Integrating Large Language Models and Geospatial Data for Territory-Aware Public Health Surveillance

Vinicius R. P. Borges[1], Arthur S. M. Freire[1], Luís P. F. Garcia[1], Anna Carolina F. Martins[1], Flávio de Barros Vidal[1], Aleteia Araujo[1], Edward T. Maia[2], Gabriel M. Veloso[2], Wagner de Jesus Martins[2], Vladimir Molchanov[3,4], Lars Linsen[3]

[1]University of Brasília and [2]Fundação Oswaldo Cruz, Brasília, Brazil
[3]University of Münster and [4]Leibniz FH, University of Applied Sciences, Hanover, Germany
Correspondence: viniciusrpb@unb.br

Abstract

Agents based on Large Language Models combined with Retrieval-Augmented Generation (RAG) can support public health surveillance by improving access to heterogeneous information sources and increasing transparency for citizens. This paper presents ongoing research on a multimodal conversational agent that integrates a RAG-based pipeline for extracting and localizing health-related information from official government public health news, epidemiological surveillance records, and geospatial datasets related to geological risk zones and wildfire-affected regions. Preliminary experiments were conducted using Llama3.1-8B and Qwen3-8B on a synthetic question-answering corpus of 500 pairs curated from official public health news issued by the Health Department of the Federal District, Brazil. Results showed that expanding the candidate pool with BM25 retrieval yields substantial gains over Vanilla RAG, while the Hybrid RAG, enriched with geographic matching and named entity reranking, further improves answer relevance and correctness. These findings highlight the potential of combining natural language understanding and geospatial reasoning to transform fragmented public health data into accessible territorial intelligence for researchers, health professionals, and citizens.

1 Introduction

The Brazilian Unified Health System (SUS) is grounded in the constitutional principles of universality, equity, and comprehensiveness, and is organized around Primary Health Care (*Atenção Primária à Saúde* - APS), which emphasizes territorialized and community-based care (Botero et al., 2025). Within this structure, Basic Health Units (UBSs) constitute the main physical infrastructure of APS and the preferred entry point to the public healthcare system, providing medical, nursing, dental, preventive, and health promotion services. In recent years, these units have increasingly adopted digital communication channels to disseminate information about vaccines, medications, and public health campaigns.

Despite its robust design, the SUS faces significant challenges, including funding constraints, fragmented information systems, and regional inequalities in infrastructure and healthcare workforce distribution (de Melo et al., 2024). These structural limitations become even more critical in regions affected by complex public health demands, such as tropical disease outbreaks (e.g., dengue), Severe Acute Respiratory Syndrome (SARS) surveillance, and environmental and climate-related events that directly impact population health, including wildfires, drought periods, and geological risk areas.

Public health surveillance in Brazil relies on territorialized monitoring practices carried out by Primary Health Care teams and Community Health Agents (ACS), who collect and analyze information about local health conditions and vulnerabilities within specific geographic and social contexts (Veríssimo et al., 2024). The collected data are subsequently analyzed by public health professionals, epidemiologists, and territorial surveillance teams, who combine structured data from official databases with unstructured community-level information gathered through territorial monitoring and social cartography (Martins et al., 2024).

However, the large volume of heterogeneous and fragmented data sources makes it difficult to identify correlations, detect emerging risks, and provide centralized information to citizens. In this context, Large Language Models (LLMs) and LLM-based agents (Plaat et al., 2026) have gained attention due to their ability to interact with external tools, databases, APIs, and heterogeneous information sources during the response generation process. Such architectures can retrieve up-to-date information and coordinate multiple processing steps to answer complex user queries, and have already

been successfully explored in other domains such as education (Chu et al., 2025) and medicine (Vatsal et al., 2026).

The current public health scenario in the Federal District, Brazil (Distrito Federal - DF), motivated us to investigate and develop an agentic architecture capable of integrating heterogeneous and geographically grounded information sources. The Federal District government publicly provides datasets related to healthcare services, primary healthcare indicators[1], and other repositories containing statistics at the census-sector level with demographic and environmental attributes. At the same time, public health facilities (including UBSs), are frequently mentioned in official government news portals reporting vaccination campaigns, healthcare task forces, and territorial interventions. Since these events and facilities naturally encode some type of georeference, their joint integration and analysis with epidemiological and environmental datasets can support contextualized situational awareness for both specialists and citizens.

This paper describes ongoing work on a multimodal conversational agent for public health surveillance that integrates a Retrieval-Augmented Generation (RAG) pipeline combining dense vector search, BM25 (Robertson et al., 2004) keyword retrieval, and entity-based reranking over a corpus of public health news articles from the Health Department of the Federal District (*Secretaria de Saúde do Distrito Federal* - SES-DF). The agent further supports interactive geospatial visualization by using official datasets, including dengue surveillance records, geological risk zones, and wildfire perimeters, enabling users to query epidemiological information in a single conversational interface.

We expect that the proposed solution allows non-specialist users to access epidemiological and geographic information through natural language interactions, without requiring expertise in data science or programming (Kwok et al., 2024). This paper provides the following contributions:

- a multimodal conversational agent that integrates natural language understanding with interactive geospatial visualization, enabling both structured map commands and public health queries over heterogeneous epidemiological and environmental datasets;
- a hybrid RAG architecture combining dense semantic retrieval, BM25 keyword matching, and LLM-based Named Entity Recognition (NER) for reranking, applied to a domain-specific corpus of public health news, alongside location and entity metadata.

