

sergeyavdovin.com

Sergey Vdovin

Lead AI Engineer

End-to-end, hands-on delivery of modern AI solutions
Top-tier (Professional) Generative AI certified

Agentic AI · LLM Systems · AWS Bedrock · Data and Analytics · PhDc

Antalya, Turkey | Open to relocation (visa sponsorship) · Remote · Direct employment · Employer of Record (EOR) · B2B contract

sergey.a.vdovin@gmail.com | +90 534 872 8703

SUMMARY

Lead AI Engineer, empowered by 20+ years in Data and Analytics, validated by a top-tier (Professional) Generative AI certification — hands-on, end-to-end delivery and ownership of modern AI solutions.

Delivering AI solutions on AWS Bedrock, including AWS AgentCore, Strands Agents, Stanford DSPy, MCP, and A2A implementations — like recent OHDSI concept-mapping agent over a 7.5M-embedding Bedrock knowledge base: first demo in 7 days, ~4¢ per mapping (10× cost reduction via prompt caching and RAG footprint minimization in the optimization cycle), and a 4× AI-vs-SME quality win improvement via LLM-as-judge evaluations.

Understand and prefer to leverage managed, serverless AI architectures for rapid production delivery, balanced with a demonstrated track record of engineering custom, containerized AI solutions for complex enterprise client environments.

To deliver value, always learning and readily crossing responsibility boundaries if needed — both into IT infrastructure and the business.

PhD candidate (PhDc), Mathematics and Computer Science.

GenAI OHDSI Concept Mapping Agent — delivery timeline showing what my delivery looks like in practice:
https://www.linkedin.com/pulse/sample-project-timeline-see-how-my-work-may-looks-like-sergey-vdovin-tzv5f

EXPERIENCE

Independent AI Product — Self-Built, Live (realtors, ex-colleagues & daily personal use) – AI Architect & Engineer – Antalya, Turkey
Apr 2022 – Present

- Architected data platform evolution from Excel PoC to production-scale AWS lakehouse (Kinesis Data Firehose, Apache Iceberg, Athena), migrating through Snowflake and BigQuery iterations.
- Integrated LLM(GPT)-powered data quality validation into the production message processing pipeline.
- Built a real-estate demand-collection subsystem (now in production use) with reverse ETL writing results back into Telegram channels, on Apache Airflow (WSL), AWS S3, AWS Lambda, AWS Glue, and AWS Secrets Manager.
- LLM-system evaluations with LangChain (LangSmith) and MLflow.
- Opened a free Antalya real-estate listing for EPAM colleagues and earned an EPAM recognition badge. The recognition led to the transition from data-heavy projects with supporting AI to AI-heavy projects with supporting data, with corresponding position titles shifting from Data Engineer to AI Engineer.
- Built the analytics layer up from Excel through Microsoft Power BI (drill-downs) and reverse ETL (the real-estate subsystem above) to a personal MCP server on AWS AgentCore Gateway over AWS Athena and Apache Iceberg. Claude mobile app via MCP summarises thousands of messages across hundreds of Telegram channels.

EPAM Systems – Lead AI Engineer – Antalya, Turkey, *Oct 2024 – Mar 2026*

Thomson Reuters – AI Engineer · EPAM Systems · Client: Reuters Editorial
Oct 2025 – Dec 2025

- LLM-systems engineering on a strategic-investment project for the Reuters Editorial team.

- Agent-to-Agent (A2A) protocol, MCP, and OAuth2 (Cognito and Auth0) MVP deployed.
- Recommended adopting an established A2A framework instead of the team's planned custom FastAPI paths and from-scratch A2A protocol implementation.
- Vetted the client's candidate PydanticAI A2A server via a Proof of Concept (PoC), found it incompatible with the project's legacy REST A2A transport requirements, and recommended Google's A2A SDK as the alternative.
- Implemented a PoC of the recommended solution using the legacy A2A REST transport and legacy polling, with end-to-end demonstration and verification through the A2A inspector.
- Re-raised the strategic question of whether the legacy REST and polling approaches were still necessary, after which the A2A SDK solution was confirmed and deployed as the current MVP with OAuth2 (Cognito and Auth0) for credential management.
- Completed an MCP-over-A2A PoC exposing the A2A implementation via FastAPI-MCP / FastMCP, tested with VS Code Copilot MCP.

