple categories. This requires nuanced understanding of epidemiological concepts and their contextual relationships.

Biomedical Alert Question Answering. The Biomedical Alert News and Disease (BAND) dataset (Fu et al. 2024) comprises $1,508$ expert-vetted outbreak reports, each paired with 30 carefully designed questions, yielding $45,240$ question-answer pairs. These questions target critical epidemiological information including pathogen, disease, symptoms, host types, geographical spread, case numbers, and affected populations. We categorize questions along two dimensions: (1) Answer format: seven open-ended questions require free-text extraction of entities, assessing the model's ability to identify key information despite linguistic variations. The remaining 23 closed-ended questions restrict responses to predefined options (e.g., "Yes/No/Cannot Infer"), testing precise information extraction and inference. (2) Geographical specificity: three questions require standardized location codes for the outbreak's country, state, and city. These geocode questions test the model's ability to normalize location mentions to standardized codes-a capability requiring both information extraction and external geographical knowledge. The remaining 27 non-geocode questions focus on other epidemiological aspects.

Multi-label Disease Event Classification. The second evaluation task involves multi-label classification of news

Source	Model	Accuracy by Question Type				
		Overall ↑	**Open ↑**	**Closed ↑**	**Geocode ↑**	**Non-geocode ↑**
		Default Template (Prompt Engineering)				
Meta	3.3-70B-Instruct	49.41	65.05	44.65	12.45	53.52
	3.1-70B-Instruct	49.54	67.08	44.21	10.30	53.90
	3.1-8B-Instruct	47.71	63.03	43.05	5.15	52.44
	3.2-3B-Instruct	57.58	64.87	55.36	5.15	63.41
	3.2-1B-Instruct	44.65	56.53	41.03	0.14	49.59
DeepSeek	R1-Distill-70B	48.31	70.26	41.63	8.15	52.77
	R1-Distill-8B	48.91	69.53	42.64	11.59	53.06
API	OpenAI GPT-4o	64.61	70.39	62.85	47.21	66.54
	DeepSeek Reasoner	56.32	67.93	52.79	24.32	59.88
		Customized Template (Prompt Engineering)				
Meta	3.3-70B-Instruct	67.51	73.51	65.68	18.31	72.98
	3.1-70B-Instruct	68.94	71.06	68.30	18.60	74.54
	3.1-8B-Instruct	59.79	63.95	58.52	7.01	65.65
	3.2-3B-Instruct	58.56	66.89	56.02	6.72	64.31
	3.2-1B-Instruct	45.84	56.41	42.62	0.57	50.87
DeepSeek	R1-Distill-70B	66.87	68.24	66.45	16.60	72.45
	R1-Distill-8B	50.36	62.29	46.73	2.15	55.71
API	OpenAI GPT-4o	79.21	84.30	77.66	54.94	81.91
	DeepSeek Reasoner	76.81	84.43	74.49	38.77	81.04
		Fine-Tuning				
Baseline	BLOOM	75.10	–	–	–	–
Meta	3.1-8B-Instruct	77.38	80.66	76.39	49.71	80.46
Ours	PandemIQ Llama	**86.07**	**88.47**	**85.33**	**56.08**	**89.40**

Table 2: Performance comparison on the BAND task across prompt engineering and fine-tuning approaches. Higher values (↑) indicating better performance. Models grouped by: (1) Prompt engineering, zero-shot using either default or custom templates without training; (2) Fine-tuning. Key comparison: w/ Domain Adaptation (PandemIQ Llama: continuous pre-training + fine-tuning) vs. w/o Domain Adaptation (Llama-3.1-8B-Instruct: fine-tuning only). Baseline from Fu et al. (2024).

snippets according to a detailed infectious disease event taxonomy (Piskorski et al. 2023). This dataset consists of $4,500$ news snippets, with each snippet assigned one or more labels from eight epidemiological categories: outbreak reporting, impact assessment, public health measures, violation, research & development, communication, support, and miscellaneous. This classification task directly tests whether PandemIQ Llama has acquired the epidemiological knowledge needed to distinguish between related event types while recognizing their potential co-occurrence.

