Accelerating medical curation via LLM labelling

Isaque V. M. Pim[1], Eduardo Adame[1], Juliano Genari[1], Felipe Adeildo[2], Daniel Csillag[1], Guilherme T. Goedert[1].

[1]Escola de Matemática Aplicada, Fundação Getulio Vargas, Rio de Janeiro, RJ, Brazil
[2]Insper, São Paulo, SP, Brazil

FGV EMAp

Background

Syndromic surveillance systems are essential for early detection of disease outbreaks and emerging health threats [2]. Traditional manual curation of clinical data is time-consuming and resource-intensive, creating critical delays in public health response. Large Language Models (LLMs) offer a promising solution to **accelerate medical data processing while maintaining accuracy**. However, deploying LLMs in healthcare settings requires addressing challenges of **reliability, interpretability, and statistical validity**. Active Learning (AL) frameworks [1, 3] can optimize the annotation process by strategically selecting the most informative samples for human review, reducing labeling effort while improving model performance. This work addresses these challenges by combining LLM-based classification with conformal prediction and active learning to create a **robust, scalable system for syndromic surveillance in real-world clinical settings**.

OLIM development

In order to make our methodology easily available, we also developed **OLIM (Open Labeller for Iteractive Machine Learning, Figures 1 and 2)**. It provides a web-based interface for collaborative text labelling, featuring role-based access, Elasticsearch-powered text filtering, and bulk export and import operations. Tightly integrated with the AL framework it prioritizes uncertain samples for annotation. Dockerized deployment supports cloud or on-premise setups.

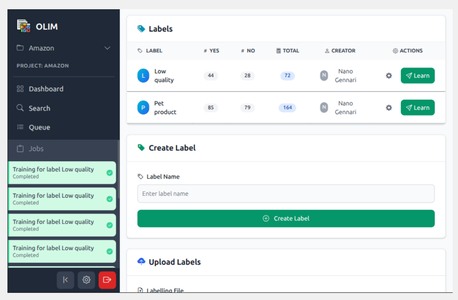

Figure 1. Label management dashboard with active learning controls.

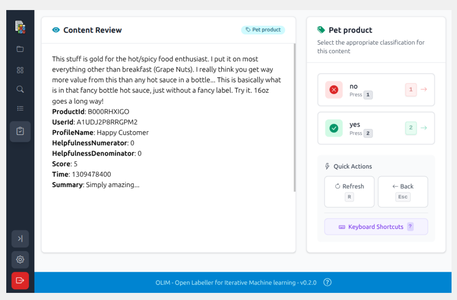

Figure 2. Interaction page for domain specialists.

Conformal Active Learning

Our framework combines conformal prediction with automated selection to **transform LLM outputs into reliable predictions with statistical guarantees**. By leveraging human annotations to train a selection model, we ensure both statistical rigor and practical utility in high-stakes applications.

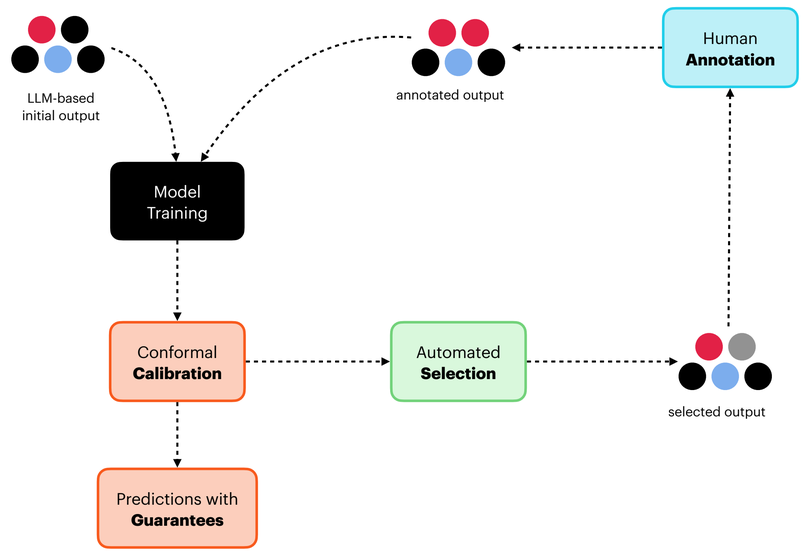

Figure 3. Active learning workflow.

The **conformal calibration** step provides formal coverage guarantees, ensuring predictions contain the true answer with user-specified confidence. **Automated selection** then filters these calibrated outputs based on quality thresholds learned from annotations.

References

[1] Juliano Genari and Guilherme Tegoni Goedert. *Mining Unstructured Medical Texts With Conformal Active Learning*. *AAAI DAI*. 2025.

[2] Kenneth D. Mandl et al. "Implementing Syndromic Surveillance: A Practical Guide Informed by the Early Experience". *Journal of the American Medical Informatics Association*. 2004.