This paper is structured as follows. Section 2 reviews related work on LLM and RAG applications for public health news and documents. Section 3 describes the modules of the proposed LLM-based agent and the heterogeneous datasets it integrates. Section 4 presents preliminary experiments and results. Section 5 discusses the limitations of the current work. Finally, Section 6 summarizes the conclusions and outlines directions for future work.

2 Related Work

LLMs have been increasingly applied in public health domains, particularly to improve access to health information and mitigate issues related to hallucinated or unreliable responses (Xu et al., 2025). Conversational agents have also been explored for several healthcare-related tasks, including public health education (Färber et al., 2025) and decision support or assistance systems (van Baal et al., 2024). We review below recent works that combine LLMs with geospatial information and the integration of datasets for health analysis.

RACCOON (Lin et al., 2025) is an open-source geocoding approach designed to automatically extract geographical coordinates from news articles to support applications like epidemic intelligence and disaster management. RACCOON adopts a RAG strategy to identify geolocations by retrieving candidate entries from the GeoNames gazetteer through a process involving country-assisted retrieval and population-based re-ranking. This information is then provided as context to an LLM to generate coordinates. Evaluation was performed using three benchmark datasets using quantitative metrics, demonstrating that RACCOON generally improves precision and consistency compared to standard LLM prompting and other baselines.

Niu et al. (2025) developed an AI-driven community application designed to solve the Local Life Information Accessibility (LLIA) problem by unifying information retrieval and personalized recommendations within the "15-minute city" framework. The proposed architecture features a three-layer RAG pipeline that combines geographic, graph-based, and semantic-vector retrieval, employing an

[1] https://info.saude.df.gov.br/sala-de-situacao/ accessed on June 10, 2026.

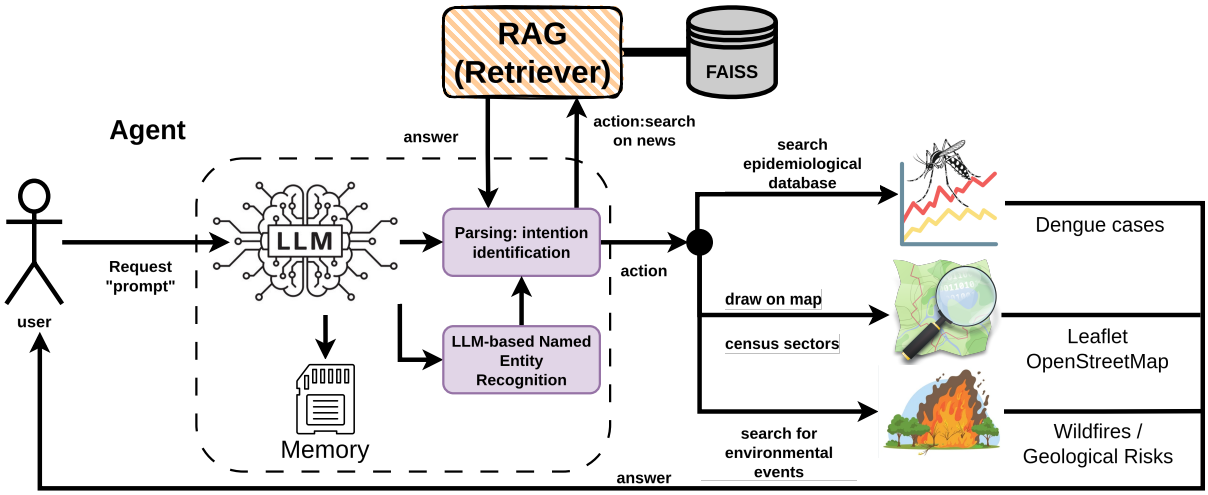

Figure 1: Overall workflow of the proposed conversational LLM-based agent.

LLM to extract spatial entities while a geospatial agent identifies geolocations via external APIs to create structured geometric representations. Evaluations conducted on a real-world benchmark of $20,000$ geotagged posts from Shenzhen show that AskNearby significantly outperforms existing LLM and map-based baselines in retrieval accuracy, ranking quality, and spatiotemporal grounding.

Chen et al. (2026) investigated how LLMs perform on urban spatial reasoning by proposing a novel geographic RAG framework (GeoRAG) to address spatial semantics. GeoRAG uses OpenStreetMap data to automatically build a spatio-temporal knowledge graph and incorporate both hierarchical and proximity-based spatial features. Experiments were performed on Qwen2.5-32B and using the CityGPT benchmark, which showed that GeoRAG significantly improved the spatial reasoning performance with accuracy gains of 11.89%, 13.25%, and 7.49% on Beijing, London, and New York datasets, respectively.

Although those works demonstrate the potential of combining LLMs with geospatial reasoning and metadata, they are tailored to specific data types, thus not being fully suitable to the heterogeneous nature of public health data. The proposed agent addresses this gap by integrating a hybrid RAG pipeline over a domain-specific news corpus with interactive geospatial layers from epidemiological and environmental datasets.

