

Iker García Ferrero

SENIOR LLM RESEARCHER · FOUNDATION MODELS · HPC & DISTRIBUTED TRAINING

✉ igarciaf896@gmail.com | 🏠 ikergarcia1996.github.io | 📄 github.com/ikergarcia1996

📄 Iker García-Ferrero | 🏠 Iker García-Ferrero

Summary

PhD in Natural Language Processing from the University of the Basque Country UPV/EHU, IXA Group, and HiTZ Basque Center for Language Technologies. I specialize in training, compressing, and optimizing large-scale foundation models for both language and vision.

I have contributed to the development of several state-of-the-art models: Hypernova-60B, a compressed LLM optimized for agentic and tool-use workflows at Multiverse Computing; FLUX.1 Krea, a diffusion image model trained in collaboration with Black Forest Labs; Latxa, an instruction-tuned LLM for Basque; GoLLIE, a 34B-parameter model achieving state-of-the-art zero-shot Information Extraction; and MedMT5, the first open-source multilingual text-to-text model for the medical domain.

Strong experience in performance optimization and HPC across academic and industry environments. At HiTZ, I trained Latxa on the Leonardo Supercomputer using up to 512 GPUs across 128 nodes. At Krea, I worked on optimizing fine-tuning pipelines for video and image diffusion models using distributed training techniques such as FSDP2 on Krea's internal GPU cluster. At Multiverse Computing, I optimized inference and fine-tuning pipelines on the company's internal GPU cluster, achieving 3x faster logit precomputation and 5x faster KL-divergence training in internal pipelines relative to their prior implementations.

Experience

Multiverse Computing

Donostia - San Sebastián, Spain

August 2025 – Present

Senior LLM Researcher (August 2026 – Present); LLM Researcher (August 2025 – July 2026)

- Developing compressed LLMs that make AI systems faster, cheaper, and more energy-efficient, including Hypernova-60B, a 60B-parameter model optimized for agentic workflows and tool use.
- Developed a steering-vector method for editing LLM knowledge and behavior, including adding or removing censorship around sensitive topics while preserving model performance.
- Optimized internal inference and fine-tuning codebases, achieving 3x faster logit precomputation and 5x faster KL-divergence training in internal pipelines relative to their prior implementations.
- Built large-scale synthetic data generation pipelines for LLM training and evaluation.

Krea.ai

San Francisco, USA

January 2025 – July 2025

Machine Learning Researcher

- Contributed to the development of the FLUX.1 Krea model, a diffusion-based image model trained in collaboration with Black Forest Labs.
- Developed a personalized image recommendation engine using user interaction data and deep learning techniques.
- Designed and trained SigLIP-style embedding models for artistic style-based image retrieval.
- Improved fine-tuning codebases for video (Wan 2.1) and image (Flux) models, focusing on training performance and distributed training techniques such as FSDP2.
- Implemented and deployed generative services for 3D models and textures, powering features used by thousands of users on the Krea platform.

University of Pennsylvania

Philadelphia, USA

October 2022 – January 2023

Visiting PhD Student

- Visiting Associate at the School of Engineering and Applied Science, Department of Computer and Information Science, Cognitive Computation Group.
- Supervised by Dan Roth.
- Implemented a Named Entity Recognition system that leverages external knowledge bases (Wikipedia and Wikidata) to improve performance in low-resource languages and multilingual settings.
- Implemented T-Projection, a state-of-the-art annotation projection method that replaces noisy word alignments with pretrained text-to-text language models for candidate generation and machine translation systems for candidate ranking.

Amazon

Barcelona, Spain

November 2019 – January 2020

Applied Scientist Intern

- Conducted research on multilingual neural models for information extraction in languages with no labeled data.
- Developed a system that uses English labeled data to automatically generate training data in other languages by combining machine translation with word-alignment-based annotation projection.

