

Thiago Tarraf Varella

LinkedIn | GitHub | Scholar | San Francisco, CA

SUMMARY

Machine Learning Research Scientist at Meta with a Princeton PhD in computational neuroscience. Builds and ships production ML at AI-infrastructure scale: the first ML system on a large GPU job-scheduling team, a transformer-based job-runtime predictor, agentic evaluation infrastructure, and TB-scale telemetry pipelines. Independently runs an autonomous ML research system covering the full loop from literature review to GPU-dispatched training runs; current track converts pretrained autoregressive models into diffusion language models targeting faster single-request inference. Published in Nature Communications Biology and PLoS Computational Biology.

EXPERIENCE

Meta

Jun 2025 - Present

Machine Learning Research Scientist · AI Control Plane (MAST) · Menlo Park, CA

- Led the introduction of ML to a large GPU job-scheduling team that previously ran on deterministic algorithms; owned the full roadmap end-to-end, from data pipelines and feature engineering through model training and API integration.
- Built a transformer-based encoder generating embeddings from heterogeneous job metadata, feeding a downstream network that predicts job runtime, data transfer, and resource availability.
- Built a statistical model for runtime upper-bound prediction at >95% accuracy enabling deadline-aware scheduling, with an expected ~15% reduction in preemptions; explored GBDT and Patch Time-Series Transformer backbones for v2.
- Built a SQL telemetry pipeline over terabytes of GPU data-transfer data feeding ML features and 8 production dashboards across AI Infra, Ads, FB, and recommendations.
- Built **agentic evaluation infrastructure** to prevent production regressions: daily automated evals for knowledge recall, adversarial robustness, and task completion; designed the scoring and added A/B testing of new skills into CI/CD.
- Traced a multi-system kill-signal chain causing orphaned jobs (waste sized at ~\$4M/year); fixed via a UI change contributing to ~\$2M in savings.

Princeton University

Jan 2017 - May 2025

Graduate Researcher (Quantitative Modeling) · Princeton Neuroscience Institute

- Built a custom **actor-critic RL environment** on a physics-based simulation, designing the reward function from acoustic error; high-exploration agents significantly outperformed imitation learning, and poorly-timed external reward corrupted the intrinsic learning signal (IEEE ICDL, 2024).
- Built a multi-agent Bayesian decision loop maximizing expected information gain to model how agents choose actions under uncertainty about a partner (*Nature Communications Biology*, 2024).
- Developed a thermodynamic landscape model quantifying behavioral transitions as trade-offs between exploration, information gain, and metabolic cost; R^2 0.86-0.96 across three species vs 0.54-0.80 for linear and recurrent baselines (*PLoS Computational Biology*, 2022).
- Built reproducible signal-processing pipelines (MATLAB / Python / R) for 100k+ longitudinal recordings (audio, EMG, EEG, ECoG); primary architect of the ML stack; 3 peer-reviewed papers; invited international talks.

INDEPENDENT RESEARCH

Autonomous ML Research System (I-DLM) Autonomous Experimentation

Independent framework that runs an ML research loop: literature review, experiment scaffolding, GPU-dispatched training runs, and metric-based hypothesis selection. Current track: I-DLM, converting a pretrained autoregressive model into a diffusion

language model with a custom training loop over a 10B-token corpus, targeting self-speculative decoding for faster single-request inference.

Python · PyTorch · wandb · Thunder Compute (A100 / A6000) · Claude Code

Photo Restoration Benchmark Frontier Model Eval

Benchmark evaluating frontier image models (GPT Image 2, InstructPix2Pix) on damaged photo restoration, with a synthetic degradation pipeline, perceptual similarity scoring, and VLM-based quality validation. Built to measure where current frontier models fail on a concrete real-world task.

Python · PyTorch · frontier-model APIs · perceptual metrics

SELECTED PUBLICATIONS & PRESENTATIONS

T. T. Varella, D. Y. Takahashi, A. A. Ghazanfar. "Active sampling in primate vocal interactions." *Nature Communications Biology*, 7, 1098 (2024).

T. T. Varella et al. "Explore-exploit trade-off in vocal learning." *IEEE ICDL* (2024). [Conference poster](#).

T. T. Varella et al. "Large language models in 12 points." *Nexo Políticas Públicas* (2024). Public-facing AI explainer for a Brazilian policy outlet.

T. T. Varella, Y. Zhang, A. A. Ghazanfar, D. Y. Takahashi. "A mechanism for punctuating equilibria during mammalian vocal development." *PLoS Computational Biology*, 18(6) (2022).

T. T. Varella, M. B. Reyes, M. S. Caetano, R. Y. De Camargo. "A model for the peak-interval task based on neural oscillation-delimited states." *Behavioural Processes*, 168 (2019).

EDUCATION

Ph.D., Psychology & Computational Neuroscience · [Princeton University](#) 2025
GPA 3.9/4.0 · Dissertation: "Vocal Adaptation in Evolutionary, Developmental, and Cognitive Timescales" · NSF Graduate Research Fellowship

B.Sc., Molecular Sciences (focus: Mathematics Applied to Neuroscience) · [University of São Paulo](#) 2018
GPA 8.9/10 · Merit-based exchange semester at Princeton (2017)

SKILLS

ML / AI:	PyTorch, NumPy, pandas, scikit-learn; actor-critic RL, RL environment & reward design; transformer encoders, GBDTs; Bayesian inference; generative models; synthetic data; multi-GPU training
LLM / Agentic:	Anthropic API, OpenRouter, Claude Code; LLM inference optimization, self-speculative decoding; tool use, skill composition, eval loops, frontier-model benchmarking
Programming:	Python (10+ yrs), SQL, MATLAB, R, C++ (modern), JavaScript, Bash
Data & Analysis:	SQL telemetry at TB scale, dashboards, statsmodels OLS, bootstrapping, time-series forecasting, anomaly detection
Engineering:	Git, trunk-based DevOps, CI/CD, code reviews, unit / E2E tests; Unix/Linux; observability
Signal processing:	Audio, EMG, EEG, ECoG; wav2vec 2.0 / AvesMeanEmbedding; Butterworth, FFT, ODE integration

OTHER

Science Communication: Host / founder, "Brains, Black Holes, and Beyond" podcast (2020-2023, Princeton researchers, 1000+ downloads); Nexo Políticas Públicas AI columnist

Awards & Training: NSF Graduate Research Fellowship (2020-2025); President's Fellowship, Princeton (2019); Princeton Psychology Service Award (2023); BAMB Barcelona (2023); RLHF, Hugging Face (2024); Deep Learning for Multi-GPUs, Princeton PICSciE (2023)

Languages: Portuguese (native), English (fluent), Spanish (working), French (working)