RetaiLLM: A Multilingual, Optimised LLM-based Chatbot System for Retail Management

Pedro José Vivancos-Vicente[1], Juan Salvador Castejón-Garrido[1], Camilo Caparrós-Laiz[2], Ronghao Pan[2], José Antonio García-Díaz[2] and Rafael Valencia-García[2]

[1]*VÓCALI SISTEMAS INTELIGENTES S.L. Parque Científico de Murcia, Carretera de Madrid km 388. Complejo de Espinardo, 30100 Murcia, Spain*

[2]*Departamento de Informática y Sistemas, Universidad de Murcia, Campus de Espinardo, 30100 Murcia, Spain*

Abstract

RetaiLLM enhances chatbot solutions powered by Large Language Models adapted for small retail businesses. Although LLMs have impressive conversational capabilities, they struggle in retail scenarios due to specific challenges. These include hallucinations, the need for additional fine-tuning to handle domain-specific vocabulary and interactions, and the computational demands of real-time deployment. These limitations can result in misinformation, loss of customer trust and increased support costs, which is especially critical for small businesses with limited resources. Furthermore, many existing chatbot systems struggle with multilingual interactions and integration with popular communication platforms. To overcome these challenges, VÓCALI has developed RetaiLLM, a project that combines LLMs with contextual information from retail management companies and online sources. Using a combination of quantisation techniques and a Retrieval-Augmented Generation approach, RetaiLLM optimises LLMs by providing a hybrid search mechanism that combines semantic vector search with fuzzy text-based retrieval. This ensures precise, contextually relevant answers, reduces hallucinations and provides a correction mechanism.

Keywords

Large Language Models, Quantization, Hallucination, Chatbot, Retrieval-Augmented Generation

1. Introduction

In the context of retail management, many small businesses lack the economic, human or technological resources required to respond quickly and effectively to customer queries. Although Large Language Models (LLMs) have impressive conversational capabilities, they are still not widely used in real-world retail scenarios due to challenges such as hallucinations, the need for additional fine-tuning and high computational requirements. Additionally, current chatbot systems often struggle to integrate seamlessly with popular communication platforms, and they do not always provide a satisfactory multilingual experience [1]. These limitations hinder the adoption of advanced conversational solutions by small and medium-sized retailers.

To address these challenges, we present RetaiLLM: a multilingual, LLM-based chatbot system optimised for small retail businesses. The project's primary goal is to develop an efficient, reliable, and customisable solution that can easily integrate into the communication channels commonly used by these businesses. This solution will provide accurate, contextually relevant answers without requiring large-scale infrastructure. RetaiLLM incorporates various techniques, such as model quantisation, Retrieval-Augmented Generation (RAG), hallucination detection, plugin-based customisation and emotion recognition. These components are implemented within a flexible, multi-platform, multi-lingual

SEPLN 2025: 41st International Conference of the Spanish Society for Natural Language Processing, Zaragoza, Spain, 23-26 September 2025.

pedro.vivancos@vocali.net (P. J. Vivancos-Vicente); juans.castejon@vocali.net (J. S. Castejón-Garrido); camilo.caparrosl@um.es (C. Caparrós-Laiz); ronghao.pan@um.es (R. Pan); joseantonio.garcia8@um.es (J. A. García-Díaz); valencia@um.es (R. Valencia-García)

https://www.vocali.net/ (P. J. Vivancos-Vicente); https://www.vocali.net/ (J. S. Castejón-Garrido); https://github.com/Smolky (J. A. García-Díaz); https://webs.um.es/valencia (R. Valencia-García)

0000-0002-5191-7500 (C. Caparrós-Laiz); 0009-0008-7317-7145 (R. Pan); 0000-0002-3651-2660 (J. A. García-Díaz); 0000-0003-2457-1791 (R. Valencia-García)

architecture equipped with a graphical interface and an administration dashboard for chatbot configuration and monitoring.

The project is divided into four main objectives: (OB1) development of a multilingual, reliable and efficient LLM-based chatbot system for retail management; (OB2) design of input/output communication interfaces; (OB3) implementation of a plugin system; and (OB4) development of an administration dashboard for chatbot creation and monitoring.

