ARIA
APPLIED RESEARCH IN ACTION

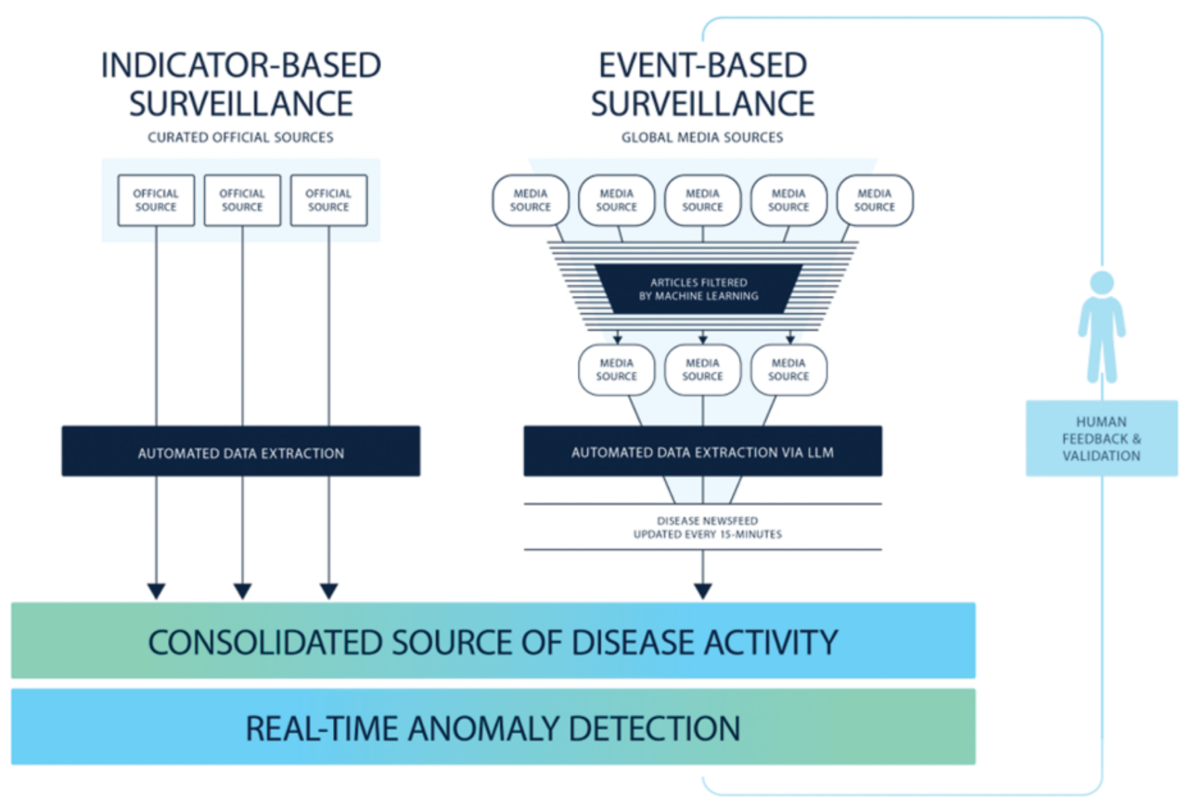
A Global Surveillance System for Infectious Diseases and Bio-Security Threats

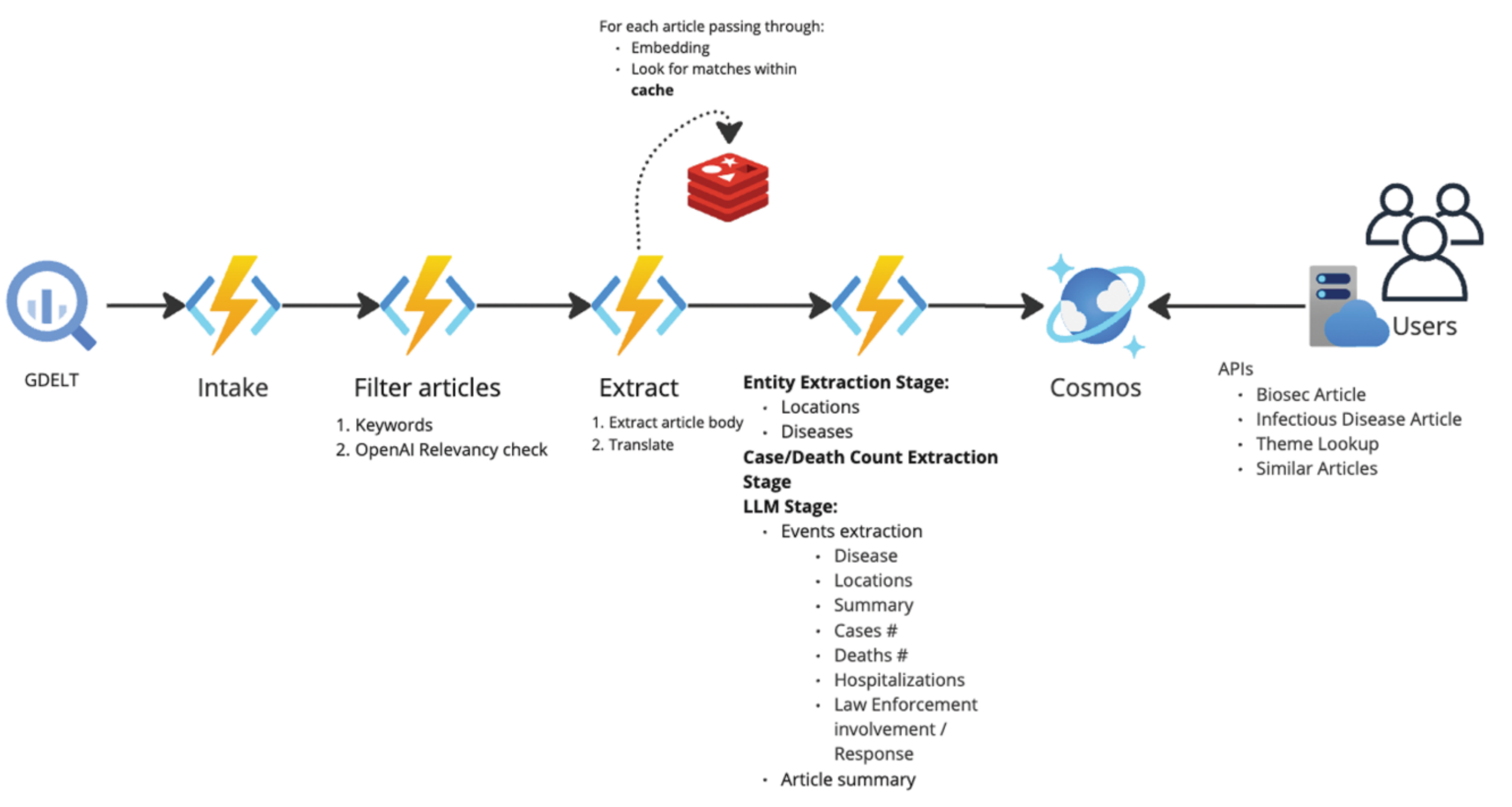
Revolutionizing Outbreak Analysis Using Large Language Models for Comprehensive Disease Event Extraction

