LREC 2026

LANLP: Bridging Ibero and Latin American NLP Communities

Networking Symposium Proceedings

Editors
Luis Chiruzzo, Pablo Gamallo, Rafael Muñoz Guillena and German Rigau Claramunt

16 May 2026

ISBN 978-2-493814-95-1

Preface

It is with great pleasure that we present the proceedings of the **Networking Symposium on Natural Language Processing in Ibero and Latin America (LANLP)**, co-located with **LREC 2026** in Palma de Mallorca. This inaugural symposium was established to fill a critical historical gap: providing a dedicated forum for researchers across the Iberian Peninsula and Latin America to share initiatives, corpora, and tools.

The Ibero-American region represents a vast and diverse linguistic landscape, encompassing major world languages such as Spanish (558M speakers) and Portuguese (267M speakers), alongside tens of millions of speakers of indigenous languages including Quechua, Guaraní, Nahuatl, and Aymara. Despite this cultural richness, over 88% of the world's languages remain unsupported by modern language technologies. This networking event directly addresses that disparity by fostering community-driven resource development and evaluation for both major and minoritized languages.

The emergence of Large Language Models (LLMs) represents a pivotal shift for our community, moving the region from the passive consumption of Northern-centric models to the active development of localized foundation models. While this era presents challenges—such as digital scarcity, algorithmic colonialism, and "language domination"—it also offers unprecedented opportunities for innovation:

- **Technological Sovereignty:** Initiatives like **ILENIA** and **ALIA** projects demonstrate the feasibility of building regional foundation models that prioritize regional specificity and language inclusion.
- **Efficiency and Equity:** Our research agenda emphasizes Small Language Models (SLMs) and synthetic data to mitigate linguistic inequalities.
- **Trustworthy AI:** The community is prioritizing privacy, accountability, and cultural alignment, ensuring that deployment in high-stakes environments like healthcare is both robust and ethically grounded.

These proceedings reflect a research agenda where linguistic diversity and technical robustness are inseparable. The success of this symposium is built upon the synergy of established ecosystems, including **SEPLN**, **PROPOR**, **AmericasNLP** and **CLARIAH-ES**. By weaving together these varied perspectives—from computer science to the digital humanities—we aim to amplify the potential of our current research capabilities on a global stage.

We hope these proceedings serve as a catalyst for long-term engagement, converting short-term networking into durable cross-regional collaboration and ensuring that Ibero-American languages thrive in the digital age.

The LANLP 2026 Organizing Committee

Eugenio Martínez Cámara, Luis Chiruzzo, Pablo Gamallo, Rafael Muñoz, and German Rigau

Organizing Committee

Organizers

Pablo Gamallo (PROPOR and Universidade de Santiago de Compostela, Spain)
Eugenio Martínez Cámara (Universidad de Jaén, Spain)
Rafael Muñoz Guillena (SEPLN and Universidad de Alicante, Spain)
Luis Chiruzzo (AmericasNLP and Universidad de la República, Uruguay)
German Rigau Claramunt (CLARIA-ES and Universidad del País Vasco, Spain)

Scientific Committee

Itziar Aldabe (Universidad del País Vasco, Spain)
Elvis Alves de Souza (Universidade de São Paulo, Brazil)
Xabier Arregi (Universidad del País Vasco, Spain)
Elena Battaner (Universidad Rey Juan Carlos, Spain)
Alba Bonet (Universidad de Alicante, Spain)
Marília Costa Rosendo Silva (Universidade de São Paulo, Brazil)
Viviana Cotik (Universidad de Buenos Aires, Argentina)
Fermín Cruz (Universidad de Sevilla, Spain)
Iria de-Dios-Flores (Universitat Pompeu Fabra, Spain)
Cristina España-Bonet (BSC, Spain)
Ainara Estarrona (Universidad del País Vasco, Spain)
Aritz Farwell (Universidad del País Vasco, Spain)
Joaquim Ferreira da Silva (Universidade Nova de Lisboa, Portugal)
Marcos Garcia (Universidade de Santiago de Compostela, Spain)
José Manuel Gómez Pérez (Expert.ai, Spain)
José ángel González, (Symanto Research, Spain/Germany)
María Grandury (EPFL, Switzerland)
Inma Hernaez (Universidad del País Vasco, Spain)
Elena Lloret (Universidad de Alicante, Spain)
Manuel Mager (Johannes Gutenberg University of Mainz, Germany)
María Teresa Martín Valdivia (Universidad de Jaén, Spain)
Paloma Martínez Fernández (Universidad Carlos III de Madrid, Spain)
Arturo Montejo (Universidad de Jaén, Spain)
Andrés Montoyo Guijarro (Universidad de Alicante, Spain)
Renato Moraes Silva (Universidade de São Paulo, Brazil)
Maite Oronoz (Universidad del País Vasco, Spain)
Manuel Palomar Sanz (Universidad de Alicante, Spain)
Thiago Pardo (Universidade de São Paulo, Brazil)
Paulo Quaresma (Universidade de évora, Portugal)
Livy Real (Universidade de São Paulo, Brazil)
Aiala Rosá (Universidad de la República, Uruguay)
Robiert Sepúlveda Torres (Universidad de Alicante, Spain)
Susana Sotelo (Universidade de Santiago de Compostela, Spain)
Luis Alfonso Ureña López (Universidad de Jaén, Spain)
Rafael Valencia García (Universidad de Murcia, Spain)

Table of Contents

Networking Symposium Program

9:00–9:30 ***Welcome and Opening Remarks***

9:30–10:30 ***Panel Session I - About the LANLP supporting organizations***

9:30–10:30 *Bridging Ibero and Latin American NLP communities*
Eugenio Martínez Cámara, Luis Chiruzzo, Pablo Gamallo, Rafael Muñoz Guillena and German Rigau

11:00–12:00 *Keynote: Opportunities & Challenges of LLMs for Ibero-Latin Languages*
Antonio Branco

12:00–13:00 **Paper Session I**

12:00–12:20 *An Oral-first Interactive Agentic System for Guaraní Speakers*
Samantha Adorno, Akshata Kishore Moharir and Ratna Kandala

12:20–12:40 *AI-TraLow: AI-Driven Translation for Low-Resource Languages and Cultures*
Antoni Oliver, Maite Melero, Felipe Sanchez-Martinez and Víctor M. Sánchez-Cartagena

12:40–13:00 *OpenCor: Latin American and Iberian Languages Open Corpora Forum*
Livy Real and Valeria de Paiva

14:00–15:20 **Paper Session II**

14:00–14:20 *MedicaLLM: LLM-Driven Speech and Language Solutions for Healthcare*
Ronghao Pan, Pedro José Vivancos-Vicente, Juan Salvador Castejón-Garrido, Tomás Bernal-Beltrán and Rafael Valencia-Garcia

14:20–14:40 *mCS-LM: Multimodal Customer Service and Incident Management Systems Based on Large Language Models*
Carlos Díaz-Morales, Marcos Checa-Rubio, Tomás Bernal-Beltrán, Ronghao Pan, David Barbáchano, María del Pilar Salas-Zárate, Mario Andrés Paredes-Valverde and Rafael Valencia-Garcia

14:40–15:00 *SAFEWORDs: un marco reproducible para anonimización conforme al RGPD y evaluación de generación en lenguas cooficiales*
Rafael Muñoz Guillena, Manuel Palomar, Elena Lloret and Nuria Fernández

15:00–15:20 *Exploration of Sentence Representations in Spanish BERT-like Models*
Gonzalo Herrera, Aiala Rosá and Luis Chiruzzo

15:20–16:00 ***Panel Session II: Strategic Network RutaMdL***

16:30–17:15 ***Hands-on LANLP***

17:15–17:45 ***Group Reports & Discussion***

17:45–18:00 ***Concluding Remarks & Next Steps***

Bridging Ibero and Latin American NLP communities

Eugenio Martínez, Luis Chiruzzo, Pablo Gamallo, Rafael Muñoz, German Rigau
CEATIC (UJA), UdelaR, CiTIUS (USC), CENID (UA), HiTZ (UPV/EHU)
emcamara@ujaen.es, luischir@fing.edu.uy, pablo.gamallo@usc.gal,
rafael@dlsi.ua.es, gernan.rigau@ehu.eus

Abstract

LANLP focuses on community-driven resource development and evaluation for Iberian languages, and diverse Latin American languages (including indigenous and minority languages). We aim to bridge regional communities to share initiatives, corpora and tools. LANLP fills this gap, fostering new contacts between Iberian and Latin American NLP research groups. The goals are to (1) highlight challenges in processing these languages, (2) share novel datasets and models, and (3) catalyze future collaborations and shared tasks. We emphasize both academic rigor and community inclusivity, encouraging contributions from established researchers and grassroots language advocates alike.

Keywords: Networking, NLP communities, digital divide

1. Introduction

We organize a Networking Symposium on Latin American NLP (LANLP), focusing on natural language processing for the diverse languages of the Iberian Peninsula and Latin America. This region includes major world languages (e.g., Spanish (∼558M speakers), Portuguese (∼267M)), as well as regional and indigenous languages. For example, Latin America alone hosts tens of millions of speakers of Quechua (∼10M), Guaraní (>6M), Nahuatl (∼2M), and Aymara (∼2M), among many others. Such languages are highly underresourced: over 88% of the world's languages remain largely unsupported by language technologies.

This networking event addresses that gap by promoting collaboration on ethically and culturally sensitive resource creation, evaluation, and novel methods for low-resource multilingual NLP in Iberian and Latin American languages and varieties. Our goal is to bring together communities (SEPLN, PROPOR, AmericasNLP, and CLARIAH-ES) to share cutting-edge research, language resources, and best practices.

LANLP focuses on community-driven resource development and evaluation for Iberian languages, and diverse Latin American languages (including indigenous and minority languages). We aim to bridge regional communities: for instance, past forums such as OpenCor[1] note that "Latin American and Iberian communities... did not have an established event" to share initiatives, corpora, and tools. LANLP tries to fill this gap, fostering new contacts between Iberian and Latin American NLP research groups.

The goals are to:

1. Highlight challenges in processing these languages;
2. Share novel datasets and models;
3. Catalyze future collaborations and shared tasks.

We emphasize both academic rigor and community inclusivity, encouraging contributions from established researchers and grassroots language advocates alike.

2. LANLP Communities

2.1. SEPLN & IberLEF

The Spanish Society for Natural Language Processing (SEPLN), founded in 1983, has played a central role in structuring the scientific and technological community for Natural Language Processing (NLP) in Spain. Its contribution goes beyond the organization of an annual conference: SEPLN has progressively consolidated a stable ecosystem for research, evaluation, dissemination, technology transfer, and collaboration among universities, research centers, public institutions, and companies. This long-term community-building role has been essential for strengthening both scientific excellence and the practical impact of language technologies in the Iberian context.

Within this ecosystem, the journal *Procesamiento del Lenguaje Natural* (PLN Journal) is a core instrument for scientific dissemination. The journal, published biannually, has become a reference venue for research in NLP in Spanish and related contexts. Its relevance is not only historical, but also institutional, as reflected in recognized quality indicators. In particular, the journal has obtained the FECYT quality seal and is indexed in ESCI (Emerging Sources Citation Index, Web of Science.

[1] https://opencor.gitlab.io/

Proceedings of LANLP: Bridging Ibero and Latin American NLP Communities, pages 1–8
16 May 2026.

2025:Q2 Linguistics, Q4 Computer Science. Artificial Intelligence). In addition, SEPLN highlights its visibility through bibliometric indicators and its role as a trusted publication channel for the community. These quality markers reinforce the journal's strategic value for the consolidation and international visibility of NLP research.

SEPLN has also strengthened the technology transfer and industry engagement dimension of the field through initiatives such as TecnoLing. Integrated into the SEPLN conference framework, TecnoLing is designed as a technology showcase and networking space that brings together companies, startups, R&D centers, and public-sector actors. Its format typically includes demonstrations, exhibitor stands, presentations, and networking sessions, explicitly aimed at facilitating knowledge transfer, identifying collaboration opportunities, and promoting the adoption of language technologies in real-world settings. TecnoLing is therefore especially important because it helps transform the conference environment into a broader innovation ecosystem, where academic advances can more easily connect with industrial needs and societal applications.

In the area of shared evaluation, the trajectory that has led to IberLEF (Iberian Languages Evaluation Forum) represents one of the most significant contributions of the SEPLN community. IberLEF emerged from earlier evaluation campaigns such as TASS and IberEval, and has evolved into a consolidated forum for the comparative evaluation of NLP systems developed by research groups and companies. Its value lies in promoting reproducible evaluation, shared datasets and benchmarks, and coordinated progress on challenges relevant to Iberian languages.

IberLEF is better understood through its scale and thematic diversity. Recent editions have included a large number of shared tasks and broad international participation, with contributions spanning a wide range of topics such as sentiment analysis, harmful content detection, inclusive language, simplification and easy-to-read text, multimodal analysis (e.g., memes), historical text processing, question answering, clickbait detection, and Portuguese-focused tasks, among others. This diversity reflects the maturity of the forum and its capacity to support both foundational and applied research across multiple languages and domains.

Taken together, SEPLN, the PLN Journal, IberLEF, and TecnoLing constitute a highly valuable infrastructure for NLP in Spain and the broader Iberian research space. They combine scientific publication, editorial quality standards, shared evaluation practices, and technology transfer mechanisms in a coherent ecosystem. This integrated structure is particularly relevant for fostering sustainable collaboration, improving research visibility, and ensuring that advances in NLP and AI can generate tangible scientific, economic, and social impact.

2.2. PROPOR

PROPOR refers both to a long-standing research community dedicated to the computational processing of Portuguese and related varieties, and to the biennial International Conference on the Computational Processing of Portuguese, which has served as its main scientific meeting since 1993. Over more than three decades, the PROPOR community and conference have provided a central forum for research on written and spoken Portuguese, fostering methodological advances, language resource development, and collaboration between academia and industry. The conference has alternated between Brazil and Portugal, reflecting the transatlantic nature of the Portuguese-speaking world, and has recently expanded its scope to more explicitly include Galician, as evidenced by the 2024 edition hosted in Santiago de Compostela, marking a significant step toward a recognition of Galician as a variety of the Lusophone diasystem. Both the community and the conference are managed by a Steering Committee that is renewed every two years and includes Portuguese, Brazilian, and Galician researchers.[2]

While PROPOR has successfully consolidated a transatlantic research community around Portuguese (and more recently Galician) technology, it has historically evolved in parallel with other Iberian and Latin American NLP communities, particularly those focused on Spanish and indigenous languages. By alternating its conference venues between Brazil, Portugal and Galicia, PROPOR has encouraged sustained collaboration between research groups on all continents, promoting the sharing of methodologies, language resources, evaluation frameworks, and technological developments. The conference and its associated community provide a stable organizational structure and a long-standing forum for scientific exchange, supporting both academic research and industrial applications. This transatlantic dimension positions PROPOR as a key hub for multilingual research, enabling studies on many language varieties of Portuguese across several continents, including not only America and Europe, but also Africa, Asia and Oceania. This contributes to the development of inclusive and interoperable language technologies for the Portuguese linguistic space.

[2] https://propor.org

2.3. AmericasNLP

AmericasNLP[3] is a workshop that has been running since 2021 with the objective of bringing together researchers of NLP and computational linguistics applied to indigenous languages throughout the Americas. The aim is to encourage and increase the visibility of work on indigenous languages of the Americas, by encouraging research on NLP, computational linguistics, corpus linguistics and speech for indigenous languages, to connect researchers and professionals from underrepresented communities and native speakers of endangered languages, and more generally to promote machine learning approaches suitable for low-resource languages.

So far there have been five editions of the workshop, with an upcoming sixth edition on the way, co-located either with one of the *CL conferences or with NeurIPS. In total, more than 100 papers have been published in the proceedings of the workshop, including system papers of the participants of the eight shared tasks organized during this time, which include: Machine Translation for low-resource indigenous languages, creation of educational materials, Automatic Speech Recognition of indigenous languages, and MT metrics for indigenous languages.

2.4. CLARIAH-ES

In the current context of digital transformation, study, research, and development in the humanities, arts and social sciences require scientific and technological infrastructures that allow for the computational processing of textual, visual, numerical and/or audio data. The CLARIAH-ES[4] (F. et al., 2024) infrastructure supports and contributes to the management and coordination in Spain of the European Research Infrastructure Consortia (ERIC) CLARIN[5] (focused on digital data and processes related to Language) and DARIAH[6] (focused on digital data and processes related to arts and humanities scholars). Although both CLARIN and DARIAH are independent ERIC infrastructures, some European countries have formed joint CLARIAH consortia. Both infrastructures promote multilingualism, digital methods, interoperability, maintenance and reuse of resources, open science, visibility and scientific cooperation in Europe, thus overcoming the fragmentation of research communities and increasing the impact of their research.

CLARIAH-ES contributes to the advancement of research in the humanities and social sciences, as well as to its strategic positioning in international projects and programs. With this in mind, CLARIAH-ES seeks to bring together research groups and initiatives that have a stake in these research infrastructures and that wish to reduce the digital divide, promoting new multidisciplinary lines of research in the humanities, arts and social sciences (and beyond) by facilitating their digital transformation with the help of language technologies.

The ambitious goals of CLARIAH-ES can only be achieved by bringing together the necessary resources in terms of data, computing facilities, and knowledge that are not available to any one research group in Spain. CLARIAH-ES is formed by a multidisciplinary group of twelve leading research centers in Language Technologies (TL), Artificial Intelligence (AI), High Performance Computing (HPC), linguistic experts in the official languages in Spain (Spanish, Catalan, Basque and Galician), and experts in digital transition in the areas of humanities, arts and social sciences.

In summary, the CLARIAH-ES research infrastructure seeks a tangible impact on society. Through greater exchange of knowledge, data, technologies, infrastructures, skills, and best practices, we aim to amplify the potential of our current research capabilities. This collaborative synergy will not only ensure the sustainability of tools and services, but also foster collaborative environments. Our aspirations transcend borders, encompassing Europe, Ibero-America, and the global stage as we endeavor to increase funding opportunities for vital infrastructures. The interdisciplinarity of the participating research groups, which includes areas of research as diverse as computer science, philology, social sciences, history, etc., ensures a broad contribution from different perspectives. By properly weaving together these perspectives, we can develop research results that are useful for promoting high-impact digital tools and artificial intelligence applications in different social science and digital humanities scenarios, including research and cultural infrastructures such as libraries and museums.

3. Related initiatives

3.1. ILENIA and ALIA

To understand the current trajectory of Ibero-American NLP, it is essential to examine large-scale institutional projects that provide the backbone for digital sovereignty. Two of the most significant initiatives in the Iberian context are **ILENIA** and **ALIA**, which focus on the structural development of language resources and foundational AI models for Spanish and the co-official languages of Spain.

The **ILENIA** project[7] is a cornerstone initiative

[3] https://americasnlp.org/
[4] http://www.chariah.es
[5] http://www.clarin.eu
[6] http://www.dariah.eu
[7] https://proyectoilenia.es/

designed to promote the digital presence and technological parity of Spain's linguistic diversity. By coordinating the efforts of leading research centers—including the BSC-CNS, HITZ Center (UPV/EHU), CiTIUS and ILG (USC), and CENID (UA)—ILENIA focuses on creating high-quality, interoperable Language Technology infrastructure. The main objectives are the following:

- **Multilingual Infrastructure:** The project facilitates the creation of advanced corpora, lexicons, and tools for Spanish, Catalan, Basque, and Galician.
- **Scientific Cooperation:** It aligns with the goals of **CLARIAH-ES** to reduce the digital divide and promote open science across the humanities and social sciences.
- **Technological Integration:** ILENIA ensures that resources follow common standards, enabling their reuse in a wide variety of AI applications, from machine translation to speech processing.

The **ALIA** project[8] (*Alianza por la Inteligencia Artificial en Español*) represents a strategic push to develop a large-scale foundation model that is culturally and linguistically grounded in the Hispanic world. The main goals are the following:

- **Digital Sovereignty:** It aims to provide a "technological public good," offering an open-source alternative that allows institutions and startups to build domain-specific applications without dependency on closed, proprietary APIs.
- **Mitigating Bias:** ALIA seeks to mitigate "algorithmic colonialism" by training models on regional legal, medical, and academic datasets rather than relying solely on translated English benchmarks.
- **Community Synergy:** ALIA benefits from the collaborative ecosystem, utilizing community-driven datasets and shared evaluation suites such as "IberoBench" (Baucells et al., 2025).

In the context of the LANLP Networking Symposium, these projects serve as successful blueprints for how institutional funding and academic expertise can be synthesized to achieve technological autonomy. The symposium aims to catalyze similar large-scale, transatlantic collaborations that ensure Ibero-American languages are central to the next generation of AI.

3.2. SomosNLP

SomosNLP[9] is a community of academic researchers, industry practitioners, and open-source contributors dedicated to creating and sharing resources that enable and accelerate the development of NLP in Ibero-American languages. The aim is to address linguistic inequity by connecting individuals from the region, collaboratively exploring unique challenges, and building the necessary open-access research infrastructure to enable rapid and relevant technological progress across the Spanish and Portuguese-speaking world. The community has been active since 2021 and was registered as a non-profit organization in 2025.

The SomosNLP initiative is built upon three core pillars: to promote open-source resource creation, to educate by providing access to high-quality content, and to connect for multidisciplinary collaboration.

The annual SomosNLP hackathon is the foundational mechanism for engagement, typically attracting over 300 registrations. These events facilitate resource creation collaboratively. The hackathons are framed as open innovation spaces where participants collectively identify, build, and contribute to open-source resources. The hackathons are open to all NLP practitioners—from industry experts to independent developers and academics—welcoming anyone regardless of their background or expertise. This collaborative model has generated significant and verifiable technical output on Hugging Face. The hackathon is supported by established companies: Hugging Face has been the gold sponsor since the first edition in 2022, and for the 2025 event, the hackathon secured sponsorship from Mistral and was recognised as a Cohere community grantee.

The education strategy focuses on making high-quality resources universally accessible. SomosNLP hosts a curated library of 70+ recorded talks and workshops on the public YouTube channel. The knowledge shared in the talks and workshops explicitly encourages open source code development, data sharing, and open methods/protocols—the essential building blocks for ethical, reproducible, and responsible NLP research.

To connect and bridge the gap between researchers across the Ibero-American community and the global NLP community, SomosNLP enables spaces for targeted networking. The SomosNLP Discord hosts over 2000 members who share resources, projects, events, relevant news, and support each other. Moreover, in 2025 SomosNLP started hosting Birds-of-a-Feather (BoF) events at major international conferences: ACL, COLM, and NeurIPS.

[8] http://alia.gob.es

[9] https://somosnlp.org

A successful example of international cross-institutional open-source collaboration led by SomosNLP is “La Leaderboard” (Grandury et al., 2025), a LLM leaderboard for Spanish varieties and languages of Spain and Latin America. This leaderboard hosts 66 benchmarks in Spanish, Catalan, Basque, and Galician, donated by 13 research groups. The paper presenting La Leaderboard was accepted at ACL Main 2025, and the web interface receives around 3.000 monthly visits.[10]

3.3. LATAM-GPT: A Foundation for Regional Digital Sovereignty

The emergence of **LATAM-GPT** represents a pivotal shift in the Ibero-American AI landscape, moving the region from passive consumption of Northern-centric models to the active development of localized foundation models. Led by Chile’s National Center for Artificial Intelligence (CENIA) and supported by a coalition of over 60 institutions across 15 countries, LATAM-GPT is designed to address the "linguistic and cultural representation gap" inherent in global Large Language Models (LLMs).

LATAM-GPT is built upon the `Llama 3.1` architecture and operates at a scale of approximately 50 billion parameters, positioning it within the same performance tier as GPT-3.5. Unlike proprietary models where training data remains opaque, LATAM-GPT’s development prioritizes regional specificity through a curated corpus:

- **Data Volume:** A massive $18 - 20.5$ TB dataset comprising over 1 billion documents.
- **Linguistic Diversity:** The model integrates high-resource languages (Spanish and Portuguese) alongside $50+$ indigenous languages such as Quechua, Mapuche, and Guaraní.
- **Infrastructure:** Initial training was conducted using high-performance computing clusters, including nodes equipped with NVIDIA H200 GPUs, facilitated by partnerships with AWS and the Data Observatory.

The primary objective of the LATAM-GPT project is to mitigate *algorithmic colonialism*—the tendency of global models to project Anglo-Saxon cultural norms and linguistic structures onto non-English speakers. By training on regional legal, medical, and academic datasets, LATAM-GPT provides:

1. **Semantic Precision:** Nuanced understanding of regional slang, idioms, and local administrative terminology.
2. **Inclusion:** Reduced tokenization fragmentation for Spanish and Portuguese, which lowers computational costs and improves response quality for regional users.
3. **Technological Autonomy:** As a "technological public good," it offers an open-source alternative for local governments and startups to build domain-specific applications without dependency on closed APIs.

3.4. LatinX in Natural Language Processing (LXNLP)

The **LatinX in Natural Language Processing (LXNLP)** workshop,[11], operating under the institutional umbrella of LatinX in AI (LXAI), serves as a primary affinity group dedicated to enhancing the visibility and professional advancement of researchers from Ibero-American and Latin American origins. This initiative addresses the systemic under-representation of LatinX scholars in core NLP venues, where historical participation from the Global South has faced significant structural and financial barriers.

3.5. Overview of the Accepted Contributions

The accepted papers offer a coherent snapshot of current research priorities at the intersection of Ibero- and Latin American NLP, while also reflecting the methodological and sociotechnical diversity of the communities involved. Taken together, they show that work in the region is not limited to adapting mainstream NLP pipelines to new languages, but is increasingly concerned with building language technologies that are community-aware, resource-conscious, multilingual, and deployable in socially or institutionally sensitive settings.

A first thematic cluster focuses on low-resource, Indigenous, and underrepresented languages, highlighting the need for approaches that move beyond simple technological transfer from high-resource settings. In this line, the oral-first interactive system for Guaraní speakers foregrounds conversational agency and culturally aligned interaction design in primarily oral language contexts, while AI-TraLow addresses machine translation for low-resource languages through a combination of curated data, linguistic resources, and efficient modeling. OpenCor complements these contributions by documenting the collective and infrastructural dimension of language technology, emphasizing the importance of open corpora, lexical resources, and the sustainability challenges behind community-driven resource creation.

[10] https://huggingface.co/spaces/la-leaderboard/la-leaderboard

[11] https://www.latinxinai.org/naacl-2024

A second cluster centers on trustworthy and domain-sensitive language technologies, especially in regulated or high-stakes environments. SAFEWORDs proposes a reproducible framework for GDPR-aligned anonymization and evaluation across Spain's co-official languages, placing privacy, replicability, and governance at the core of LLM assessment. In parallel, mCS-LM and MedicaLLM explore applied multilingual and multimodal systems for customer service, incident management, and healthcare, respectively. Both contributions stress the importance of grounded generation, structured outputs, validation layers, and interoperability when deploying language technologies in real-world organizational contexts.

The program also includes work on foundational language modeling and linguistic representation, as illustrated by the study on sentence representations in Spanish BERT-like models. By analyzing how different layers encode syntactic and semantic information, this contribution provides a more fine-grained understanding of representational behavior in Spanish-language models, offering insights that are relevant not only for benchmarking, but also for downstream system design and evaluation.

Overall, the accepted papers reveal three broader tendencies. First, they confirm that resource creation, evaluation, and deployment must be treated as interconnected challenges, especially for languages and varieties that remain underrepresented in mainstream NLP. Second, they highlight a growing emphasis on responsible and context-aware AI, including privacy, accountability, interpretability, and cultural alignment. Third, they show that the field is advancing simultaneously at multiple levels: from infrastructures and corpora, to model analysis, to translation, speech, and domain-specific applications. In this sense, the program reflects a research agenda in which linguistic diversity, technical robustness, and societal relevance are increasingly inseparable.

4. Expected Outcomes

LANLP Networking Symposium at LREC 2026 is designed to be a catalyst for long-term comunity engagement. We anticipate several key outcomes categorized into academic, technical, and socio-economic domains.

4.1. Scientific and Technological Autonomy and Policy Influence

The symposium aims to provide the scientific backing needed for digital sovereignty.

We aim at generating a policy brief for regional science and technology funding agencies regarding the strategic importance of developing sovereign AI infrastructure to avoid over-reliance on external proprietary APIs.

4.2. Formalization of a Regional Research Roadmap

A primary expected outcome is the publication of a *LANLP 2026 White Paper*. This document will synthesize the findings from the panel discussions and breakout sessions to establish a "Community Priority List." This roadmap can identify:

- **Linguistic Priorities:** Identifying specific dialects and indigenous languages in urgent need of digitization and resource creation.
- **Benchmarking Standards:** Establishing standardized evaluation protocols for Ibero-American NLP tasks, ensuring that regional models are measured against culturally relevant metrics rather than just translated English benchmarks.
- **Sustainability:** Establishing a permanent digital directory of Ibero-American NLP labs to facilitate easier cross-border internship and PhD exchange programs.

4.3. Consolidation of Open-Source Language Resources

By leveraging the co-location with LREC, we expect to accelerate the release and documentation of several key datasets.

- **Instruction-Tuning Corpora:** We expect to release the next iteration of open-source, instruction-tuned datasets specifically for regional Spanish and Portuguese variations.
- **Interoperability:** Ensuring that new resources from the region are integrated into global repositories such as the ELRA catalogs and the Hugging Face Hub, increasing their visibility to the global scientific community.