3 Proposed LLM-based Agent

Figure 1 illustrates the steps of the proposed conversational agent, which is integrated with an LLM that parses the user's inquiry to determine further actions. These actions reflect real-world information needs of both public health researchers, who require epidemiological and territorial analysis, and citizens searching for information about the conditions or updates over nearby health facilities.

The agent processes user requests through a four-stage pipeline. In the first stage, the input query is analyzed by an LLM that simultaneously detects the language of the request, currently supporting Portuguese and English, and extracts named entities corresponding to locations, organizations, and events (e.g., vaccination campaigns).

In the second stage, a command parsing module classifies the intent of the query into one of seven predefined action categories: drawing administrative boundary polygons; revealing census-sector grids; searching for points of interest; geocoding addresses; displaying thematic layers (e.g., dengue case distribution, geological risk zones or wildfire perimeters); answering analytical questions about epidemiological data; and routing public health questions to the knowledge retrieval pipeline.

The intent classification is performed by the LLM set at zero temperature, which receives the user query alongside a dynamically assembled prompt that lists the available action categories and their activation conditions. Geographic references such as administrative region names and health facility identifiers are also included in the prompt, ensuring the LLM maps locations provided by users to the correct spatial vocabulary. The LLM then generates a structured JSON object that includes not only the intent label but also the extracted parameters, such as the target region, the facility category, or a specific unit identifier, so that a single

inference determines what the agent should do and whether there are constraints associated with geographic and thematic references.

In the third stage, the action is dispatched to the appropriate module, which proceeds with its operations and returns a structured output. Currently, the agent supports four modules: a geospatial renderer, an external location service, a tabular data processing module, and the RAG pipeline for consulting the public health news corpus.

In the final stage, the agent assembles the module output and uses an LLM to generate a response in natural language in the same language as the original query, optionally accompanied by a map update reflecting the geographic entities identified during the interaction.

3.1 Public Health and Geospatial Data Sources

Primary Care Health Units

To support both retrieval filtering and map-based interactions, the proposed approach relies on a curated set of location references specific to the Federal District, Brazil. For primary healthcare facilities, the system uses an official dataset published by the Brazilian Ministry of Health[2] containing all UBSs registered in the country, from which the 199 units located in the Federal District were extracted, including their geographic coordinates, addresses, and neighborhoods.

Census Sectors

Administrative boundaries were obtained from the official census geometries provided by the Brazilian Institute of Geography and Statistics (*Instituto Brasileiro de Geografia e Estatística - IBGE*), covering all census sectors of the Federal District organized by Administrative Region (RA). This dataset allows the system to dynamically render territorial polygons and match location mentions in user queries against known administrative subdivisions.

Public health news

The knowledge base was constructed by scraping the official news portal of the SES using Selenium, collecting articles published from January 2023 onward. Each article's textual content, publication date, title, and metadata were extracted. Articles were then filtered to retain only those containing location references matching the neighborhoods and administrative regions covered by the UBS dataset described above, yielding a total of 1,628 news articles.

Epidemiological Datasets

The proposed agent incorporates epidemiological surveillance datasets published by the SES-DF, including dengue records from 2025 through April 2026. The dataset comprises approximately 16,000 notified cases, together with demographic, clinical, temporal, and geographic attributes derived from the Brazilian notifiable disease surveillance system (SINAN). These records are aggregated by census sector and integrated with the official territorial geometries of the Federal District, enabling the generation of dynamic geospatial layers for visualization and territorial analysis of dengue transmission patterns. In addition, the system integrates environmental and territorial datasets related to wildfire occurrences, geological risk areas, and demographic indicators, allowing the joint analysis of epidemiological and environmental factors that may affect population health.

3.2 LLM-based Named Entity Recognition

To improve retrieval precision, the agent incorporates an NER module that prompts an LLM to extract three entity categories from user queries prior to retrieval: *locations* (e.g., administrative regions and health facilities), *organizations* (e.g., health secretariats, programs, and agencies), and *events* (e.g., outbreaks, vaccination campaigns, and public holidays). The extracted entities are returned as a structured object that augments both the retrieval query and the re-ranking step, boosting the relevance score of document chunks whose geographic and institutional scope aligns with the entities identified in the query.

3.3 Hybrid RAG

The Hybrid RAG extends conventional retrieval typically adopted in RAG systems by incorporating entity-aware augmentation derived from both document metadata and named entities extracted from the user query. This approach follows recent strategies that integrate structured attributes into retrieval pipelines (Dadopoulos et al., 2025; Wang et al., 2025), combining dense vector search with sparse keyword matching to improve retrieval robustness for geographically specific queries.

The proposed pipeline is composed of three main stages:

[2] https://cnes.datasus.gov.br/

- **Index construction:** Each document is segmented into overlapping textual units and enriched with structured metadata, including the geographic scope of the article (i.e., the administrative regions mentioned) and the types of health facilities referenced. These enriched representations are encoded using a sentence embedding model and stored in a dense vector index. Simultaneously, a sparse BM25 index is built over the same chunk corpus, enabling keyword-level matching as a complement to semantic similarity. As a result, chunks associated with specific locations or facility types become retrievable through both semantic and lexical signals.