Prompt Engineering Model Selection. We conducted evaluations across model architectures, parameter scales, and development generations: Llama family (1B/3B/8B/70B) spanning versions 3.1-3.3, open-source DeepSeek models, and commercial models (OpenAI GPT-4o and DeepSeek Reasoner) accessed via API.

Prompting Templates. We compared two distinct prompt templates for each task. *Default prompts* present questions exactly as formulated in the original benchmarks. *Customized prompts* incorporate epidemiological context and structured reasoning guidance. For BAND, we tailored prompts to each question type, supporting both precise entity extraction and contextual reasoning. For event classification, we enriched prompts with detailed category descriptions according to the annotation guidelines.

Fine-tuning. Building upon our prompt engineering, we implemented task-specific fine-tuning to compare two adaptation paradigms: (1) fine-tuning our domain-adapted PandemIQ Llama, which underwent continuous pre-training on the Pandemic Corpus, and (2) conventional fine-tuning of Llama-3.1-8B-Instruct without domain adaptation. We fine-tuned each model separately on both downstream tasks using parameter-efficient techniques. We employed Low-Rank Adaptation (LoRA) (Hu et al. 2022), which injects trainable low-rank decomposition matrices into frozen pre-trained weights. To assess adaptation capacity effects, we varied rank parameters across $r \in \{8, 16, 32, 64, 128, 256\}$, with scaling parameter $\alpha = 2r$. Our implementation extended beyond standard approaches by applying adaptation to both attention components (query, key, value, output projections) and feed-forward networks (down_proj and up_proj), allowing for more parameter updates throughout the network while maintaining computational efficiency.

Training Details. We used hyperparameter configurations optimized for each benchmark. For BAND, we used a learning rate of 5×10^{-5}, 7 training epochs, and gradient accumu-

lation of 4 steps with a per-device batch size of 4. For Event, we employed a lower learning rate of 2×10^{-5} with 3 training epochs and no gradient accumulation. Both configurations utilized cosine learning rate decay with 3% warmup ratio and zero weight decay. All fine-tuning experiments were conducted using PEFT library with the SFT Trainer on a single node with 8 A100 (80GB) GPUs. For all rank variations in our ablation study, we maintained these task-specific configurations constant, modifying only the LoRA rank parameter and its corresponding scaling factor.

Results

We compared our approach with different baseline methods, including prompt engineering evaluations across Llama variants, DeepSeek variants, and commercial APIs, as well as fine-tuning outcomes for PandemIQ Llama and Llama-3.1-8B-Instruct on question answering and event classification tasks.

BAND Question Answering Task

On the BAND QA benchmark, we report overall accuracy across all 30 questions and specific accuracy for question subcategories (open-ended vs. closed-ended; geocoded vs. non-geocoded). As shown in Table 2 and Figure 2, PandemIQ Llama demonstrates consistent superiority across all evaluation dimensions when compared to alternative approaches: prompt engineering with GPT-4o (our best zero-shot result), fine-tuning without domain adaptation (Llama-3.1-8B-Instruct), and the published BLOOM benchmark. PandemIQ Llama achieves 86.07% overall accuracy, outperforming GPT-4o by 6.86%, BLOOM by 10.97%, and fine-tuned Llama-3.1-8B-Instruct by 8.69%.

Analysis of prompt engineering across model variants revealed three key patterns: (1) model size correlates positively with accuracy, (2) our epidemiologically-informed prompts significantly outperformed default templates, and (3) commercial models (particularly GPT-4o) excelled compared to open-source alternatives, especially on geolocation-related questions. Despite template optimizations, all prompt-engineered models still underperformed PandemIQ Llama.

Figure 3 shows the performance of fine-tuning across different rank configurations for both Llama-3.1-8B-Instruct and PandemIQ Llama. Two clear patterns are evident: (1) accuracy improves with increasing rank values for both models, and (2) PandemIQ Llama outperforms Llama-3.1-8B-Instruct at every rank configuration. The performance gap between models is most pronounced for geocoded questions. Based on these results, we selected rank 256 for subsequent cross-model comparisons.

