# The #Somos600M Project: Generating NLP resources that represent the diversity of the languages from LATAM, the Caribbean, and Spain

**María Grandury**
SomosNLP
mariagrandury@somosnlp.org

## Abstract

We are 600 million Spanish speakers. We launched the #Somos600M Project because the diversity of the languages from LATAM, the Caribbean and Spain needs to be represented in Artificial Intelligence (AI) systems. Despite being the 7.5% of the world population, there is no open dataset to instruction-tune large language models (LLMs), nor a leaderboard to evaluate and compare them. In this paper, we present how we have created as an international open-source community the first versions of the instruction and evaluation datasets, indispensable resources for the advancement of Natural Language Processing (NLP) in our languages.

## 1 Introduction

We are 600 million Spanish speakers[1] according to the 2023 Yearbook of the Cervantes Institute (Fernández, 2023), i.e., 7.5% of the global population. However, we lack native instruction-tuned LLMs – models fine-tuned on a collection of NLP tasks formatted as instructions. This fine-tuning improves their zero-shot capabilities (Wei et al.) and adaptability, relevant for AI alignment, chat-like interactions and retrieval augmented generation (RAG) applications. The #Somos600M Project[2], led by SomosNLP[3], aims to create the necessary resources to fine-tune and evaluate these language models.

Spanish is the official language in 22 countries, which implies the existence of a great number of geographic varieties or dialects and influences the performance of language models (Bogantes et al., 2016; Castillo-lópez et al., 2023).

Moreover, in these countries there are many co-official languages from completely different language families like Quechua in Latin America (LATAM) and Euskera in Spain. The scarcity of resources in these languages hinders the development of NLP applications (Hedderich et al., 2021), which worsens the socioeconomic situation of these communities and the risk of extinction of some of these languages (Mager et al., 2018).

With the #Somos600M Project, we emphasize the importance of representing this diversity in LLMs. The initial objectives of the project are:

- **Creation of an open instruction dataset:** A set of input-output pairs that represent various tasks, include different Spanish varieties and co-official languages, and enable the fine-tuning of LLMs to follow instructions.
- **Establishment of an open leaderboard:** Standardizing the evaluation of generative LLMs in Spanish and co-official languages by creating an open impartial leaderboard.

All the generated resources are open-source.[4]

## 2 Previous Work

Since the release of the first Spanish pre-trained language model, BETO (Cañete et al., 2020), there has been a notable increment in the availability of open Spanish resources (Annex A) thanks to various initiatives.[5] There are also shared tasks and projects that aim to create resources in Spanish dialects (Guevara-Rukoz et al., 2020; Hernandez Mena and Meza Ruiz, 2022), indigenous languages of LATAM (Pendas et al., 2023; Ebrahimi et al., 2023), and co-official languages of Spain (Etxaniz et al., 2024; Gonzalez-Aguirre et al., 2024).

However, when it comes to instructions there is a void. Since 2020, we have seen a trend to fine-tune language models using English natural language instructions (Longpre et al., 2023) and there are more than 2,000 "instruct" datasets in the Hugging Face Hub[6]. To the best of our knowledge, in total in

[1] Combining natives and foreign language learners.
[2] somosnlp.org/somos600m
[3] SomosNLP.org is a community of Spanish speakers whose mission is to achieve fair representation of Spanish and co-official languages in the digital world.
[4] huggingface.co/somosnlp
[5] hf.co/spaces/somosnlp/spanish-nlp-initiatives
[6] huggingface.co

our languages there are 227k instructions in Catalan and only 10k originally created in Spanish, by the AINA[7] and ILENIA[8] projects. This forces the Spanish-speaking community to translate and validate English datasets[9] or use multilingual machine translated datasets (Singh et al., 2024).

We proposed the creation of instructions as the task for the 2024 edition of the SomosNLP Hackathon[10]. The general goal of this recurring international online event is the creation of open-source NLP resources in Spanish and co-official languages, encouraging projects with societal impact related to the Sustainable Development Goals. In 2022, we invited the community to fine-tune Transformer architecture models (Vaswani et al., 2017), and in 2023, once again using LoRA-type techniques (Hu et al., 2022), resulting in the publication of interesting projects (Serrano et al., 2022; Vásquez-Rodríguez et al., 2022).

To further strengthen our contribution, we acknowledge the growing need for standard evaluation benchmarks in the realm of Spanish language models. There is a new leaderboard, ODESIA[11], with 15 bilingual Spanish/English discriminative tasks, and another one, CLUB, for Catalan[12]. Regarding text generation, we highlight recent assessments of LLMs' knowledge (Conde et al., 2024; Martínez et al., 2023). Our proposal is to create an open leaderboard that evaluates different capabilities of generative models (e.g., topic knowledge, information extraction, linguistic proficiency, ethical aspects) in our languages and serves as a reference for the Spanish-speaking scientific community.

## 3 The Project

To create a large instruction dataset and a generative LLM leaderboard, we have launched several initiatives: an instruction generation hackathon, a dataset collection campaign, and an effort to translate and validate English evaluation datasets.

### 3.1 Instruction generation

During the SomosNLP 2024 Hackathon, the initial version of the large open instruction dataset was created. The participants had to generate synthetic instruction datasets for the later fine-tuning of an LLM with up to 7B parameters with QLoRA-like techniques (Dettmers et al., 2023). Given the scarcity of resources across all topics, each team was free to choose the theme of their project. The hackathon was open to everyone (Annex B), regardless of prior NLP knowledge, and targeted individuals with both technical and linguistic backgrounds, encouraging interdisciplinary teams.

The teams had access to computing and storage resources, example notebooks, mentorship sessions, workshops, and talks[13] throughout March 2024 up to April 10th, in addition to visibility and prizes to continue developing their NLP skills.

### 3.2 Dataset collection

In addition to generating new resources, reusing existing ones is crucial. Hence, we launched a dataset collection campaign, with a focus on Spanish dialects and co-official languages. Training datasets will be transformed into question-answer pairs (Keskar et al., 2019), while evaluation datasets will be included in the generative LLM leaderboard.

### 3.3 Translation validation

The Open LLM Leaderboard (Beeching et al., 2023) stands as one of the most popular English LLM leaderboards, and some of its constituent datasets were machine-translated as part of the Okapi project (Dac Lai et al., 2023). In collaboration with Hugging Face and Argilla, we launched a community effort for native Spanish speakers to validate these translations. We also joined the international initiative Data Is Better Together (DIBT)[14] to validate the translation of 500 prompts, in order to include Spanish in the corresponding future multilingual leaderboard.

## 4 Results

We present the results with respect to both objectives of the #Somos600M Project.

### 4.1 Instruction datasets

18 projects were presented to the hackathon, resulting in a total of 2,333,052 examples created, summing up to 324 MB of data (Annex C).

We highlight the high number of countries represented in the chosen topics (e.g., Colombian Aeronautical Regulation, Refugee Legal Assistance, Peruvian Constitution, international traditional recipes), as well as the project on Guarani

[7] projecteaina.cat
[8] proyectoilenia.es
[9] hf.co/datasets/somosnlp/somos-clean-alpaca-es
[10] somosnlp.org/hackathon
[11] leaderboard.odesia.uned.es
[12] club.aina.bsc.es
[13] somosnlp.org/eventos
[14] github.com/huggingface/data-is-better-together

culture. Most teams focused on text models, except for one delving into the various accents of rural Spain. A significant amount of data was generated in the healthcare and legal sectors (Figure 1). We also note projects involving clear and inclusive language rewriting, clickbait news summarization, and sustainability text detection.

