

LLM-Enhanced Cloud PBX-IP for Real-Time Emergency Call Triage: A Preliminary Evaluation

Matheus Vasconcelos Silveira, Rodrigo Mesquita de Vasconcelos and Walter da Cruz Freitas Júnior

Abstract—Response time in emergency services is critical for public safety and life preservation. Traditional human-based triage systems may suffer from latency, scalability limitations, and susceptibility to human error. This work proposes and evaluates a cloud-based Private Branch Exchange over IP (PBX-IP) system integrated with Large Language Models (LLMs) for automated emergency call triage. The architecture combines Voice over IP (VoIP), Speech-to-Text (STT), and Retrieval-Augmented Generation (RAG) to enable real-time classification and routing of calls. Experimental results show an accuracy of 83.40% and an average response time of 2.4 seconds, demonstrating the feasibility of the proposed approach.

Keywords—LLM, VoIP, PBX-IP, RAG, Emergency Systems.

I. INTRODUCTION

Modern communication systems are increasingly incorporating artificial intelligence to enhance automation, scalability, and decision-making capabilities. Emerging paradigms such as fifth- and sixth-generation mobile networks (5G/6G) are inherently driven by the integration of artificial intelligence, enabling intelligent, adaptive, and low-latency processing mechanisms required to support real-time services and mission-critical applications [1], [2].

Emergency call systems are a fundamental component of public safety infrastructure, where response time plays a decisive role in life preservation and risk mitigation. In traditional systems, the triage process is performed manually by human operators, which may introduce delays, inconsistencies, and errors, especially in high-demand scenarios or when dealing with distressed or non-cooperative callers. These limitations highlight the need for intelligent and automated approaches capable of assisting or partially replacing human decision-making in critical situations.

In emergency scenarios, communication is often affected by stress, incomplete information, and informal language usage. These factors make it difficult for traditional rule-based systems to correctly interpret user intent, increasing the likelihood of misclassification or delayed response. Additionally, real-world emergency calls frequently involve ambiguous or coded language, requiring contextual interpretation beyond simple keyword matching. This reinforces the need for intelligent systems capable of semantic understanding and adaptive learning.

Recent advances in Large Language Models (LLMs) have significantly improved machine capabilities in natural language understanding, enabling applications such as intent

recognition, classification, and contextual reasoning [3]. These models, based on the Transformer architecture, are capable of processing sequential data with high contextual awareness, making them suitable for interpreting complex and ambiguous human language [6].

In parallel, Voice over IP (VoIP) technologies have become the backbone of modern telecommunication systems, enabling flexible, scalable, and cost-efficient communication infrastructures [4]. Systems based on Private Branch Exchange over IP (PBX-IP), such as Asterisk, allow programmable call routing and seamless integration with external services, which makes them suitable platforms for deploying intelligent communication solutions [7].

Furthermore, Retrieval-Augmented Generation (RAG) has emerged as a promising approach to enhance LLM performance by incorporating external knowledge sources during inference. This technique improves contextual accuracy and reduces hallucinations, which is particularly relevant in safety-critical applications [5].

A. Related Work

The integration of artificial intelligence into communication systems has been widely investigated, particularly in the context of next-generation networks. AI-driven approaches have been proposed to optimize network management, automate decision-making, and improve the efficiency of communication services in 5G and 6G environments [1], [2].

In the telecommunication domain, VoIP-based systems have been extensively used for call management and routing due to their flexibility and programmability [4], [7]. Traditional approaches for call triage and routing are typically based on predefined rules or keyword matching, which limits their ability to handle unstructured, ambiguous, or context-dependent human language.

With the advent of LLMs, several works have explored their application in tasks such as text classification, conversational systems, and intent detection, demonstrating strong performance in interpreting informal and incomplete inputs [3], [6]. More recently, RAG-based approaches have been introduced to enhance LLM capabilities by combining parametric and non-parametric knowledge, enabling better generalization and adaptability [5].

However, despite these advances, the application of LLMs in real-time telecommunication systems is still an emerging field, where existing solutions are mostly restricted to text-based or offline scenarios. Recently, a study by Venkateshperumal et al. (2025) [8] demonstrated that integrating LLMs

with RAG can significantly improve the reconstruction of fragmented speech, allowing for efficient prioritization of emergency calls under severe VoIP network degradation. However, this approach depends on independent indexers (e.g., FAISS) and resource-intensive cloud infrastructures. While the more direct integration of LLM-based triage into lightweight and low-cost PBX-IP architectures capable of meeting real-time voice processing requirements, such as low latency, remains unexplored, especially in developing regions.

