Luis Miguel Montoya

AI Safety Researcher | Research Engineer

luismmonh@gmail.com · (+57) 3117180665 · Medellín, Colombia
github.com/lmmontoya-ai · www.luismmontoya.dev · linkedin.com/in/luismmontoya

Summary & Core Skills

Summary: Empirical AI safety researcher working on when model behavior can be predicted, monitored, and trusted. First-author paper at the ICML 2026 Mechanistic Interpretability Workshop on unverbalized evaluation awareness, a chain-of-thought monitorability failure mode. Currently a Pivotal Research Fellow studying whether LLM agents genuinely model the behavior of other agents they interact with or oversee. Three years building production ML systems before moving into safety research full time.
Core areas: Chain-of-thought monitorability and faithfulness; evaluation awareness; alignment evaluations; behavioral and activation-level monitoring; model organisms; scalable oversight; AI control; mechanistic interpretability.
Methods: Linear probes; activation and path patching; activation steering; on-policy trace resampling; automated behavioral evaluations; LLM autoraters; TransformerLens and NNsight; logit lens; counterfactual edits.
Training & inference: PyTorch; JAX; HF Transformers/Accelerate; vLLM distributed inference; PEFT/LoRA/QLoRA; DPO; DeepSpeed; experiment harnesses.
Engineering: Reproducible pipelines; seed control; evaluation harnesses; dataset sanitation; monitoring workflows; profiling and debugging; Git/GitHub CI.
Programming & languages: Python (advanced); Bash; Spanish (native); English (C1).

Publications

- **Towards Measuring and Detecting Unverbalized Evaluation Awareness. L. M. Montoya***, R. Haskins*, B. Adams (*equal contribution). *ICML 2026 Mechanistic Interpretability Workshop* (poster). Evaluation-conditioned behavior survives complete suppression of verbalized awareness; cue-masked linear probes recover eval/deploy context from suppressed generations. Code.
- **Warm-Starting Pretrained Transformers for Long-Context Test-Time Training. L. M. Montoya**, V. K. Madisetti. Manuscript in preparation, 2026.

Research Experience

Pivotal Research Fellow, Self- and Other-Modeling in LLM Agents **2026/06 – Present**
Pivotal Research Fellowship, mentored by Damiano Fornasiere and Gaël Gendron (LawZero) *Remote*

- Studying whether LLM agents can predict other models' next actions in shared multi-agent environments, decomposing forecasts into target-specific modeling, self-projection, and role or base-rate components; oversight schemes assume the overseer can model the overseen, and self-projecting monitors fail precisely on targets unlike themselves.
- Built a behavioral measurement harness where every predictor condition (model, role, information visibility) is evaluated against identical frozen environment snapshots, comparing sampled forecast and target action distributions (Jensen–Shannon divergence, total variation, mode agreement) against self-prediction, anonymized-context, shuffled-history, and self-projection controls.
- Running the predictor–target matrix across social-simulation, workplace, and AI-control environments (SHADE-Arena, ControlArena adapters), where a safety monitor observes an executor that may pursue a covert side objective.

Independent Researcher, Unverbalized Evaluation Awareness **2025/10 – Present**
Originated in MATS 9.0 exploration phase (Neel Nanda stream); continued with R. Haskins and B. Adams (U. Canterbury) *Remote*

- First-author paper accepted at the ICML 2026 Mechanistic Interpretability Workshop on evaluation-conditioned behavior that persists when verbalized awareness is suppressed, a concrete chain-of-thought monitorability failure mode.
- Suppressed all detected verbalizations of evaluation awareness via on-policy trace resampling (Thought Branches); the eval/deploy behavior gap survived at +7.0pp in a Llama-Nemotron-49B model organism and +65.3pp in a Qwen3.5 cross-family replication, at zero detected verbalization.
- Trained cue-masked linear probes on non-coding eval/deploy contrasts that transfer across domains and recover the context from verbalization-suppressed generations (response AUROC 0.81 and 0.71), an activation-level monitoring signal for cases where reading the reasoning trace is insufficient.
- Built the full experimental stack: model-organism fine-tuning, vLLM distributed inference, on-policy resampling, LLM autorater auditing with an adversarial second-auditor sensitivity check, probe training, and statistical testing.

Graduate Researcher, Long-Context Language Model Training **2025/08 – Present**
Georgia Institute of Technology (OMSCS), advised by Prof. Vijay K. Madisetti (IEEE Fellow) *Remote*

- First-author manuscript on warm-starting pretrained transformers for long-context TTT-E2E; controlled multi-scale study (125M and 760M) comparing four long-context training routes and separating faster training from final-quality gains.
- Warm-starting reduced marginal training cost by 7.4× while the final-quality gap halved from 125M to 760M; implemented the full reproduction stack in JAX on 8× H200 nodes, with per-position NLL and RULER/S-NIAH retrieval analyses.

GamesBench, Open-Source Benchmark **2025/09 – Present**
github.com/lmmontoya-ai/GamesBench *Open Source*

- Benchmark for long-horizon planning in tool-calling LLM agents, using deterministic games (Towers of Hanoi, Sokoban) to measure multi-step reasoning and action sequencing beyond single-turn tasks.

Industry Experience

Senior Data Scientist **2026/04 – Present**
UP.LABS / PartsPulse (via UP.CODE) *Remote / Medellín, Colombia*

- Developing models and decision-support systems for an aftermarket parts-intelligence platform, integrating demand, pricing, inventory, and behavioral signals for OEM and dealer workflows.

Data Science Engineer **2024/11 – 2026/04**
Mercado Libre *Medellín, Colombia*

- Built an LLM-assisted review system for risky transactions combining structured and unstructured signals; defined the intent taxonomy, curated seed data, and built an offline evaluation harness with human review in the loop.
- Deployed real-time anomaly detection on credit time-series features, capturing ∼70% of emerging account-takeover and device-theft patterns; improved detection precision by 10% using behavioral biometrics.

Earlier Data Science Roles **2022/08 – 2024/11**
Sistecrédito; Seguros SURA; Línea Directa *Medellín, Colombia*

- Credit default prediction (∼80% accuracy), corporate health-plan pricing with explainable dashboards, Double ML for dynamic pricing, and a Temporal Fusion Transformer that improved sales-forecast accuracy by 10%; introduced Línea Directa's first MLOps framework.

Education

M.S. in Computer Science, Machine Learning **2025/08 – 2027 (Expected)**
Georgia Institute of Technology *Remote*

- Completed: Reinforcement Learning; Introduction to Research (CS-8903). In progress: Machine Learning (CS-7641) and graduate research with Prof. Vijay K. Madisetti.

B.S. in Physics **2019/07 – 2023/07**
Universidad EIA *Medellín, Colombia*

- GPA: 4.29/5.00 (top 10%). Thesis on language models for venture-funding prediction in Latin America.

Fellowships, Grants & Awards

- **Pivotal Research Fellowship** (2026): competitive AI safety research fellowship; project on self- and other-modeling in LLM social agents.
- **MATS 9.0** (2025): exploration phase scholar in Neel Nanda's mechanistic interpretability stream.
- **Coefficient Giving Career Development Grant** (2025): competitive grant funding graduate tuition and research compute for AI safety work.