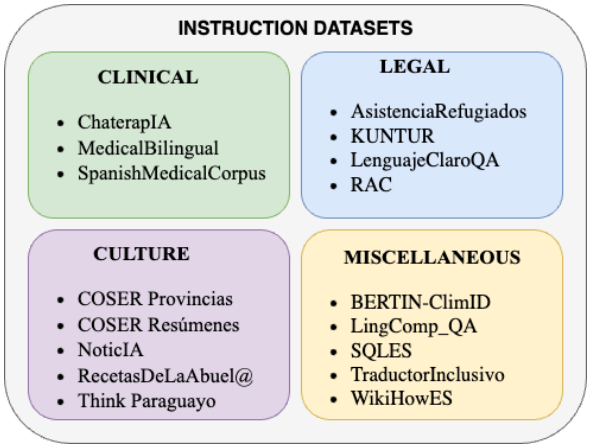


Figure 1: Instruction datasets generated during the #Somos600M Hackathon grouped by domain.

### 4.2 Evaluation datasets

In the first collection campaing round, we received the donation of 5 evaluation datasets manually annotated by experts and, in the second one, 14 more, adding new languages (Figure 2). Together with the translations, they form the first version of the open generative LLM leaderboard (Annex D).

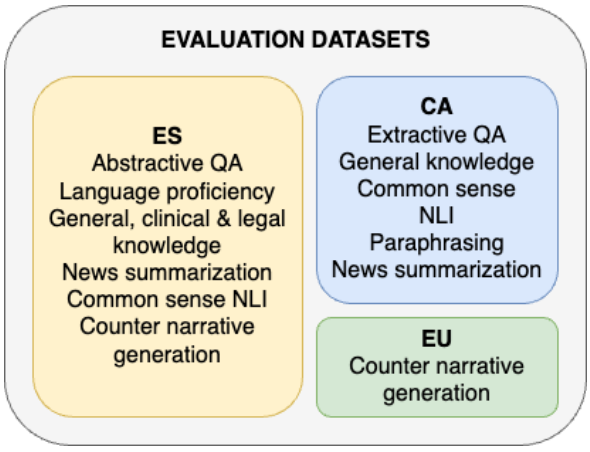


Figure 2: Tasks and languages (ES: Spanish, CA: Catalan and EU: Euskera) of the evaluation datasets of the first version of the open generative LLM leaderboard.

In the validation of the Okapi translations, a total of 61 persons participated, covering a 60% of ARC-C (Clark et al., 2018), 15% of HellaSwag (Zellers et al., 2019) and 15% of MMLU (Hendrycks et al., 2021).[15] Moreover, with the support of 37 persons, 100% of DIBT's prompts were validated, which positioned Spanish as the first language to fully validate its translation.[16]

## 5 Discussion

We are very proud of how the community answered our call. Generating 2 million instructions and gathering 22 evaluation datasets is a great advancement for instruction-tuned LLMs in our languages.

Concerning the hackathon, we are pleased to see that the number of datasets generated tripled last year's. We confirm the usefulness of the libraries `distilabel`, `Argilla` and `transformers` to fine-tune LLMs with instructions synthetically created and manually reviewed. We highlight that a couple of teams also created their own annotation spaces and asked the community for collaboration.

In the translation campaigns, we observed that most of the validation was conducted by 10% of the individuals. For teams interested in organizing similar efforts (and for our next iteration), we recommend: 1) writing a clear annotation guide and enabling a feedback channel to iterate and improve the guide, 2) sharing an instructional video, and 3) creating a visualization of the initiative's progress to motivate and give visibility to the contributors.

## 6 Conclusion

The hackathon, the collection campaign, and the annotation efforts have enabled us to create the initial versions of the large instruction dataset and the open generative LLM leaderboard.

We are going to keep collaborating with entities from LATAM, the Caribbean, and Spain to organize hackathons focused on specific topics, varieties, and languages, scale up the collection campaign to create the most inclusive dataset possible, and expand the leaderboard by including evaluations of ethical (e.g., biases, hate speech) and linguistic (e.g., language variety adequacy) aspects, as well as other co-official languages.

The generated resources are open; we invite entities with greater computing power to use them for training (with our support, if desired) high-quality LLMs that are open, inclusive, and native.

[15] hf.co/spaces/somosnlp/BenchmarkAnnotationDashboard
[16] hf.co/spaces/DIBT/PromptTranslationMultilingualDashboard

## Acknowledgments

We thank all hackathon participants for their efforts. Thanks to their work, we now have the first version of a diverse instruction dataset. We thank Hugging Face for sponsoring the compute and storage resources, LenguajeNaturalAI, Cálamo & Cran, and SaturdaysAI for providing prizes to motivate the participants, and LatinX in AI for inviting us to present our work to the LatinX in NLP workshop. Thank you also to the speakers for sharing their knowledge with the community.

With respect to the leaderboard, we thank Hugging Face and Argilla for co-organizing the translation validation efforts, and all the volunteers who participated in the annotation process. We are also thankful to the Instituto de Ingeniería del Conocimiento (IIC) de la Universidad Autónoma de Madrid (UAM), LenguajeNaturalAI, Grupo de Internet de Nueva Generación (GING) de la Universidad Politécnica de Madrid (UPM), Centro Vasco de Tecnología de la Lengua (HiTZ) and Barcelona Supercomputing Center (BSC) for donating high-quality evaluation datasets.

Finally, we extend our heartfelt gratitude to all those who generously volunteer their time to support our mission of democratizing NLP for the Spanish-speaking community.

## References

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Preprint*, arXiv:2404.05590.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard. `https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard`.

Jaione Bengoetxea, Yi-Ling Chung, Marco Guerini, and Rodrigo Agerri. 2024. Basque and Spanish Counter Narrative Generation: Data Creation and Evaluation. Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING).

Diana Bogantes, Eric Rodríguez, Alejandro Arauco, Alejandro Rodríguez, and Agata Savary. 2016. Towards lexical encoding of multi-word expressions in Spanish dialects. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2255–2261, Portorož, Slovenia. European Language Resources Association (ELRA).

Galo Castillo-lópez, Arij Riabi, and Djamé Seddah. 2023. Analyzing zero-shot transfer scenarios across Spanish variants for hate speech detection. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 1–13, Dubrovnik, Croatia. Association for Computational Linguistics.

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*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

Javier Conde, Miguel González, Nina Melero, Raquel Ferrando, Gonzalo Martínez, Elena Merino-Gómez, José Alberto Hernández, and Pedro Reviriego. 2024. Open source conversational llms do not know most spanish words. *Preprint*, arXiv:2403.15491.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv e-prints*, pages arXiv–2307.

Rubén de la Fuente, Sergio Chicón, and Marta F. Gómez. 2024. Lenguaje claro dataset.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219, Toronto, Canada. Association for Computational Linguistics.

Alvaro Hidalgo Eduardo Muñoz, Teresa Martín. 2024. Asistenciarefugiados.

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. *Preprint*, arXiv:2403.20266.

David et al. Fernández. 2023. *El español en el mundo. Anuario del Instituto Cervantes 2023*. Instituto Cervantes.

Miguel López Pérez Imanuel Rozenberg Josué Sauca Gaia Quintana Fleitas, Andrés Martínez Fernández-Salguero. 2024. Traductor inclusivo.

Iker García-Ferrero and Begoña Altuna. 2024. Noticia: A clickbait article summarization dataset in spanish. *Preprint*, arXiv:2404.07611.

Gabriela Zuñiga Gerardo Huerta. 2024. Dataset for bertin-climid: Bertin-base climate-related text identification.

Iakes Goenaga, Aitziber Atutxa, Koldo Gojenola, Maite Oronoz, and Rodrigo Agerri. 2023. Explanatory argument extraction of correct answers in resident medical exams. *Preprint*, arXiv:2312.00567.

Aitor Gonzalez-Aguirre, 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 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. European Language Resources Association and the International Committee on Computational Linguistics.