EPAM Systems Internal – AI Architect & Engineer · GenAI OHDSI Concept Mapping Agent
Jul 2025 – Sep 2025

- POC/DEMO of agentic AI on AWS Bedrock.
- Architected a three-stream delivery approach (AI, UI, infrastructure) approved by the project VP, then owned end-to-end execution when volunteer team members rotated off.
- Achieved SME approval on first prototype in 2 days, shipped full demo in 7 days using Bedrock Agent Classic over 7.5M-embedding Aurora knowledge base with serverless architecture.
- Reached 90% evaluation-set coverage by day 17 at 12 cents per mapping, also proposing the API Gateway reference architecture for the infrastructure stream. Reached 100% coverage and 4 cents per mapping by day 24 via reduced RAG metadata footprint and Python post-processing.
- Migrated from Bedrock Agent Classic to Strands Agents on AgentCore (ReAct loop), the 3rd day after AWS announced AgentCore, addressing critical requirements (no installation on macOS, long-term credential management) via MCP calls through AgentCore Gateway, Cognito, and Aurora KB upgraded to pgvector 0.8.0 for metadata filtering.
- Prompt caching: Bedrock Agents did not support prompt caching, despite documented support. I verified across the Python SDK and the REST API. AWS Support's eventual account-level-switch suggestion fix came after I'd migrated to Strands, which solved it natively and future-proofed the stack.
- Rebuilt the absorbed streams in VBA / PowerQuery (parallel workers + xlwings rewrite).
- Improved AI-vs-SME win rate from 0.4 to 1.7 (4×) via Bedrock evaluations and LLM-as-judge, with 83% caching for 10× cost reduction at 4 cents per mapping via Strands hooks and pipe-delimited RAG format.

EPAM Systems Internal – AI Architect & Engineer · RWE Prometheus
Oct 2024 – Aug 2025

- AI assistant development for the OHDSI cohort and concept set generation in RWE Prometheus.
- The AI feature in the Prometheus platform (OMOP CDM) generates Prometheus query JSON from plain-text study design descriptions, with data access remaining within the platform. Productized version was deployed to Takeda, Dec 2024 – May 2025 (see separate position).
- Built the first prototype during the same week as the joining interview: LangChain Python with prompt management, OpenAI streaming, and a FastHTML interface inspired by OpenAI Canvas and Visual Studio Copilot, then extended to AWS Bedrock Prompt Flows in week two.
- After the prototype, I became overall responsible for the AI part of the project, retaining the AI stream end-to-end as the team grew.
- Interviewed incoming team members and distributed scope as the team grew, splitting the work into a prototyping and experimentation stream and a dedicated deployment stream that I took solo. The deployment stream set up the later dedicated client deployment to Takeda (see separate position).

- Designed the project-specific LLM-system evaluation pipeline on AWS Bedrock Prompt Flows using AWS Athena, AWS Lambda, and dbt over AWS Glue Data Catalog, with experiment analytics in Excel pivots and a fixed-format Power BI Paginated report (SSRS).
- Reached 100% PrimaryCriteria match on a data subset using LLaMA 3.1 70B (second demo video), which surfaced that the original 1000 description-cohort sample pairs did not reflect real user behavior and triggered the shift to 61 SME-sourced test cases.
- Built a Stanford DSPy proof-of-concept with evaluation and led its deployment to an external client.

