

Vimalkumar Patel

Agentic AI Platforms | LLM Evaluation | Large-Scale Compliance Infrastructure

New York, NY | +1 312-459-8254 | vimalkumar.patel1987@gmail.com | <https://linkedin.com/in/vimalkumarpatel>

SUMMARY

Software engineer with 15+ years of experience, including 7 years at Meta, building regulatory compliance and agentic AI platforms used by 2,300+ engineers. Owns technical direction for LLM evaluation and observability infrastructure, delivers 90%+ performance improvements on core systems, and removes thousands of hours of manual operational work annually. Operates across engineering, product, UX research, and program management to align multi-team initiatives.

TECHNICAL SKILLS

Languages: Python, Java, Hack, PHP, JavaScript, SQL, Shell

AI and LLM Engineering: Agentic AI workflows, AI agents, LLM evaluation, LLM-as-a-Judge, prompt engineering, LLM observability, continuous evaluation pipelines, drift detection, ground-truth dataset design

Frameworks and APIs: Spring Boot, Spring MVC, React, GraphQL, REST APIs, Microservices, Hibernate, JPA, JUnit, Mockito, WireMock

Infrastructure and Tools: AWS, Docker, Jenkins, Git, Mercurial, Maven, SonarQube, Splunk, Postman, Jira

Data: SQL, NoSQL, PostgreSQL, Oracle, query optimization, caching strategies

Practices: System design, distributed systems, API and platform architecture, technical roadmapping, SLA modeling, capacity planning, performance engineering, SLI/SLO definition, mentorship, Agile

EXPERIENCE

Meta | Senior Software Engineer | New York, NY | Aug 2019 - Aug 2026

AI Agent Platforms and LLM Evaluation (2024 - 2026)

- Built the **LLM observability and continuous evaluation pipeline** that gates regulatory AI responses, implementing drift detection through a hybrid topology of code-based and **LLM-as-a-Judge** graders.
- Deployed **AI agents** that automated Regulatory Inquiry Response workflows end to end, **cutting manual response processing to under 15 minutes** per inquiry.
- Proposed and implemented a **multi-provider LLM architecture** across Google, Anthropic, and OpenAI models for every agent in the Regulatory Inquiry workflow, replacing a single-provider dependency that hard-failed on upstream outages and quota exhaustion; **absorbed 2 provider incidents in its first month**.
- Pre-qualified fallbacks by sweeping **23 temperature and token-limit configurations across 6 models and 3 providers**, scored against ground truth authored by Legal Compliance partners; added **SLI instrumentation** that surfaced provider degradation for operator-triggered failover, plus concurrency controls preventing quota overdraw on shared API keys.
- Engineered **access-controlled ground-truth datasets** and a standing validation gate that scored them against their source corpus, using edit distance and LCS to classify mismatch type and **LLM-as-a-Judge** to grade severity; separating formatting drift from substantive error delivered a **2x improvement in dataset quality**.
- Shipped parallel multi-file processing and abstracted file-format handling, **expanding end-to-end agentic workflow coverage by 29%** across non-standard enterprise payloads.
- Built a unified review inbox for regulatory inquiries that **resolved over 50% of cases within 14 days**, beating the team target.
- Ranked in the **top 2% company-wide** for AI tool adoption; built an AI-powered weekly recap feature for internal bots.

Compliance Workflow and Scheduling Platforms (2019 - 2024)

- Architected and shipped the **Federation Platform Scheduler**, replacing manual regulatory compliance task creation across **65K+ employees and 2,300+ on-call rotations** with priority, availability, and SLA-constrained dispatch; delivered **7-day advance visibility**, eliminated **2 dedicated on-call engineers per shift** and **2 engineer-weeks of manual work every month**. Optimized its greedy heuristic, which replaced a generalized datacenter load-balancing service, into code the team could debug and own.
- Designed the **federation contracts** that partner teams conform to when federating tasks onto the core platform, enabling multiple teams to execute work in parallel.
- Built **task ownership**, a core federation platform module that reconciles static assignment against time-varying rotation state to resolve each task to its accountable owner, whether an individual, an on-call rotation, a team, or a shared queue; reached **61.1% adoption with 47K+ ownership confirmations** and eliminated manual triage overhead.
- Built the queue management system for the **Mobilizer federation platform**, which handles **67K+ compliance tasks per month**, improving query performance by **95% (28s to 800ms)** and cutting memory footprint by **90%**.