Adriana Guevara-Rukoz, Isin Demirsahin, Fei He, Shan-Hui Cathy Chu, Supheakmungkol Sarin, Knot Pipatsrisawat, Alexander Gutkin, Alena Butryna, and Oddur Kjartansson. 2020. Crowdsourcing Latin American Spanish for low-resource text-to-speech. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6504–6513, Marseille, France. European Language Resources Association.

Michael Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. pages 2545–2568.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Carlos Daniel Hernandez Mena and Ivan Vladimir Meza Ruiz. 2022. Creating Mexican Spanish language resources through the social service program. In *Proceedings of the 2nd Workshop on Novel Incentives in Data Collection from People: models, implementations, challenges and results within LREC 2022*, pages 20–24, Marseille, France. European Language Resources Association.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Instituto de Ingeniería del Conocimiento. 2024a. Abstractive question-answering in spanish (aquas) dataset.

Instituto de Ingeniería del Conocimiento. 2024b. Retrieval-augmented-generation and queston-answering in spanish (ragquas) dataset.

Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Unifying question answering, text classification, and regression via span extraction. *Preprint*, arXiv:1904.09286.

LenguajeNaturalAI. 2024a. Humorqa.

LenguajeNaturalAI. 2024b. Medicalexpertes.

LenguajeNaturalAI. 2024c. Spalawex.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: designing data and methods for effective instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Montoya Dylan-Bermúdez Daniel Lopez Dionis, Garcia Alvaro. 2024. Spanishmedicallm.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Gonzalo Martínez, Javier Conde, Pedro Reviriego, Elena Merino-Gómez, José Alberto Hernández, and Fabrizio Lombardi. 2023. How many words does chatgpt know? the answer is chatwords. *Preprint*, arXiv:2309.16777.

Andrea Morales-Garzón, Oscar A. Rocha, Sara Benel Ramirez, Gabriel Tuco Casquino, and Alberto Medina. 2024. Recetasdelaabuel@.

Begoña Pendas, Andres Carvallo, and Carlos Aspillaga. 2023. Neural machine translation through active learning on low-resource languages: The case of Spanish to Mapudungun. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 6–11, Toronto, Canada. Association for Computational Linguistics.

David Alonso Quispe Castillo. 2024a. Kuntur: Asistencia legal en perú.

David Alonso Quispe Castillo. 2024b. Wikihowes.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2024. Meta4xnli: A crosslingual parallel corpus for metaphor detection and interpretation. *Preprint*, arXiv:2404.07053.

Edison Jair Bejarano Sepulveda, Nicolai Potes Hector, Santiago Pineda Montoya, Felipe Ivan Rodriguez, Jaime Enrique Orduy, Alec Rosales Cabezas, Danny Traslaviña Navarrete, and Sergio Madrid Farfan. 2024. Towards enhanced rac accessibility: Leveraging datasets and llms. *Preprint*, arXiv:2405.08792.

Alejandro Vaca Serrano, David Betancur Sánchez, Alba Segurado, Guillem García Subies, and Álvaro Barbero Jiménez. 2022. Biomedia: A complete voice-to-voice generative question answering system for the biomedical domain in spanish. In *North American Chapter of the Association for Computational Linguistics Conference: LatinX in AI (LXAI) Research Workshop*.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *Preprint*, arXiv:2402.06619.

Laura Vásquez-Rodríguez, Pedro-Manuel Cuenca-Jiménez, Sergio Morales-Esquivel, and Fernando Alva-Manchego. 2022. A benchmark for neural readability assessment of texts in Spanish. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 188–198, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

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.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jorge Zamora Rey, Mario Crespo Miguel, and Isabel Moyano Moreno. 2024. Lingcomp_qa, un corpus educativo de lingüística computacional en español.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *Preprint*, arXiv:1905.07830.

## A Spanish resources in the Hugging Face Hub

Even though the number of open-source NLP resources in Spanish on the Hugging Face Hub is increasing, the gap between Spanish and English remains substantial (Figure 3).

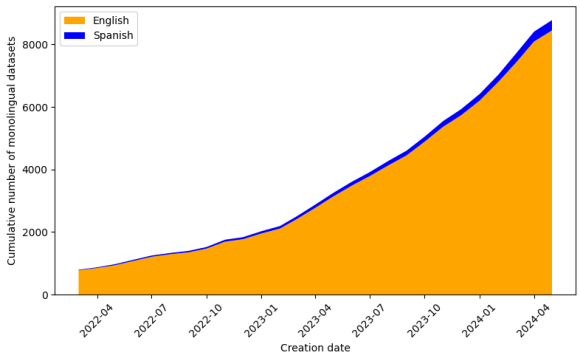


Figure 3: Cumulative number of monolingual English (orange) and Spanish (blue) datasets in the Hugging Face Hub over time until May 13 2024.

## B Hackathon participants

651 people from 29 countries registered for the hackathon, of which 222 were interested exclusively in attending the talks. It is worth noting that at least[17] 14% of the individuals had already participated in one of the previous editions, almost 50% of the participants were in LATAM (Table 1) and less than 40% self-identify as non-male (Table 2), these are numbers we would like to increase in future editions. With respect to the background and occupation of the participants, more than 40% thought they had a fundamental NLP level before starting the hackathon (Table 3) and most of them are "developers", "engineers" or "data scientists" (Figure 4).

| LATAM | Spain | Other |
|---|---|---|
| 46% | 32% | 22% |

Table 1: Localization of the #Somos600M Hackathon participants, might not match the nationality.

| Female | Male | NB | NR |
|---|---|---|---|
| 22% | 60% | 1% | 17% |

Table 2: Self-identified gender of the #Somos600M Hackathon participants: "female", "male", "NB" (no binary) or "NR" (no response).

| Fund. | Intermediate | Advanced | NR |
|---|---|---|---|
| 40% | 31% | 12% | 17% |

Table 3: Self-assigned NLP level of the Somos600M Hackathon participants: "fundamental", "intermediate", "advanced", "NR" (no response).

[17]The questions were optional.

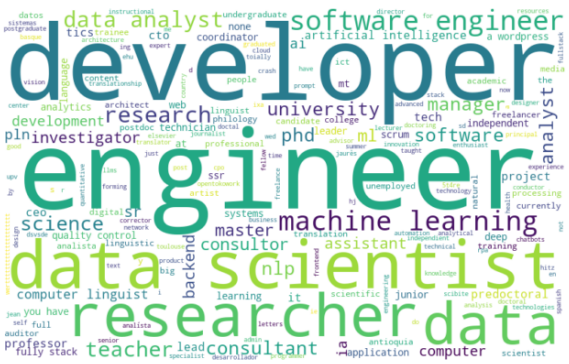


Figure 4: Wordcloud of the occupations of the Hackathon #Somos600M participants translated to English.

## C Instruction dataset details

We enumerate the datasets created during the #Somos600M Hackathon (Table 4): **AsistenciaRefugiados**: Legal assistant for refugees or individuals in political asylum (Eduardo Muñoz, 2024), **BERTIN-ClimID**: BERTIN-Base Climate-related text Identification (Gerardo Huerta, 2024), **ChaterapIA**: Therapeutic conversations, **COSER Provincias**: Identification of provinces from oral interviews in rural Spain, **COSER Resúmenes**: Summaries of oral interviews in rural Spain, **KUNTUR**: Legal assistance in Peru (Quispe Castillo, 2024a), **LenguajeClaroQA**: Simplification of administrative language (de la Fuente et al., 2024), **LingComp_QA**: Educational corpus about computational linguistics (Zamora Rey et al., 2024), **NoticIA**: Clickbait Article Summarization Dataset (García-Ferrero and Altuna, 2024), **RAC**: Colombian Aeronautical Regulation (Sepulveda et al., 2024), **RecetasDeLaAbuel@**: Recipes from Spanish-speaking countries (Morales-Garzón et al., 2024), **SMC**: Spanish Medical Corpus (Lopez Dionis, 2024), **SQLES**: Interact with an SQL database in Spanish, **ThinkParaguayo**: Get to know the Guaraní culture, **TraductorInclusivo**: Rewriting of texts using inclusive language (Gaia Quintana Fleitas, 2024), and **WikiHowES** (Quispe Castillo, 2024b).