Takeda – AI Engineer · EPAM Systems · Client: Takeda · RWE Study-Design Automation
Dec 2024 – May 2025

- Real-World Evidence (RWE) study-design automation deployment at Takeda. Solo deployment of the RWE Prometheus product, where I held overall responsibility for the AI part at EPAM (see separate EPAM position), with Takeda RWD (Real-World Data) databases registered in the platform. Genie, the AI assistant inside Prometheus, generates Prometheus query JSON for cohort and concept-set design from plain-text study descriptions, with data access remaining inside the platform.
- Agreed the MVP deployment scope at the director level with Takeda, finalizing a 4-stage environment topology (dev and test inside EPAM with PR-driven flow, plus two Takeda-side environments both on the Takeda dev account, sufficient for the MVP).
- Pioneered EPAM's first deployment into Takeda's new AWS infrastructure (the new AWS environment was nascent inside Takeda itself).
- Carried the Genie MVP solo into Takeda's environment, installed and configured at the client.

EPAM Systems – Lead Software Engineer (Data) – Antalya, Turkey, *Sep 2022 – Oct 2024*

Bayer – Data Engineer · EPAM Systems · Client: Bayer AG
Jun 2024 – Jul 2024

- Brought in as a dbt (Data Build Tool) specialist for the EPAM-managed Bayer engagement.
- Proposed and implemented a dbt-native approach for handling Source-to-Target Mapping (STTM) files for dbt code generation: files loaded as seeds, transformed via dbt, and generated objects exposed via marts.
- Introduced dbt-project-evaluator for managing DWH architecture.
- Identified gaps and missing requirements in the STTM files and proposed a partial generative-AI (LLM) remediation path.
- Continued explaining dbt to the management hierarchy, referencing prior public-channel discussions.

GoDaddy – Lead Software Engineer (Data), EPAM Systems · Client: GoDaddy
Sep 2022 – Apr 2024

- FinTech platform re-implementation and Deferred Revenue (DR) closeout. SOX-compliant DR Mart on Microsoft SQL Server, sitting between the eCommerce upstream and the Accounting / FinTech downstream. The technical role expanded into selected managerial duties on the customer side during ongoing transformations.
- Executed the formal Deferred Revenue closeout over the final 8 months of the engagement
- Investigated and fixed operational issues in the DR Mart, identified longer-term remediation, and shipped several formal Change Orders with code review.
- Migrated the team's error-investigation methodology from SQL scripts and source-system tracing into Excel templates with Power Query and Azure Data Studio SQL Notebooks.
- Introduced Sankey diagrams for deviation investigation in multilayered data flows, and on-premises H2O AutoML for root-cause analysis, identifying hundreds of influential order attributes.

    - Implemented an on-premises SQL and C++ static data-lineage code analysis and testing, covering more than 8,000 database objects via dbt, with partial dbt code generation in Python.

Sberbank – Direction Lead in Analytics/Planning, Technology Division for Finance – Moscow, Russia
Feb 2020 – Apr 2022

Sberbank – Senior Technology Expert, Technology Division for Finance
Nov 2019 – Feb 2020

- Continue to deliver value by building analytical solutions on Microsoft Data Platform with further growth in DWA (Data Warehouse Automation), Open-Source Data and Analytics stacks, and Data Science areas. One of the solutions is now the main candidate for the company-wide data source for macroeconomic indicators.

MERLION – BI/DWH Team Lead – Moscow, Russia
Oct 2018 – Nov 2019

- Shaping the Data Analytics landscape in the largest Russian dealer in information technologies, electronics, household appliances, stationery, and office furniture. Architecture, Planning, Deployment, Monitoring.
- Power BI (both cloud and Power BI Report Server) promotion and implementation.
- Power BI Premium load testing, Power BI Report Server, Power BI Gateway.
- Data Warehouse Automation (DWA) promotion, selection processes, including contacts, and tailored presentations for business; pilot project with WhereScape.
- Data Catalog by Deloitte promotion, maturity level 2 implementation: Confluence + Bitbucket.
- SQL Server column-store indexes promotion and implementation.
- Layered Data Warehouse architecture promotion and implementation.
- SSAS ROLAP promotion and implementation.