This project is a collaboration between VÓCALI, a company specialised in Natural Language Processing (NLP) and speech technologies, and the TECNOMOD research group at the University of Murcia. RetaiLLM is funded by the CDTI and the European Union (ERDF) under the project code IDI-20240115.

2. Background information

Recent advances in LLMs have significantly improved the performance of conversational systems [2]. Models such as ChatGPT, LLaMA [3] or Gemma [4] demonstrate strong capabilities in understanding and generating natural language. However, their adoption in real-world applications, especially in small business environments such as retail, remains limited due to their high resource requirements, lack of domain adaptation, and the risk of generating incorrect or fabricated content [5].

To make these technologies more accessible, several optimisation techniques have been proposed [6]. One of the most important is quantization, which reduces the model size and memory consumption by lowering numerical precision (e.g., to 8-bit or 4-bit formats). Combined with QLoRA (Quantized Low-Rank Adaptation) [7], models can be fine-tuned with minimal computational cost, making them viable for deployment on local servers or mid-range GPUs [8].

Another key approach is RAG, which improves response accuracy by incorporating relevant external knowledge during inference. By combining semantic vector search with traditional keyword-based retrieval, RAG enables chatbots to provide contextual and factual responses without retraining the model. In some cases, this process can be enhanced by using structured knowledge representations such as ontologies [9, 10].

One of the biggest disadvantages of LLMs is their tendency to generate hallucinated content [11, 12]. Instead of relying on larger models to verify responses, our approach focuses on detecting and correcting factual errors by comparing the chatbot's responses with retrieved documents, improving reliability in customer-facing scenarios.

3. System architecture

In short, the system is a modular, multilingual chatbot platform designed to support small retail businesses using state-of-the-art natural language processing techniques. It connects to popular messaging services via customisable interfaces, processes user queries with optimised LLMs and enhances responses using a RAG mechanism. To ensure fluency and factual accuracy, it incorporates modules for hallucination detection and response regeneration. The system also includes an emotion recognition module that can classify user input into different emotional states, providing valuable insight into customer satisfaction. The system's architecture also supports plug-ins that can be used to trigger actions such as reservations or order tracking, and it provides a dashboard for chatbot configuration and usage monitoring. The system's overall structure is shown in Figure 1.

3.1. Efficient LLM-Based Chatbot System

This module comprises two components: the I/O interfaces and the LLMs. To ensure seamless integration with popular communication platforms such as Telegram, Facebook and email, the system includes a RESTful API with specific endpoints for each channel. These endpoints convert incoming messages into a consistent internal format. Using a webhook-based architecture, external services can forward user messages to the system, which processes them asynchronously and returns responses adapted