5. Towards a common Research Agenda

The LANLP community is unified by a commitment to advancing the state of Natural Language Processing through the lens of Ibero-American linguistic and cultural specificity. Our research agenda is organized around four primary pillars that address the unique challenges of our community.

5.1. Linguistic Resource Infrastructure and Evaluation

At the core of the LANLP mission is the expansion of the digital footprint for regional languages. This includes:

- **Primary Resource Creation:** Development of high-fidelity corpora, lexicons, and multimodal annotations for the Iberian and Latin American languages, with a focus on capturing both text and speech nuances.
- **Morphosyntactic Analysis:** Specialized research into the analysis and tagging of morphologically rich and under-documented languages (e.g., Basque, Mapudungun, Bribri), utilizing frameworks such as Universal Dependencies.
- **Tailored Benchmarking:** Designing evaluation metrics and benchmarks that move beyond translated English datasets to reflect the authentic linguistic realities of our languages and cultures.

5.2. New Developments in the Era of Large Language Models

The community explores the intersection of localized needs and global modeling trends, specifically:

- **Model Efficiency and Equity:** Investigating Small Language Models (SLMs), synthetic data generation, and strategies to mitigate "language domination" and the digital scarcity affecting Ibero-American varieties.
- **Transfer Learning and Multilinguality:** Advancing cross-lingual representations and embedding methods that bridge the gap between high-resource Spanish and Portuguese and the diverse minority languages of the continent.
- **Translation and Generation:** Optimizing Machine Translation (MT) and generation for regional variations and low-resource indigenous languages like Quechua, Aymara, and Nahuatl.

5.3. Dialectology, Contact, and Domain-Specific NLP

Ibero-America provides a unique laboratory for studying language in contact and social context:

- **Dialectal Variation and Code-Switching:** Developing robust methods for identifying and handling regional dialects and complex language contact, such as Spanish-Portuguese code-mixing and Spanish-Indigenous language intersection.
- **Speech and Audio Innovation:** Tailoring ASR and TTS systems to the prosody and phonology of Latin American Spanish, Brazilian Portuguese, and indigenous oral traditions.
- **Applied Socio-NLP:** Tackling domain-specific tasks—including sentiment analysis and hate-speech detection—within the specific socio-political contexts of Latin American digital media.

5.4. Ethics, Governance, and Participatory Methods

Recognizing the deep cultural stakes of our work, the LANLP community prioritizes:

- **Participatory Research:** Implementing community-driven methods, including crowdsourcing and citizen science, to ensure data collection is representative and inclusive.
- **Data Sovereignty and Rights:** Addressing the ethics of indigenous language rights, data governance, and the sustainability of resources within the framework of fair and transparent AI.
- **Digital Humanities:** Applying NLP to historical texts and cultural heritage to preserve and analyze the vast literary legacy of the Ibero-American space.

6. Conclusions

LANLP is positioned as a networking venue explicitly designed to bridge Iberian and Latin American NLP communities, with a strong emphasis on under-resourced and minoritized languages, ethical and culturally sensitive practices, and community-driven resource development and evaluation. The overview of existing ecosystems (e.g., SEPLN, CLARIAH-ES, PROPOR and Americas-NLP) makes clear that substantial expertise, resources, and infrastructures already exist on both sides of the Atlantic, but that they remain only partially connected in terms of shared benchmarks, interoperable resource pipelines, and sustained coordination mechanisms.

A key takeaway is that the most impactful contribution of LANLP is likely to be as a *connective layer* rather than a replacement for established venues: a place where communities align on common practices for resource documentation, evaluation design, and responsible collaboration, while still preserving the identity and strengths of each network. In this sense, evaluation-oriented communities and shared-task initiatives can provide reusable protocols for reproducibility and comparability, whereas

infrastructure-oriented initiatives can ensure long-term sustainability through standards, metadata, and interoperability.

At the same time, the current community map is necessarily incomplete. Several communities and initiatives are still emerging or are not yet fully represented, and the long tail of languages and varieties in Iberia and Latin America cannot be covered by any single event. This highlights an immediate agenda for LANLP: (i) broaden participation to include additional networks, grassroots language advocates, and regionally grounded initiatives; (ii) encourage shared evaluation suites and task designs that explicitly account for dialectal variation and code-switching; and (iii) promote governance models that support ethical data collection, long-term maintenance, and transparent reuse.

Overall, LANLP can catalyze durable cross-regional collaboration by prioritizing three outcomes: interoperable language resources, evaluation and benchmarking practices tailored to low-resource realities, and sustained community links that convert short-term networking into long-term joint work.

Acknowledgments

This work has been partially supported by the Strategic Networks CLARIAH-ES (RED2024-154077-E) and RutaMdL (RED2024-154067-E) and the Ministerio para la Transformación Digital y de la Función Pública - Funded by EU – NextGenerationEU within the framework of the project "Desarrollo de Modelos ALIA". The authors also thank Aritz Farwell and Ainara Estarrona for helping with the organization of the symposium.

References

Irene Baucells, Javier Aula-Blasco, Iria de Dios-Flores, Silvia Paniagua Suárez, Naiara Perez, Anna Salles, Susana Sotelo Docio, Júlia Falcão, Jose Javier Saiz, Robiert Sepulveda Torres, Jeremy Barnes, Pablo Gamallo, Aitor Gonzalez-Agirre, German Rigau, and Marta Villegas. 2025. IberoBench: A benchmark for LLM evaluation in Iberian languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10491–10519, Abu Dhabi, UAE. Association for Computational Linguistics.

Carreras F., Estarrona A., Farwell A., Iruskieta M., Marco M., Melero M., Montejo A., Rigau G. Riaño D., Romero D., Ros S., Sánchez E., and Sousa X. 2024. Clariah-es: Strategic network for the integration in the european research infrastructures in social sciences and humanities. In *SEPLN-CEDI-PD 2024: Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations, CEUR Workshop Proceedings Vol-3729*, La Coruña, Spain. SEPLN.

María Grandury, Javier Aula-Blasco, Júlia Falcão, Clémentine Fourrier, Miguel González Saiz, Gonzalo Martínez, Gonzalo Santamaria Gomez, Rodrigo Agerri, Nuria Aldama García, Luis Chiruzzo, Javier Conde, Helena Gomez Adorno, Marta Guerrero Nieto, Guido Ivetta, Natàlia López Fuertes, Flor Miriam Plaza-del Arco, María-Teresa Martín-Valdivia, Helena Montoro Zamorano, Carmen Muñoz Sanz, Pedro Reviriego, Leire Rosado Plaza, Alejandro Vaca Serrano, Estrella Vallecillo-Rodríguez, Jorge Vallego, and Irune Zubiaga. 2025. La leaderboard: A large language model leaderboard for Spanish varieties and languages of Spain and Latin America. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32482–32524, Vienna, Austria. Association for Computational Linguistics.

An Oral-first Interactive Agentic System for Guaraní Speakers

Samantha Adorno, Akshata Kishore Moharir, Ratna Kandala
University of Kansas, Independent Researcher, University of Kansas
samantha.adorno00@gmail.com, akshatankinshore5@gmail.com, ratnanirupama@gmail.com

Abstract

Artificial intelligence systems are often presented as universal, yet their interaction paradigms remain predominantly text-first, limiting alignment with primarily oral languages and communicative practices. Using Guaraní, an official and widely spoken language of Paraguay, as a motivating case, this work examines how language support risks remaining symbolic when spoken interaction is reduced to a speech-to-text interface. We explore an oral-first, multi-agent framing in which turn-taking, repair, shared context, and governance are treated as core components of interaction rather than peripheral features. By separating language understanding from the conversation state and permission mechanisms, the architecture makes conversational structure and control explicit, enabling reasoning over interaction dynamics rather than isolated commands. Framing conversational coordination as a cognitively motivated reasoning problem over shared state connects insights from human dialogue to the design of AI systems that are more interpretable and responsive in oral and low-resource settings.

Keywords: Conversational AI, Oral-first Interaction, Low-resource Languages, Diglossia, Guaraní, Culturally Grounded AI, Multi-Agent Systems, Indigenous Data Governance

1. Introduction

Most AI systems and everyday interaction with machines remain oriented around text input, such as keyboards, menus, and form-like interfaces. Voice features are increasingly common, but in many deployments they function primarily as a spoken front-end to a text-first pipeline (transcribe, parse, respond) rather than as sustained conversation with turn-taking, clarification, and repair. Voice assistants such as Amazon Alexa illustrate this interaction style: wake word, short request, single response. Such systems frequently struggle with interruptions, response timing, and turn coordination (Skantze, 2021), and breakdown handling often places the burden on users to rephrase or restart (Alghamdi et al., 2024).

Conversation, however, is coordinated action. People manage understanding through grounding and shared context, using clarification and repair to maintain alignment (Clark and Brennan, 1991). Turn-taking is central to this coordination, with measurable cross-cultural variation in timing (Stivers et al., 2009). When systems cannot manage turn-taking and repair, voice interaction collapses into brittle command-and-control behavior, even when speech recognition and synthesis are available (Skantze, 2021; Alghamdi et al., 2024). These limitations disproportionately affect low-resource languages, many of which are primarily oral and lack extensive written or dialogue resources (Turin, 2012).

Rather than adapting oral languages to text-centric systems, oral interaction should be treated as a first-class design starting point. Orality-grounded HCI argues that assumptions imported from literate settings can fail in oral cultures, where knowledge organization relies on narrative structure, repetition, and socially distributed memory (Sherwani et al., 2009). Although recent projects have improved speech recognition for underrepresented speech communities, such as Aboriginal English in Australia (Hutchinson, 2025; The University of Western Australia, 2025), most systems remain structured around text-first interaction assumptions and lack explicit mechanisms for dialogue state tracking, repair, and consent. Prior work on Guaraní speech recognition has similarly highlighted the challenges of low-resource settings: competition results covering Guaraní alongside other Indigenous languages of the Americas show poor performance largely attributable to data scarcity (Ebrahimi et al., 2023), and targeted efforts to fine-tune models such as Whisper specifically for Guaraní further underscore the gap between off-the-shelf multilingual models and the needs of the language (Acevedo Zarza et al., 2024).

We use Guaraní as a motivating case. Guaraní is one of Paraguay's two official languages and is widely used in daily life (Organization of American States (OAS), 1992). While most speakers regularly use Guaraní, Spanish, or both (Instituto Nacional de Estadística (INE), Paraguay, 2024), digital systems continue to privilege Spanish for interaction and disambiguation. The challenge is therefore not only recognition or translation, but supporting multi-turn interaction, shared context, and repair in the language users actually speak.

To address this, we propose an oral-first assistant architecture that separates language understanding, conversation state, action execution, and governance into interacting agents. A Multi-Agent

Proceedings of LANLP: Bridging Ibero and Latin American NLP Communities, pages 9–14

System (MAS) framework enables specialization while keeping interfaces explicit and inspectable. Prior work shows that decomposing conversational task-solving into specialized agents improves performance relative to monolithic models (Becker, 2024), and that turn-taking and repair require dedicated state tracking distinct from generation or execution (Chen et al., 2025).

2. Case Study: Guaraní in Paraguay

Interface design in Paraguay operates within a context of *diglossia*: a stable separation of language functions across domains, with a "High" (H) code used in formal and written contexts and a "Low" (L) code used in everyday speech (Ferguson, 1959; Fishman, 1967). In practice, Guaraní predominates in oral and community settings, while Spanish dominates literacy, bureaucracy, and public-facing written systems (Ito, 2012). Because most digital interaction is mediated through written interfaces-menus, forms, and error messages-systems implicitly privilege Spanish literacy. Even when Guaraní is supported, interaction often defaults to Spanish at moments requiring verification or correction, creating a *domain mismatch* between everyday reasoning and digital infrastructure (Ito, 2012). Internet use in Paraguay has grown substantially, rising from 61.1% in 2017 to 81.6% in 2024 among those aged 10+ (Instituto Nacional de Estadística (INE), Paraguay, 2025). However, increased connectivity has not translated into broader Guaraní integration in digital services. Mainstream platforms remain optimized for Spanish literacy and symbolic input, meaning language support often remains nominal rather than functionally embedded in interaction. This tension between everyday oral practice and text-centric digital systems motivates the need for interaction models that treat spoken coordination, repair, and shared context as primary rather than secondary design considerations.

3. Proposed Architecture: An Oral-First Multi-Agent System

Figure 1 illustrates the proposed system architecture, showing how Guaraní voice input is routed through a pipeline of specialized agents before any action is executed.

We propose an *oral-first* architecture that treats speech as the primary interaction modality and makes multi-turn context, repair, and governance explicit. The system orchestrates six specialized agents:

- **Speech Interface Agent:** Performs Voice Activity Detection and turn segmentation, using pause duration and timing cues to distinguish floor-holding from turn completion (Skantze, 2021; Stivers et al., 2009). This is critical for Guaraní, where brief pauses and glottal stops (*puso*) can occur within words and should not be misinterpreted as turn endings (Estigarribia, 2020).

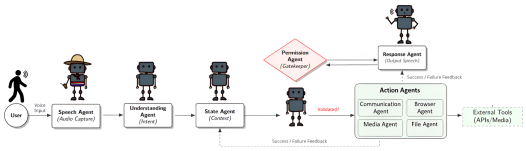


Figure 1: Proposed Oral-First Multi-Agent Architecture. Voice input is routed through speech capture, intent understanding, conversation state tracking, and a permission and governance layer before tool execution. The Speech Agent (far left) is depicted in traditional Paraguayan attire to symbolize the system's grounding in local cultural identity.

- **Guaraní Understanding Agent:** Interprets Guaraní and Jopará speech by mapping utterances to structured intents and slots (e.g., `PLAY_MUSIC`, `OPEN_TAB`). It is designed to operate on community-verified speech data to capture culturally specific references and loanword usage.
- **Conversation State Agent:** Maintains dialogue memory across turns, enabling resolution of ellipsis and implicit references.
- **Permission & Governance Agent:** Operates as a sovereign mediation layer between interpretation and execution. It evaluates requested actions against user consent settings and community-defined privacy norms (e.g., default "do not store audio") before allowing downstream execution.
- **Response Agent:** Generates conversational responses grounded in the current dialogue state, which are checked by the Permission & Governance Agent before being delivered to the user.
- **Action Agents:** Domain-specific executors (e.g., Media or Browser agents) responsible for carrying out approved actions through external tools or APIs.

Table 1 illustrates a sample multi-turn exchange, demonstrating how the agents coordinate to handle context resolution and repair across turns.

3.1. Training Data and Community Alignment

A central question for the Guaraní Understanding Agent is how to obtain authentic, community-verified speech rather than relying on synthetic

User (Guaraní)	System Action & Agent Logic
"Che ahenduse purahei" (I want to listen to music)	1. **Speech Agent** captures audio. 2. **Understanding Agent** identifies intent: `PLAY_MUSIC`. 3. **Permission Agent** verifies "Music" is a safe category; approves. 4. **Media Agent** selects a popular playlist and starts playback.
Response Agent	*"Oĩ porã"* (Ok / It is good)
"Nda che gustái" (No, I don't like [this])	1. **Understanding Agent** detects negative sentiment/rejection. 2. **Conversation State Agent** resolves the implicit object: "this" = *current song*. Updates intent to `SKIP`. 3. **Media Agent** executes `NEXT_TRACK`.

Table 1: Multi-turn interaction demonstrating context resolution and repair.

translations. Two initiatives illustrate the available resources. Mozilla Common Voice (Guaraní, `gn`) provides a crowdsourced open speech resource that supports multilingual acoustic modeling (Ardila et al., 2020), and community-led efforts such as *Aikuaa*, organized by *El Surtidor* through collaborative "mingas," further capture *Jopará* usage and conversational variation often absent from formal datasets (JournalismAI, 2025). However, both resources consist primarily of read-aloud sentences and neither contains the task-oriented or multi-turn utterances this system requires.

Addressing this gap requires a dedicated collection effort. We plan to organize community recording sessions modeled on the Aikuaa minga format, in which native speakers are prompted with realistic task scenarios (e.g., controlling media, opening a browser tab, asking follow-up questions) and asked to respond naturally in Guaraní or Jopará. Critically, prompts will be presented orally rather than as written text, to avoid literacy-based anchoring of responses. Sessions will be designed to capture: (1) *task-oriented utterances* covering the intent categories the system supports; (2) *multi-turn exchanges* that include ellipsis, implicit reference, and repair; and (3) *code-switching patterns* reflecting everyday mixed-language use. All recordings will be collected under explicit community consent protocols aligned with the Permission & Governance Agent's design, with speakers retaining control over storage and reuse of their voice data.

3.2. Evaluation Criteria

Evaluating an oral-first architecture requires metrics beyond accuracy that capture context, sentiment, and privacy in multi-turn interaction. We consider four dimensions of conversational success:

- **Task Success Rate (TSR):** Measures the percentage of multi-turn goals completed successfully, capturing whether the Conversation State Agent maintains dialogue coherence and whether the Understanding Agent correctly interprets evolving intents across turns.
- **Repair Success Rate:** Measures conversational resilience by evaluating how often the system recovers from errors, such as misheard words or misunderstood intents, without requiring users to restart their task.
- **Perceived Sovereignty:** Assesses whether users trust that their voice data remains under their control. This qualitative metric evaluates confidence that audio is not stored or reused without consent and requires community-centered, ethnographic evaluation methods.
- **Latency:** Evaluates whether system response timing aligns with Guaraní conversational tempo, avoiding both premature interruptions that violate turn-taking norms and prolonged silences that disrupt conversational flow.

4. Discussions and Limitations

The proposed architecture highlights the potential of oral-first language technology, but several challenges extend beyond technical implementation.

4.1. Standardization vs. Lived Reality

A persistent challenge for Guaraní language technology is the gap between institutional standardization and everyday speech. Although Guaraní is an official language supported by bodies such as the Academia de la Lengua Guaraní under the Ley de Lenguas (Secretaría de Políticas Lingüísticas (Paraguay), 2010; Secretaría de Políticas Lingüísticas, [n. d.]), daily use frequently involves *jopará* and code-switching (Mortimer, 2006; Estigarribia, 2015; Kellert and Tyagi, 2025). As a result, standardized forms may diverge from lived usage. Oral-first systems should therefore treat variation as expected input and prioritize communicative intent over enforcing a single normative register (Mortimer, 2006; Kellert and Tyagi, 2025).

4.2. The Data Bottleneck is Specifically Conversational

Beyond general data scarcity, oral-first systems lack conversational audio capturing turn-taking, repair, and shared context. Many endangered languages remain primarily oral and underrepresented in digital resources (Turin, 2012). For Guaraní, existing corpora and scraping-based methods mainly support text-based evaluation rather than spontaneous multi-turn speech (Chiruzzo et al., 2022; Góngora et al., 2021). Speech recognition results from the AmericasNLP shared task further confirm that performance on Guaraní lags significantly behind higher-resource languages (Ebrahimi et al., 2023), while recent ASR work fine-tuning Whisper for Guaraní demonstrates both progress and the continued need for domain-appropriate conversational data (Acevedo Zarza et al., 2024). Future efforts should prioritize community-led collection of conversational data (JournalismAI, 2025).

4.3. Governance and Perceived Control

Oral interfaces raise governance challenges because speech data is inherently identifiable and easily repurposed. Indigenous data governance frameworks emphasize community benefit, control, and accountability (Carroll et al., 2020). This motivates separating execution from a dedicated permission and privacy layer that mediates consent and data retention, including explicit "do not store audio" defaults. Such separation supports ethical commitments while increasing perceived user control and trust in multi-turn interaction (Carroll et al., 2020; Alghamdi et al., 2024).

5. Conclusion

This work contributes to culturally grounded AI by treating conversation as a core computational structure rather than an interface layer. When interaction models fail to reflect how language is practiced, language support risks remaining symbolic rather than operational. Using Guaraní as a motivating case, we outline a multi-agent architecture that elevates turn-taking, repair, and shared context to first-class system components. By separating language understanding from explicit permission and governance mechanisms, the architecture makes conversational reasoning, control, and accountability inspectable and modular. This framing moves beyond universal, text-first assumptions toward interaction models that reflect human communicative coordination and sociolinguistic reality. More broadly, the work argues that equitable and aligned AI systems must reason over conversation as it is lived, particularly in oral and low-resource settings, rather than adapting those settings to inherited text-centric paradigms.

Santiago Rubén Acevedo Zarza, Mateo Andrés Fidabel Gill, Christian Daniel von Lücken Martínez, and Diego Pedro Pinto Roa. 2024. Desarrollo de un sistema de reconocimiento del habla en guaraní: Evaluación de variantes del modelo Whisper y técnicas de mejora de datos. *JAIIO, Jornadas Argentinas de Informática* 10, 1 (2024), 158–166. doi:10.1145/3641234

Essam Alghamdi, Martin Halvey, and Emma Nicol. 2024. System and User Strategies to Repair Conversational Breakdowns of Spoken Dialogue Systems: A Scoping Review. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces* (Luxembourg, Luxembourg) *(CUI '24)*. Association for Computing Machinery, New York, NY, USA, Article 28, 13 pages. doi:10.1145/3640794.3665558

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 4218–4222. https://aclanthology.org/2020.lrec-1.520/

Evan Becker. 2024. Multi-Agent Large Language Models for Conversational Task-Solving. arXiv preprint arXiv:2410.22932v1. https://arxiv.org/abs/2410.22932

Stephanie Russo Carroll et al. 2020. The CARE Principles for Indigenous Data Governance. *Data Science Journal* (2020).

Siyuan Chen et al. 2025. Multi-Party Conversational Agents: A Survey. arXiv preprint arXiv:2505.18845v1. https://arxiv.org/abs/2505.18845

Luis Chiruzzo, Santiago Góngora, Aldo Alvarez, Gustavo Giménez-Lugo, Marvin Agüero-Torales, and Yliana Rodríguez. 2022. Jojajovai: A Parallel Guarani-Spanish Corpus for MT Benchmarking. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache,

Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 2098–2107. https://aclanthology.org/2022.lrec-1.226/

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in Communication. In *Perspectives on Socially Shared Cognition*. American Psychological Association. https://www.cs.cmu.edu/~illah/CLASSDOCS/Clark91.pdf

Abteen Ebrahimi, Manuel Mager, Adam Wiemerslage, Pavel Denisov, Arturo Oncevay, Danni Liu, Sai Koneru, Enes Yavuz Ugan, Zhaolin Li, Jan Niehues, Monica Romero, Ivan G. Torre, Tanel Alumäe, Jiaming Kong, Sergey Polezhaev, Yury Belousov, Wei-Rui Chen, Peter Sullivan, Ife Adebara, Bashar Talafha, Alcides Alcoba Inciarte, Muhammad Abdul-Mageed, Luis Chiruzzo, Rolando Coto-Solano, Hilaria Cruz, Sofía Flores-Solórzano, Aldo Andrés Alvarez López, Ivan Meza-Ruiz, John E. Ortega, Alexis Palmer, Rodolfo Zevallos, Kristine Stenzel, Thang Vu, and Katharina Kann. 2023. Findings of the Second AmericasNLP Competition on Speech-to-Text Translation. In *Proceedings of the NeurIPS 2022 Competitions Track (Proceedings of Machine Learning Research, Vol. 220)*. PMLR, 217–232. https://proceedings.mlr.press/v220/ebrahimi23a.html

Bruno Estigarribia. 2015. Jopará and Guaraní in Paraguay (discussion of contact and mixed speech). Cited in Guaraní NLP work as evidence for contact-driven variation..

Bruno Estigarribia. 2020. *A Grammar of Paraguayan Guaraní*. UCL Press. https://uclpress.co.uk/book/a-grammar-of-paraguayan-guarani/

Charles A. Ferguson. 1959. Diglossia. *Word* 15, 2 (1959), 325–340. doi:10.1080/00437956.1959.11659702

Joshua A. Fishman. 1967. Bilingualism with and without Diglossia; Diglossia with and without Bilingualism. *Journal of Social Issues* 23, 2 (1967), 29–38. doi:10.1111/j.1540-4560.1967.tb00573.x

Santiago Góngora, Nicolás Giossa, and Luis Chiruzzo. 2021. Experiments on a Guarani Corpus of News and Social Media. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, Manuel Mager, Arturo Oncevay, Annette Rios, Ivan Vladimir Meza Ruiz, Alexis Palmer, Graham Neubig, and Katharina Kann (Eds.). Association for Computational Linguistics, Online, 153–158. doi:10.18653/v1/2021.americasnlp-1.16

Ben Hutchinson. 2025. A partnership with The University of Western Australia to improve speech technology for Aboriginal and Torres Strait Islander people's voices. Google Australia Blog. https://blog.google/intl/en-au/company-news/technology/a-partnership-to-improve-speech-technology-for-first-nations-voices/

Instituto Nacional de Estadística (INE), Paraguay. 2024. Día Internacional de la Lengua Materna: Diversidad lingüística en Paraguay. https://www.ine.gov.py/noticias/2298/dia-internacional-de-la-lengua-materna-diversidad-linguistica-en-paraguay Household language-use reporting based on EPHC 2023. Accessed 2026-02-02.

Instituto Nacional de Estadística (INE), Paraguay. 2025. 8 de cada 10 personas utiliza internet en Paraguay (EPH 2017–2024). Reports 81.6% internet use among population aged 10+ in 2024, up from 61.1% in 2017; includes noted exclusions. Accessed 2026-02-02.

Hiroshi Ito. 2012. With Spanish, Guaraní lives: a sociolinguistic analysis of bilingual education in Paraguay. *Multilingual Education* 2, 1 (2012), 6. doi:10.1186/2191-5059-2-6

JournalismAI. 2025. Guarani AI: When building language tech means building community. https://www.journalismai.info/blog/5fcm6ayykhqq7564kbvt9nw92wwmy9 Documents community "mingas" and Guaraní audio dataset efforts. Accessed 2026-02-02.

Olga Kellert and Nemika Tyagi. 2025. Where and How Do Languages Mix? A Study of Spanish-Guaraní Code-Switching in Paraguay. In *Proceedings of the Workshop on Computational Approaches to Linguistic Code-Switching*. Association for Computational Linguistics.

Katherine Mortimer. 2006. Guaraní Académico or Jopará? Educator Perspectives and Ideologies on Language in Paraguay. Often cited for discussions of standardization, literacy, and ideologies around Guaraní/Spanish mixing..

Organization of American States (OAS). 1992. Paraguay's Constitution of 1992 with Amendments through 2011. PDF. https://www.oas.org/ext/Portals/33/Files/Member-States/Parag_intro_textfun_eng_1.pdf English translation; consolidated text with amendments through 2011.

Secretaría de Políticas Lingüísticas. [n. d.]. Academia de la Lengua Guaraní. https://spl.gov.py/es/academia-de-la-lengua-guarani/.

Secretaría de Políticas Lingüísticas (Paraguay). 2010. Ley Nº 4251/2010: Ley de Lenguas (texto bilingüe). PDF. https://spl.gov.py/files/legal/Ley%204251%20-%20bilingue.pdf

Jahanzeb Sherwani, Nosheen Ali, Carolyn Penstein Rosé, and Roni Rosenfeld. 2009. Orality-Grounded HCID: Understanding the Oral User. *Information Technologies & International Development* 5, 4 (2009), 37–49. https://itidjournal.org/index.php/itid/article/download/422/422-1096-2-PB.pdf

Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language* 67 (2021), 101178. doi:10.1016/j.csl.2020.101178

Tanya Stivers, N. J. Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, G. Hoymann, Federico Rossano, Jan P. de Ruiter, Kyung-Eun Yoon, and Stephen C. Levinson. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences* 106, 26 (2009), 10587–10592. doi:10.1073/pnas.0903616106

The University of Western Australia. 2025. First Nations people to benefit from inclusive technology partnership. UWA News. https://www.uwa.edu.au/news/article/2025/february/first-nations-people-to-benefit-from-inclusive-technology-partnership

Mark Turin. 2012. Voices of vanishing worlds: Endangered languages, orality, and cognition. *Análise Social* 205, 47 (2012). https://www.researchgate.net/publication/262778986_Voices_of_vanishing_worlds_Endangered_languages_orality_and_cognition

AI-TraLow: AI-Driven Translation for Low-Resource Languages and Cultures

Antoni Oliver[†], **Maite Melero**[‡], **Felipe Sánchez-Martínez**[◇], **Víctor M. Sánchez-Cartagena**[◇]
[†]Universitat Oberta de Catalunya (UOC),
[‡]Barcelona Supercomputing Center (BSC), [◇]Universitat d'Alacant (UA)
aoliverg@uoc.edu, maite.melero@bsc.es, {fsanchez,vm.sanchez}@ua.es

Abstract

In this paper, we present AI-TraLow, a project dedicated to advancing AI-driven translation for low-resource languages and cultures. The research is structured around three primary objectives: firstly, the development of advanced data curation techniques designed to refine parallel corpora and detect machine-generated content; secondly, the exploration of integrating linguistic resources—such as dictionaries and grammatical rules—directly into model prompts and fine-tuning techniques to enhance translation precision; and thirdly, the mitigation of hardware constraints through knowledge distillation to produce efficient models viable for standard desktop environments. By targeting specific linguistic groups, including Iberian varieties (Aranese, Aragonese, Asturian and Eonavian), Mayan languages, and languages of vulnerable migrant communities, AI-TraLow seeks to foster linguistic diversity and digital inclusion. Ultimately, this initiative delivers open-source tools and models that ensure cultural heritage is both preserved and accessible within the contemporary digital landscape.