Figure 4 presents a fine-grained analysis across nine epidemiological dimensions. PandemIQ Llama demonstrates the most substantial gains in traditionally challenging categories: location identification (10.51%), case number extraction (7.94%), and symptom recognition (9.03%). Vulnerable population identification tasks (elderly, animal workers, and healthcare workers) show consistent improvements of 5.58%-9.23%. In questions about disease and host type identification that were already approaching ceiling performance

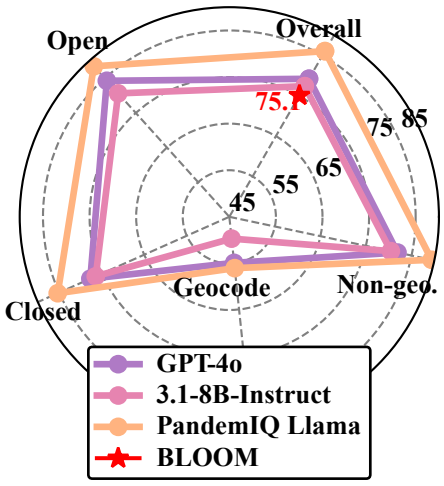

Figure 2: Performance comparison of training paradigms on the event classification task.

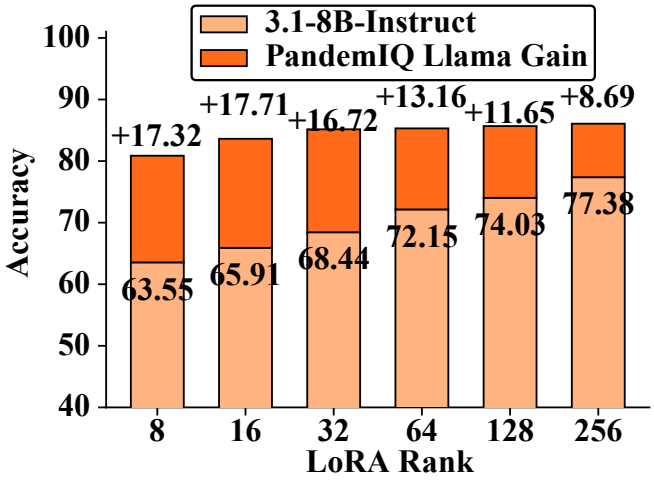

Figure 3: Performance on BAND across LoRA ranks.

with Llama-3.1-8B-Instruct, our model maintains high accuracy.

Multi-label Classification Task

On the event classification benchmark, we evaluate performance using Macro-F_1, Micro-F_1, and Weighted-F_1 scores across all eight event categories. As shown in Table 3 and Figure 5, PandemIQ Llama demonstrates consistent superiority across all evaluation metrics when compared to alternative approaches: prompt engineering with DeepSeek Reasoner (our strongest zero-shot baseline), fine-tuning without domain adaptation in pre-training (Llama-3.1-8B-Instruct), and the published RoBERTa benchmark. PandemIQ Llama achieves state-of-the-art performance with a Macro-F_1 score of 79.80%, outperforming DeepSeek Reasoner by 8.81%, RoBERTa by 3.80%, and fine-tuned Llama-3.1-8B-Instruct

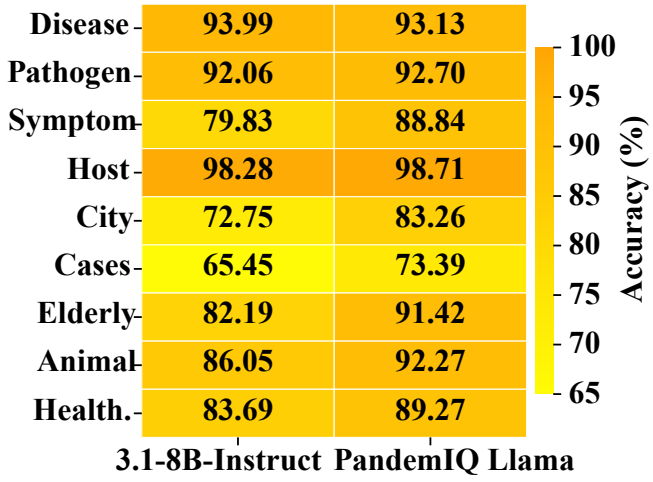

Figure 4: Accuracy on BAND across nine dimensions: disease, pathogen, symptoms, host, city, cases (numbers), elderly, animal (workers), and health (healthcare workers).

by 5.29%.

