



Valentin Lapchevskiy

AI-Native Product Engineer — full-stack + sistemas LLM / agentes, de extremo a extremo

Entrego productos completos de extremo a extremo — backend, frontend y la capa de agentes LLM — orquestando agentes de programación con el rendimiento de un equipo.

CONTACTO

Argentina · UTC-3

Remoto · horario de EE. UU.

v.lapchevskiy@gmail.com

Telegram @gen4sp

Upwork — Top Rated

App in the Air ↗

DESTACADOS

- **7M+** descargas · destacada por Apple
- **\$300K+** · 61 proyectos · Top Rated
- **5** sistemas de IA + investigación (2024–26)
- Entrega agent-native

CÓMO TRABAJO

- Convierto un alcance difuso en **sistemas entregados** — asumo la ambigüedad de principio a fin.
- **Primero el sistema:** trazo el camino y los trade-offs antes de escribir código.
- **Prototipador incansable** — convierto ideas en experimentos que funcionan, rápido.
- Calma ante requisitos cambiantes; **bajo overhead**, colaboración remota.

HABILIDADES

IA / LLM

Agentic systems (tool-calling, OODA), Vercel AI SDK, OpenAI · Anthropic · Gemini, RAG / pgvector, multimodal (vision, STT), structured outputs, cost-aware

RESUMEN

Ingeniero de producto AI-native que entrega sistemas completos de extremo a extremo — backend, frontend y la capa de agentes LLM — orquestando agentes de programación con el rendimiento de un equipo. 15 años construyendo productos: cofundador de **App in the Air** (**7M+ descargas**, destacada por Apple) e ingeniero independiente, **Top Rated** en Upwork con **\$300K+ en 61 proyectos** (~100% de éxito). Enfoque reciente: sistemas con agentes en producción, pipelines LLM multimodales e interpretabilidad de IA.

PROYECTOS DESTACADOS

Plataforma B2B de colaboración marca-creador

2026 – act.

Contrato · Ingeniero senior full-stack / IA · *bajo NDA*

- Diseñando un **motor de workflows basado en Temporal** y un **control plane de proveedores de IA** (cost ledger, rate limiting, circuit breakers, reintento de webhooks) que impulsa una plataforma multidominio.
- Establecí un **monorepo orientado a dominios** (núcleo de plataforma portable + dominios de negocio) con smoke gates P0–P5 y observabilidad completa (Grafana / Loki / Alloy).
- Construyendo un **sistema de generación con LLM basado en skills** (18+ skills de dominio — enriquecimiento de marca, generación de briefs/guiones, antifraude) con guardrails y evaluaciones de calidad.

Sistema de soporte con agentes LLM

2026

Contrato · Ingeniero de IA · *empresa de infraestructura GPU B2B, bajo NDA*

- Construí un **sistema de soporte de extremo a extremo con agentes LLM** (Vercel AI SDK, bucle OODA con tool-calling) en Telegram / Email / Web — clasificación de intención, extracción de campos y ticketing automatizado; entregado en < 2 meses (3 prototipos → producción).
- Desarrollé **entrada multimodal** (visión + Whisper STT) y memoria contextual sobre PostgreSQL/pgvector + Redis.
- Implementé un **bucle de evaluación auto-mejorable** (45+ casos reales → refinamiento de prompts con LLM → gating de regresión en CI) e infraestructura de producción (Docker, Langfuse, colas de reintento, apagado controlado).

LLM engineering (cost ledgers, per-tier guards), LLM eval & 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.

DATOS / ML

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

INFRA

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

ADEMÁS

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

EDUCACIÓN E IDIOMAS

Máster en Ciencias Cognitivas

Higher School of Economics (2018–2020).

Licenciatura en Informática Aplicada

Russian University of Cooperation (2000–2005).

Inglés (profesional) · Ruso (nativo) · Español (B1, en progreso).

Sistema de memoria corporativa auditable

2026

Contrato · I+D de IA · *empresa de infraestructura GPU B2B, bajo NDA*

- Diseñé y prototipé un **sistema de memoria corporativa auditable** — capa raw inmutable + capa estructurada recreable, con procedencia de extremo a extremo desde cualquier salida de IA hasta su fuente.
- Construí **extracción de knowledge-graph con LLM** (entidades / relaciones / hechos) con lentes por dominio configurables y type guards sobre PostgreSQL/pgvector, expuesto vía MCP — entregado como prototipo para desarrollo posterior.