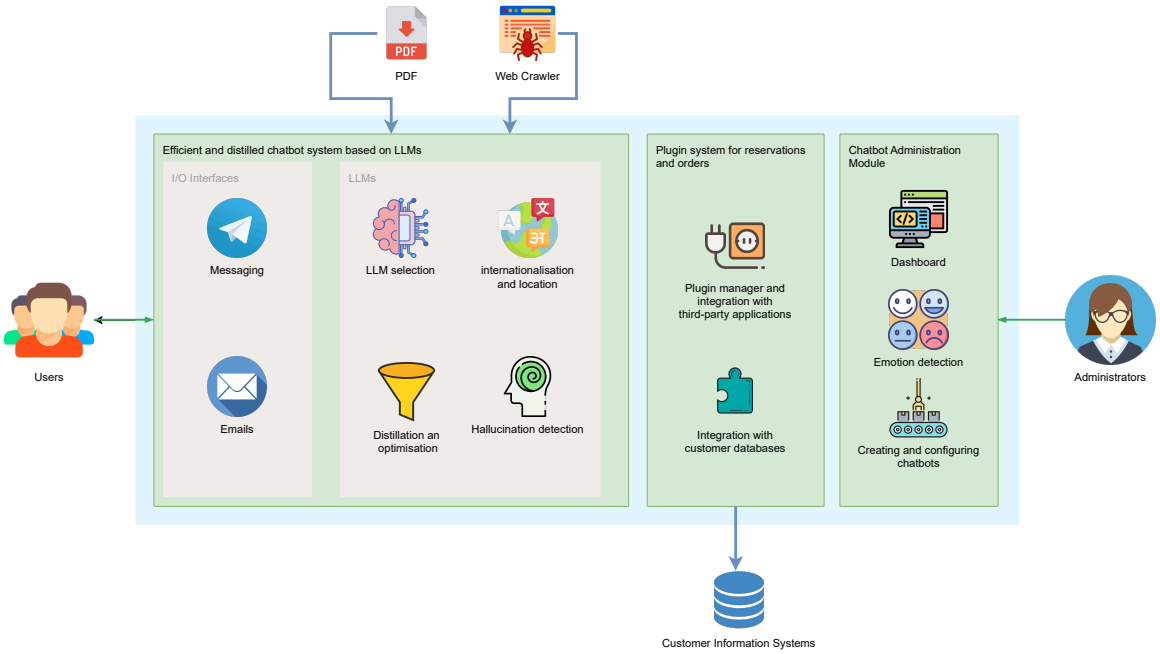

Figure 1: System architecture.

to each platform's format. The LLM component, on the other hand, is responsible for generating coherent, context-aware responses based on user input and retrieved information. This component is divided into four main functional blocks: (1) model selection, where appropriate instruction-tuned LLMs are selected based on performance and resource constraints; (2) internationalisation, which ensures multilingual support through language detection and translation; (3) distillation and optimisation, which reduce model size and improve inference speed through quantization and QLoRA; and (4) hallucination detection and correction, which aims to identify and mitigate factual inconsistencies in the model's responses.

Firstly, model selection identifies models that strike a balance between response quality and computational efficiency, rendering them suitable for use in small business environments. To this end, we evaluated several instruction-tuned LLMs that support multilingual input and are compatible with quantisation techniques.

The selected candidates include Gemma 2 (9B and 2B) and LLaMA (3.1 8B and 3.2 3B) models, which are optimised to follow natural language instructions and adapt to different communication contexts. In this case, an exhaustive study was carried out to select these models. This included analysing the disk space requirements and performance of the models in different hardware configurations, as well as evaluating their response time on a specific dataset. This allowed us to identify models that met the technical requirements of the available hardware while maintaining acceptable efficiency in a production scenario. These models are integrated into a conversation module that enables the system to retain and reuse information from previous interactions with users. Prompting strategies have therefore been evaluated and adapted for each communication channel. For example, longer and more formal messages are generated for email, while shorter and more direct messages are generated for instant messaging. To improve the relevance and factual accuracy of the answers generated, the RAG module conducts a hybrid search of a document collection, combining semantic similarity via sentence embeddings, such as Distiluse Base Multilingual, Paraphrase Multilingual, All MiniLM and Multi-QA, with a fuzzy keyword search based on edit distance. The retrieval engine, implemented via Elasticsearch, supports both dense vector and raw text indexing. User queries are encoded and compared using k-nearest neighbours with cosine similarity to retrieve the most relevant passages. These are then ranked and incorporated into

the query, enabling the LLM to generate more accurate, contextualised responses.

Secondly, to facilitate multilingual communication, the system incorporates an internationalisation module that utilises automatic language identification and translation. Incoming messages are identified and translated into a reference language for processing. Once the response has been generated, the system ensures it is delivered in the user's original language by carrying out a consistency check and, if necessary, performing a final translation. This strategy guarantees clear and consistent communication, regardless of the user's language.