Keywords: Large Language Models, Machine Translation, Low-Resource Languages, Data Curation, Linguistic Integration, Knowledge Distillation

1. Introduction

The rapid advancement of Large Language Models (LLMs) has fundamentally transformed the landscape of Machine Translation (MT). Despite these breakthroughs, languages with limited digital resources continue to lag behind due to the heavy data dependency of modern architectures. The AI-TraLow project (*AI-Driven Translation for Low-Resource Languages and Cultures*) was established to tackle these disparities by hypothesizing that the emergent few-shot capabilities and multimodal nature of LLMs can be leveraged to bridge the resource gap, even when traditional parallel data is scarce.

AI-TraLow is structured as a coordinated project with a total duration of three years, having officially launched on September 1, 2025. The initiative is powered by a consortium of three institutions, each leading a specialized subproject:

- **Universitat d'Alacant (UA):** is the coordinating institution and leads the subproject *Curation and Exploitation of Heterogeneous Resources for Translating Low-resource Languages* (CEHR-TraLowLa). This unit focuses on data acquisition, curation, and the integration of linguistic information into LLM.
- **Barcelona Supercomputing Center (BSC):** Responsible for the subproject *Multimodal Large Language Models for Translating Low-Resource Languages* (MLLM4TRA). Their research centers on building multimodal models and developing strategies for adapting LLMs to low-resource scenarios.
- **Universitat Oberta de Catalunya (UOC):** Manages the subproject *Large Language Models for Translating Low-Resource Romance Languages of the Iberian Peninsula*. This team focuses on the development of linguistic resources and the rigorous evaluation of translation systems.

Ultimately, AI-TraLow aims to empower marginalized linguistic communities—ranging from the Iberian Peninsula (Aranese, Aragonese, Asturian, Eonavian) to Mayan varieties and migrant languages such as Wolof or Amazigh—ensuring they remain viable and visible in the contemporary digital era.

2. Related Work

The AI-TraLow project is situated at the intersection of LLMs and MT for low-resource scenarios. The following areas represent the current scientific landscape and the specific gaps this initiative aims to address.

2.1. LLMs in Machine Translation

Recent advancements in LLMs, particularly those based on the Transformer decoder-only architecture, have demonstrated remarkable capabilities in translation tasks (Alves et al., 2024; Xu et al., 2024; Yang et al., 2023). However, as highlighted in the project's rationale, these models often suffer from the "curse of multilinguality," where model capacity

Proceedings of LANLP: Bridging Ibero and Latin American NLP Communities, pages 15–20

is diluted across numerous languages, leading to suboptimal performance for those with limited digital presence. AI-TraLow builds upon the potential of instruction tuning (Alves et al., 2024) to specialize these models for neglected linguistic varieties.

2.2. Data Curation and Synthetic Content

A primary bottleneck for low-resource MT is the prevalence of noisy and machine-translated content in web-crawled data. This project will create high-quality datasets-encompassing text and, where applicable, image data, refined through state-of-the-art filtering and curation tools. Because the increasing presence of machine-generated text on the web poses a risk of "model collapse," a central focus of our work is the development of robust classifiers to distinguish between human-authentic and synthetic content. By ensuring that datasets are representative of human-authored language, and by employing innovative LLM-based approaches to clean existing parallel corpora, we will establish a reliable foundation for advancing translation technologies in data-scarce scenarios.

2.3. Integration of Heterogeneous Linguistic Resources

While current LLMs rely heavily on massive datasets, they often underutilize high-quality symbolic resources. Research has shown that providing in-context linguistic descriptions can significantly aid the learning of endangered languages. AI-TraLow extends this by exploring a hybrid methodology: integrating linguistic resources—such as dictionaries and transfer rules from *Apertium* (Forcada et al., 2011)—directly into both model prompts and fine-tuning techniques to teach models "unseen" languages on the fly (Tanzer et al., 2024; Zhang et al., 2024a,b; Hus and Anastasopoulos, 2024; Elsner et al., 2024).

2.4. Multimodal and Efficient Modeling

The state of the art in multimodality suggests that moving away from traditional subword tokenization can benefit languages with non-standardized orthographies. By employing pixel-based models (Salesky et al., 2024; Caglayan et al., 2019) and byte-to-byte architectures (Choe et al., 2019; Xue et al., 2022; Clark et al., 2022), AI-TraLow addresses the limitations of fixed vocabularies. Finally, to ensure the usability of these models in standard desktop environments, the project draws on knowledge distillation and efficiency strategies to mitigate the hardware constraints typically associated with large-scale models.

3. Objectives, methodology, and work plan

3.1. Objectives

The main objective of this project is to advance the development of MT systems for low-resource languages using decoder-only LLMs, thereby enabling these languages to join the wave of adoption of LLMs, specifically for MT. This global objective can be broken down into several specific objectives:

O1 Obtain and curate high-quality resources for low-resource languages.

O2 Leverage linguistic information to improve the translation of low-resource languages with LLMs.

O3 Build large language models tailored for translating a subset of the Mayan languages.

O4 Develop advanced methods for image-to-text translation of low-resource languages with multimodal LLMs.

O5 Develop advanced strategies for adapting LLMs for translating low-resource languages

O6 Build multimodal LLMs for the translation of low-resource languages spoken by migrants in vulnerable situation.

O7 Develop advanced methods for the automatic enhancement of existing linguistic resources used in rule-based MT and the generation of synthetic data with them.

O8 Build LLMs for the translation of low-resource languages of the Iberian Peninsula.

O9 Evaluation and comparison of encoder-decoder translation models and decoder-only models for translating low-resource languages of the Iberian Peninsula.

3.2. Methodology

The project is structured around a three-phase iterative cycle, which will be executed twice throughout its duration. These phases comprise: (i) data compilation and curation, (ii) research into the training and fine-tuning of LLMs for MT, and (iii) the release of the high-performance models.

During the *data compilation* phase, the consortium will develop new corpora for the target languages while simultaneously refining existing datasets. The generation of new resources will involve diverse methodologies, including document scanning and optical character recognition, alongside targeted web crawling. To safeguard data integrity, the project will develop techniques to detect and filter machine-generated or automatically translated content. Additionally, this stage encompasses the acquisition of visual data to support

image-to-text translation tasks and the creation of novel evaluation benchmarks tailored to the languages under study.

The second phase is dedicated to the exploration of LLM training and fine-tuning strategies for translation. The research team will employ established methodologies while remaining adaptable to emerging techniques within the field. A key focus will be investigating the integration of linguistic resources (e.g. dictionaries and grammars) with monolingual and parallel corpora. Furthermore, we will assess the impact of synthetic data on model performance and the efficacy of multimodal inputs in mitigating translation ambiguity. To this end, a specialized task is devoted to the development of multimodal LLMs. Throughout this process, models will be continuously monitored using a suite of automatic evaluation metrics.

In the final phase, the most effective models from the preceding stage will undergo rigorous human evaluation to ensure translation quality. Once validated, all models and tools will be released to the public under open-source licenses.

3.3. Work plan

The project is structured into five interconnected work packages (WPs). **WP1** is dedicated to the acquisition and curation of linguistic resources, with a specific emphasis on developing automated mechanisms to detect machine-generated or synthetically translated content, thereby ensuring the integrity of the training data. **WP2** explores the integration of linguistic resources (dictionaries, translation rules, grammar books) to augment the translation capabilities of LLMs in low-resource scenarios. **WP3** investigates multimodal architectures to facilitate text translation directly from visual inputs, a strategy aimed at mitigating contextual disambiguation challenges in MT. **WP4** focuses on the design of robust strategies for training and adapting LLMs; this includes advanced fine-tuning methodologies, the generation of high-quality synthetic corpora, and the distillation of knowledge from massive models into compact, efficient encoder-decoder architectures. Finally, **WP5** will apply the methodologies and tools derived from previous WPs to the development and release of optimized LLMs for a targeted subset of languages, thereby maximizing the project's impact and fostering digital inclusion within the respective linguistic communities.

WP1. Data acquisition and curation

This work package focuses on supplying resources for training and fine-tuning LLMs for low-resource language pairs. We will pursue several lines of work: (i) *corpus generation:* scanning existing books in extremely low-resource languages and conducting optical character recognition; and gathering text and image data in low-resource languages; (ii) *data quality enhancement:* researching methods for the automatic detection of text generated or translated automatically, and utilizing LLMs to clean existing parallel corpora; and (iii) *induction of linguistic resources:* developing methods for automatically generating linguistic resources that can be used later for augmenting training data (WP4) or for in-context learning (WP2).

WP2. Leveraging linguistic information in LLMs for translation

Low-resource languages often lack the substantial amounts of monolingual and bilingual data required to train competitive MT systems. Leveraging explicit linguistic information presents a promising approach to enhancing translation quality for these languages. In this WP, we will focus on exploiting underutilized resources, such as components from existing rule-based MT systems and monolingual dictionaries. Additionally, we will develop advanced methods to integrate a wide range of linguistic resources, including grammar books, into LLMs, and perform an analysis of the semantic representations to better understand which resources are useful and why.

WP3. Image-to-text translation of low-resource languages with multimodal LLMs

Low-resource languages frequently face tokenization challenges and are underrepresented in state-of-the-art LLMs (Petrov et al., 2024), resulting in significant performance disparities when accurately representing all languages (Ali et al., 2023). Pixel-based models have shown potential in overcoming tokenization challenges and improving cross-lingual transfer. Resolving ambiguous situational translations is a practical application of pixel-based models that involves translating source sentences that are directly influenced by the image given as input to the model. This work package focuses on: (i) investigating the suitability of pixel-based translation models for low-resource languages, and (ii) improving the reasoning abilities of multimodal LLMs for dealing with ambiguous situational translations that require visual cues to resolve it.

WP4. Strategies for training and adapting LLMs for translating low-resource languages

This WP focuses on developing and evaluating advanced methodologies to optimize LLMs for effective machine translation in data-scarce scenarios. A key priority is the development of gender-inclusive MT systems. Given that low-resource datasets are often sourced from uncurated web content,

they frequently reflect and amplify historical gender biases. Following the project's ethical roadmap, this WP will investigate fine-tuning techniques and debiasing prompts to ensure that translations do not perpetuate harmful stereotypes. Additionally, this WP explores an array of innovative, scalable, and resource-efficient solutions—such as knowledge distillation into compact architectures—that will then be applied in WP5

WP5. Building models for translating low-resource languages

In this WP, we will develop LLMs for translating a subset of the languages of interest to the project. We will apply, when possible, the methods developed and the lessons learnt in the previous WPs.

The resulting models will undergo a comprehensive evaluation, where they will be compared against encoder-decoder models in terms of error typology and characteristics of the produced translations. This comparison will determine whether training/fine-tuning encoder-decoder systems or LLMs is more effective for the languages of interest, and which strategies work best. The findings will guide future efforts.

4. Resources released during the project

The project focuses on three primary low-resource language groups: Mayan languages,[1] Romance languages of the Iberian Peninsula (Aranese, Aragonese, Asturian, and Eonavian), and languages spoken by migrant communities in Spain (Wolof, Amazigh, and Pashto). The main objective is to develop and release LLM-based translation models for select languages within each of these clusters.

In addition to these models, the project will produce the following group-specific resources:

Mayan languages

- Digitized dictionaries and descriptive grammars.
- Curated corpora of literary and web-crawled text.
- Translations of the FLORES+ (Goyal et al., 2022) evaluation dataset into K'iche', Kaqchikel, Q'eqchi', and Mam.

[1]The following languages will be addressed: Achi, Awakateko, Ch'orti', Chuj, Kaqchikel, Itza', Ixil, Jakalteko, Q'eqchi', Q'anjob'al, Akateko, Mam, Mopan, Poqomam, Poqomchi', K'iche', Sipakapense, Sakapulteko, Tekiteko, Tzeltal, Tz'utujil, Uspanteko, and Yucatec Maya.

Romance languages of the Iberian Peninsula

- Expanded Apertium lexical resources.
- Enhanced versions of the FLORES+ evaluation dataset.
- New translations for the NTREX (Federmann et al., 2022) and WMT24++ evaluation benchmarks.

Migrant community languages

- Multimodal image-text datasets designed specifically for pixel-based translation models.

5. Current Status

Although AI-TraLow officially launched in September 2025, several key milestones have already been achieved across the project's work packages:

- WP1. Data Acquisition: We have compiled and started the preprocessing of a comprehensive set of monolingual and bilingual Apertium dictionaries for Asturian, Aragonese, and Aranese. Simultaneously, we have completed the scanning and OCR of 29 Mayan dictionaries. We have developed a new precise method for automatically detecting machine translated text (García-Romero et al., 2025) and studied the ability of bilingual speakers to distinguish between human and machine translated text (García-Romero et al., 2026). To ensure high-quality evaluation, the human translation of the following datasets is currently underway: FLORES+, NTREX, and WMT24++ into Asturian, Aragonese, Aranese, and Eonavian; and FLORES+ into K'iche', Kaqchikel, Q'eqchi' and Mam.
- WP2. Integration of Linguistic Information: the challenges of adding dictionaries to the prompt have been deeply analysed and methods based on *Group Relative Policy Optimization* to inject terminology have been explored (Garcia Gilabert et al., 2025).
- WP3. Multimodality: preliminary experiments have established initial baselines for emoji-based disambiguation in vision-language models, providing a foundation for pixel-based translation tasks.
- WP4. Training Strategies: we devised and extensively evaluated methods for knowledge distillation leveraging multiple translations from the initial (teacher) model (Galiano-Jiménez et al., 2025).

6. Conclusions

The AI-TraLow project represents a strategic effort to ensure that low-resource languages are not left behind in the current era of Large Language Models. Beyond the technical development of translation systems, this initiative establishes a scalable paradigm for linguistic preservation by treating linguistic knowledge and multimodal signals as essential anchors for neural architectures.

Our approach shifts the focus from purely data-driven methods to a hybrid model where AI-driven curation and symbolic expertise safeguard the integrity of minority languages. Ultimately, AI-TraLow is committed to the principles of Open Science. By releasing high-performance, efficient models and curated tools, we aim to empower local communities and researchers, ensuring that cultural heritage and linguistic diversity can actively flourish and remain visible within the contemporary digital landscape.

Acknowledgements

Coordinated project funded by the Spanish Ministry of Science, Innovation and Universities (MICIU), the Spanish Research Agency and the European Regional Development Fund (FEDER) through R+D+i project grants PID2024-158157OB-C31, PID2024-158157OB-C32 and PID2024-158157OB-C33.

Bibliographical References

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Schulze Buschhoff, Charvi Jain, Alexander Arno Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, Stefan Kesselheim, and Nicolas Flores-Herr. 2023. Tokenizer choice for LLM training: Negligible or crucial?

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, Volume 1 (Long and Short Papers)*, pages 4159–4170, Minneapolis, Minnesota.

Dokook Choe, Rami Al-Rfou, Mandy Guo, Heeyoung Lee, and Noah Constant. 2019. Bridging the gap for tokenizer-free language models.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Micha Elsner et al. 2024. Shortcomings of llms for low-resource translation: Retrieval and understanding are both the problem. In *Proc. of the Ninth Conference on Machine Translation*, pages 1332–1354.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proc. of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25:127–144.

Aarón Galiano-Jiménez, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, and Víctor M Sánchez-Cartagena. 2025. Multi-hypothesis distillation of multilingual neural translation models for low-resource languages. *arXiv preprint arXiv:2507.21568*.

Javier Garcia Gilabert, Carlos Escolano, Xixian Liao, and Maite Melero. 2025. Terminology-constrained translation from monolingual data using GRPO. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1335–1343, Suzhou, China. Association for Computational Linguistics.

Cristian García-Romero, Miquel Esplà-Gomis, and Felipe Sánchez-Martínez. 2025. Automatic machine translation detection using a surrogate multilingual translation model. *arXiv preprint arXiv:2511.02958*.

Cristian García-Romero, Miquel Esplà-Gomis, and Felipe Sánchez-Martínez. 2026. When translations surprise: Human awareness of predictability in translations. In *Proceedings of the 15th biennial Language Resources and Evaluation Conference (in press)*, Palma de Mallorca, Spain.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Jonathan Hus and Antonios Anastasopoulos. 2024. Back to school: Translation using grammar books. In *Proc. of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20207–20219, Miami, Florida, USA.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2024. Language model tokenizers introduce unfairness between languages. *Advances in Neural Information Processing Systems*, 36.

Elizabeth Salesky, Philipp Koehn, and Matt Post. 2024. Benchmarking visually-situated translation of text in natural images. In *Proc. of the Ninth Conference on Machine Translation*, pages 1167–1182, Miami, Florida, USA.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages.

Chen Zhang, Xiao Liu, Jiuheng Lin, and Yansong Feng. 2024a. Teaching large language models an unseen language on the fly.

Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024b. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15654–15669, Bangkok, Thailand.

OpenCor: Latin American and Iberian Languages Open Corpora Forum

Livy Real[1,2], **Valeria de Paiva**[3]
[1] Instituto Kunumi, Belo Horizonte, Brazil
[2] Instituto de Computação – Universidade Federal do Amazonas, Manaus, Brazil
[3] Topos Institute, Berkeley, USA
livy@kunumi.com, valeria@topos.institute

Abstract

The availability of open resources and corpora is a fundamental requirement for research in Natural Language Processing (NLP) and Computational Linguistics; however, languages spoken in Latin America and the Iberian Peninsula, particularly indigenous, minority, and regional varieties, remain structurally under-resourced and under-represented. This paper presents a historical account of OpenCor (Latin American and Iberian Languages Open Corpora Forum), a community-driven initiative created to promote, document, and discuss open corpora and lexical resources for these languages. Conceived as a collaborative forum rather than a competitive evaluation venue, OpenCor focuses on data creation, licensing practices, sustainability, and community building. Between 2018 and 2024, OpenCor was organized as a recurring workshop co-located with major conferences, fostering dialogue across countries, institutions, and linguistic traditions. By documenting the initiative's motivations, organizational trajectory, submission trends, and the diversity of resources presented, this paper aims to preserve institutional memory, highlight the often-invisible labor of corpus development, and provide a reference for future initiatives dedicated to openness and linguistic diversity.

Keywords: Language Resources, Corpora, Latin American Languages, Iberian Languages

1. Introduction

The availability of open corpora and lexical resources is a central requirement for the development of Natural Language Processing (NLP) and Computational Linguistics research. While major languages benefit from a growing ecosystem of openly licensed datasets, tools, and benchmarks, languages spoken in Latin America and the Iberian Peninsula have historically faced structural challenges related to visibility, funding, infrastructure, and coordination. These challenges are particularly acute for under-resourced languages, regional varieties, and languages with limited institutional support.

In this context, **OpenCor** (Latin American and Iberian Languages Open Corpora Forum) emerged as a community-driven initiative aimed at mapping, discussing, and promoting open corpora and lexical resources for these languages. Rather than focusing on system comparison or shared tasks, OpenCor was conceived as a forum for resource presentation, discussion of licensing practices, sustainability, and community building. Between 2018 and 2024, OpenCor was organized as a recurring workshop co-located with major conferences in the field in both Latin America and the Iberian Peninsula, fostering dialogue across countries, institutions, and linguistic traditions.

This paper presents a historical report of the OpenCor initiative, covering its motivations, organizational trajectory, meetings, and the evolution of the resources discussed within the forum. By documenting this history, we aim to preserve institutional memory, highlight community efforts that are often invisible in traditional evaluation venues, and provide a reference for future initiatives focused on openness and linguistic diversity. The project maintains a public web presence at https://opencor.gitlab.io/.

This paper makes four principal contributions. First, it presents a documented historical timeline of OpenCor from 2018 to 2024. Second, it provides an analysis of submission patterns and participation trends across editions. Third, it offers a curated overview of the open corpora showcased through the initiative, highlighting their linguistic, methodological, and regional diversity. Finally, it reflects on the structural challenges involved in sustaining open linguistic infrastructures, situating OpenCor within broader debates about research recognition, funding asymmetries, and long-term maintenance.

2. OpenCor Motivation

The primary motivation behind OpenCor is the absence of a consolidated, openly accessible list of corpora and lexical resources for Latin American and Iberian languages. Although many resources exist, information about them is fragmented across personal webpages, institutional repositories, or conference proceedings. This fragmentation makes it difficult for researchers, students, and practitioners to discover existing datasets, assess their licensing conditions, and reuse them responsibly. The initiative is inspired by

Proceedings of LANLP: Bridging Ibero and Latin American NLP Communities, pages 21–28

Linguateca[1], established in 1998 as a distributed resource center for the computational processing of Portuguese, with a mission centered on facilitating access to existing resources, collaboratively developing new ones, and organizing shared evaluation efforts (Santos, 2003).

The name *OpenCor* was chosen with a deliberate double meaning. It refers to *Open Corpora*, the central focus of the initiative, but also to *cor*, the Latin word for *heart*. This second meaning serves as a metaphor for openness, generosity, and inclusion. A common saying across the Iberian Peninsula and Latin America describes something as 'as big as a mother's heart, where there is always room for one more'. Historically, many workshops and shared tasks in NLP prioritize system performance and competition, which can inadvertently marginalize discussions about data creation, annotation labor, licensing, and long-term maintenance. OpenCor was designed to be: first, a list of open resources; second, a complementary space, a forum explicitly dedicated to resources rather than models, and to collaboration rather than competition.

Another important motivation was the production of a historical timeline of open resources and community efforts. By tracking meetings, submissions, and presented corpora over time, OpenCor functions as a living record of how openness has evolved in the regions. The OpenCor website was therefore not only an organizational hub but also a public archive documenting editions, programs, calls for papers, and links to resources.

The initial discussions that led to OpenCor took place during a Linguistic Data Consortium (LDC)[2] meeting in Mexico City, sponsored by the University of Pennsylvania, in 2018. At that moment, participants reflected on the structural asymmetry in the production and consolidation of linguistic infrastructures: although many corpora and initiatives concerned Latin American and Iberian languages, centralized documentation and resource aggregation were often located outside the regions where these languages are spoken. This observation reinforced the need for a regionally grounded forum dedicated to mapping, documenting, and sustaining open resources from within the community itself. OpenCor is also motivated by the recognition that many researchers in Latin America and the Iberian Peninsula face significant funding constraints. The workshop consistently aimed to lower barriers to participation by allowing "workshop-only" registration, using free submission platforms, and inviting speakers who were already attending the host conference or able to participate without additional financial support. Finally, OpenCor meetings always allowed for 'non-archival' submissions, aiming to gather the community that had already published their works in another venue but were interested in discussing and sharing their resources.

OpenCor focuses on corpora that are freely accessible and reusable, thereby supporting transparency and cumulative research. It prioritizes resources relevant to the languages spoken in Latin America and the Iberian Peninsula, including indigenous and minority languages that are often underrepresented in mainstream NLP research. At the same time, it seeks to engage a broad scholarly community, encompassing researchers in NLP, linguistics, education, and related fields.

The initiative serves two complementary purposes:

1. As a **curated, evolving list of open corpora**, improving discoverability and reuse of existing resources;
2. As a **forum and meeting point** for corpus creators and users, providing space for discussion of methodologies, annotation practices, and long-term maintenance issues.

3. OpenCor Development

OpenCor has been active since the late 2010s and has been organized as a recurring forum, often co-located with established conferences and workshops in Computational Linguistics and NLP. By 2024, the project had reached its fifth edition, reflecting sustained community interest and ongoing relevance.

OpenCor has evolved from a small, informal effort into a recognized venue for presenting and discussing corpus-related work that often does not fit comfortably within traditional publication formats.

Across all editions, OpenCor was organized as a community-driven workshop with a strong commitment to openness and accessibility. EasyChair was consistently used as the submission and review platform due to its free availability for small meetings. Invited speakers were selected based on their contributions to open resources and their ability to participate without dedicated travel funding.

The first OpenCor workshop[3] was held in 2018 in Canela, Brazil, co-located with PROPOR, the International Conference on Computational Processing of Portuguese. The workshop was organized following an invitation from Prof. Aline Villavicencio and established the core format for subsequent editions. Of the eight accepted papers, three of them (almost half) were not presented due to a lack of funding, a challenge that would recur in later editions. The invited speaker for this first edition was

[1] https://linguateca.pt/

[2] https://www.ldc.upenn.edu/

[3] https://opencor.gitlab.io/opencor-2018/

Year	Event	Location	Format	Submitted	Accepted	Keynote
2018	PROPOR	Canela (BR)	In-person	10	8	A. Medina Urrea
2019	PLAGAA	Guanajuato (MX)	In-person	9	7	Fernanda López
2020	PROPOR	Évora (PT)	Hybrid	9	7	Marcos Garcia
2021	STIL	Online	Online (full day)	10	10	Valeria de Paiva
2024	PROPOR	Santiago (ES)	In-person	11	7	–

Table 1: Conference statistics by year

Alfonso Medina Urrea (El Colegio de México, Red Temática de Tecnologías del Lenguaje, Mexico), who presented on Mexican corpora and their applications (Medina Urrea, 2018). Importantly, funding from PROPOR itself enabled the participation of the invited speaker, making this the only edition of OpenCor with direct financial support for this purpose.

In 2019, OpenCor[4] was held in Guanajuato, Mexico, as part of the 4th PLAGAA, *Taller Mexicano de Detección de Plagio y Análisis de Autoría*[5]. In this edition Adrian Pastor López Monroy (Centro de Investigación en Matemáticas, Mexico) contributed as the local chair. Among the submissions, two focused on Portuguese and five on Spanish or other languages. The invited speaker was Fernanda López (Universidad Nacional Autónoma de México), who presented CLOE (*Corpus de Lengua Oral del Español*) (López-Escobedo and Solorzano-Soto, 2016).

The 2020 edition[6] took place in Évora, Portugal, co-located with PROPOR. The invited speaker was Marcos Garcia (Universidade de Santiago de Compostela, Galicia), who presented an overview of open language resources for Galician (García and Crespo-Otero, 2022). This edition marked an important moment in connecting resource-building efforts across the Iberian Peninsula with those in Latin America, particularly among closely related Ibero-Romance languages such as Portuguese and Galician.

Originally planned to take place in Fortaleza, Brazil, OpenCor 2021[7] was held fully online due to the COVID-19 pandemic and co-located with STIL, the *Symposium in Information and Human Language Technology*. For the first time, OpenCor was held as a full-day workshop. The invited speaker was Valeria de Paiva (Topos Institute), who spoke about open linguistic resources for Brazilian Portuguese (Rademaker et al., 2017; de Paiva et al., 2016). The online format enabled broader participation across countries, partially mitigating travel-related funding barriers, while also introducing new organizational challenges typical of virtual events.

After a hiatus, OpenCor returned in 2024 (OpenCor at PROPOR 2024). Due to the density of the main conference program, no invited speaker was scheduled. The continued interest, reflected in the number of submissions, demonstrated the sustained relevance of the forum.

3.1. OpenCor Trends

Across its editions from 2018 to 2024, OpenCor received a total of 49 paper submissions, of which 39 were accepted, resulting in an overall acceptance rate of approximately 80%. Acceptance rates varied across years, ranging from 64% in 2024 to 100% in 2021, the latter reflecting the inclusive character of the fully online edition held during the COVID-19 pandemic.

Rather than reflecting low selectivity, these acceptance rates are indicative of the workshop's curatorial and community-building role. Submissions were evaluated primarily on their contribution to the documentation, creation, or dissemination of open corpora and lexical resources, with an emphasis on clarity, reusability, and licensing transparency. In several editions, a non-negligible number of accepted papers could not be presented due to travel and funding constraints, particularly in in-person events, underscoring structural barriers faced by researchers working on language resources in Latin America and the Iberian Peninsula.

The relatively stable number of submissions across editions further suggests sustained community interest, despite the absence of dedicated funding, travel grants, or shared-task incentives. These trends support the interpretation of OpenCor as a long-term community infrastructure effort, rather than a competitive evaluation venue.

Despite the geographic scale and academic diversity of both focused regions, the absolute number of submissions to OpenCor has remained relatively modest across editions. This should not be interpreted as a lack of interest or research activity, but rather as a reflection of structural participation constraints. Limited access to travel funding continues to be a major barrier for many researchers

[4] https://opencor.gitlab.io/program-2019/

[5] https://sites.google.com/view/plagaa2019

[6] https://opencor.gitlab.io/opencor-2020/

[7] https://comissoes.sbc.org.br/ce-pln/stil2021/programSTIL2021.pdf

in the region, particularly for international conferences. Although the 2021 edition was held fully online and achieved full acceptance, the exceptional conditions of the pandemic make it difficult to treat this edition as representative evidence that online venues alone are sufficient to address participation asymmetries. Additionally, submission patterns show a higher concentration of contributions from Brazil and Mexico, which are the countries of affiliation of the workshop organizers, suggesting that local academic networks and proximity to organizers play a significant role in community engagement when broader institutional support is lacking.

4. Open Corpora List

A central outcome of OpenCor has been the accumulation of a curated list[8] of open corpora and lexical resources for Latin American and Iberian languages. Building on the tradition of initiatives such as Linguateca, OpenCor similarly emphasizes documentation, accessibility, and community-driven reuse. The resources presented across editions are heterogeneous in scope, modality, language coverage, and level of maturity.

Rather than enforcing a fixed taxonomy, OpenCor emphasized descriptive presentations, allowing authors to discuss data collection methodologies, annotation schemes, licensing decisions, and known limitations. Over time, the OpenCor website accumulated links to these resources, gradually forming a distributed list of open datasets.