Volvo Group – Senior Business Intelligence Consultant – Wroclaw, Poland
Apr 2018 – Jul 2018

- SSAS, SSIS, SSRS, DB Engine/DWH - Renault Trucks MS BI maintenance and enhancements projects. See the second recommendation letter.

International Institute for Applied Systems Analysis (IIASA) – Young Scientists Summer Program Participant – Laxenburg, Austria
Jun 2006 – Aug 2006

- Research: "Using Data Mining for System Behavior Prediction" – early ML work on predictive analytics and pattern recognition algorithms for complex systems. The article is part of the dissertation (in Russian): http://sergeyavdovin.com/ContentPage/ArtWalMart.

Numerous companies and businesses related to databases and analytics – Russia
1999 – 2018

- After the scientific center school, I showed outstanding results in physics and mathematics and was promoted by the radio-technical faculty dean to a scientific company, where my database journey began and later naturally transitioned to analytics and AI. Some details are available on my site:http://sergeyavdovin.com/ContentPage/Resume.

SKILLS

Leadership & Technical Strategy

- Educated and upskilled management and technical staff in modern data/AI tools, including dbt, Stanford DSPy, and LLM frameworks.
- Championed the transition to structured onboarding and participatory knowledge transfer, replacing passive approaches.

- Led the creation of a Data & Analytics Center of Excellence (CoE) at Sberbank, organizing regular knowledge-sharing sessions and establishing best practices.
- Delivered business and technical presentations on DWA tools, Gartner frameworks, and modern BI/Analytics architectures.

Technical Skills

- **AI & Machine Learning:** Generative AI, Agentic AI, LLM Systems, RAG, AI Agents, Foundation Models, Vector Databases, Prompt Engineering, LLMOps, AI Evaluation
- **Cloud Platforms & Infrastructure:** AWS Bedrock, Amazon Bedrock AgentCore, AWS Lambda, AWS Glue, AWS Athena, Apache Iceberg, Kinesis Data Firehose, Aurora, pgvector, CI/CD, Docker
- **Programming & Frameworks:** Python, LangChain, LangSmith, Stanford DSPy, FastAPI, dbt, Apache Airflow, MCP, Agent-to-Agent (A2A), MLflow
- **Data and Analytics:** Big Data, Data Analytics, Apache Iceberg, Power BI, Snowflake, BigQuery, Microsoft SQL Server, SSAS

Industry Sectors

- Healthcare & Life Sciences (EPAM, Bayer, Takeda)
- Payment Systems & Fintech (GoDaddy billing, telepayural.ru)
- Media & Publishing (Reuters Editorial)
- Banking & Financial Services (Sberbank, Ural Bank of the Saving Bank of Russia, Bank of Russia)
- Automotive & Transportation (Volvo Group IT, Car incident management end-to-end self-employed development)
- Public Sector & Government (Administration of Kirovskii district, Bank of Russia)
- IT & Consulting Services (EPAM, incoma.ru, independent consulting)
- System Integration (Program Systems Institute, incoma.ru)
- Retail & FMCG (Merlion, Citilink, Presshouse, Trading firm Tom UPI)
- Telecommunications (vikmaster.ru)
- Scientific & Research Institutions (IIASA, Scientific Energy RnD Center Komatek, Russian Academy of Sciences, University Festival Spring UPI Participants registration system)
- Entrepreneurship (Self-employed, custom solution delivery)

References, Publications

- EPAM recommendation letter 2026, Volvo Group IT recommendation letter 2018 - available upon request.
- LinkedIn Recommendations: https://www.linkedin.com/in/sergey-vdovin-46935223/details/recommendations
- TechNet interview 2015:
  “Interview with a SQL Server Business Intelligence Expert and Wiki Ninja - Sergey Vdovin”
  https://web.archive.org/web/20220814165354/https://techcommunity.microsoft.com/t5/wiki-ninjas-blog/interview-with-a-sql-server-business-intelligence-expert-and/ba-p/945775
- Article (PhD materials) 2006:
  “Using Data Mining for Explanation and Prediction of Systems Behavior: Wal-Mart in the U.S.” (Russian)
  http://sergeyavdovin.com/ContentPage/ArtWalMart
- Project delivery timeline (sample) 2026: "Sample Project Timeline — How My Contribution May Look" — GenAI OHDSI Concept Mapping Agent development milestones, costs, and AI-vs-SME win rate.
  https://www.linkedin.com/pulse/sample-project-timeline-see-how-my-work-may-looks-like-sergey-vdovin-tzv5f

