# Pierre Colombo

(+33) 650118518 • colombo.pierre@gmail.com • Websites:

Experienced researcher and entrepreneur in NLP/LLMs. Currently **Chief Science Officer at Equall** & **Former Associate Professor at Université Paris Saclay (i.e., Maître de Conférences)**, with a proven track record in leading research teams and driving innovation in academia and startups.

## Experience

✓ **Co-founder - Chief Science Officer** — **Paris, France / NYC, USA**
*Equall* — *July 2023 – Present*

Co-founded Equall, initially focused on building legal LLM (e.g SaulLM NeurIPS 2024 - Le Monde du Droit).

- **LLM Training:** Built **from scratch** large-scale data pipelines; curated and processed 500B+ tokens of legal data. Adapted LLMs for the legal domain on a 512-GPU setup, including large-scale continuous pretraining, post-training.
- **Product Development:** Developed and deployed legal agent and chat systems **from scratch**. Designed multi-step legal workflows with a focus on explainability, citation accuracy, and performance.
- **Cross-functional Execution:** Developed core backend patterns (e.g., Celery, RabbitMQ, GraphQL); collaborated closely with frontend teams and legal experts to align product behavior and define APIs.
- **Strategic Contributions:** Crafted technical narratives, participated in investor meetings & technical diligence.

✓ **Associate Professor – Maître de Conférence** — **Paris, France**
*Université Paris-Saclay (CentraleSupélec)* — *Sept 2022 – Sept 2025*

Founded and led the first NLP group at CS, with a focus on large-scale, open-source LLM research and training.

- **Leading large scale research (20+ people):** Lead cutting-edge research in training and evaluating *open-source LLMs*, with an emphasis on multilinguality, safety, and real-world deployment.
- **Primary advisor to 6 PhD students** (Maxime Darrin '21–, Manuel Faysse '22–, Nicolas Boizard '23–, Hippolyte Gisserot-Boukhlef '23–, Duarte Alves '23–, N Guerreiro (deF '25)) and 3 MSc students.
- **Obtained 5 major compute grants** (Jean Zay, MareNostrum, Adastra) - $6.1M in equivalent compute resources
- **Raised $1M in industrial/PhD co-funding** from partners (e.g Illuin Technology, Diabolocom, Artefact, DataIA).
- **Co-organizer of ACL 2024** (Book Chair) — NLP top-tier conferences.
- **Delivered over 700 hours of classes** across CentraleSupélec, Telécom Paris, EPFL, and IP Paris.
- **Invited Professor at ILLS** (MILA, Université de Montréal / McGill - July - November 2022).

✓ **Postdoctoral Researcher** — **Paris, France**
*CNRS-L2S (CentraleSupélec)* — *Nov. 2022 – July 2023*

- Worked with two PhD students on adversarial attack and OOD detection for vision with publications in TMLR 2023.
- OOD detection for text (published at NeurIPS 2022, ACL 2022) and new benchmarking techniques (NeurIPS 2022).

✓ **PhD Student** — **Paris, France**
*IBM* — *Nov. 2018 – Nov. 2021*

- Published first-author papers in top AI/NLP conferences, including ACL, EMNLP, and NeurIPS.
- Awarded **Best Student Paper** at AAAI 2022.
- Conducted research on disentangled representations, text generation, and multimodal sentiment analysis.

✓ **Research Internships on NLP/Biology** — **USA - Switzerland**
*Various* — *2017 – 2018*

- *Disney Research* (2018, 6 months) - Lab Associate, LA, USA NAACL 2018 and patent publication.
- *IBM Research* (2017/2018, 6 months) - Research Intern, Zürich, Switzerland - Computational Biology.
- *Swisscom AI Lab* (2017, 6 months) - Research Student, Lausanne, Switzerland. Retrieval Augmented Chatbot.

## Education

✓ **PhD in Computer Science** — **Telécom Paris, Institut Polytechnique de Paris**
*Paris, Switzerland* — *2018 – 2021*

✓ **MSC in Computer Science** — **EPFL**
*Lausanne, Switzerland* — *2016 – 2018*

✓ **MSC in Applied Mathematics** — **CentraleSupelec**
*Paris, France* — *2014–2018*

# Research

## Awards

★ **Best Student Paper Award (Acceptance Rate: 15% | #1 out of 9020 papers.)** **AAAI 2022**
*Award for: InfoLM: A New Metric to Evaluate Summarization & Data2Text Generation.* *2022*