In this context, this work proposes a system that integrates LLMs with a cloud-based PBX-IP infrastructure to perform automated triage of emergency calls in real time. The proposed architecture combines VoIP communication, Speech-to-Text (STT) processing, LLM-based semantic analysis, and Retrieval-Augmented Generation (RAG) to improve classification accuracy and contextual understanding.

The main contributions of this work are:

- Integration of VoIP systems with LLM-based semantic processing for the real-time triage of emergency calls;
- Implementation of an adaptive learning mechanism via RAG, enabling contextual enhancement and error correction through a feedback loop;
- Development of a highly resource efficient and low cost architecture based on open-source tools (e.g., Linux, Asterisk, PostgreSQL) capable of running on limited cloud environments;
- Experimental validation using simulated emergency scenarios, demonstrating the system's capability to act with high accuracy and low latency.

II. SYSTEM DEVELOPMENT

A. Architecture Overview

To address the challenges of real-time emergency service triage, we developed a modular pipeline that receives, transcribes, and semantically routes a phone call. The proposed system integrates the following key components:

- Asterisk PBX-IP for call management;
- Speech-to-Text (STT) for audio transcription;
- LLM for semantic classification;
- PostgreSQL database with vector-based retrieval.

VoIP systems convert voice signals into digital packets, enabling efficient transmission over IP networks [4]. In this architecture, the system operates as a processing pipeline in which incoming VoIP calls are first handled by the PBX-IP system and then forwarded to the STT module for real-time transcription. The transcribed text is subsequently analyzed by the LLM, which performs semantic classification enhanced by RAG. Finally, the system routes the call to the appropriate emergency service.

The Speech-to-Text (STT) module plays a critical role in the system, as it converts raw audio signals into textual representations that are subsequently processed by the LLM. The quality of this transcription directly impacts classification performance, since errors in speech recognition may propagate to the semantic analysis stage.

In real-world scenarios, STT systems must deal with noise, incomplete sentences, and variations in speech patterns, which

can degrade transcription accuracy and affect downstream processing.

This modular design allows each component to be independently optimized, improving scalability and enabling future deployment of local models.

Fig. 1 illustrates the overall architecture of the proposed system and the interaction between its main components.

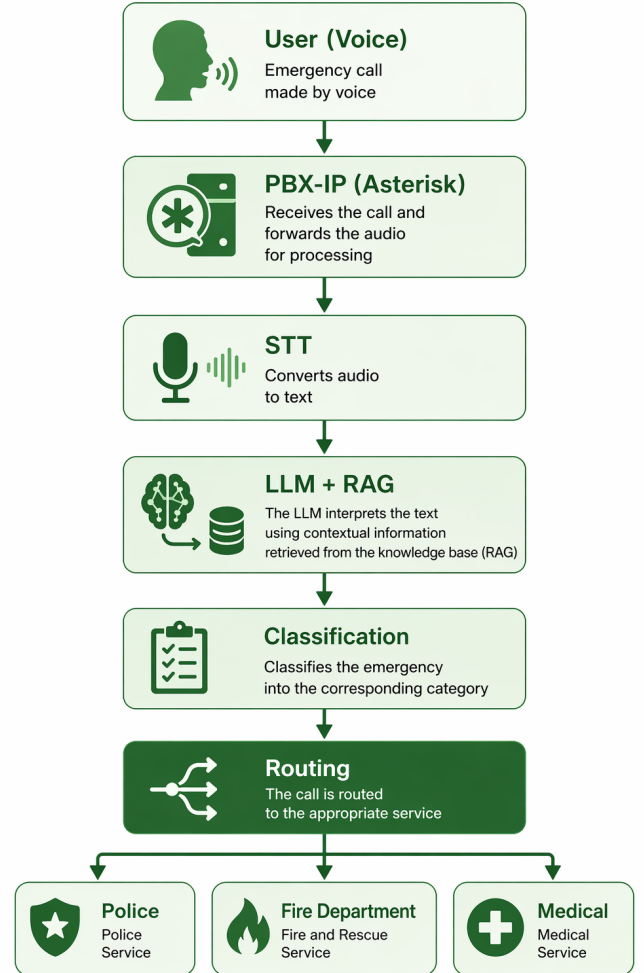


Fig. 1. System architecture for real-time emergency call triage.