- **Query augmentation:** At retrieval time, the user query is processed by the NER module (Section 3.2). The extracted entities are appended to the query terms used for BM25 retrieval (Robertson et al., 2004), expanding the lexical representation of the query with domain-specific vocabulary. This augmentation increases the likelihood of retrieving chunks that share geographic or institutional scope with the query, even when the phrasing differs from the document text.

- **Geographic-aware re-ranking:** Rather than selecting the top-K candidates based only on semantic similarity, the method applies a re-ranking step that integrates semantic, lexical, geographic, and entity-matching signals. For a given query q and document segment d, let $\hat{s}_{\text{dense}}(d, q)$ and $\hat{s}_{\text{bm25}}(d, q)$ denote the min-max normalized dense and BM25 scores, let $g(d, q) \in \{0, 1\}$ denote a binary geographic matching score computed by normalized substring matching between the query location and the neighborhoods and administrative regions retained during corpus filtering (Section 3.1), and let $\text{ner}(d, q) \in [0, 1]$ denote the fraction of epidemiological entities extracted from the query that appear in d. If no location is detected in the query, $g(d, q) = 0$ for all candidates. The final score is computed as:

$$\begin{aligned} \text{score}(d, q) &= W_{\text{dense}} \cdot \hat{s}_{\text{dense}}(d, q) \\ &+ W_{\text{bm25}} \cdot \hat{s}_{\text{bm25}}(d, q) \\ &+ W_{\text{geo}} \cdot g(d, q) \\ &+ W_{\text{ner}} \cdot \text{ner}(d, q) \end{aligned} \tag{1}$$

  The weights W_{dense}, W_{bm25}, W_{geo}, and W_{ner} balance the contribution of each signal in the final score.

3.4 Question Answer Corpora Construction

To evaluate the proposed RAG pipeline in a domain-specific and location-aware setting, we constructed a synthetic Question Answer (QA) corpus grounded in the public health news knowledge base $\mathcal{B}$. While questions written by humans/experts remain the gold standard, producing them is a time-consuming task that requires extensive review of large collections of news articles to ensure proper coverage (Schmidt et al., 2024). As a result, generating synthetic test sets has become a widely adopted alternative (Ushio et al., 2023; Filice et al., 2025), in which an LLM produces questions directly from the reference documents.

With the support of a public health expert, we designed a prompt to simulate the style of questions commonly asked by target users. QA pairs were generated from the public health news corpus using Claude Haiku 4.5 (`claude-haiku-4-5-20251001`), prompted in Portuguese to produce factual, interpretive, community-impact, and service-identification questions. Each question was required to explicitly contain at least one location name mentioned in the source article (e.g., "Ceilândia"), and pairs that did not satisfy this constraint were discarded during post-processing. The locality check was performed by verifying, via case-insensitive substring matching, whether the generated question contained at least one of the location names associated with its source article. In total, 933 QA pairs were initially generated, of which 283 were discarded for failing the locality check (30.3% discard rate). The pairs were then shuffled and split into a validation set $\mathcal{V}$ of 150 samples, used to refine prompts and the agent, and a final evaluation set $\mathcal{T}$ of 500 samples. The size of this corpus is comparable to other synthetic QA datasets for domain-specific tasks, where related works have used between 100 (van Elburg et al., 2026) and 400 (Thulke et al., 2025) samples for evaluation.

4 Preliminary Experiments

We conducted preliminary experiments to evaluate the proposed RAG strategy integrated into the agent using the public health news corpus. All experiments were performed on an Intel Xeon Gold 5220R CPU and an NVIDIA Tesla V100S GPU

with 32 GB of VRAM, running a 64-bit Linux operating system. The proposed pipeline was implemented in Python 3.10 using the latest versions of Transformers 4.60, Scikit-learn 1.8.0, LangChain 1.2.17, Ollama 0.6.2, and Pandas 3.0.2. The source code, the prompts used in all phases, and the corpus are publicly available in our repository[3].

Two open-source LLMs were evaluated in combination with the RAG approaches: Qwen3-8B (Yang et al., 2025) and Llama3.1-8B (Grattafiori et al., 2024). These models were selected because they can run locally, avoiding inference costs and allowing private deployments within public health institutions.

The quality of the generated answers was evaluated using both automated similarity metrics and LLM-based assessments. For each question in the QA corpus, the generated response was compared with its corresponding reference answer through standard reference-based evaluation metrics widely used in the literature. We considered METEOR, ROUGE, and BLEU, which measure lexical overlap between generated and reference answers (Banerjee and Lavie, 2005). In addition, semantic evaluation metrics were also employed, namely BERTScore-F1 (Zhang et al., 2020) and BLEURT-20 (Sellam et al., 2020), where higher scores indicate greater semantic similarity to the reference answer.

We also incorporate an LLM-as-a-judge evaluation to assess aspects that are not fully captured by surface-level similarity. For this purpose, we adopt the RAGAS framework (Es et al., 2024) and consider the following metrics, all ranging from 0 to 1, where higher values indicate better performance. Answer Relevance (Rel) measures how well the generated response addresses the user query. Faithfulness (Faith) evaluates the degree to which the generated answer is supported by the retrieved context. Answer Correctness (Corr) assesses the agreement between the generated and reference answers. Context Precision measures the relevance of the retrieved context to the query, while Context Recall evaluates how well the retrieved context covers the information required to answer the query. To avoid evaluation bias, the LLM judges employed in this stage are different from those used to generate answers in the RAG pipelines.

