Santiago Martínez Novoa

Machine Learning Researcher | Multilingual NLP & Trustworthy LLM Evaluation

s.martinezn@uniandes.edu.co | LinkedIn | GitHub | Website

Research Interests

My research focuses on making language models reliable for the people and languages they serve worst. I study multilingual NLP and cross-lingual transfer for low-resource languages, and how to evaluate models where labeled data is scarce: deciding when a model is good enough without labels, LLM-as-judge and rubric-based evaluation, and culturally grounded benchmarks. I apply this in high-stakes settings such as health and education, where I also build production machine learning systems for academic advising and student success.

Education

Universidad de los Andes — Bogotá, Colombia
M.Sc. Information Engineering — *2025 – 2026 (To be conferred)*
GPA: 4.78/5.00

Universidad de los Andes — Bogotá, Colombia
B.Sc. Systems and Computing Engineering — *2021 – 2024*
GPA: 4.58/5.00, ***Cum Laude***
Minor in French

Publications

Nguyen, Q. P., Gaschi, F., Anugraha, D., **Martínez Novoa, S.**, Lee, E.-S. A. "Online Language Adaptive Sampling for Better Distributed Cross-lingual Gains." *Findings of EMNLP 2026, Budapest, Hungary, pp. 36873–36891* (2026). DOI: 10.18653/2026.findings-emnlp.1724.

Prieto, J. C., **Martínez Novoa, S.**, Gómez Mesa, L., Manrique, R. "SciBETO: A Domain-Specific RoBERTa Model Trained on Spanish Scientific Texts." *EACL 2027* (2026, under review).

Martínez Novoa, S., Melo Barrero, F., Ruiz Giraldo, M., Velásquez Marín, S., Carvajal Chaves, N., Hernández Hoyos, M., Reyes, J. P. "Centralizing the Student Lens: Design and Evaluation of a Learning Analytics Dashboard for Academic Advising in Higher Education." *WEEF (IFEES & GEDC) 2026, Cartagena de Indias, Colombia* (2026, to appear).

Martínez Novoa, S., Gómez Mesa, L., Prieto, J., Manrique, R. "MeSHClass-ES and AnatEM-ES: Open Resources for Spanish Biomedical NLP." *BioNLP 2026 (ACL Workshop), San Diego, California, pp. 617–629* (2026). DOI: 10.18653/v1/2026.bionlp-1.49.

Soto Parada, E. G., **Martínez Novoa, S.**, Klopstock Triana, N., Fuquen Florez, D. A., Peña, F. C., Hernández Hoyos, M., Reyes, J. P. "No Student Left in Limbo: Multi-Level Clustering for Success Prediction in a Latin American University." *AIED 2026, WideAIED Track, Communications in Computer and Information Science, Springer, pp. 63–77* (2026). DOI: 10.1007/978-3-032-29791-4_5.

Martínez Novoa, S., Ibáñez, M. C., Gómez Mesa, L., Kramer, J. "Classification of Chest XRay Diseases through image processing and analysis techniques." (2026). arXiv:2601.10913.

Martínez Novoa, S., Rozo Fajardo, N., González Vargas, D. A., Bedoya Figueroa, N. "Efficient Multilingual Dialogue Processing via Translation Pipelines and Distilled Language Models." (2026). arXiv:2601.09059.

Martínez Novoa, S. "Detection and Segmentation of Malignant Melanoma Regions in Dermoscopic Images Using Machine Learning." *Undergraduate thesis, Universidad de los Andes* (2024). Available at: Uniandes

Repository.

Research Experience

Lee Language Lab, Ontario Tech University May 2026 – Present
Mitacs Globalink Research Intern - Collaborating Researcher *Oshawa, Canada*

- Co-authored a study on online adaptive language sampling for better distributed cross-lingual transfer, accepted to Findings of EMNLP 2026.
- Lead its direct continuation, studying when to stop cross-lingual realignment: more realignment is not always better for transfer, and low-resource languages offer no labeled data to tell when to stop; paper under review at ACL Rolling Review (ARR).
- Co-develop M2-ALIGN, studying where multilingual alignment should live in vision-language models: how to extend visual grounding to 11 languages, including low-resource ones, without eroding the multilingual reasoning the base models already have; ongoing work on culturally grounded evaluation (CVQA).
- Contributed to a PRISMA systematic literature review of multilingual, multimodal, and multicultural conversational agents.

