

Max Butorin

FORWARD DEPLOYED AI ENGINEER / AI SOLUTIONS ARCHITECT

contact@maxbutorin.dev | linkedin.com/in/max-butorin | maxbutorin.dev | github.com/AbsoluteMode

OPEN TO RELOCATION | REMOTE | INTERNATIONAL

SUMMARY

Forward Deployed AI Engineer / AI Solutions Architect with 5+ years of experience turning ambiguous business problems into production AI systems.

Delivered 20+ AI solutions across LLMs, RAG, AI agents, document intelligence, speech, developer tooling, and high-load inference - from customer discovery and technical scoping through architecture, implementation, evaluation, deployment, and production iteration.

Hands-on full-stack engineer with experience building complex backend systems, web products, macOS and Android applications, as well as leading ML engineering teams. Comfortable working directly with customers and domain experts, translating business constraints into technical architecture and measurable quality gates.

Experienced in balancing model quality, latency, reliability, infrastructure cost, and delivery speed across cloud and on-prem environments.

EXPERIENCE

AI Lead · MR Group

Aug 2025 - 2026 · Remote

Major real estate developer focused on residential and commercial properties.

- Owned end-to-end delivery of Construction AI, from customer discovery and business-process analysis through technical scoping, architecture, implementation, evaluation, and production rollout; led 4 ML engineers and managed a 3-person delivery squad.
- Designed and delivered an on-prem document-intelligence platform processing 2,000+ checklist items across 30+ PDFs under a 65 GB VRAM constraint, reducing expert workload from 90-126 to 15-21 person-days per month (~6x).
- Introduced a separate validation gate with confidence-based routing, achieving ~99.9% accuracy on the high-confidence tier covering ~30% of checklist items.
- Built the production application around the AI system, including SSO, document upload, asynchronous processing, progress tracking, source-grounded answers, expert review workflows, and Telegram notifications.
- Designed and launched an AI code-review system in 3 weeks; during its first month it completed 578 reviews, identified 44 confirmed critical defects, and produced 600+ confirmed findings.
- Delivered an AI meeting assistant with transcription, diarization, speaker identification, structured summaries, task export, and conversational Q&A over meeting history.
- Worked directly with stakeholders and domain experts throughout discovery, validation, and rollout, converting operational requirements into technical decisions and measurable acceptance criteria.

Machine Learning Engineer · BP

Apr 2025 - Aug 2025 · Remote

Global integrated energy company.

- Built a RAG observability and health-monitoring pipeline for an enterprise MVP connected to ~12,000 SharePoint sources, monitoring ingestion, indexing, freshness, and context delivery.
- Designed evaluation methodology for a unified RAG platform, measuring retrieval quality, answer accuracy, and end-to-end task success.
- Created regression test suites against fixed benchmarks to validate pipeline changes before production releases.
- Identified SharePoint synchronization drift, stale indexes, and multimodal ingestion failures, improving data freshness and context completeness.

Machine Learning Engineer · VkusVill

Oct 2023 - Apr 2025 · Saint Petersburg

Grocery and food retail company focused on fresh food and online delivery.

- Designed and delivered an AI semantic search system serving 500-1,000 users and 1,000-3,000 daily queries, reducing search-to-cart time from 54 to 35 seconds.
- Optimized high-load model inference through quantization, reducing model size from 30 GB to 13 GB while reaching 0.6 s generation latency at 60 RPS on 4x A100 GPUs.
- Preserved retrieval quality after quantization with only 4% hit-rate degradation and p95 latency of 0.7 s.
- Worked across the full delivery stack, from ML architecture and inference optimization to backend integration and production operation.

Machine Learning Engineer · YooMoney

Jan 2022 - Oct 2023 · Saint Petersburg

Fintech and online payment service.

- Built an end-to-end agentic support system handling 2,000-6,000 daily requests, reducing manual workload by 20% and automating repetitive cases at 99%+ confidence.
- Developed a CRM operator-assist system that placed the correct script in the top three with 96% probability and reduced resolution time from 1.5 minutes to 37 seconds; rolled out across 3 teams.
- Delivered an AI meeting-notes assistant with Jira integration, supporting 30+ users and 50+ meetings per week.

AI PRODUCTS & CONSULTING

AI Solutions Architect / Senior AI Engineer · Independent

2025 - Present · Remote

End-to-end AI delivery across customer discovery, architecture, PoC/MVP development, evaluation, deployment, and production iteration.

- Built Whytap, an AI voice-input and meeting-intelligence product for macOS and Windows, combining real-time speech processing, terminology-aware recognition, meeting intelligence, and MCP-based access to organizational knowledge.
- Created Session Recall, an open-source memory infrastructure for Claude Code, Codex, and Cursor that builds a unified searchable index across coding-agent sessions and exposes semantic retrieval through MCP.
- Designed Storm, a multi-model software-engineering workflow coordinating several coding agents across analysis, planning, implementation, testing, debugging, and independent review.
- Migrated an LLM evaluation workload to the Batch API, reducing inference cost by 50% while decreasing answer variance from 60% to ~6% against an expert-curated golden dataset.
- Designed GPU infrastructure for speech analytics processing 50,000 minutes/day, reducing expected infrastructure cost by ~30% while preserving capacity to scale toward 100,000 minutes/day.

AI Consultant · TheNext

Sep 2025 - Present · Remote

AI product team developing a B2B trend-monitoring platform.

- Designed an LLM pipeline processing 10,000+ daily calls and generating relationship graphs for regional trend reports.
- Built an AI trend-monitoring platform analyzing 40+ public sources and translated customer workflows into AI architecture, evaluation criteria, and production implementation.

TECHNICAL EXPERTISE

AI / LLM LLMs, AI Agents, Multi-Agent Systems, RAG, Retrieval, Reranking, Context Engineering, Memory Systems, LLM Evaluation, Validation Gates, Fine-Tuning, Quantization

BACKEND & FULL-STACK Python, FastAPI, Flask, Django, TypeScript, JavaScript, REST APIs, WebSockets, PostgreSQL, Redis, Kafka, RabbitMQ

AI TOOLING PyTorch, Hugging Face, vLLM, LangChain, LangGraph, Langfuse, RAGAS, OpenAI API, Anthropic API, MCP

INFRASTRUCTURE AWS, GCP, Azure, Azure OpenAI, Azure AI Search, Docker, Kubernetes, CI/CD, Qdrant, Milvus

PRODUCT ENGINEERING Web Applications, Backend Systems, macOS Applications, Android Applications, Production Integrations

EDUCATION & LANGUAGES

B.S., Oil & Gas Engineering

Tyumen State Oil and Gas University · 2022

Russian - Native

English - Professional Working Proficiency