[3] Yu Xia et al. "From Selection to Generation: A Survey of LLM-based Active Learning". *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. July 2025.

Performance of LLM-based classification

We augmented standard Active Learning with a **zero-shot LLM classifier**, using carefully designed prompts to generate preliminary labels on a subset of data. This dataset, comprising 100,000 clinical records from São João do Meriti (Jan-Apr 2025), was accessed via a partnership with the **Rio de Janeiro State Health Department (SES-RJ)**. Our primary goal is **measles surveillance** through detection of core symptoms: fever, cough, coryza, exanthema, and conjunctivitis For evaluation of the LLM classification, an initial subset of the data was manually labeled for each symptom. For classification, we first selected a model from a pool of open-weight options based on performance on a validation set and technical constraints (i.e., model size). The top-performing model was Qwen2.5-1.5B.

We employed two prompting strategies for symptom classification:

1. **Simple Prompt**: A direct instruction for the model to output 'present' or 'absent' for a given symptom based on the clinical note.
2. **Knowledge-Enhanced Prompt**: This method involved injecting a dictionary of curated medical knowledge definitions into the prompt context to ground the model's reasoning in clinical terminology.

Symptom	Accuracy	Precision	F1	Sensitivity	N
Conjunctivitis	0.911	0.837	0.854	0.872	157
Coryza	0.829	0.750	0.822	0.909	76
Exanthema	0.971	0.981	0.972	0.964	103
Fever	0.915	0.828	0.896	0.976	224
Cough	0.984	1.000	0.981	0.964	126

Table 1. Performance of the LLM on our manually labeled dataset for validation. Inference for all large language models (LLMs) was performed on an NVIDIA RTX 6000 Ada Generation GPU with 48 GB of VRAM.

Classification via LLMs offers key advantages over classical NLP by resolving ambiguous cases through contextual reasoning. This is evidenced by errors from the Simple Prompt that were corrected by the Knowledge-Enhanced approach (translated example):

> Female patient [...] **presenting with a maculopapular rash** on the upper limbs, [...]. The patient is at the appointment with her mother,[...] **not presenting with a rash**.

Here, the LLM correctly attributes the rash to the daughter, while a simple keyword search would be misled by the negation applied to the mother. This demonstrates the LLM's superior ability to handle linguistic complexity, especially for rare symptoms like exanthema.
We defined a suspected measles case based on the standard clinical criteria: a patient presenting with **fever, exanthema, and at least one of the following: conjunctivitis, coryza, or cough**. The resulting time series of suspected cases, smoothed with a moving average filter, is shown below.

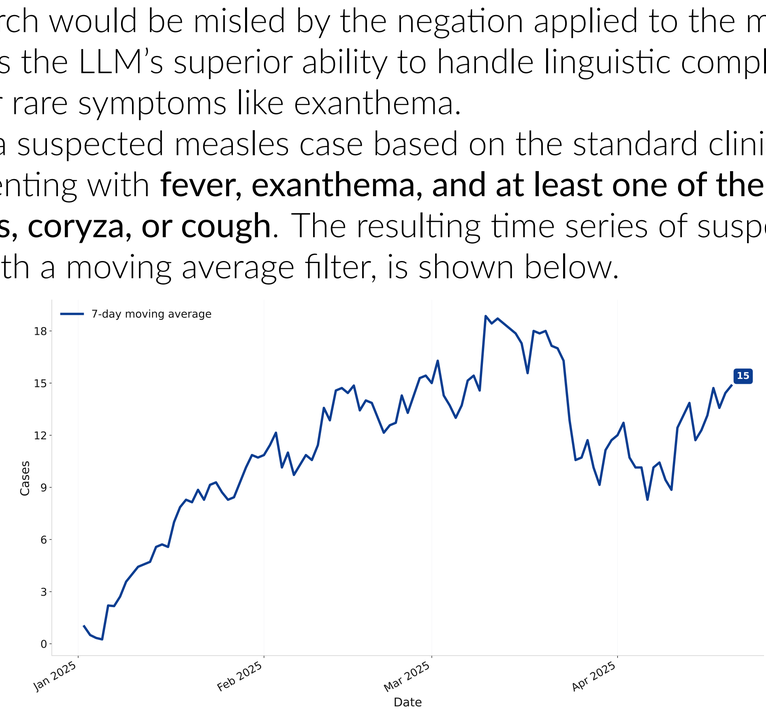

Figure 4. Trend of suspected measles cases over time (7-day moving average).

Discussion

Our results demonstrate that Large Language Models (LLMs) significantly improve symptom detection in clinical notes. Future work will focus on the following directions:

- **Expanding the knowledge base**: medical knowledge to enhance LLM prompting.
- **Incorporating spatial data**: enabaling outbreak analysis and cluster detection.
- **Domain-specific refinement**: Fine-tuning LLMs on Brazilian clinical notes.

https://gitlab.com/nanogennari/olim