

Yevgen Sharyi

I build voice-agent systems that run in production for US companies.

Most recent: a complete voice sales system for a US logistics company. AI agents take inbound and outbound calls, qualify leads and hand them to the sales team; around them I built the telephony, CRM sync, orchestration, the data layer and an admin panel for leads and calls. Before that, eight years in digital marketing, where automating lead handling led me to engineering.

yevgensharyi@gmail.com t.me/yevgensharyi yevhensharyi.site

01 Skill map

The roadmap.sh AI Engineer roadmap, every topic. A filled dot means I have used the topic in real work and can show where; a ring is in progress, a dashed circle is next.

01 Foundations	● What is an AI engineer ● AI engineer vs ML engineer ◌ AI vs AGI ● LLMs ● Inference ◌ Training ● Embeddings ● Vector databases ● AI agents ● RAG ● Prompt engineering ● Common terminology ● Impact on product development ● Roles and responsibilities
02 Pre-trained models	● Benefits of pre-trained models ● Limitations and considerations ● OpenAI models ● Anthropic Claude ● Google Gemini ◌ Azure AI ◌ AWS SageMaker ● Hugging Face models ◌ Mistral AI ◌ Cohere ◌ Replicate ● Capabilities and context length ● Cut-off dates and knowledge
03 OpenAI platform	● Chat completions API ● Writing prompts ● Maximum tokens ● Token counting ● Pricing considerations ● Managing tokens ◌ OpenAI playground ◌ Fine-tuning
04 AI safety and ethics	● Prompt injection attacks ● Security and privacy concerns ◌ Bias and fairness ● Understanding AI safety issues ◌ OpenAI moderation API ◌ End-user IDs in prompts ● Adversarial testing ● Robust prompt engineering ● Know your customers and use cases ● Constraining inputs and outputs ● Safety best practices
05 Open source AI	● Open vs closed source models ● Popular open source models ● Hugging Face hub ◌ Hugging Face tasks ● Finding open source models ◌ Inference SDK ◌ Transformers.js ◌ Ollama ◌ Ollama models ◌ Ollama SDK
06 Embeddings and vector databases	● What are embeddings ● Semantic search ◌ Data classification ◌ Recommendation systems ◌ Anomaly detection ● OpenAI embedding models ● OpenAI embeddings API ◌ Embedding pricing ◌ Sentence transformers ◌ Open source embeddings ● Vector database purpose ● Chroma ◌ Pinecone ◌ Weaviate ◌ FAISS ◌ LanceDB ◌ Qdrant ● Supabase pgvector ◌ MongoDB Atlas ● Indexing embeddings ● Similarity search
07 RAG	● RAG use cases ● RAG vs fine-tuning ◌ Chunking ● Embedding step ● Vector database step ● Retrieval process ● Generation step ● Implementing with SDKs directly ◌ LangChain ◌ LlamaIndex ◌ OpenAI Assistant API
08 AI agents	● Agent use cases ● Agent prompt engineering ◌ ReAct prompting ● Manual implementation ● OpenAI functions and tools ◌ Assistant API agents
09 Multimodal AI	● Multimodal use cases ● Image understanding ● Image generation ◌ Video understanding ● Audio processing ● Text to speech ● Speech to text ● OpenAI vision API ◌ DALL-E API ● Whisper API ● Hugging Face multimodal models ◌ LangChain multimodal ◌ LlamaIndex multimodal
10 Development tools	● AI code editors ● Code completion tools

02 Selected work

What each system does, what it runs on, and the decisions behind it. Client names are withheld under NDA.

01 US logistics company / Owner, end to end

/ Nov 2025 - now • IN PRODUCTION

Voice sales system

The whole sales loop around the voice agents: inbound and outbound AI calls, CRM sync, orchestration, a Postgres data layer, automatic call review, and an admin panel where the sales team works with leads, calls and agent settings.

ElevenLabs Agents Claude n8n Supabase Deno

Next.js Twilio Pipedrive

- 01 Two generations: a first version with its own test suite and prompt iterations, then the current system from the tender breakdown to live customer calls, with server tools, orchestration, database, admin panels and a separate development environment.
- 02 Prices, discounts and order state are decided on the server. The agent can only ask through tools, so nothing a caller says can change a price.
- 03 Every tool call is idempotent by its call id, so a retry from the voice platform never creates a second order.
- 04 Before each outbound call the system checks the do-not-call list and the lead's local calling window; a lead is claimed atomically in Postgres, so two workers never dial the same person.
- 05 A second model reviews every call and may report an error only with a quote from the transcript; critical cases go to the team in Telegram.