## Spotlight Projects

- **Large Language Model Pretraining**
  - ➢ **EuroLLM** (9B & 1.7B): LLMs for EU languages (competitive with LLAMA3) - **Project Co-Lead**
    - *Individual Contributor*: Data filtering, pretraining codebase, evaluation frameworks
    - *Scale*: 3M H100 hours (~$3M equivalent), 15+ researchers, 5,000B tokens.
    - *Impact*: Competitive multilingual and multimodal LLM offering for all European languages; featured in OSOR and Slator, Research paper.
  - ➢ **CroissantLLM**: Bilingual French-English SLM - **Project Lead** - **Individual Contributor**
    - *Individual Contributor*: Data filtering, post-training, evaluation, pretraining architecture
    - *Scale*: 300k A100 hours (~$300k equivalent), 10+ researchers, 3,000B tokens.
    - *Impact*: Featured in Usine Digitale, TMLR 2025 paper, model with 50k+ downloads
  - ➢ **EuroBERT**: General-purpose encoder for European languages - **Project Lead**
    - *Individual Contributor*: Data filtering pipeline design and implementation
    - *Scale*: 500k MI300A and MI250 hours (~$500k equivalent), 10+ researchers, 5,000B tokens.
    - *Output*: COLM 2025 paper, model with 30k+ downloads
- **Large Language Model Domain Adaptation**
  - ➢ **SaulLM Series**: First LLM family designed explicitly for legal text comprehension and generation, with models ranging from 7B to 141B parameters - **Main Individual contributor & Project Lead**
    - *Individual Contributor*: Legal data filtering, post-training, evaluation, pretraining codebase
    - *Compute*: 300k MI300 hours (~$300k equivalent), 8x22B MoE architecture, 500B legal tokens
    - *Impact*: State-of-the-art proficiency in understanding and processing legal documents; featured in Le Monde du Droit, NeurIPS 2024 paper, models with 30k+ downloads
  - ➢ **TowerLLM**: State-of-the-art translation language model - **Project Co-Lead**
    - *Impact*: First AI model to surpass GPT-4o in machine translation tasks; featured in AiThority
    - *Output*: COLM 2024 paper (Oral), model release (10k downloads).
- **Multimodal Learning**
  - ➢ **ColPali**: Efficient document retrieval with vision language models - **Project Lead**
    - *Innovation*: Novel multimodal approach for document understanding and retrieval
    - *Output*: ICLR 2025 paper, widely adopted model by **NVIDIA, COHERE, AMAZON**.

## Patent

- A Modi, M Kapadia, Douglas A Fidaleo, J Kennedy, W Witon and **P Colombo**, US Patent 16,226,166, *A framework for Affective Conversational System*

## Selected Publications

*Underlined authors are my PhD students. If not the first author, I was the main advisor for theses projects.*

- **P Colombo**, C Clavel and P Piantanida. InfoLM: *A New Metric to Evaluate Summarization & Data2Text Generation.* Oral AAAI 2022 Granted the Outstanding Student Paper Award
- **P Colombo**, N Noiry, E Irurozki, S Clemencon. *What are the best Systems? New perspectives on NLP benchmarking* NeurIPS 2022
- N Guerreiro, D Alves, J Waldendorf, B Haddow, A Birch, **P Colombo**, A Martins. *Hallucinations in Large Multilingual Translation Models* TACL 2023
- N Guerreiro, R Rei, D van Stigt, L. Coheur, **P Colombo**, A Martins *xcomet: Transparent Machine Translation Evaluation through Fine-grained Error Detection* Oral TACL 2024

- **P Colombo**, TP Pires, M Boudiaf, D Culver, R Melo, C Corro, AFT Martins, et al. *SaulLM-54B & SaulLM-141B: Scaling Up Domain Adaptation for the Legal Domain* NeurIPS 2024
- D Alves, J Pombal, N Guerreiro, P Martins, J Alves, A Farajian, **P Colombo** et al. *Tower: An Open Multilingual Large Language Model for Translation-Related Tasks* Oral COLM 2024
- M Faysse, H Sibille, T Wu, B Omrani, G Viaud, C Hudelot, **P Colombo** *Colpali: Efficient document retrieval with vision language models.* ICLR 2025
- P Martins, P Fernandes, J Alves, N Guerreiro, R Rei, D Alves, J Pombal, A Farajian, M Faysse, M Klimaszewski, **P Colombo**, et al. *Eurollm: Multilingual Language Models for Europe*
- N Boizard, K El-Haddad, C Hudelot, **P Colombo** *Towards Cross-Tokenizer Distillation: the Universal Logit Distillation Loss for LLMs*, TMLR 2025
- N Boizard, H Gisserot-Boukhlef, D Alves, et al. M Faysse, M Peyrard, N M Guerreiro, P Fernandes, R Rei, **P Colombo**, *EuroBERT: Scaling Multilingual Encoders for European Languages* COLM 2025
- M Faysse, P Fernandes, N Guerreiro, A Loison, D Alves, C Corro, et al., **P Colombo**. *CroissantLLM: A Truly Bilingual French-English Language Model* TMLR 2025

