

Nilay Baranwal

nilaybaranwal@gmail.com | github.com/alphasr | linkedin.com/in/nilaybaranwal | Prague, Czech Republic

Forward Deployed Engineer | AI Engineer

Full stack and AI engineer with 7 years of experience embedding with users, turning a vague problem into a working system, and keeping it running in production. Sat with traders at Barclays to design and ship a React/GraphQL pre-trade dashboard they use daily, and now building Hive Mind, an internal MCP server that connects AI agents to the bank's pre-trade systems; alongside that, built four AI products from zero as side projects, deploying each into real operating environments: Prague nightclubs on event night, marketing teams inside their own analytics stacks, and short-term rental operations run by AI agents. Builds agentic systems end to end: agent harnesses, tool and MCP design, memory and context engineering, evals and guardrails, and the integrations that connect them to the customer's data. Comfortable where the work is customer-facing, integration-heavy and underspecified: scoping onsite, prototyping in days, then folding what was learned back into the product. Deep in TypeScript/React/Next.js, Node.js, Python, Java Spring, and cloud-native architectures.

Professional Experience

Barclays Investment Group – Senior Full Stack Software Engineer, May 2021 – Present

Markets Pre-Trade Analysis (embedded with the trading desk)

- Building **Hive Mind**, Barclays' internal Model Context Protocol (MCP) server that gives AI agents and assistants governed access to pre-trade tooling, internal market-data services and team knowledge, so desks plug their agents into one service instead of building bespoke integrations
- **Auth:** OAuth 2.1 with PKCE against the corporate identity provider (SSO), short-lived per-user tokens propagated end to end so every tool call runs under the caller's own entitlements, least-privilege tool scopes, no shared service credentials, and a full audit trail of who called which tool, with what arguments and result
- **Harness and tool design:** MCP tool annotations (readOnlyHint, destructiveHint, idempotentHint, openWorldHint) driving the agent harness, so read-only tools run freely while state-changing calls hit human-in-the-loop approval gates; typed JSON Schema inputs and outputs, structured errors agents can recover from, and pagination and response-size limits to protect the context window
- **Principles:** read-only by default, deny by default on anything unannotated, stateless and idempotent operations, versioned tool contracts, a thin MCP layer over the same service layer the UI uses, OpenTelemetry tracing per tool call, and an eval and regression suite checking tool selection and argument accuracy before each release
- **Stakeholder and customer interaction:** primary technical point of contact for traders, desk heads and business stakeholders; run regular demos, discovery and feedback sessions, translate trading workflows into technical requirements, manage expectations on scope and timelines, and present trade-offs and release plans to non-technical audiences
- Embedded directly with traders and business stakeholders to spec and lead a React/GraphQL pre-trade dashboard, running requirement sessions on the desk and shipping iteratively against live feedback, improving decision speed by 40% and delivering \$2M in annual cost savings
- Diagnosed and fixed a critical market-data bottleneck reported by users mid-session: optimized queries and introduced a Redis caching layer, cutting API response times by 60% and directly improving the trader experience
- Owned the feature end-to-end in a regulated environment: integration with internal market-data services, release, monitoring and frontline support for the desk

TEDxKanke – Lead Full Stack Developer (Remote), May 2019 – February 2022

Event Management Platform Development

- Acted as the sole technical partner to a non-technical organising team: defined product scope, built and launched a full-stack event platform from scratch, growing registrations by 250% and significantly increasing organic reach

Nationale-Nederlanden – Junior Software Engineer, May 2018 – August 2018

PeopleSoft HRMS Development

- Delivered 8 PeopleCode modules for the HR management system serving 3,000+ employees across European regional offices, gathering requirements directly from regional HR teams
- Optimized Oracle SQL queries, improving report generation performance by 40% & reducing database load during peak usage

Commerzbank GMBH – Junior Software Engineer, January 2018 – April 2018

Core Banking Applications Development

- Contributed to Spring Boot banking application development serving 100,000+ retail & corporate banking customers daily
- Established Git workflow best practices, reducing merge conflicts by 50% & improving team collaboration across 15 developers

Education

Czech Technical University in Prague	B.Tech Software Engineering	October 2016 – June 2022
--------------------------------------	-----------------------------	--------------------------

Side Projects

Social Line: AI Social Listening Platform www.socialline.app, March 2026 – Present

- Built and shipped an AI social listening SaaS end-to-end: tracks brand, product, competitor and campaign keywords across Facebook, Instagram, X, LinkedIn, Reddit, YouTube and TikTok, with pre-scored mentions landing in one feed in under two minutes
- Designed the scoring and triage pipeline (LLM sentiment classification, estimated reach, auto-categorisation into customer-defined categories, and a reputation score) that turns a raw mention stream into a queue ranked by reach at risk
- Shipped a Model Context Protocol (MCP) server over OAuth 2.0 and scoped API keys so customers' existing AI agents drive keywords, mentions, triage and runs through the same stateless service layer as the dashboard and REST API, with no bespoke integration to maintain per account
- Ran the full loop myself: onboarding calls, pricing and packaging across four tiers, a no-signup interactive demo built to shorten the evaluation cycle, and public-sector outreach to the Government of Bihar on social listening for flood and relief management

AI-Native Short-Term Rental Operations: Agentic Property Management, September 2026 – Present

