Vishalan Gharat

Engineering Leader, Applied AI | Agentic AI Systems, Evaluation, MCP, AI Platform and Governance

BrowserStack | Ex-Morgan Stanley | Ex-Oracle | IIM Ahmedabad (AGMP)

Mumbai, India (Remote) | +91-8898741571 | vishalangharat@gmail.com | linkedin.com/in/vishalan | github.com/vishalan | vishalan.com

SUMMARY

Applied AI engineering leader with 10 years in software and 3 years leading AI teams, taking agentic AI systems and LLM agents from proof of concept to production deployment across 4 BrowserStack products. Owns a multi-provider LLM budget of about $1M a year (OpenAI, Anthropic, Google Vertex AI, AWS Bedrock) and makes model choices on evals, cost and latency trade-offs tied to business impact. Works cross-functionally with product, security and enterprise customers; led ISO/IEC 42001 certification and a customer-facing AI audit for a global bank.

KEY AI PROJECTS

- **Agent-ready platform and MCP (AI Transformation).** Wrote the engineering spec for making BrowserStack's products usable by AI agents: a reusable harness and integration patterns over product APIs with agent identity, deny-by-default guardrails, an MCP contract registry and a model-agnostic gateway. Set governance for the central LangGraph agent platform (5 review gates, 4 risk tiers, OWASP MCP and Agentic Top 10); took over the MCP server and grew it to 18+ tools across 4 products with remote MCP (OAuth).
- **Percy visual review agent (multimodal LLM, GA Oct 2025).** Ran a model bake-off on evals, cost and latency (GPT-5-mini at 82% coverage and 85% accuracy vs 76% and 75% for Gemini 2.5 Flash), a 10% to 100% canary rollout, and a provider switch that cut daily inference cost by roughly 75-90%. In the quarter after GA, agent adoption rose from 86% to 91% and product ARR more than doubled.
- **Low-code test automation agent.** Drove the pivot to a visually grounded browser agent built on the open-source Magnitude framework, validated on an LLM-as-judge testbed: AI step stability 85-87% to 98%, first-action latency 25s to 14s, three month customer retention 61% to 79%; shipped India data residency for a BFSI customer.
- **LLM platform review and gateway.** Audited the AI platform from source and found 5 paths to model providers with only 1 governed; ran a build-vs-buy review of LLM gateways (OpenRouter, LiteLLM, Portkey, Helicone) and launched an agent engine squad (planned at 9 engineers) to own inference, guardrails, memory and the provider gateway.
- **Test management AI (RAG over structured and unstructured data).** Semantic test selection on embeddings and pgvector, shipped as a Jira app (put through the patent process); an eval-first PRD-to-test-case pipeline (LLM-as-judge) whose prompt iterations raised coverage from 83.4% to 90.4%.

PROFESSIONAL EXPERIENCE

BrowserStack | Mumbai, India (Remote)

Senior Engineering Manager, AI | *Apr 2026 - Present*

- Lead AI engineering for the AI Features org (low-code test automation, Percy visual testing, the agent engine squad); led Phase 1 of merging the accessibility rule engine, owned by ~26 engineers, into the AI org.
- Own the org's LLM budget and vendor strategy: provider-level usage tracking, revenue per AI dollar by product, and volume discounts negotiated with cloud model providers while avoiding single-provider lock-in.
- Prototyped and shipped the AI Spend Console hands-on (cost analytics): Anthropic, OpenAI and Vertex AI spend across 14 datasets and 13 months, role-based access and a statistical anomaly digest, queryable through MCP and API.
- Led ISO/IEC 42001 certification of the AI management system, from gap assessment and risk register to a pass (Jun 2026).

Engineering Manager, AI Platform and AI Features | *Oct 2023 - Mar 2026*

- Built the AI squads' operating model from scratch in the CTO's Office: team types and growth stages, a RACI across model, platform and feature layers, GA criteria for AI features, and production-readiness gates (alerts, health metrics, feature fencing, uptime definition).
- Helped scale the AI org from 12-15 to 50-60 engineers in 12 months, adding 3 EMs; designed the ML system design interview round and an AI onboarding curriculum (LLMs, RAG, MCP) taken by 35+ engineers.
- Ran LLMOps for production agents: multi-provider fallback, feature-level rate limiting, event capacity planning and provider switchover during an AWS outage; my team shipped an RCA agent on Claude Sonnet 4 in a week.
- Led AI security and compliance: SOC 2 requirements for the AI platform, the first internal AI pentest (OWASP LLM Top 10), Zero Trust for AI repos and data, a customer-facing AI audit for a global bank, and data-processing records.
- Set the AI org's first engineering OKRs and an AI-assisted engineering program (est. 25% efficiency gain, 35% AI-written code over 56 tickets); a Claude Code cohort prototyped a cloud browser in 2 days vs an 8-week estimate.
- Grew leaders: mentored 3 peer EMs and grew two engineers into squad leads (one earned a CEO Hall of Fame award and promotion); manager eNPS 80+ in most cycles.