Naveen Thangavelu

Annie En-Shiun Lee
ACADEMIC SUPERVISOR

Tim Lambertus
INDUSTRY SUPERVISOR


INDICATOR-BASED SURVEILLANCE
CURATED OFFICIAL SOURCES
OFFICIAL SOURCE
OFFICIAL SOURCE
OFFICIAL SOURCE
AUTOMATED DATA EXTRACTION
EVENT-BASED SURVEILLANCE
GLOBAL MEDIA SOURCES
MEDIA SOURCE
MEDIA SOURCE
MEDIA SOURCE
MEDIA SOURCE
MEDIA SOURCE
ARTICLES FILTERED BY MACHINE LEARNING
MEDIA SOURCE
MEDIA SOURCE
MEDIA SOURCE
AUTOMATED DATA EXTRACTION VIA LLM
DISEASE NEWSFEED UPDATED EVERY 15-MINUTES
HUMAN FEEDBACK & VALIDATION
CONSOLIDATED SOURCE OF DISEASE ACTIVITY
REAL-TIME ANOMALY DETECTION



PROJECT SUMMARY

Our research study focuses on extracting infectious diseases and bio-security threat-related events across a diverse array of news articles spanning over 65 languages and 200+ locations. Timely intervention is paramount, as any delays in response can lead to devastating consequences. To illustrate, had decisive action been taken merely a week earlier during the COVID-19 pandemic, the global mortality rate could have been curtailed significantly by 44.1% [1]. This task presents a multifaceted challenge: not merely discerning specific details such as types of diseases, geographic locations, and chronological data but also comprehending the nuanced context in which these events unfold. While existing Named Entity Recognition solutions partially address this challenge by capturing individual entities of interest, our methodology utilizes Large Language models and few-shot prompting [2] to reframe the extraction problem as a text completion challenge. To enhance the accuracy of our results, we have employed prompting techniques like Chain of Thought [3] and Chain of Verification [4] aimed at mitigating hallucinations. We've curated an evaluation benchmark dataset comprising articles from diverse geographical locations encompassing a rich spectrum of disease types and established evaluation metrics to refine our methods. Our results show a substantial performance improvement in extracting disease names, achieving a 14% increase. Our ongoing work includes efforts in Parameter-Efficient fine-tuning [5] and Retrieval Augmented Generation [6] to address hallucinations. In strengthening global surveillance, we are taking a vital step towards building a resilient world capable of responding promptly to future pandemics, saving lives globally.

REFERENCES

[1] Piovani D, Christodoulou MN, Hadjidemetriou A, Pantavou K, Zaza P, Bagos PG, Bonovas S, Nikolopoulos GK. Effect of early application of social distancing interventions on COVID-19 mortality over the first pandemic wave: An analysis of longitudinal data from 37 countries. J Infect. 2021 Jan;82(1):133-142. doi: 10.1016/j.jinf.2020.11.033. Epub 2020 Dec 1. PMID: 33275956; PMCID: PMC7706420.

[2] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., . . . Amodei, D. (2020). Language Models are Few-Shot Learners. ArXiv. /abs/2005.14165

[3] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. ArXiv. /abs/2201.11903

[4] Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2023). Chain-of-Verification Reduces Hallucination in Large Language Models. ArXiv. /abs/2309.11495

[5] Hu, E. J., Shen, Y., Wallis, P., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. ArXiv. /abs/2106.09685

[6] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. ArXiv. /abs/2005.11401

bluedot

Computer Science
UNIVERSITY OF TORONTO

Master of Science in Applied Computing