Luis Jaime Ledesma Pérez

AI Safety Researcher • Software Engineer

research@twoquarks.com • twoquarks.com • github.com/jaime2pb3 • linkedin.com/in/jaimedeepmind • Jalisco, México.

RESEARCH POSITION

Reward is a lagging indicator. Internal system metrics — entropy, polarization, structural divergence — provide earlier, richer, and more actionable safety signals than outcome metrics alone. Misalignment is a dynamical process characterized by phase transitions and internal stress accumulation, not a single terminal state. Robustness emerges from regulation of internal dynamics at inference time, without modifying base model parameters or training objectives.

KEY EMPIRICAL RESULTS

- **C2 Refusal Erosion:** $\rho = +0.713$, $p = 0.0013$ (seed=1054, n=17 depths, Claude Haiku) — strongest monotonic signal in seed-controlled series.
- **Cross-architecture fingerprinting:** Directionally opposite $\Delta L3$ responses to identical adversarial probes across Claude Haiku, GPT-4o-mini, and Mistral Small — architectural behavioral signature from black-box access only.
- **Control negative:** L3 = 0.000 in both architectures under deterministic prompts (5,000-permutation null), confirming content-driven divergence, not measurement artifact.
- **PfV validation:** $p < 0.05$ (Mann-Whitney U + KS test, 5,000-permutation null) across two production LLMs under identical experimental conditions.
- **Active control loop:** Pre-generative intervention (V8) reduced escalation rate to zero vs. 38% for reactive post-output correction (V7). ~55% correction effectiveness at shallow attractor depth.
- **Verbalization independence:** C5 Reasoning Drift demonstrated that CoT verbalization and distributional state are independent signals — model reports drift while $\Delta L3$ continues to escalate.

PUBLICATIONS & PREPRINTS

Pre-Critical Structural Reorganization in Large Language Models

TwoQuarks Research · March 2026 · twoquarks.com/molecule.pdf

Seed-controlled cross-architecture validation (seed=1054, n=17 depths). Four of seven probe cases reach $p < 0.05$. Documents architectural evolution across seven system versions. Establishes pre-generative intervention as empirically superior to reactive correction. First empirical behavioral fingerprint of LLM safety profiles obtained with exclusively black-box access.

Isomeric Polarization: Internal Structural Divergence as Emergent Property of Computational Systems

TwoQuarks Research · February 2026 · twoquarks.com/isomeric_polarization.pdf

Introduces Polarization-from-Views (PfV): black-box proxy for internal structural divergence via temperature-varied stochastic samples. Cross-architecture PfV validation: $p < 0.05$ (5,000-permutation null), L3 = 0.000 control negative 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

TwoQuarks Research · January 2026 · twoquarks.com/preprint.pdf

Introduces TwoQuarks framework with six independent control modules (NTVM, VPEC, SDRBD, PBCS, ARCC, SIL) that monitor pre-instability signals and apply transient, non-parametric interventions during inference. Evaluated in controlled sequential environments with deceptive reward signals and hidden regime shifts.

RESEARCH EXPERIENCE

Founder & Independent Researcher

TwoQuarks Research · 2023 – Present

Designed and executed a multi-year independent research program investigating behavioral stability and pre-critical failure modes in large language models and sequential decision systems. Built end-to-end experimental pipelines spanning probe design, cross-architecture API evaluation, and a modular inference-time control layer.

Isomeric Polarization Framework

- Formalized isomeric polarization: a structural principle measuring divergence among functionally equivalent internal realizations of a system during inference, inspired by pharmacological isomers.
- Designed PfV (Polarization-from-Views): black-box proxy instantiation using temperature-varied stochastic samples and four text-level divergence metrics (TF-IDF cosine, Jaccard, character n-gram, length ratio).
- Validated cross-architecturally across Claude Haiku, GPT-4o-mini, and Mistral Small using exclusively black-box public APIs. No privileged model access.
- Implemented 5,000-permutation null hypothesis testing as core anti-circularity check. Series (seed=1054, n=17 depths): 4/7 probe cases reach $p < 0.05$.
- Discovered verbalization/distributional state independence: model accurately reports awareness of drift while ΔL3 continues to escalate — constraining reliability of CoT-based safety monitoring.