Prompt engineering experiments across model variants showed patterns that were similar to our BAND findings: (1) model scale correlates strongly with event categorization ability, (2) customized prompts enriched with annotation guidelines substantially improved performance across all models, and (3) the reasoning-optimized model (DeepSeek Reasoner) demonstrated superior discriminative capabilities for nuanced event types. Overall, all prompt-engineered approaches remained less effective than PandemIQ Llama.

Fig. 6 illustrates classification performance across LoRA rank configurations for both Llama-3.1-8B-Instruct and PandemIQ Llama. Two clear trends are apparent: (1) F_1 scores improve progressively with higher rank values for both models, and (2) PandemIQ Llama consistently outperforms Llama-3.1-8B-Instruct across the entire rank spectrum. Based on these findings, we adopted rank 256 for all subsequent evaluations.

Fig. 7 shows performance across all 8 event categories. PandemIQ Llama outperforms the vanilla fine-tuned Llama-3.1-8B-Instruct in 6 of 8 event categories, with the largest improvements in categories requiring specialized epidemiological knowledge. For *Support*, we achieve 8.75% higher F_1, effectively recognizing financial assistance mechanisms and resource allocation patterns during pandemic responses. In *Impact*, PandemIQ Llama shows an 8.52% improvement, accurately identifying diverse societal disruptions caused by disease outbreaks. The model also demonstrates stronger performance in *Measure* (7.35%) and *Miscellaneous* (6.87%), reflecting enhanced recognition of technical language related to preventive policies and interventions. Notably, gains are more modest in *Research* and *Reporting* categories, while performance remains comparable to Llama-3.1-8B-Instruct in the *Violation* category. These category-specific patterns confirm that domain adaptation enhances classification performance for event types that rely on specialized pandemic terminology and contextual understanding.

Source	Model	Macro-F_1 ↑	Micro-F_1 ↑	W-F_1 ↑
	Default Template (Prompt Engineering)			
Meta	3.3-70B-Instruct	48.08	47.49	48.73
	3.1-70B-Instruct	48.08	49.20	48.93
	3.1-8B-Instruct	43.57	39.32	38.99
	3.2-3B-Instruct	41.48	35.90	38.99
	3.2-1B-Instruct	19.55	24.74	18.98
DS	R1-Distill-70B	64.54	64.60	65.48
	R1-Distill-8B	48.91	46.83	50.43
API	OpenAI GPT-4o	53.61	52.55	53.98
	DS Reasoner	57.55	58.17	59.64
	Customized Template (Prompt Engineering)			
Meta	3.3-70B-Instruct	63.45	62.68	64.85
	3.1-70B-Instruct	59.37	60.15	59.96
	3.1-8B-Instruct	57.44	52.61	57.30
	3.2-3B-Instruct	49.63	47.41	46.60
	3.2-1B-Instruct	13.16	22.16	13.08
DS	R1-Distill-70B	64.54	64.60	65.48
	R1-Distill-8B	62.01	60.52	62.54
API	OpenAI GPT-4o	66.68	66.78	67.89
	DS Reasoner	70.99	69.78	72.05
	Fine-Tuning			
Baseline	RoBERTa	76.00	76.00	76.00
Meta	3.1-8B-Instruct	74.51	74.98	74.86
Ours	PandemIQ Llama	**79.80**	**80.48**	**80.45**

Table 3: Performance comparison on the event classification task across prompt engineering and fine-tuning approaches. Higher values (↑) indicate better performance. Models grouped by: (1) Prompt engineering, zero-shot using either default or customized templates without training; (2) Fine-tuning. Key comparison: w/ Domain Adaptation (PandemIQ Llama: continuous pre-training + fine-tuning) vs. w/o Domain Adaptation (Llama-3.1-8B-Instruct: fine-tuning only).