It is important to note that OpenCor does not collect linguistic data. The corpora described here correspond exclusively to the resources voluntarily submitted by participants of the workshops and related events. As a consequence, the set of languages represented in OpenCor should not be interpreted as a deliberate or comprehensive coverage of the languages of Latin America and the Iberian Peninsula. Many relevant languages are not present, including, for example, Basque, Tikuna, and Guarani Kaiowá, simply because no resources for these languages were submitted to the initiative.

OpenCor's list, despite being vastly incomplete, brings together a wide variety of open corpora and lexical resources for Latin American and Iberian languages, including both newly developed datasets and resources created before the initiative. Inclusion in the list has not been restricted to works presented at OpenCor meetings: authors may also request the addition of their resources independently, through a public submission form, available from the website of the initiative.

This author-initiated list reflects OpenCor's emphasis on community inclusion, visibility, and reuse. As of 2025, the list contains 40 resources. The contributions showcased at OpenCor cover a wide range of corpus types, beginning with traditional written and spoken corpora for Spanish (Sánchez Fernández and Medina Urrea, 2020) and Portuguese (Santos et al., 2018; Okano et al., 2020). These include both general-purpose and domain-specific collections, such as product reviews (Real et al., 2019; dos Santos Silva et al., 2024), oil and gas domain (Freitas et al., 2023) or standardized reference corpora (Crespo et al., 2023; Mendes, 2024).

Beyond language-specific corpora, the program highlights regional and national resources from Latin America and the Iberian Peninsula, addressing linguistic variation and language diversity (Pichel Campos et al., 2019). These include resources for Catalan (Rodriguez-Penagos et al., 2021) and multilingual survey questionnaires (Zavala-Rojas et al., 2021), reflecting the initiative's commitment to capturing regional diversity. Complementing these text-level datasets are structured lexical resources, such as high-coverage morphological lexicons (e.g. MorphoBR (Alencar et al., 2018)) and terminological databases (e.g. PetrolNer (Coelho and de Freitas, 2020)), as well as oral language collections (Junior et al., 2024; Othero and Ayres, 2014) that provide specialized annotation for phenomena such as emotional children's speech (Pérez-Espinosa et al., 2020), thereby broadening the functional range of available linguistic data.

A core concern of OpenCor is representation of underrepresented varieties and minority languages. Examples include corpora documenting indigenous languages, for instance, the Wixarika-Spanish Parallel Corpus (Mager et al., 2018), which provides aligned text material for an indigenous Mexican language, and ongoing efforts to build annotated parallel corpora for Nheengatu (de Alencar, 2023), the lingua franca in the Amazon basin, alongside resources for Sri Lanka Portuguese (Silva and Trigo, 2022) and Southern Quechua (Cardenas et al., 2018). Finally, the initiative also embraces non-spoken and multimodal language resources, such as multimedia corpora developed for applications like acoustic interaction research (Rascon et al., 2018) and sign language translation (Núñez-Reyes, 2016), underscoring OpenCor's engagement with multiple modalities of linguistic data.

This inclusive approach has brought visibility to corpora often marginalized in mainstream NLP venues, particularly work on Indigenous languages of Latin America and regional languages of the Iberian Peninsula. It highlights not only the existence of such data, but also the labor, linguistic diversity, and communities involved in their creation and maintenance. Notably, several of these

[8] https://opencor.gitlab.io/corpora/

corpora were first presented at OpenCor meetings during their early stages of development and were officially released at a later time.

5. Difficulties and Challenges

Despite its successes, OpenCor faces persistent structural challenges. These include limited and unstable funding, particularly for minority and indigenous language resources, where financial precarity directly affects both scope and continuity. The initiative also contends with the coordination costs inherent in geographically dispersed and institutionally diverse communities, which require sustained collaboration across national and disciplinary boundaries. Beyond initial creation, the long-term sustainability of corpora remains a significant concern, encompassing issues of hosting, licensing clarity, documentation, and ongoing updates. Finally, the labor involved in corpus development continues to suffer from insufficient academic recognition, as data creation and curation are often undervalued in hiring, promotion, and research assessment frameworks.

These difficulties are not unique to OpenCor, but reflect broader systemic issues in open scientific infrastructure. Funding has been the most persistent challenge faced by OpenCor. Apart from the initial support in 2018, the workshop never had dedicated funding for travel grants, invited speakers, or infrastructure. A funding request submitted to NAACL (NAACL Emerging Regions fund) in 2020 to support the 2021 edition was denied. As a result, participation often depends on authors' ability to self-fund or already be present at the host conference. While options such as "workshop-only" registration help reduce costs, several accepted papers across editions were not presented due to financial constraints.

These challenges highlight the broader structural issue that initiatives focused on data, openness, and linguistic diversity often struggle to secure funding compared to model-centric or commercially oriented research agendas. These challenges are not unique to OpenCor but reflect structural dynamics within the Latin American NLP ecosystem. To better situate OpenCor within this landscape, the following section examines related regional initiatives and their organizational models.

6. More Latin American Initiatives

OpenCor is not an isolated effort, but part of a landscape of community-driven initiatives dedicated to NLP and open linguistic resources in Latin America. Over the last decade, the region has witnessed the emergence of multiple forums, collectives, and research networks addressing structural gaps in visibility, infrastructure, and representation within global NLP research. Rather than forming a coordinated ecosystem, these initiatives typically operate independently, often without sustained interaction or mutual awareness. This fragmentation reflects structural constraints and contributes to the difficulty of sustaining long-term initiatives, many of which emerge, transform, or disappear over time.

Within this broader and fragmented landscape, we highlight a small number of initiatives of which we are aware, selected for their relevance to the development, visibility, and organization of language-related research in Latin America. This selection is not intended to be exhaustive, but rather illustrative of different models of community organization and engagement in the region. Although these initiatives differ in scope, structure, and objectives from OpenCor, we include them to situate our work within this wider regional context and to acknowledge that it forms part of an ongoing collective effort rather than a standalone endeavor.

Within the NLP community, one visible initiative is AmericasNLP[9], a workshop series focused on Indigenous languages of the Americas and often co-located with major international conferences such as ACL and NAACL. The venue promotes the development of corpora and shared datasets in low-resource settings and brings together researchers from universities in Latin America and Europe alongside participants affiliated with large technology companies. This configuration, also present in smaller venues, such as the Workshop on NLP for Indigenous Languages of Lusophone Countries[10], reflects a broader strategic interest in expanding multilingual AI systems to historically underrepresented languages. At the same time, the growing involvement of major industry actors raises questions about governance, ownership, and long-term control of linguistic resources, particularly where corporate infrastructures intersect with community-based language initiatives.

In Brazil, the Brazilian Symposium on Information and Human Language Technology[11] (STIL) constitutes probably the longest-running and most consolidated NLP venue in Latin America. Organized under the Brazilian Computing Society since 2003, STIL has historically served as a primary outlet for the presentation of open corpora, treebanks, and lexical resources for Brazilian Portuguese, while also welcoming work on Spanish and other languages. The 4th OpenCor edition was co-located with STIL, reflecting the overlap in community and goals, while maintaining distinct scopes.

[9] https://github.com/AmericasNLP

[10] https://sites.google.com/view/illc-nlp-2024/home

[11] https://sites.google.com/view/ce-pln/eventos/stil

Beyond formal academic venues, grassroots and independent collectives have played a decisive role in advancing open linguistic infrastructures. In Mexico, *Comunidad Elotl*[12] exemplifies a bottom-up model centered on the development of free and open-source tools and corpora for Indigenous languages such as Nahuatl, Otomí, and Mixtec. Similarly, in the Andean region, initiatives such as *Siminchik*[13] and *Siminchikkunarayku*[14] have focused on building open speech corpora for Quechua (Cardenas et al., 2018) and other native languages of Peru (Zevallos et al., 2022), often relying on community participation and crowdsourcing to overcome institutional limitations. Interestingly, many Latin American researchers focused on native language preservation are nowadays working with underrepresented languages of Spain or even with open data, but in other European countries, which may reflect structural constraints whereby researchers committed to endangered languages and open data struggle to find institutional spaces in Latin America that enable sustained, long-term, and continuous research efforts.

At a broader regional level, initiatives such as Khipu[15], a Latin American meeting on Artificial Intelligence from 2019, and institutions like CENIA[16] (*Centro Nacional de Inteligencia Artificial*) in Chile contribute to cross-national collaboration and infrastructure building. While their scope extends beyond language technologies, these initiatives reinforce a regional culture of openness, collaboration, and data sharing that directly benefits NLP research. Likewise, open data forums such as *Abrelatam Con Datos*[17] have shaped the policy and civic discourse around data as a public good, creating favorable conditions for open linguistic resources development since 2013.

Open data and data reuse are also central concerns for communities operating outside academic NLP itself, particularly those engaged with public information, transparency, and political accountability in Latin America. One such initiative is Coda.Br[18] (Conferência Brasileira de Jornalismo de Dados e Métodos Digitais), which occupies a space distinct from Computational Linguistics research while engaging directly with practices and debates around openness and data access. Organized since 2015, Coda.Br brings together journalists, civic technologists, and professionals interested in politics, transparency, and the use of open data. A distinctive feature of the conference is its commitment to regional decentralization, with recent editions held annually in the Amazon region, fostering participation beyond traditional academic and technological centers. Although not focused on language resources or NLP research, Coda.Br contributes to a broader culture of open data and critical engagement with public information that intersects with the conditions under which data, including linguistic data, are created, accessed, and sustained in Latin America.

Within this landscape, OpenCor occupies a complementary and specialized role. Rather than serving as a general NLP venue, it provides a dedicated forum for the documentation, discussion, and dissemination of open corpora and lexical resources for Latin American and Iberian languages. By focusing explicitly on data, licensing, and sustainability, OpenCor contributes to the regional ecosystem by addressing a critical but often under-recognized layer of NLP research infrastructure.

Taken together, these initiatives reveal a regional ecosystem characterized by decentralization, fragmentation, and limited structural funding. At the same time, they demonstrate that Latin America is actively shaping the development of NLP and lexical resources, rather than merely consuming technologies produced elsewhere. Researchers, collectives, and institutions across the region have built open corpora, organized conferences, and sustained collaborative networks, often under significant structural constraints. Within this context, OpenCor occupies a specific infrastructural niche: a forum dedicated not to model performance, but to the documentation and sustainability of open corpora. Its persistence across editions, despite the absence of stable funding, illustrates both the vitality and the precarity of data-centered initiatives in Latin America.

7. Conclusion

This paper has presented a historical account of OpenCor, a community-driven forum dedicated to open corpora and lexical resources for the languages of Latin America and the Iberian Peninsula. Between 2018 and 2024, OpenCor functioned as a space for documenting resources, discussing licensing and sustainability practices, and fostering exchange among researchers working on under-resourced, regional, and minority languages.

By consolidating information about meetings, submissions, and the resources presented across editions, this paper contributes to the preservation of institutional memory around an initiative whose outcomes extend beyond traditional publication metrics. In particular, it makes visible forms of labor, such as data collection, annotation, curation,

[12] https://elotl.mx/
[13] https://watuchi.org/
[14] https://siminchikkunarayku.pe/
[15] https://khipu.ai/
[16] https://cenia.cl/
[17] 'Open Latin America with Data'; https://abrelatam.org
[18] https://escoladedados.org/coda/

and maintenance, that are often underrepresented in mainstream NLP venues.

Rather than positioning OpenCor as a model to be replicated or sustained indefinitely, this historical record highlights both the possibilities and the structural constraints faced by the community-led efforts centered on openness and linguistic diversity. As such, the paper aims to serve as a reference point for researchers, organizers, and institutions interested in understanding how open language resource initiatives have emerged, operated, and evolved within the specific conditions of Latin America and the Iberian Peninsula.

7.1. Limitations

This paper provides a brief historical account of the OpenCor initiative and does not aim to be an exhaustive survey of all open corpora efforts in Latin America and the Iberian Peninsula. While we sought to reference relevant NLP groups, workshops, and initiatives, the landscape of language resource development in the region is broad, heterogeneous, and continuously evolving, with many efforts occurring outside established academic venues. The initiatives discussed also reflect, in part, the academic networks and geographic contexts in which OpenCor was organized, which may result in the under-representation of some communities or languages.

8. Bibliographical References

Leonel Figueiredo de Alencar, Bruno Cuconato, and Alexandre Rademaker. 2018. MorphoBR: An Open Source Large-Coverage Full-Form Lexicon for Morphological Analysis of Portuguese. *Texto Livre: Linguagem e Tecnologia*, 11(3):1–25.

Ronald A. Cardenas, Rodolfo Zevallos, Reynaldo Baquerizo, and Luis Camacho. 2018. Siminchik: A speech corpus for preservation of southern quechua. In *Proceedings of the Workshop "Improving Social Inclusion Using NLP: Tools, Methods and Resources" co-located with the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 21–28, Miyazaki, Japan. European Language Resources Association (ELRA).

Leonardo Gularte Coelho and Larissa Astrogildo de Freitas. 2020. Construção do corpora calendário brasileiro de saúde. In *Anais da Semana Integrada de Inovação, Ensino, Pesquisa e Extensão (SIIEPE)*, Pelotas, Brazil. Paper presented at SIIEPE 2020, Universidade Federal de Pelotas.

Maria Clara Ramos Morales Crespo, Maria Lina de Souza Jeannine Rocha, Mariana Lourenço Sturzeneker, Felipe Ribas Serras, Guilherme Lamartine de Mello, Aline Silva Costa, Mayara Feliciano Palma, Renata Morais Mesquita, Raquel de Paula Guets, Mariana Marques da Silva, Marcelo Finger, Maria Clara Paixão de Sousa, Cristiane Namiuti, and Vanessa Martins do Monte. 2023. Carolina: a general corpus of contemporary brazilian portuguese with provenance, typology and versioning information.

Leonel Figueiredo de Alencar. 2023. Yauti: A tool for morphosyntactic analysis of nheengatu within the universal dependencies framework. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 135–145, Porto Alegre, RS, Brazil. Sociedade Brasileira de Computação (SBC).

Valeria de Paiva, Livy Real, Hugo Gonçalo Oliveira, Claudia Freitas, Alexandre Rademaker, and Alberto Simões. 2016. An overview of portuguese wordnets. In *Proceedings of the 8th Global WordNet Conference*, Bucharest, Romania.

Lucas Nildaimon dos Santos Silva, Ana Cláudia Zandavalle, Carolina Francisco Gadelha Rodrigues, Tatiana da Silva Gama, Fernando Guedes Souza, Phillipe Derwich Silva Zaidan, Alice Florencio Severino da Silva, Karina Soares, and Livy Real. 2024. Repro: A benchmark dataset for opinion mining in brazilian portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese – Vol. 1*, pages 432–440, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.

Cláudia Freitas, Elvis de Souza, Maria Clara Castro, Tatiana Cavalcanti, Patrícia Ferreira da Silva, and Fábio Corrêa Cordeiro. 2023. Recursos linguísticos para o pln específico de domínio: o petrolês. *Linguamática*, 15(2):51–68.

Marcos García and Alfredo Crespo-Otero. 2022. A targeted assessment of the syntactic abilities of transformer models for galician-portuguese. In *Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21–23, 2022, Proceedings*, volume 13208 of *Lecture Notes in Computer Science*, pages 46–56. Springer.

Isaac Junior, Gabriela Wick-Pedro, Cláudia Barros, and Oto Vale. 2024. Roda viva boundaries: an overview of an audio-transcription corpus. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*.

Fernando López-Escobedo and Jorge Solorzano-Soto. 2016. Propuesta de clasificación de un banco de voces con fines de identificación forense. *Linguamática*, 8(1):33–41.

Manuel Mager, Diónico Carrillo, and Ivan Meza. 2018. Probabilistic finite-state morphological segmenter for wixarika (huichol) language. *Journal of Intelligent & Fuzzy Systems*, 34(5):3081–3087.

Alfonso Medina Urrea. 2018. Pluricentric languages: New perspectives in theory and description. *Nueva Revista de Filología Hispánica*, 66(1):211–214.

Amália Mendes. 2024. The reference corpus of contemporary portuguese: Corpus design and case study on discourse markers. In Miguel Calderón Campos and Gael Vaamonde, editors, *Linguistic Corpora and Big Data in Spanish and Portuguese*, pages 145–178. Walter de Gruyter GmbH.

Alba Núñez-Reyes. 2016. Agrupamiento de textos cortos en dominios cruzados. Technical report, Repositorio Institucional, Instituto de Lingüística, Universidad Autónoma Metropolitana.

Emerson Yoshiaki Okano, Zebin Liu, Donghong Ji, and Evandro Eduardo Seron Ruiz. 2020. Fake news detection on fake.br using hierarchical attention networks. In *Computational Processing of the Portuguese Language*, pages 143–152, Cham. Springer International Publishing.

Gabriel de Ávila Othero and Mônica Rigo Ayres. 2014. Anotação morfológica automática de corpus de língua falada: desafios ao Aelius. *Texto Livre: Linguagem e Tecnologia*, 7(2):44–60.

José Ramón Pichel Campos, Pablo Gamallo Otero, and Iñaki Alegria Loinaz. 2019. Measuring diachronic language distance using perplexity: Application to English, Portuguese, and Spanish. *Natural Language Engineering*, 26(4):433–454.

Humberto Pérez-Espinosa, Juan Martínez-Miranda, Ismael Espinosa-Curiel, Josefina Rodríguez-Jacobo, Luis Villaseñor-Pineda, and Himer Avila-George. 2020. Iesc-child: An interactive emotional children's speech corpus. *Computer Speech & Language*, 59:55–74.

Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria de Paiva. 2017. Universal Dependencies for Portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 197–206, Pisa, Italy. Linköping University Electronic Press.

Caleb Rascon, Ivan V. Meza, Aldo Millan-Gonzalez, Ivette Velez, Gibran Fuentes, Dennis Mendoza, and Oscar Ruiz-Espitia. 2018. Acoustic interactions for robot audition: A corpus of real auditory scenes. *The Journal of the Acoustical Society of America*, 144(5):EL399–EL403.

Livy Real, Marcio Oshiro, and Alexandre Mafra. 2019. B2w-reviews01: an open product reviews corpus. In *Proceedings of the Symposium in Information and Human Language Technology (STIL 2019)*, Brazil. Dataset and paper presented at STIL 2019.

Carlos Rodriguez-Penagos, Carme Armentano-Oller, Marta Villegas, Maite Melero, Aitor Gonzalez, Ona de Gibert Bonet, and Casimiro Carrino Pio. 2021. The catalan language club.

Manuel Alejandro Sánchez Fernández and Alfonso Medina Urrea. 2020. Hacia el etiquetado de estados informativos en el corpus periodístico del noroeste de méxico (copenor). *Signos Lingüísticos*, 16(31):164–181.

Diana Santos. 2003. Relatório linguateca: Relatório relativo ao período 2000–2003. Technical report, Linguateca.

Diana Santos, Cláudia Freitas, and Eckhard Bick. 2018. Obras: a fully annotated and partially human-revised corpus of brazilian literary works in public domain. *In CorLex*.

C. Silva and L. Trigo. 2022. *PtLanka*. CLUP – Centro de Linguística da Universidade do Porto, Porto.

Diana Zavala-Rojas, Danielly Sorato, Lidun Hareide, and Knut Hofland. 2021. [MCSQ] Multilingual Corpus of Survey Questionnaires. *Meta: Journal des traducteurs*.

Rodolfo Zevallos, Luis Camacho, and Nelsi Melgarejo. 2022. Huqariq: A multilingual speech corpus of native languages of Peru for Speech recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5029–5034, Marseille, France. European Language Resources Association.

MedicaLLM: LLM-Driven Speech and Language Solutions for Healthcare

Ronghao Pan[2], **Pedro José Vivancos-Vicente**[1], **Juan Salvador Castejón-Garrido**[1], **Tomás Bernal-Beltrán**[2], **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
ronghao.pan@um.es, pedro.vivancos@vocali.net,
juans.castejon@vocali.net, tomas.bernalb@um.es, valencia@um.es

Abstract

Although healthcare documentation is increasingly dependent on speech-based clinical interactions, general-purpose Automatic Speech Recognition (ASR) and Large Language Models (LLMs) lack the domain adaptation, structured control and interoperability guarantees required in regulated medical environments. These limitations often result in transcription errors, hallucinated content, and limited alignment with standardized coding systems. This paper introduces MedicaLLM, a multilingual, end-to-end framework integrating domain-adapted ASR, LLM-based structured report generation, and ontology-driven semantic enrichment within a modular architecture for clinical documentation. MedicaLLM combines medical interview transcription with structured report generation, summarization, and error correction; Named Entity Recognition (NER); and Medical Entity Linking (MEL) to align with standards such as SNOMED-CT and ICD-10. Deployed as a secure software as a service (SaaS) platform with REST API integration, MedicaLLM aims to reduce the administrative burden, improve the quality of documentation, and enhance semantic interoperability across healthcare systems, all while maintaining computational efficiency and clinical reliability.

Keywords: Automatic Speech Recognition, Large Language Models, Clinical Documentation, Medical Report Generation, Named Entity Recognition, Medical Entity Linking, Multilingual Healthcare AI

1. Introduction

Healthcare systems are currently undergoing an accelerated digital transformation, driven by the increasing demand for efficiency, accuracy, and interoperability in the clinical documentation processes (Holmgren et al., 2024). The growing administrative burden faced by healthcare professionals, together with the need for precise and standardized medical records, has highlighted the limitations of traditional documentation workflows. Automatic Speech Recognition (ASR) (Blackley et al., 2019; Ng et al., 2025) and Large Language Models (LLMs) (Shool et al., 2025; Thirunavukarasu et al., 2023), represents a strategic opportunity to modernize clinical practice while maintaining high standards of safety, compliance and linguistic precision.

Despite recent advances in Transformer-based architectures such as BERT (Devlin et al., 2019), GPT-based models (Brown et al., 2020), LLaMa-3 (Grattafiori et al., 2024), Qwen-3 (Yang et al., 2025), Whisper (Radford et al., 2023) and related variants, directly adopting these technologies in the healthcare domain remains challenging. Clinical language is characterized by highly specialized terminology, domain-specific abbreviations, multilingual variability (notably Spanish and Catalan), and strict regulatory requirements (Gu et al., 2021; Carrino et al., 2022). Generic ASR and LLM systems often struggle to accurately transcribe medical consultations, correctly interpret technical terminology, or generate structured medical reports aligned with standardized ontologies such as SNOMED-CT and ICD-10 (Hu et al., 2024). Furthermore, concerns about data privacy, computational costs, robustness in noisy clinical environment, and model hallucinations further complicate their deployment in real-world healthcare settings (Kim et al., 2025).

The MedicalLLM project addresses these challenges by developing an end-to-end Artificial Intelligence (AI) pipeline. This pipeline integrates domain-specific ASR systems with LLM-based modules that are specialized for generating structured reports, recognizing named entities, and linking medical entities. Specifically designed for a multilingual healthcare environment in Spanish and Catalan, the system incorporates fine-tuned Transformer-based models, medical-domain datasets, and normalization mechanisms aligned with international clinical standards. The project aims to automate and optimize key stages of clinical documentation by combining speech processing, semantic understanding, summarization, and ontology mapping within a unified architecture.

The proposed architecture is organized into modular components that separate data acquisition, transcription, semantic processing, normalization, and presentation layers. First, an end-to-end ASR module processes medical interviews and dictations, ensuring high transcription accuracy under

Proceedings of LANLP: Bridging Ibero and Latin American NLP Communities, pages 29–37

realistic hospital acoustic conditions. Next, LLM-based modules transform raw transcripts into structured medical reports through context-aware summarization and error correction mechanisms. Then, NER and MEL components extract relevant clinical entities, such as symptoms, diagnoses, medications, and procedures, and map them to standardized terminologies to ensure interoperability and regulatory compliance. The system is deployed via a scalable software as a service (SaaS) platform with secure backend processing, user-friendly interfaces, and API-based integration with electronic health record (EHR) systems.

Beyond technological innovation, the project places strong emphasis on reliability, privacy and responsible AI deployment. Data governance mechanisms compliant with General Data Protection Regulation (GDPR) and healthcare regulations, hallucination migation strategies, structured output validation, and human-in-the-loop validation workflows are integrated into the system design. Furthermore, optimization techniques such as quantization and parameter-efficient fine-tuning are explored to enable deployment in resource-constrained healthcare environments.

The project (CPP2024-011574) is funded by the Spanish National Research Agency (AEI) through the Colaboración público-privada call. The consortium members are VOCALI SISTEMAS INTELIGENTES S.L., a company with extensive experience in medical ASR solutions, and the TECNOMOD research group at the Universidad de Murcia, specialized in NLP, LLMs and semantic technologies.

Currently, the platform is being developed and validated through structured work packages covering multilingual resource creation, ASR adaptation, LLM fine-tuning, ontology alignment, SaaS development, and integrated system validation. The expected outcome is a Technology Readiness Level (TRL) 6 prototype capable of operating in real clinical environments, reducing administrative workload, improving documentation quality, and enhancing the interoperability of medical records across healthcare systems.

In summary, MedicaLLM proposes a comprehensive, multilingual and domain-adapted AI framework that bridges the gap between state-of-the-art speech and language technologies and the practical demands of modern healthcare systems, contributing to more efficient, standardized and scalable clinical workflows.

2. Background Information

In recent years, advances in AI, particularly in Natural Language Processing (NLP) and ASR, have significantly transformed the way textual and auditory data are processed. Transformer-based architectures such as BERT (Devlin et al., 2019), LLaMa-3 (Grattafiori et al., 2024), GPT3-family models (Brown et al., 2020), Qwen-3 (Yang et al., 2025), Whisper (Radford et al., 2023), HuBERT (Hsu et al., 2021) and Wav2Vec 2.0 (Baevski et al., 2020) have demonstrated remarkable capabilities in contextual language modeling, speech-to-text transcription, and generative reasoning. These technologies have reached near-human performance in general-domain tasks and have been widely adopted in applications such as virtual assistants, automated transcription services, machine translation, and conversational systems (Blackley et al., 2019; Ng et al., 2025; Shool et al., 2025; Thirunavukarasu et al., 2023).

However, the healthcare domain presents unique linguistic, operational and regulatory challenges that limit the direct applicability of generic AI models. Clinical language is highly specialized and characterized by domain-specific terminology, acronyms, abbreviations, implicit contextual references, and heterogeneous discourse styles (Gu et al., 2021; Carrino et al., 2022). Medical interviews involve spontaneous dialogue between healthcare professionals and patients that often combining colloquial expressions with technical vocabulary (Kuligowska et al., 2023). In contrast, medical dictations and clinical reports are more structured and formal, with a greater emphasis on terminology. This variability poses significant challenges for speech recognition and language understanding systems (Hodgson and Coiera, 2016).

Although state-of-the-art ASR models have achieved strong performance in controlled environments, their accuracy tends to degrade in domain-specific contexts such as hospitals, where background noise, overlapping speech, and speaker variability are common (Lamy et al., 2018). Moreover, generic ASR systems often misrecognize specialized medical terminology, resulting in transcription errors that can have critical implications for diagnosis, treatment planning, and clinical documentation (Zuchowski and Göller, 2022). There is a significant unmet need for domain-adapted ASR systems that are fine-tuned using multilingual medical datasets, particularly in Spanish and Catalan.

Concurrent with the development of ASR, LLMs have precipitated a paradigm shift in NLP by facilitating sophisticated functionalities, including text summarization, information extraction, reasoning, and structured generation (Brown et al., 2020). In the medical domain, LLMs exhibit considerable promise in automating clinical documentation, generating structured reports from raw transcripts, extracting relevant medical entities, and mapping them to standardized ontologies such as SNOMED-CT and ICD-10 (Thirunavukarasu et al., 2023). Spe-

cialized variants such as BioBERT (Lee et al., 2020) and ClinicalBERT (Huang et al., 2019) have exhibited enhancements in biomedical text mining tasks. However, the majority of high-performing generative models are predominantly trained on English data, thereby constraining their robustness in other languages.

The presence of multiple languages in a given context introduces a degree of complexity that must be addressed. While Spanish is extensively incorporated in NLP resources, Catalan is underrepresented in substantial medical datasets. The paucity of high-quality annotated corpora for medical speech and clinical text in these languages imposes constraints on the fine-tuning and evaluation of domain-adapted transformer models. Additionally, healthcare systems mandate strict adherence to data protection regulations (e.g., GDPR), interoperability standards, and traceability requirements, which generic AI deployments frequently neglect to address adequately.

Another critical challenge pertains to reliability and robustness. LLMs have been observed to generate hallucinations, defined as outputs that are either factually inaccurate or fabricated, particularly when operating outside the parameters of their training distribution or when tasked with generating structured medical content, as evidenced by recent (Huang et al., 2025). In healthcare environments, where precision and accountability are essential, such risks must be mitigated through controlled generation strategies, structured output constraints, validation layers, and human-in-the-loop supervision mechanisms (Kim et al., 2025).

From a technological maturity perspective, many AI-based healthcare documentation solutions are still in the experimental or early prototype stage. Integrating ASR, LLM-based summarization, NER, and MEL into a unified, scalable, and interoperable pipeline poses significant research and engineering challenges. Additionally, the computational requirements for deploying large transformer models can be prohibitive for small and medium-sized healthcare institutions. This makes model optimization techniques, such as quantization, parameter-efficient fine-tuning, and resource-aware deployment, essential for real-world adoption.

In this context, there is a clear need for an end-to-end framework tailored to multilingual healthcare environments. This framework should combine domain-specific speech recognition, structured language generation, semantic normalization, interoperability with clinical coding systems, secure data governance, and scalable deployment via SaaS architectures. Addressing these gaps will not only reduce administrative burden and improve documentation efficiency but also enhance data quality, interoperability, and clinical decision support in modern healthcare systems.