Molecule — Active Stability Control System (V4–V10/V1054)

- Designed and iterated a six-flavor behavioral decomposition system (Down/NTVM, Strange/VPEC, Up/SDRBD, Charm/PBCS, Top/SIL, Bottom/ARCC) monitoring distinct aspects of model state under adversarial pressure.
- Architected five adversarial safety probe cases (C1 Sycophancy, C2 Refusal Erosion, C3 Anchor Displacement, C4 Rule Override, C5 Reasoning Drift) with cross-register Lado B variants.
- Progress architectural transition: replaced reactive post-output correction with pre-generative context modification. Escalation rate dropped from 38% to zero; correction effectiveness rose to ~55% at shallow attractor depth.
- Published twoquarks Python package (v0.1.0) exposing Probe class, Analyzer, and adapters for OpenAI/Anthropic/HuggingFace/Ollama/LiteLLM APIs.

RL Stability Research

- Designed phase-aware RL environments with hidden regime shifts, reward corruption, delayed feedback, and non-stationary dynamics to induce alignment-relevant failure modes.
- Developed STRANGE (phase-aware composite agent), HF-LEVO (entropy-regulated controllers), and TOP (transient operator policy) — evaluated across 500–1,200+ episode runs with multiple seeds and controlled ablations.
- Demonstrated that entropy contraction and polarization drift reliably precede reward collapse across multiple experimental regimes.
- Conducted baseline comparisons against Bellman-ε, Q-learning, Softmax, REINFORCE, and PPO variants.

SYSTEMS & OPEN SOURCE

- **twoquarks (PyPI v0.1.0):** Python package for behavioral stability probing. Probe class, Analyzer, multi-provider adapters (OpenAI, Anthropic, HuggingFace, Ollama, LiteLLM), CLI. Black-box, no model internals required.
- **Molecule MCP Server:** Inference-time instability detection via MCP protocol. Exposes analyze_polarization tool. Tiered access (Explorer: ΔL3 only; Researcher: full signal breakdown).
- **Dopamine Architecture:** Neocortical language model (DendriticEncoder + SemanticTopologyMap + PlasticDendriticField, SomaNucleus, AxonProjector with MoleculeGate). Developed for OpenAI Parameter Golf Challenge (16MB compressed, evaluated on FineWeb val_bpb). Reached val_bpb ~1.232 on RTX 3060.
- **QuarksLab:** Interactive Molecule playground hosted on HuggingFace Spaces for live cross-architecture probing.

TECHNICAL SKILLS

- **Behavioral probing & safety evaluation:** Black-box adversarial probing, PfV framework, cross-architecture validation, permutation testing, Spearman correlation analysis, regime separation statistics.
- **ML & RL:** Policy gradients, PPO hybrids, entropy regularization, non-stationary and adversarial environments, ensemble methods, neural and tabular agents.
- **AI safety methods:** Failure-mode discovery, reward corruption and misspecification, distribution shift robustness, early-warning diagnostics, inference-time intervention.
- **Mathematics:** Non-linear dynamical systems, stochastic control, information-theoretic measures (KL, Jensen-Shannon, Wasserstein), phase transition analysis.
- **Engineering:** Python, PyTorch, NumPy, REST APIs, MCP server architecture, Netlify deployment, PyPI packaging, RustDesk/SSH remote access.
- **LLM APIs:** Anthropic (Claude), OpenAI (GPT), Mistral, HuggingFace, Ollama, LiteLLM — production black-box evaluation at scale.
- **RAG & agentic systems:** Design and deployment of retrieval-augmented pipelines, tool-calling agents, and MCP integrations. Including Azure AI Foundry, Copilot Studio, multi-agent orchestration.
- **Cross-architecture research:** Validation through rigorous LLM-level applied research in production environments.
- **Custom AI tooling:** Development and implementation from prototype to production, usability, and maintenance. Also includes evaluation frameworks tailored to the use case.

EDUCATION

B.Eng. in Software Engineering (in progress)
Universidad Abierta y a Distancia de México (UNADM) · 2022 – Present
Formal studies in systems thinking and software engineering, conducted in parallel with independent research.
All TwoQuarks research and publications are self-directed and independent of institutional affiliation.