

Gabriel Jacob Perin

[!\[\]\(919a2cb85b99741a73c0c31a427236a8_img.jpg\) gabrieljp@usp.br](mailto:gabrieljp@usp.br)

[!\[\]\(666e09182d4cd268646ea700ea60dcdf_img.jpg\) gabjp](https://orcid.org/0000-0002-1343-832X)

[!\[\]\(c3d993ca47bfe2a953c700506ce31fa0_img.jpg\) Google Scholar](https://scholar.google.com/citations?user=0000-0002-1343-832X&hl=en)

Research Interests

My research focuses on three interconnected fronts: **Large Language Models (LLMs)**, **Geometric Deep Learning (GDL)**, and **AI for Science (AI4Science)**. I aim to develop methods that are efficient (both data- and compute-wise), interpretable, safe, and robust. Within AI4Science, I am particularly interested in applications across astronomy, materials science, and healthcare, among other scientific domains.

Education

BS University of São Paulo, Computer Science

March 2021 - December 2025

- GPA: 9.2/10
- **Coursework:** Machine Learning, Optimization, Algorithms, Discrete Mathematics.
- **Thesis:** An Introduction to Geometric Deep Learning on Sets, Graphs, and Grids.

MS University of São Paulo, Computer Science

March 2026 - December 2027 (Expected)

- Advisor: Prof. Nina S. T. Hirata.

Research Experience

University of São Paulo, Research Assistant

SP, BR

February 2022 - April 2025

- Advisor: Prof. Nina S. T. Hirata.
- Developed machine learning models (Random Forests, CNNs, and self-supervised methods) for classifying astronomical objects (galaxies, stars, quasars) in the S-PLUS survey, under the co-advising of Prof. Claudia L. M. de Oliveira.
- Designed few-shot retinal disease classification pipelines using meta-learning algorithms (Reptile) with CNN and Vision Transformer backbones.

University of Texas at Austin, Research Visitor

TX, USA

April - July 2024

- Advisor: Prof. Zhangyang "Atlas" Wang.
- Researched model merging techniques, with a focus on applications to Large Language Models (LLMs).
- Investigated robustness of safety alignment under benign and malicious fine-tuning scenarios.
- Collaborated with the group remotely (May 2023 – April 2025) in addition to the official research visit (Apr – Jul 2024).

IBM Research, Intern

SP, BR

November 2024 - November 2025

- Manager: Dr. Mathias Steiner.
- Worked on Machine Learning Interatomic potentials (MLIPs) and Foundation Models for Material Science ([pos-egnn](#) ).

Teaching Experience

University of São Paulo, Teaching Assistant

SP, BR

March 2025 - June 2025

- Advisor: Prof. Nina S. T. Hirata.
- Course: Introduction to Machine Learning (undergraduate)
- Assisted students by answering questions and provided detailed feedback through grading assignments.

Publications

AD-VF: LLM-Automatic Differentiation Enables Fine-Tuning-Free Robot Planning from Formal Methods Feedback	2026
Y. Yang, J. Hong, G. Jacob Perin , Z. Fan, L. Yin, Z. Wang, U. Topcu ICRA 2026	
LoX: Low-Rank Extrapolation Robustifies LLM Safety Against Fine-tuning	2025
G. Jacob Perin , R. Chen, X. Chen, N. S. T. Hirata, Z. Wang, J. Hong COLM 2025	
Extracting and understanding the superficial knowledge in alignment	2025
R. Chen, Gabriel Jacob Perin , X. Chen, X. Chen, Y. Han, N. S. T. Hirata, J. Hong, B. Kailkhura NAACL 2025	
Few-shot Retinal Disease Classification on the Brazilian Multilabel Ophthalmological Dataset	2024
G. Jacob Perin , Nina S. T. Hirata SIBGRAPI 2024	
RankMean: Module-Level Importance Score for Merging Fine-tuned LLM Models	2024
G. Jacob Perin , X. Chen, S. Liu, B. Kailkhura, Z. Wang, B. Gallagher ACL 2024 - Findings (short paper)	
The Fourth S-PLUS Data Release: 12-filter photometry covering 3000 square degrees in the Southern Hemisphere	2024
F. R. Herpich, ..., G. Jacob Perin , et al. Astronomy and Astrophysics (A&A)	
Combinação de Dados Tabulares e Imagens para a Classificação de Objetos Astronômicos	2023
G. Jacob Perin , L. Nakazono, C. Mendes de Oliveira, N. S. T. Hirata SIBGRAPI 2023, Workshop of Undergraduate Works (WUW)	

Awards and Grants

FAPESP Scholarship for Research Experience for Undergraduates, October 2022 - November 2024
Highlight of the Intermediate Phase - International Symposium of Scientific and Technological Initiation of the University of São Paulo (SIICUSP), 2023
FAPESP International Fellowship for Research Experience for Undergraduates (BEPE), April - July 2024

Event Presentations

LoX: Low-Rank Extrapolation Robustifies LLM Safety Against Fine-tuning	2025
G. Jacob Perin , R. Chen, X. Chen, N. S. T. Hirata, Z. Wang, J. Hong Poster, COLM 2025	
Few-shot Retinal Disease Classification on the Brazilian Multilabel Ophthalmological Dataset	2024
G. Jacob Perin , N. S. T. Hirata Oral, SIBGRAPI 2024	
RankMean: Module-Level Importance Score for Merging Fine-tuned LLM Models	2024
G. Jacob Perin , X. Chen, S. Liu, B. Kailkhura, Z. Wang, B. Gallagher Poster, ACL 2024 (digital)	

G. Jacob Perin, L. Nakazono, C. Mendes de Oliveira, N. S. T. Hirata

Poster, SIBGRAPI/WUW 2023 & SIICUSP 2023

Others

Languages: Portuguese (native), English (fluent), Spanish (basic)

L.E.A.R.N founding member - Machine Learning group advised by Prof. Nina S. T. Hirata

Technical Skills: Python (Pytorch, Pytorch Lightning, TensorFlow, Pandas, Scikit-learn), C++, Git, Bash, SQL