FLAG Research Lab - A.I. and Machine Learning, Universidad de los Andes Jan 2025 – Present
Evaluation and Enhancement of Natural Language Models in Spanish *Bogotá, Colombia*

- Research how to improve natural language models for Spanish medical contexts, focusing on text classification and named entity extraction tasks under Professor Ruben Manrique Ph.D.
- Develop and train specialized models from scratch for specific Spanish medical domain tasks.
- Conduct comparative evaluations between highly-trained models and large language models (LLMs) using prompting techniques such as zero-shot and few-shot learning.

Augmented Artificial Intelligence Initiative, Universidad de los Andes Jan 2025 – Jun 2025
Generative Artificial Intelligence for Research - Research Assistant *Bogotá, Colombia*

- Designed an LLM-as-judge evaluation architecture that scores response quality and grades student work against rubrics, using Prometheus and BERTScore.
- Built locally deployed RAG systems and custom AI agents that answer domain-specific research questions.

College of Computing and Information Science, Cornell University Jun 2024 – Aug 2024
Summer Research Intern *Ithaca, United States*

- Designed a Data Discovery architecture to explore and modify datasets in various formats under Professor Sainyam Galhotra Ph.D.
- Developed the first version of the architecture including the upload and scoring through semantic search of multiple databases.

Work Experience

EdTech Node, Universidad de los Andes Mar 2026 – Present
Data Scientist *Bogotá, Colombia*

- Lead the unit's adoption of institutional data governance and quality policies: establishing data ownership and origin with source units, building column-level definitions, and defining quality rules to audit, correct, and monitor its data products.
- Manage the unit's data end to end, from source systems to published products: the institutional information behind "No Estás Solx," the academic advising platform, and university-wide data products such as academic percentiles and student dropout indicators; the platform team received the 2026 Dejar Huella Award.
- Lead the team's research line, which has produced two peer-reviewed publications (AIED 2026 and WEEF 2026) and a journal article in preparation.

- Design, validate, and deploy machine learning models to predict student dropout and academic risk.

Vice Dean's Office of Student Affairs, Universidad de los Andes Jan 2025 – Mar 2026
Data Scientist - Graduate Research Assistant *Bogotá, Colombia*

- Enabled the university to save approximately $100K USD annually by replacing an external academic counseling platform.
- Co-implemented a risk detection system that alerts stakeholders about over 1000 critical student cases.
- Designed and implemented the unit's complete data architecture in Databricks, following a medallion architecture and fully automating data processing and institutional information storage pipelines.
- Optimized ETL pipeline performance by 87% (reduced processing time from 15 to 2 minutes) by migrating from Pandas to Polars for data retrieval and storage from Azure Blob Storage.
- Processed academic data and developed Power BI dashboards under the guidance of Professor Marcela Hernández Hoyos, Ph.D. and Professor Juan Pablo Reyes Gómez, Ph.D., delivering actionable insights on student performance and institutional metrics.

Precia PPV S.A Aug 2024 – Dec 2024
Machine Learning Intern *Bogotá, Colombia*

- Co-developed and trained a Multi-layer Perceptron (MLP) that reduced the calculation time of a critical financial metric by 70%, giving clients significantly faster access to preliminary estimates.
- Presented our work at the Analytics Forum 2025 on model optimization and infrastructure architecture, including the use of MLPs, AWS pipelines, and data transformation strategies.
- Designed and implemented end-to-end AWS infrastructure, including robust ETL pipelines integrating multiple data sources (databases, FSx servers), and automated model deployment using AWS Glue and Lambda functions.

Vice Dean's Office of Student Affairs, Universidad de los Andes Aug 2024 – Dec 2024
Data Scientist - Undergraduate Research Assistant *Bogotá, Colombia*

- Co-implemented the ETL process that moves institutional data from Azure Blob Storage into a relational database, with integrity tests validating the multidimensional model against the original data.

