

Slobodan Milic

FULL-STACK + AI ENGINEER | REACT, TYPESCRIPT, PYTHON, LLM PRODUCT FEATURES

Available

Belgrade, Serbia

SKILLS

Python, JavaScript, TypeScript, Java, Go, C#, PHP, React, Vue, Tailwind CSS, Storybook, FastAPI, Node.js, Express, NestJS, LLM APIs, RAG, embeddings, vector search, eval harnesses, PostgreSQL, Docker, CI/CD, feature flags, inference monitoring, access control, tool-calling APIs, webhooks, accessible AI UI patterns

ABOUT

Full-stack engineer shipping AI-assisted product features inside existing B2B apps: copilots, summarization, extraction, and doc search from React UI through Python and Node APIs.

Defines eval harnesses, review queues, and rollout guardrails before LLM features reach production users. Vector stores, prompt versioning, and cost dashboards evolve alongside regular product releases.

End-to-end ownership from schema and API contracts through accessible UI, with AI features shipped on the same cadence as core product work.

EXPERIENCE

Senior Software Engineer · HTEC

Mar 2022 – Present

Belgrade, Serbia

- Shipped React copilot panels embedded in 18 admin workflows, serving 420 weekly LLM-assisted queries with review queues and audit logs.
- Shipped FastAPI and Node.js RAG endpoints over 320 internal runbooks with p95 latency under 380 ms and tenant-scoped access control.
- Rolled out TypeScript feature flags routing 12% of users through AI summarization with offline eval gates before full launch.
- Integrated vector search into 5 self-serve dashboards, cutting analyst lookup time by 40% versus keyword search alone.
- Coordinated schema, API, and UI releases for 9 AI-assisted export flows with contract tests on every deploy.

Software Engineer · Execom

Jun 2019 – Feb 2022

Novi Sad, Serbia (Remote)

- Released customer-facing React portal features with embedded ticket summarization for 3,200 accounts across 11 countries.
- Prototyped Python FastAPI extraction services lifting auto-routing accuracy from 62% to 79% on 400 monthly dispute forms.
- Wrapped OpenAI calls behind caching layers cutting duplicate-token spend by 34% across 3 internal admin tools.
- Shipped Node.js and Rails APIs unified with PostgreSQL read models powering 120K daily calls and 3 LLM widgets.
- Authored prompt regression suites and human-review paths before 2 LLM search widgets reached 900 queries/week.

Software Developer · Levi9

Jan 2017 – May 2019

Belgrade, Serbia

- Introduced React commerce admin with AI-assisted catalog tagging suggestions adopted by 4 merchandising squads during peak season.
- Integrated Python recommendation hooks into PHP checkout flows serving 2.1M monthly sessions with A/B guardrails.
- Shipped Vue operational dashboards with natural-language filters over 85K SKU rows, p95 loads under 1.2s.
- Added TypeScript API layers connecting 4 storefront event streams to embedding pipelines without blocking checkout latency budgets.
- Partnered with data science on offline eval harnesses before promoting 3 ML-driven product sort experiments.

Junior Systems Engineer · Comtrade

Aug 2013 – Dec 2016

Belgrade, Serbia

- Maintained internal React tools with doc-search copilots over 1,400 runbooks, shortening onboarding for 30+ developers.
- Released lightweight Python FastAPI assistants for 2 DNS and certificate workflows with structured audit trails.
- Integrated LLM-powered ticket triage into 3 Ionic field apps with human override on every high-risk classification.
- Shipped Node.js webhook handlers feeding embedding indexes used by 6 support squads during incident response.
- Documented prompt versioning and rollback playbooks adopted before 2 production AI pilots reached hosting operations teams.

EDUCATION

University of Belgrade · MSc, Software Engineering

Sep 2012 – Jun 2014

University of Belgrade · BSc, Computer Science

Sep 2008 – Jun 2012

PROJECTS

Product Copilot Shell

React

Launched a React sidebar copilot with streaming answers, citation chips, and role-scoped retrieval over 320 runbooks adopted by 18 admin workflows. Authored chunking rules, tenant filters, and citation scoring for runbook retrieval serving 420 weekly power-user queries at sub-400 ms p95. Added human-review queues and prompt regression suites blocking 8% of low-confidence answers before the governed-export pilot reached production.

AI Feature Rollout Console

TypeScript

Released a TypeScript feature-flag console with staged cohorts, offline eval gates, and one-click rollback for 3 summarization experiments in admin workflows. Integrated token and latency dashboards for 3 LLM providers so product owners could compare cost per successful task weekly. Wired contract tests across 9 AI-assisted export flows so schema, API, and UI changes shipped in the same release train.

Support Workflow Summarizer

React

Composed a React triage sidebar that condensed 12-field support forms into 3 reviewable summaries for 420 account managers during a 6-week pilot. Instrumented per-tenant token budgets and weekly spend alerts so product owners could cap 3 pilot widgets before a broader admin rollout. Drafted reviewer override flows and exportable prompt-version ledgers satisfying audit requests during a 10-week compliance review window.