3. System Architecture

The MedicaLLM platform is designed as a modular end-to-end pipeline that integrates domain-adapted ASR and LLMs to support multilingual (Spanish and Catalan) clinical documentation workflows. As described in the Figure 1, the global SaaS system is organized into four main modules: (1) ASR for medical interviews and dictation; (2) medical report generation with summarization and error correction; (3) NER and MEL with normalization to clinical standards; and (4) user interfaces and service communication. The architecture separates perception (speech-to-text), clinical reasoning and structuring (LLM-based generation), semantic enrichment (NER/MEL), and interaction layers (SaaS UI and APIs), enabling independent optimization and scalable deployment in real healthcare environments.

At runtime, a healthcare professional records or uploads audio from either a doctor–patient interview or a clinician dictation. Audio is securely transmitted to the ASR module, which produces a transcript while preserving domain terminology and, in interview scenarios, applying speaker diarization to distinguish patient and clinician turns. The resulting text is routed to the report generation module, where an LLM transforms unstructured transcripts into structured clinical documentation (e.g., symptoms, diagnosis, treatment plan), incorporating summarization and context-aware error correction to mitigate transcription artifacts. Next, the NER/MEL module extracts relevant clinical entities and links them to standardized terminologies (e.g., SNOMED-CT, ICD-10), enabling interoperability and downstream integration with EHR systems. Finally, the outputs are presented to the user through the SaaS interface, where clinicians can validate and edit results, and optionally export them through secure API integrations.

All core services are conceived to operate under a SaaS deployment model, supporting both real-time and offline workflows, and enabling integration with external hospital systems via RESTful APIs. The following subsections describe each module in more detail.

3.1. ASR Module for Medical Interviews and Dictations

The ASR module constitutes the perceptual backbone of the MedicaLLM architecture. It objective is to convert spoken medical interactions into accurate textual representations while preserving clinical terminology, speaker structure and contextual coherence.

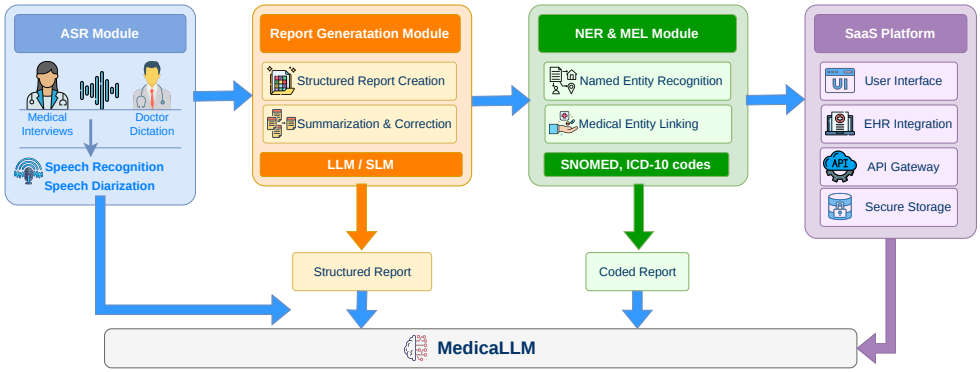


Figure 1: Overview of the modules that conform the MedicaLLM system architecture.

Unlike generic speech recognition systems, the MedicaLLM ASR component is specifically adapted to the medical domain. It addresses two distinct but complementary use cases:

1. **Medical interviews:** spontaneous, multi-speaker interactions between clinician and patient, characterized by turn-taking, colloquial expressions, incomplete sentences, interruptions and embedded technical terminology.

2. **Medical dictation:** structured, terminology-dense speech produced by healthcare professionals, often including acronyms, abbreviations and highly domain-specific expressions.

The module will be built upon transformer-based end-to-end ASR architectures (e.g., Whisper, Wav2Vec 2.0, or HuBERT variants), which will be fine-tuned using multilingual domain-specific datasets generated within the project. These datasets will include manually transcribed real and simulated consultations, augmented audio reflecting hospital noise conditions, and diverse accents to ensure robustness across clinical settings.

For interview scenarios, speaker diarization mechanisms will be integrated to distinguish between clinician and patient turns. This separation is essential for downstream structured report generation, where the attribution of symptoms and statements must be correctly contextualized. For dictation scenarios, additional preprocessing steps will be applied, including phonetic normalization of acronyms and abbreviation expansion modeling, enabling the system to handle expressions such as letter-by-letter pronunciations and abbreviated clinical terminology.

The fine-tuning strategy will evaluate multiple adaptation configurations, including encoder freezing, full fine-tuning, and gradual unfreezing, optimizing the trade-off between domain adaptation and the preservation of general acoustic representations. Performance will be evaluated using Word Error Rate (WER), domain-specific terminology accuracy, and Real-Time Factor (RTF), ensuring both transcription precision and operational feasibility.

The output of this module will be a timestamped, optionally diarized transcript that will serve as structured input for the LLM-based report generation module.

3.2. LLM-Based Medical Report Generation, summarization and Error Correction

The report generation module transforms raw ASR transcripts into structured, standardized, and clinically meaningful documentation. This process is expected to significantly reduce the administrative workload for healthcare professionals and improve consistency in clinical reporting.

The module will leverage multilingual LLMs and SLMs (e.g., Gemma-3, Mistral, Phi-4, and Qwen-3) adapted to the medical domain through supervised fine-tuning and instruction tuning on curated clinical datasets. This process will incorporate transcripts, structured reports, and annotated data to ensure alignment with clinical writing conventions.

This component performs three tightly integrated functions:

3.2.1. Structured Report Generation

The system transforms unstructured conversational transcripts into organized reports divided into predefined sections such as, chief complaint, history of present illness, physical examination findings, diagnosis and treatment plan.

This structuring will be achieved through prompt conditioning, instruction tuning, and template-guided decoding strategies, ensuring compliance with standardized medical documentation formats.

3.2.2. Abstractive and Extractive summarization

The module uses hybrid summarization techniques. Abstractive summarization will condense long transcripts into coherent, concise clinical narratives, while extractive mechanisms will ensure that critical factual elements, such as drug names, dosages, and lab results, are explicitly preserved. This dual strategy is expected to reduce information loss while maintaining readability.

Different advanced approaches will be progressively evaluated to support this objective. These approaches include encoder–decoder architectures, such as mT5 and mBART, for multilingual abstractive summarization. Subsequently, LLMs will be assessed to enhance contextual coherence, domain adaptation, and structured generation capabilities.

3.2.3. Context-Aware Error Correction

Given that ASR systems may introduce transcription errors, especially in terminology-dense contexts, the LLM incorporates contextual correction mechanisms. By leveraging domain knowledge encoded during fine-tuning, the model identifies probable inconsistencies and corrects medical terminology when contextual evidence supports such correction. This step enhances reliability and reduces post-editing workload.

Prompt engineering strategies (zero-shot, few-shot, and chain-of-thought prompting) will be evaluated to improve structured generation under computational constraints. Additionally, model optimization techniques such as quantization and QLoRA will be explored to enable deployment in resource-constrained healthcare infrastructures.

The output of this module is a structured, linguistically refined clinical report ready for semantic enrichment.

3.3. NER and Medical Entity Linking for Normalization

The semantic enrichment layer enhances interoperability and machine-readability of the generated reports by extracting and normalizing clinically relevant entities.

This module consists of two sequential components:

3.3.1. Named Entity Recognition (NER)

The NER subsystem identifies entities such as symptoms, diagnoses, medications, procedures, date, dose of the medication, organization, laboratory tests, and among others.

Two complementary strategies will be evaluated:

- **LLM-based extraction via prompt engineering**, enabling flexible zero-shot or few-shot recognition in low-resource contexts.
- **Fine-tuned encoder-based models** (e.g., BERT, RoBERTa, XLM-R variants) trained on manually annotated corpora, optimized for computational efficiency.

Performance is measured using precision, recall and F1-score, ensuring clinical-grade entity extraction accuracy.

3.3.2. Medical Entity Linking (MEL)

Following entity extraction, the MEL subsystem will map surface-level terms to standardized ontology entries such as SNOMED-CT and ICD-10 codes. Transformer-based embedding approaches (e.g., SapBERT-style representations) will be used to generate contextual embeddings for candidate concepts, which will be retrieved through similarity search mechanisms.

To address ambiguity and near-synonymous terminology, a re-ranking layer will be incorporated to refine candidate selection using context-sensitive scoring. This step is critical in clinical scenarios where subtle semantic differences may correspond to distinct diagnostic codes.

The output will be an enriched clinical report annotated with standardized codes, enabling seamless integration into EHR systems and supporting downstream analytics.

3.4. SaaS Platform, User Interfaces and Service Communication

The final layer of the architecture corresponds to the SaaS deployment infrastructure and user interaction components. The platform is designed to operate as a scalable cloud-based service while supporting integration with on-premise hospital systems when required.

3.4.1. Backend Layer

The backend orchestrates the full operational lifecycle of the system, coordinating ASR processing, LLM inference, semantic enrichment workflows, dataset management, and model versioning within a unified service layer. It manages task scheduling, resource allocation, logging, and inter service communication, ensuring that each module, including speech recognition, report generation, and entity normalization, operates cohesively within the overall pipeline.

Secure storage mechanisms are integrated at multiple levels, including encrypted data at rest and in transit, role based access control, audit logging, and traceability of model outputs and edits. The

infrastructure is designed to comply with GDPR requirements and healthcare data protection standards, supporting anonymization or pseudonymization strategies where appropriate, as well as controlled data retention policies.

The architecture supports both synchronous real time and asynchronous batch processing workflows. In real time scenarios, such as live dictation or consultation transcription, low latency inference pipelines ensure timely feedback to clinicians. In asynchronous mode, larger audio files or bulk documentation tasks can be processed through batch execution, optimizing computational resource utilization. This dual capability enables flexible deployment across diverse healthcare environments, from small clinics to large hospital networks.

3.4.2. Frontend Interfaces

User interfaces provide healthcare professionals with intuitive and user-centered tools that support the full clinical documentation workflow. Through the interface, users can securely record consultations in real time or upload previously recorded audio files. They are able to review automatically generated transcripts with timestamp navigation, compare diarized speaker segments when applicable, and detect potential transcription inconsistencies. The interface also enables validation of structured clinical reports, allowing clinicians to examine how information has been organized into predefined sections such as symptoms, diagnosis and treatment plan. In addition, users can inspect extracted clinical entities along with their associated standardized codes, verify their correctness within context, and perform manual corrections when necessary. Before final submission, professionals can edit, refine and formally approve the documentation to ensure clinical accuracy and completeness.

Visualization dashboards complement these tools by providing operational transparency. They offer real-time and aggregated metrics on transcription quality (e.g., error rates and processing latency), report generation performance, entity extraction statistics, and dataset usage indicators. These monitoring capabilities support quality assurance, facilitate model evaluation and version comparison, and enable data-driven optimization of the overall system.

3.4.3. API and Integration Layer

RESTful APIs enable integration with third-party hospital systems, EHRs, and external services. An API gateway manages authentication, access control and secure communication between frontend and backend services.

The SaaS architecture is designed following microservices principles, enabling containerized deployment, horizontal scalability and independent updating of ASR, LLM and NER components. This modular design ensures maintainability, resilience and adaptability to evolving clinical requirements.

3.5. Implementation Status and Validation

To clarify the distinction between contributions and ongoing work, this section details the implementation status of each component, the datasets used, and the validation strategy adopted in the MedicaLLM project.

MedicaLLM is currently under development within a structured work plan and is expected to reach a TRL 6 prototype, validated in realistic clinical environments. The system follows a modular design, where different components are at different stages of maturity:

- **Data resources (In progress)**: Multilingual datasets in Spanish and Catalan have been created, including real and simulated doctor–patient conversations, medical dictations, and structured clinical reports. These datasets are manually transcribed and annotated by experts, forming gold-standard corpora for ASR, NER, and MEL tasks.
- **ASR module (under development)**: Baseline transformer-based ASR systems (e.g., Whisper, Wav2Vec 2.0) have been selected and are currently being adapted through domain-specific fine-tuning using the generated datasets. Optimization strategies such as encoder freezing and gradual fine-tuning are under evaluation.
- **LLM-based report generation (under development)**: Initial pipelines for structured report generation, summarization, and error correction have been implemented using multilingual LLMs. Ongoing work focuses on instruction tuning, prompt optimization, and alignment with clinical documentation standards.
- **NER and MEL modules (under development)**: Annotated datasets for entity recognition and linking have been completed. Model development is ongoing, combining fine-tuned encoder-based models and LLM-based extraction approaches, along with ontology alignment to SNOMED-CT and ICD-10.
- **Integrated SaaS platform (planned integration phase)**:The full end-to-end integration of all modules into a scalable SaaS platform, including API-based interoperability with EHR systems, is currently under development and will be validated in later project stages.

The operational pipeline follows a well-defined sequence. First, audio is acquired either from medical interviews or clinician dictations. The signal is processed by the ASR module, which generates a transcript and optionally performs speaker diarization to distinguish between patient and clinician turns. The resulting text is then transformed by the LLM-based module into structured clinical reports, incorporating summarization and context-aware error correction. Subsequently, NER and MEL components extract relevant clinical entities and map them to standardized terminologies such as SNOMED-CT and ICD-10. The normalized output is then presented to clinicians through the user interface, where validation and editing take place before final export and integration into EHR systems. This design ensures a clear separation between perception, reasoning, semantic normalization, and user interaction layers.

The system relies on domain-specific datasets generated within the project, addressing the scarcity of medical resources in Spanish and Catalan. These datasets include transcribed doctor–patient conversations (both real and simulated), clinical dictations, structured medical reports, and synthetic audio generated via text-to-speech to improve robustness under diverse acoustic conditions. In addition, manually annotated corpora have been created for NER and MEL tasks, enabling supervised training and evaluation of semantic extraction and normalization components. All data collection and processing follow strict GDPR-compliant protocols, including anonymization procedures and expert validation.

Given the risks associated with generative models in clinical contexts, MedicaLLM incorporates multiple complementary hallucination mitigation strategies. These include template-guided and structured generation, which constrains outputs to predefined clinical sections, as well as hybrid extractive–abstractive summarization techniques that preserve critical factual information while ensuring readability. Furthermore, context-aware error correction mechanisms leverage domain knowledge during LLM inference to detect and resolve inconsistencies. Ontology grounding through MEL aligns generated outputs with standardized medical terminologies, and human-in-the-loop validation ensures that all outputs are reviewed before clinical use. Together, these mechanisms transform hallucination mitigation from a conceptual objective into an operational component of the system.

Human validation constitutes a central element of the workflow. Through the SaaS interface, clinicians can review transcripts and structured reports, inspect extracted entities and their associated standardized codes, and perform manual corrections where necessary. Users are able to edit, refine, and formally approve the final report, ensuring clinical accuracy and completeness. Only validated outputs are exported to EHR systems, guaranteeing reliability, traceability, and regulatory compliance.

Although large-scale experimental results are still under development, the evaluation framework is already defined. The ASR module is assessed using metrics such as WER, domain-specific terminology accuracy, and RTF. The NER component is evaluated through precision, recall, and F1-score, while MEL performance is measured in terms of linking accuracy and ranking quality. At the system level, evaluation includes clinician validation time, correction rate, and usability indicators.

4. Conclusions and Further Work

MedicaLLM integrates the core technological components required for end-to-end multilingual clinical documentation, including a domain-adapted ASR module for medical interviews and dictation, LLM-based structured report generation with summarization and error correction capabilities, a semantic enrichment layer for NER and MEL, and a secure SaaS platform enabling user interaction and EHR integration. The architecture combines speech perception, structured generation, ontology alignment and human-in-the-loop validation within a modular and scalable microservices framework tailored to healthcare environments.

Currently, the platform is being developed and validated through structured work packages covering multilingual resource creation, ASR adaptation, LLM fine-tuning, ontology alignment, SaaS development, and integrated system validation. These activities include the systematic construction and annotation of domain-specific datasets, the progressive adaptation of speech and language models to clinical terminology, and the iterative evaluation of transcription accuracy and structured report generation quality. The validation framework assesses the structural and semantic consistency of automatically generated medical reports, as well as the precision and recall of entity extraction and normalization components. Particular emphasis is placed on minimizing transcription artifacts, reducing clinically unsafe hallucinations, and ensuring full alignment with standardized medical coding systems and interoperability requirements.

Future work will concentrate on several complementary research and development directions. First, we plan to enhance domain robustness through adaptive fine-tuning strategies that incorporate continual learning mechanisms, enabling the ASR and LLM components to evolve as new clinical terminology, interaction styles and specialty-specific vocabularies emerge. Second, we aim to explore tighter integration between report genera-

tion and ontology normalization, investigating joint modeling approaches where structured generation is directly constrained by standardized terminologies during decoding. Finally, we will investigate automated feedback loops between clinician corrections and model refinement pipelines. By incorporating validated post-edits into incremental training cycles, the system will progressively improve its structured generation accuracy and semantic consistency.

Acknowledgements

This research is part of the research project MedicaLLM (CPP2024-011574) funded by MICIU/AEI /10.13039/501100011033 and the European Regional Development Fund (ERDF EU/FEDER UE)- a way of making Europe.

5. Ethical Considerations and Limitations

The data used for system validation and adaptation consist of enterprise-provided incident records, policy documents and internally generated test cases. All materials are handled within controlled corporate environments in accordance with applicable data protection regulations. No personal or sensitive information is publicly released as part of this research.

The system is designed to operate within enterprise infrastructures, ensuring that multimodal inputs (including text, images and audio) are processed under organizational data governance policies. When required, anonymization and access control mechanisms are applied to prevent unauthorized exposure of customer information.

6. Bibliographical References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Suzanne V Blackley, Jessica Huynh, Liqin Wang, Zfania Korach, and Li Zhou. 2019. Speech recognition for clinical documentation from 1990 to 2018: a systematic review. *Journal of the american medical informatics association*, 26(4):324–338.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Casimiro Pio Carrino, Joan Llop, Marc Pàmies, Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Joaquín Silveira-Ocampo, Alfonso Valencia, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. Pretrained biomedical language models for clinical NLP in Spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 193–199, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

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

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Tobias Hodgson and Enrico Coiera. 2016. Risks and benefits of speech recognition for clinical documentation: a systematic review. *Journal of the american medical informatics association*, 23(e1):e169–e179.

A Jay Holmgren, Julia Adler-Milstein, and Nate C Apathy. 2024. Electronic health record documentation burden crowds out health information exchange use by primary care physicians: Article examines electrnoic health record documentation burden. *Health Affairs*, 43(11):1538–1545.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou,

Zehan Li, Xiaoqian Jiang, Zhiyong Lu, et al. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9):1812–1820.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, Lizhou Fan, Eugene Park, Tristan Lin, Joonsik Yoon, Wonjin Yoon, Maarten Sap, Yulia Tsvetkov, Paul Liang, Xuhai Xu, Xin Liu, Daniel McDuff, Hyeonhoon Lee, Hae Won Park, Samir Tulebaev, and Cynthia Breazeal. 2025. Medical hallucination in foundation models and their impact on healthcare. *medRxiv*.

Karolina Kuligowska, Maciej Stanusch, and Marek Koniew. 2023. Challenges of automatic speech recognition for medical interviews-research for polish language. *Procedia Computer Science*, 225:1134–1141.

Manuel Lamy, Rúben Pereira, João C Ferreira, José Braga Vasconcelos, Fernando Melo, and Iria Velez. 2018. Extracting clinical information from electronic medical records. In *International Symposium on Ambient Intelligence*, pages 113–120. Springer.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Joel Jia Wei Ng, Eugene Wang, Xinyan Zhou, Kevin Xiang Zhou, Charlene Xing Le Goh, Gabriel Zheng Ning Sim, Hiang Khoon Tan, Serene Si Ning Goh, and Qin Xiang Ng. 2025. Evaluating the performance of artificial intelligence-based speech recognition for clinical documentation: a systematic review. *BMC medical informatics and decision making*, 25(1):236.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Sina Shool, Sara Adimi, Reza Saboori Amleshi, Ehsan Bitaraf, Reza Golpira, and Mahmood Tara. 2025. A systematic review of large language model (llm) evaluations in clinical medicine. *BMC Medical Informatics and Decision Making*, 25(1):117.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

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

Matthias Zuchowski and Aydan Göller. 2022. Speech recognition for medical documentation: an analysis of time, cost efficiency and acceptance in a clinical setting. *British Journal of Healthcare Management*, 28(1):30–36.

mCS-LM: Multimodal Customer Service and Incident Management Systems Based on Large Language Models

Carlos Díaz-Morales[1], Marcos Checa-Rubio[1], Tomás Bernal-Beltrán[2], Ronghao Pan[2], David Barbáchano[1], María del Pilar Salas-Zárate[3,4], Mario Andrés Paredes-Valverde[3,4], Rafael Valencia-García[2]

[1] PANEL SISTEMAS INFORMÁTICOS S.L., C. de Josefa Valcárcel, 9, Cdad. Lineal, 28027 Madrid, Spain
[2] Departamento de Informática y Sistemas, Universidad de Murcia, Campus de Espinardo 30100 Murcia, Spain
[3] Tecnológico Nacional de México/I.T.S. Teziutlán, Fracción I y II, Teziutlán 73960, Mexico
[4] Medio Melon, Av. Coyoacán 1435, Del Valle, Delegación Benito Juárez, Ciudad de México 03100 Mexico
carlos.diaz@panel.es, marcos.checa@panel.es, tomas.bernalb@um.es, ronghao.pan@um.es, david.barbachano@panel.es, maria.sz@teziutlan.tecnm.mx, mario.pv@teziutlan.tecnm.mx, valencia@um.es

Abstract

Customer service and incident management increasingly rely on multimodal evidence, combining text, images and audio. However, general-purpose models lack domain grounding, structured output control and reliability guarantees required in regulated enterprise environments, often leading to hallucinated responses and limiting their practical deployment. This paper presents mCS-LM, a multilingual multimodal framework that integrates Large Language Models (LLMs), Visual Language Models (VLMs), Audio Language Models (ALMs) and Retrieval-Augmented Generation (RAG) within a modular and traceable architecture tailored to customer service and incident management. The system introduces complementary processing flows: (i) perception modules for visual and audio understanding aligned with LLM-based reasoning, and (ii) structured report generation from multimodal evidence through supervised fine-tuning using QLoRA and efficient adaptation techniques. To mitigate hallucinations and improve factual reliability, the framework incorporates vector databases and multimodal RAG pipelines that retrieve domain-specific knowledge from external corporate sources. Formal structural schemas and validation mechanisms enforce output consistency and syntactic correctness. The platform is deployed as a web-based system with REST API integration, enabling scalable multimodal interaction across channels such as instant messaging, email and web chat. Experimental results demonstrate that multimodal generative models can be specialized for structured, domain-constrained enterprise tasks while maintaining computational viability and robustness.

Keywords: Multimodal Conversational Systems, Customer Service Automation, Large Language Models, Visual Language Models, Audio Language Models, Retrieval-Augmented Generation, Hallucination Mitigation

1. Introduction

Customer service and incident management have undergone significant digital transformation in recent years, driven by the need to handle large volumes of interactions efficiently while maintaining quality, personalization and multilingual support (Balasubramanian et al., 2018). In sectors such as insurance, e-commerce, telecommunications and logistics, customer interactions increasingly involve multimodal evidence, including textual descriptions, images documenting damages and audio recordings (Cui et al., 2017). Small and medium-sized enterprises (SMEs), however, often lack the technical and computational resources required to deploy advanced AI-driven solutions capable of processing and reasoning over such heterogeneous data (Luccioni et al., 2023).

Recent advances in Large Language Models (LLMs) (Zhao et al., 2023), Visual Language Models (VLMs) (Gan et al., 2022) and Audio Language Models (ALMs) (Su et al., 2025) have demonstrated strong capabilities in conversational reasoning, visual understanding and speech processing. Nevertheless, the direct adoption of these models in regulated enterprise environments remains challenging (Eling and Lehmann, 2018). General-purpose models are prone to hallucinations, struggle to produce domain-constrained structured outputs, require costly computational infrastructure and often lack integration with external corporate knowledge sources (Stoeckli et al., 2018; Lewis et al., 2020). Furthermore, existing solutions typically address text, image or audio processing in isolation, rather than providing a unified multimodal framework (Shumanov and Johnson, 2021; Xu et al., 2017). This gap limits the development of reliable and controllable multimodal assistants suitable for real-world customer service and incident management workflows.

This paper presents the mCS-LM project, a multilingual (including Spanish and English) multimodal framework designed to integrate LLMs, VLMs, ALMs and Retrieval-Augmented Generation (RAG) within a modular and scalable architecture tailored to customer service and incident management. The system combines perception modules for text, image and audio processing with

Proceedings of LANLP: Bridging Ibero and Latin American NLP Communities, pages 38–45

structured generation mechanisms and multimodal RAG pipelines supported by vector databases to systematically mitigate hallucinations and enhance factual grounding. In addition, formal structural schemas, output validation layers and controlled generation constraints enforce syntactic correctness, domain alignment and semantic consistency, providing explicit safeguards against unreliable responses in high-stakes enterprise environments. The project consortiun is formed by Panel Sistemas S.L. (Spain), TECNOMOD research group at the University of Murcia (Spain), Medio Melón S.A. (Mexico) and Instituto Tecnológico Teziutlán (Mexico), as part of an International Technological Cooperation Project with unilateral certification and monitoring (reference UNI-20240017).

The resulting platform is deployed as a web-based system with REST API integration, enabling seamless interaction across channels such as instant messaging, email and web chat, while supporting multilingual and multimodal communication in enterprise environments. As detailed in Section 3, the architecture is organized into modular components that separate perception, reasoning, retrieval and generation processes, allowing flexible adaptation to different customer service domains such as insurance, e-commerce, telecommunications and logistics.

The overall architecture integrates specialized multimodal processing modules (see Section 3.1) responsible for handling text, image and audio inputs through LLM, VLM and ALM components, whose outputs are aligned within a unified semantic representation space. These modules are connected to a multimodal RAG layer (see Section 3.2), which integrates external corporate knowledge sources, such as policy documents, procedural manuals and internal databases, stored in vector databases for efficient semantic search over domain-specific corporate knowledge sources. This retrieval layer conditions generation and reinforces factual grounding.

To ensure reliability in regulated enterprise contexts, the system combines multiple complementary mechanisms for hallucination mitigation (see Section3.3). Beyond grounding responses through RAG, a dedicated hallucination control module introduces intention-aware routing and controlled generation strategies. This module performs prior intent classification and interaction-type identification, conditioning the response generation process and explicitly constraining the model's output space according to domain-specific schemas and predefined prompts. Together with validation layers and controlled decoding mechanisms, these safeguards mitigate hallucinations and prevent syntactic or semantic inconsistencies.

The web interface (Section 3.4) provides user-facing interaction and REST API services within a scalable microservices architecture, facilitating cloud-ready deployment and seamless integration into existing enterprise infrastructures. From the end-user perspective, customers interact directly with the multimodal conversational layer powered by LLM, VLM and ALM components, receiving automated responses and structured incident reports. In parallel, company agents access a dedicated dashboard focused on operational management. This interface provides real-time analytics and configurable Key Performance Indicators (KPIs) related to registered incidents, detected intents, resolution status and escalation cases. It also enables agents to monitor system activity, review AI-generated outputs and intervene in complex or high-risk interactions requiring human supervision. This dual-layer design separates automated conversational processing from operational oversight, ensuring both scalability and controlled human involvement.

The platform is currently in its final stages of development and is undergoing validation within real operational environments in collaboration with customer service professionals from the participating companies. These domain experts evaluate the system using representative real-world scenarios, including both common and complex incident cases, allowing iterative refinement of multimodal understanding, structured report generation and hallucination control mechanisms.

This practitioner-driven validation process ensures that the framework aligns with practical workflow requirements, domain-specific constraints and the most critical edge cases encountered in day-to-day operations. Within this scope, the present work emphasizes the design and integration of the proposed architecture, while more comprehensive quantitative evaluation, including benchmarking against baseline models, is part of ongoing work as the system continues to evolve within real operational settings.

2. Background Information

The development of advanced customer service and incident management systems is driven by recent progress in Natural Language Processing (NLP) (Vaswani et al., 2017; Devlin et al., 2019) and multimodal Artificial Intelligence, including vision (He et al., 2016) and audio (Yang et al., 2025). LLMs have demonstrated remarkable capabilities in natural language understanding, contextual reasoning and text generation, enabling more flexible and conversational interfaces compared to traditional rule-based systems (Brown et al., 2020; Achiam et al., 2023). Transformer-based architectures have become the dominant paradigm, supporting multilingual interaction and domain adap-

tation through fine-tuning and in-context learning strategies (Radford et al., 2021; Li et al., 2023). However, their deployment in regulated enterprise environments introduces challenges related to hallucinations, computational cost and controllability (Balasubramanian et al., 2018; Huang et al., 2025).

In parallel, VLMs extend language reasoning to visual inputs, integrating image encoders with language decoders to enable multimodal understanding and generation (Radford et al., 2021). These models allow the interpretation of images such as damaged products, infrastructure incidents or technical documentation, transforming visual evidence into structured textual representations (Li et al., 2023). Similarly, ALMs leverage pretrained speech encoders such as Whisper or Wav2Vec to transcribe and interpret spoken interactions (Su et al., 2025), supporting voice-based incident reporting and customer support scenarios (Gung et al., 2023). Despite these advances, multimodal models are typically optimized for open-ended tasks rather than domain-constrained structured generation required in enterprise workflows.