Education

Doctor of Philosophy (PhD), Natural Language Processing

Donostia - San Sebastián, Spain

University of the Basque Country UPV/EHU

January 2020 – February 2025

- Title: Acquisition and exploitation of Cross-Lingual knowledge
- Advisors: German Rigau, Rodrigo Agerri

Master's Degree, Natural Language Processing

Donostia - San Sebastián, Spain

University of the Basque Country UPV/EHU

September 2018 – September 2019

- Master thesis: A Common Semantic Space for Monolingual and Cross-Lingual Meta-Embeddings
- Advisors: German Rigau

Bachelor's Degree, Computer Science

Donostia - San Sebastián, Spain

University of the Basque Country UPV/EHU

2014 – 2018

- Bachelor thesis: Study of Word Embeddings and methods of generating Meta Embeddings
- Advisors: German Rigau

Safety for Whom? Boundary-Aware Self-Distillation for Controlled LLM Safety Refusal

arXiv · 2026 (preprint)

Alejo López-Ávila, [Iker García-Ferrero](#), Jezabel Garcia, Antonio Tiene, Román Orús.

Paper / publication

2026

Quantization-Aware Healing: A Practical Recipe for Recovering Compressed, 4-Bit LLMs

arXiv · 2026 (preprint)

Bakbergen Ryskulov, [Iker García-Ferrero](#), David Montero, David Jansen, Ali Hashemi, Jezabel R. Garcia, Antonio Tiene, Román Orús.

Paper / publication

2026

Efficient Knowledge Distillation for LLMs: Offline Top-K Logits and a Fused Chunked KL Loss

arXiv · 2026 (preprint)

Bakbergen Ryskulov, [Iker García-Ferrero](#), David Montero, David Jansen, Ali Hashemi, Jezabel R. Garcia, Antonio Tiene, Román Orús.

Paper / publication

2026

Refusal Steering: Fine-grained Control over LLM Refusal Behaviour for Sensitive Topics

LREC · 2026

[Iker García-Ferrero](#), David Montero, Roman Orus.

Paper / publication

2026

FactORes: Fact-checking with an Evidence-based Open Resource in Spanish

LREC · 2026

Nagore Bravo, Jaione Bengoetxea, [Iker García-Ferrero](#), Alba Bonet Jover, Estela Saquete, Rodrigo Agerri.

2026

Discriminative Benchmarking of Spanish Language Models: Findings from the ODESIA Challenge 2024

PLN · 2026

Alejandro Benito Santos, Roser Morante, Adrián Ghajari Espinosa, [Iker García-Ferrero](#), Robiert Sepúlveda Torres, German Rigau, Rodrigo Agerri, et al.

Paper / publication

2026

Releasing Open Weights for FLUX.1 Krea

Krea Blog · 2025

Sangwu Lee, Titus Ebbecke, Erwann Millon, Will Beddow, Le Zhuo, [Iker García-Ferrero](#), Liam Esparraguera, Mihai Petrescu, Gian Saß, Gabriel Menezes, Victor Perez.

Paper / publication

2025

Instructing LLMs for Low-Resource Languages: A Systematic Study for Basque

EMNLP · 2025

Oscar Sainz, Naiara Perez, Julen Etzaniz, Joseba Fernandez de Landa, Itziar Aldabe, [Iker García-Ferrero](#), Aimar Zabala, Ekhi Azurmendi, German Rigau, Eneko Agirre, Mikel Artetxe, Aitor Soroa.

Paper / publication

2025

GuideX: Guided Synthetic Data Generation for Zero-Shot IE

Findings of ACL · 2025

Neil De La Fuente, Oscar Sainz, [Iker García-Ferrero](#), Eneko Agirre.

Paper / publication

2025

Cross-Lingual Transfer for Low-Resource Natural Language Processing

Doctoral thesis · 2025

[Iker García-Ferrero](#)

Paper / publication

2025

Data Contamination Report from the 2024 CONDA Shared Task

CONDA Workshop · 2024

Oscar Sainz, [Iker García-Ferrero](#), Alon Jacovi, Jon Ander Campos, Yanai Elazar, Eneko Agirre, Yoav Goldberg.

