Gaurav Koley

+1 857 318 6502 | @gaurav@koley.in | LinkedIn | GitHub | Website | San Francisco, CA

Work Experience

Klarity San Francisco, CA

Senior Software Engineer, AI *June 2025 – Present*

- Architected centralized multi-provider LLM execution (OpenAI, Anthropic, Vertex/Gemini, Bedrock) for retrieval, structured-output, and agent workflows, with routing, fallback/retry, cost budgets, and tracing.
- Drove eval-driven model migration using multi-judge benchmarks, human-alignment analysis, and A/B rollout with replay logging, cutting per-token cost by over 90% while holding quality — **$1.5M annualized savings**.
- Designed SQL-backed LLM usage and cost analytics across features, workspaces, and providers, turning raw logs into finance-grade reconciliation and automated model-rebalance recommendations.
- Built the real-time multimodal AI Companion 0→1 — WebRTC (Daily.co), multi-screen capture, Celery pipelines, Python/FastAPI, and TypeScript/React; owned production reliability and enterprise incident response.
- Built the paved road for shipping and operating AI features safely: regression gates, AI pre-review, PR-readiness agents, and triage agents that turn logs, telemetry, product analytics, and code context into incident dossiers.

Beek Health Remote

Data Science Consultant (Part-time) *June 2024 – Jan 2025*

- Designed and shipped a production **RAG** pipeline grounding clinical predictions in curated medical literature — document indexing, embedding retrieval, and answer-grounding — reducing false positives by **23%**.
- Optimized end-to-end model serving (batching, caching, async I/O), cutting inference latency from **22s to 1.4s** and enabling real-time use; built a disease-prediction model on patient history and biomarkers (87% validation accuracy).

Node Finance Remote

Full Stack Engineer (Part-time) *Aug 2022 – Jan 2023*

- Built the backend for Node's Wallet SDK and consumer wallet app (React, React Native, Next.js), optimizing transaction processing and data management for a seamless mobile experience.
- Grew the consumer wallet app from **300 to 5,000 monthly active users** as a core part of a small product team.

Smaran Boston, USA

Co-Founder *Oct 2020 – Jun 2022*

- Architected a GraphQL/FastAPI multi-format ingestion pipeline (text, HTML, PDF, image, video) with OCR and embedding-based semantic tagging, powering a unified semantic-search experience — built pre-GPT, before general-purpose LLMs were available.
- As sole engineer, shipped a cross-platform Vue.js PWA (offline caching, push, Web Share Target) and established CI/CD and TDD discipline to refactor safely without a review partner.

Microsoft R&D India Bengaluru, India

Software Development Engineer *Jul 2019 – Aug 2021*

- Built a **Java Spring + Apache JMeter + Docker + Azure Functions** scale-testing framework for Microsoft's OmniChannel Customer Support platform, isolating bottlenecks and gating production rollouts via performance baselines.
- Scaled the platform from **10 to 50K Daily Active Agents**; delivered Agent-Supervisor analytics (React + ASP.NET) enabling HP onboarding at 10K DAA.

Research Experience

Boston University Boston, USA

Graduate Research Assistant *Mar 2023 – May 2025*

- **SALM** — Built a framework making LLM-driven multi-agent social simulations practically feasible: cut token usage by **73%** while preserving behavioral coherence, with an attention-based memory system achieving **80% cache-hit rates** for long-horizon agent runs.
- **Agent distillation via actor-critic** — Distilled multi-agent reasoning patterns into smaller models using actor-critic networks, enabling cheaper inference without losing strategic behavior.

- **Algorithmic-platform simulation harness** — Designed a large-scale agent-based simulation of social platforms with swappable recommender engines and RL-trained agents, producing a reproducible evaluation environment for ranking and engagement policies.
- **Custom model adaptation & evaluation** — Deployed custom Llama 2/3 and BERT variants for domain-specific tasks; designed few-shot/zero-shot evaluation protocols comparing outputs against expert-coded ground truth (human-in-the-loop).

Open Source

Manas — Open-source Python framework for building LLM-powered applications: agent architectures, tool integration, task decomposition, structured outputs, and dynamic multi-agent workflows.

Pdfvuer — A PDF viewer component for Vue.js built on Mozilla's PDF.js (widely used in the Vue ecosystem).

Selected Publications & Preprints

Koley, G. & Thiruvengadam, A. (2025). "LLM Agents as Programmable Subjects: Assays and Benchmarks for Agentic Behavior and Alignment". *Preprints*.

Koley, G. (2025). "SALM: A Multi-Agent Framework for Language Model-Driven Social Network Simulation". *arXiv preprint arXiv:2505.09081*.

Koley, G. & Digrajkar, S. (2025). "A Simulation Framework for Studying Recommendation-Network Co-evolution in Social Platforms". *arXiv preprint arXiv:2512.10106*.

Koley, G., Deshmukh, J., & Srinivasa, S. (2020). "Social Capital as Engagement and Belief Revision". *Social Informatics: 12th International Conference, SocInfo 2020*. **Best Paper Award.**

Koley, G. & Rao, S. (2018). "Adaptive Human-Agent Multi-Issue Bilateral Negotiation using the Thomas-Kilmann Conflict Mode Instrument". *22nd IEEE/ACM International Symposium on Distributed Simulation and Real Time Applications*.

Skills

AI Systems: Multi-provider LLM routing & orchestration, RAG / retrieval pipelines, agentic workflows & tool use, structured outputs, model evaluation (multi-judge, human-alignment, A/B), distillation, fine-tuning, prompt engineering, cost/latency optimization

Developer Platform: CI/CD, automated regression & coverage gates, AI pre-review of diffs, internal triage agents, AI-in-SDLC tooling (Cursor / Claude / Codex)

Backend & Data: Python (FastAPI, Celery, PyTorch, Hugging Face), Node.js, GraphQL, SQL / PostgreSQL (pgvector), MongoDB, Elasticsearch, Apache Lucene, Redis, API design, data & ingestion pipelines, Java (Spring)

Frontend & Product: TypeScript, React / Next.js, real-time WebRTC (Daily.co), product analytics, customer-facing AI workflows

Infra & Observability: AWS, GCP / Vertex AI, Azure, Docker, distributed tracing, PostHog, Sentry, CloudWatch

Education

Boston University — Boston, USA
M.S. in Data Sciences (awarded Aug 2025) — *Sep 2021 – Jun 2025*
Ph.D. program, Computing and Data Sciences — completed research, departed with M.S.

International Institute of Information Technology, Bangalore — Bengaluru, India
M.Tech. in Information Technology; Minor in Data Science — *Aug 2014 – Jul 2019*
B.Tech. in Information Technology