Japiim Health: Democratizing Access to Brazilian Arbovirus Data via Retrieval-Augmented Generation on WhatsApp

Ana Rita Marega Gonçalves[1], Nicolas de Sousa Maia[1], Melissa de Sousa Maia[1], Marina Carvalho Arruda Barreto[1], Robson Parmezan Bonidia[2], André de Carvalho [1]

[1]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
Av. Trabalhador São-carlense, 400, 13566-590, São Carlos-SP, Brasil

[2]Departamento de Computação – Universidade Tecnológica Federal do Paraná
Av. Alberto Carazzai, 1640 — Centro, Cornélio Procópio–PR, 86300-000, Brasil

Abstract. *Although the "Sistema de Informação de Agravos de Notificação" is fundamental for epidemiological surveillance in Brazil, its complex tabular format and the lack of contextual interpretation in existing visualization tools create significant barriers to data accessibility. In this context, we present Japiim, a multimodal chatbot integrated with WhatsApp, developed to democratize access to national microdata on arboviruses. Based on a Retrieval-Augmented Generation (RAG) architecture and the Sabiá-4 language model, Japiim processes granular health records, transforming them into intuitive texts, voice responses, and dynamic charts. A comparative analysis with general-purpose language models (ChatGPT, Gemini, Claude) reveals that Japiim mitigates quantitative hallucinations and imposes strict safety limits, avoiding the generation of inappropriate medical advice. By combining a widely adopted and low-friction communication platform with a foundation in reliable data, Japiim empowers health managers and citizens with precise and immediate information on public health.*

1. Background

Historically, dengue has been one of the most persistent and complex public health challenges in Brazil [Barkhad et al. 2026]. For decades, disease has imposed recurrent epidemic cycles, placing a continuous burden on the healthcare system. In 2024, in Brazil, more than 6.5 million cases were registered, resulting in approximately 190,000 hospitalizations and over 6,300 deaths, according to SINAN (*Sistema de Informação de Agravos de Notificação*) data. The alarming numbers reflect the chronic difficulty in controlling the disease, a challenge with multiple causes, such as climate change, the persistence of vector breeding sites, and the limited effectiveness of government initiatives [Gurgel-Gonçalves et al. 2024].

Arboviruses, such as Dengue, Chikungunya, and Zika, require mandatory notification [Ministério da Saúde 2025], generating a continuous stream of SINAN data. Consequently, this system serves as a fundamental tool for public health, enabling the monitoring of the temporal and spatial evolution of diseases, the identification of vulnerable populations, and the formulation of evidence-based public policies. The availability of this information in time is crucial for strengthening epidemiological surveillance, optimizing resource allocation, and promoting societal engagement. Despite the importance of SINAN, data accessibility remains a significant challenge due to the structural rigidity of existing platforms [Fernandez et al. 2025].

Although current systems provide comprehensive tabular data, they lack contextual interpretation capabilities, forcing health managers to perform complex manual processing that delays strategic actions and distances civil society from information [Wu et al. 2026]. The gap is not addressed by governmental visualization initiatives, such as the Ministry of Health's Arbovirus Monitoring Dashboard, which assumes levels of data literacy and technical proficiency beyond the realities of many users while offering no direct interface for specific and localized queries [Henriques et al. 2024].

Moreover, the inequality in digital access and the low technological literacy among significant portions of the Brazilian population further deepen this barrier [Comini et al. 2024]. As a result, epidemiologically valuable data become inert assets, highlighting the need to move beyond static visualizations toward approaches capable of interpreting epidemiological context and responding intuitively to specific demands, a gap for which Artificial Intelligence (AI), especially Large Language Models (LLMs), emerges as a promising alternative.

In this context, when applied to epidemiological surveillance, AI presents great potential to support public management and Primary Health Care, contributing to improved efficiency and resource allocation in health services [Qin and Tong 2025]. Even so, its application in sensitive areas requires reliability and credibility: beyond conversational capabilities, models must be grounded in official data sources to avoid the "hallucinations" commonly associated with generative systems.

Similarly, for these tools to effectively impact public health, they must be accessible, eliminating technical barriers that still distance managers and citizens from epidemiological information [Chiat et al. 2024]. The main contribution of this work, therefore, lies in the development of a multimodal RAG-based framework integrated with ubiquitous messaging platforms, aimed at bridging the gap between complex governmental microdata repositories and the immediate need for intuitive epidemiological insights for non-technical users.

2. Japiim: Democratizing Access for All

This work presents Japiim Health[1], a solution designed to overcome barriers to the interpretation of SINAN data. Japiim consists of a fully integrated WhatsApp chatbot that uses the RAG technique to process the Brazilian arbovirus database, covering all states and municipalities. Its differential lies in the ability to transform complex tabular data into natural dialogues and on-demand graphic reports. Unlike static dashboards, the tool democratizes information by delivering accurate epidemiological analyses directly to the user's smartphone, allowing both managers and the population to understand the local health scenario without the need for advanced technical literacy.

