Eva: An LLM-based Multilingual Voice-agent Network for Restaurant Operations

Demonstration Track

Zhiwei (Tony) Qin
foreva.ai (Eva)
San Jose, CA, USA
tonyqin@foreva.ai

Jianming Zhou
foreva.ai (Eva)
San Jose, CA, USA
jz@foreva.ai

ABSTRACT

Eva is a voice AI agentic system automating restaurant phone operations with individual agents for tasks like order placement and a global agent for multi-restaurant settings. It uses a hierarchical multi-agent architecture with various agent technologies, demonstrating LLM applications for improved efficiency and service in the restaurant industry.

KEYWORDS

LLM, voice AI, multi-agent, restaurant operations, multilingual

ACM Reference Format:
Zhiwei (Tony) Qin and Jianming Zhou. 2025. Eva: An LLM-based Multilingual Voice-agent Network for Restaurant Operations: Demonstration Track. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

The restaurant industry is facing a critical challenge of rising personnel cost and human labor shortage, leading to under-utilized business phones that many restaurants rely on getting phone orders and interacting with their customers. Nearly half of the US adult population call restaurants to order food and reserve tables. Therefore, there are out-sized needs for automating the restaurant phone operations. In response to that, we have built and deployed Eva (demo: https://youtu.be/UpL1qkC2dlE), an LLM-based multilingual voice AI agentic system for restaurants. Eva is designed to handle restaurants' inbound phone calls autonomously via open conversations. This short paper focuses on describing the architecture, technology, and functionality of Eva. In particular, agent evaluation [6] is not discussed here due to space constraint but is an important active research topic nevertheless.

Recent research has explored various applications of Large Language Models in customer-facing scenarios, e.g., customer support [1], supermarket robot [8], and task-driven conversational agents [10, 11]. Use cases in restaurant phone call scenarios present significant complexity in reasoning and planning, which current commercial foundation LLMs still fall short [5] and active research has been tackling [4].

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

2 GENERAL ARCHITECTURE

The architecture of a voice AI agentic system like Eva integrates three key components: Automatic Speech Recognition (ASR), Large Language Models (LLM), and Text-to-Speech (TTS) synthesis.

ASR utilizes deep learning models (e.g., Deepgram[2]) to transcribe spoken languages into text. This component serves as the initial interface between the caller and Eva.

The LLMs power the cognitive layer that houses the agents, processing the ASR output and generating contextually appropriate responses. These models are customized and fine-tuned for the restaurant domain, enabling them to align with the restaurant operational tasks. We focus on this cognitive layer in this paper, where we leverage a basket of foundation models from OpenAI and Anthropic to build the agents and agentic flows.

TTS transforms the LLM's text output into natural-sounding speech, employing neural network-based models such as those developed by ElevenLabs[3].

The architecture also includes a telephony layer and a dialogue management framework that coordinates conversation flow and interfaces with external systems like restaurant-side merchant Apps and point-of-sales systems.

3 STORE PHONE AGENTIC SYSTEM

The phone agentic system primarily consists of a group of task-driven agents, responsible for order placement, table reservation, catering request handling, and general inquiries. We focus on describing order placement, which has the highest complexity among all.

The *ordering agent* is responsible for carrying out a conversation with the customer to collect their order details, answer clarification questions, and notify the customer for any correction needed with respect to the constraints set by the menu. After the ordering agent completes the call, the *evaluation agent* examines the entire conversation to extract the required information for downstream components in the workflow. In particular, it generates the order in compliance to the menu, self reflects on the output, and makes any necessary corrections so as to achieve a high success rate of ordering. The store phone agentic system is illustrated in Figure 1.

3.1 Knowledge Representation and Retrieval

A key difference between the ordering agent and a general chatbot is that the agent is designed to complete a specific task, i.e., accurately collecting and placing the customer order through conversation, and that the agent possesses full knowledge about the specific menu and rules of the restaurant. Each menu has its own specificity

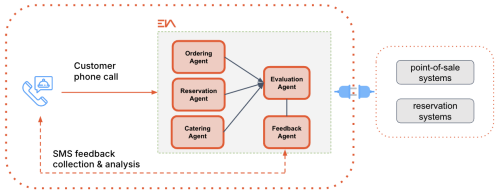

Figure 1: Store phone agentic system.

in the item names, modifier structure and constraints. We have developed a standard knowledge representation layer that presents the information optimally to the LLMs to understand. The output of this layer is stored in a vector database to facilitate a retrieval-augmented generation (RAG) pipeline [7].

3.2 Supervised Fine-tuning

In situations where a smaller and less capable model is required due to cost or speed considerations, supervised fine-tuning is an effective approach to reach a working between speed/cost and task quality. The key question here is how to generate high-quality training examples for fine-tuning the smaller models. In our case, these examples are in the form of conversations between user messages (prompts) and agent responses. Inspired by reinforcement learning and AlphaGo, we utilize self-play based on GPT4o to generate the desired training examples.

3.3 State Tracking and Planning

Fine-tuning is best for refining agent behavior, but to improve the in-conversation planning capability of the agent, we have designed and implemented a specialized real-time state tracking and planning module for our conversational agents. It enables the agents to effectively handle tasks that require collecting and processing information over multiple turns, such as table reservation and ordering food items that have compatibility constraints.