LLM-only and Vanilla RAG approaches were implemented as baselines, in which the RAG settings share a common set of base hyperparameters with the Hybrid RAG. Such hyperparameter values were determined by manually inspecting retrieval results and generated responses on the validation set $\mathcal{V}$. Documents were split into chunks of $1,000$ characters with an overlap of 200 characters, using a recursive character text splitter that prioritizes paragraph boundaries as natural language separators. The retrieval step returns the top-7 most relevant chunks for each query. All document chunks and queries were encoded using the `alfaneo/bertimbau-base-portuguese-sts` sentence embedding model (Souza et al., 2020), suitable for Portuguese texts, and stored in a FAISS vector database.

For the Hybrid RAG, a pool of 100 candidate chunks is first retrieved by combining dense, BM25, and geographic signals, which is then reranked using LLM-extracted entity overlap before truncating to the top-7 chunks passed to the generative model. To isolate the contribution of the geographic and NER-based reranking signals from the effect of expanding the candidate pool, we introduced Vanilla-BM25 as a controlled ablation baseline, which retrieves from the same pool of 100 candidates using only dense and BM25 signals, without geographic or entity-based reranking. For the Hybrid RAG, the combination weights were set empirically to $W_{\text{dense}} = 0.5$, $W_{\text{bm25}} = 0.25$, $W_{\text{geo}} = 0.1$, and $W_{\text{ner}} = 0.15$ by manually inspecting retrieval results on the validation set $\mathcal{V}$.

Table 1 presents the performance of the RAG approaches according to the classic evaluation metrics. LLM-only produced the weakest results across all metrics, confirming that the LLMs alone cannot generate accurate answers for this domain without retrieved context. The results demonstrate that adding BM25 to the retrieval pipeline further improved performance compared to LLM-only and Vanilla RAG, with Qwen3-8B achieving the best BLEU and ROUGE scores. The Hybrid RAG, which further incorporates geographic and NER-based reranking signals, achieved the best METEOR (0.3718 with Llama3.1-8B) and BLEURT (0.3146) and BERTScore-F1 (0.7429) with Qwen3-8B. However, the improvements in semantic similarity to the reference responses achieved by territory-aware reranking are marginal compared to those obtained with Vanilla-BM25.

For the RAGAS framework metrics, we used GPT-4o-mini and DeepSeek as judges, allowing

[3] https://gitlab.com/gvic-unb/ijcai-ecai-fsg-2026

RAG	LLM	BLEU	ROUGE	METEOR	BLEURT	BERTScore-F1
LLM-only	Llama3.1-8B	3.8603	0.1397	0.1430	0.1416	0.6685
	Qwen3-8B	4.3566	0.1310	0.0999	0.1311	0.6635
Vanilla	Llama3.1-8B	8.0172	0.2319	0.3108	0.2681	0.6995
	Qwen3-8B	11.1084	0.2723	0.3035	0.2783	0.7287
Vanilla-BM25	Llama3.1-8B	9.9466	0.2678	0.3669	0.3028	0.7169
	Qwen3-8B	**13.1023**	**0.2986**	0.3416	0.3130	0.7398
Hybrid	Llama3.1-8B	10.5764	0.2701	**0.3718**	0.3048	0.7180
	Qwen3-8B	12.9183	0.2976	0.3400	**0.3146**	**0.7429**

Table 1: Lexical and semantic metrics for the evaluated RAG approaches combined with different LLMs. Best overall values for each metric are shown in bold.

Judge	RAG	LLM	Rel	Corr	Faith	Recall	Precision
DeepSeek	LLM-only	Llama3.1-8B	0.1642	0.2335	0.1905	-	-
		Qwen3-8B	0.1709	0.2285	0.0803	-	-
	Vanilla	Llama3.1-8B	0.6238	0.5028	0.7651	0.5211	0.3782
		Qwen3-8B	0.6820	0.5256	0.7881	0.5214	0.3847
	Vanilla-BM25	Llama3.1-8B	0.7348	0.5919	0.8047	0.7229	0.6084
		Qwen3-8B	0.7798	0.5964	0.8139	0.7198	0.6018
	Hybrid	Llama3.1-8B	0.7245	0.5915	**0.8148**	**0.7384**	0.6064
		Qwen3-8B	**0.8074**	**0.6083**	0.8087	0.7382	**0.6202**
GPT-4o-mini	LLM-only	Llama3.1-8B	0.1576	0.2795	0.1760	-	-
		Qwen3-8B	0.1562	0.2442	0.1336	-	-
	Vanilla	Llama3.1-8B	0.5790	0.4898	0.7787	0.6135	0.6134
		Qwen3-8B	0.6374	0.5253	0.7953	0.6120	0.6055
	Vanilla-BM25	Llama3.1-8B	0.6782	0.5501	**0.8458**	0.7540	0.8040
		Qwen3-8B	0.7450	0.5811	0.8328	0.7499	0.8072
	Hybrid	Llama3.1-8B	0.6805	0.5513	0.8380	**0.7784**	**0.8193**
		Qwen3-8B	**0.7694**	**0.5936**	0.8157	0.7678	0.8180

Table 2: RAGAS metrics per judge for each combination of RAG approach and LLM architecture. Best values are highlighted separately for each judge. Recall and Precision are undefined for the LLM-only setting because no retrieval step is involved.

us to compare RAG strategy performance for two different types of evaluators. Both judges were accessed via their respective APIs. The results of RAGAS metrics are shown in Table 2. The LLM-only setting produced the lowest scores for all metrics and in both judges.