03 Voice AI product for US small businesses

/ Architecture and backend / May - Jul 2026

AI receptionist from a website

A visitor enters a website address. The system reads the site, builds a business profile and shows a phone number; the voice agent that answers already knows the business and talks to the caller as its receptionist.

ElevenLabs Claude OpenAI n8n Supabase pgvector

Twilio Next.js

- 01 One persistent voice agent and a small pool of phone numbers, routed by the dialed number when the call starts, instead of a new agent and number per visitor.
- 02 Site to profile in two passes: one model reads several pages with web fetch, a second model checks the result; a fallback model keeps the demo reaching a call when the first path fails.
- 03 A knowledge base per business in pgvector; abuse limits, consent and a budget cap designed into the first version.
- 04 Architecture, the orchestration backend, the admin panel, and a frontend contract for a second developer.

02 US business-jet charter operator

/ Technical control, fixes, outreach

/ May 2026 - now • IN PRODUCTION

Pilot sourcing and outreach

Finds, verifies and contacts business-jet pilots for a charter operator. An external contractor built the first pipeline; I checked it against real vacancies, fixed what was wrong and built the email outreach.

Python Streamlit ChromaDB OpenAI Apollo.io

FAA registry Mailgun

- 01 Checked the returned candidates by hand against the vacancy requirements and traced the irrelevant matches to a stale scraping cache; after the fix the next run returned no irrelevant candidates.
- 02 Outreach from the client's own domain with DKIM and DMARC; replies land in the recruiting inbox and get classified; sending stays off until someone switches it on.
- 03 Security review of the contractor's API plan: model output passed to eval(), personal data behind unauthenticated endpoints and an open mail relay, each rated critical with a fix.
- 04 Contractor work accepted against a written list of acceptance questions and independent testing before the second payment.

04 Own product, two-person team / Lead developer

/ Jul - Aug 2026 • IN TESTING

Voice interview trainer

A spoken mock interview against a real job posting for Ukrainian IT candidates, scored, with a session report.

Expo React Native TypeScript Gemini Live

Vertex AI Supabase Cloud Run RevenueCat

- 01 Real-time voice over a raw WebSocket to Gemini Live: voice activity detection tuned for interviews, session resumption, transcription hints for Ukrainian.
- 02 Sessions kept dropping under the edge-function runtime limit, so the voice relay moved to a container service with one connection for the whole interview.
- 03 Row-level security in every migration, an SSRF guard on the job-URL fetch, and a CV with a prompt injection in the test set.
- 04 Expo mobile app with a RevenueCat subscription paywall.

05 Own product / Solo / Aug 2026

Telegram Mini App with payments

Tarot readings sold for Telegram Stars: a grammY bot, a Mini App with a 3D deck, and interpretations generated in Ukrainian over a deck drawn by image models.

TypeScript Deno grammY React three.js Supabase
Vertex AI ComfyUI Vercel

- 01 Model chosen by a blind comparison of candidates, with a fallback model behind it.
- 02 Payments idempotent on the provider's charge id; user identity taken only from Telegram's own update context, never from message text.
- 03 Row-level security and grants written in the same migration as each table.
- 04 Deck art from Gemini image models with a FLUX ControlNet fallback; a short-video pipeline on a rented GPU with LTX, Remotion and ffmpeg.

06 Real-estate agency, Kyiv / Sole developer

/ Apr 2026 - now ● IN DAILY USE

Claude Code workspace for a real-estate agency

Claude Code skills that search property portals, value apartments and rentals, draft listings and build PDF presentations for a small agency.

Claude Code Python Playwright SQLite

- 01 Portal search through official APIs where they exist and Playwright where they do not; valuation with several statistical methods side by side, plus rental scoring.
- 02 A file CRM in JSON and Markdown plus SQLite, readable by the agents and by the agency staff.
- 03 Listing forms are filled the way a person would fill them and are never published automatically.
- 04 In daily use by the agency since April 2026.

03 Claude Code

How the systems above get built.

I build with Claude Code: agents write the code, and I own the architecture, the acceptance checks, security, cost and the systems in production.

PROCESS

- 01 Scope

Which process to automate, and what result will show it worked. Cost is measured before anything scales.
- 02 Spec

Modules and phases with a definition of done, and acceptance questions written before the build starts.
- 03 Build with agents

Claude Code writes the code inside the guardrails below; every branch gets a read-only review.
- 04 Run in production

Test on real calls, then watch quality after launch through automatic call review and alerts.