2.1. Data Acquisition and Processing

Data acquisition was performed using the *PySUS* library to interface with the DATASUS infrastructure. We extracted SINAN microdata covering the period 2017 to the present for three main arboviral diseases: Dengue, Zika, and Chikungunya. The raw dataset consists of granular notification records, where each entry represents a single suspected

[1]https://japiimbr.inteligentehub.com.br/

clinical case. To ensure consistency, we developed a custom decoding pipeline based on the official data dictionary to standardize numerically encoded categorical variables. Moreover, our analysis incorporates all suspected notifications rather than limiting the scope to confirmed cases. This approach captures the immediate demand on the healthcare system and facilitates early-warning signals before final diagnostic confirmation. From these processed data, we generated a series of structured visual and textual artifacts:

- **Annual Seasonality Analysis:** Line charts illustrating the intra-annual variation of case volume to identify seasonal peak periods.
- **Longitudinal Trend Analysis:** Multi-year time series visualizing the number of cases and outbreak cycles across the full period.
- **Spatio-Temporal Heatmaps:** Dynamic density maps, in the form of videos, showing the geospatial evolution of the disease, aggregated by municipality (`ID_MUNICIP`).
- **Demographic Profiling:** Age-sex pyramids depicting the distribution of the reported patient population.
- **Monthly Epidemiological Aggregates:** A structured summary that details incidence rates, mortality, and demographic stratification (sex and age group) aggregated monthly for each year.
- **Severe Dengue & Comorbidity Profile:** A specialized report on Dengue focusing on severe clinical manifestations. It analyzes the correlation between severe outcomes, specific comorbidities, and mortality rates.
- **National Symptomatology Analysis:** A statistical overview of clinical symptoms reported in the national territory, providing a frequency distribution of symptoms.

2.2. RAG Pipeline Architecture

To ensure response accuracy and mitigate the risk of hallucinations, we implemented an RAG architecture that grounds the LLM's reasoning in our processed SINAN artifacts. The pipeline, illustrated in Figure 1, follows a multi-stage retrieval strategy designed to balance relevance with latency. The inference process consists of four distinct stages:

1. **Input Guardrails and Orchestration:** Upon receiving a user query, the system employs a dynamic orchestration layer powered by Sabiá-4, rather than relying on static keyword lists or regex-based filters. This LLM-based gatekeeper performs semantic intent classification to determine the appropriate workflow: routing the query to the RAG pipeline for epidemiological analysis, triggering a geolocation-based search for healthcare facilities, or discarding the input if it falls outside the predefined public health scope. This approach ensures high robustness against linguistic variations and prevents the model from processing irrelevant or unsafe prompts that could lead to off-scope medical advice.
2. **Context-Aware Retrieval:** Unlike standard semantic search, our retriever incorporates user-specific context metadata. The vector store was implemented using PostgreSQL with the pgvector extension, orchestrated through the langchain_community library. This architecture enables a unified data ecosystem where relational metadata (state, municipality, disease type) and high-dimensional embeddings are stored in the same database. The system executes a similarity search based on cosine distance to retrieve an initial candidate pool of approximately 20 documents that strictly match the user's geographic and disease-specific metadata constraints.

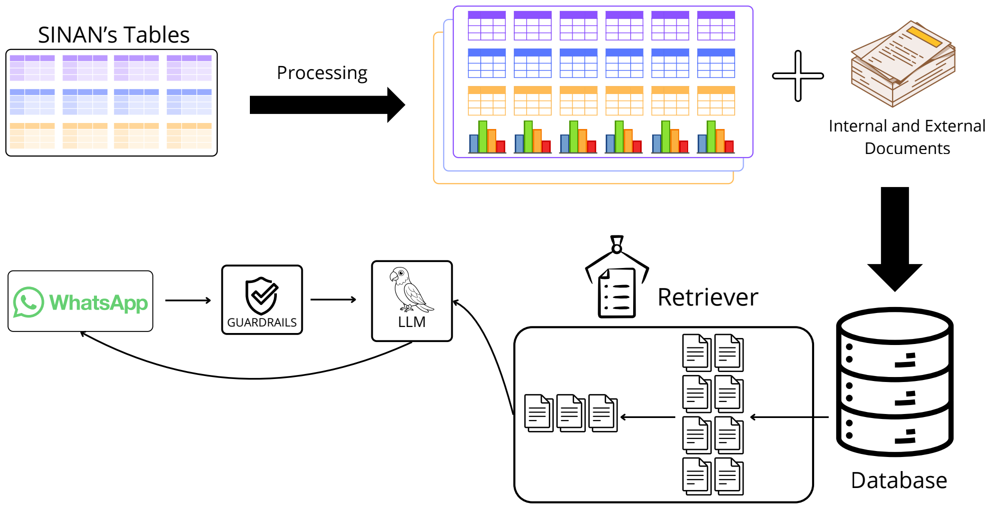

Figura 1. Japiim Architecture

3. **Reranking:** To optimize the context window, we employ a reranking model that scores the relevance of the initial candidates against the user's specific question. The system selects the top-k textual reports for analytical depth and the top-k visual artifacts to provide interpretative support.
4. **Augmented Generation:** The selected text context is appended to the system prompt, and the final response is generated and delivered to the user. After this, the user receives a list of the top-k visual artifacts to select which ones to see.