It can also be seen that Vanilla-BM25 outperformed Vanilla RAG by a large margin across all RAGAS metrics. Hybrid RAG further improved over Vanilla-BM25 in answer relevance and correctness with both judges, with Qwen3-8B achieving the highest relevance scores overall, obtaining 0.8074 with DeepSeek and 0.7694 with GPT-4o-mini. When evaluated by GPT-4o-mini, Hybrid RAG with Llama3.1-8B achieved the highest context recall (0.7784) and precision (0.8193), suggesting that geographic and NER-based reranking slightly improves the alignment between retrieved context and user queries.

However, Vanilla-BM25 outperformed Hybrid RAG in faithfulness with both judges, particularly with Llama3.1-8B, achieving 0.8458 against 0.8380 with GPT-4o-mini, indicating that the reranking signals occasionally promote contextually relevant but less faithful chunks. Across both judges and models, the results consistently show that Hybrid RAG leads in relevance and correctness, while the differences from Vanilla-BM25 remain modest, suggesting that BM25 retrieval accounts for the majority of the gains over Vanilla RAG in this domain.

Regarding the judges, DeepSeek commonly assigned higher relevance and lower faithfulness scores compared to GPT-4o-mini when assessing

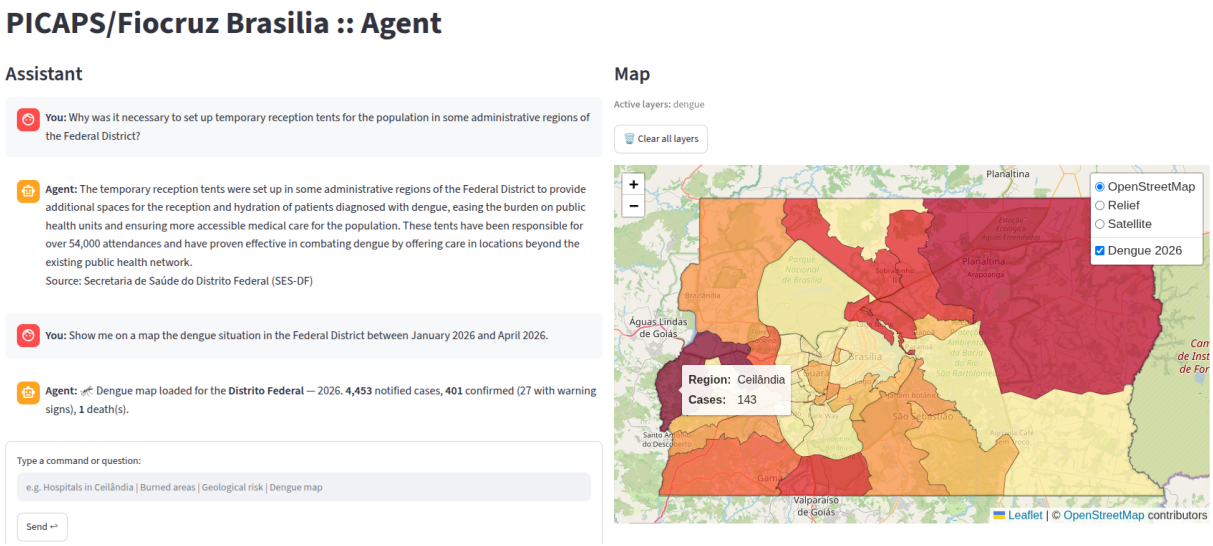

Figure 2: Use case illustrating user interaction with the conversational LLM, including map visualization and access to tabular epidemiological data related to dengue in Brasilia, Distrito Federal.

response quality. GPT-4o-mini tends to be more strict when assessing faithfulness and precision, in which DeepSeek only showed to be more demanding in relevance. The analysis over the generative models led us to verify that Qwen3-8B outperformed Llama3.1-8B in relevance and correctness across all RAG configurations and both judges, while Llama3.1-8B showed modest advantages in faithfulness and recall in some configurations. These results suggest that Qwen3-8B is better suited for news article texts in the public health domain, likely due to stronger alignment with the writing style of the corpus, which combines journalistic language with administrative and epidemiological terminology in Portuguese.

4.1 Use Case

To validate the proposed agent, we developed a prototype graphical user interface using the proposed Hybrid RAG with Qwen3-8B as the underlying agent. Figure 2 illustrates the interface, which is composed of an interactive map and a conversational panel through which users can query the agent.

In the first interaction, we provided the agent with a question from the validation set, asking about the reasons behind the installation of temporary reception tents across administrative regions of the Federal District. The agent retrieves and summarizes information from a relevant news article, reporting that the objective of such action was to ease the burden on public health units and ensure more accessible medical care for the population.

In the second interaction, the user instructs the agent to display the dengue situation in the Federal District from January 2026 to April 2026. The agent renders a heatmap layer showing the spatial distribution of dengue cases by administrative region. These interactions demonstrate the agent's ability to integrate structured epidemiological data with geospatial visualization, delivering natural language responses that support public health surveillance tasks.