- Built an operations platform for bulk compliance workstream management that reached **100% organizational adoption** and eliminated **52+ hours of manual overhead monthly**.
- Redesigned the company-wide task reassignment workflow, cutting annual operational overhead **from thousands of hours to roughly 400** and reducing per-manager effort from hours to **under 2 minutes** per request.
- Defined and rolled out **SLA commitments** across the platform, improving task owner discovery by **3 days** and first-action time by **5 days**; a companion ownership accuracy initiative saved roughly **530 hours of task transfer time monthly** and improved time-to-last-owner by **8%**.
- Built an extensible **notifications framework** with engagement tracking; in a pilot it reached **30%+ engagement**, lifted assignment ownership confirmation **6.7%**, and raised task completion **3.8%**.

Technical Direction, Performance, and Mentorship

- Set technical direction for **2 of 3 orphaned products** the team inherited in H2 2025, sunsetting Insights and driving **Federation Goals** to a KTLO-ready handoff that **freed 3 engineers** onto agentic AI work.
- Modernized **Federation Goals** ahead of a framework deprecation that would have silently corrupted half-yearly compliance planning across every Meta org; migrated **25 search filters** to current standards, cut **page latency 93.8% (9.4s to 580ms)** and **user dissatisfaction 93.3% (149 to 10 users)**, and resolved GraphQL out-of-memory failures under load (**P99 12.6s to 2.2s, memory down 87%**).
- Ran the **Engineering Excellence maturity program** for a team of 10-15 engineers, and authored the **technical roadmap** for insights tooling that partner teams executed against.
- Founded an **internal technical community** for the org, running **15+ monthly brown bags** on agentic AI, performance tuning, and architecture patterns.
- Mentored **4 interns** end to end, scoping their technical roadmaps and running calibration cycles; onboarded **5+ senior engineers** across teams and contributed to hiring interviews.
- Resolved **4 high-severity production incidents** and defined **SLIs and SLOs** for 3 product surfaces with automated alerting.

Target | Lead Software Engineer | San Francisco, CA | May 2019 - Aug 2019

- Established code quality standards and led API design, integration, and testing using JaCoCo, SonarQube, and Spring Cloud Contracts.

Wells Fargo | Senior Software Engineer, AI Chatbot | San Francisco, CA | Nov 2017 - May 2019

- Engineered a **12.5x performance improvement** for a core conversational AI platform, **cutting latency from 5+ seconds to 400ms** via an in-memory database and an asynchronous batch-backup transcript architecture.
- Built the **extensibility model** for the conversational platform, a modular framework for injecting new machine learning models, intent mapping, and multi-domain banking APIs.
- Implemented a caching layer across the conversation stack that **cut intent resolution time by 50%**.
- Designed the **system integration layer** in Spring MVC, orchestrated external API connections that embedded core banking systems into conversational AI platforms, and standardized RESTful API specifications using YAML and Swagger.

Capital One | Senior Software Engineer, Online Account Opening | San Francisco, CA | Mar 2016 - Oct 2017

- **Increased customer approvals 33%** by optimizing identity matching pipelines and scale-testing integration layers with external vendors.
- Led a **database modernization initiative** using the AWS Schema Conversion Tool to migrate transactional schemas from **Oracle 11g to PostgreSQL**, and prototyped decomposition of legacy monoliths into **microservices** with Spring Boot and Docker.
- Integrated an event-driven audit and logging framework using distributed JMS queues, SLF4J, and Splunk, capturing cross-layer errors and audit-trail events for **KYC regulatory compliance**.
- Scaled automated test coverage **from 45% to 90% in 2.5 weeks** using JUnit, Mockito, WireMock, and SonarQube.

Wells Fargo | Java Developer, Shared Enrollment Business Services | San Francisco, CA | Nov 2014 - Mar 2016

- Designed, implemented, and integrated REST APIs and XDS interfaces across Wells Fargo products using Spring MVC, JMS messaging, BPM tools, SQL, and PL/SQL stored procedures.

Yelp Inc. | Software Programmer, Security and PII Leak Detection | San Francisco, CA | 2014

CME Group | Intern, Post-Trade Technology and QA Tools | Chicago, IL | 2013

Accenture | Senior Programmer, Telecom OSS for Telekom Austria | India | 2010 - 2012

Open Strides | Programmer, Java/J2EE CRM | India | 2009 - 2010

EDUCATION

Master of Science (M.S.), Computer Science - University of Illinois at Chicago | May 2014

Bachelor of Engineering (B.E.), Computer Engineering - Sarvajani College of Engineering and Technology | 2005 - 2009