

Gustavo Pinto, Ph.D.

AI Engineer • Applied AI Research • LLM Systems

✉ mail@gustavopinto.org

in in/ghlp

🌐 gustavopinto

🌐 https://gustavopinto.org

🎓 Scholar: 5,200+ cites, h-index 40

📍 Belém, Pará, Brazil • available for remote work



Profile

- Computer scientist with 15+ years bridging **applied research and production engineering**. I turn research hypotheses into working artifacts — LLM agents, RAG and GraphRAG pipelines, evaluation harnesses, benchmark frameworks — that run in production *and* are reported in peer-reviewed venues. **Staff AI Engineer** at iFood and **Assistant Professor** at UFPA; 100+ publications (ICSE, TSE, CACM, EMSE — Qualis A1), 5,200+ citations, h-index 40.

Employment

2026–now ► **Staff AI Engineer — iFood Tech** (Remote, Brazil)

- Building *Niago*, a post-sales AI agent handling consumer-credit interactions at scale: LLM-powered conversational flows behind a pluggable provider layer, with safe defaults when customer data is missing. Stack: Python (FastAPI), React, PostgreSQL, Redis, cloud services and CI/CD for scalable deployment.
- Fine-tuned a GPT model to classify Niago's responses, replacing brittle heuristics with a task-specific classifier.
- Built the evaluation stack — a golden dataset that regression-tests LLM responses across prompt and model changes, making quality shifts measurable before release.
- Instrumented LLM trace observability (Langfuse) and APM (Datadog), making prompt regressions and latency issues debuggable in production.

2021–2026 ► **Zup Innovation** (Remote, Brazil) — **Staff AI Engineer** (2023–2026), **Head of Research** (2021–2023)

- Led a team of five researchers; defined the research agenda and turned it into functional proofs-of-concept validated with engineering and product teams. Early adopter of LLMs inside the company (alongside encoder-only transformers and classical ML), before they went mainstream.
- Built and maintained a Retrieval-Augmented Generation system in Python (LangChain, FastAPI, Flask, Docker), with vector databases (pgvector, ChromaDB) storing high-dimensional embeddings for retrieval at scale.
- Created the **benchmark framework** used to evaluate StackSpot AI agents — the reference harness for measuring agent quality across releases.
- Engineered agentic workflows (LangChain, LangGraph, StackSpot AI) over open and closed LLMs/SLMs, with API orchestration, cloud services and CI/CD for scalable deployment.
- Prototyped end-to-end AI use cases — ingestion, embedding generation, retrieval, downstream evaluation — validating feasibility before hand-off to product.

2017–now ► **Assistant Professor — Federal University of Pará (UFPA)** (Belém, Brazil)

- Published 100+ research papers, most in highly competitive venues (ICSE, TSE, EMSE, CACM), with artifacts and replication packages released alongside them.
- Supervised 40+ students (4 PhD, 10 MSc, 10+ capstone, 10+ undergraduate research); raised 800k+ BRL as PI and 2M+ BRL as co-PI for R&D on AI for software engineering.
- Editor-in-Chief of JSERD (2020–2022), leading a board of around 30 associate editors; CNPq Research Productivity Fellow since 2020 (renewed in 2023 and 2026).

Technical Skills

Python & SE ► Python (primary, 15+ years), Git and code review, automated testing, Docker, CI/CD, FastAPI/Flask REST services. Also Java, JavaScript/TypeScript, SQL.

ML & DL ► PyTorch, Hugging Face Transformers (fine-tuning and post-training of open models), scikit-learn; encoder-only transformers for classification, clustering and embeddings; classical ML pipelines.

LLM systems ► RAG and hybrid GraphRAG pipelines; agentic workflows (LangChain, LangGraph, StackSpot AI); prompt and context engineering; open and closed providers behind a pluggable layer; SLM/LLM trade-off analysis for cost, latency and quality. Post-training of open models: instruction fine-tuning and direct preference optimisation (DPO) to align model behaviour with task-specific criteria.

