

Berk Gökden

AI/ML & Data Engineer | AI Agents · Voice AI · RAG · Production ML

Amsterdam, Netherlands · Remote (EU) Dutch citizen · EU work authorization berk@primaxiom.ai +31 62 719 89 52 berkgokden.com

linkedin.com/in/berkgokden github.com/bgokden huggingface.co/Berk

Professional Summary

Full-stack AI/ML & Data Engineer who builds and ships GenAI applications and production ML systems end-to-end, fast and cost-efficiently, using Claude and coding agents to accelerate delivery. Freelance AI/ML engineer since 2021, delivering for clients including Klarna (5,000+ employees), Booking.com, and ING, and founder of PrimAxiom Labs, an Amsterdam venture building and operating AI-driven software products. Experienced across multi-agent systems, RAG pipelines, LLM fine-tuning and applications, recommendation systems, NLP, and data infrastructure. Led engineering at Vamp.io (acquired by CircleCI) managing a 12-person team. Open-source contributor (llama.cpp sampler, vector search) with expertise in Python, PySpark, LangChain, Hugging Face, scikit-learn, and distributed systems.

Highlights

- Built an internal AI platform for **5,000+ employees** at Klarna — **40%** efficiency gains and lower cost per meeting
- **Design, build, and deploy production AI agent systems** end-to-end — multi-agent orchestration, MCP tools, evaluation, and guardrails, from prototype to live systems
- Cut infrastructure cost **90%+** with an AI load-prediction optimizer at Vamp.io (acquired by CircleCI)
- **15+ years of production software engineering** — distributed systems, microservices, and Kubernetes across Go, Python, and TypeScript
- Ship **CI-tested open-source AI tooling** — MCP servers, an NER data generator, and a voice meeting-intelligence agent — plus published multilingual NER models and vector-search/LLM contributions

Professional Experience

Machine Learning Engineer (Contract)

2026 - Present

[Booking.com](https://www.booking.com)

Amsterdam, Netherlands

- Build AI agents and automated workflows on production systems
- Run agent and LLM evaluation and prompt optimization to improve accuracy, latency, and cost
- Integrate MCP (Model Context Protocol) for tool and data connectivity

Founder

2026 - Present

[PrimAxiom Labs](https://www.primaxiom.ai)

Amsterdam, Netherlands

- Founder of an Amsterdam technology venture building and operating AI-driven software products — from AI agents and voice to data platforms
- Build production AI systems in-house: multi-agent systems, voice agents, RAG pipelines, and GenAI applications
- Design agentic workflows using LangGraph, Claude Agent SDK, MCP, and custom orchestrators over OpenAI, Anthropic, and open-weight models
- Build multi-agent systems for complex task decomposition, with short-term (working) and long-term (vector-backed) agent memory
- Build voice agents integrating speech-to-text, LLM, and text-to-speech (Retell) with telephony and low-latency tuning
- Ship fast using Claude and coding agents with agent-driven development workflows
- Technologies: Python, Go, TypeScript, Claude Code, LangChain, LangGraph, vLLM, Retell, PostgreSQL+pgvector, FAISS, Kubernetes, AWS, GCP

Senior AI Engineer (Contract)

Sep 2023 - Dec 2025

Klarna

Amsterdam, Netherlands

- Built and deployed an AI platform serving 5,000+ employees with multiple production agent systems: meeting intelligence, conversational voice agents, automated feedback, and internal search
- Built a voice meeting-intelligence agent — speech-to-text transcription with LLM summarization and structured extraction of decisions and action items — reducing cost per meeting and improving meeting efficiency (~40%)
- Deployed GenAI applications using LangChain with OpenAI and Anthropic LLMs under enterprise agreements with custom RAG implementations
- Conducted systematic evaluation of LLM performance across different models (OpenAI, Anthropic, Hugging Face models such as Llama, Mistral) and parameters, iterating on prompts, RAG configurations, and model selection to optimize for accuracy, latency, and cost
- Fine-tuned LLMs and embedding models from Hugging Face for domain-specific terminology and internal knowledge bases
- Implemented vector search using PostgreSQL+pgvector, FAISS, and Annoy for cost-effective semantic search
- Built and deployed Slack-integrated AI agents connected to internal workflows and a RAG knowledge base — employees query company knowledge and trigger actions conversationally, in the tool they already use
- Deployed multi-agent systems and automated workflows that saved teams hours of manual tracking each week
- Shipped these systems via CI/CD to a containerized internal platform on AWS
- Technologies: Python, TypeScript, Go, LangChain, OpenAI, Anthropic, Hugging Face, RAG, PostgreSQL+pgvector, FAISS, Docker, CI/CD, AWS

Senior Software Engineer - Data (Contract)