Thirdly, in order to ensure that LLMs can be used efficiently in environments with limited resources, we applied quantisation and QLoRA. However, these techniques have drawbacks, such as longer training times and an increased risk of hallucinations due to quantisation noise. We tested the models using two hardware setups: one based on an AMD EPYC MILAN 7313 CPU with an NVIDIA RTX 4090 GPU and another using an Intel Xeon 6530 CPU with an NVIDIA L40S GPU. Quantising the models to 8-bit and 4-bit significantly reduced memory usage and increased generation speed, with the smallest models showing the most significant gains. For instance, the LLaMA and Qwen variants experienced notable speed increases at 4-bit quantisation, particularly on high-performance hardware. While larger models benefited from reduced size, their speed improvements were more modest, likely due to their greater architectural complexity. Hardware differences also played a significant role: the NVIDIA L40S GPU system consistently outperformed the RTX 4090 setup, particularly with larger models and under heavier loads. These results emphasise the importance of balancing model size, quantisation level, and hardware capabilities when selecting LLMs for the efficient deployment of chatbots.

The process begins with intent classification, for which Transformer-based models are used that have been trained on domain-specific YAML datasets. Each intent is then linked to a prompt and a set of decoding parameters, such as temperature, top-k or top-p. These parameters reduce the likelihood of hallucinated outputs. Intents such as price enquiries or business hours are treated with stricter settings to prioritise factual accuracy. Once the LLM has generated a response, the system uses a custom Named Entity Recognition (NER) module developed for this project to extract relevant entities such as dates, locations, prices and phone numbers. These entities are then cross-checked against the retrieved RAG context, which contains verified business information from the vector database. If any discrepancies are found, the LLM regenerates the response using the corrected context and constraints in an iterative process until all critical entities align with the factual data. While formal benchmarking of the consistency module is ongoing, preliminary internal testing has shown promising results in terms of detecting and correcting factual errors.

Finally, a two-step detection and correction mechanism has been developed. First, a model trained with transformer-based architectures analyses the user query to determine whether a high degree of factual precision is required (e.g. when asking for prices or contact information). In such cases, the system adjusts the decoding parameters of the LLM, reducing the temperature to 0.25, the top-k to 30, and the top-p to 0.7, in order to promote more accurate and deterministic responses. For standard scenarios, more flexible values are employed (temperature 0.7, top-k 80 and top-p 1.0) to prioritise fluency. After generation, a named entity comparison is performed between the model output and the retrieved context. If any are found, the response is regenerated iteratively until it matches the source information. Finally, a consistency check is performed on the multilingual output to ensure that the translations retain the factual content, and then the final response is delivered to the user.

3.2. Plugin System for Reservations and Orders

To extend the system's functionality, a plugin architecture was implemented that enables administrators to activate or deactivate specific actions, such as booking reservations, retrieving order details or managing surveys. To support these real-world interactions, the chatbot engine's intent detection model was enhanced to recognise a broader set of user intents linked to plugin-triggered actions. When a relevant intent is detected, the system prompts the user for any missing information and interacts with external services to complete the task. The architecture is fully extensible, enabling new plugins to be easily integrated as future requirements arise.

These plugins incorporate customer-specific information using database adapters that connect to external client databases and import their contents into the vector database. This allows structured business data, such as product catalogues or customer records, to be included in the context used when generating responses. These connections are configurable, and administrators can add or remove them based on the specific needs of each deployment.

An example of this system can be seen in Figure 2, where the administrator can configure a user authentication plugin. Through a dedicated settings panel, the administrator is able to select which attributes should be requested from users when they need to log in via the chat interface. Available options include name, phone number, national ID, email address, and date of birth.

