

Evgenii Perminov

APPLIED AI / RETRIEVAL ENGINEER

RAG systems · Hybrid retrieval · Evaluation · Distributed backends

perminov.eugene@gmail.com	+372 5368 5244	Tallinn, Estonia
github.com/PerminovEugene	linkedin.com/in/Eugene-Perminov	evgeniiperminov.dev

PROFESSIONAL SUMMARY

Software engineer with 11 years in production systems, specializing in retrieval and LLM applications. Eight of those years were spent on Node.js and TypeScript distributed backends, including a roughly 100-service Kafka platform on PostgreSQL. I build ingestion and chunking pipelines, embeddings, hybrid dense and BM25 search, reranking, and the evaluation harnesses that measure them, and I own the Kubernetes, GitOps and observability the services run on.

CORE SKILLS

- AI & Retrieval:** RAG, ingestion pipelines, chunking strategies, embeddings, dense retrieval, BM25, hybrid retrieval, RRF, DBSF, weighted score fusion, cross-encoder reranking, retrieval evaluation, Recall@K, MRR, nDCG, Qdrant, Ollama, OpenAI API, LangChain, LangGraph, MCP
- Backend:** Node.js, TypeScript, Python, NestJS, Express, FastAPI, REST, GraphQL, gRPC, WebSockets, microservices, event-driven architecture
- Data & Messaging:** PostgreSQL, Redis, MongoDB, Kafka, Debezium, RabbitMQ, Dramatiq, SQLAlchemy, TypeORM
- Infrastructure:** Kubernetes, Docker, Docker Compose, Argo CD, GitOps, GitHub Actions, AWS, Prometheus, Grafana, Loki, Tempo
- Frontend:** React, Next.js, Vite, Redux, Zustand, Tailwind, Material UI, Svelte

PROFESSIONAL EXPERIENCE

Principal Software Engineer

Clairbytes OÜ | Tallinn, Estonia | Feb 2025 – Present

- Offer Inspector** – SaaS, in production
- Built and launched Offer Inspector, a full-stack salary comparison SaaS using Next.js, React, NestJS, PostgreSQL and Go workers, covering authentication, billing, tax logic, localization and production infrastructure.
 - Designed an AST-based tax engine for salary calculations across Estonia and Poland, supporting multiple legal forms, net-to-gross calculations, employer costs, paid leave and effective salary comparisons.
 - Set up GitOps-based deployment on a self-hosted k3s cluster with Argo CD, blue-green releases, automated database migrations and encrypted secrets.
 - Added application and infrastructure observability using OpenTelemetry, Prometheus, Grafana, Loki and Tempo, including database metrics, tracing, dashboards and Slack alerts.

- Embeddorium** – open-source retrieval-engineering workbench
github.com/PerminovEugene/Embeddorium
- Built a distributed ingestion pipeline with Dramatiq, RabbitMQ, PostgreSQL outbox events, idempotent jobs and bounded retries.
 - Separated canonical datasets from indexing pipelines, allowing the same corpus to be tested with different chunkers and embedding models while keeping experiment configurations reproducible.
 - Implemented dense, BM25 and hybrid retrieval with RRF, DBSF, weighted score fusion and optional cross-encoder reranking.
 - Built a chunker-independent evaluation system using document character spans, with Recall@K, MRR, nDCG, Hit Rate and Passage Coverage metrics.
 - Added auto-discovered plugins for parsers, chunkers, fetchers, filters and model providers.
 - Built an MCP server exposing provider, dataset preparation, indexing and staged search workflows to AI agents.
- Stack:** Python, FastAPI, Dramatiq, RabbitMQ, PostgreSQL, Qdrant, Ollama, OpenAI API, LangChain, LangGraph, MCP, React, Vite, Docker Compose

PROFESSIONAL EXPERIENCE (CONTINUED)

Senior Software Engineer

[Katana Cloud Inventory](#) | Tallinn, Estonia | Remote | Aug 2021 – Sep 2024

- Built and scaled Node.js and TypeScript services inside a roughly 100-service, Kafka-driven event platform on PostgreSQL, serving manufacturing customers worldwide.
- Re-architected manufacturing-planning calculations to improve execution time and reduce calculation errors.
- Delivered cross-team features end to end: public API gateway, webhooks, sales orders and the inventory planner.
- Tuned database schemas and queries to improve performance across key services.
- Improved CI/CD pipelines with faster test runs and more reliable deployments.
- Led technical investigations, estimation and epic planning across teams.

Senior Software Engineer

[Huawei R&D Lab](#) | Saint Petersburg, Russia | Remote | Aug 2020 – Jun 2021

- Built a VS Code-based IDE for C/C++ developers using the VS Code API, Node.js, TypeScript and Svelte.
- Implemented a gRPC remote-debugging server and the packaging and delivery pipeline for the extension.

Full-Stack Developer

[Diversido](#) | Remote | May 2018 – Aug 2020

- Built Node.js, Express and NestJS backends on PostgreSQL with React, Redux and TypeScript frontends for client products.
- Delivered GraphQL APIs, real-time features over WebSockets, and containerized deployments with Docker on AWS.
- Worked with clients and product teams to refine requirements and keep several projects delivering reliably in parallel.

Full-Stack Developer

[Ecombix](#) | Saint Petersburg, Russia | Jul 2017 – Apr 2018

- Built a TypeScript and RabbitMQ integration layer syncing the platform with Shopify, Magento and WooCommerce.
- Worked across the Angular 2 frontend and the Node.js and MongoDB services behind the integration layer.

Software Engineer

[7Bits](#), [Avelix](#), [Burning Buttons](#) | Omsk, Russia | Sep 2014 – Mar 2017

- Built a Facebook chatbot on Node.js, Express and PostgreSQL, with Stripe integrated for payment processing.
- Contributed front-end features to banking and healthcare (FHIR) products using React and Redux.

OPEN SOURCE

[messy-folder-reorganizer-ai](#)

Local-first Rust CLI that uses a local LLM and embeddings to match file names and content signals against a target folder taxonomy. Rust, Ollama, embeddings. 2024–2025.

EDUCATION

[Dostoevsky Omsk State University](#)

Bachelor's Degree in Applied Mathematics and IT | 2011–2015

LANGUAGES

English — Advanced

Russian — Native

Estonian — Beginner