## Technical Skills

- **Programming & Frameworks:** Python, PyTorch, NumPy, Pandas, Matplotlib, spaCy, NLTK, Prodigy
- **Cloud & Infrastructure:** AWS, Google Cloud Platform, Microsoft Azure, Docker, Redis, RabbitMQ
- **Large scale training (over 256 GPUs)** on NVIDIA (H100, A100, V100, K80) and AMD GPUs (e.g., MI250) and APU (e.g., MI300A).
- **Languages** English (fluent), French (native), German (upper intermediate)

## Academic Duties

- Reviewing for: ARR, ACL, EMNLP, NAACL, AAAI, NeurIPS, ICML, ICLR
- Talks (2019–2024): presentations at NeurIPS, ACL, EMNLP, AAAI and various academic/industry venues (e.g., Bouygues Telécom, Paris-Saclay, Telécom Paris, DataIA, IRTSystemX, ILLS, Datacraft) on LLM adaptation, safe AI, evaluation & representation learning.
- Open Source:
  - Apache – *EuroLLM*: EuroLLM-9B and EuroLLM-1.7B - Generic LLM encoders for European languages, with performance in part with Llama 3.
  - MIT – *SaulLM*: A MOE-LLM optimized for legal tasks: SaulLM-141B - SaulLM-54B, Saul-7B-v1.
  - MIT – *ColPali*: ColPali - A multimodal model for documentary tasks.
  - MIT – *CroissantLLM*: CroissantLLM - A bilingual (French-English) generic LLM.
  - CC-BY-NC-4.0 – *TowerLLM*: TowerLLM - LLM for translation-related tasks.
  - MIT – *EuroBERT*: EuroBERT (2B - 600M - 200M) - General purpose encoder.

## Other Publications

### Safe AI (Out Of Distribution | Adversarial Attacks | Hallucinations)

- M Darrin, G Staerman, EDC Gomes, JCK Cheung, P Piantanida, **P Colombo**. *Unsupervised Layer-wise Score Aggregation for Textual OOD Detection.* AAAI 2024
- NM Guerreiro, **P Colombo**, P Piantanida, AFT Martins. *Optimal Transport for Unsupervised Hallucination Detection in Neural Machine Translation. ACL 2023*
- M Darrin, P Piantanida, **P Colombo** . *Rainproof: An Umbrella To Shield Text Generators From Out-Of-Distribution Data. EMNLP 2023*
- M Picot, N Noiry, P Piantanida, **P Colombo**. *Adversarial Attack Detection Under Realistic Constraints*
- M Picot, G Staerman, F Granese, N Noiry, F Messina, P Piantanida, **P Colombo**. *A Simple Unsupervised Data Depth-based Method to Detect Adversarial Images.*

- M Picot, F Granese, G Staerman, F Messina, M Romanelli, P Piantanida, **P Colombo**. *A Halfspace-Mass Depth-Based Method for Adversarial Attack Detection.* TMLR 2023
- **P Colombo**, G Staerman, N Noiry, P Piantanida. *Beyond Mahalanobis Distance for Textual OOD Detection.* NeurIPS 2022
- **P Colombo**, M Picot, F Granese, N Noiry, G Staerman, P Piantanida. *Toward Stronger Textual Attack Detectors. EMNLP 2023*
- E Gomes, **P Colombo**, N Noiry, G Staerman, P Piantanida. *A Functional Perspective on Multi-Layer Out-of-Distribution Detection.*

## Fairness with Large Models

- **P Colombo**, N Noiry, G Staerman, P Piantanida. *A Novel Information Theoretic Objective to Disentangle Representations for Fair Classification* Findings of AACL 2023
- G Pichler*, **P Colombo***, M Boudiaf*, G Koliander, P Piantanida. *KNIFE: Kernelized-Neural Differential Entropy Estimation.* (Oral) ICML 2022
- **P Colombo**, G Staerman, N Noiry, P Piantanida. *Learning Disentangled Textual Representations via Statistical Measures of Similarity.* (Oral) ACL 2022
- **P Colombo**, C Clavel and P Piantanida. *A Novel Estimator of Mutual Information for Learning to Disentangle Textual Representations* (Oral) ACL 2021
- H Jalalzai*, **P Colombo***, C Clavel, E Gaussier, G Varni, E Vignon, and A Sabourin. *Heavy-tailed representations, text polarity classification & data augmentation.* NeurIPS 2020