4.2 Expert Feedback

To provide initial qualitative evidence of the agent's usability, two public health researchers interacted with the prototype in a structured 10-minute session. The participants had 7 and 2 years of experience in public health, respectively. Each participant first evaluated five queries from the validation set $\mathcal{V}$ using a five-point Likert agreement scale, with the original news articles available for reference. They then formulated three free-form queries and rated the response quality using the same scale. Finally, participants completed two tasks associated with the map visualization involving dengue cases and UBS location queries.

The assessment revealed strong agreement with the agent's responses to the synthetic news-based queries. For free-form public health queries, participants reported satisfactory agreement, though one participant noted incomplete information regarding dengue cases in specific administrative regions. On the map-based tasks, both participants correctly identified the region with the highest dengue con-

centration and the UBS locations rendered after querying the agent.

Participants also provided constructive feedback on the prototype's limitations. Both highlighted the agent's ability to provide concise responses and recognized the usefulness of the map-chatbot integration. They also reported limitations concerning the visibility of small geographic areas on the map, and incomplete responses for hospital specialty queries and aggregated epidemiological queries. These results suggest that the natural language interface enables experts to access territorial health information in a centralized and visually supported manner and can support territorial health queries in practice.

5 Limitations

Despite the promising results, this ongoing research has some limitations. First, the absence of prior work on question answering over public health news with location-aware queries makes it difficult to establish a consolidated baseline. Thus, our results serve as an initial reference for the field.

Second, the synthetic QA corpus was generated using a proprietary LLM, which introduces potential biases in the types and phrasings of questions produced. Related to this, the quality of the corpus is sensitive to prompt design, which may result in questions that do not fully reflect the linguistic diversity and domain specificity of real-world user queries in public health contexts.

Third, the evaluation pipeline employs synthetic QA pairs grounded in the same news sources used by the RAG pipeline. In practice, users may ask questions that go beyond the scope of the news article knowledge base, which can lead to hallucinations and imprecise responses. Although we varied question styles across four types and employed distinct LLMs for generation, retrieval, and evaluation, expert-authored ground truth remains absent and is left for future work.

Finally, the qualitative evaluation was conducted with the participation of only two public health researchers, which is insufficient for drawing strong conclusions about usability for the broader target population, including community health agents and citizens. A large-scale user study is planned as future work.

6 Conclusions

This paper introduced a conversational geospatial intelligence system to support public health surveillance in the Federal District, Brazil. The proposed agent interprets natural language queries, retrieves information using hybrid RAG over public health news articles, combining dense retrieval, BM25, geographic scoring, and LLM-based entity re-ranking, and returns contextualized responses alongside interactive map visualizations.

Preliminary experiments showed that expanding the candidate pool with BM25 retrieval yields substantial gains over Vanilla RAG, while the proposed Hybrid RAG further improves answer relevance and correctness, as consistently observed across both LLM-judges. Moreover, a preliminary qualitative evaluation with two public health researchers using a prototype implementing the agent provided initial evidence of the system's usefulness and usability. The proposed system plays an important role in bridging the gap between fragmented public health data sources and communities, researchers, and healthcare professionals.

Several directions are planned for future work. First, we intend to enrich the UBS knowledge base with service information from two sources: patient reviews from Google Reviews and open data from government transparency portals, covering aspects such as available specialties, opening hours, and waiting times. Second, we plan to improve the agent by enabling it to route queries to different language models depending on the type of document or task. For instance, using a model specialized in epidemiological data for surveillance questions and another for tabular data.

Acknowledgments

The authors would like to thank *Fundação de Apoio à Pesquisa do Distrito Federal (FAPDF)* - Grant #00193-00001398/2024-21 for funding this research. The authors also acknowledge the computational resources provided by Advanced Micro Devices, Inc. through the AMD University Program's AI & HPC Cluster, which contributed to the development of this work.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Pro-*

ceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72.

Dany Alexandra Morales Botero, Rafaele Bonfim, Karina Fonseca, Rubia Laine de Paula Andrade-Gonçalves, Aline Aparecida Monroe, and Fátima Morales. 2025. Artificial intelligence in brazilian primary health care: scoping review. *Revista Brasileira de Enfermagem*, 78(suppl 3):e20240363.

Tao Chen, Shengfa Miao, Shuangfeng Cai, Houcheng Liu, Yongkang Mu, Kaiming Yu, Hualong Deng, Xin Jin, Qian Jiang, Hua Jiang, Puming Wang, Dinan Ma, and Ahmed Zahir. 2026. GeoRAG: A geographic retrieval augmented generation framework based on urban spatio-temporal knowledge graph. In *Computational Linguistics and Natural Language Processing*, pages 130–140, Singapore. Springer Nature Singapore.

Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jingheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S. Yu, and Qingsong Wen. 2025. LLM agents for education: Advances and applications. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13782–13810, Suzhou, China. Association for Computational Linguistics.

Michail Dadopoulos, Anestis C. Ladas, Stratos Moschidis, and Ioannis C. Negkakis. 2025. Metadata-driven retrieval-augmented generation for financial question answering. *ArXiv*, abs/2510.24402.