## D Evaluation dataset details

The evaluation datasets included in the first version of the generative LLM leaderboard (Table 5) are: **AQuAS**: Abstractive Question-Answering in Spanish (Instituto de Ingeniería del Conocimiento, 2024a) and **RagQuAS**: Retrieval-Augmented-Generation and Question-Answering in Spanish (Instituto de Ingeniería del Conocimiento, 2024b), donated by the Instituto de Ingeniería del Conocimiento (IIC) of the Universidad Autónoma de Madrid (UAM); **SpaLawEx**: Spanish Law School Access Exams (LenguajeNaturalAI, 2024c), **MedicalExpertES**: Diagnosis and treatment of clinical cases in Spanish (LenguajeNaturalAI, 2024b) and **HumorQA**: White humour joke classification (LenguajeNaturalAI, 2024a), donated by LenguajeNaturalAI; **TELEIA**: Test de Español como Lengua Extranjera para Inteligencia Artificial, donated by the Next Generation Internet Group (GING) of the Universidad Politécnica de Madrid (UPM); **Meta4XNLI**: A Crosslingual Parallel Corpus for Metaphor Detection and Interpretation (Sanchez-Bayona and Agerri, 2024), **NoticIA**: A Clickbait Article Summarization Dataset in Spanish (García-Ferrero and Altuna, 2024), **MedExpQA**: Multilingual Benchmarking of Large Language Models for Medical Question Answering (Alonso et al., 2024), **CasiMedicos-SQUAD**: Explanatory Argument Extraction of Correct Answers in Resident Medical Exams (Goenaga et al., 2023) and **CONAN-EUS**: Basque and Spanish Parallel Counter Narratives Dataset (Bengoetxea et al., 2024) donated by the Basque Center for Language Technology (HiTZ); **CatalanQA**: Extractive QA in Catalan, **TE_ca**: Textual Entailment in Catalan, **caBREU**: Article Summarization in Catalan, **XNLI_ca**: Cross-lingual Natural Language Inference in Catalan, **COPA_ca**: Choice Of Plausible Alternatives in Catalan, **PAWS_ca**: Paraphrase Adversaries from Word Scrambling in Catalan, **XQUAD_ca**: Cross-lingual Question Answering Dataset in Catalan and **WNLI_ca**: Winograd NLI dataset, created or professionally translated with funding of the Project AINA and donated by the Barcelona Supercomputing Center (BSC) (Gonzalez-Aguirre et al., 2024).

| Dataset | # Examples | MB | Domain | Country |
|---|---|---|---|---|
| AsistenciaRefugiados | 10707 | 20.7 | Legal | ES, MX, VE + |
| BERTIN-ClimID | 3680 | 1.63 | Sustainability | PE, ES |
| ChaterapIA | 1000 | 2.30 | Psychology | ES |
| COSER Provincias | 1150 | 0.22 | Rural culture | ES (oral) |
| COSER Resúmenes | 230 | 1.08 | Rural culture | ES (oral) |
| KUNTUR | 2075 | 0.73 | Legal | PE |
| LenguajeClaroQA | 4094 | 1.72 | Legal admin. | ES |
| LingComp_QA | 1004 | 0.35 | Comp. Linguistics | ES |
| MedicalBilingual | 8138 | 12.8 | Clinical | Mix |
| NoticIA | 850 | 3.41 | Press | ES |
| RAC | 24478 | 1.84 | Legal | CO |
| RecetasDeLaAbuel@ | 20221 | 42.4 | Gastronomy | ES, MX, PE, AR+ |
| SpanishMedicalCorpus | 2136490 | 48.5 | Clinical | ES, CL |
| SQLES | 81 | 0.40 | Programming | - |
| Think Paraguayo | 1498 | 0.19 | Guarani culture | PY |
| TraductorInclusivo | 4196 | 0.40 | Miscellaneous | ES, AR, MX, CL, CR + |
| WikiHowES | 113160 | 186 | Miscellaneous | Mix |
| Total | 2,333,052 | 324.67 | - | - |

Table 4: Instruction datasets generated by the teams participating in the #Somos600M Hackathon, available at huggingface.co/somosnlp. We excluded the dataset versions adapted to specific formats for model training (e.g. Gemma). The countries are represented by their corresponding code ISO 3166-1 alfa-2, the symbol "+" means that there are other countries represented in the corpus with a lower percentage.

| Dataset | Language | Domain | Task |
|---|---|---|---|
| AQuAS | ES | Miscellaneous | Abstractive QA |
| RagQuAS | ES | Miscellaneous | RAG and Abstractive QA |
| HellaSwag_es | ES | Miscellaneous | Commonsense NLI |
| MMLU_es | ES | Miscellaneous | Multiple choice QA |
| TELEIA | ES | Language proficiency | Multiple choice QA |
| Meta4XNLI | ES | Language proficiency | NLI |
| HumorQA | ES | Language proficiency | Classification |
| ARC-C_es | ES | Science | Multiple choice QA |
| NoticIA | ES | Press | Summarization |
| SpaLawEx | ES | Legal | Multiple choice QA |
| MedicalExpertES | ES | Clinical | Open QA |
| MedExpQA | ES | Clinical | Multiple choice QA |
| CasiMedicos-SQUAD | ES | Clinical | Extractive QA |
| CONAN-EUS | ES, EU | Hate speech | Counter narrative generation |
| CatalanQA | CA | Miscellaneous | Extractive QA |
| TE_ca | CA | Miscellaneous | NLI |
| XNLI_ca | CA | Miscellaneous | NLI |
| WNLI_ca | CA | Miscellaneous | NLI |
| COPA_ca | CA | Miscellaneous | Commonsense Reasoning |
| PAWS_ca | CA | Miscellaneous | Paraphrasing |
| XQUAD_ca | CA | Miscellaneous | Extractive QA |
| caBREU | CA | Press | Summarization |

Table 5: Datasets of the first version of the generative LLM leaderboard, that includes tasks in Spanish (ES), Catalan (CA) and Euskera (EU) Corpus and evaluates abstractive and extractive QA, general, clinical and legal knowledge, common sense reasoning, natural language inference (NLI) and language proficiency.

# Proyecto #Somos600M: Generación de recursos de PLN que representen la riqueza de las lenguas de LATAM, El Caribe y España

**María Grandury**
SomosNLP
mariagrandury@somosnlp.org

## Abstract

Somos 600 millones de hispanohablantes. Lanzamos el Proyecto #Somos600M porque necesitamos que la riqueza de nuestras lenguas esté representada en los sistemas de Inteligencia Artificial (IA). A pesar de ser el 7.5% de la población mundial, no contamos con un gran corpus de instrucciones abierto que nos permita adaptar grandes modelos de lenguaje generativos, ni con una tabla de clasificación estándar para evaluarlos y compararlos. En este artículo presentamos cómo hemos creado en comunidad la primera versión de los corpus de instrucciones y de evaluación, recursos imprescindibles para el avance del Procesamiento de Lenguaje Natural (PLN) en nuestras lenguas.

## 1 Introducción

Somos cerca de 600 millones de hispanohablantes[1] según el Anuario del Instituto Cervantes 2023 (Fernández, 2023), es decir, el 7.5% de la población mundial. Sin embargo, no contamos con grandes modelos de lenguaje (LLM, del inglés *Large Language Model*) propios adaptados para seguir instrucciones (o *prompts*). Esta adaptación mejora la versatilidad de los modelos (Wei et al.), importante para el alineamiento de la IA y aplicaciones de tipo conversacional y RAG (*Retrieval Augmented Generation*). El Proyecto #Somos600M, liderado por SomosNLP[2], tiene por objetivo crear los recursos necesarios para adaptar y evaluar estos modelos.