B. Implementation Details and Reproducibility

To ensure reproducibility and meet real-time processing constraints, the proposed architecture was built using a combination of open-source PBX infrastructure and cloud-based APIs. The VoIP call ingestion is managed by Asterisk PBX-IP. To extract and process the live audio stream, we utilized the Asterisk External Application Protocol (AEAP). A custom middleware was developed in Node.js (adapted and configured from Asterisk's official AEAP implementation scripts) to bridge the PBX-IP and the external cognitive services.

The Node.js middleware first streams the audio to the Google Speech-to-Text (STT) API for real-time transcription. Once transcribed, the text is routed to the semantic classification engine. For this core task, we employed OpenAI's

gpt-4o-mini model via API. This specific LLM was selected due to its optimal balance between fast inference speeds and cost-effectiveness, which is paramount for high-volume, low-latency triage scenarios [3].

To mitigate model hallucinations and ground the classification in domain-specific operational guidelines, the Retrieval-Augmented Generation (RAG) pipeline was implemented natively using a PostgreSQL database equipped with the pgvector extension [5]. During inference, the Node.js backend generates embeddings for the incoming transcription and performs a cosine similarity search within PostgreSQL to retrieve the Top-K most relevant historical emergency cases.

This retrieved context is dynamically injected into the gpt-4o-mini prompt. This architectural choice allows the system to seamlessly adapt to regional slang and coded language without the need for resource-intensive model fine-tuning. The final output is structured to include a summarized description of the emergency and the predicted service category (Police, Fire, Medical, or General).

III. RESULTS

To evaluate the proposed system, a dataset comprising 247 simulated emergency calls was compiled. This dataset was constructed using both real-world dispatcher interaction patterns and unscripted contributions from volunteer participants, who were given the open-ended instruction to ‘call this number and report an emergency.’ Conducted entirely in Portuguese, these calls were processed under controlled PBX-IP conditions to replicate diverse levels of semantic complexity and signal quality, thereby testing the system’s robustness against typical VoIP degradation.

This experimental design enables the evaluation of the system under different levels of semantic complexity and signal quality, providing a comprehensive assessment of its robustness in real-world scenarios.

The experiments were organized into three groups:

- Group A: Direct and explicit emergency calls;
- Group B: Context-dependent and ambiguous calls;
- Group C: Degraded or incomplete audio scenarios.

A. Group A: Direct Calls

Instead of relying solely on keyword identification, the integration with the LLM enabled a deeper semantic analysis. In semantically well-structured sentences with strong verbs and explicit intent, such as “there is a fire happening” or “I have just been robbed”, the system exhibited excellent performance in call routing, correctly directing them to the Fire Department and Police sectors, respectively.

Table I presents the results for Group A, consisting of well-structured sentences with clear semantic meaning.

B. Group B: Contextual and Ambiguous Calls

In this group, the use of RAG improved performance by enabling the model to leverage contextual similarity from previously observed cases. Without contextual augmentation, the LLM tends to misclassify indirect or non-explicit phrases.

TABLE I
RESULTS FOR GROUP A.

Spoken Phrase	Transcribed Phrase	Correct Sector	Predicted Sector
My house is on fire	My house is on fire	Fire Department	Fire Department
My mother is having a heart attack	My mother is having a heart attack	Emergency Medical Services	Emergency Medical Services
I see smoke coming out of a store	I see smoke coming out of a store	Fire Department	Fire Department
A serious accident on the road	A serious accident on the road	Emergency Medical Services	Emergency Medical Services
My baby swallowed a coin	My baby swallowed a coin	Emergency Medical Services	Emergency Medical Services
My husband hit me	My husband hit me	Police	Police

However, when supported by retrieved examples, the model is able to infer hidden intent and correctly interpret coded or ambiguous messages.

In practical scenarios, emergency calls may involve disguised or indirect communication patterns, in which users intentionally avoid explicit language. For example, individuals under distress may use coded phrases to request help without alerting a potential aggressor. In such cases, the system was able to correctly infer the underlying intent and classify the call appropriately.