Discussion and Conclusion

PandemIQ Llama shows notable improvements across two demanding biosurveillance tasks. It achieves 86.07% accuracy on question answering, outperforming GPT-4o by 6.86% and fine-tuned Llama 3.1-8B-Instruct by 8.69%. On event classification, it reaches a Macro-F_1 score of 79.80%, surpassing DeepSeek Reasoner by 8.81% and a task-tuned Llama baseline by 5.29%. These gains are most pronounced in areas that require deeper epidemiological insight. In question answering, the model shows the strongest improvements in outbreak characterization, specifically in identifying locations, extracting case counts, and recognizing symptoms. For event classification, it performs particularly well in categories like *Support* and *Impact*, which involve understanding resource needs and broader societal effects during health emergencies.

These results suggest that integrating epidemiological

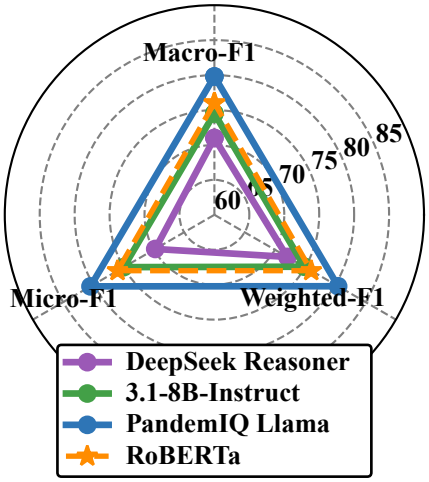

Figure 5: Performance comparison of training paradigms on the event classification task.

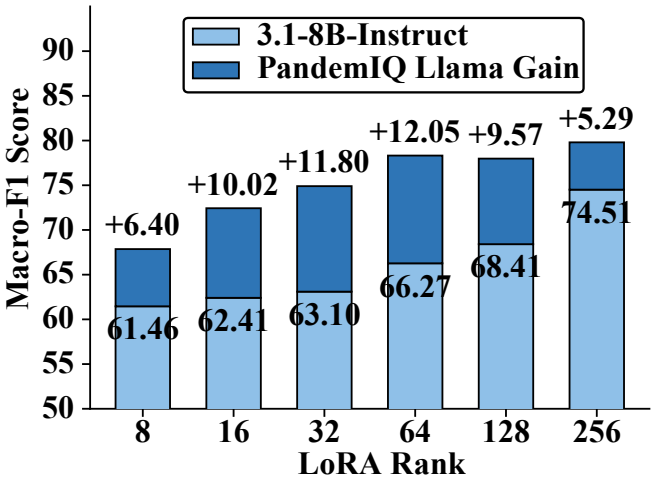

Figure 6: Performance on EVENT across LoRA ranks.

knowledge during pretraining leads to broader and more meaningful improvements than those achieved through *prompt engineering* or *task-specific fine-tuning* alone. Rather than simply excelling at isolated benchmarks, PandemIQ Llama demonstrates the ability to generalize across a wide range of biosurveillance tasks. This kind of generalization is critical for real-world public health needs, as it enables more accurate outbreak detection, faster risk assessment, and better-informed intervention strategies.

Beyond the benchmarks reported here, PandemIQ Llama is powering a deployed biosurveillance platform called Biothreats Emergence, Analysis and Communications Network (BEACON)[1], where it is used in a variety of real-world applications. These include extracting epidemiological data from unstructured reports, assessing event risk, and generating structured summaries for expert users. Its performance in these operational settings underscores its practical value for outbreak analysts, decision-makers, and public health professionals.