For the generation component of our pipeline, we selected **Sabiá-4**, a LLM developed by Maritaca AI[2]. Unlike generic multilingual models that are often biased towards English-centric training data, Sabiá-4 is specifically engineered for high-fidelity performance in Brazilian Portuguese.

2.3. User Experience (UX) and Accessibility

The user interface was guided by the principles of inclusivity and low-friction access. We recognized that for a public health tool to be effective in Brazil, it must meet the population where it is already. Consequently, the system is deployed through **WhatsApp**, the most widely used communication platform in the country. This strategic choice eliminates the need for users to download new applications or learn complex interfaces, democratizing access to epidemiological data.

To further enhance accessibility and engagement, we implemented several key UX features. Recognizing that literacy barriers and visual impairments can hinder access to information, our agent supports full voice interaction through multimodal capabilities. Users can send voice questions and receive audio responses, ensuring that the tool is accessible to those who cannot or prefer not to type. The system also features dynamic intent detection through a layer capable of detecting context shifts.

[2]https://www.maritaca.ai/

If a user wishes to change their target location or inquire about a different disease mid-conversation, the intent detector perceives this and asks the user which new context is wanted. Given that trust is paramount in health applications, we adopted a privacy-first approach. The interaction begins with a transparent Terms of Service agreement, clearly explaining that while conversation data are processed for the answer, no Personally Identifiable Information (PII) is stored, ensuring compliance and user safety.

3. Results and Discussion

To evaluate the performance of Japiim against general-purpose LLMs, we curated a comprehensive evaluation dataset. The queries used for this assessment consisted of a mixture of general knowledge questions about arbovirus diseases and specific numerical questions. To ensure transparency and reproducibility, the complete set of questions, the corresponding ground truth references, and the raw responses generated by each evaluated LLM are available in the supplementary spreadsheet[3]. The comparative performance of Japiim, ChatGPT, Gemini, and Claude is summarized in Table 1. The evaluation was conducted using three primary metrics to capture both the qualitative relevance and quantitative accuracy of the models, based on the RAGAS framework [Es et al. 2024] for the evaluation of LLMs:

- **Factual Correctness:** This metric represents the overall accuracy of the responses. For general questions on diseases, the score is calculated using the traditional Ragas framework methodology. However, because the standard Ragas evaluation struggles to accurately assess pure numerical answers, we applied a distinct numeric factual correctness method for quantitative questions. We use this blended approach to show the overall score of factual correctness across the diverse set of questions.
- **Numeric Factual Correctness:** This metric strictly evaluates the precision of numerical data retrieval. It is calculated as $1 - error$ (comparing the numeric response with the truth of the ground). A strict threshold is applied: the score is only awarded if the error is less than 5%; if the error equals or exceeds this threshold, the resulting score is 0.0.
- **Answer Relevancy:** This metric assesses how directly and appropriately the model addresses the user's prompt, independent of the factual ground truth.

Tabela 1. Performance Comparison of LLMs across Evaluation Metrics

Model	Factual Correctness	Numeric Factual Correctness	Answer Relevancy
Japiim	**0.501**	**0.689**	**0.941**
ChatGPT-5.1	0.164	0.000	0.821
Gemini 3 Flash	0.195	0.000	0.884
Claude 4.6 Sonnet	0.193	0.000	0.654

As demonstrated by the results, general-purpose LLMs are capable of achieving non-zero scores in overall Factual Correctness, largely due to their ability to correctly

[3]https://encurtador.com.br/gaXd

answer broader, general knowledge questions regarding diseases. However, their performance completely fails when evaluated on Numeric Factual Correctness, where ChatGPT, Gemini, and Claude all scored 0.0. This stark contrast highlights a critical limitation: while generic models can converse accurately about general medical concepts, they cannot reliably retrieve or provide real epidemiological numbers without a grounded domain-specific architecture.

In contrast, Japiim significantly outperformed baseline models in all accuracy metrics. Using its specialized RAG pipeline anchored in official public health data, Japiim achieved a numerical fact correctness of 0.689 and the highest overall fact correctness (0.501). Consequently, for tasks requiring specific numerical data and exact health statistics, Japiim performs better and is recommended over generic counterparts.

