



Valentin Lapchevskiy

AI-Native Product Engineer — full-stack + LLM / agent systems, end-to-end

I ship full products end-to-end — backend, frontend, and the LLM-agent layer — by orchestrating AI coding agents at team-level throughput.

CONTACT

Argentina · UTC-3

Remote · US-hours overlap

v.lapchevskiy@gmail.com

Telegram @gen4sp

Upwork — Top Rated

App in the Air ↗

HIGHLIGHTS

- **7M+** installs · Apple-featured app
- **\$300K+** · 61 projects · Top Rated
- **5** AI systems + research (2024–26)
- Agent-native delivery

HOW I WORK

- Turn vague scope into **shipped systems** — own ambiguity end-to-end.
- **Systems-first**: map the path and trade-offs before writing code.
- **Relentless prototyper** — turn ideas into working experiments, fast.
- Calm under shifting requirements; **low-overhead** remote collaboration.

SKILLS

AI / LLM

Agentic systems (tool-calling, OODA), Vercel AI SDK, OpenAI · Anthropic · Gemini, RAG / pgvector, multimodal (vision, STT), structured outputs, cost-aware LLM engineering (cost ledgers, per-tier guards), LLM eval &

SUMMARY

AI-native product engineer who ships full systems end-to-end — backend, frontend, and the LLM-agent layer — by orchestrating AI coding agents at team-level throughput. 15 years building products: co-founder of **App in the Air** (**7M+ installs**, featured by Apple) and an independent engineer, **Top Rated** on Upwork with **\$300K+ across 61 projects** (~100% success). Recent focus: production agentic systems, multimodal LLM pipelines, and AI interpretability.

SELECTED WORK

B2B Brand-Creator Marketing Platform

2026 – present

Contract · Senior Full-Stack / AI Engineer · *under NDA*

- Architecting a **Temporal-based workflow engine** and an **AI provider control plane** (cost ledger, rate limiting, circuit breakers, webhook replay) powering a multi-domain platform.
- Established a **domain-driven monorepo** (portable platform core + business domains) with P0–P5 smoke gates and full observability (Grafana / Loki / Alloy).
- Building a **skills-based LLM generation system** (18+ domain skills — brand enrichment, brief/script generation, anti-fraud) with guardrails and quality evals.

Agentic Support System

2026

Contract · AI Engineer · *B2B GPU-infrastructure company, under NDA*

- Built an end-to-end **LLM-agent support system** (Vercel AI SDK, OODA tool-calling loop) across Telegram / Email / Web — intent classification, field extraction, automated ticketing; delivered in < 2 months (3 prototypes → production).
- Engineered **multimodal intake** (vision + Whisper STT) and context-aware memory over PostgreSQL/pgvector + Redis.
- Implemented a **self-improving eval loop** (45+ real cases → LLM-driven prompt refinement → regression gating in CI) and production infra (Docker, Langfuse, retry queues, graceful shutdown).

observability (Langfuse), agent-driven dev; interpretability, tensor networks.

BACKEND

Node.js / TypeScript, Fastify / Express, Temporal.io, Redis / BullMQ, PostgreSQL (Drizzle, pgvector), MongoDB.

FRONTEND

React, Vue / Nuxt, TypeScript, Tailwind, design systems, FSD.

MOBILE

React Native, Flutter, Electron.

DATA / ML

Python, PyTorch, scikit-learn, data pipelines, scraping.

INFRA

Docker, CI/CD, observability (Prometheus / Grafana / Jaeger / OTel), GCP / AWS.

ALSO

Solidity / Web3; creative tech (TouchDesigner, GLSL/HLSL).

EDUCATION & LANGUAGES

Graduate studies, Cognitive Science

Higher School of Economics (2018–2019).

BSc, Applied Informatics

Russian University of Cooperation (2000–2005).

English (professional) · Russian (native) · Spanish (B1, in progress).

Auditable Corporate-Memory System

2026

Contract · AI R&D · B2B GPU-infrastructure company, under NDA

- Designed and prototyped an **auditable corporate-memory system** — immutable raw layer + re-creatable structured layer, giving end-to-end provenance from any AI output back to its source.
- Built **LLM knowledge-graph extraction** (entities / relations / facts) with pluggable per-domain lenses and type guards over PostgreSQL/pgvector, served via MCP — handed off as a prototype for further development.

Viracle — AI Video-Virality Analytics

2025 – 2026

Founder / Solo Engineer

- Architected an **event-driven backend** (Fastify, Temporal.io, 4 microservices) that scrapes, analyzes, and scores short-form video (TikTok / Reels / Shorts).
- Built **multimodal video analysis** on Gemini structured outputs (hook, pacing, CTA, niche, sentiment) and a **7-stage Python ML pipeline** (correlation, clustering, anomaly detection → content directives).
- Built production-grade infra: PostgreSQL/Drizzle (20+ tables), Redis Streams, full observability (Prometheus / Grafana / Jaeger), billing & auth.

Noracle — AI Dream-Interpretation SaaS

2026

Founder / Solo Engineer · Live demo ↗

- Shipped a **live, monetized** Telegram AI product — dream interpretation plus an evolving per-user personality profile (Telegram Stars billing; **38% D30 retention** among paying users).
- Built a **multi-stage LLM pipeline** — agentic mode routing, multi-provider image generation, per-tier cost guards, structured outputs (~\$0.02/dream).
- Executed a **zero-downtime, idempotent migration** of 3,485 users / 6,340 records with cold-backup safety.

Independent AI Research — Interpretability & Efficient Architectures

2025 – 2026

Self-directed

- 200+ experiments on extracting **exact, verifiable logic** from trained networks (lossless NN → formula → Verilog), vs. approximate methods (SHAP/LIME).
- Notable results: a **40-parameter** recurrent Boolean cell solves dense parity-128 (100% held-out, 5/5 seeds) where a GRU baseline stays at chance; exact **ANF → Verilog** extraction with 100% model/formula agreement; a **0.08%-parameter** Boolean coprocessor gives a frozen GPT-2 multi-digit addition it cannot do zero-shot.
- Rigorous methodology (multi-seed, Bonferroni correction, 3 verification rounds) with a documented falsified-hypotheses log.

GRAB-A-WORD-II — Real-time Multiplayer Word Game

2024

Founder / Solo Engineer

- Built a solo **full-stack real-time multiplayer** game (Nuxt/Vue, Socket.io, Redis/BullMQ) with RU/EN localization and Telegram integration.
- Created a modular design system (Storybook, 26 components) and an Apify + OpenAI dictionary pipeline across a 10-module monorepo.

EXPERIENCE

Independent Full-Stack Engineer & Product Consultant

2015 – present

Upwork · Top Rated

- **\$300K+ earned across 61 projects / 4,446 hours, ~100% success;** clients primarily in the US, Canada, and Europe.
- 30+ MVPs and proofs-of-concept; deep involvement in product decisions and technical advising.

Co-founder · Product & Full-Stack Engineer

2011 – 2014

Empatika — App in the Air

- Co-founded a travel-assistant app and owned the product for its first three years — pivoted it from a location-based chat to airport travel assistance and grew it **0 → 600K users** on under \$15K of marketing.
- Featured by Apple — **Best New Apps in 120 countries**, Editors' Choice, preinstalled across Apple Stores; the app has since passed **7M+ installs**.

Creative Technologist · Visual Programmer

2008 – 2015

INTY · TNT Broadcast Network

- Interactive installations and real-time visual systems (TouchDesigner, Ventuz, GLSL, Arduino).