Gustavo Alves de Melo, Maria Gabriela Mendonça Peixoto, Maria Cristina Angélico Mendonça, Marcel Andreotti Musetti, André Luiz Marques Serrano, and Lucas Oliveira Gomes Ferreira. 2024. Performance measurement of brazilian federal university hospitals: an overview of the public health care services through principal component analysis. *Journal of Health Organization and Management*, 38(3):351–371.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 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, St. Julians, Malta. Association for Computational Linguistics.

Simone Filice, Guy Horowitz, David Carmel, Zohar Karnin, Liane Lewin-Eytan, and Yoelle Maarek. 2025. Generating Q&A benchmarks for RAG evaluation in enterprise settings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 469–484, Vienna, Austria. Association for Computational Linguistics.

Fernanda Färber, Julia Dollis, Pedro Ribeiro, Iago Brito, Rafael Sousa, and Arlindo Galvão Filho. 2025. Iaramed: A women's healthcare chatbot for portuguese speakers. In *Anais do XXV Simpósio Brasileiro de Computação Aplicada à Saúde*, pages 931–942, Porto Alegre, RS, Brasil. SBC.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kin On Kwok, Tom Huynh, Wan In Wei, Samuel YS Wong, Steven Riley, and Arthur Tang. 2024. Utilizing large language models in infectious disease transmission modelling for public health preparedness. *Computational and Structural Biotechnology Journal*, 23:3254–3257.

Jonathan Lin, Aditya Joshi, Hye-young Paik, Tri Dung Doung, and Deepti Gurdasani. 2025. RACCOON: A retrieval-augmented generation approach for location coordinate capture from news articles. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1123–1127.

Wagner de Jesus Martins, Edward Torres Maia, Antonia Sheila Gomes Lima, Osvaldo Peralta Bonetti, Lilian Silva Gonçalves, João Paulo Almeida Brito da Silva, Daniel Fernandes Barbosa, and Mauro Niskier Sanchez. 2024. Plataforma de inteligência cooperativa com a atenção primária à saúde (picaps): soluções tecnocientíficas em saúde digital no enfrentamento da COVID-19 e outras crises. *Ciência & Saúde Coletiva*, 29(10):e01622023.

Luyao Niu, Zhicheng Deng, Boyang Li, Nuoxian Huang, Ruiqi Liu, and Wenjia Zhang. 2025. AskNearby: An LLM-based application for neighborhood information retrieval and personalized cognitive-map recommendations. In *Proceedings of the 8th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, GeoAI '25, pages 22–32, New York, NY, USA. Association for Computing Machinery.

Aske Plaat, Max van Duijn, Niki Van Stein, Mike Preuss, Peter van der Putten, and Kees Joost Batenburg. 2026. Agentic large language models, a survey. *Journal of Artificial Intelligence Research*, 84.

Stephen Robertson, Hugo Zaragoza, and Michael Taylor. 2004. Simple BM25 extension to multiple weighted fields. In *Proceedings of the thirteenth ACM international conference on Information and Knowledge Management*, pages 42–49.

Maximilian Schmidt, Andrea Bartezzaghi, and Ngoc Thang Vu. 2024. Prompting-based synthetic data generation for few-shot question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13168–13178, Torino, Italia. ELRA and ICCL.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

David Thulke, Jakob Kemmler, Christian Dugast, and Hermann Ney. 2025. Listen to the context: Towards faithful large language models for retrieval augmented generation on climate questions. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 245–259, Vienna, Austria. Association for Computational Linguistics.

Asahi Ushio, Fernando Alva-Manchego, and Jose Camacho-Collados. 2023. An empirical comparison of LM-based question and answer generation methods. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14262–14272, Toronto, Canada. Association for Computational Linguistics.

Simon T van Baal, Suong TT Le, Farhad Fatehi, Antonio Verdejo-Garcia, and Jakob Hohwy. 2024. Testing behaviour change with an artificial intelligence chatbot in a randomized controlled study. *Journal of Public Health Policy*, 45(3):506–522.

Jonas van Elburg, Peter van der Putten, and Maarten Marx. 2026. Can we evaluate RAGs with synthetic data? In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 177–192, Cham. Springer Nature Switzerland.

Shubham Vatsal, Harsh Dubey, and Aditi Singh. 2026. Agentic AI in healthcare and medicine: A seven-dimensional taxonomy for empirical evaluation of LLM-based agents. *IEEE Access*, 14:4840–4863.

Larissa Veríssimo, Helen Rainey, Roberta Lindemann, and Anne Hendry. 2024. Are community health agents the link to integrating care? Lesson from Brazil. *Journal of Integrated Care*, 32(1):45–51.

Zihan Wang, Zihan Liang, Zhou Shao, Yufei Ma, Huangyu Dai, Ben Chen, Lingtao Mao, Chenyi Lei, Yuqing Ding, and Han Li. 2025. InfoGain-RAG: Boosting retrieval-augmented generation through document information gain-based reranking and filtering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7190–7204, Suzhou, China. Association for Computational Linguistics.

Shan Xu, Zhaokun Yan, Chengxiao Dai, and Fan Wu. 2025. MEGA-RAG: a retrieval-augmented generation framework with multi-evidence guided answer refinement for mitigating hallucinations of LLMs in public health. *Frontiers in Public Health*, 13:1635381.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.