

Max Butorin

AI SOLUTIONS ARCHITECT / SENIOR AI ENGINEER

✉ contact@maxbutorin.dev [in linkedin.com/in/max-butorin](https://www.linkedin.com/in/max-butorin)
△ maxbutorin.dev [gh github.com/AbsoluteMode](https://github.com/AbsoluteMode)

• OPEN TO RELOCATION · REMOTE



SUMMARY

AI Solutions Architect and Senior AI Engineer with 5+ years of experience designing, building, and deploying production AI/ML systems.

Delivered 20+ AI solutions across LLMs, RAG, AI agents, document intelligence, speech, evaluation, and developer tooling. Combines business-process analysis and solution architecture with hands-on engineering, production deployment, and measurable quality gates.

Led a team of 4 ML engineers and owned architecture, delivery, evaluation methodology, code review, and stakeholder communication.

EXPERIENCE

AI Lead · MR Group

Aug 2025 – 2026 · Remote

Major Russian real estate developer focused on residential and commercial properties.

- Led a team of 4 ML engineers and managed a 3-person delivery squad for the Construction AI project, owning solution architecture, delivery, evaluation methodology, code review, and stakeholder communication.
- Led development of Construction AI — an on-prem document-intelligence platform processing 2,000+ checklist items across 30+ PDFs within a 65 GB VRAM constraint, reducing manual workload from 90–126 to 15–21 person-days per month (~6×).
- Implemented a separate validation gate with confidence-based routing, achieving ~99.9% accuracy on the high-confidence tier covering ~30% of checklist items.
- Built a production web service with SSO, document upload, progress tracking, source-grounded answers, and Telegram notifications, enabling asynchronous expert review.
- Designed and launched an AI code reviewer in 3 weeks; in its first month it completed 578 reviews, identified 44 confirmed critical defects, and produced 600+ confirmed findings.
- Delivered an AI meeting assistant: transcription, diarization, speaker identification, structured summaries, task export, and Q&A over meeting history.

Machine Learning Engineer · BP

Apr 2025 – Aug 2025 · Remote

Global integrated energy company.

- Built a RAG health-monitoring pipeline for an MVP connected to ~12,000 SharePoint sources, tracking ingestion, indexing, data freshness, and context delivery.
- Developed evaluation logic for a unified RAG platform, measuring retrieval quality, answer accuracy, and end-to-end task success.
- Created regression test suites on fixed benchmarks, validating RAG pipeline changes before release.
- Detected SharePoint sync drift, stale indexes, and multimodal ingestion failures, strengthening data freshness and context completeness.

Machine Learning Engineer · **VkusVill**

Oct 2023 – Apr 2025 · Saint Petersburg

Russian grocery and food retail company focused on fresh food and online delivery.

- Designed and delivered an AI smart-search system over product composition for 500–1,000 users and 1,000–3,000 daily queries, improving search-to-cart time from 54 to 35 seconds.
- Optimized model serving through quantization, reducing model size from 30 GB to 13 GB and generation latency to 0.6 s at 60 RPS on 4× A100.
- Preserved retrieval quality after quantization, keeping hit-rate degradation at 4% and p95 latency at 0.7 s.

Machine Learning Engineer · **YooMoney**

Jan 2022 – Oct 2023 · Saint Petersburg

Russian fintech and online payment service.

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

INDEPENDENT PRACTICE

AI Solutions Architect / Senior AI Engineer · **Independent AI Practice**

2025 – Present · Remote

End-to-end AI systems: business-process analysis, solution architecture, PoC/MVP, evaluation, deployment, production iteration.

- Built whytap — a voice-input and meeting-intelligence product for macOS and Windows: fast voice input, automatic meeting notes, terminology-aware speech recognition, and meeting knowledge exposed to internal AI assistants through MCP.
- Created Session Recall — open-source shared memory for Claude Code, Codex, and Cursor: one searchable index across coding-agent sessions, with semantic search exposed through MCP.
- Developed Storm — a multi-model software-engineering workflow running analysis, planning, implementation, review, and debugging through several coding agents in parallel, consolidated by an explicit review stage.
- Consulting: moved an LLM evaluation system to the Batch API (–50% cost) and cut answer variance from 60% to ~6% on an expert golden set; sized GPU infrastructure for speech analytics (50k min/day) 30% cheaper, with headroom to 100k minutes.

AI Consultant · **TheNext**

Sep 2025 – Present · Remote

AI product team developing Sferyx B2B, a trend-monitoring platform for food, beverage, and beauty markets.

- Designed an LLM pipeline processing 10,000+ daily calls, generating relationship graphs for regional trend reports.
- Built an AI trend-monitoring system analyzing 40+ public sources, accelerating trendwatching for food and beverage clients.

SKILLS

CORE	LLMs, AI Agents, Multi-Agent Systems, Context Engineering, Memory Systems
RAG & EVAL	RAG, Retrieval, Reranking, Chunking, LLM Evaluation, Validation Gates
SPEECH	Speech-to-Text, Diarization, Speaker Identification, Meeting Intelligence
MODELING	Machine Learning, Fine-Tuning, Quantization, PyTorch, vLLM
TOOLING	Python, Hugging Face, LangChain, LangGraph, Langfuse, RAGAS, OpenAI API, Anthropic API, MCP
INFRA & DATA	Qdrant, Milvus, Azure OpenAI, Azure AI Search, Docker, Git

EDUCATION

B.S., Oil & Gas Engineering

Tyumen State Oil and Gas University · 2022

LANGUAGES

Russian — Native · **English** — B2