To address factual inconsistencies and hallucinations, RAG architectures have gained prominence (Lewis et al., 2020). By combining generative models with external knowledge retrieval mechanisms supported by vector databases and semantic search, RAG systems improve grounding and ensure access to up-to-date domain-specific information. Additionally, model optimization techniques such as quantization, distillation and parameter-efficient fine-tuning methods like LoRA (Hu et al., 2022) and QLoRA (Dettmers et al., 2023) have been proposed to reduce computational requirements while maintaining performance, facilitating deployment in SMEs environments.

Although significant progress has been made in text, vision and speech processing independently, comprehensive frameworks that integrate LLMs, VLMs, ALMs and RAG within a unified, controllable and enterprise-ready architecture remain limited. Existing solutions often lack explicit mechanisms for structured output enforcement, hallucination mitigation and seamless integration into operational customer service infrastructures. This gap motivates the development of a modular multimodal framework capable of combining perception, retrieval, controlled generation and deployment scalability within a single system.

3. System Architecture

Figure 1 shows the overall architecture of the system. The platform is organized into five main modules. The first module corresponds to the multimodal processing layer (see Section 3.1), which is responsible for handling text, image and audio inputs through domain adjusted LLM, VLM and ALM models. This module performs perception, encoding and semantic alignment of heterogeneous data sources, transforming multimodal inputs into unified representations suitable for downstream reasoning.

The second module focuses on the RAG pipeline (see Section 3.2), which connects the generative models with external corporate knowledge bases. Through vector databases and semantic search mechanisms, this layer retrieves relevant domain-specific information, such as policy documents and procedural manuals, to ground responses and enhance factual consistency.

The third module addresses structured output and hallucination control (see Section 3.3). In addition to enforcing domain-specific schemas, controlled decoding strategies and validation mechanisms, this module incorporates an intention-aware classification and routing layer that conditions the response generation process before decoding takes place. By identifying the interaction type and constraining the generation space through predefined prompts and domain-aligned templates, the system explicitly restricts the model's behavior, reducing the likelihood of producing unsupported, out-of-scope or fabricated content. These combined safeguards ensure syntactic correctness, semantic consistency and reliable responses in regulated enterprise contexts.

The fourth module corresponds to the incident management system interface (see Section 3.4), which provides the user-facing conversational layer and REST API services. This component enables seamless integration with enterprise communication channels such as instant messaging, email and web chat, supporting multilingual (including Spanish and English) and multimodal interaction workflows.

The fifth module corresponds to the operational dashboard (see Section 3.5), designed for company agents and supervisors. This interface provides real-time analytics and configurable KPIs related to registered incidents, detected intents, resolution status and escalation cases. It enables monitoring of system activity, review of AI-generated outputs and human intervention in complex or high-risk interactions. By separating conversational automation from operational oversight, the architecture ensures scalability while maintaining controlled human involvement in enterprise workflows. The following subsections describe these modules in more detail.

The platform is designed to be offered under a Software-as-a-Service (SaaS) model, allowing flexible adoption by SMEs. Basic functionality supports multimodal interaction and structured incident reporting, while advanced features, such as extended knowledge base integration, customiz-

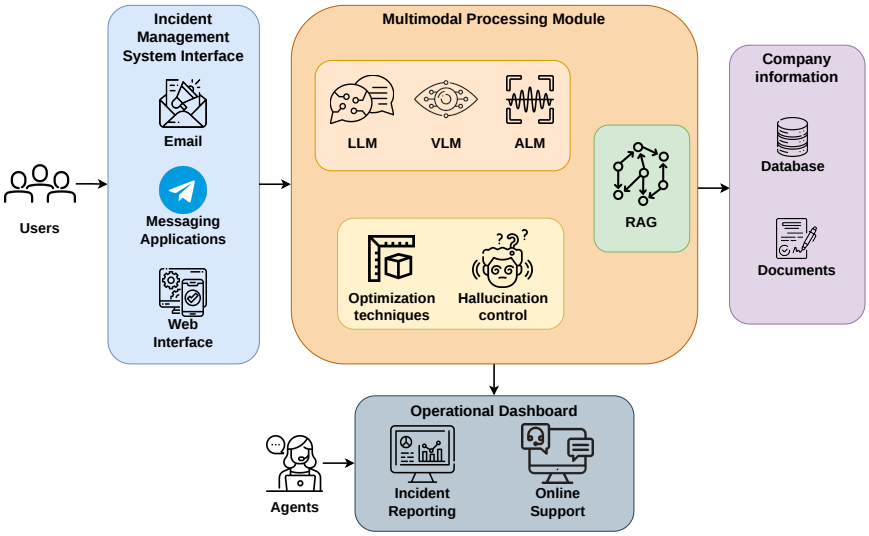


Figure 1: Overview of the modules that conform the mCS-LM system architecture.

able workflows and enterprise-level scalability, can be configured according to organizational requirements. All services and software components are containerized using Docker, enabling modular deployment, horizontal scalability and flexible orchestration depending on performance constraints and infrastructure availability.

3.1. Multimodal Processing Modules

The multimodal processing layer constitutes the perception and semantic encoding backbone of the mCS-LM architecture. It is responsible for handling heterogeneous inputs (text, images, and audio) through specialized domain-adapted models, enabling unified downstream reasoning and structured response generation.

To ensure computational viability in enterprise environments, all multimodal components, including LLMs, VLMs and ALMs, are optimized through techniques such as quantization and parameter-efficient fine-tuning using QLoRA. These strategies reduce hardware requirements while preserving performance, facilitating deployment in SMEs infrastructures.

Textual interactions are processed through LLMs fine-tuned for customer service and incident management tasks. Domain adaptation is achieved through parameter-efficient fine-tuning strategies, including QLoRA, allowing the system to specialize in policy interpretation, claim management and incident reporting scenarios while maintaining computational efficiency.

Visual inputs are handled by VLMs to transform visual evidence, such as photographs of damages, defective products or infrastructure incidents, into structured textual representations. The VLM component operates under two complementary processing strategies. In the descriptive flow, visual content is first converted into a controlled textual description that is subsequently processed by the LLM reasoning module, separating perception from linguistic inference. In the generative flow, a fine-tuned VLM produces domain-specific structured outputs (e.g., JSON incident reports) from raw images. This dual design enables both modular reasoning over image-derived descriptions and end-to-end structured generation aligned with predefined reporting schemas.

Audio interactions are processed through ALMs to extract contextual embeddings from spoken inputs. These embeddings are projected into a shared semantic space aligned with the LLM decoder, allowing the system to perform transcription, intent understanding and response generation from voice-based interactions. This design supports multilingual speech inputs and enables seamless integration of audio within multimodal conversational workflows.

To ensure coherent multimodal reasoning, the outputs of the LLM, VLM and ALM components are aligned within a unified semantic representation space. This alignment facilitates late-fusion integration strategies, enabling the system to combine textual, visual and acoustic information before retrieval and controlled generation stages. By separating perception from reasoning and maintaining modular specialization of each modality, the architecture achieves both flexibility and scalability across diverse enterprise domains.

3.2. Retrieval-Augmented Generation Pipeline

The RAG pipeline enhances the generative capabilities of the multimodal models by grounding responses in external domain-specific knowledge sources. Instead of relying exclusively on the parametric knowledge encoded within the LLM, the system dynamically retrieves relevant information from corporate repositories before response generation.

Domain documents, including policy manuals, procedural guidelines and internal databases, are preprocessed, segmented and embedded into vector representations using sentence-level embedding models. These embeddings are stored in vector databases that enable efficient semantic search through similarity-based retrieval mechanisms. When a user query is received, its representation is used to retrieve the most relevant contextual fragments, which are then incorporated into the generation prompt.

To ensure that retrieved information remains up-to-date in dynamic enterprise environments, the system supports incremental updates of the vector database. New or modified documents can be embedded and indexed without requiring full re-indexing, enabling efficient maintenance of the knowledge base. Additionally, document-level metadata, such as timestamps and versioning information, is incorporated into the retrieval process to prioritize more recent and relevant content.

When potentially conflicting information is retrieved (e.g., legacy documents versus updated protocols), the system applies a ranking strategy that combines semantic similarity, document recency and source reliability. This prioritization ensures that the most current and authoritative information is used to condition generation, while the structured output module enforces consistency in the final response delivered to the user.

This retrieval layer conditions the decoding process, reinforcing factual consistency and reducing the likelihood of hallucinated or outdated responses. By integrating structured corporate knowledge into the generation workflow, the system ensures that responses remain aligned with current policies, operational procedures and domain constraints.

The architecture also supports multimodal retrieval scenarios, allowing textual queries to be associated with information derived from images or audio metadata when available. This design extends traditional text-only RAG approaches and enables unified grounding across heterogeneous enterprise data sources.

3.3. Structured Output and Hallucination Control

Ensuring reliability in regulated enterprise environments requires explicit mechanisms to control the behavior of generative models. Beyond retrieval-based grounding, the mCS-LM framework incorporates a dedicated structured output and hallucination control module designed to constrain generation and reduce unsupported or out-of-scope responses.

The control pipeline introduces a hierarchical intention-aware classification process prior to response generation. Each incoming user message is first processed by a global intent classifier that performs a high-level categorization of the interaction (e.g., conversational intent, domain-specific action, policy-related claim, out-of-scope request or unspecified issue). This initial classification determines whether the interaction can be directly mapped to a predefined response strategy or requires further specialized analysis.

Messages assigned to intermediate categories are routed to a second-stage classifier tailored to the corresponding interaction type. For instance, conversational intents are processed by a dialogue-oriented classifier, while domain-specific actions related to incident management or policy handling are redirected to dedicated domain classifiers. This two-step classification strategy reduces task complexity, isolates interaction types and provides finer control over subsequent generation stages.

The outcome of this hierarchical classification process conditions the prompt selection and generation strategy used by the LLM responsible for producing the final response. Instead of generating outputs directly from the raw input, the model operates under predefined domain-aligned templates and constrained decoding configurations associated with the detected interaction type. This intention-aware routing mechanism explicitly restricts the model's output space, reducing the likelihood of hallucinated, ambiguous or semantically inconsistent responses.

In addition, structured generation schemas are enforced for tasks requiring formal outputs, such as incident reports or claim summaries. Validation layers verify syntactic correctness and schema compliance before responses are delivered to the user or forwarded to downstream systems. Together, hierarchical intent classification, controlled prompt selection, constrained decoding and schema validation form a multi-layer safeguard strategy that enhances reliability, interpretability and domain alignment in enterprise workflows.

3.4. Incident Management System Interface

The incident management system interface constitutes the user-facing interaction layer of the mCS-LM framework. It enables customers to report incidents, submit supporting evidence and receive automated assistance through multimodal conversational workflows.

The interface supports text, image and audio inputs across multiple communication channels, including web chat, email and instant messaging platforms. Incoming interactions are routed to the appropriate multimodal processing modules, triggering the perception, retrieval and controlled generation pipeline described in previous sections. The system manages conversational context, session state and interaction history to ensure coherent multi-turn exchanges.

Through REST API services, the interface can be integrated into existing enterprise infrastructures, customer portals or third-party applications. This API-driven design allows flexible orchestration of requests, enabling automated ticket creation, structured incident reporting and interaction logging within corporate management systems.

Multilingual support is provided at both the understanding and generation levels, allowing customers to interact in different languages without requiring manual intervention. By abstracting the complexity of the underlying multimodal and retrieval components, the interface delivers a seamless conversational experience while maintaining alignment with domain constraints and enterprise policies.

3.5. Operational Dashboard

The operational dashboard provides the management and supervision layer of the mCS-LM framework, designed for company agents and administrative personnel. While the incident management system interface enables automated multimodal interaction with end users, the dashboard focuses on operational oversight, monitoring and human intervention when required.

The dashboard offers real-time visibility into registered incidents, detected intents, resolution status and escalation cases through configurable KPIs. These analytics allow supervisors to monitor system performance, identify high-risk or ambiguous interactions and assess workload distribution across incident categories. Interaction-level metadata, including intent classifications and retrieval traces, can be inspected to enhance transparency and traceability.

In addition to monitoring capabilities, the dashboard enables human-in-the-loop intervention. Agents can review AI-generated outputs, validate structured reports, modify responses when necessary and manually escalate complex or sensitive cases. This supervision layer ensures that automated decision-making remains aligned with organizational policies and regulatory requirements.

The operational dashboard is integrated within the same microservices architecture as the conversational interface, allowing synchronized access to interaction logs, structured reports and knowledge base updates. By separating conversational automation from operational management, the system achieves scalability without sacrificing control, accountability or human expertise in critical enterprise workflows.

4. Conclusions and Further Work

The current version of mCS-LM integrates the core modules required for multimodal incident management, including domain-adapted LLM, VLM and ALM components, a RAG pipeline, a structured output and hallucination control module and dedicated interfaces for both end users and operational agents. The architecture combines perception, retrieval, controlled generation and human-in-the-loop supervision within a modular and scalable microservices framework suitable for enterprise environments.

The platform is currently undergoing validation in real operational settings with customer service professionals from participating companies. This validation phase focuses on assessing multimodal understanding accuracy, structured report consistency and the effectiveness of hierarchical intent classification in reducing hallucinations and out-of-scope responses. In this context, the present work emphasizes the design and integration of the proposed architecture, while more comprehensive quantitative evaluation, including benchmarking against baseline models, is part of ongoing work as the system continues to evolve within real operational settings.

Future work will focus on several directions. First, we plan to enhance the multimodal RAG pipeline by incorporating more advanced cross-modal retrieval strategies, enabling tighter alignment between textual queries and visual or audio-derived knowledge representations, as highlighted in recent studies on cross-modal RAG systems (Abootorabi et al., 2025). Second, we will explore adaptive intent classification mechanisms based on continual learning, allowing the hallucination control module to evolve as new interaction patterns emerge, in line with emerging research on continual learning for generative AI frameworks (Wang et al., 2024). Third, we aim to investigate automated feedback loops between the operational dashboard and the generation modules, leveraging agent corrections to refine prompts and improve structured output quality over

time, a strategy that aligns with recent proposals for integrating human feedback into generative systems (Sun et al., 2023).

These developments will strengthen the robustness, scalability, and practical applicability of multimodal conversational systems in regulated enterprise contexts. Additionally, future work will explore the extension of the proposed framework to low-resource settings, including regional languages and dialectal variations, leveraging its modular and language-agnostic design to support broader linguistic diversity and robustness in real-world multilingual scenarios.

Acknowledgements

This work is being funded by CDTI and the European Regional Development Fund (ERDF EU-/FEDER UE)-a way of making Europe, through project mCS-LM IDI-20250122.

5. Ethical Considerations and Limitations

The data used for system validation and adaptation consist of enterprise-provided incident records, policy documents and internally generated test cases. All materials are handled within controlled corporate environments in accordance with applicable data protection regulations. No personal or sensitive information is publicly released as part of this research.

The system is designed to operate within enterprise infrastructures, ensuring that multimodal inputs (including text, images and audio) are processed under organizational data governance policies. When required, anonymization and access control mechanisms are applied to prevent unauthorized exposure of customer information.

6. Bibliographical References

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. 2025. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16776–16809.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ramnath Balasubramanian, Ari Libarikian, and Doug McElhaney. 2018. Insurance 2030—the impact of ai on the future of insurance. *McKinsey & Company*, pages 1–10.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Lei Cui, Shaohan Huang, Furu Wei, Chuanqi Tan, Chaoqun Duan, and Ming Zhou. 2017. Superagent: A customer service chatbot for e-commerce websites. In *Proceedings of ACL 2017, system demonstrations*, pages 97–102.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Martin Eling and Martin Lehmann. 2018. The impact of digitalization on the insurance value chain and the insurability of risks. *The Geneva papers on risk and insurance-issues and practice*, 43(3):359–396.

Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. 2022. Vision-language pre-training: Basics, recent advances, and future trends. *arXiv preprint arXiv:2210.09263*.

James Gung, Emily Moeng, Wesley Rose, Arshit Gupta, Yi Zhang, and Saab Mansour. 2023. Natcs: Eliciting natural customer support dialogues. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9652–9677.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang,

Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

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*, 33:9459–9474.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2023. Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of machine learning research*, 24(253):1–15.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Michael Shumanov and Lester Johnson. 2021. Making conversations with chatbots more personalized. *Computers in human behavior*, 117:106627.

Emanuel Stoeckli, Christian Dremel, and Falk Uebernickel. 2018. Exploring characteristics and transformational capabilities of insurtech innovations to understand insurance value creation in a digital world. *Electronic markets*, 28(3):287–305.

Yi Su, Jisheng Bai, Qisheng Xu, Kele Xu, and Yong Dou. 2025. Audio-language models for audio-centric tasks: A survey. *arXiv preprint arXiv:2501.15177*.

Xin Sun, Jos A Bosch, Jan De Wit, and Emiel Krahmer. 2023. Human-in-the-loop interaction for continuously improving generative model in conversational agent for behavioral intervention. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 99–101.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2024. A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5362–5383.

Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 3506–3510.

Chih-Kai Yang, Neo S Ho, and Hung-yi Lee. 2025. Towards holistic evaluation of large audio-language models: A comprehensive survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10155–10181.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

SAFEWORDs: un marco reproducible para anonimización conforme al RGPD y evaluación de generación en lenguas cooficiales

Rafael Muñoz[1], **Manuel Palomar**[1], **Elena Lloret**[1], **Nuria Fernández**[1]
[1] CENID-Universidad de Alicante (UA)
rafael@dlsi.ua.es, mpalomar@dlsi.ua.es
elloret@dlsi.ua.es, nuria.fernandez@ua.es

Abstract

Los Grandes Modelos de Lenguaje (LLMs) abren oportunidades para el Procesamiento del Lenguaje Natural (PLN) en contextos institucionales, si bien plantean riesgos críticos en entornos regulados y multilingües, especialmente en lo relativo a protección de datos personales, trazabilidad de decisiones y equidad entre lenguas con distinta disponibilidad de recursos. Presentamos SAFEWORDs, proyecto que acaba de iniciarse en el marco del proyecto coordinado *"HumanAIze"* (Plan Nacional de Inteligencia Artificial 2025, España), que propone un marco reproducible de *privacy-by-design* y *ethics-by-design* para la evaluación y alineación de LLMs en las lenguas oficiales de la Península Ibérica (español, catalán, valenciano, gallego y euskera). El marco integra: (i) anonimización automática conforme al RGPD, con protocolos explícitos de detección de fuga residual y verificación adversarial; (ii) transformación orientada a la accesibilidad textual y al lenguaje claro; y (iii) evaluación en el dominio biomédico, donde la sensibilidad de los datos y la precisión terminológica exigen mecanismos adicionales de control generativo. Desde el punto de vista metodológico, se comparan configuraciones *zero-shot* y *few-shot*, y se documentan prompts, hiperparámetros y recursos para facilitar la replicabilidad y la gobernanza de recursos. Además de sintetizar resultados de referencia de la literatura para contextualizar métricas y órdenes de magnitud esperables, el trabajo discute implicaciones éticas y limitaciones del enfoque propuesto. La propuesta se alinea con las líneas de trabajo de SEPLN y con los objetivos de LANLP, al establecer protocolos transferibles para el desarrollo de tecnologías lingüísticas confiables en ecosistemas caracterizados por variación dialectal y lenguas infrarepresentadas.

Keywords: modelos de lenguaje; anonimización; RGPD; lenguaje claro; dominio biomédico; lenguas ibéricas; evaluación multilingüe; IA confiable

1. Introducción

El desarrollo reciente de los Grandes Modelos de Lenguaje (LLMs) ha ampliado significativamente el alcance del Procesamiento del Lenguaje Natural (PLN), posibilitando aplicaciones avanzadas en contextos institucionales, administrativos y científicos. Sin embargo, su integración en entornos europeos regulados exige garantizar principios de robustez, legalidad, transparencia y respeto a los derechos fundamentales, especialmente en lo que concierne a la protección de datos personales y la equidad lingüística.

El proyecto HumanAIze (Plan Nacional de Inteligencia Artificial 2025) aborda estos retos desde la perspectiva de una inteligencia artificial centrada en el ser humano, con especial atención a las lenguas de la Península Ibérica: español, catalán, valenciano, gallego y euskera. Este enfoque reconoce la diversidad lingüística del territorio y la necesidad de evaluar el comportamiento de los modelos de lenguaje en escenarios multilingües con distintos grados de disponibilidad de recursos.

En este marco, SAFEWORDs constituye la contribución específica del equipo de la Universidad de Alicante dentro de HumanAIze. Su objetivo es desarrollar y validar un marco metodológico para la evaluación y alineación de LLMs en contextos regulados y multilingües, tomando como referencia tres casos de uso representativos:

1. **Anonimización automática conforme al RGPD**, aplicada a textos administrativos, jurídicos y sanitarios en lenguas peninsulares.

2. **Transformación orientada a la mejora de la accesibilidad textual**, con especial atención a escenarios de lenguaje claro en documentos institucionales.

3. **Aplicación en el dominio biomédico**, donde la sensibilidad de los datos y la precisión terminológica requieren mecanismos robustos de generación y control de salida.

Estos tres casos de uso comparten un núcleo metodológico común: la necesidad de evaluar sistemáticamente el comportamiento de los LLMs bajo restricciones normativas y lingüísticas explícitas, comparando distintas configuraciones experimentales y analizando riesgos como la fuga de información, la inconsistencia interlingüística o la degradación semántica.

SAFEWORDs se concibe, por tanto, no como una aplicación aislada, sino como un marco experimental orientado a la evaluación reproducible y la gobernanza de recursos lingüísticos en entornos de alto impacto social.

Proceedings of LANLP: Bridging Ibero and Latin American NLP Communities, pages 46–54

2. Trabajo Relacionado

El marco propuesto en SAFEWORDs se sitúa en la intersección de cuatro líneas de investigación activas: (i) alineación normativa de modelos generativos, (ii) anonimización automática y reconocimiento de entidades nombradas, (iii) generación controlada y accesibilidad textual en dominios especializados, y (iv) desarrollo reciente de modelos de lenguaje para lenguas ibéricas.

2.1. LLMs y alineación normativa

La literatura reciente ha puesto de relieve los riesgos estructurales asociados al despliegue de modelos fundacionales en contextos regulados. Bender et al. (2021) identifican problemas sistémicos relacionados con sesgos y falta de control en modelos generativos de gran escala. Desde la perspectiva de la seguridad, Carlini et al. (2021) demostraron empíricamente que estos modelos pueden memorizar y revelar datos sensibles presentes en los datos de entrenamiento, con implicaciones directas para el cumplimiento del Reglamento General de Protección de Datos (RGPD) (Parlamento Europeo and Consejo de la Unión Europea, 2016) y de la Ley Orgánica de Protección de Datos y Garantía de los Derechos Digitales (Jefatura del Estado (España), 2018).

Estas preocupaciones resultan especialmente relevantes en entornos institucionales europeos, donde la generación automática debe evaluarse bajo criterios explícitos de robustez, legalidad y minimización del riesgo.

2.2. Anonimización y reconocimiento de entidades

La relación entre los Grandes Modelos de Lenguaje (LLMs) y la anonimización de datos atraviesa una fase de transformación dual entre 2024 y 2026. Por un lado, los LLMs han demostrado ser herramientas de anonimización avanzadas capaces de superar las limitaciones de los métodos tradicionales basados en reglas o Reconocimiento de Entidades Nombradas (NER) simple, gracias a su capacidad para comprender el contexto y realizar una redacción adaptativa de información personal identificable (PII) (Staab et al., 2024; Ponomarenko et al., 2026). Investigaciones recientes destacan el auge de la "anonimización adversarial", donde se utilizan modelos para identificar inferencias indirectas que podrían conducir a la reidentificación, permitiendo así proteger datos que parecen anónimos pero que son vulnerables ante la capacidad deductiva de modelos avanzados (Miranda et al., 2024; Zamroz and Morozov, 2024).

Sin embargo, el despliegue de estas tecnologías enfrenta desafíos críticos, principalmente el equilibrio entre privacidad y utilidad. Mientras que técnicas como la Privacidad Diferencial (DP) y el uso de Modelos de Lenguaje Pequeños (SLMs) locales ofrecen garantías de seguridad superiores y cumplimiento regulatorio (como GDPR o HIPAA en entornos clínicos), suelen conllevar una degradación en la calidad de las respuestas o un aumento en los costos operativos (Garza et al., 2025; Yang et al., 2024). Además, persiste una preocupación significativa sobre la "capacidad inferencial" de los modelos propietarios cerrados, lo que ha impulsado el desarrollo de marcos de trabajo de código abierto y arquitecturas *self-hosted* que permiten procesar datos sensibles sin exponerlos a nubes de terceros, garantizando así una soberanía de datos real en sectores altamente regulados (Jonnagaddala and Wong, 2025; Mancera et al., 2025).

La desidentificación automática en textos clínicos ha sido ampliamente estudiada mediante modelos neuronales supervisados (Dernoncourt et al., 2017). No obstante, la incorporación de LLMs generativos introduce nuevos desafíos, como la generación inconsistente de sustitutos léxicos o la persistencia de información residual que permita la reidentificación indirecta.

2.3. Generación controlada y accesibilidad en el dominio biomédico

La simplificación textual automática ha sido extensamente estudiada en inglés (Alva-Manchego et al., 2020; Shardlow, 2014). En español, trabajos recientes publicados en *Procesamiento del Lenguaje Natural* han abordado la simplificación en contextos de salud mediante modelos basados en transformers. En particular, el ajuste fino de modelos BART para la simplificación de textos sanitarios (Alarcón et al., 2023) y la construcción de corpus comparables para la evaluación en salud (Campillos-Llanos et al., 2022) evidencian avances significativos en generación controlada en dominio especializado.

SAFEWORDs se sitúa en esta línea de investigación, integrando accesibilidad textual y generación en el dominio biomédico dentro de un marco unificado de evaluación normativa y lingüística.

2.4. Lenguas ibéricas y evaluación multilingüe

El desarrollo de infraestructuras multilingües para lenguas peninsulares ha sido impulsado, entre otras iniciativas, por El proyecto ILENIA[1] ha impulsado el desarrollo de una familia de modelos de lenguaje entrenados sobre grandes volúmenes

[1] https://proyectoilenia.es

de datos: SALAMANDRA (34 idiomas) (Gonzalez-Agirre et al., 2025), AITANA (valenciano) (GPLSI – Language and Information Systems Group, University of Alicante, 2026), LATXA (euskera) (Etxaniz et al., 2024), CARVALHO (gallego) (Gamallo et al., 2024) y ANIA (catalán) (González-Agirre et al., 2024). Estos modelos sientan una base sólida para el desarrollo de sistemas instruidos en tareas concretas con resultados excelentes en tareas como generación, traducción y su aplicación en casos de usos reales.

3. Marco metodológico

SAFEWORDs (UA) contribuye a HumanAIze con un marco metodológico de *ethics-by-design* y *privacy-by-design* para el desarrollo, control y evaluación de LLMs en las lenguas oficiales de España. Este marco operacionaliza directrices y umbrales ético-legales en forma de *checklists* y puntos de control aplicables a datos, modelos y procedimientos de evaluación (WP3–WP4), tal como se ilustra en la Figura 1.

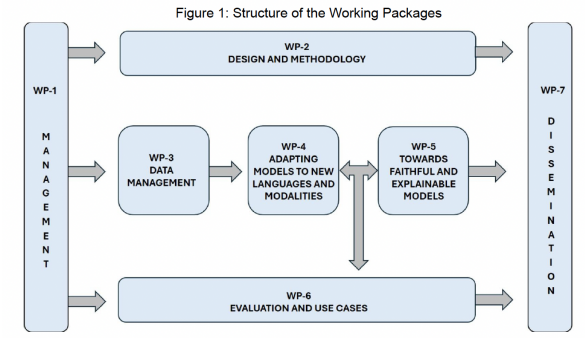


Figure 1: Paquetes de trabajo del proyecto HumanAIze.

El marco se implementa como un **pipeline trazable** que conecta cuatro etapas: (i) adquisición y curación de datos, (ii) anonimización y mitigación de riesgos de reidentificación, (iii) alineación mediante instrucciones y evaluación comparativa, y (iv) verificación previa a la liberación de recursos y modelos, bajo umbrales estrictos de privacidad y equidad.

En coherencia con el planteamiento del proyecto, la evaluación se diseña explícitamente en escenarios **zero-shot** y **few-shot**, comparando distintas configuraciones de prompts y verificando el cumplimiento del RGPD y el EU AI Act, dado que la verificación automática completa sigue siendo un problema abierto en el estado del arte.

3.1. Diseño transversal: configuraciones y reproducibilidad

Para cada caso de uso se comparan, como mínimo, las siguientes configuraciones:

- **Zero-shot**: instrucciones directas con restricciones normativas explícitas.
- **Few-shot**: incorporación de ejemplos representativos, multilingües cuando sea posible.
- **Ablaciones de guardrails**: con y sin módulos de control (p. ej., validadores de formato, listas de exclusión léxica, detectores de entidades nombradas).

La reproducibilidad se garantiza mediante:

- Versionado de datos, prompts y configuraciones experimentales.
- Registro sistemático de hiperparámetros (modelo, temperatura, longitud máxima de contexto, etc.).
- Firma y trazabilidad de recursos (datos, prompts, salidas) como parte del control de procedencia y cadena de custodia.