May 2023 - Sep 2023

Booking.com

Amsterdam, Netherlands

- Supported cloud migration from on-premise to AWS, facilitating data quality during infrastructure modernization
- Developed data validation services and pipelines using microservices architecture
- Improved data quality through a governance framework, meeting Booking's 99% pipeline-reliability requirement
- Validated millions of booking records daily across large-scale data processing systems
- Technologies: DBT, Apache Spark SQL, Snowflake, Hadoop, Java, Python, Argo, Kubernetes, AWS

Senior Data Engineer (Contract)

Jul 2022 - Apr 2023

ING

Amsterdam, Netherlands

- Automated environmental sustainability data processing for ING Wholesale Banking, reducing processing time from weeks to hours
- Built reliable daily pipeline tracking sustainability metrics for hundreds of companies used in credit score calculations, replacing manual Excel workflows
- Developed framework and tools to help data engineers build production pipelines faster
- Solved complex entity resolution challenges tracking company histories through mergers, closures, and restructuring
- Technologies: Python, Apache Spark, PySpark, Kedro, Docker, Airflow, Kubernetes (OpenShift)

Big Data Engineer / ML Engineer (Freelance)

Sep 2021 - Feb 2022

VodafoneZiggo

Utrecht, Netherlands

- Built predictive ML models to detect TV/internet box failures before they occurred, analyzing heat patterns and device telemetry
- Automated proactive customer support by triggering service calls when boxes showed failure patterns, reducing box-related complaints by ~70%, measured through call-center metrics
- Developed forecasting models to predict volume of failing boxes for capacity planning and inventory management
- Delivered ML pipelines to production in collaboration with data scientists
- Technologies: AWS (S3, EMR, Glue, Athena, DynamoDB, Lambda), Python

Machine Learning Engineer

Sep 2020 - Aug 2021

DPG Media Netherlands

Amsterdam, Netherlands

- Built search re-ranking system for DPG Media magazine app using user history while avoiding filter bubbles
- Developed notification optimization models tracking user engagement patterns to improve message interactions
- Built user data pipeline merging clickstream data from multiple websites and apps across DPG Media acquisitions to create unified user profiles for ad targeting
- Implemented ML models including classification, clustering, and decision trees for user segmentation, personalization, and content recommendations
- Won company hackathon by building vector-based recommendation system using Universal Sentence Encoder for article enrichment and fast search, first transformer-based system deployed at DPG Media
- Developed streaming applications using Apache Kafka (AWS MSK) and Spark Streaming for real-time data processing
- Technologies: Python, scikit-learn, pandas, numpy, Apache Kafka, Spark, Delta Lake, AWS (S3, DynamoDB, EKS), Airflow, MLflow, Databricks (evaluated), FastAPI, Flask, Terraform, Universal Sentence Encoder

Senior Software Engineer

Caspar AI

Mar 2020 - Sep 2020

Rotterdam, Netherlands

- Built AI sensor system for elderly care facilities in US, monitoring falls, controlling lighting/heating, and alerting nurses
- Set up on-premise Kubernetes clusters inside buildings for privacy compliance, ensuring no user data left facilities
- Deployed deep learning models for image recognition on edge servers, processing data locally for security
- Built data pipeline sending only operational metrics to cloud while keeping sensitive data on-premise
- Technologies: Python, Kubernetes (Kubespray), TensorFlow Serving, Nginx, AWS (S3, EC2, MSK), Kafka

Engineering Lead

Vamp.io (acquired by CircleCI)

Oct 2017 - Mar 2020

Amsterdam, Netherlands

- Led engineering team of 12 people building cloud-native AIOps platform for canary releasing and microservices deployment
- Built distributed systems and container orchestration infrastructure on Kubernetes
- Developed AI load prediction algorithm for Vamp Continuous Cloud Optimizer, optimizing CPU and memory usage and reducing infrastructure costs by over 90%
- Company subsequently acquired by CircleCI
- Technologies: Go, Kubernetes, Docker, microservices architecture, machine learning

Senior Software Engineer

Zoover / Weeronline

Sep 2016 - Sep 2017

Amsterdam, Netherlands

- Rebuilt weather data ETL pipeline for Weeronline app/website, reducing radar image and rain prediction latency from 45 minutes to under 1 minute
- Rapidly built fraud detection system using Apache Spark analyzing review patterns and user behavior to identify and flag fraudulent reviews on Zoover platform
- Set up CI/CD pipelines with Kubernetes (GKE), CircleCI, and GitLab using Scala

Developer

SAP SE

Mar 2015 - Sep 2016

Istanbul / Waldorf, Germany

- Developed SAP Vora in-memory query engine integrated with Apache Spark for Big Data analytics on Hadoop
- Built Docker-based deployment system for running Spark clusters in cloud environments
- Worked on database internals and cloud service integrations for SAP Vora