BrowserStack | Mumbai, India

Team Lead, Live Product Experience | *Apr 2021 - Sep 2023*

- Led the squad owning Live (real-device testing, the highest-usage product), the shared MySQL RDS fleet and free developer tools, with 100% OKR completion and every customer SLA met.
- Designed company-wide exception routing via CODEOWNERS: alerts down 85% (118 to 17 in consecutive two-month windows) and MTTA from about 35 hours to 1h 40m; built an AutoOps bot for automated alert diagnosis.
- Ran the MySQL 5.6 to 5.7 upgrade benchmark by replaying a week of production slow queries (about 120K queries an hour); cut disk IOPS 30% through tuning instead of provisioned IOPS at 3-4x the cost.
- Gave a large US bank browser-based Mac access in a day and 4 dedicated Macs in four weeks, averting churn.
- Cut infrastructure cost 35% for free developer tools and 12% for Live Product Experience.

Senior Software Engineer, Live (real-device streaming) | *Dec 2019 - Apr 2021*

- Co-led the macOS nodeRTC rollout (12% faster connect time, fewer reconnects for 12.7% of sessions) and built the socket-based Android screen capturer for the WebRTC m89 upgrade.
- Upgraded streaming to WebRTC m78 (+20% FPS, -40% interaction freezes) and added multi-tunnel TURN failover.
- Ops owner for Live within two months; led engineering for a new Pro pricing plan (new-purchase revenue +9-14% in test).

Morgan Stanley (investment bank) | Mumbai, India

Senior Associate, Data Strategy, Private Wealth Management | *Jan 2019 - Jan 2020*

- Architected a data fabric on Apache Ignite over an RDS source of 100M+ transaction records, cutting average API response time from 1.7s to 0.5s.
- Delivered a sales-process orchestration dashboard (Angular, MongoDB, Spring Boot) with a team of 4 that automated 95% of manual activities.

Oracle | Mumbai, India

Applications Developer 1 and 2, Oracle Banking Platform | *Aug 2016 - Dec 2018*

- Built a decision-tree credit-score prediction model on 20+ customer attributes with 86% accuracy, halving third-party credit API calls and cost.
- Led a team of 3 to deliver a pluggable REST framework serving 3,000+ core-banking domain services.
- Delivered a credit monitoring dashboard on Oracle Data Integrator and OBIEE aggregating 1B+ data points across limits and collateral management.

SKILLS

- **LLM applications and agents:** large language models (LLMs), generative AI, NLP, agentic workflows, tool calling, Model Context Protocol (MCP), retrieval-augmented generation (RAG), embeddings and vector search (pgvector), multimodal models, prompt engineering, browser and computer-use agents, LangGraph
- **Evaluation and LLMOps:** evaluation frameworks and golden sets, LLM-as-judge, model bake-offs, canary rollouts, production deployment, guardrails, prompt tracing (Braintrust), LLM gateways, provider fallback and rate limiting, inference cost and latency optimization, AI observability
- **Models and platforms:** OpenAI GPT, Anthropic Claude, Google Gemini and Vertex AI, AWS Bedrock, Azure OpenAI; evaluated open weight small language models (Qwen, DeepSeek, Gemma, Phi); self-hosted local LLM inference
- **AI governance and security:** responsible AI, ISO/IEC 42001, SOC 2, OWASP LLM and MCP Top 10, Zero Trust, AI risk and impact assessment, data residency
- **Engineering:** Python, Java, JavaScript/Node.js, Ruby on Rails, Spring Boot, cloud architecture (AWS, GCP), Docker, Kubernetes, Kafka, Redis, MySQL, PostgreSQL, MongoDB, BigQuery, Apache Ignite, WebRTC, distributed systems
- **Leadership:** AI strategy and transformation, organizational AI adoption and change management, cross-functional leadership, org design, scaling teams and hiring, OKRs, LLM budget and vendor negotiation, executive stakeholder management, agentic coding with Claude Code

EDUCATION

Accelerated General Management Programme, IIM Ahmedabad | 2023 - 2024

Bachelor of Engineering, Computer Science, University of Mumbai | 2012 - 2016

CERTIFICATIONS AND AWARDS

- Deep Learning Specialization (DeepLearning.AI); Machine Learning (Stanford Online)
- Guided 4 AI inventions from my teams, incl. AI test selection, through BrowserStack's patent process, from disclosure to review with patent counsel (2024)
- BrowserStack Hackathon winner (2023); Performance Excellence Awards at BrowserStack (2020) and Oracle