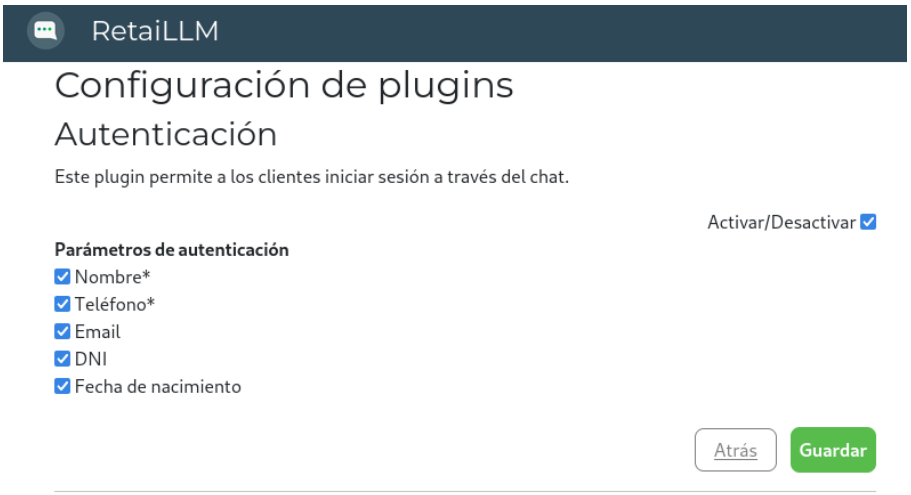

Figure 2: Authentication plugin configuration example.

3.3. Chatbot Administration Module

The platform includes an administration dashboard offering a centralised interface for managing chatbot configurations and monitoring usage. Administrators can customise prompts, datasets, plugins and communication channels, and access usage analytics such as conversation volume, reservation activity and word cloud visualisations. Calendar-based views and detailed statistics support effective supervision and ongoing system optimisation.

To analyse user behaviour and detect satisfaction levels, a transformer-based text emotion classification module is included. It has been trained on various corpora, including EmoContext [13], LiSSS [14], EMOVO [15] and MELD [16] among other datasets [17]. The model estimates the emotional content of each user message. These emotions are then integrated into the conversation logs and used to generate aggregated user sentiment statistics for administrators.

Finally, there is a graphical user interface (GUI) that enables administrators to efficiently configure and deploy chatbot instances. Through this interface, users can customise plugin access, select input/output channels (e.g. messaging apps, email or the web) and define the data sources for each chatbot. Once configuration is complete, the system automatically generates the deployment settings required for each selected interface, enabling tailored chatbot instances to be launched with minimal effort.

4. Conclusions and further work

This project marks a significant milestone in the development of flexible, efficient, multilingual chatbot systems designed to meet the needs of small businesses. RetaiLLM combines optimised LLMs with a modular architecture that includes plug-in integration, emotion detection, hallucination control and multilingual support, enabling seamless deployment across multiple communication channels.

Its extensibility and configuration interface make RetaiLLM suitable for a wide range of real-world scenarios, reducing the technical barriers to adoption in sectors such as retail.

In future iterations of the system, we plan to enhance the chatbot's emotional intelligence by incorporating multimodal emotion detection. This involves combining textual and prosodic features from voice inputs [18, 19]. This will enable the system to better understand user intent and satisfaction, particularly in voice-based interactions. Furthermore, we intend to enhance the customisation of LLM prompts via user feedback loops, facilitating dynamic prompt adjustment based on actual usage. Other developments will include extending plugin functionality to support more complex business workflows and integrating external knowledge graphs to improve response generation in domain-specific scenarios.

In future stages, we will conduct a quantitative evaluation of RetailLLM. To achieve this, we will conduct thorough benchmarking of its multilingual capabilities and hallucination mitigation strategies, and analyse the efficiency gains derived from using retrieval-augmented generation (RAG) and quantisation techniques. We also plan to conduct comparative studies with existing LLM-based retail assistants to validate our approach. To promote reproducibility and encourage adoption within the research and development community, we also intend to publish implementation details, including model configurations, dataset sizes and key hyperparameters.

Future work will focus on exploring strategies to ensure compliance with data protection regulations, such as the European GDPR. This is particularly important when RetaiLLM is integrated with private customer databases or internal retail systems. This will involve adopting techniques for data anonymisation and pseudonymisation, implementing access control mechanisms and establishing secure data processing pipelines. We also intend to examine the implications of storing conversational histories, implementing user consent protocols and ensuring transparency in automated decision-making. These are all key aspects of the ethical deployment of LLM-based solutions in real-world retail environments.

Acknowledgments

This work is being funded by CDTI and the European Regional Development Fund (FEDER / ERDF) through project RetaiLLM IDI-20240115.