CERTIFICATIONS

AWS Certified Generative AI Developer - Professional (2026)

https://www.credly.com/badges/2af85d9e-ca48-431e-9a52-e0ac76f32d30

AWS Certified Generative AI Developer - Professional Early Adopter (2026)

https://www.credly.com/badges/5931a213-b416-440f-bf1f-dff78c54d4dc

AWS Certified AI Practitioner (2026)

https://www.credly.com/badges/c204729f-4338-4a6e-a81c-f8428ca32713

IBM Python for Data Science and AI (2019)

https://www.credly.com/badges/12e62bc3-3a1e-4539-b3c9-8067d6d02397

EDUCATION

PhD Candidate (ABD) - Mathematical Modeling, Digital Methods, and Complexes of Programs
2008 (4 years) Ural Branch of the Russian Academy of Sciences, Dynamical Systems Department. Part of Dissertation: "Using Data Mining for System Behavior Prediction" - early ML work on predictive analytics and pattern recognition algorithms for complex systems, IIASA, Austria.

Engineer Diploma- Computing machinery, complexes, systems, and networks
2004 (5.5 years) Ural State Technical University, Radiotechnical Faculty

High School - Scientific university center (Ural lyceum), class with in-depth study of physics and mathematics
1998 (2 years)

Verification: https://www.credly.com/badges/2af85d9e-ca48-431e-9a52-e0ac76f32d30



AWS Certified Generative AI Developer - Professional

This badge was issued to Sergey Vdovin
Date issued: March 19, 2026 | Expires: March 19, 2029
Issued by Amazon Web Services Training and Certification
Certification earners have the practical knowledge needed to implement GenAI solutions into production environments using AWS technologies. They demonstrated the ability to design and build solutions using various GenAI architectures, integrate FMs into applications and business workflows, and implement agentic AI solutions. They are able to apply prompt engineering techniques, optimize for cost and performance, and configure in line with security, governance, and Responsible AI practices.

Skills

AI | AI Agents | AI Ethics | AWS | AWS Certification | AWS Cloud | Amazon Web Services | Amazon Web Services (AWS) | Artificial General Intelligence | Artificial Intelligence (AI) | Artificial Intelligence Applications | Cloud Architecture | Cloud Certification | Cloud Data | Cloud Infrastructure | Cloud Migration | Cloud Services | Configuration Management | Cost Management | Cost Optimized Compute Design | Cost Reduction | Engineering | Foundation Models | Generative AI | Governance | Machine Learning (ML) | Security

Certification | Advanced

Certification Criteria

- Successfully passed the AWS Certified Generative AI Developer – Professional exam.

Verification: https://www.credly.com/badges/5931a213-b416-440f-bf1f-dff78c54d4dc



AWS Certified Generative AI Developer - Professional Early Adopter

This badge was issued to Sergey Vdovin
Date issued: March 19, 2026
Issued by Amazon Web Services Training and Certification
Earners of this badge are among the first 5,000 individuals to earn AWS Certified Generative AI Developer – Professional and are an Early Adopter of this certification.

Skills

AI and ML on AWS - Foundational | AWS | AWS Certification | AWS Cloud | Amazon Web Services | Artificial Intelligence (AI) | Cloud Certification | Foundation Models | Generative AI | Machine Learning (ML)

Experience

Earning Criteria

- Earners of this badge are among the first 5,000 individuals to earn the AWS Certified Generative AI Developer – Professional certification.