3.2. Pipeline de privacidad y verificación RGPD/EU AI Act

Adoptamos un enfoque **privacy-by-design** en el que cada corpus de entrada es analizado mediante reconocedores multilingües para detectar:

- Identificadores directos (correos electrónicos, identificadores fiscales, números de teléfono, direcciones postales, etc.).
- *Contextual cues* que pueden facilitar ataques de reidentificación por enlace (p. ej., la combinación de cargo profesional, localidad y fecha de evento).

Cuando resulta técnicamente viable, las referencias a datos personales son:

- **eliminadas** mediante purga directa, o
- **reemplazadas** por **sustitutos sintéticos** (*surrogates*) generados bajo presupuestos de privacidad diferencial, preservando la utilidad textual (p. ej., coherencia morfológica de género y número) y registrando el proceso para garantizar la trazabilidad.

El pipeline se registra de forma que permita cumplir con las obligaciones de trazabilidad y transparencia establecidas en el marco regulatorio vigente.

3.2.1. Protocolo de detección de *fuga residual* y verificación de cumplimiento

Definimos **fuga residual** (*residual leakage*) como la presencia en las salidas del modelo de información que permita: (i) identificar directa o indirectamente a una persona física, o (ii) recuperar datos personales mediante inferencias o ataques de enlace. La evaluación combina tres niveles de verificación:

(1) Verificación automática

- Detección post-generación mediante herramientas de NER y expresiones regulares multilingües orientadas a información de identificación personal (PII).
- Validadores de formato: patrones de DNI/NIE, números de teléfono, correos electrónicos y códigos postales.
- Heurísticas de enlace: detección de co-ocurrencias infrecuentes que puedan resultar identificadoras (p. ej., combinación única de profesión, micro-localización y fecha).

(2) Verificación adversarial Se generan *prompts de ataque* diseñados para inducir la revelación de PII:

- Consultas de recuperación directa ("¿quién es la persona X?", "proporciona el nombre completo").
- Reformulaciones orientadas a forzar la generación de detalles identificadores ("amplía la información anterior").
- Ataques por consistencia: generación múltiple de la misma entrada para detectar divergencias reveladoras.

(3) Auditoría humana Muestreo estratificado de salidas (por lengua, dominio y tipo de PII) con los objetivos de:

- Identificar falsos negativos de los detectores automáticos.
- Evaluar el riesgo de enlace y reidentificación contextual en casos límite.

Métricas de evaluación Se reportan: (i) tasa de fuga residual, (ii) tasa de falsos negativos del detector de PII, (iii) severidad de la fuga (PII directa frente a inferencial), y (iv) degradación de utilidad textual (comparación entre muestras anonimizadas y originales mediante métricas automáticas y valoración humana).

3.3. Caso de uso A: Lenguaje claro en el dominio administrativo-jurídico

HumanAIze define explícitamente que la producción de lenguaje claro se abordará mediante la transformación de rasgos lingüísticos en **tres niveles**: discursivo, morfosintáctico y léxico, y que los desarrollos resultantes se incorporarán a la herramienta **arText** (da Cunha, 2022) como vía de transferencia tecnológica.

Diseño y evaluación

- Corpus anotados manualmente para medir la adecuación de las salidas generadas.
- Evaluación automática basada en precisión y recall sobre transformaciones etiquetadas.
- Estudios con usuarios reales para contrastar la claridad percibida y la comprensión efectiva de los textos transformados.

3.4. Caso de uso B: Anonimización para la preservación de la privacidad

Este escenario evalúa técnicas de anonimización asistidas por LLMs sobre textos legales y administrativos en lenguas peninsulares. Se comparan distintas configuraciones de prompt en modalidades zero-shot y few-shot, con verificación explícita del cumplimiento de los requisitos de privacidad establecidos por el RGPD y el EU AI Act.

El diseño experimental contempla:

- Comparativa de técnicas (enmascarado, sustitución léxica y generación de sustitutos sintéticos) y sus posibles combinaciones.
- Evaluación del cumplimiento normativo, verificando que las salidas no contengan datos personales no anonimizados.
- Introducción de documentos con alta densidad de información sensible como casos de estrés del sistema (*adversarial inputs*).

3.5. Caso de uso C: Dominio biomédico

La evaluación en el dominio biomédico se realiza sobre corpus anotados y validados por expertos clínicos. Se incorporan métricas específicas del dominio, protocolos de *test* rigurosos y análisis comparativo con *benchmarks* establecidos en la literatura, incluyendo *shared tasks* relevantes cuando aplique.

4. Resultados de referencia en la literatura

Dado que SAFEWORDs se presenta en este trabajo como una contribución metodológica y de diseño evaluativo en el marco de HumanAIze, incluimos resultados de referencia procedentes de la literatura reciente (principalmente para español y lenguas cooficiales en España) con el fin de contextualizar el tipo de métricas y los órdenes de magnitud esperables en las tareas relacionadas con nuestros casos de uso. Esta sección no reporta resultados propios del proyecto, sino que sintetiza *baselines* y recursos publicados disponibles en el momento de la escritura de este trabajo.

En español, el desarrollo de modelos de lenguaje específicos ha avanzado de forma notable con iniciativas como ILENIA[2] en el que se pueden encontrar una gran diversidad d edatasets, modelos texto y voz, y demostradores. En tareas de accesibilidad en salud, se han publicado enfoques basados en ajuste fino de BART/mBART con resultados expresados en términos de SARI, junto con evidencias de mejora en escalas de legibilidad estándar. La construcción de corpus comparables de simplificación médica en español ha permitido, asimismo, disponer de *benchmarks* de referencia a escala considerable.

4.1. Implicaciones para SAFEWORDs

Estos resultados respaldan dos decisiones metodológicas centrales del equipo UA en SAFEWORDs: (i) la necesidad de protocolos de evaluación que combinen métricas automáticas y verificación cualitativa en tareas de accesibilidad y dominio biomédico, y (ii) la viabilidad de enfoques adaptados a bajo recurso en lenguas cooficiales mediante generación de datos sintéticos y evaluación con métricas estandarizadas.

En fases posteriores del proyecto, SAFEWORDs utilizará estos resultados y órdenes de magnitud como referencias para diseñar comparativas y análisis de robustez bajo restricciones normativas (anonimización conforme al RGPD) y lingüísticas (español, catalán/valenciano, gallego y euskera).

5. Recursos, gestión de datos y liberación de recurso

En coherencia con los principios de HumanAIze, SAFEWORDs (UA) no se limita a evaluar modelos, sino que contribuye a un ciclo completo de *data governance* y liberación controlada de activos, con el objetivo de facilitar la adopción científica e institucional de tecnologías lingüísticas en las lenguas oficiales de España. Este planteamiento se articula en torno a dos pilares: (i) un repositorio multilingüe y versionado de datos de tarea e instrucción, y (ii) un conjunto de umbrales y listas de verificación en materia de privacidad y equidad que actúan como compuertas de calidad (*quality gates*) antes de la reutilización o difusión de cualquier activo.

5.1. Repositorio multilingüe y versionado (datos de tarea e instrucción)

HumanAIze establece WP3 como el núcleo de gestión de datos del proyecto, con un portfolio balanceado de corpus (desde *crawls* web filtrados hasta colecciones específicas de dominio) y un pipeline que registra procedencia, licencia y propiedades estadísticas, garantizando la trazabilidad y la comparabilidad longitudinal entre versiones.

Antes de liberar cualquier activo del proyecto (datos, instrucciones, modelos o herramientas), se verifica que cumple umbrales estrictos de privacidad y equidad, como parte del proceso de gobernanza versionada definido en WP3 y alineado con las directrices del WP2.

Dentro de este marco, SAFEWORDs contribuye de forma explícita a:

- Establecer y curar un **repositorio multilingüe, versionado y auditado** de datos de tarea e instrucción que cubra dominios prioritarios y las lenguas oficiales de España (T3.1, T3.3).
- Verificar que **cada activo** cumple umbrales estrictos de privacidad y equidad antes de su uso en entrenamiento o evaluación (T3.4, T3.5 y tareas relacionadas en WP5).

5.2. Privacidad, trazabilidad y control de riesgo

La memoria del proyecto define un enfoque **trustworthy-by-design** y **privacy-by-design** en el que cada corpus de entrada es auditado automáticamente y complementado con revisiones humanas en casos límite, traduciendo los requisitos del RGPD y del EU AI Act a *checklists* operativas para la adquisición, retención, archivo y acceso controlado a datos.

Este marco incorpora mecanismos explícitos para:

- Mitigar riesgos de reidentificación mediante pruebas de estrés (p. ej., *membership inference*, *attribute inference* y *record linkage*) y, cuando sea necesario, aplicar técnicas de pseudonimización fina o re-muestreo dirigido.
- Garantizar la trazabilidad y la transparencia mediante el registro completo del pipeline de

[2] https://proyectoilenia.es/recursos-modelos-datasets/

Trabajo (tarea)	Recurso / escenario	Resultados reportados
ILENIA (modelos de lenguaje)	Familia de modelos de lenguaje en español y lenguas cooficiales	Publicación de modelos y recursos de referencia para en la web del proyecto y hugginface
Alarcón et al. (simplificación en salud)	Ajuste fino de BART/mBART para simplificación de textos sanitarios en español	SARI: 59.7 (mBART fine-tuned); 29.74 (mBART preentrenado en generación de resúmenes); mejora de legibilidad según escala Inflesz
Campillos-Llanos et al. (corpus biomédico)	CLARA-MeD: corpus comparable de simplificación médica en español	24.298 pares de textos (profesional vs. simplificado); $>$96M tokens

Table 1: Resultados de referencia de la literatura relacionados con los casos de uso de SAFEWORDs. Esta tabla no reporta resultados propios del proyecto.

procesamiento, facilitando el cumplimiento de las obligaciones regulatorias aplicables a sistemas de IA de propósito general.

HumanAIze establece como principio que los recursos del proyecto (datos, modelos y software) sean **reutilizables, reproducibles y debidamente documentados**, y que se distribuyan mediante plataformas de recursos digitales relevantes para la comunidad investigadora, facilitando tanto el escrutinio científico como la adopción institucional o industrial.

En el caso de SAFEWORDs, la liberación de recursos está condicionada a la superación de controles previos de privacidad y equidad, alineados con las directrices del WP2 y el régimen de trazabilidad del WP3.

6. Consideraciones éticas y limitaciones

En HumanAIze, los aspectos éticos, legales y sociales no se tratan como una capa adicional posterior, sino como un componente *trustworthy-by-design* que se traduce en guías, recomendaciones y listas de verificación operativas que condicionan la adquisición de datos, el entrenamiento, la evaluación y la diseminación de resultados.

SAFEWORDs (UA) se alinea con este enfoque codificando principios y umbrales en *checklists* aplicables a todo el ciclo de vida de los activos del proyecto —privacidad, no discriminación, supervisión humana y sostenibilidad computacional— y verificando su cumplimiento antes de cualquier uso o liberación.

6.1. Privacidad, RGPD y EU AI Act

HumanAIze explicita que, desde agosto de 2024, el EU AI Act (Reglamento 2024/1689) complementa el RGPD con obligaciones graduadas en función del nivel de riesgo del sistema, y adopta una postura *privacy-by-design* a lo largo de todo el ciclo de vida. En la práctica, todo corpus entrante se escanea con reconocedores multilingües para identificar identificadores directos (correos electrónicos, identificadores fiscales) y señales contextuales susceptibles de facilitar ataques por enlace.

Cuando resulta técnicamente viable, las referencias personales se eliminan o se reemplazan mediante sustitutos sintéticos generados bajo presupuestos de privacidad diferencial; el proceso se registra íntegramente para satisfacer los requerimientos de trazabilidad y transparencia del EU AI Act.

Como medida adicional de control del riesgo, el proyecto contempla pruebas de estrés de reidentificación (incluyendo *membership inference*, *attribute inference* y *record linkage*) y comparativas emparejadas entre muestras anonimizadas y originales para cuantificar la pérdida de utilidad textual. Este conjunto de prácticas guía el caso de uso de anonimización de SAFEWORDs sobre textos legales y administrativos, con el objetivo explícito de que ningún identificador personal persista tras la ingesta y de que los activos resultantes sean defendibles bajo el marco regulatorio vigente.

6.2. Equidad, sesgos y evaluación culturalmente informada

HumanAIze establece un programa sistemático de auditoría y mitigación de sesgos en paralelo al desarrollo de modelos. Se prevé la construcción de un *benchmark* multilingüe (lenguas oficiales de España e inglés) para evaluar estereotipos, desequilibrios de representación y diferencias de rendimiento entre grupos demográficos, con anotación por especialistas y revisión comunitaria orientada a capturar señales culturales que típicamente no aparecen en conjuntos de datos de orientación anglocéntrica.

La evaluación se operacionaliza mediante métricas diagnósticas más allá de la exactitud agregada (p. ej., paridad demográfica condicional, consistencia bajo sustituciones contrafactuales y medidas de daño interseccional), y se integra como compuerta de calidad: regresiones significativas

en estas métricas bloquean la promoción de versiones de modelo. Cuando se detectan sesgos, se contemplan estrategias de mitigación mediante reponderación de datos, debiasing basado en modelo y calibración *post-hoc*, documentando los compromisos entre equidad, utilidad y coste computacional.

6.3. Supervisión humana y recogida de retroalimentación

El proyecto incorpora directrices explícitas sobre la interacción humana y la recogida de retroalimentación (*feedback*) como parte del marco de recomendaciones de WP2, con el objetivo de garantizar un enfoque centrado en el ser humano a lo largo de todo el proyecto.

En SAFEWORDs, estas directrices condicionan tanto la generación de datos de alineación e instrucción como la evaluación humana en escenarios sensibles (administrativo-jurídico y biomédico), evitando mecanismos de recogida de retroalimentación que sean coercitivos, sesgados o no representativos de la diversidad de usuarios.

6.4. Sostenibilidad y control de huella computacional

HumanAIze incorpora como requisito la minimización de la huella de carbono y el diseño de prácticas de *green AI* (WP2), promoviendo la eficiencia computacional y la reutilización de modelos existentes siempre que no se comprometan los objetivos científicos del proyecto.

SAFEWORDs adopta estas restricciones como parte de su diseño experimental, priorizando configuraciones y protocolos reproducibles sobre iteraciones de entrenamiento que no aporten evidencia metodológica adicional.

6.5. Limitaciones

El marco presentado tiene varias limitaciones inherentes que deben tenerse en cuenta al interpretar sus resultados:

- **Verificación incompleta de privacidad.** Aunque se aplican auditorías automáticas, pruebas adversariales y ataques de reidentificación, no existe garantía absoluta de ausencia de riesgo residual en contextos abiertos y multifuente; el riesgo se minimiza pero no se elimina.
- **Cobertura desigual por lengua y dominio.** La disponibilidad de datos y recursos varía considerablemente entre el español y las lenguas cooficiales, lo que puede comprometer la comparabilidad entre configuraciones y obliga a diseños de evaluación cuidadosamente estratificados.
- ***Trade-off* entre utilidad y cumplimiento normativo.** Determinados escenarios pueden requerir técnicas de pseudonimización fina o re-muestreo dirigido para preservar la fidelidad factual tras la anonimización; la pérdida de utilidad asociada debe monitorizarse sistemáticamente.
- **Dependencia de la evaluación humana.** La auditoría culturalmente informada y la supervisión humana son componentes esenciales del marco, pero introducen variabilidad interanotador y costes de escalado que deben gestionarse mediante protocolos de anotación robustos y muestreo estratificado.

Finalmente, en línea con los principios de HumanAIze, todos los recursos del proyecto (datasets, modelos y software) serán abiertos, reutilizables y debidamente documentados; en SAFEWORDs, cualquier liberación estará condicionada a superar los umbrales estrictos de privacidad y equidad definidos por el propio marco del proyecto.

7. Conclusiones

Este trabajo ha presentado SAFEWORDs como contribución al proyecto HumanAIze, orientada a operacionalizar un marco *trustworthy-by-design* para Grandes Modelos de Lenguaje en las lenguas oficiales de España (español, catalán, valenciano, gallego y euskera). SAFEWORDs articula un enfoque metodológico integrado que combina: (i) alineación normativa y verificación de cumplimiento del RGPD y el EU AI Act, (ii) control generativo y evaluación comparativa en configuraciones *zero-shot* y *few-shot*, y (iii) un esquema reproducible de auditoría de fuga residual, detección de sesgos y trazabilidad a lo largo del ciclo de vida de los activos.

La propuesta se valida conceptualmente en tres casos de uso de alto impacto social: (1) anonimización para la preservación de privacidad en textos administrativos, jurídicos y sanitarios, (2) mejora de la accesibilidad textual, incluyendo lenguaje claro en documentación institucional, y (3) aplicación en el dominio biomédico, donde la sensibilidad de los datos y la precisión terminológica exigen protocolos de evaluación más estrictos. Frente a aproximaciones centradas exclusivamente en métricas de rendimiento, el marco propuesto prioriza criterios operativos de legalidad, robustez y equidad multilingüe como condiciones necesarias para el despliegue responsable de sistemas basados en LLMs.

Como trabajo futuro, SAFEWORDs avanzará en: (i) la consolidación del repositorio multilingüe

versionado de datos de tarea e instrucción, (ii) la ampliación de los protocolos de auditoría adversarial y las pruebas de reidentificación, (iii) la evaluación sistemática por lengua y dominio con participación de expertos, y (iv) la preparación de activos (datasets, prompts, herramientas y, cuando proceda, modelos) para su liberación condicionada a umbrales estrictos de privacidad y equidad, reforzando la transferibilidad y reutilización científica en línea con los objetivos de HumanAIze.

Declaración ética

El proyecto garantiza el cumplimiento del RGPD en todas las fases de tratamiento de datos, minimiza los riesgos de reidentificación mediante auditorías automáticas y humanas, y evita prácticas extractivas en comunidades lingüísticas minoritarias. Los datos experimentales utilizados en el desarrollo del marco han sido previamente anonimizados conforme a los protocolos descritos en este trabajo.

Agradecimientos

Este trabajo ha sido financiado por el Ministerio de Ciencia, Innovación y Universidades (España) en el marco de la convocatoria 2025 de Proyectos de Investigación en el Ámbito de la Inteligencia Artificial (AIA2025), referencia **AIA2025-163322-C63**.

8. References

Rodrigo Alarcón, Paloma Martínez, and Lourdes Moreno. 2023. Tuning bart models to simplify spanish health-related content. *Procesamiento del Lenguaje Natural*, (70):111–122.

Fernando Alva-Manchego et al. 2020. Asset: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *ACL*, pages 4668–4679.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 610–623.

Leonardo Campillos-Llanos, Ana R. Terroba Reinares, Sofía Zakhir Puig, Ana Valverde-Mateos, and Adrián Capllonch-Carrión. 2022. Building a comparable corpus and a benchmark for spanish medical text simplification. *Procesamiento del Lenguaje Natural*, (69):189–196.

Nicholas Carlini et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Iria da Cunha. 2022. Un redactor asistido para adaptar textos administrativos a lenguaje claro. *Procesamiento del Lenguaje Natural*, 69:39 – 49.

Franck Dernoncourt et al. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, 24(3):596–606.

Julen Etxaniz, Oscar Sainz, Naiara Perez, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. Latxa: An open language model and evaluation suite for basque. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand. Association for Computational Linguistics.

Pablo Gamallo, Pablo Rodríguez, Diana Santos, Salvador Sotelo, N. Miquelina, S. Paniagua, D. Schmidt, I. de Dios-Flores, P. Quaresma, D. Bardanca, J. R. Pichel, V. Nogueira, and Senén Barro. 2024. A galician-portuguese generative model. In *EPIA 2024 – Portuguese Conference on Artificial Intelligence (Conference Proceedings)*.

A. Garza et al. 2025. Prvl: Quantifying the capabilities and risks of large language models for pii redaction. *arXiv preprint arXiv:2508.05545*.

Aitor González-Agirre, Montserrat Marimon, Carlos Rodriguez-Penagos, Javier Aula-Blasco, Irene Baucells, Carme Armentano-Oller, Jorge Palomar-Giner, Baybars Kulebi, and Marta Villegas. 2024. Building a data infrastructure for a mid-resource language: The case of catalan. In *Proceedings of the LREC-COLING 2024 Conference*, pages 2556–2566.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, Mario Mina, Iñigo Pikabea, Adrián Rubio, Alexander Shvets, Anna Sallés, Iñaki Lacunza, Jorge Palomar, Júlia Falcão, Lucía Tormo, Luis Vasquez-Reina, Montserrat Marimon, Oriol Pareras, Valle Ruíz-Fernández, and Marta Villegas. 2025. Salamandra technical report. arXiv:2502.08489.

GPLSI – Language and Information Systems Group, University of Alicante. 2026. Aitana-6.3b (model card). https://huggingface.co/gplsi/Aitana-6.3B. Accessed: 2026-02-13.

Jefatura del Estado (España). 2018. Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales. BOE núm. 294, de 06/12/2018. Referencia: BOE-A-2018-16673. Entrada en vigor: 07/12/2018. ELI: https://www.boe.es/eli/es/lo/2018/12/05/3/con.

J. Jonnagaddala and M. C. Wong. 2025. Strategies for privacy-preserving ehr analysis using llms.

J. Mancera et al. 2025. Pba-llm: Privacy- and bias-aware nlp using named-entity recognition (ner). *arXiv preprint arXiv:2507.02966*.

Michele Miranda, Elena Sofia Ruzzetti, Andrea Santilli, et al. 2024. Preserving privacy in large language models: A survey on current threats and solutions. *arXiv preprint arXiv:2408.05212*.

Parlamento Europeo and Consejo de la Unión Europea. 2016. Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos (Reglamento general de protección de datos). DO L 119, 4.5.2016, pp. 1–88. CELEX: 32016R0679. Aplicación desde 25/05/2018.

V. Ponomarenko et al. 2026. Capid: Context-aware pii detection for question-answering systems. *arXiv preprint arXiv:2602.10074*.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1).

Robin Staab, Mark Vero, Mislav Balunović, et al. 2024. Large language models are advanced anonymizers. *arXiv preprint arXiv:2402.13846*.

Tianyu Yang, Xiaodan Zhu, and Iryna Gurevych. 2024. Robust utility-preserving text anonymization based on large language models. *arXiv preprint arXiv:2407.11770*.

P. I. Zamroz and Yu. V. Morozov. 2024. Large language models and personal information: security challenges and solutions through anonymization. *Komp'ûternì sistemi ta merežì*.

Exploration of Sentence Representations in Spanish BERT-like Models

Gonzalo Herrera, Aiala Rosá, Luis Chiruzzo
Instituto de Computación Facultad de Ingeniería
Universidad de la República, Uruguay
{gonzalo.herrera, aialar, luischir}@fing.edu.uy

Abstract

Transformer-based language models, ubiquitous in modern NLP, generate internal representations (embeddings) of words and sentences. Yet, systematic comparisons of embedding strategies from various models remain limited. In this work, we evaluate Spanish embeddings from several BERT-like models (BETO, multilingual BERT, XLM-RoBERTa, ROUBERTa) to understand their syntactic and semantic capabilities across layers. We propose sentence-level analogy tests to probe generalization. Results suggest tasks like verb negation or word reordering perform best with embeddings from earlier layers, while nuanced semantic distinctions—such as agent or patient gender—are better captured by deeper layers. Our findings provide guidelines for embedding strategies and offer a foundation for further NLP research.

Keywords: BERT, embeddings, evaluation, Spanish, Analogies, Similarity

1. Introduction

The task of finding numerical vectors to represent language information has a long history: the concept of word embeddings emerged from the idea of assigning a vector representation to each word. That vector lies in a space whose dimensionality is significantly smaller than the vocabulary size, with the extra constraint that similar words should have close vectors in this space. Collobert and Weston (2008) proposed the approach of pre-training embeddings with unlabeled text. This approach was extended to concatenate additional features to represent extra information, in particular the position of the word in the sentence and its distance to the main verb (Collobert et al., 2011). Later, two new methods were proposed to create word embeddings: word2vec (Mikolov et al., 2013a) and GloVe (Pennington et al., 2014). However, these approaches do not allow a representation to adapt based on the context the word appears in.

The introduction of the transformer architecture Vaswani et al. (2017) enabled the development of contextualized language models. Transformer-based models leverage information across the sentence to create contextual embeddings. This led to the creation of two main families of models: generative models such as GPT, which use decoder-only transformers (Radford et al., 2018), and BERT and BERT-like models, which use encoder-only transformers (Devlin et al., 2019). Encoder-only models produce contextualized vector representations (embeddings) that encode syntactic and semantic information.

Understanding the structure and properties of these embeddings is particularly important in retrieval-based applications such as semantic search and Retrieval-Augmented Generation (RAG) (Lewis et al., 2020). For RAG systems, documents are divided into chunks, which are encoded as vectors and stored in a database. Given a query, its embedding is used to retrieve the most relevant documents, which are passed to a generative model. The effectiveness of the retrieval system depends directly on the quality of the embeddings. Therefore, reliable intrinsic methods to evaluate embedding representations are essential.

Language models are commonly evaluated indirectly, through benchmarks that measure their performance on downstream tasks. While this approach aligns well with how decoder-only transformer models are commonly used, it does not fully reflect the way encoder-only models are employed for retrieval systems. This motivates the need for intrinsic evaluation methods that specifically analyze the semantic and syntactic properties of sentence embeddings. This becomes even more relevant when we consider models for languages other than English, which have been less extensively studied. Furthermore, languages such as Spanish are more morphologically rich than English, introducing additional inflectional variations that embeddings must capture to perform optimally.

In this work, we study how to intrinsically evaluate embeddings produced by BERT-like models for Spanish. We evaluate the models with two sets of tests: a semantic textual similarity test and a new sentence-level analogy task inspired by the word analogies test previously used in word embeddings (Mikolov et al., 2013b). We define a set of controlled syntactic and semantic transformations we can apply to different entities in the sentence

Proceedings of LANLP: Bridging Ibero and Latin American NLP Communities, pages 55–65
16 May 2026.

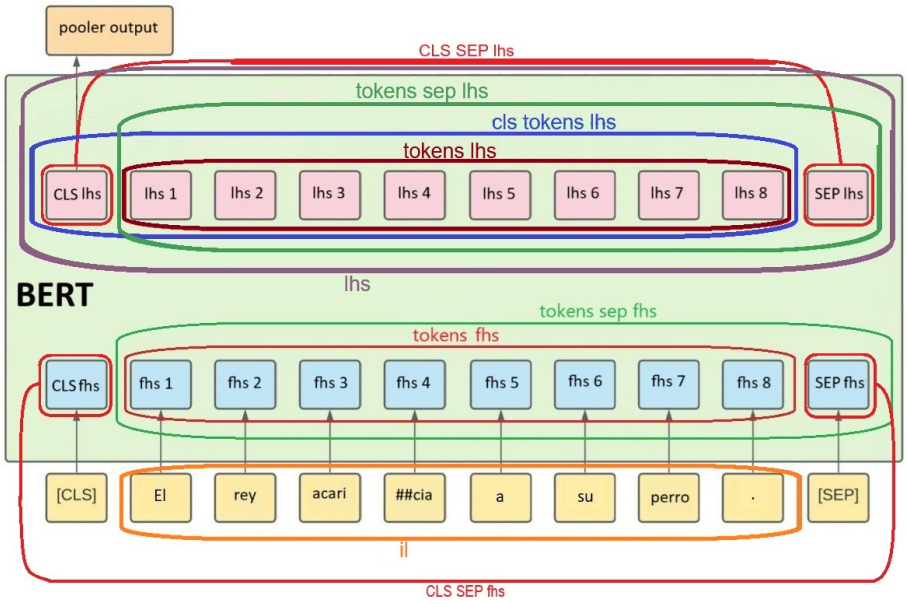


Figure 1: Structure and layers involved in BERT embeddings. The input layer is at the bottom; then the first hidden state (fhs) that consists of the input embeddings + sequence embeddings + positional embeddings; then all the other layers until the last hidden state (lhs); finally there is the pooler output that, in BERT's case is a function that takes the CLS token from the last hidden state. This diagram also shows which transformer activations we considered for defining the different word and sequence embeddings we use in this work.

and evaluate whether such transformations can be transferred to other sentences by operating with the vector representations of the sentences. We call this test the "Analogy Tests using Sentences", and will be made public so other researchers can use it as a benchmark.

This paper is organized as follows. In section 2 we analyze language model evaluation and how it has been done for BERT-like models and LLMs. In section 3 we describe the models we tested and how we extracted the sentence embeddings for the tests. In section 4 we detail all the experiments and tests we did and our proposed evaluation method. In section 5 we show the results we obtained. In section 6 we provide the final conclusions and future work.

2. Related Work

There has been a lot of work when it comes to evaluating language models. Broadly, we can classify the evaluation into two main categories: extrinsic and intrinsic evaluations.

Extrinsic evaluations consist of evaluating the language model based on its performance when used to solve different tasks, also called downstream tasks. BERT and RoBERTa (Liu et al., 2019b) models are usually evaluated with benchmarks like GLUE (Wang et al., 2018), SQuAD (Rajpurkar et al., 2016) and SWAG (Zellers et al., 2018). Other works explored the potential these models have before and after fine-tuning them for specific tasks such as semantic textual similarity and natural language inference (Choi et al., 2021).

Intrinsic evaluations on the other hand try to evaluate the internal representation the model has constructed. Word analogies (Mikolov et al., 2013b) and word similarities (van der Maaten and Hinton, 2008) are two of the main ways that were used to test traditional word embeddings.