El español es lengua oficial en 22 países, lo que implica la existencia de una gran cantidad de variedades geográficas o dialectos, que influyen en el rendimiento de los modelos de PLN (Bogantes et al., 2016; Castillo-lópez et al., 2023).

Además, en estos países se hablan otras lenguas cooficiales de diferentes familias, como el quechua en LATAM y el euskera en España. La escasez de recursos en estas lenguas dificulta el desarrollo de modelos de lenguaje (Hedderich et al., 2021), lo que empeora la situación socioeconómica de estas comunidades y el riesgo de extinción de algunas de estas lenguas (Mager et al., 2018).

Con el proyecto #Somos600M hacemos hincapié en la representación de las variedades del español y las lenguas cooficiales en los recursos de PLN. Los objetivos iniciales del proyecto son:

- **Crear un gran corpus abierto de instrucciones**: Un conjunto de pares pregunta-respuesta que represente las variedades del español y lenguas cooficiales y permita adaptar modelos que sigan instrucciones.
- **Crear una tabla de clasificación abierta**: Estandarizar la evaluación de modelos de lenguaje generativos en español y lenguas cooficiales mediante la creación de una tabla de clasificación abierta e imparcial.

Todos los recursos generados son abiertos.[3]

## 2 Antecedentes

Desde la publicación del primer modelo de lenguaje pre-entrenado en español, BETO (Cañete et al., 2020), hemos visto un aumento del número de recursos en español y lenguas cooficiales disponibles (Anexo A) gracias a varias iniciativas.[4]. También hay talleres en congresos y proyectos para crear recursos en dialectos (Guevara-Rukoz et al., 2020; Hernandez Mena and Meza Ruiz, 2022), lenguas originarias de LATAM (Pendas et al., 2023; Ebrahimi et al., 2023), y lenguas cooficiales de España (Etxaniz et al., 2024).

En lo referente a instrucciones, desde 2020 hemos visto una tendencia a adaptar LLMs utilizando instrucciones en inglés (Longpre et al.,

[1] Suma de los grupos de dominio nativo, competencia limitada y aprendices de lengua extranjera.

[2] SomosNLP.org es una comunidad de hispanohablantes cuya misión es conseguir una justa representación del español y lenguas cooficiales en el mundo digital.

[3] huggingface.co/somosnlp

[4] hf.co/spaces/somosnlp/spanish-nlp-initiatives

2023) Sin embargo, hasta donde sabemos existen 227k instrucciones en catalán y 10k originales en español, creadas por el Proyecto ILENIA[5]. Esto obliga a la comunidad hispanohablante a utilizar instrucciones traducidas automáticamente (Singh et al., 2024).

Así, propusimos la creación de instrucciones como tarea del Hackathon SomosNLP 2024[6]. El objetivo general recurrente de este evento internacional en línea es la creación de recursos abiertos de PLN en español y lenguas cooficiales, con enfoque en impulsar proyectos con impacto social. En 2022 hicimos una llamada a la comunidad para adaptar modelos con arquitectura Transformer (Vaswani et al., 2023) mediante fine-tuning y en 2023 con técnicas tipo LoRA (Hu et al., 2022), resultando en la publicación de proyectos interesantes (Serrano et al., 2022; Vásquez-Rodríguez et al., 2022).

Reforzamos nuestra contribución atendiendo a la necesidad creciente de evaluar nuestros modelos. Existe una nueva tabla clasificación para modelos discriminativos, ODESIA[7], y una para catalán, CLUB[8]. Respecto a la generación de texto, destacamos recientes evaluaciones del conocimiento de los LLMs (Conde et al., 2024; Martínez et al., 2023). Nuestra propuesta es crear una tabla de clasificación abierta que evalúe diferentes capacidades de los modelos generativos (e.g., dominio de un tema, extracción de información, adecuación lingüística, aspectos éticos) y sirva de referencia para la comunidad científica hispanohablante.

## 3 El Proyecto

Para crear un gran corpus de instrucciones y una tabla de clasificación de modelos generativos hemos lanzado varias iniciativas: un hackatón de generación instrucciones, una campaña de recolección de corpus y la traducción y validación de corpus de evaluación en inglés.

### 3.1 Generación de instrucciones

Aprovechamos el Hackathon SomosNLP 2024 para crear la primera versión del gran corpus abierto de instrucciones. La tarea de los equipos era generar instrucciones sintéticas para la posterior adaptación de un LLM de hasta 7B con técnicas tipo QLoRA (Dettmers et al., 2023). Dada la escasez de recursos en todas las temáticas, dejamos a elección de cada equipo el tema concreto de su proyecto. El hackatón estaba abierto a todo el mundo (Anexo B), sin requerimientos de conocimientos previos de PLN, y dirigido a personas tanto de formación técnica como lingüística, animando la creación de equipos interdisciplinares.

Los equipos participantes tuvieron acceso durante el mes de marzo de 2024 y hasta el 10 de abril a recursos de computación y almacenamiento, tutoriales, mentorías, talleres y charlas[9], además de visibilidad y premios para seguir formándose.

### 3.2 Recolección de corpus

Además de generar nuevos recursos, es importante reutilizar los existentes. Lanzamos una campaña de recolección de corpus[10], con especial foco en las diferentes variedades del español y lenguas cooficiales. Las bases de datos de entrenamiento se utilizarán para generar pares pregunta-respuesta (Keskar et al., 2019). Las de evaluación se incluirán en la tabla de clasificación de modelos generativos.

### 3.3 Validación de traducciones

La *Open LLM Leaderboard* (Beeching et al., 2023) es una de las tablas de clasificación más populares para evaluar LLMs en inglés y algunas de las bases de datos que la constituyen fueron traducidas automáticamente como parte del proyecto Okapi (Dac Lai et al., 2023). Lanzamos con Hugging Face y Argilla una iniciativa para validar en comunidad dichas traducciones. También nos sumamos a la iniciativa internacional *Data Is Better Together* (DIBT) [11] para validar la traducción de 500 instrucciones y que el español forme parte de la correspondiente futura tabla de clasificación multilingüe.

## 4 Resultados

Exponemos los resultados respecto a los dos objetivos del Proyecto #Somos600M.

### 4.1 Corpus de instrucciones

Se presentaron al hackatón 18 proyectos y en total se crearon 2,333,052 ejemplos, que se traducen en 324 MB de datos (Anexo C).

Destacamos la gran variedad de países representados en los proyectos (e.g., Reglamento Aeronáutico Colombiano, Asistencia a refugiados, Constitución del Perú, recetas típicas), así como el proyecto sobre cultura guaraní. La mayoría de los equipos se

[5] proyectoilenia.es
[6] somosnlp.org/hackathon
[7] leaderboard.odesia.uned.es
[8] club.aina.bsc.es
[9] somosnlp.org/eventos
[10] somosnlp.org/donatucorpus
[11] github.com/huggingface/data-is-better-together

centraron en modelos de texto, excepto un proyecto que se enfocó en las diferentes maneras de hablar en la España rural. Se generaron gran cantidad de datos de los sectores salud y legal (Figura 1). También destacamos los proyectos de reescritura con lenguaje claro e inclusivo, resumen de noticias clickbait y detección de textos sobre sostenibilidad.