Declaration on Generative AI

During the preparation of this work, the authors used DeepL for grammatical and spelling correction. After using this tool, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] G. Nikhil, D. R. Yeligatla, T. C. Chaparala, V. T. Chalavadi, H. Kaur, V. K. Singh, An Analysis on Conversational AI: The Multimodal Frontier in Chatbot System Advancements, in: 2024 Second International Conference on Inventive Computing and Informatics (ICICI), IEEE, 2024, pp. 383–389.

[2] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., A survey of large language models, 2025. URL: https://arxiv.org/abs/2303.18223. arXiv:2303.18223.

[3] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[4] G. Team, M. Riviere, et al., Gemma 2: Improving open language models at a practical size, 2024. URL: https://arxiv.org/abs/2408.00118. arXiv:2408.00118.

[5] S. Johnson, D. Hyland-Wood, A primer on large language models and their limitations, 2024. URL: https://arxiv.org/abs/2412.04503. arXiv:2412.04503.

[6] L. Donisch, S. Schacht, C. Lanquillon, Inference optimizations for large language models: Effects, challenges, and practical considerations, 2024. URL: https://arxiv.org/abs/2408.03130. arXiv:2408.03130.

[7] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLoRA: Efficient Finetuning of Quantized LLMs, 2023. URL: https://arxiv.org/abs/2305.14314. arXiv:2305.14314.

[8] A. Chavan, R. Magazine, S. Kushwaha, M. Debbah, D. Gupta, Faster and Lighter LLMs: A Survey on Current Challenges and Way Forward, 2024. URL: https://arxiv.org/abs/2402.01799. arXiv:2402.01799.

[9] J. M. Ruiz-Sánchez, R. Valencia-García, J. T. Fernández-Breis, R. Martıínez-Béjar, P. Compton, An approach for incremental knowledge acquisition from text, Expert Systems with Applications 25 (2003) 77–86. URL: https://www.sciencedirect.com/science/article/pii/S0957417403000083. doi:https://doi.org/10.1016/S0957-4174(03)00008-3.

[10] J. A. García-Díaz, M. Cánovas-García, R. Valencia-García, Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in Latin America, Future Generation Computer Systems 112 (2020) 641–657.

[11] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, ACM Transactions on Information Systems 43 (2025) 1–55. URL: http://dx.doi.org/10.1145/3703155. doi:10.1145/3703155.

[12] Z. Xu, S. Jain, M. Kankanhalli, Hallucination is inevitable: An innate limitation of large language models, 2025. URL: https://arxiv.org/abs/2401.11817. arXiv:2401.11817.

[13] A. Chatterjee, K. N. Narahari, M. Joshi, P. Agrawal, SemEval-2019 task 3: EmoContext contextual emotion detection in text, in: J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, S. M. Mohammad (Eds.), Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 39–48. URL: https://aclanthology.org/S19-2005/. doi:10.18653/v1/S19-2005.

[14] J.-M. Torres-Moreno, L.-G. Moreno-Jiménez, LiSSS: A toy corpus of Spanish Literary Sentences for Emotions detection, 2020. URL: https://arxiv.org/abs/2005.08223. arXiv:2005.08223.

[15] G. Costantini, I. Iaderola, A. Paoloni, M. Todisco, EMOVO corpus: an Italian emotional speech database, in: N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 3501–3504. URL: https://aclanthology.org/L14-1478/.

[16] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations, 2019. URL: https://arxiv.org/abs/1810.02508. arXiv:1810.02508.

[17] A. Salmerón-Ríos, J. A. García-Díaz, R. Pan, R. Valencia-García, Fine grain emotion analysis in Spanish using linguistic features and transformers, PeerJ Computer Science 10 (2024) e1992.

[18] R. Pan, J. A. García-Díaz, M. Á. Rodríguez-García, R. Valencia-García, Spanish MEACorpus 2023: A multimodal speech–text corpus for emotion analysis in Spanish from natural environments, Computer Standards & Interfaces 90 (2024) 103856.

[19] R. Pan, J. A. García-Díaz, M. Á. Rondríguez-García, F. García-Sánchez, R. Valencia-García, Overview of EmoSPeech at IberLEF 2024: Multimodal Speech-text Emotion Recognition in Spanish, Procesamiento del Lenguaje Natural 73 (2024) 359–368.