PRACTICES

- Context engineering

Rules per project in CLAUDE.md, persistent memory between sessions, credentials routed by folder with direnv, and MCP servers scoped per project, so an agent only sees its own database.
- Guardrails with hooks

PreToolUse hooks that block writes to production hosts and destructive commands and ask before anything ambiguous; a session-start check that warns when privacy settings drift.
- Skills as procedures

Repeatable work packaged as skills: property search and valuation for an agency, weekly management reports, n8n workflow builds, telephony and CRM operations that default to read-only.
- Agent teams and review gates

Agent teams designed per project, parallel agents for research and implementation, a fact-checker before client deliverables, and a read-only reviewer on every branch.
- MCP and verification

Supabase in read-only mode, Playwright for end-to-end checks and screenshots, Context7 for current library docs, Jira and n8n, plus an own MCP server that indexes n8n nodes.
- Workflow as code

n8n workflows written in TypeScript with the official SDK, validated over MCP and pushed only to allow-listed workflow ids; database changes only through versioned migrations.

04 Experience

Nov 2025 - now

AI engineer, technical lead for AI automation

US technology holding: logistics software, voice AI products, client projects

I own the voice-agent direction: I design, build and run the agents, the orchestration and the data layer, and I check contractor deliveries before they are accepted.

- Voice sales system for the logistics business, end to end, in two generations: the current one went from the tender breakdown to live customer calls, with admin panels and a parallel development environment.
- Read-only audit of an existing production setup of voice agents and n8n workflows: documented it and ranked the risks, among them unauthenticated webhooks and API tokens stored in the clear.
- Pilot sourcing for a charter operator: technical control over an external contractor, acceptance before payment, hands-on fixes and the email stack.
- AI receptionist product: architecture, orchestration backend, vector knowledge base, admin panel, and a frontend contract for a second developer.
- Aircraft sales and sourcing projects: checked specs and vendor proposals against live FAA registry data; my verdict put a build on hold until the market signals are validated.
- Delivery estimates for new AI products, benchmarked against published AI-productivity studies; weekly reports to the director in English and Russian.

2024 - now

Independent automation developer

Small businesses, a partner studio, own products

From 2024, n8n workflows and AI agents for small businesses. Since late 2025, own products built solo or in a pair, each on Supabase with row-level security and migrations from the first commit.

- n8n workflows with conditional logic, data processing and API calls; integrations between CRMs, Telegram, Instagram, email services and analytics.
- AI agents and chatbot funnels that qualify leads, segment audiences and answer clients from the first contact to payment.
- Voice interview trainer and the Telegram Mini App, above.
- Invoicing app for a partner studio: features, written pull-request reviews, and invoice import from a PDF or photo in a single model call.
- Browser voice agent with a lip-synced avatar on Gemini Live native audio, server-executed tools and visemes computed from the audio spectrum.
- Real-estate platform prototype on Twenty CRM, n8n, LangGraph and a custom MCP server, with FreePBX and Fonoster telephony trials.

2016 - 2024

Digital marketer

Two furniture companies, Kyiv

Marketing end to end: strategy, content, paid ads, landing pages and analytics. Automating lead handling there is what led me to engineering.

- Paid campaigns on Facebook, Instagram and Google Ads, a landing page per campaign, analytics with GA4, UTM tags and CRM reports per channel.
- Instagram chatbots that took inquiries, qualified leads and sent reminders, so clients got an answer in minutes instead of hours.
- Connected the sales team to Instagram inquiries, so every lead went through one pipeline.
- Short-form video for Reels, TikTok and Shorts: scripts, shooting and editing.

05 Skills

What AI engineer vacancies ask for most in September 2026, and where I have done it.

The full toolset by layer follows.