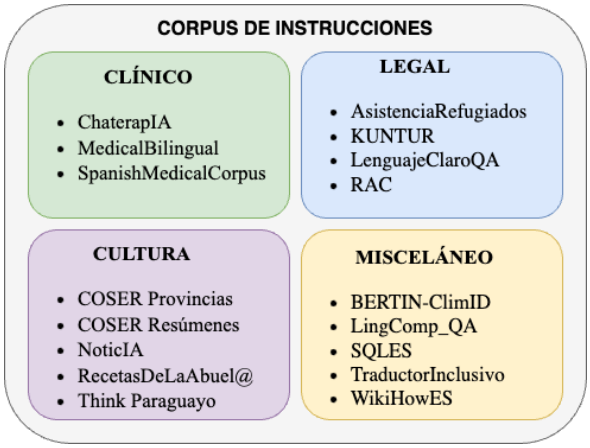


Figure 1: Corpus de instrucciones generados durante el Hackathon #Somos600M agrupados por dominio.

### 4.2 Corpus de evaluación

Con la primera ronda de la campaña de recolección conseguimos la donación de 5 corpus de evaluación anotados manualmente por especialistas y, con la segunda, 14 más. Combinados con las traducciones constituirán la primera tabla de clasificación de LLMs generativos en español (Figura 2).

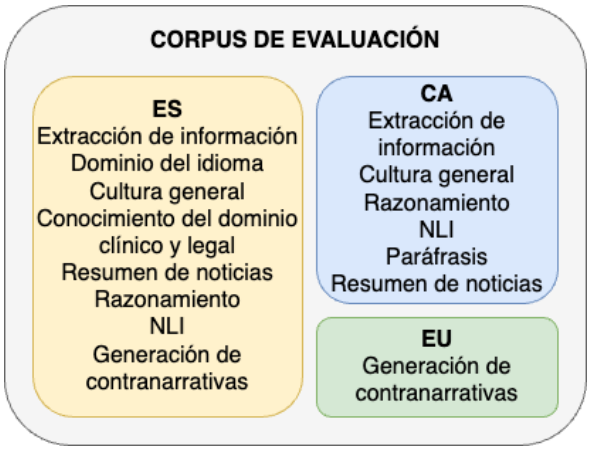


Figure 2: Corpus que constituyen la primera versión de la tabla de clasificación de modelos de lenguaje generativos.

En la validación de las traducciones de Okapi participaron un total de 61 personas y se cubrió un 60% de ARC-C (Clark et al., 2018), 15% de HellaSwag (Zellers et al., 2019) y 15% de MMLU (Hendrycks et al., 2021). [12]. Con el apoyo de 37 personas se validó el 100% de las instrucciones de DIBT, convirtiendo al español en el primer idioma en validar al completo su traducción. [13]

## 5 Discusión

Nos llena de satisfacción ver la gran respuesta de la comunidad a nuestra llamada. La creación de 2 millones de instrucciones y la recolección de 22 corpus de evaluación es un gran avance para los LLMs en nuestras lenguas.

Respecto al hackathon, nos alegra ver que el número de bases de datos generados triplica el del año pasado. Confirmamos la utilidad de las librerías `distilabel`, `Argilla` y `transformers` para el desarrollo de LLMs adaptados con instrucciones sintéticas y revisadas manualmente.

En las campañas de anotación observamos que la mayor parte de la validación fue hecha por un 10% de las personas. Para esfuerzos similares recomendamos: 1) escribir una guía de anotación clara, habilitar un canal de comunicación y utilizar las dudas para iterar y mejorar la guía, 2) compartir un vídeo de ejemplo y 3) crear una visualización del progreso de la iniciativa para motivar y dar visibilidad a las personas voluntarias.

## 6 Conclusión

El hackatón, la campaña de recolección y los esfuerzos de anotación nos han permitido crear las primeras versiones del gran corpus de instrucciones y la tabla de clasificación de LLMs generativos.

Vamos a continuar aunando esfuerzos con entidades de LATAM, el Caribe y España para organizar más hackatones enfocados en temas, variedades o lenguas específicas, escalar la campaña de recolección para crear un corpus lo más inclusivo posible, y extender la tabla de clasificación incluyendo evaluaciones de aspectos éticos (sesgos y discurso de odio) y lingüísticos (e.g. adecuación de la variedad de la lengua generada), así como otras lenguas cooficiales.

Los recursos generados son abiertos[14], invitamos a entidades con mayor poder de computación a utilizarlos para entrenar (con nuestro apoyo, si desean) grandes modelos de lenguaje generativos abiertos, de calidad, inclusivos y nativos.

[12] hf.co/spaces/somosnlp/BenchmarkAnnotationDashboard
[13] hf.co/spaces/DIBT/PromptTranslationMultilingualDashboard
[14] huggingface.co/somosnlp

## Agradecimientos

Agradecemos el esfuerzo de todas las personas participantes en el hackatón, gracias a su trabajo contamos ahora con la primera versión de un gran corpus de instrucciones diverso. Damos gracias a Hugging Face por patrocinar los recursos de computación y a LenguajeNaturalAI, Cálamo & Cran y SaturdaysAI por patrocinar premios para motivar a los equipos y a LatinX in AI por invitarnos a presentar los proyectos al workshop LatinX in NLP. Gracias también a todas las personas que compartieron su conocimiento con la comunidad en ponencias y talleres.

Respecto a la tabla de clasificación, damos las a Hugging Face y Argilla por co-organizar los esfuerzos de validación de traducciones y a todas las personas voluntarias que participaron en la anotación. También agradecemos las donaciones de bases de datos de evaluación de calidad al Instituto de Ingeniería del Conocimiento (IIC), LenguajeNaturalAI, Grupo de Internet de Nueva Generación (GING) de la Universidad Politécnica de Madrid (UPM), Centro Vasco de Tecnología de la Lengua (HiTZ) y Barcelona Supercomputing Center (BSC).

Para finalizar, damos las gracias de corazón a todas las personas que voluntariamente ofrecen su tiempo para apoyar nuestra misión de democratizar el PLN para la comunidad hispanohablante.

## References

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Preprint*, arXiv:2404.05590.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard. `https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard`.

Jaione Bengoetxea, Yi-Ling Chung, Marco Guerini, and Rodrigo Agerri. 2024. Basque and Spanish Counter Narrative Generation: Data Creation and Evaluation. Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING).

Diana Bogantes, Eric Rodríguez, Alejandro Arauco, Alejandro Rodríguez, and Agata Savary. 2016. Towards lexical encoding of multi-word expressions in Spanish dialects. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2255–2261, Portorož, Slovenia. European Language Resources Association (ELRA).

Galo Castillo-lópez, Arij Riabi, and Djamé Seddah. 2023. Analyzing zero-shot transfer scenarios across Spanish variants for hate speech detection. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 1–13, Dubrovnik, Croatia. Association for Computational Linguistics.

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*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

Javier Conde, Miguel González, Nina Melero, Raquel Ferrando, Gonzalo Martínez, Elena Merino-Gómez, José Alberto Hernández, and Pedro Reviriego. 2024. Open source conversational llms do not know most spanish words. *Preprint*, arXiv:2403.15491.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv e-prints*, pages arXiv–2307.

Rubén de la Fuente, Sergio Chicón, and Marta F. Gómez. 2024. Lenguaje claro dataset.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219, Toronto, Canada. Association for Computational Linguistics.

Alvaro Hidalgo Eduardo Muñoz, Teresa Martín. 2024. Asistenciarefugiados.

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. *Preprint*, arXiv:2403.20266.

David et al. Fernández. 2023. *El español en el mundo. Anuario del Instituto Cervantes 2023*. Instituto Cervantes.

Miguel López Pérez Imanuel Rozenberg Josué Sauca Gaia Quintana Fleitas, Andrés Martínez Fernández-Salguero. 2024. Traductor inclusivo.

Iker García-Ferrero and Begoña Altuna. 2024. Noticia: A clickbait article summarization dataset in spanish. *Preprint*, arXiv:2404.07611.

Gabriela Zuñiga Gerardo Huerta. 2024. Dataset for bertin-climid: Bertin-base climate-related text identification.