[1]https://beaconbio.org

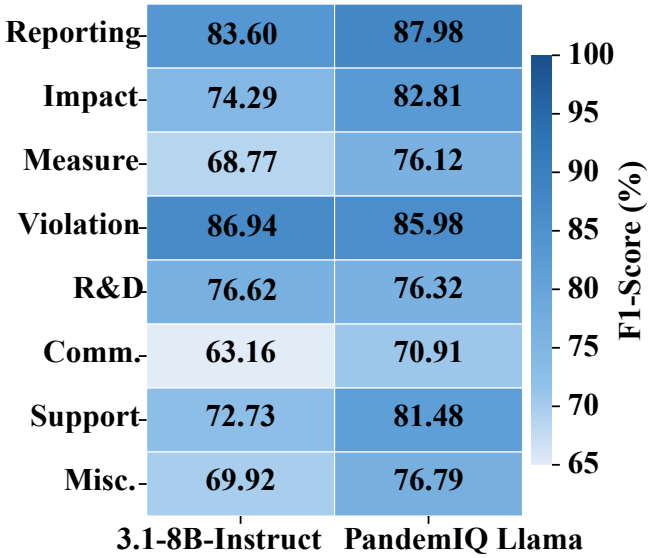

Figure 7: Performance across eight event categories: Reporting, Impact, Measure, Violation, R&D (Research), Comm. (Communication), Support, and Misc. (Miscellaneous).

By embedding domain knowledge into the foundation model itself, PandemIQ Llama helps close the gap between general-purpose AI and the specialized needs of public health. Just as BioBERT and ClinicalBERT advanced biomedical and clinical NLP, PandemIQ Llama is intended to serve as a foundational model for the public health NLP community. We have released the model publicly to support further research and development, allowing others to apply it to emerging biosurveillance challenges and build tools that enhance global pandemic preparedness. Ultimately, this work aims to support more effective monitoring and response to infectious disease threats, contributing to improved public health outcomes worldwide.

Social Impact: Pandemic intelligence has been underserved by existing NLP tools because it requires specialized domain expertise. This work fills this critical gap by providing the first dedicated domain-adapted LLM specifically designed for pandemic intelligence applications. The urgent societal need for such capabilities has been highlighted by recent global health challenges, making our work not just academically interesting but practically essential for future pandemic preparedness. The BEACON platform, powered by our model, has been launched and now serves over 100 government and multilateral public health organizations and users across 154 countries.

Ethics Statement

The model's knowledge cutoff (April 15, 2024) constrains its applicability to emerging infectious diseases that appear

afterward. Like all LLMs, PandemIQ Llama may hallucinate factually incorrect information – a particular concern in biosurveillance contexts where misinformation could impact public health decision-making. We recommend human expert oversight when our model is used directly or fine-tuned for downstream tasks, as inherited biases and limitations may persist through adaptation.

Acknowledgments

The research was partially supported by the NSF under grants CCF-2200052, IIS-1914792, ECCS-2317079, DEB-2433726, and NAIRR240274, the NIH under grant UL54 TR00413, Nvidia, the Gates Foundations, the Tianqiao and Chrissy Chen Foundation, Boston University, and other funders of the Biothreats Emergence Analysis & Communications Network (BEACON).

References

Alsentzer, E.; Murphy, J. R.; Boag, W.; Weng, W.-H.; Jin, D.; Naumann, T.; and McDermott, M. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*.

Bastani, H.; Drakopoulos, K.; Gupta, V.; Vlachogiannis, I.; Hadjichristodoulou, C.; Lagiou, P.; Magiorkinis, G.; Paraskevis, D.; and Tsiodras, S. 2021. Efficient and targeted COVID-19 border testing via reinforcement learning. *Nature*, 599(7883): 108–113.

BlueDot Inc. 2025. BlueDot: The World's Most Trusted Infectious Disease Intelligence. Accessed: 2025-08-01.

Brownstein, J. S.; Rader, B.; Astley, C. M.; and Tian, H. 2023. Advances in artificial intelligence for infectious-disease surveillance. *New England Journal of Medicine*, 388(17): 1597–1607.

Chen, Q.; Du, J.; Hu, Y.; Keloth, V. K.; Peng, X.; Zheng, A.; Xie, Q.; and Xu, H. 2024. Multimodal Large Language Models in Health Care: Applications, Challenges, and Future Outlook. *Journal of Medical Internet Research*, 26: e59505.

EPIWATCH. 2025. EPIWATCH: AI-driven epidemic early warning system. Accessed: August 2, 2025.