After the release of the BERT and BERT-like models, there has been an interest in understanding the internal representation of words captured by transformers. Many works were done in order to understand the information captured by these models. One foundational work looked at how the representation of words is modified for words with multiple meanings, concluding that the embeddings of a word are clustered in separate groups based on context (Reif et al., 2019). Another similar approach is to look at how similar the representation of words remains in sentences and how they change when contextualized (Ethayarajh, 2019). Other works on polling the internal structure of different neural models exist (Liu et al., 2019a), many indicating differences in representation between syntax and semantics that can be found in different layers (Jawahar et al., 2019; Tenney et al., 2019), and the encoding of social biases (Bommasani et al., 2020). Yun et al. (2021) presents a visualization tool based on dictionary learning to better understand the inner works of transformer

Name	Layer / tokens averaged	Tokens for the example sentence
`il`	All word tokens — *input layer*	"El" + "rey" + "acari" + "##cia" + "a" + "su" + "perro" + "."
`tokens fhs`	All word tokens — *first hidden state*	fhs 1 + fhs 2 + ... + fhs 8
`tokens sep fhs`	Word tokens + [SEP]/</s> — *first hidden state*	fhs 1 ... fhs 8 + fhs SEP
`cls lhs`	[CLS]/<s> token — *last hidden state*	lhs CLS
`sep lhs`	[SEP]/</s> token — *last hidden state*	lhs SEP
`tokens lhs`	All word tokens — *last hidden state*	lhs 1 + ... + lhs 8
`cls sep lhs`	[CLS]/<s> + [SEP]/</s> — *last hidden state*	lhs CLS + lhs SEP
`cls tokens lhs`	[CLS]/<s> + word tokens — *last hidden state*	lhs CLS + lhs 1 ... lhs 8
`tokens sep lhs`	Word tokens + [SEP]/</s> — *last hidden state*	lhs 1 ... lhs 8 + lhs SEP
`lhs`	All tokens — *last hidden state*	lhs CLS + lhs 1 ... lhs 8 + lhs SEP
`pooler output`	Pooler output vector	pooler output

Table 1: Embedding representations evaluated. Column 1 lists the shorthand names used throughout the paper; Column 2 specifies which layer and token subset each name refers to; Column 3 illustrates the actual vectors selected for the sentence "El rey acaricia a su perro." (*The king pets his dog.*) as tokenized by BETO-cased.

based language models. Most recently, Bonino et al. (2025) did a comprehensive mathematical analysis of the representations of the attention layers in these models. All of these works, while truly insightful about the inner workings of BERT-like models, do not provide a way to compare them and their vector representations at sentence level. Furthermore, these works are for English, and the exploration of these same representations in Spanish with its morphological particularities has not been explored thoroughly.

3. Models

In this work we tested four variants of the BERT model and five variants of the RoBERTa model. All models were downloaded from HuggingFace.

For evaluating BERT-like models, we chose BETO (Cañete et al., 2020), both cased and uncased variants. BETO is a Spanish specialized model. The other two BERT models we tried are the multilingual BERT models, both cased and uncased.

In the case of RoBERTa, we picked FacebookAI's multilingual model, xlm-RoBERTa (Conneau et al., 2019), both the base model and the large model, two versions of xlm-RoBERTa fine-tuned for Spanish, a base model (Pandya et al., 2021) and a large model (Lange et al., 2021). Lastly, we also tried a RoBERTa model trained specifically over Spanish news text, called ROUBERTa (Filevich et al., 2024). All xlm-RoBERTa models are cased models while ROUBERTa is an uncased model.

The names we used in tables are the following: *beto-cased*, *beto-uncased*, *mbert-cased*, *mbert-uncased*, *xlm-roberta-base*, *xlm-roberta-large*, *xlm-roberta-base-sp*, *xlm-roberta-large-sp* and *rouberta*

3.1. Embedding Extraction

For each model we extract embeddings from different layers and combine different tokens from the layer. We perform different tests using the representation of the input layer of the models, the representation of the first hidden layer, the last hidden layer, and the pooler output vector. When more than one token is involved, we calculate the centroid of the tokens to get the sentence embeddings. Fig. 1 shows each extraction format from a BERT example. Table 1 presents a reference summary with representation name, how it is calculated and an example for each.

4. Description of the Experiments

We performed two different tests in order to evaluate the representations that can be extracted from the models at both the syntactic and semantic levels. The first experiment is a semantic textual similarity test, using the vector representation of the sentences and cosine similarity. The second group is an analogy test using sentences, where we created a small dataset to test the vector representation of different sentences after applying certain transformations.

4.1. Semantic Textual Similarity

The Semantic Textual Similarity (STS) test identifies whether or not two sentences share similar meanings. For this test, we used the Semeval 2015 dataset for Spanish (Agirre et al., 2015), which includes 500 sentence pairs from news articles and 251 pairs from Wikipedia. Each pair is annotated by humans in a range from 0 to 4. To create the gold standard, four annotators scored the similarity of each pair and the label is the average score.

We feed the model both sentences and calculate the cosine similarity for the vectors of both. After that, we calculate the correlation coefficient

Verb	Agents	Patients
acariciar *(to pet)*	hombre, rey, mujer, reina, *(man, king, woman, queen)*	perro, gato, perra, gata *(dog, cat, female dog, female cat)*
comer *(to eat)*	hombre, rey, mujer, reina, perro, gato, perra, gata *(man, king, woman, queen, dog, cat, female dog, female cat)*	pescado, fruta *(fish, fruit)*
llegar *(to arrive)*	hombre, rey, mujer, reina, perro, gato, perra, gata *(man, king, woman, queen, dog, cat, female dog, female cat)*	No patient
dormir *(to sleep)*	hombre, rey, mujer, reina, perro, gato, perra, gata *(man, king, woman, queen, dog, cat, female dog, female cat)*	No patient
pasear *(to walk)*	hombre, rey, mujer, reina *(man, king, woman, queen)*	perro, gato, perra, gata *(dog, cat, female dog, female cat)*
perseguir *(to follow)*	hombre, rey, mujer, reina, perro, gato, perra, gata *(man, king, woman, queen, dog, cat, female dog, female cat)*	hombre, rey, mujer, reina, perro, gato, perra, gata *(man, king, woman, queen, dog, cat, female dog, female cat)*

Table 2: List of agents and patients for each verb. For some verbs, a patient cannot be included given how the sentences were constructed. All agents and patients are also used in their plural form.

Variation	Verbs	Number of experiments
Negation	acariciar, comer, pasear, perseguir *(to pet, to eat, to walk, to follow)*	6000
Gender change of the agent	acariciar, comer, pasear, perseguir *(to pet, to eat, to walk, to follow)*	6000
Number change of the agent	acariciar, comer, pasear, perseguir *(to pet, to eat, to walk, to follow)*	6000
Gender change of the patient	acariciar, pasear, perseguir *(to pet, to walk, to follow)*	4500
Number change of the patient	acariciar, comer, pasear, perseguir *(to pet, to eat, to walk, to follow)*	6000
Temporal modifier reordering	llegar, dormir *(to arrive, to sleep)*	3000

Table 3: List of verbs for each transformation.

between the computed cosine similarity and the human-assigned gold standard scores. In addition, we plot scatter-plots to validate the correlations and get a better understanding of the results.

4.2. Analogy Tests using Sentences

This test is inspired on the word analogies test, but with sentence embeddings. For these tests, we created a controlled synthetic dataset to validate if the models are able to capture different types of morphologic, syntactic and semantic transformations.

4.2.1. Dataset Creation

The dataset was automatically generated using a python script that specified how to combine the different words to construct each sentence. To create it we selected a few verbs that could fit the different transformations we intended to test. For each verb we selected a few nouns to work as agents of the verb, a few to work as the patient of the verb if any applied, and which transformations we would test. For all the verbs with a patient we have sentences in the active voice and in the passive voice. The full list of agents and patients for each verb are shown in Table 2. All of the verbs were conjugated into the past simple, present simple and future simple to simplify the creation and evaluation. Limiting the dataset like this allows us to have a controlled dataset where we are sure every sentence is grammatically correct. The final dataset consists of 6528 sentences.

The transformations we selected are: negation, gender or number change in the agent or patient, and temporal modifier reordering. For all this, we created a collection of sentences and extended them with all the transformations that we could apply, as long as the sentence still made sense. In Table 3 we show which verbs were chosen for each transformation.

Table 4 shows the transformation for the sentence "El rey acaricia a su perro." (*The king pets his dog*) for all the transformations except temporal modifier reordering since this sentence does not have a temporal modifier. For the temporal modifier reordering we took sentences with one of three possible temporal modifiers: "en la mañana" (*in the morning*), "en la tarde" (*in the evening*) and "en la noche" (*at night*). We take the temporal modifier which may go either at the beginning or the end of the sentences, for example: "El hombre duerme en la mañana." (*The man sleeps in the morning*) would be changed into "En la mañana el hombre duerme." (*In the morning the man sleeps*).

4.2.2. Experiments

We made experiments for all the transformations we explained previously, this was done in three different sets:

– all four sentences were in the active voice

Variation	Active voice	Passive voice
Base sentence	El rey acaricia al perro. *(The king pets the dog.)*	El perro es acariciado por el rey. *(The dog is pet by the king.)*
Negation	El rey *no* acaricia al perro. *(The king does not pet the dog.)*	El perro *no* es acariciado por el rey. *(The dog is not pet by the king.)*
Gender change of the agent	*La reina* acaricia al perro. *(The queen does not pet the dog.)*	El perro es acariciado por *la reina*. *(The dog is not pet by the queen.)*
Number change of the agent	*Los reyes* acarician al perro. *(The kings do not pet the dog.)*	El perro es acariciado por *los reyes*. *(The dog is pet by the kings.)*
Gender change of the patient	El rey acaricia a la *perra*. *(The king pets the female dog.)*	*La perra* es acariciada por el rey. *(The female dog is pet by the king.)*
Number change of the patient	El rey acaricia a los *perros*. *(The king pets the dogs.)*	*Los perros* son acariciados por el rey. *(The dogs are pet by the king.)*

Table 4: Syntactic transformations used in our dataset. Starting from a base sentence in both active and passive voice, five transformations are applied: negation, gender/number change of the agent, and gender/number change of the patient. There are also composed transformations, for this example "La reina no acaricia al perro." (*The queen does not pet the dog.*) is the negation and the agent gender change, we also had other sentences such as "Las reinas acarician al perro." (*The queens pet the dog.*) and even "Las reinas no acarician a las perras." (*The queens do not pet the female dogs.*) in the dataset.

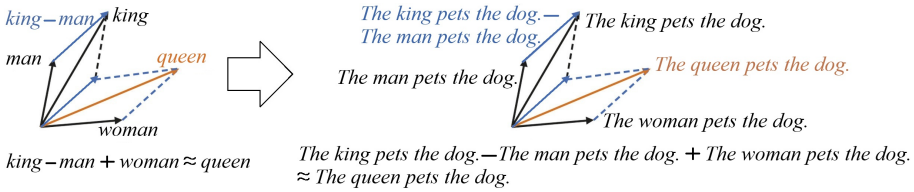


Figure 2: Vector operation. This figure illustrates the $A' - A + B \approx B'$ vector calculation for the traditional method and one example for our proposed method with full sentences. In our method we adapt all words needed for the sentences but no more than that.

- all four sentences were in the passive voice
- two sentences (A and A') were in active voice and the other two (B and B') were in passive voice.

This way we can see if we get different results when we are using the active voice versus the passive voice. We can also see if there is an impact when mixing both passive and active voice in the calculation.

For the experiments, we made sure the same verb was used for both A and B to limit variability. Even though we used the same verb, A and B might have different time conjugations. For each combination of transformation and verb we randomly drew 500 pairs of sentences for the active voice experiments, 500 pairs for the passive voice experiments, and 500 for the mixed voice experiments: 250 where A and A' were in the active voice while B and B' were in the passive voice, and 250 where A and A' were in the passive voice while B and B' were in the active voice. The final number of experiments for each transformation can be seen in Table 3.

4.2.3. Results calculation

We use the embedding of three of the sentences to evaluate whether the vector of the fourth sentence can be recovered. Given two sentences, A and B and their transformed counterparts A' and B' respectively, we compute the vector operation A' - A + B. The resulting vector is then compared against the embeddings of all the sentences in the dataset using cosine similarity. The result is considered correct every time B' is in the top-k most similar sentences. We calculated the results for top-1, top-3 and top-5. However, top-1 accuracy is zero in most of our experiments. Top-3 accuracy generally follows the same trend as top-5 accuracy, but the absolute values remain very low, which makes comparative analysis less informative.

For this reason, we focus our discussion on top-5 accuracy. This allows for a more meaningful comparison across models.

In Fig. 2, we show a transformation of the classic example king/man/woman/queen, and one with full sentences.

5. Results

Due to space limitations, we highlight the best results or those we consider most interesting to analyze.

5.1. Semantic Textual Similarity

As previously mentioned, we present the results of the STS test based on two analyses: a correla-

Model	beto cased	beto uncased	mbert cased	mbert uncased	xlm roberta base	xlm roberta base sp	xlm roberta large	xlm roberta large sp	rouberta
tokens sep fhs	0.499	0.478	0.493	0.493	0.466	**0.466**	**0.434**	**0.444**	0.424
sep lhs	**0.673**	0.527	0.407	0.603	0.028	0.114	0.295	0.135	0.335
tokens lhs	0.581	0.604	**0.616**	0.574	**0.470**	0.238	0.426	0.265	0.607
cls tokens lhs	0.585	0.604	0.615	0.572	0.465	0.235	0.427	0.263	**0.617**
cls sep lhs	0.660	0.528	0.307	**0.627**	0.049	0.109	0.086	0.195	0.429
tokens sep lhs	0.587	**0.604**	0.615	0.575	0.465	0.240	0.425	0.270	0.607

Table 5: Semantic Textual Similarity – Pearson correlation between cosine similarity and the human-ranked gold standard (higher is better). We only show the embedding extraction methods that achieve a maximum in at least one model, omitting those that are outperformed for all the models

tion analysis to get a numeric comparison of the models, and a graphical analysis to get a better understanding of the results.

5.1.1. Correlation analysis

Table 5 shows the Pearson correlation between the gold standard and the cosine similarity for each sentence pair. Most of the best results are found in the last hidden state. This is to be expected since later layers should be able to better capture contextual information as seen in previous studies (Hewitt and Manning, 2019; Jawahar et al., 2019).

BETO outperforms all the other models in the cased variation. It is interesting to see that the option with the best results is the representation of the SEP token in the last hidden state, outperforming the CLS representation for more than ten points. This suggests that relying solely on the CLS token to calculate the pooler output in the original version of BERT might not always be optimal.

Another interesting observation is that the ROUBERTa model has the best performance of all the RoBERTa models for this test, outperforming all of them with a big margin and being near most of the BERT variants.

5.1.2. Graphical analysis

When comparing figures 3a, 3b, and 3c, we can see that while BETO and ROUBERTa's similarity scores range from 0.5 to 1, Multilingual-BERT's scores range from 0.8 to 1. This indicates that although relative similarity increases alongside the gold standard, the absolute similarity values are not a reliable measure. Interestingly, ROUBERTa shows a better spatial representation than Multilingual BERT even though both have almost the same correlation between the similarity and the gold standard. This is probably due to ROUBERTa being specifically trained using news data and the dataset having 500 sentence pairs from newswire.

This reaffirms that these models already have a lot of syntactic and semantic information even without finetuning, as seen in previous studies.

5.2. Analogy Tests

We present a summarization for each transformation and group of experiments: when all the sentences are in active voice, when all the sentences are in passive voice, and when two of them are in passive voice and the other two are in active voice (which we call mixed-voices). We also present the results of specific transformations in more detail.

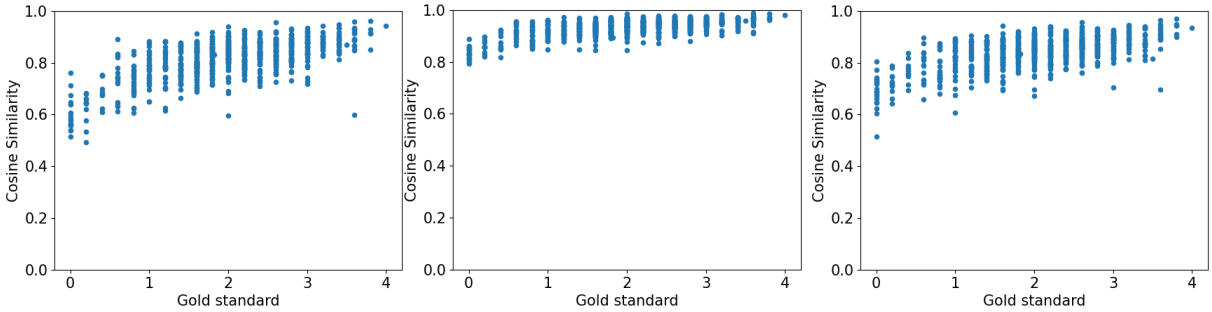
5.2.1. Agent vs Patient

Before analyzing the results, it is important to note that the agent takes the role of the subject of the sentence in the active voice, while it takes the role of the complement in the passive voice. In the same way, the patient shifts from being the complement in the active voice to the subject in the passive voice. This distinction is important because changes in the subject affect the verb to maintain grammatical agreement.

While there are good results for Number of patient in active voice, as seen in the top rows of Table 6, this is not seen in passive voice (middle rows of the table). In the same way, the Number of agent gets somewhat good results (albeit not as good as Number of patient) in passive voice while getting poor results in active voice. In a similar trend we get more balanced results in the active to passive voice test, as seen in the bottom rows. These results indicate that the models are better at changing the number of the complement of the sentence in comparison than changing the number of the subject of the sentence. This makes sense since changing the number of the subject of the sentence also changes the conjugation of the verb, something that does not happen when changing the number of the complement.

The results for gender transformations are better compared to the number transformations. Changing the gender of the subject does not impact the verb in the active voice, and has a smaller impact in the passive voice compared to changing the number of the subject. This also explains why we see better results in the active voice compared to the passive voice.

The result that we find the most interesting is how the models have the best performance in the



(a) **BETO Cased SEP LHS** (b) **mBERT uncased CLS + SEP LHS** (c) **ROUBERTa CLS + tokens LHS**

Figure 3: Each point represents a pair of sentences. In the X axis we have the human-assigned similarity score while in the Y axis we have the cosine similarity between the vectors.

Experiment		beto cased	beto uncased	mbert cased	mbert uncased	xlm roberta base	xlm roberta base sp	xlm roberta large	xlm roberta large sp	rouberta
Active voice	Negation	14.70	47.45	34.90	28.35	**52.80**	47.00	52.35	44.05	45.30
	Gender of patient	11.60	10.00	**47.20**	8.33	2.00	19.13	12.40	18.87	10.47
	Number of patient	42.80	49.65	30.55	37.90	20.65	20.95	16.85	33.95	**61.60**
	Gender of agent	23.60	**39.45**	17.85	7.15	4.90	13.30	11.05	10.80	12.05
	Number of agent	0.15	1.40	0.05	0.05	1.45	2.15	**2.80**	1.05	0.00
Passive voice	Negation	1.65	41.80	14.70	13.60	**53.15**	49.95	53.10	41.70	28.00
	Gender of patient	0.07	5.93	**30.47**	0.53	0.07	11.13	3.73	1.87	0.87
	Number of patient	2.70	**7.15**	3.65	1.65	0.15	2.55	2.25	1.00	0.25
	Gender of agent	31.55	**46.45**	31.75	30.85	4.30	10.60	22.35	25.60	25.70
	Number of agent	42.85	33.95	12.25	26.00	4.35	14.10	26.20	**43.70**	28.05
Mixed active-passive	Negation	8.45	52.25	24.20	22.45	**53.70**	47.45	53.35	41.40	37.70
	Gender of patient	3.47	10.20	**60.27**	4.93	1.07	18.80	8.53	14.00	9.53
	Number of patient	22.30	**34.20**	18.00	21.70	7.80	16.75	8.60	17.45	31.05
	Gender of agent	28.35	**48.20**	26.65	16.10	4.85	15.25	19.50	22.15	15.35
	Number of agent	9.20	18.20	4.80	7.60	2.45	7.20	15.10	**19.70**	6.75

Table 6: Top-5 results for each model for experiments using only active voice, only passive voice, or two in active and two in passive voice. The values show the percentage of times a sentence appeared in the top 5 among its nearest vector neighbors. Results were obtained in different layers depending on the model and task.

mixed-voices tests as seen when comparing results in the three sections of Table 6 for the gender change tests. This goes against what we see in the number change tests. Another surprising result is that the only model that performs well for the gender of patient test is the Multilingual-BERT cased model. Firstly, this model was not fine-tuned for Spanish and we expected it to be one of the least performing models. Secondly, it did not matter if the transformation was in active, passive, or active to passive voices, since, as previously stated, the patient becomes the object or subject of the verb based on which voice we are using. This was also noted for the gender of the agent, where BETO uncased was the model that got the best results regardless of the voice transformation being used. These results prompt further analysis of these kinds of transformations in the future to get a better understanding of the models.

When we analyze these results at the layer level we can see how the information is best represented in the last hidden state (top rows in Table 7). In fact, the best results use SEP token, followed by CLS token. This indicates that this semantic information is better summarized there for all the models we tested, coinciding with the results seen in the STS test.

5.2.2. Negation

The results for the negation of the verb were consistently low for all of the models. This indicates that negation may not be linearly represented in sentence embeddings. However, the best results were seen in the first layers, be it the input layer or the first hidden state as shown in the middle rows of Table 7. This indicates that the changes were captured primarily at word level rather than at the semantic level. It would be interesting to see if newer BERT-like models perform better since new models have been released for the specific purpose of embedding generation in RAG (Wang et al., 2024).

5.2.3. Temporal modifier reordering

The bottom rows of Table 7 show both BETO and Multilingual-BERT got 100% in their uncased vari-

	Representation	beto cased	beto uncased	mbert cased	mbert uncased	xlm roberta base	xlm roberta base sp	xlm roberta large	xlm roberta large sp	rouberta
Gender of Agent task	tokens il	0.12	0.13	1.50	0.02	0.10	0.17	0.10	0.28	8.28
	tokens fhs	0.02	0.08	0.82	0.00	0.02	0.07	0.02	0.13	3.58
	tokens sep fhs	0.02	0.03	0.73	0.00	0.03	0.08	0.03	0.15	3.53
	cls lhs	21.20	41.82	21.90	10.63	4.10	5.92	12.15	17.53	**14.33**
	sep lhs	**26.98**	**44.70**	22.68	16.70	4.05	11.67	17.08	11.08	9.72
	tokens lhs	21.43	38.42	22.47	5.62	3.43	8.43	16.52	16.97	4.83
	cls sep lhs	25.75	43.50	22.18	**16.78**	4.23	**13.05**	**17.20**	15.37	11.98
	cls tokens lhs	22.53	39.45	22.70	5.97	3.75	8.90	16.15	**17.83**	6.82
	tokens sep lhs	22.25	39.47	22.70	6.12	3.80	8.92	16.55	16.95	6.18
	lhs	23.02	40.08	**22.87**	6.32	4.20	9.47	16.38	17.65	7.30
	pooler output	21.72	42.18	18.22	4.50	**4.25**	5.90	11.65	17.47	13.93
Negation task	tokens il	4.32	13.40	**24.45**	18.43	**53.22**	**48.13**	**52.93**	**42.38**	0.00
	tokens fhs	**8.22**	46.17	22.20	**21.47**	22.92	20.35	22.75	21.85	36.68
	tokens sep fhs	7.43	**47.17**	19.65	19.42	21.80	17.80	21.82	20.85	**37.00**
	cls lhs	0.02	0.30	0.28	0.35	7.17	7.75	0.00	1.33	0.38
	sep lhs	0.85	0.38	0.23	1.70	4.78	14.55	0.00	1.78	1.08
	tokens lhs	0.02	0.83	0.97	1.90	26.53	14.25	0.00	0.70	1.03
	cls sep lhs	0.13	0.28	0.18	1.63	5.85	11.68	0.00	1.48	0.58
	cls tokens lhs	0.00	0.67	0.85	1.68	16.10	12.63	0.00	0.80	0.98
	tokens sep lhs	0.07	0.70	0.85	1.82	14.38	13.07	0.00	0.70	1.07
	lhs	0.10	0.60	0.82	1.80	7.65	11.50	0.00	0.78	1.05
	pooler output	0.02	0.22	0.70	0.57	5.32	8.07	0.00	1.58	0.43
Temporal Modifier Reordering task	tokens il	6.43	**100.00**	**8.27**	**100.00**	**85.20**	**85.17**	**76.87**	**62.57**	42.93
	tokens fhs	**6.90**	**100.00**	1.10	**100.00**	57.00	57.03	67.97	50.87	**91.00**
	tokens sep fhs	6.60	**100.00**	1.10	**100.00**	57.30	57.40	68.03	51.47	**91.00**
	cls lhs	0.20	0.43	0.20	0.20	12.40	0.17	20.17	0.07	0.00
	sep lhs	0.67	0.40	0.00	6.80	10.03	0.00	24.23	0.00	0.00
	tokens lhs	0.00	6.57	0.03	0.60	37.63	1.53	32.63	0.00	0.00
	cls sep lhs	0.60	0.43	0.07	5.77	11.23	0.03	23.37	0.00	0.00
	cls tokens lhs	0.03	5.90	0.03	0.67	38.83	1.57	32.77	0.00	0.03
	tokens sep lhs	0.03	6.20	0.07	0.93	39.27	1.10	34.07	0.00	0.03
	lhs	0.07	5.50	0.10	0.93	42.03	1.27	34.03	0.00	0.03
	pooler output	0.03	0.30	0.63	1.90	11.63	0.20	19.30	0.10	0.00

Table 7: Top-5 accuracy for Gender of Agent, Negation, and Temporal Modifier Reordering tasks. Scores of all the experiments: the active-voice, passive-voice, and mixed-voice evaluations. In bold is the best result for each model. The best result across all the variants is underlined.

ants. This is expected since they got those results in the input layer and first hidden state. In these layers, embeddings are the same for both models since the sentence is composed of the same words and word order does not matter, making the task trivial. Interestingly, the capitalization of words was enough to make the cased versions of these models perform poorly. This result was unexpected since the sentences maintained the same meaning and we expected them to have similar representations in higher layers. RoBERTa models on the other hand got pretty good results. These are cased models, with the exception of ROUBERTa, so their much better performance in the task when compared to BETO and Multilingual-BERT is something to note. However, as already seen for the negation tests, the best results were obtained in the input layer and first hidden state, with performance deteriorating in the last hidden state and output. Interestingly, both Multilingual-RoBERTa models got good results in the last layers. As they are not fine-tuned for Spanish, we did not expect them to be the only models to be able to perform well.

An interesting analysis would be to test all other hidden states to determine whether performance deteriorates gradually across layers or drops suddenly after a specific one. Another possible analysis would be to test whether we can change one temporal modifier to another — for example changing “in the morning” to “at night” or adding a temporal modifier to a sentence without one.

6. Conclusions and Future Work

We proposed a new intrinsic approach to test and compare BERT-like language models with respect to their capabilities to retain syntactic and semantic information at the sentence level. The tests yield interesting results that enable us to compare different models in terms of their ability to preserve specific syntactic and semantic information across sentences.

Our experiments show that different types of transformations are not equally captured. While gender and number changes were partially captured, the negation transformation was not well represented in later layers.

However, this framework is only an initial proposal and needs to be further developed to provide

a more robust comparison of models. Nonetheless, we expect this approach to raise awareness of the need for improved methods to evaluate transformer-based language models, thereby enabling more accurate comparisons in their current applications.

Future work includes expanding the range of semantic transformations, such as verb tense variants, adjective modifications, synonym-antonym substitutions, and active to passive voice. We also would like to explore sentence embedding models and how those compare to the results we found. Additionally, we plan to extend the dataset with more naturalistic sentences constructed from real-world examples.

7. Bibliographical References

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.

Matteo Bonino, Giorgia Ghione, and Giansalvo Cirrincione. 2025. The geometry of bert.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Hyunjin Choi, Judong Kim, Seongho Joe, and Youngjune Gwon. 2021. Evaluation of bert and albert sentence embedding performance on downstream nlp tasks. In *2020 25th International conference on pattern recognition (ICPR)*, pages 5482–5487. IEEE.

R. Collobert and J. Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *International Conference on Machine Learning, ICML*.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Juan Pablo Filevich, Gonzalo Marco, Santiago Castro, Luis Chiruzzo, and Aiala Rosá. 2024. A language model trained on uruguayan Spanish news text. In *Proceedings of the Second International Workshop Towards Digital Language Equality (TDLE): Focusing on Sustainability @ LREC-COLING 2024*, pages 53–60, Torino, Italia. ELRA and ICCL.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Lukas Lange, Heike Adel, and Jannik Strötgen. 2021. Boosting transformers for job expression extraction and classification in a low-resource setting. In *Proceedings of The Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings.

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, 33:9459–9474.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tomas Mikolov, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *Proceedings of Workshop at ICLR*, 2013.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Hariom A. Pandya, Bhavik Ardeshna, and Dr. Brijesh S. Bhatt. 2021. Cascading adaptors to leverage english data to improve performance of question answering for low-resource languages.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. *Advances in Neural Information Processing Systems*, 32.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *Proceedings of ICLR 2019*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024. Text embeddings by weakly-supervised contrastive pre-training.

Zeyu Yun, Yubei Chen, Bruno Olshausen, and Yann LeCun. 2021. Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 1–10, Online. Association for Computational Linguistics.

8. Language Resource References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Author Index