Iakes Goenaga, Aitziber Atutxa, Koldo Gojenola, Maite Oronoz, and Rodrigo Agerri. 2023. Explanatory argument extraction of correct answers in resident medical exams. *Preprint*, arXiv:2312.00567.

Aitor Gonzalez-Aguirre, 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 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. European Language Resources Association and the International Committee on Computational Linguistics.

Adriana Guevara-Rukoz, Isin Demirsahin, Fei He, Shan-Hui Cathy Chu, Supheakmungkol Sarin, Knot Pipatsrisawat, Alexander Gutkin, Alena Butryna, and Oddur Kjartansson. 2020. Crowdsourcing Latin American Spanish for low-resource text-to-speech. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6504–6513, Marseille, France. European Language Resources Association.

Michael Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. pages 2545–2568.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Carlos Daniel Hernandez Mena and Ivan Vladimir Meza Ruiz. 2022. Creating Mexican Spanish language resources through the social service program. In *Proceedings of the 2nd Workshop on Novel Incentives in Data Collection from People: models, implementations, challenges and results within LREC 2022*, pages 20–24, Marseille, France. European Language Resources Association.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Instituto de Ingeniería del Conocimiento. 2024a. Abstractive question-answering in spanish (aquas) dataset.

Instituto de Ingeniería del Conocimiento. 2024b. Retrieval-augmented-generation and queston-answering in spanish (ragquas) dataset.

Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Unifying question answering, text classification, and regression via span extraction. *Preprint*, arXiv:1904.09286.

LenguajeNaturalAI. 2024a. Humorqa.

LenguajeNaturalAI. 2024b. Medicalexpertes.

LenguajeNaturalAI. 2024c. Spalawex.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: designing data and methods for effective instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Montoya Dylan-Bermúdez Daniel Lopez Dionis, Garcia Alvaro. 2024. Spanishmedicallm.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Gonzalo Martínez, Javier Conde, Pedro Reviriego, Elena Merino-Gómez, José Alberto Hernández, and Fabrizio Lombardi. 2023. How many words does chatgpt know? the answer is chatwords. *Preprint*, arXiv:2309.16777.

Andrea Morales-Garzón, Oscar A. Rocha, Sara Benel Ramirez, Gabriel Tuco Casquino, and Alberto Medina. 2024. Recetasdelaabuel@.

Begoña Pendas, Andres Carvallo, and Carlos Aspillaga. 2023. Neural machine translation through active learning on low-resource languages: The case of Spanish to Mapudungun. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 6–11, Toronto, Canada. Association for Computational Linguistics.

David Alonso Quispe Castillo. 2024a. Kuntur: Asistencia legal en perú.

David Alonso Quispe Castillo. 2024b. Wikihowes.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2024. Meta4xnli: A crosslingual parallel corpus for metaphor detection and interpretation. *Preprint*, arXiv:2404.07053.

Edison Jair Bejarano Sepulveda, Nicolai Potes Hector, Santiago Pineda Montoya, Felipe Ivan Rodriguez, Jaime Enrique Orduy, Alec Rosales Cabezas, Danny Traslaviña Navarrete, and Sergio Madrid Farfan. 2024. Towards enhanced rac accessibility: Leveraging datasets and llms. *Preprint*, arXiv:2405.08792.

Alejandro Vaca Serrano, David Betancur Sánchez, Alba Segurado, Guillem García Subies, and Álvaro Barbero Jiménez. 2022. Biomedia: A complete voice-to-voice generative question answering system for the biomedical domain in spanish. In *North American Chapter of the Association for Computational Linguistics Conference: LatinX in AI (LXAI) Research Workshop*.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *Preprint*, arXiv:2402.06619.

Laura Vásquez-Rodríguez, Pedro-Manuel Cuenca-Jiménez, Sergio Morales-Esquivel, and Fernando Alva-Manchego. 2022. A benchmark for neural readability assessment of texts in Spanish. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 188–198, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jorge Zamora Rey, Mario Crespo Miguel, and Isabel Moyano Moreno. 2024. Lingcomp_qa, un corpus educativo de lingüística computacional en español.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *Preprint*, arXiv:1905.07830.

## A Recursos de español en el Hub de Hugging Face

Aunque el número de recursos abiertos de PLN en español en el Hub de Hugging Face esté aumentando, la brecha entre el español y el inglés sigue siendo inmensa (Figura 3).

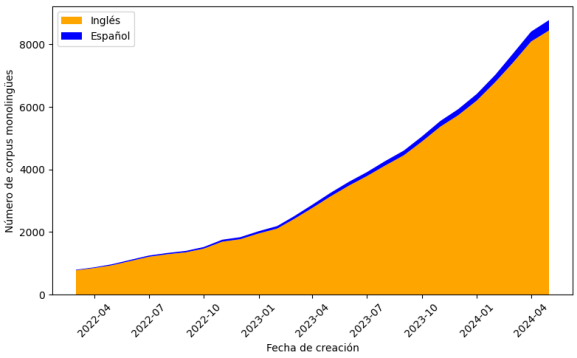


Figure 3: Evolución del número de corpus monolingües en inglés (naranja) y español (azul) en el Hub de Hugging Face hasta el 13 de mayo de 2024.

## B Participantes del hackatón

651 personas de 29 países se registraron para el hackatón, de las cuales 222 estaban interesadas únicamente en asistir a las charlas. Cabe destacar que como mínimo[15] el 14% de las personas ya habían participado en alguna de las dos ediciones anteriores, casi el 50% se conectaba desde LATAM (Tabla 1) y menos del 40% auto-identifica su género como no-masculino (Tabla 2), nos gustaría incrementar estos tres números en próximas ediciones. Respecto al nivel y la ocupación, más del 40% de las personas afirmaron tener un nivel de PLN fundamental antes de comenzar el hackatón (Tabla 3) y la mayor parte son desarrolladores/as, ingenieros/as o científicos/as de datos (Figura 4).

| LATAM | España | Otros |
|---|---|---|
| 46% | 32% | 22% |

Table 1: Localización de las personas participantes en el Hackathon #Somos600M, puede no coincidir con la nacionalidad.

| Femenino | Masculino | NB | SR |
|---|---|---|---|
| 22% | 60% | 1% | 17% |

Table 2: Género con el que se auto-identifican las personas participantes en el Hackathon #Somos600M: "femenino", "masculino", "NB" (no binario) o "SR" (sin respuesta).

## C Detalle de los corpus del hackatón

Enumeramos los corpus creados durante el Hackathon #Somos600M, detallados en la Tabla 4:

[15] Las preguntas no son de obligatoria respuesta.

| Fund. | Intermedio | Avanzado | SR |
|---|---|---|---|
| 40% | 31% | 12% | 17% |

Table 3: Nivel auto-asignado de las personas participantes en el Hackathon #Somos600M: "fundamental", "intermedio", "avanzado", "SR" (sin respuesta)"

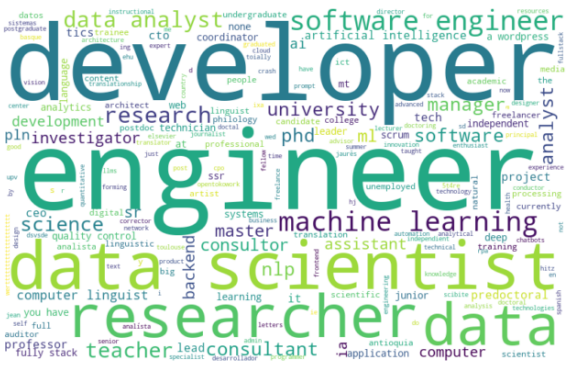


Figure 4: Nube de palabras representando las ocupaciones de las personas que asistieron al Hackathon #Somos600M, previamente traducidas al inglés para combinar términos equivalentes.

