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

Empirical AI safety researcher and research engineer focused on predicting and detecting model behavior for oversight. Co-first author of an ICML 2026 Mechanistic Interpretability Workshop paper on unverbalized evaluation awareness. Pivotal Research Fellow measuring when language models can forecast other models' actions and whether their forecasts reveal goal inference beyond simple behavioral baselines. Build reproducible experiments from fine-tuning, long-context training, and distributed inference through behavioral evaluation, activation probing, and statistical analysis; industry ML experience since 2022.
Research: Alignment evaluations; behavioral forecasting of LLM agents; model organisms; evaluation awareness; scalable oversight; AI control; chain-of-thought monitorability; behavioral and activation-level monitoring; mechanistic interpretability.
Methods: Controlled behavioral evaluations; on-policy trace resampling; linear probes; SAE latent ablation; activation/path patching and steering; LLM autorater auditing; bootstrap and permutation tests; TransformerLens; NNsight.
Training & inference: PyTorch; JAX; Hugging Face Transformers/Accelerate; vLLM; PEFT/LoRA/QLoRA; DPO; DeepSpeed.
Engineering: Reproducible pipelines; evaluation harnesses; dataset validation; monitoring; profiling; 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). 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, Behavioral Forecasting in LLM Agents **2026/06 – Present**
Pivotal Research Fellowship and extension; mentors: Damiano Fornasiere and Gaël Gendron (LawZero) *Remote*

- Designed controlled studies of agent forecasting across office, contracting, and engineering simulations; the main panel compared nine forecasters predicting three target models at identical frozen decisions. Research report.
- Built repeated-sampling evaluations with noise-corrected Jensen–Shannon divergence, paired information interventions, and interaction-clustered confidence intervals to separate target behavior, forecast error, and forecaster self-projection.
- Tested situation, history, and private-goal information. A situation-plus-action-count baseline showed no detectable error difference from model forecasts in Office or Contracting and lower error in Engineering.
- Used controlled goal assignments and new-situation tests to distinguish behavioral effects from goal inference. In Office transfer tests, real-goal examples improved forecasts but showed no measurable advantage beyond action counts from the same examples.
- Showing earlier target reasoning reduced Contracting forecast error by 0.046 bits (95% CI for reduction: 0.030–0.057); replacing it with summaries removed the gain. Reasoning showed no detectable benefit in the other two settings.

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

- Co-first-authored research showing that evaluation-conditioned behavior can persist with no detected verbalization of evaluation awareness, identifying a failure mode for monitoring based on explicit awareness statements.
- Used on-policy trace resampling (Thought Branches) to suppress detected awareness verbalizations. The evaluation/deployment behavior gap remained at +7.0 percentage points in a Llama-Nemotron-49B model organism and +65.3 points in a Qwen3.5 cross-family replication.
- Trained cue-masked linear probes on non-coding evaluation/deployment contrasts; probes transferred across domains and recovered context from suppressed generations (response AUROC 0.81 and 0.71), providing an activation-level monitoring signal.
- Built the full experimental stack: model-organism fine-tuning, vLLM distributed inference, on-policy resampling, probe training, and statistical testing. Audited LLM autoraters with an adversarial second-auditor sensitivity check.

Research Experience (continued)

Independent Study, User Conditioning in GLM-5.3-Flash **2026/08 – 2026/09**
github.com/lmmontoya-ai/glm53-user-conditioning *Remote*

- Ran Transluce's user-awareness task on GLM-5.3-Flash with a preregistered confirmatory run (34,400 rows). The model stated slightly lower confidence to famous AI figures than generic fame predicts (interaction −0.65 pp, 95% CI [−1.21, −0.10]; about 0.06 SD), replicating the discovery run, with no measurable change in the yes/no answer. An explicit model-assessment cue moved confidence the other way (+1.08 pp, 95% CI [0.71, 1.46]).

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**
Vantora (formerly UP.LABS) / PartsPulse, contractor *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*

- GPA: 3.80/4.00. Completed: Machine Learning; Reinforcement Learning; Artificial Intelligence; Introduction to Research; graduate research with Prof. Vijay K. Madisetti (CS-8903). In progress: Deep Learning; Robotics: AI Techniques.

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 behavioral forecasting and goal inference in LLM agents; selected to continue in the extension program.
- **MATS 9.0** (2025): exploration phase scholar in Neel Nanda's mechanistic interpretability stream; application project on whether secret-word knowledge in Taboo Gemma-2-9B-IT is localized or distributed (SAE feature ablation, secret-direction subtraction).
- **Coefficient Giving Career Development Grant** (2026): competitive grant funding graduate tuition and research compute; 25k USD over 18 months.