- Building an AI-native short-term rental management operation in which AI agents run guest messaging, review responses, issue triage, cleaning and maintenance dispatch, and owner reporting, targeting roughly half the fee of a traditional property manager
- Architected the agent layer on top of a PMS (Beds24) as system of record: event-driven via PMS webhooks with shared-secret validation, tool calls against reservations, calendars, messaging and pricing APIs, and a rate-limit-aware scheduler for the API credit budget
- Designed human-in-the-loop escalation and approval gates for high-risk actions (refunds, pricing changes, guest disputes), with per-property memory of house rules, SOPs and guest history feeding each agent's context
- Scoped a custom owner portal with multi-property sub-accounts, revenue reporting and payout calculations, plus GDPR and EUR/CZK localisation for a Prague pilot; evaluated Hostaway, Guesty and Calry's unified PMS API as the path to serving other operators

Orvio.ai: Generative Engine Optimization (GEO) Platform orvio.ai, August 2025 – October 2026

- Co-founded Orvio Labs s.r.o. and built the product end-to-end: defined scope, shipped the Next.js MVP, and grew it into a platform that gets businesses cited in ChatGPT, Gemini, Perplexity and Google AI Overviews
- Deployed into each customer's own stack (GA4, Search Console and CRM connectors) to attribute AI citations and exposure back to pipeline, so buyers could see measurable impact rather than a vanity score
- Built an empirical source-graph engine measuring citation frequency by domain across a frozen prompt panel, a defensible alternative to competitors' unverifiable composite visibility scores
- Led enterprise scoping and delivery directly with clients on multi-market AI-visibility programmes and published case studies, and ran the GDPR groundwork for the crawling pipeline (legitimate interest assessment, Art. 14 transparency, DPIA, source-URL provenance)

Partymaker: Nightlife Ticketing & Event Operations partymaker.app, January 2025 – Present

- Co-founded and built a ticketing and event operations platform for promoters running 500–3,000 capacity events: checkout, door check-in, guestlist/friendlist management and organiser expense settlement
- Deployed on-site on event nights, running offline-tolerant QR door scanning, real-time roster sync across scanners, and live triage with door staff, then fed the failure modes observed at the door straight back into the roadmap
- Shipped an automated payout and settlement layer reconciling ticket revenue and shared costs per event across currencies; designed a cashless drink-token system and modelled commission, surcharge and top-up pricing against incumbents (Weezevent/Weeztix), aligned to PSD2 surcharging rules and agent-vs-principal VAT treatment
- Exposed the platform as an MCP server so organisers run guestlists, expenses, tasks and payout rollups conversationally through an AI agent; live with Prague venues and labels including Ankali and Harmony Rec

Automated Crypto Trading Platform, August 2023 – February 2024

- Engineered a Solana trading bot with 92% success rate processing 500+ daily blockchain transactions via Web3.js, with secure OAuth authentication and real-time Firebase sync supporting 100+ concurrent active traders

SkillNavigator: Professional Services Marketplace, August 2023 – September 2025

- Developed an Angular marketplace connecting 1,000+ users with geo-location matching and real-time chat on a serverless Firebase architecture, cutting infrastructure costs by 85% at a 99.9% uptime SLA

Skills and Tools

- **AI Engineering & Agents:** LLM application development (Anthropic/Claude, OpenAI, Gemini), agent harnesses and agent loops (planner/executor, subagents, tool routing), multi-agent orchestration, Claude Agent SDK, OpenAI Agents SDK, LangChain/LangGraph, Vercel AI SDK, Model Context Protocol (MCP) server and tool design (enterprise and customer-facing), structured outputs & function calling, human-in-the-loop approvals and guardrails
- **Memory, Context & Retrieval:** agent memory systems (short-term, long-term and persistent file-based memory, summarisation and context compaction), context engineering, prompt engineering, RAG, embeddings, vector search (pgvector, Pinecone), hybrid search and reranking, chunking strategies
- **Evals & LLM Ops:** eval harnesses (golden datasets, LLM-as-judge, regression suites), output scoring, LLM observability and tracing (Langfuse, LangSmith), prompt versioning, cost and latency optimisation (prompt caching, model routing, batching), sentiment analysis & NLP classification, AI workflow automation (n8n), Generative Engine Optimization (GEO/AEO), AI-assisted development (Claude Code, Cursor)
- **Forward Deployment & Solutions:** customer discovery and onsite scoping, rapid prototyping, technical pre-sales and demos, solution architecture, customer data integration, onboarding and enablement, production debugging with users in the room, incident triage and support, requirements translation between business and engineering, working solo in ambiguity
- **Product & Engineering:** full stack development, 0 → 1 delivery, JavaScript, TypeScript, Java, Python, React, Next.js, Angular, Node.js, Tailwind CSS, RESTful APIs, GraphQL, webhooks, microservices, SQL/NoSQL (PostgreSQL, Redis, Firebase), system architecture, API integration
- **Cloud & Platform:** AWS, Google Cloud, Azure, Vercel, Railway, CI/CD, Docker, Kubernetes, Infrastructure as Code, serverless, observability & monitoring, test automation, unit/integration testing, deployment automation
- **Payments, Data & Compliance:** Stripe & payment orchestration, checkout and payout flows, PSD2 surcharging, VAT agent-vs-principal treatment, GDPR (LIA, DPIA, Art. 14 transparency), GA4 & Search Console, CRM connectors, PMS APIs (Beds24, Hostaway, Guesty), Web3.js/Solana
- **Delivery & Leadership:** Agile, Scrum, feature prioritisation, pricing & packaging, stakeholder management, technical leadership, team mentoring, roadmap execution, technical documentation
- **Languages:** English (fluent), Hindi. Based in Prague, EU work authorisation, available to travel to customer sites