

George (Gosha) Mareev

AI Engineer

RAG · Agents · On-prem LLM Platforms

Moscow (GMT+3) · hello@mareevapps.com · [GitHub](#) · mareevapps.com

PROFILE

AI engineer with commercial AI/ML experience since June 2025 on an enterprise platform team. I deploy and operate LLM systems inside a secured perimeter, build RAG pipelines and internal assistants, and develop access control, evaluation, observability and prompt-level audit. English B2.

EXPERIENCE

Global FMCG company | AI Engineer (AI/ML Solutions Engineer)

Jun 2025 - present · Moscow

- Deployed and maintain a corporate AI platform: LiteLLM as a single model gateway, Open WebUI, local and OpenAI-compatible models, isolated services and Microsoft Entra ID SSO. Adoption: 120 enterprise users, 10 teams served.
- Designed and develop internal RAG assistants: document ingestion, chunking, embeddings, metadata-aware retrieval, ranking, grounded refusals and cited answers; SharePoint synchronization through Microsoft Graph. Two domain assistants share a common core.
- Built operational log shipping via Fluent Bit into Loki and a separate prompt-audit transport to SIEM, with credential masking and Prometheus request and latency metrics.
- Implemented a configurable JSON-driven scenario system so operational priorities, diagnostics and rules can be updated without code changes or an application release.
- Maintain access control, deployment isolation and recurring answer-quality evaluation. Platform guardrail policy and SIEM detection rules are in progress.

SELECTED PROJECTS

Secure LLM Platform Reference | 2026

Runnable public reference using synthetic data: Open WebUI, LiteLLM, pgvector, scoped RAG, EN/RU privacy checks and durable audit. Core RAG checks: 32/32 in each of three runs; expanded checks: 125/128. Methodology and limitations: [versioned reports](#).

Women's Clothing Reviews | NLP analysis | 2025

Sentiment and topic modelling: TextBlob, GPU-accelerated BERTopic, interactive Streamlit dashboard and PDF reports.

Data Science Salaries | ETL and analytics | 2025

Kaggle ingestion, Airflow orchestration, PostgreSQL data mart and Superset visualizations; Docker Compose, Python and SQL.

Personal labs | mareevapps.com

Neural Bonsai: browser MLP training, pruning and model inspection in TypeScript. ApsisDB: embedded Go database for cyclic telemetry. Generative typography and a shader lab.

CORE SKILLS

Stack: Python · FastAPI · Docker · Postgres/pgvector · LiteLLM · Open WebUI · Loki · nginx · Linux

LLM & RAG: chunking, embeddings, retrieval, reranking, cited answers, evaluation, ChromaDB, FAISS, Ollama, OpenAI-compatible APIs; privacy and injection checks in the public reference.

Backend & operations: Flask, Pydantic, REST APIs, Bash, Git, Prometheus, Fluent Bit, SSO / Entra ID, SIEM audit transport.

ML & data: scikit-learn, PyTorch, BERTopic, pandas, NumPy, SQL, Airflow, Streamlit, Superset.

EDUCATION & ADDITIONAL

Moscow City University - MSc in progress, Business Analytics and Big Data; BSc, Education, major in Computer Science.

Part-time teaching: School No. 1533 "LIT", Moscow - Computer Science teacher since Sep 2023; programming, algorithms and competitive-programming coaching.

Award: 1st place, "Technologies of Success" olympiad, Photonics track - team lead.

Languages: Russian - native; English - B2.