Software Engineer

T2 Software

Nov 2013 - Mar 2015

Ankara, Turkey

- Developed full-stack solutions for major Turkish telecom operator handling 100M+ subscriber records
- Built customer service systems including provider transfer functionality completing migrations in under 10 minutes
- Built service-oriented architecture and improved backend performance
- Negotiated with stakeholders to improve API design
- Technologies: Java, JavaScript, Oracle SQL

Software Engineer

Aselsan

Sep 2011 - Nov 2013

Ankara, Turkey

- Developed real-time simulation system for radar test laboratory testing missile guidance systems
- Built flight scenario simulator commanding radar hardware while logging and visualizing test results
- Implemented distributed architecture using multi-threading and reflective memory in C/C++ on Linux

Open Source Contributions & Projects

Focused open-source samples extracted from production work — clean, tested, CI-backed reference implementations of larger systems.

voxbrief — Meeting-Intelligence Voice Agent

github.com/bgokden/voxbrief

Local meeting-intelligence agent: Whisper transcription → structured brief (summary, decisions, action items with owners, topics) → Q&A grounded in the transcript. Ships as a library, CLI, and MCP server with pluggable STT/LLM backends and structured (JSON-schema) output.

personalens — Persona-based UI/Product Review

github.com/bgokden/personalens

"A code review, but for UX": drives Playwright and a vision LLM to review a URL from each persona's point of view — scoring the page, listing problems, and tracking scores over time, with a CI merge gate. Pluggable local-first vision backends (Ollama, Qwen2.5-VL); library + CLI.

MCP (Model Context Protocol) server that gives AI agents and assistants (Claude Desktop, Claude Code, Cursor) a multilingual place-extraction tool — cities, countries, regions, and landmarks across 13 languages — backed by the mDeBERTa/ONNX models below.

synthspan — Synthetic NER Data Generatorgithub.com/bgokden/synthspan

Open-source Python library and CLI that generates synthetic labeled data for NER / token classification — from templates + gazetteers or local-LLM few-shot with structured (JSON-schema) output — with balancing and span-preserving augmentation. CI-tested against llama.cpp and Ollama; the data-generation approach behind the place-extractor models.

Multilingual Place Extractor Models

huggingface.co/Berk

Multilingual place and entity extraction models built on mDeBERTa for token classification across 13 languages, published on Hugging Face with ONNX-optimized versions for fast inference.

veri - Vector Search in Go

github.com/bgokden/veri

Vector search engine built in Go for ML systems. High-throughput feature serving with low latency. 54+ GitHub stars.

llama-constrain - Constrained LLM Generation

github.com/bgokden/llama-constrain

Developed custom sampler in llama.cpp enabling strict constrained sampling in LLMs. Library for constrained text generation with Large Language Models for structured output using GGUF models.

Annoy - Vector Search Library (Contributor)

github.com/spotify/annoy

Contributor to Spotify's approximate nearest neighbors library for vector search. Built custom implementations for production use cases.

Technical Skills

AI/ML (High Priority)

- **Agents:** Multi-agent orchestration, agent memory (short-term working + long-term vector-backed), RAG, MCP, tool use, LangChain/LangGraph, Claude Agent SDK, Claude Code & coding agents
 - **Eval & safety:** Agent evaluation & observability (Arize, Phoenix), guardrails, PII redaction, prompt-injection defense (OWASP LLM Top 10)
- **Models:** LLMs (OpenAI, Anthropic, Gemini, Qwen, Llama, Mistral), fine-tuning, embeddings, encoder models (BERT, DeBERTa), Whisper, Cloud GenAI (Azure OpenAI, AWS Bedrock)
 - **Inference:** vLLM, llama.cpp/GGUF, constrained/structured decoding (grammars, JSON schema)
 - **Voice & data:** Voice agents (Retell), synthetic data generation (NER), NLP & recommendation, scikit-learn, pandas, PyTorch, TensorFlow

Languages

- Python, Go, TypeScript, SQL, Scala, Java

Data & Vector Search

- **Vector:** pgvector, FAISS, Annoy (contributor), Pinecone, Qdrant
- **Data:** ETL/ELT, streaming, feature stores, warehousing, analytics

Infrastructure & Architecture

- Kubernetes, Docker, Terraform (IaC), CI/CD (GitHub Actions), microservices, distributed systems, AWS, GCP

Education

Master of Science - Electronics & Electrical Engineering (Not Completed)

Bilkent University, Ankara

2010 - 2012

Focus: Machine Learning applied to signal/image/audio processing. Developed classification models using SVMs, including full ML pipeline: data collection, feature engineering, model development, testing, and validation. Technologies: MATLAB, Python, Support Vector Machines. Worked as Research and Teaching Assistant.

Bachelor of Science - Electronics & Electrical Engineering

Bilkent University, Ankara

2006 - 2010

Spoken Languages

- **Turkish:** Native
- **English:** Full Professional Proficiency