Call2Instruct: Automated Pipeline for Generating Q&A Datasets from Call Center Recordings for LLM Fine-Tuning

Alex Echeverria, Sávio Salvarino Teles de Oliveira, Fernando Marques Federson[1]

[1] Instituto de Informática – Universidade Federal de Goiás
CEP 74690-900 – Goiânia – GO – Brazil

{federson@ufg.br}

Abstract. *The adaptation of Large Language Models (LLMs) to specialized domains critically depends on high-quality instructional datasets. A significant bottleneck exists in generating Question-Answer (Q&A) datasets from noisy, unstructured sources such as call center audio recordings. This work presents Call2Instruct, a novel end-to-end automated pipeline that integrates five sequential modules: audio processing (diarization, noise suppression, ASR), text processing (normalization, anonymization), semantic extraction and vectorization, Q&A dataset generation using embedding-based similarity matching, and dataset validation through LLM fine-tuning.*

1. Introduction

Large Language Models (LLMs) have revolutionized artificial intelligence, generating substantial interest in fine-tuning techniques to adapt these models to specific tasks or domains. A critical component of successful fine-tuning is the availability of high-quality instructional datasets, typically in Question-Answer (Q&A) format. Despite advances in both general LLM fine-tuning techniques and isolated stages of call center data processing, a critical gap remains: the development of integrated, automated systems capable of efficiently transforming raw conversational audio into instructional datasets ready for supervised LLM training. Recent studies demonstrate the effectiveness of LLM in specific tasks, such as conversation summarization and insight extraction from transcripts via document retrieval. However, these approaches remain fragmented, addressing steps such as audio capture, transcription, or data adaptation in isolation, without offering end-to-end solutions. The ability to automatically transform audio recordings into structured Q&A pairs, ready for LLM fine-tuning, represents an area with significant room for contribution, which this paper addresses through the Call2Instruct pipeline.

2. Methodology

The methodology encompasses an operational flow structured into five sequential and interdependent data transformation stages: **Audio Acquisition and Preprocessing:** Collection of recorded calls, application of channel separation techniques (diarization), utilization of denoising algorithms, identification and removal of non-relevant segments, and Automatic Speech Recognition (ASR). **Textual Preprocessing and Cleaning:** Punctuation restoration and correction of transcription artifacts, removal of disfluencies (hesitations, repetitions), and spontaneous speech artifacts to enhance clarity; detection and anonymization of Personally Identifiable Information (PII), ensuring privacy compliance

with regulations such as GDPR and LGPD. **Semantic Information Extraction and Vectorization:** Automatic detection of text segments corresponding to the main customer requests (demands) and agent solutions (responses), rewriting the identified segments to ensure objectivity and clarity, and generating vector representations (embeddings) stored in a vector database for efficient semantic similarity-based searches. **Q&A Dataset Generation:** Semantic similarity search for each customer demand vector, pairing demands with retrieved responses to form Q&A instances and structuring pairs in formats suitable for Instruct Fine-Tuning of LLMs with natural language instructions. **Generated Dataset Validation:** Implementation of automatic checks for coherence, redundancy, and completeness, along with functional validation through fine-tuning experiments using the instruction-formatted dataset.

3. Results

The Call2Instruct pipeline was implemented and validated using over 3,000 dual-channel call recordings from a telecommunications call center. The audio preprocessing stage employed Demucs-based noise suppression, K-Means clustering for IVR segment identification and removal, and the Whisper Large model optimized with Faster Whisper and LoRA for high-fidelity transcription with clear speaker attribution. ASR-induced artifacts were systematically rectified using filtering heuristics, followed by rigorous PII redaction, which confirmed the complete removal of sensitive information while preserving semantic integrity. Customer intent distillation into canonical question format achieved a 100% success rate through iterative prompt optimization, improved from an initial 53% baseline. Vector representations (1536-dimensional) for all 3,120 customer intents and agent responses were generated using text-embedding-ada-002 and indexed in Elasticsearch. For each intent vector, k-Nearest Neighbors search (k=3) retrieved the most semantically similar agent responses, which were synthesized into coherent answers through LLM-based refinement. Intrinsic quality assessment revealed exceptional structural and semantic integrity: an automated quality heuristic identified only 2 pairs (0.06%) as potentially invalid from 3,120 generated Q&A pairs, resulting in a 99.94% automated acceptance rate. This high internal consistency confirms that the pipeline reliably generates structurally sound and contextually relevant instruction-tuning data. Extrinsic validation through the fine-tuning of the Llama 2 7B model using the Lamini framework demonstrated a marked improvement in domain-specific competence. When presented with simulated telecommunications queries, the fine-tuned model consistently produced contextually appropriate and factually relevant responses, contrasting sharply with the generic answers from non-specialized base models.

4. Conclusion

The significance of this work lies in addressing the critical bottleneck of creating high-quality, domain-specific datasets by offering a systematic path to unlock the potential of inherently noisy call center data for Instruct Fine-Tuning. The end-to-end approach integrating audio processing, text normalization, semantic extraction, and validation represents a comprehensive solution. While this work establishes a promising foundation for the automated generation of instructional datasets from call center audio, substantial opportunities remain for enhancements that can lead to even more capable and useful LLMs in practical customer interaction applications. Questions regarding the methodology, implementation, and results obtained can be directed to the authors.