Traditional keyword-based approaches, or LLMs without contextual augmentation, would likely fail to achieve this level of semantic abstraction, highlighting the critical role of RAG in enhancing semantic understanding in scenarios where explicit keywords are absent.

To further analyze system performance across different scenarios, Table II presents a comparison of precision, recall, and F1-score for each experimental group.

TABLE II
PERFORMANCE COMPARISON ACROSS EXPERIMENTAL GROUPS.

Group	Samples	Precision	Recall	F1-score
Group A	103	96.3%	96.1%	0.961
Group B	71	87.9%	70.4%	0.769
Group C	73	89.4%	78.0%	0.826
Total	247	88.8%	83.4%	0.848

The results indicate that Group A achieved the highest performance due to its structured and explicit nature. Group B presented lower recall due to semantic ambiguity, while Group C maintained moderate performance despite degraded inputs, highlighting the robustness of the approach.

C. Adaptive Learning via RAG

Table III presents an example of the model’s learning behavior.

TABLE III
LEARNING BEHAVIOR EXAMPLE.

Scenario	Transcribed Phrase	Correct Sector	Predicted Sector
First attempt	roof fell onto the car	Fire Department	General
Second attempt	I am trapped in the garage	Fire Department	Fire Department

After incorporating feedback into the RAG database, the system correctly classified similar future inputs, further demonstrating the adaptive nature of the approach.

This adaptive capability is enabled by a dynamic knowledge base that evolves through an automated feedback loop. After each call, user feedback is collected, and in cases of misclassification, the correct sector is incorporated into the vector database. This updated information is then retrieved in future queries through the RAG mechanism, allowing the system to progressively improve its performance without requiring retraining of the LLM.

Initially, due to the absence of sufficient contextual information, the system incorrectly classified the input. However, after incorporating feedback, subsequent similar inputs were correctly routed, demonstrating the effectiveness of the continuous learning mechanism.

To further evaluate the impact of the RAG module, a comparative analysis was conducted against a zero-shot LLM baseline (architecture without contextual augmentation), as shown in Table IV.

TABLE IV
PERFORMANCE COMPARISON WITH AND WITHOUT RAG.

Scenario	Precision	Recall	F1-score
Group B (without RAG)	98.1%	70.0%	0.798
Group B (with RAG)	89.4%	66.0%	0.738
Group C (without RAG)	93.5%	72.0%	0.784
Group C (with RAG)	93.2%	82.0%	0.861

The results indicate that the effectiveness of RAG depends on the quality and representativeness of the knowledge base. In Group B, the use of RAG led to a decrease in performance, likely due to the limited size and coverage of the dataset, which introduced irrelevant contextual information.

In contrast, for Group C, RAG improved recall and F1-score, suggesting that contextual augmentation is particularly beneficial in scenarios with incomplete or degraded input.

This behavior suggests that RAG performance is highly dependent on the quality and diversity of the underlying dataset. A limited knowledge base may introduce bias or incorrect associations during vector retrieval, negatively impacting LLM predictions, especially in scenarios involving slang or complex contextual expressions.

D. Group C: Degraded Audio

Table V presents the most challenging scenarios, involving incomplete or corrupted speech.

TABLE V
RESULTS FOR GROUP C.

Spoken Phrase	Transcribed Phrase	Correct Sector	Predicted Sector
Help... he is going to hit me	help he is going to hit me	Police	Police
Grandmother... face... paralyzed	face... paralyzed	Emergency Medical Services	Emergency Medical Services
Baby... swallowed... coin	baby swallowed coin	Emergency Medical Services	Emergency Medical Services
Pole... fell... fire	pole fell fire	Fire Department	Fire Department
Hole... car fell	hole car fell	Fire Department	General

The results show that the system remains functional under degraded conditions, but accuracy decreases when critical semantic information is lost. These cases highlight the dependency on STT quality.

This degradation highlights a key limitation of end-to-end systems that rely on sequential processing stages. Errors introduced at the STT level reduce the quality of the input provided to the LLM, limiting its ability to correctly infer semantic intent.

This observation reinforces the importance of improving robustness at the signal processing level, particularly in noisy or degraded communication environments.