3.4 Multilingualism

Leveraging the multilingualim of the foundation LLMs, Eva is able to interact with restaurant customers in multiple languages of their choice. Our architecture enabling multilingualism closely resembles the mechanism observed in human beings. Only the conversational agents generate responses in a specific (non-English) language, while the evaluation agent and knowledge representation, which handle workflows, remain in English. This leads to a high-accuracy and lean framework scalable in the number of languages to support.

4 GLOBAL AGENT

Building on top of the store-based agentic systems described in the previous section, we have further designed and developed a *hierarchical* multi-agent system that allows a user to freely search and navigate among all restaurants, whether they have adopted Eva or not. A key differentiator of the *global agent* from a traditional yellowpage or telephone router is that the store agents of those restaurants already onboarded with Eva are native to the global agent, allowing customers to directly placing their orders (or reservations and any other store-related tasks) at those stores within the same phone call without the need of any call transfer.

4.1 Hierarchical Agents

In this hierarchical architecture, the global agent acts as the searcher and router, and the store agents described in the previous sections perform store-specific tasks. The message-passing and handoff mechanism of our system resembles that of Swarm [9], facilitating seamless task delegation among agents.

4.2 Local Business Search

Not every restaurant has adopted Eva. To make the global agent more helpful to the user, we have augmented the agent with web search capability for local businesses. This is realized through a function call that queries the Google Maps API for the geographical information specified by the user. Since those restaurants retrieved through web search have not adopted Eva yet, and hence, Eva is unable to directly place an order at those restaurants within the call. However, the agent is able to transfer the user's call to a restaurant of those on demand via the telephony layer, allowing the user to speak immediately to the personnel without having to make another call.

5 ORDER WORKFLOW

After the evaluation agent generates the final order, the information goes through a series of steps in a post-conversation workflow to reach the restaurant personnel as well as to close the customer feedback loop.

5.1 Confirmation and Payment

The order information is sent to the restaurant-side merchant App, which is essentially a lightweight menu management system allowing the restaurant personnel to update the menu and receive notification on incoming phone orders. The order total is computed and the order content is then either printed out or digitally sent to the kitchen to process.

The order content and total are used to generate the order confirmation that is sent to the customer via SMS along with a secure payment link. After the customer pays, the restaurant is further notified on the merchant App to proceed with cooking.

5.2 SMS Feedback

To close the feedback loop, we have developed an SMS feedback agent responsible for proactively collecting user feedback after both successful and aborted ordering phone calls. The feedback agent interacts with customers via SMS natural conversations, this agent is critical for surfacing production issues and guiding generation of more targeted fine-tuning examples.

6 DISCUSSIONS

Eva's deployment highlights the practical value of LLMs and AI agents in real-world customer-facing scenarios, particularly for restaurants. By efficiently managing tasks like orders and reservations, Eva enhances operational efficiency and has the potential to redefine service standards and customer expectations in the industry.

REFERENCES

[1] Biagio Antonelli and Gonzalo Cordova. 2024. Leveraging Large Language Models to build a low-resource customer support chatbot in a three-sided marketplace. *EasyChair Preprint* (2024).

[2] Deepgram. 2024. Deepgram (https://deepgram.com/product/speech-to-text). (2024). https://deepgram.com/product/speech-to-text

[3] ElevenLabs. 2024. ElevenLabs (https://elevenlabs.io/docs/speech-synthesis/models). (2024). https://elevenlabs.io/docs/speech-synthesis/models

[4] Atharva Gundawar, Mudit Verma, Lin Guan, Karthik Valmeekam, Siddhant Bhambri, and Subbarao Kambhampati. 2024. Robust Planning with LLM-Modulo Framework: Case Study in Travel Planning. *arXiv preprint arXiv:2405.20625* (2024).

[5] Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Kaya Stechly, Mudit Verma, Siddhant Bhambri, Lucas Saldyt, and Anil Murthy. 2024. LLMs Can't Plan, But Can Help Planning in LLM-Modulo Frameworks. *arXiv preprint arXiv:2402.01817* (2024).

[6] Sayash Kapoor, Benedikt Stroebl, Zachary S Siegel, Nitya Nadgir, and Arvind Narayanan. 2024. AI agents that matter. *arXiv preprint arXiv:2407.01502* (2024).

[7] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[8] Chandran Nandkumar and Luka Peternel. 2024. Enhancing Supermarket Robot Interaction: A Multi-Level LLM Conversational Interface for Handling Diverse Customer Intents. *arXiv preprint arXiv:2406.11047* (2024).

[9] OpenAI. 2024. Swarm: An educational framework exploring ergonomic, lightweight multi-agent orchestration. (2024). https://github.com/openai/swarm

[10] Jesús Sánchez Cuadrado, Sara Pérez-Soler, Esther Guerra, and Juan De Lara. 2024. Automating the Development of Task-oriented LLM-based Chatbots. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*. 1–10.

[11] Guangzhi Sun, Xiao Zhan, and Jose Such. 2024. Building Better AI Agents: A Provocation on the Utilisation of Persona in LLM-based Conversational Agents. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*. 1–6.