## Multimodal Large Language Model

- **P Colombo**, E Chapuis, M Labbeau and C Clavel. *Improving Multimodal fusion via Mutual Dependency Maximisation.* (Oral) EMNLP 2021
- A Garcia*, **P Colombo***, S Essid, F d'Alché-Buc, and C Clavel. *From the token to the review: A hierarchical multimodal approach to opinion mining* EMNLP 2019

## Large Language Model

- H Gisserot-Boukhlef, R Rei, E Malherbe, C Hudelot, **P Colombo**, N M Guerreiro *Is Preference Alignment Always the Best Option to Enhance LLM-Based Translation? An Empirical Analysis* (Oral) WMT 2024
- **P Colombo**, TP Pires, M Boudiaf, D Culver, R Melo, C Corro, AFT Martins, et al. *SaulLM-7B: A pioneering Large Language Model for Law.*
- H Gisserot-Boukhlef, M Faysse, E Malherbe, C Hudelot, **P Colombo** *Towards Trustworthy Reranking: A Simple yet Effective Abstention Mechanism* TMLR 2024
- A Himmi, G Staerman, M Picot, **P Colombo**, NM Guerreiro *Enhanced Hallucination Detection in Neural Machine Translation through Simple Detector Aggregation* EMNLP 2023
- **P Colombo***, V. Pellegrin*, M Boudiaf, I. Ben Ayed, M Tami, et al. *Transductive Learning for Textual Few-Shot Classification in API-based Embedding Models* EMNLP 2023
- **P Colombo** among 20+ authors. *The BigScience Corpus A 1.6 TB Composite Multilingual Dataset.* (Oral) NeurIPS 2022
- Over 400 authors with **P Colombo**. *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model* JMLR 2024

## Evaluation of AI Systems

- A Himmi, E Irurozki, N Noiry, S Clemencon, **P Colombo** *Towards More Robust NLP System Evaluation: Handling Missing Scores in Benchmarks.* Findings of EMNLP 2023
- **P Colombo**, M Peyrard, N Noiry, R West, P Piantanida *The Glass Ceiling of Automatic Evaluation in Natural Language Generation.* Findings of AACL 2023
- C Chung **P Colombo**, F Suchanek, C Clavel. *Of Human Criteria and Automatic Metrics: A Survey and Benchmark of the Evaluation of Story Generation.* (Oral) COLLING 2022

- **P Colombo**, G Staerman, C Clavel and P Piantanida. *Automatic Text Evaluation through the Lens of Wasserstein Barycenters.* (Oral) EMNLP 2021
- G Staerman, P Mozharovskyi, **P Colombo**, S Clémençon, F d'Alché-Buc. *A Pseudo-Metric between Probability Distributions based on Depth-Trimmed Regions.* TMLR 2024

## Conversational AI

- **P Colombo** with over 80+ authors. *Nl-augmenter: A framework for task-sensitive natural language augmentation* NEJLT 2023
- **P Colombo***, E Chapuis*, M Labeau, C Clavel. *Code-switched inspired losses for generic spoken dialog representations.* EMNLP 2021
- **P Colombo**, C Yang, G Varni, C Clavel. *Beam search with bidirectional strategies.* ICNLSP 2021
- E Chapuis*, **P Colombo***, M Manica, M Labeau, C Clavel. *Hierarchical pre-training for sequence labelling in spoken dialog* Findings of EMNLP 2020
- T Dinkar*, **P Colombo***, M Labeau, C Clavel. *The importance of fillers for text representations of speech transcripts.* EMNLP 2020
- **P Colombo***, E Chapuis*, M Manica, E Vignon, G Varni, C Clavel. *Guiding attention in sequence-to-sequence models for dialogue act prediction.* (Oral) AAAI 2020
- **P Colombo***, W Witon*, A Modi, J Kennedy, M Kapadia. *Affect-driven dialog generation.* NAACL 2019

## Others

- (PhD Thesis) **P Colombo** *Learning to represent and generate text using information measures*
- W Witon*, **P Colombo***, A Modi, M Kapadia. *Disney at IEST 2018: Predicting emotions using an ensemble.* Workshop WASSA@EMNLP 2018