Furthermore, Japiim achieved the highest Answer Relevancy score (0.941). It is important to note that Answer Relevancy does not measure accuracy against the ground truth, but rather how well and contextually appropriately the models structure their answers to the questions. The fact that Japiim outperformed the other models in this metric indicates that its specialized prompt and retrieval design not only grounds its facts but also helps it generate more focused, relevant, and structurally coherent answers compared to the broader responses provided by the other LLMs.

Another limitation of generic LLMs refers to patient safety. Models like Gemini and Claude provided direct medical recommendations, suggesting specific medications. This behavior is dangerous in encouraging self-medication, involving risks such as incorrect self-diagnosis and adverse reactions. Such responses incorrectly position the systems as sources of therapeutic guidance, transferring clinical risk to the machine. In contrast, Japiim imposes strict safety limits (guardrails), establishing a clear boundary between epidemiological information and personalized advice. The tool actively instructs the user to seek the healthcare network, keeping healthcare professionals as the only appropriate authorities for diagnosis and treatment.

4. Conclusion

Japiim represents an advance in the democratization of Brazilian epidemiological data by reducing the gap between the technical complexity of SINAN and the need for rapid and understandable access to health information. Moreover, the system contributes to combating misinformation and expands the potential use of epidemiological data in decision-making processes. Integration with WhatsApp, which supports voice and text interactions, reduces barriers to digital and technological literacy, allowing more inclusive access to SUS information without the need for specialized training or additional applications. In this way, Japiim brings epidemiological information closer to the daily lives of the population and strengthens democratic access to public health knowledge. Finally, the project presents itself as a promising approach to support epidemiological awareness, strengthen public transparency, and contribute to more accessible and user-centered models for health data management in Brazil.

Acknowledgments

This work was funded by the International Development Research Center, the Foreign, Commonwealth and Development Office, and Sabiá Data Lake.

Referências

Barkhad, A., Santos, G., R C Campos, S., M A Braz, L., P Freitas, L., Zinszer, K., de Souza, R., Waldron, I., Loeb, M., Luna, E., and Mbuagbaw, L. (2026). Knowledge, attitudes, practices and perceptions of the eco-bio-social determinants of dengue transmission in são paulo, brazil: A mixed-methods study. *Trop Med Internation Health*, 31(1):58–79.

Chiat, J. et al. (2024). Artificial intelligence, chatgpt, and other large language models for social determinants of health: Current state and future directions. *Cell Reports Medicine*, 5(1):101356–101356.

Comini, N., Gozzi, N., and Perra, N. (2024). Bridging brazil's digital divide: How internet inequality mirrors income gaps. World Bank Blogs. Accessed: 2026-05-18.

Es, S., James, J., Anke, L. E., and Schockaert, S. (2024). Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations*, pages 150–158.

Fernandez, M., Pinto, H. A., Fernandes, L. M. M., Oliveira, J. A. S. d., Lima, A. M. F. d. S., Santana, J. S. S., and Chioro, A. (2025). Interoperability in universal healthcare systems: insights from brazil's experience integrating primary and hospital health care data. *Frontiers in Digital Health*, Volume 7 - 2025.

Gurgel-Gonçalves, R., Oliveira, W. K. d., and Croda, J. (2024). The greatest dengue epidemic in brazil: Surveillance, prevention, and control. *Rev Soc Bras Med Trop*, 57.

Henriques, C. M. P., Moura, N. F. O., and Souza, P. B. (2024). Challenges and lessons from the covid-19 pandemic for health surveillance in brazil: reflections on technologies, models, and system organization. *Rev Bras Epidemiol*, 27:e240049.

Ministério da Saúde (2025). Portaria gm/ms no 6.734, de 18 de março de 2025. Disponível em: ¡https://www.in.gov.br/en/web/dou/-/portaria-gm/ms-n-6.734-de-18-de-marco-de-2025-620767223¿.

Qin, H. and Tong, Y. (2025). Opportunities and challenges for large language models in primary health care. *J Primar Care & Commun Health*, 16.

Wu, E., Balise, R., Katz, B., Harris, D., Bullard, M., Fareed, N., Larochelle, M., and Villani, J. (2026). Building public health data dashboards: Tutorial playbook. *JMIR public health and surveillance*, 12:e83157.

AI Usage Declaration

The authors declare that Generative Artificial Intelligence tools (LLMs) were used solely to improve the clarity and textual quality of the manuscript.

Work Origin Description

The study is the result of an Undergraduate Research project focused on the development of the Japiim Health platform for the democratization of public health data in Brazil. Japiim Health is also one of the initiatives linked to the AutoAI Hub (https://autoaipandemics.icmc.usp.br/) and the AI4PEP network (https://ai4pep.org/), which is an international partnership encompassing 16 nations of the Global South.