E. Overall Performance

In addition to classification accuracy, system latency was evaluated to assess its suitability for real-time applications.

The STT module presented an average latency of approximately 300 ms, mainly influenced by Voice Activity Detection (VAD). The RAG component required approximately 320 ms for vector-based retrieval operations. The LLM processing time was the most significant, averaging 1800 ms due to API response time.

Overall, the total system latency was approximately 2420 ms, which is considered acceptable for real-time emergency triage scenarios, as it enables timely decision-making without introducing significant delays in call routing. This latency reflects the combined processing time of the STT, RAG, and LLM components, demonstrating the feasibility of integrating these technologies in practical communication systems. The overall system performance, including classification effectiveness across different scenarios, is summarized in Table VI, providing a comprehensive view of the model's behavior under varying conditions.

A sector-based analysis was also conducted, as shown in Table VII.

The results show that the General category presents significantly lower precision, indicating that it is frequently overused

TABLE VI
CONFUSION MATRIX

Actual \ Predicted	Police	Fire	Medical	General
Police	65	2	3	10
Fire	1	52	3	13
Medical	2	0	68	5
General	1	1	0	21

TABLE VII
PERFORMANCE COMPARISON BY SECTOR.

Sector	Samples	Precision	Recall	F1-score
Police	80	94.2%	81.2%	0.872
Fire	69	94.5%	75.3%	0.838
Medical	75	91.8%	90.6%	0.912
General	23	42.8%	91.3%	0.583

in ambiguous scenarios. In contrast, Police and Medical categories exhibit higher precision and recall due to clearer semantic patterns.

Out of the 247 calls, 41 were misclassified, resulting in an overall accuracy of 83.40%. Additionally, the average total processing delay of the proposed architecture was 2.4 seconds.

Overall, the results demonstrate that the system is more reliable in scenarios with clear semantic structure, while still maintaining acceptable performance in ambiguous cases through the use of RAG. However, robustness under degraded signal conditions remains a challenge for future work.

IV. CONCLUSIONS

This work demonstrated the feasibility of integrating LLMs with PBX-IP systems for real-time emergency call triage.

The results demonstrate that the integration of AI-driven models with telecommunication infrastructures can significantly enhance the efficiency and scalability of emergency response systems. Furthermore, combining VoIP infrastructure with AI-based language models enables efficient classification and routing of emergency calls, aligning with the trend of intelligent communication systems [1], [2].

The incorporation of RAG proved essential to improve contextual understanding and reduce classification errors, especially in ambiguous scenarios [5].

Despite promising results, limitations remain, including dependency on external APIs, infrastructure constraints, and sensitivity to degraded audio input.

Future work will focus on deploying quantized local models for both LLM and STT tasks to reduce reliance on external cloud APIs, minimize latency, and resolve triage bottlenecks. Furthermore, edge processing will ensure complete privacy and security for citizens' data, in strict compliance with the Brazilian General Data Protection Law (LGPD). Additionally, efforts will include expanding the RAG knowledge base and validating the proposed architecture in real-world environments, enabling operational data collection and continuous refinement of the system.

The integration of AI with communication systems represents a key direction for next-generation intelligent networks.

REFERENCES

- [1] W. Jiang et al., "The road towards 6G: A comprehensive survey," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 334–366, 2021, doi: 10.1109/OJCOMS.2021.3057679.
- [2] B. Ji et al., "A survey of computational intelligence for 6G: Key technologies, applications and trends," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 10, pp. 7145–7154, Oct. 2021, doi: 10.1109/TII.2021.3062255.
- [3] W. X. Zhao et al., "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.
- [4] J. F. Kurose and K. W. Ross, *Computer Networking: A Top-Down Approach*, 6th ed. Boston, MA, USA: Pearson, 2013.
- [5] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [6] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017.
- [7] L. Madsen et al., *Asterisk: The Definitive Guide*, 4th ed. Sebastopol, CA, USA: O'Reilly Media, 2013.
- [8] D. Venkateshperumal, R. A. Rafi, S. Ahmed, and A. Khokhar, "LLM-Augmented VoIP for Emergency Call Triage," *Measurement: Digitalization*, vol. 2–3, p. 100010, Aug. 2025, doi: 10.1016/j.measdig.2025.100010.