Viracle — Analítica de viralidad de vídeo con IA

2025 – 2026

Fundador / Ingeniero en solitario

- Diseñé un **backend orientado a eventos** (Fastify, Temporal.io, 4 microservicios) que rastrea, analiza y puntúa vídeo corto (TikTok / Reels / Shorts).
- Construí **análisis multimodal de vídeo** sobre structured outputs de Gemini (hook, ritmo, CTA, nicho, sentimiento) y un **pipeline de ML en Python de 7 etapas** (correlación, clustering, detección de anomalías → directrices de contenido).
- Construí infraestructura de nivel producción: PostgreSQL/Drizzle (20+ tablas), Redis Streams, observabilidad completa (Prometheus / Grafana / Jaeger), facturación y autenticación.

Noracle — SaaS de interpretación de sueños con IA

2026

Fundador / Ingeniero en solitario · Demo en vivo ↗

- Lancé un **producto de IA en vivo y monetizado** en Telegram — interpretación de sueños + un perfil de personalidad por usuario que evoluciona (facturación con Telegram Stars; **38% de retención D30** entre usuarios de pago).
- Construí un **pipeline LLM multi-etapa** — enrutado agéntico de modos, generación de imágenes multi-proveedor, guardas de coste por tier, structured outputs (~\$0.02/sueño).
- Ejecuté una **migración idempotente sin downtime** de 3.485 usuarios / 6.340 registros con respaldo en frío.

Investigación independiente en IA — interpretabilidad y arquitecturas eficientes

2025 – 2026

Autodirigido

- 200+ experimentos sobre extracción de **lógica exacta y verificable** de redes entrenadas (NN → fórmula → Verilog sin pérdida), frente a métodos aproximados (SHAP/LIME).
- Resultados destacados: una celda booleana recurrente de **40 parámetros** resuelve parity-128 denso (100% en held-out, 5/5 semillas) donde un GRU se queda en el azar; extracción exacta **ANF → Verilog** con 100% de concordancia fórmula/modelo; un coprocesador booleano con el **0,08% de los parámetros** da a un GPT-2 congelado la suma multidígito que no puede hacer zero-shot.
- Metodología rigurosa (multi-seed, corrección de Bonferroni, 3 rondas de verificación) con un registro documentado de hipótesis refutadas.

GRAB-A-WORD-II — Juego de palabras multijugador en tiempo real

2024

Fundador / Ingeniero en solitario

- Construí en solitario un **juego multijugador full-stack en tiempo real** (Nuxt/Vue, Socket.io, Redis/BullMQ) con localización RU/EN e integración con Telegram.
- Creé un sistema de diseño modular (Storybook, 26 componentes) y un pipeline de diccionarios (Apify + OpenAI) en un monorepo de 10 módulos.

EXPERIENCIA

Ingeniero full-stack independiente y consultor de producto

2015 – act.

Upwork · Top Rated

- **\$300K+ generados en 61 proyectos / 4.446 horas, ~100% de éxito;** clientes principalmente en EE. UU., Canadá y Europa.
- 30+ MVPs y pruebas de concepto; participación profunda en decisiones de producto y asesoría técnica.

Cofundador · Ingeniero de producto y full-stack

2011 – 2014

Empatika — App in the Air

- Cofundé una app de asistente de viajes y dirigí el producto sus primeros tres años — pivote de un chat por geolocalización a la asistencia en aeropuertos y crecimiento de **0 → 600K usuarios** con menos de \$15K de marketing.
- Destacada por Apple — **Best New Apps en 120 países**, Editors' Choice, preinstalada en las Apple Store; hoy supera los **7M+ de descargas**.

Creative Technologist · Visual Programmer

2008 – 2015

INTY · TNT Broadcast Network

- Instalaciones interactivas y sistemas visuales en tiempo real (TouchDesigner, Ventuz, GLSL, Arduino).