Technical Skills (continued)

- Evaluation** ▶ Benchmark framework design for AI agents; golden datasets and regression testing of LLM responses across prompt/model changes; LLM-as-a-judge and human-in-the-loop protocols; observability with Langfuse and Datadog APM.
- Data at scale** ▶ PostgreSQL (DDL, DML, indexing, query tuning) and AWS RDS; vector databases (pgvector, ChromaDB); Databricks and Kafka for data ingestion; embedding generation, batch inference, reproducible experiment infrastructure.
- Spoken** ▶ Portuguese (native), English (fluent), French (beginner).

Selected Publications

Selection below: recent *Qualis A1* venues and AI/ML/LLM work. Full list at <https://gustavopinto.org/publications/> and *Google Scholar*.

- 2026** ▶ D. Naiff, R. Alves, **G. Pinto**. “Beyond Vector Search: Comparing Classical RAG with Hybrid GraphRAG for Climate Science Q&A”. *ENIAC*, 2026. [RAG/GraphRAG]
- ▶ T. Rocha, L. Teran, M. Mota, C. de Souza, K. Gama, **G. Pinto**. “Challenges and Enablers: Remote Work for People with Disabilities in Software Development Teams”. *ICSE 2026 — SEIS Track*. [Qualis A1]
- ▶ L. A. Costa, E. Dias, D. Ribeiro, A. Fontão, **G. Pinto**, R. P. dos Santos, A. Serebrenik. “Supporting Software Engineering Managers in Talent Retention: The Brazilian Context”. *Empirical Software Engineering (EMSE)*, 2026. [Qualis A1]
- 2025** ▶ R. Ferreira, G. da Silva, C. Rocha, **G. Pinto**. “Semantic Clustering of Civic Proposals: A Case Study on Brazil’s National Participation Platform”. *ENIAC*, 2025. [NLP for civic participation]
- 2024** ▶ **G. Pinto**, C. de Souza, J. B. Neto, A. Souza, T. Gotto, E. Monteiro. “Lessons from Building StackSpot AI: A Contextualized AI Coding Assistant”. *ICSE 2024 — SEIP Track*, Lisbon, Portugal. [Qualis A1, RAG in production]
- 2023** ▶ **G. Pinto**, I. Cardoso-Pereira, D. M. Ribeiro, D. Lucena, A. Souza, K. Gama. “Large Language Models for Education: Grading Open-Ended Questions Using ChatGPT”. *SBES 2023 — Education Track*. [Best Paper Award]
- 2021** ▶ E. Dias, P. Meirelles, F. Castor, I. Steinmacher, I. Wiese, **G. Pinto**. “What Makes a Great Maintainer of Open Source Projects?”. *ICSE 2021*, Madrid, Spain. [ACM SIGSOFT Distinguished Paper], Qualis A1
- 2020** ▶ **G. Pinto**, B. Miranda, S. Dissanayake, M. d’Amorim, C. Treude, A. Bertolino. “What is the Vocabulary of Flaky Tests?”. *17th International Conference on Mining Software Repositories (MSR)*, Seoul, South Korea, 2020. [Classical ML over test code; Qualis A2]

Education

- 2011–2015** ▶ **Ph.D. in Computer Science** — Informatics Center, **Federal University of Pernambuco (UFPE)**. Thesis on energy efficiency of software systems.
- 2008–2010** ▶ **M.Sc. in Computer Science** — **Federal University of Paraná (UFPR)** • **B.Sc.** — **Centro Universitário do Pará (Cesupa)**, 2004–2007.

Awards, Service and Community

- Awards** ▶ **MSR Ric Holt Early Career Achievement Award** (2022) • **ACM SIGSOFT Distinguished Paper**, ICSE 2021 • **JSS Paper of the Year** 2019 • top 2% most-cited scientists in Software Engineering (Elsevier, 2024) • best-paper awards at SBES’23, SBES’20, ICSE-SEET’19, MSR’17 and VEM’17.
- Service** ▶ General Chair of CBSOft 2028; Editor-in-Chief of JSERD (2020–2022); 70+ program committees; keynotes at SEMLA 2026 (Polytechnique Montréal), BOTSE 2026 and VEM 2024; 20+ invited talks. Author of the **ML4SE** newsletter (<https://ml4se.substack.com>) on ML and LLMs for developers.
- Outside work** ▶ Black belt in Brazilian jiu-jitsu.