ASKED FOR IN MOST AI ENGINEER ROLES

LLM APIs and model choice	Claude, OpenAI and Gemini in production. Models chosen by blind comparison, each with a fallback; reasoning-token cost measured before output limits are set. ClaudeOpenAI Responses APIGeminiVertex AI
Agents and tool calling	Server-side tools behind voice agents, idempotent by call id, with the server as the last word on prices; a hand-built tool loop over a live WebSocket, no framework. ElevenLabs Agentsfunction callingGemini Live
RAG and vector search	A knowledge base per business in pgvector, semantic candidate search in ChromaDB, a sales knowledge base in RAG mode. RAG chosen over fine-tuning on purpose. pgvectorChromaDBOpenAI embeddings
Evals and quality	Every production call scored by an LLM judge that must quote the transcript; regression runs with two agents calling each other; test suites by category; blind model comparisons. LLM-as-judgeregression runstest suites
Prompt engineering and structured output	Production prompts with mapped call scenarios and tagged sources the agent may quote; strict JSON schemas with per-field rules against hallucination. Structured OutputsJSON Schema
Guardrails and AI security	Caller text treated as untrusted data; injection cases in the test sets; security reviews of AI code, among them model output passed to eval() and personal data behind unauthenticated endpoints; TCPA and DNC rules enforced in code. prompt injectionrow-level securitySSRF guardsTCPA and DNC
MCP and agentic coding	Claude Code as the main environment with agent teams, hooks and skills; MCP servers for Supabase, Playwright, Context7, Jira and n8n, and one written in-house. Claude CodeMCP
Cost and latency	Cost measured per call minute and per request before scaling; quotas, rate limits and output budgets; the latency of the webhook path measured before deciding where tools run. usage ledgerrate limitsmodel fallback
Shipping to production	Live systems with a separate development environment, versioned migrations, feature flags without a redeploy, dry-runs before writes and rollback checkpoints; call review and alerts for monitoring. DockerCloud RunVercelGitHub Actions
Python and TypeScript	The systems run on Python (sourcing pipeline, portal scraping, valuation scripts) and TypeScript (edge functions, admin panels, n8n workflows as code). The code is written with Claude Code; I specify, review and test it, with front-end fundamentals from a mentored EPAM program in 2022. PythonTypeScriptDenoNext.js

ASKED FOR IN VOICE AI AND AUTOMATION ROLES

Voice AI pipeline	Speech to model to speech with ElevenLabs Agents and Gemini Live: turn-taking, barge-in, voice activity tuning, session resumption; telephony on Twilio with SMS, A2P 10DLC, toll-free verification and answering-machine detection. ElevenLabsGemini LiveTwilioVAD
Workflow automation and integrations	n8n since 2024, now written as code with the SDK and pushed only to allow-listed workflows; webhooks, retries and error paths; CRM, telephony, email, messaging and payment integrations. n8nwebhooksPipedriveTelegram Bot APIMailgunStripe webhooks
Data layer on Postgres	Supabase Postgres with row-level security from the first migration, least-privilege roles, atomic claims with row locks, state machines in CHECK constraints, scheduled jobs in pg_cron. PostgresSupabaseRLSpg_cron
Multimodal	Invoice OCR in a single model call, image generation with a local fallback model, audio pipelines, generated short video on rented GPUs. Gemini image modelsFLUXLTX videoComfyUI
Product and stakeholders	Business requirements turned into specs with acceptance checks, contractor deliveries accepted against them, weekly reports to management in English and Russian.
TOOLSET BY LAYER	
MODELS AND APIS	Claude Opus, Sonnet and HaikuAnthropic web_fetchOpenAI Responses APIGPT-4.1 family Structured Outputstext-embedding-3-smallGemini Flash and ProGemini Live Gemini TTS and image modelsVertex AIPerplexity sonar
VOICE	ElevenLabs Agentsserver tools and webhookslanguage presetsTwilio numbers and SMS A2P 10DLCtoll-free verificationanswering-machine detectionGemini Live WebSocket VAD and barge-inwhisperKokoroSimli
RETRIEVAL AND DATA	Supabase PostgresRLS and migrationspg_cron and VaultEdge Functions on Denopgvector ChromaDBElevenLabs KB in RAG modeSQLite
ORCHESTRATION AND AGENTS	n8n workflow-as-code@n8n/workflow-sdkn8n MCPClaude Code agent teamshooks and skills MCP: Supabase, Playwright, Context7, JiraLangGraph (prototype)
APPS AND INFRASTRUCTURE	TypeScriptPythonReactNext.jsExpo and React NativeVitegrammYTelegram Mini Apps three.jsVercelCloud RunDockerGitHub ActionsRunPodMailgunPipedrivePlaywright
GENERATIVE MEDIA	ComfyUILTX videoFLUX ControlNet via mfluxRemotionffmpeg

06 Background

LANGUAGES	Ukrainian and Russian, native. English is the working language with the US team: specs, reports and vendor calls.
EDUCATION	Master's degree, Department of Electronic Systems, Donbas State Technical University, 2008-2013. Front-End Development Program with a mentor, EPAM UpSkill and Strategeast, June to November 2022, diploma ID 913874.
CODE	Client code is covered by NDA, so my GitHub repositories (evgensharyy) stay private. Instead of code I offer an architecture walkthrough of the systems above, demos of my own products, or a live build with Claude Code.
LINKEDIN	placeholder: profile URL