Fu, Z.; Zhang, M.; Meng, Z.; Shen, Y.; Buckeridge, D.; and Collier, N. 2024. BAND: biomedical alert news dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18012–18020.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Kim, J.; and Ahn, I. 2021. Infectious disease outbreak prediction using media articles with machine learning models. *Scientific Reports*, 11(1): 4413.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

MacIntyre, C. R.; Chen, X.; Kunasekaran, M.; Quigley, A.; Lim, S.; Stone, H.; Paik, H.-y.; Yao, L.; Heslop, D.; Wei, W.; et al. 2023. Artificial intelligence in public health: the potential of epidemic early warning systems. *Journal of International Medical Research*, 51(3): 03000605231159335.

Mollalo, A.; Mao, L.; Rashidi, P.; and Glass, G. E. 2019. A GIS-based artificial neural network model for spatial distribution of tuberculosis across the continental United States. *International Journal of Environmental Research and Public Health*, 16(1): 157.

Müller, M.; Salathé, M.; and Kummervold, P. E. 2023. Covid-twitter-bert: A natural language processing model to analyse Covid-19 content on twitter. *Frontiers in Artificial Intelligence*, 6: 1023281.

Parums, D. V. 2023. Infectious disease surveillance using artificial intelligence (AI) and its role in epidemic and pandemic preparedness. *Medical science monitor: international medical journal of experimental and clinical research*, 29: e941209–1.

Piskorski, J.; Stefanovitch, N.; Doherty, B.; Linge, J. P.; Kharazi, S.; Mantero, J.; Jacquet, G.; Spadaro, A.; Teodori, G.; et al. 2023. Multi-label Infectious Disease News Event Corpus. In *Text2Story@ ECIR*, 171–183.

Ramchandani, A.; Fan, C.; and Mostafavi, A. 2020. Deepcovidnet: An interpretable deep learning model for predictive surveillance of Covid-19 using heterogeneous features and their interactions. *IEEE Access*, 8: 159915–159930.

Reich, N. G.; McGowan, C. J.; Yamana, T. K.; Tushar, A.; Ray, E. L.; Osthus, D.; Kandula, S.; Brooks, L. C.; Crawford-Crudell, W.; Gibson, G. C.; et al. 2019. Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the US. *PLoS Computational Biology*, 15(11): e1007486.

Shastri, S.; Singh, K.; Kumar, S.; Kour, P.; and Mansotra, V. 2020. Time series forecasting of Covid-19 using deep learning models: India-USA comparative case study. *Chaos, Solitons & Fractals*, 140: 110227.

Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S. S.; Wei, J.; Chung, H. W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972): 172–180.

Sundermann, A. J.; Chen, J.; Kumar, P.; Ayres, A. M.; Cho, S. T.; Ezeonwuka, C.; Griffith, M. P.; Miller, J. K.; Mustapha, M. M.; Pasculle, A. W.; et al. 2022. Whole-genome sequencing surveillance and machine learning of the electronic health record for enhanced healthcare outbreak detection. *Clinical Infectious Diseases*, 75(3): 476–482.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Xie, Q.; Chen, Q.; Chen, A.; Peng, C.; Hu, Y.; Lin, F.; Peng, X.; Huang, J.; Zhang, J.; Keloth, V.; Zhou, X.; Qian, L.; He, H.; Shung, D.; Ohno-Machado, L.; Wu, Y.; Xu, H.; and Bian, J. 2024. Me LLaMA: Foundation Large Language Models for Medical Applications. *arXiv preprint arXiv:2402.12749*.

Yang, J.; Walker, K. C.; Bekar-Cesaretli, A. A.; Hao, B.; Bhadelia, N.; Joseph-McCarthy, D.; and Paschalidis, I. C. 2024. Automating biomedical literature review for rapid drug discovery: Leveraging GPT-4 to expedite pandemic response. *International Journal of Medical Informatics*, 105500.

Zhang, L.; Guo, W.; and Lv, C. 2024. Modern technologies and solutions to enhance surveillance and response systems for emerging zoonotic diseases. *Science in One Health*, 3: 100061.

Zhao, Y.; Gu, A.; Varma, R.; Luo, L.; Huang, C.-C.; Xu, M.; Wright, L.; Shojanazeri, H.; Ott, M.; Shleifer, S.; et al. 2023. PyTorch FSDP: Experiences on Scaling Fully Sharded Data Parallel. *Proceedings of the VLDB Endowment*, 16(12): 3848–3860.