Digital Transformation, Universidad de los Andes Sep 2023 – May 2024
Backend Developer *Bogotá, Colombia*

- Developed OpenAI-powered backend services with Python and Azure Functions, documenting the project's infrastructure so the team could build on it.

Research Center of the Faculty of Engineering, Universidad de los Andes Aug 2023 – Dec 2023
Frontend Developer *Bogotá, Colombia*

- Co-developed the Angular frontend of a faculty research project, including dynamic role-based access control.

Digital Transformation, Universidad de los Andes Jun 2023 – Aug 2023
Cross-Cultural Liaison and Software Developer *Bogotá, Colombia*

- Interpreted between directors and 5 exchange students while helping integrate OpenAI and Azure Functions into an interactive chatbot.

Teaching Experience

Universidad de los Andes Jun 2025 – Oct 2026
Graduate Teaching Assistant (LLM-Based Applications & Agent Development) *Bogotá, Colombia*

- Graduate TA across three semesters for two sequential 8-week master's courses, 'LLM-Based Applications' (RAG systems) and 'Agent Development', supporting over 200 master's students.
- Created the tutorials, projects, and course materials for the RAG systems and agent development modules,

built on LangChain, LangGraph, LangSmith, and RAGAS.

Systems and Computing Engineering Department, Universidad de los Andes Jan 2024 – Dec 2024
Undergraduate Research Assistant (CupiTaller Programming Support Center) *Bogotá, Colombia*

- As part of the 7-person team coordinating CupiTaller, the university's programming support center, supervised over 110 tutors and TAs and designed the projects for the introductory course of over 1,100 students per semester.

Systems and Computing Engineering Department, Universidad de los Andes Aug 2021 – Aug 2023
Undergraduate Teaching Assistant - Programming Tutor (Programming Support Center) *Bogotá, Colombia*

- Tutored over 100 students in Python programming fundamentals, 5 hours a week for two years at the university's programming support center.

Universidad de los Andes Jan 2022 – Dec 2022
Undergraduate Teaching Assistant (Introduction to Programming) *Bogotá, Colombia*

- Assisted the professor in programming labs of about 40 students per semester, grading the course projects.

Universidad de los Andes Jan 2023 – Jun 2023
Undergraduate Teaching Assistant (Fundamentals of Technological Infrastructure) *Bogotá, Colombia*

- Graded exams and coursework while guiding about 40 students through the fundamentals of technological infrastructure.

Honors and Awards

Dejar Huella Award Aug 2026
Universidad de los Andes

- Universidad de los Andes recognition for team impact on the university community. Awarded for "No Estás Solx," a student-built peer-support platform serving 56,000+ active students with ~1,000 daily connections to institutional support.

Mitacs Globalink Research Internship May 2026
Ontario Tech University

- Competitive international Mitacs award funding a 12-week research internship at the Lee Language Lab, Ontario Tech University, under the supervision of Professor Annie En-Shiun Lee, Ph.D.

Cum Laude Apr 2025
Universidad de los Andes

- Distinction awarded to graduates with a GPA in the top 3% of their faculty at Universidad de los Andes over the last five years.

Summer Undergraduate Research Fellowship (SURF) Jun 2024
Cornell University

- Financial support awarded to participate in the Summer Undergraduate Research Fellowship (SURF) program at Cornell University.

Talks and Presentations

Accelerating Credit Value Adjustment with Machine Learning. *Analytics Forum 2025 - Universidad de los Andes*, Bogotá, Colombia. May 2025. [Web page]

Skills

Languages

- Spanish (Native)
- English (Advanced)
- French (Conversational)

Technical Skills

- *Programming Languages*: Python, SQL, JavaScript, TypeScript
- *Machine Learning & AI*: Scikit-learn, Keras, PyTorch, TensorFlow
- *LLM & Natural Language Processing*: Transformers, BERT, LangChain, LangGraph, LangSmith, Ollama
- *Data Processing*: Pandas, Polars, Databricks
- *Cloud & DevOps*: AWS, Azure Functions, Azure Blob Storage
- *Databases*: PostgreSQL, MongoDB, ChromaDB, Pinecone