Paper / publication

2024

HiTZ@Disargue: Few-shot Learning and Argumentation to Detect and Fight Misinformation in Social Media

SEPLN-CEDI-PD · 2024

Rodrigo Agerri, Jeremy Barnes, Jaione Bengoetxea, Blanca Calvo Figueras, Joseba Fernandez de Landa, [Iker García-Ferrero](#), Olya Toporkov, Irune Zubiaga.

Paper / publication

2024

Uncovering Social Changes of the Basque Twitter Community during COVID-19

SIGUL · 2024

Joseba Fernandez de Landa, [Iker García-Ferrero](#), Ander Salaberria, Jon Ander Campos.

Paper / publication

2024

NoticiaA: A Clickbait Article Summarization Dataset in Spanish

PLN · 2024

[Iker García-Ferrero](#), Begoña Altuna.

Paper / publication

2024

Prepare to be Amazed: NoticiaA, the Spanish Clickbait Dataset Transforming the Way We Read News

CLARIN Annual Conference · 2024

Begoña Altuna, [Iker García-Ferrero](#).

Paper / publication

2024

MedicalMT5: An Open-Source Multilingual Text-to-Text LLM for the Medical Domain

LREC-COLING · 2024

[Iker García-Ferrero](#), Rodrigo Agerri, Aitziber Atutxa Salazar, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johana Ramirez-Romero, German Rigau, Jose Maria Villa-Gonzalez, Serena Villata, Andrea Zaninello.

Paper / publication

2024

GoLLIE: Annotation Guidelines improve Zero-Shot IE

ICLR · 2024

Oscar Sainz, [Iker García-Ferrero](#), Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, Eneko Agirre.

Paper / publication

2024

This is not a Dataset: A Large Negation Benchmark to Challenge LLMs

EMNLP · 2023

[Iker García-Ferrero](#), Begoña Altuna, Javier Alvez, Itziar Gonzalez-Dios, German Rigau.

Paper / publication

2023

T-Projection: High Quality Annotation Projection for Sequence Labeling Tasks

Findings of EMNLP · 2023

[Iker García-Ferrero](#), Rodrigo Agerri, German Rigau.

Paper / publication

2023

NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination

Findings of EMNLP · 2023

Oscar Sainz, Jon Ander Campos, [Iker García-Ferrero](#), Julen Etxaniz, Oier Lopez de Lacalle, Eneko Agirre.

Paper / publication

2023

IXA/Cogcomp at SemEval-2023 Task 2: Context-enriched Multilingual NER using KBs

SemEval · 2023

[Iker García-Ferrero](#), Jon Ander Campos, Oscar Sainz, Ander Salaberria, Dan Roth.

Paper / publication

2023

HiTZ@ Antidote: Explainable AI for Digital Medicine

SEPLN · 2023

Rodrigo Agerri, Iñigo Alonso, Aitziber Atutxa, Ander Berrondo, Ainara Estarrona, [Iker García-Ferrero](#), Iakes Goenaga, Koldo Gojenola, Maite Oronoz, Igor Perez-Tejedor, German Rigau, Anar Yeginbergenova.

Paper / publication

2023

Acquisition and Exploitation of Cross-Lingual Knowledge

CEUR Workshop · 2022

[Iker García-Ferrero](#).

Paper / publication

2022

Model Transfer or Data Transfer? Cross-Lingual Sequence Labelling in Zero-Resource Settings

Findings of EMNLP · 2022

[Iker García-Ferrero](#), Rodrigo Agerri, German Rigau.

Paper / publication

2022

Benchmarking Meta-embeddings: What Works and What Does Not

Findings of EMNLP · 2021

[Iker García-Ferrero](#), Rodrigo Agerri, German Rigau.

Paper / publication

2021

A Common Semantic Space for Monolingual and Cross-Lingual Meta-Embeddings

arXiv · 2020

[Iker García-Ferrero](#), Rodrigo Agerri, German Rigau.