**AsistenciaRefugiados**: Asistente legal para personas en situación de refugio o asilo político (Eduardo Muñoz, 2024), **BERTIN-ClimID**: BERTIN-Base Climate-related text Identification (Gerardo Huerta, 2024), **ChaterapIA**: Conversaciones terapéuticas, **COSER Provincias**: Indentificación de provincias a partir de entrevistas orales de la España rural, **COSER Resúmenes**: Resumen de entrevistas orales de la España rural, **KUNTUR**: Asistencia legal en textos jurídicos de Perú (Quispe Castillo, 2024a), **LenguajeClaroQA**: Simplificación de lenguaje administrativo (de la Fuente et al., 2024), **LingComp_QA**: Corpus educativo de lingüística computacional en español (Zamora Rey et al., 2024), **NoticIA**: Resumen de Noticias Clickbait (García-Ferrero and Altuna, 2024), **RAC**: Reglamento Aeronáutico Colombiano (Sepulveda et al., 2024), **RecetasDeLaAbuel@**: Corpus de recetas de países hispanoamericanos (Morales-Garzón et al., 2024), **SMC**: Spanish Medical Corpus (Lopez Dionis, 2024), **SQLES**: Interactúa con una base de datos SQL en español, **ThinkParaguayo**: Conoce la cultura guaraní, **TraductorInclusivo**: Reescritura de textos en español utilizando lenguaje inclusivo (Gaia Quintana Fleitas, 2024), y **WikiHowES** (Quispe Castillo, 2024b).

## D Detalle de los corpus de evaluación

A continuación se enumeran los corpus de evaluación incluidos en la primera versión de la tabla de clasificación de modelos generativos. **AQuAS**: Abstractive Question-Answering in Spanish (Instituto de Ingeniería del Conocimiento, 2024a) y **RagQuAS**: Retrieval-Augmented-Generation and Question-Answering in Spanish (Instituto de Ingeniería del Conocimiento, 2024b), donados por el Instituto de Ingeniería del Conocimiento (IIC) de la Universidad Autónoma de Madrid (UAM); **SpaLawEx**: Exámenes de acceso a la abogacía (LenguajeNaturalAI, 2024c), **MedicalExpertES**: Diagnóstico y tratamiento de casos clínicos en español (LenguajeNaturalAI, 2024b) y **HumorQA**: Clasificación de bromas de humor blanco (LenguajeNaturalAI, 2024a), donados por LenguajeNaturalAI; **TELEIA**: Test de Español como Lengua Extranjera para Inteligencia Artificial, donado por el Grupo de Internet de Nueva Generación (GING) de la Universidad Politécnica de Madrid (UPM); **Meta4XNLI**: A Crosslingual Parallel Corpus for Metaphor Detection and Interpretation (Sanchez-Bayona and Agerri, 2024), **NoticIA**: Resumen de Noticias Clickbait (García-Ferrero and Altuna, 2024), **MedExpQA**: Multilingual Benchmarking of Large Language Models for Medical Question Answering (Alonso et al., 2024), **CasiMedicos-SQUAD**: Explanatory Argument Extraction of Correct Answers in Resident Medical Exams (Goenaga et al., 2023) and **CONAN-EUS**: Basque and Spanish Parallel Counter Narratives Dataset(Bengoetxea et al., 2024) donados por el Centro Vasco de Tecnología de la Lengua (HiTZ); **CatalanQA**: Extractive QA in Catalan, **TE_ca**: Textual Entailment in Catalan, **caBREU**: Article Summarization in Catalan, **XNLI_ca**: Cross-lingual Natural Language Inference in Catalan, **COPA_ca**: Choice Of Plausible Alternatives in Catalan, **PAWS_ca**: Paraphrase Adversaries from Word Scrambling in Catalan, **XQUAD_ca**: Cross-lingual Question Answering Dataset in Catalan and **WNLI_ca**: Winograd NLI dataset, donados por el Proyecto AINA, en concreto por el Barcelona Supercomputing Center (BSC) (Gonzalez-Aguirre et al., 2024).

| Corpus | Nº ejemplos | MB | Dominio | País(es) |
|---|---|---|---|---|
| AsistenciaRefugiados | 10707 | 20.7 | Legal | ES, MX, VE + |
| BERTIN-ClimID | 3680 | 1.63 | Sostenibilidad | PE, ES |
| ChaterapIA | 1000 | 2.30 | Psicología | ES |
| COSER Provincias | 1150 | 0.22 | Cultura rural | ES (oral) |
| COSER Resúmenes | 230 | 1.08 | Cultura rural | ES (oral) |
| KUNTUR | 2075 | 0.73 | Legal | PE |
| LenguajeClaroQA | 4094 | 1.72 | Legal admin. | ES |
| LingComp_QA | 1004 | 0.35 | Lingüística Comp. | ES |
| MedicalBilingual | 8138 | 12.8 | Clínico | Mix |
| NoticIA | 850 | 3.41 | Prensa | ES |
| RAC | 24478 | 1.84 | Legal | CO |
| RecetasDeLaAbuel@ | 20221 | 42.4 | Gastronomía | ES, MX, PE, AR+ |
| SpanishMedicalCorpus | 2136490 | 48.5 | Clínico | ES, CL |
| SQLES | 81 | 0.40 | Programación | - |
| Think Paraguayo | 1498 | 0.19 | Cultura guaraní | PY |
| TraductorInclusivo | 4196 | 0.40 | Misceláneo | ES, AR, MX, CL, CR + |
| WikiHowES | 113160 | 186 | Misceláneo | Mix |
| Total | 2,333,052 | 324.67 | - | - |

Table 4: Corpus de instrucciones creados por los equipos participantes en el Hackathon #Somos600M, disponibles en huggingface.co/somosnlp. Se excluyen las versiones de los corpus adaptadas a formatos para el entrenamiento de un modelo específico (e.g. Gemma). Los países están representados por su correspondiente código ISO 3166-1 alfa-2, el signo "+" indica que hay más países representados en el corpus en menor proporción.

| Corpus | Lengua | Dominio | Tarea |
|---|---|---|---|
| AQuAS | ES | Misceláneo | Extracción de información |
| RagQuAS | ES | Misceláneo | RAG y extracción de información |
| HellaSwag_es | ES | Misceláneo | Commonsense NLI |
| MMLU_es | ES | Misceláneo | Preguntas de opción múltiple |
| TELEIA | ES | Dominio idioma | Preguntas de opción múltiple |
| Meta4XNLI | ES | Dominio idioma | NLI |
| HumorQA | ES | Dominio idioma | Clasificación |
| ARC-C_es | ES | Ciencia | Preguntas de opción múltiple |
| NoticIA | ES | Prensa | Resumen de texto |
| SpaLawEx | ES | Legal | Preguntas de opción múltiple |
| MedicalExpertES | ES | Clínico | Preguntas de respuesta abierta |
| MedExpQA | ES | Clínico | Preguntas de opción múltiple |
| CasiMedicos-SQUAD | ES | Clínico | Extracción de información |
| CONAN-EUS | ES, EU | Discurso de odio | Generación de contranarrativas |
| CatalanQA | CA | Misceláneo | Extracción de información |
| TE_ca | CA | Misceláneo | NLI |
| XNLI_ca | CA | Misceláneo | NLI |
| WNLI_ca | CA | Misceláneo | NLI |
| COPA_ca | CA | Misceláneo | Razonamiento lógico |
| PAWS_ca | CA | Misceláneo | Paraphrasing |
| XQUAD_ca | CA | Misceláneo | Extracción de información |
| caBREU | CA | Prensa | Resumen |

Table 5: Corpus que constituyen la primera versión de la tabla de clasificación de modelos de lenguaje generativos que incluye tareas en español (ES), catalán (CA) y euskera (EU) y evalúa la capacidad de extracción de información, cultura general, conocimiento en los dominios legal y clínico, razonamiento lógico y dominio del idioma.