

# Jean Mercat

Senior Research Scientist — Foundation Model Research

Mountain View, CA | [jean.mercat@gmail.com](mailto:jean.mercat@gmail.com) | +1 (669) 250-6406  
[jean-mercet.netlify.app](http://jean-mercet.netlify.app) | [Google Scholar](#) | [GitHub](#) | [LinkedIn](#)

## SUMMARY

---

Five years pre-training generative models at Toyota Research Institute, with a prior PhD in machine learning for motion forecasting. Specialize in **transformer pre-training**: model architecture, data filtering and scaling, and evaluation; applied across LLMs, VLMs, and VLAs.

Looking to push the frontier of automation with a generalization-first approach.

## EXPERIENCE

---

**Senior Research Scientist, Machine Learning**, Toyota Research Institute  
*Los Altos, CA*

Jul 2021 – Present

Technical lead for the machine learning pre-training team. Drive research on model architecture, data scaling, and evaluation methodology, with multi-node training runs up to 7B parameters / 1T tokens (PyTorch on AWS SageMaker).

- **VLA Foundry.** Co-first author and primary contributor on a unified open-source codebase for LLM, VLM, and VLA training (8-person team, 5 core contributors). Single codebase across modalities, HF backbone swapping (vision, vision-language, language), multi-node training to 128 GPUs / 16 nodes, integrated LBM-simulator evaluation, and shipped data, pre-trained weights, and tutorials. 300+ GitHub stars in the first week;
- **Data and scaling.** Contributed to training code and data-filtering exploration on [OpenLM](#), used in [LLM scaling](#) (ICLR 2025) and [DataComp-LM](#) (NeurIPS 2024); contributed to the evaluation infrastructure for [OpenThoughts](#) ([Evalchemy](#)), ran and interpreted evaluations (notably the impact of inference-time thinking length), and authored the [OpenThoughts3 – 1,000 models evaluated](#) analysis and accompanying [HF data explorer](#).
- **Multimodal pre-training.** Led an ICLR 2025 first-author study showing that introducing the visual modality earlier in LLM training (vs. fully separate stages with LR cooldown) improves both LLM retention and downstream VLM quality, and that introducing the smallest highest-quality datasets too early is detrimental. Evaluated all checkpoints on LLM and VLM benchmarks to characterize forgetting across ablations.
- **Architecture.** Designed and implemented a method to convert pre-trained Transformer LLMs into sub-quadratic linear-attention models; demonstrated poor long-context performance of linear attention. Open-weights releases ([Mamba-7B](#), [Mistral-SUPRA](#)). First author, COLM 2024.
- **Mentorship.** Lead cross-functional collaborations with university partners, and mentor interns.

**Ph.D., Machine Learning for Motion Forecasting**, Renault & Université Paris-Saclay  
*Paris, France*

Sep 2017 – Mar 2021

- Machine learning models (Multi-head attention, RNNs, CNNs, VAEs) for vehicle trajectory forecasting; built data pipelines and evaluation tooling end-to-end.
- Won **1st place** twice in the [Argoverse Motion Forecasting Challenge](#) ([NeurIPS 2019](#), [CVPR 2020](#)), against industry and academic teams (Waymo, Uber, Alibaba, etc.);

**Data Engineer**, Inria  
*Bordeaux, France*

2017

Built a C++/PostgreSQL pipeline for medical-imaging (DICOM) data; cross-disciplinary work between CS and medicine.

**Research Engineer**, Université de Bordeaux  
*Bordeaux, France*

2016

Implemented image-processing and anisotropic-annealing simulations in C++/Qt; taught a 22h Fortran tutorial at ENSEIRB-MatMeca; co-author in [Macromolecular Rapid Communications](#).

## SELECTED PUBLICATIONS

---

Full list: [Google Scholar](#). Selected, recent, LLM-focused:

- **VLA Foundry: A Unified Framework for Training Vision-Language-Action Models.** *Mercat et al.*, 2026. [https://tri-ml.github.io/vla\\_foundry/](https://tri-ml.github.io/vla_foundry/)
- **Should VLMs Be Pre-trained with Image Data?** S. Keh, J. Mercat, et al. (co-first authors), ICLR 2025. [arXiv:2503.07603](#)
- **OpenThoughts: Data Recipes for Reasoning Models.** ICLR 2026. [arXiv:2506.04178](#)
- **DataComp-LM: In Search of the Next Generation of Training Sets for Language Models.** NeurIPS 2024. [arXiv:2406.11794](#)
- **Language Models Scale Reliably with Over-Training and on Downstream Tasks.** NeurIPS 2024. [arXiv:2403.08540](#)
- **Linearizing Large Language Models.** *Mercat et al.*, COLM 2024. [arXiv:2405.06640](#)
- **A Systematic Study of Data Modalities and Strategies for Co-training Large Behavior Models for Robot Manipulation.** RSS 2026. [arXiv:2602.01067](#)

## TECHNICAL SKILLS

---

**Models:** LLM, VLM, VLA pre-training and fine-tuning; linear-attention; multimodal training; evaluation methodology.

**Systems:** Multi-node distributed training on AWS SageMaker (PyTorch, FSDP); large-scale data curation and processing.

**Other:** Scientific computing (C++, MPI, HPC); technical leadership and intern mentoring.

## OPEN SOURCE

---

**Frameworks (co-first author / primary contributor):** [VLA Foundry](#), [Linear OpenLM](#).

**Released models / demos:** [VLA Foundry collection](#), [Mamba-7B](#), [Mistral-SUPRA](#), [OpenThoughts data explorer](#).

**Other code contributions:** [OpenLM](#), [Evalchemy](#).

## EDUCATION

---

**Ph.D., Machine Learning** — Université Paris-Saclay & Renault 2017 – 2021  
*Road Scene Motion Forecasting with Deep Sequence Models*

**M.Eng., Scientific Computing** — ENSEIRB-MatMeca, Bordeaux INP 2012 – 2015

## REFERENCES

---

References (past colleagues, past managers and PhD advisors) available on request.