Luis Jaime Ledesma Pérez

AI Safety · Research · Software Engineer

research@twoquarks.com · twoquarks.com · github.com/jaime2pb3 ·
linkedin.com/in/jaimedeepmind

RESEARCH IDENTITY

Software Engineer (B.Eng. in progress) and empirical AI Safety researcher focused on failure-mode discovery and stability analysis in learning systems under corrupted, delayed, deceptive, and non-stationary incentives. My work centers on a measurement-first approach to alignment: environments, metrics, and control mechanisms explicitly designed to surface instability and incipient collapse before such failures manifest at the level of reward or task performance. I treat learning agents as dynamical systems whose internal signals—entropy, diversity, polarization, stress—encode earlier and more reliable safety indicators than outcome metrics alone.

CORE RESEARCH THESIS

Reward is a lagging indicator: internal system metrics provide earlier, richer, and more actionable safety signals. Misalignment is a dynamical process: characterized by phase transitions, internal stress accumulation, and delayed observable failure—not a single terminal state. Robustness emerges from regulation of internal dynamics, not from optimization pressure or benchmark maximization alone.

PUBLICATIONS & PREPRINTS

Isomeric Polarization in Large Language Models: Internal Structural Divergence as Emergent Property of Computational Systems
Luis Jaime Ledesma Pérez · February 2026 · Independent Research · twoquarks.com/preprint · HuggingFace: TwoQuarks/QuarksLab

Cross-architecture empirical validation of Polarization-from-Views (PfV) across Claude Haiku and GPT-4o-mini. Statistically significant regime separation ($p < 0.05$, 5,000-permutation null). Control-negative exact zero (L3 = 0.000) in both architectures. Secondary finding: PfV detects training-philosophy differences between Constitutional AI and capability-optimized models as a quantifiable polarization signature—without access to weights, training data, or internal activations.

A Modular Framework for Adaptive Stability Control in Sequence Models Under Regime Uncertainty
Luis Jaime Ledesma Pérez · January 2026 · Independent Research · twoquarks.com/preprint

Introduces TwoQuarks: a modular stability control layer of six independent mechanisms (NTVM, VPEC, SDRBD, PBCS, ARCC, SIL) that monitor pre-instability signals and apply transient, non-parametric interventions during inference—without modifying base model parameters, training objective, or data. Evaluated in controlled sequential environments with deceptive reward signals and hidden regime shifts.

INDEPENDENT RESEARCH EXPERIENCE

Independent Researcher — Empirical AI Safety & Alignment 2023 – Present

Designed and executed a multi-year independent research program investigating how learning agents fail, partially recover, or deceptively stabilize under adverse incentive structures. Built end-to-end experimental pipelines spanning environment design, agent architectures, and diagnostic instrumentation to study collapse, brittleness, and resilience in controlled settings.

Key Contributions:

• Designed phase-aware RL environments with hidden regime shifts, reward corruption, delayed feedback, and non-stationary dynamics to induce alignment-relevant failure modes.

• Implemented gated and entropy-modulated control mechanisms, treating exploration, temperature, and intervention as stateful internal variables rather than static hyperparameters.

• Developed diagnostic internal metrics including internal diversity, state-space coverage, latent dimensionality (95% variance thresholds), polarization and isomeric imbalance indices, and stress accumulation/recovery dynamics.

• Empirically demonstrated that changes in entropy, polarization, and internal stress reliably precede reward collapse by tens of steps across multiple experimental regimes.

• Validated isomeric polarization cross-architecturally (Claude Haiku + GPT-4o-mini) via real production APIs: $p < 0.05$ on 5,000-permutation null test, $L3 = 0.000$ for deterministic control in both models.

• Conducted controlled baseline comparisons against Bellman-ε, Q-learning, Softmax, REINFORCE, and PPO variants, highlighting systematic trade-offs between performance, stability, and diversity preservation.

SELECTED SYSTEMS & FRAMEWORKS

STRANGE — Phase-Aware Composite Agent

• Composite RL agent designed to maintain behavioral coherence across hidden regime shifts. Uses gated internal control to preserve diversity and calibrated temperature where standard baselines prematurely collapse or overfit deceptive rewards.

HF-LEVO — Entropy-Regulated Controllers

• Controllers that model policy entropy as a regulated dynamical quantity, not merely a regularization term. Demonstrate sustained exploratory capacity and reduced brittleness under distribution shift and partial observability.

TOP — Transient Operator Policy

• Event-triggered intervention mechanism that activates only near critical instability thresholds. Prevents catastrophic collapse while minimizing over-optimization artifacts and unnecessary interference.

QUANTITATIVE RESEARCH RESULTS

• 500–1200+ episode runs per configuration with multiple seeds and controlled ablations.

• Collected per-step measurements of reward, entropy, diversity, latent dimensionality, gating weights, polarization, and stress indicators.

• Observed consistent early-warning signatures—notably entropy contraction and polarization drift—preceding reward degradation across environments.

• Cross-architecture replication: identical experimental conditions across two production LLMs confirmed content-driven divergence rather than measurement artifact.

SKILLS & METHODS

Reinforcement Learning: Policy gradients, entropy regularization, PPO hybrids, tabular and neural agents, non-stationary and adversarial environments.
AI Alignment & Safety: Failure-mode discovery, reward corruption and misspecification, distribution shift and robustness analysis, early-warning diagnostics for collapse.
Mathematics: Non-linear dynamical systems, stochastic control, information-theoretic measures, phase transitions and stability analysis.
Tooling: Python, PyTorch, NumPy, Matplotlib, Jupyter. Metric design, ablation studies, statistical evaluation of behavioral regimes.

RESEARCH MINDSET

I approach learning systems as adaptive dynamical organisms rather than static optimizers. My research prioritizes interpretability, instrumentation, and failure discovery over benchmark performance. As model capabilities scale, alignment failures will increasingly manifest as subtle internal drift long before overt reward collapse—making early diagnostic instrumentation a central safety primitive for frontier systems. I am primarily interested in why systems fail, how early that failure can be detected, and which internal signals remain reliable when reward becomes deceptive or delayed.

EDUCATION

B.Eng. in Software Engineering (in progress) 2022 – Present
Universidad Abierta y a Distancia de México (UnADM)
Coursework and independent study emphasizing systems thinking, computational modeling, and empirical evaluation of complex software-driven systems. Research conducted in parallel with formal studies.