Paper / publication

2020

Estudio de Word Embeddings y métodos de generación de Meta Embeddings

Bachelor's thesis · 2018

[Iker García-Ferrero](#).

Paper / publication

2018

Itzulpen Automatikoko Sistemen Analisia: Genero Alborapenaren Kasua

IkerGazte · 2021

Ander Salaberria, Jon Ander Campos, [Iker García-Ferrero](#), Joseba Fernandez de Landa.

Paper / publication

2021

Twitterreko Euskal Komunitatearen Eduki Azterketa Pandemia Garaian

IkerGazte · 2021

Joseba Fernandez de Landa, [Iker García-Ferrero](#), Ander Salaberria, Jon Ander Campos.

Paper / publication

2021

Other Contributions

Co-organizer of the 1st Workshop on Data Contamination at ACL 2024

The 62nd Annual Meeting of the Association for Computational Linguistics

Oscar Sainz, [Iker García-Ferrero](#), Eneko Agirre, Jon Ander Campos, Alon Jacovi, Yoav Goldberg, Yanai Elazar

<https://conda-workshop.github.io/>

2024

Grants and Awards

Exponential Fellow

The Exponential Fellowship: A fellowship program that aims to find brightest builders in Spain to send on a life-changing experience to the US. <https://www.goexponential.org/>

2025

Best Project Award in the SomosNLP 2024 Hackathon: Clickbait Article Summarization in Spanish.

SomosNLP: <https://sornoslp.org/hackathon>

2024

Prize for the most relevant research for the development of the Basque Country

The paper “Twitterreko Euskal Komunitatearen Eduki Azterketa Pandemia Garaian” was awarded the prize for the most relevant research for the development of the Basque Country at IkerGazte 2021 by Udabiltza

2021

Basque Government PhD grant (4 years)

Basque Government

2019

Grants for university master’s degrees at the UPV/EHU

University of the Basque Country UPV/EHU, 3000€

2018

Languages

Spanish

Native Speaker

English

Full Professional Proficiency

Basque

Basic Competence

Academic Activity

Keynote talk at the AI4HF project consortium meeting

Medical mT5: LLMs for the medical domain

Universitat de Barcelona - Spain

2024

Instructor in the Introduction to Language Technology Applications course: (2nd edition, 2022)

<http://www.ix.eus/iltapp/>

Donostia - San Sebastián, Spain

2022

Instructor in the Introduction to Language Technology Applications course: (1st edition, 2021)

<http://www.ix.eus/iltapp/>

Donostia - San Sebastián, Spain

2021

Talk at LITHME Training school 2021: Introducing HiTZ Group

<https://lithme.eu/training-school/>

Ávila, Spain

2021

Skills

Programming	Python, Bash, Git, Linux, Docker.
Deep Learning Frameworks	PyTorch, Nvidia NeMo, TorchVision, Hugging Face (Transformers, Diffusers, Datasets, Accelerate, PEFT, TRL), PyTorch Lightning.
LLMs & Foundation Models	Pre-training, instruction tuning, RLHF/DPO, knowledge distillation, quantization, pruning, LoRA/QLoRA, model merging, activation/steering vectors, RAG, agentic workflows and tool use, evaluation & benchmarking.
Generative Vision	Diffusion models (FLUX, Stable Diffusion, Wan), image & video generation, fine-tuning, CLIP/SigLIP embedding models, image retrieval.
Distributed Training & HPC	Multi-GPU & multi-node training, FSDP/FSDP2, DeepSpeed, Megatron-LM, mixed precision (FP16/BF16/FP8), Slurm, supercomputing (CINECA Leonardo, up to 512 GPUs), cloud GPU clusters.
Inference & Serving	vLLM, SGLang, llama.cpp, Ray, speculative decoding.
MLOps & Tooling	Weights & Biases, TensorBoard, MLflow, Hugging Face Hub, experiment tracking, synthetic data pipelines, evaluation harnesses.