

Om Borda

AI/ML Engineer

LLM, RAG & Agentic Systems

Würzburg, Germany · +49 155 6034 3000 · omborda2002@gmail.com

linkedin.com/in/omborda · github.com/omborda2002 · omborda2002.showdata.online



Authorized to work in Germany · Available immediately for full-time employment · Master's thesis in progress, expected submission January 2027

PROFILE

AI/ML Engineer focused on production LLM, RAG and recommendation systems. At LikeTik I build hybrid retrieval, LLM re-ranking and multimodal video-analysis pipelines across approximately 6,800 creators, 209,000 videos and 19,500 products, using Qdrant, Whisper, Qwen-VL and self-hosted GPU inference. Currently completing an MSc in Artificial Intelligence at THWS Würzburg-Schweinfurt, with a thesis on guardrails and sensitive-data leakage in multi-agent systems.

PROFESSIONAL EXPERIENCE

AI/ML Engineer (Working Student)

LikeTik (Axinity GmbH & Co. KG)

Jul 2025 – Present

Würzburg, Germany

- Built a hybrid recommendation system matching approximately 19,500 products to approximately 6,800 analysed creators, combining multilingual embeddings, Qdrant hybrid vector search, metadata filtering and an LLM re-ranking layer returning 5–10 ranked recommendations per creator.
- Designed the scoring and segmentation layer across eight dimensions – engagement, audience fit, brand fit, reach, discovery, trend fit, cost efficiency and shop conversion – with embedding-based clustering for creator segments and content patterns.
- Built a multimodal video-analysis pipeline over approximately 209,000 videos using OpenAI Whisper (Large) for speech-to-text, Qwen-VL for frame-level scene understanding and PyTesseract OCR; extracted signals feed back into the recommendation features.
- Enforced structured outputs across LLM components with Pydantic response models and JSON Schema, using function calling, tool calling, schema validation and fallback handling on validation failure.
- Developed production AI APIs with Flask and FastAPI, using MongoDB as the operational database and Qdrant for embeddings, vector search and hybrid retrieval, with batch processing, error handling, logging and data-quality validation.
- Deployed containerised services with Docker and Kubernetes on Hetzner Cloud, serving a self-hosted Qwen 7B on an NVIDIA A100 40 GB GPU; supported pipelines designed for approximately 184,000 creator profiles and more than one million videos.

Data Analytics & Machine Learning Engineer

Bigscal Technologies Pvt. Ltd. · Part-time alongside studies

Jun 2023 – Feb 2025

Gujarat, India

- Built a customer churn prediction pipeline in Python with pandas, scikit-learn and XGBoost, engineering features on activity trends, session duration, failed-payment ratio and support-ticket frequency, with weekly retraining.
- Queried and processed data from relational client databases using SQL, consolidating structured metadata into the feature pipeline and reporting layer.
- Developed a sales analytics and forecasting dashboard in Streamlit with KPI cards, regional heatmaps and regression-based forecasts.

Full Stack Developer

Greendotslab Software Solutions · Part-time alongside studies

May 2022 – Jun 2023

Gujarat, India

- Built and deployed more than ten client web applications end to end with Node.js, React and REST APIs, with modular architecture and load balancing for production use.

PROJECTS

Stock Research Agent · LangGraph, LangChain, Groq, Qdrant, Supabase (PostgreSQL), FastAPI equity.showdata.online

- Designed and deployed a five-node LangGraph agent pipeline producing source-cited equity research reports across six global exchanges: RAG over SEC 10-K and 10-Q filings via Qdrant hybrid search, real-time news sentiment and a self-correcting Critic Agent with confidence gating.
- Applied hallucination mitigation via grounding, source citations, structured outputs and schema validation; metadata modelled in PostgreSQL (Supabase) with SQL and FastAPI.

Gemma-2B Reasoning Model · PyTorch, LoRA, PEFT, bitsandbytes, Hugging Face Transformers Google Tunix Hack on Kaggle

- Co-developed a reasoning-tuned Gemma-2B in a team of three, owning the inference pipeline; fine-tuned with LoRA (rank 32, alpha 64) on approximately 570k samples from MetaMath, OpenThoughts, Medical-O1, Bespoke-Stratos and GSM8K, with 4-bit NF4 quantization.

Medical Image XAI & Model Compression · PyTorch, ResNet18, MobileNetV2, Grad-CAM, LIME GitHub

- Fine-tuned ResNet18 for pneumonia detection on chest X-rays with Grad-CAM and LIME explainability; compressed a MobileNetV2 variant via structured pruning and INT8 quantization from 8.7 MB to 4.4 MB, a 49% reduction.

UAV Waterfowl Detection · YOLOv8, PyTorch, OpenCV, thermal imagery GitHub

- Trained a YOLOv8 object detector on thermal UAV imagery, reaching 86.44% mAP@0.5 and 93.21% precision on 83 test images with 1,411 ground-truth annotations.

EDUCATION

Technische Hochschule Würzburg-Schweinfurt (THWS)

Mar 2025 – Jan 2027 (expected)

Master of Science in Artificial Intelligence

Würzburg, Germany

- Master's thesis (in progress): *Security Failure Propagation in Multi-Agent Coding Systems*, supervised by Prof. Dr. Ivan Yamshchikov – measuring how secret and PII leakage propagates across coder, tester and reviewer agents in a LangGraph pipeline, and evaluating a four-layer guardrail ensemble (gitleaks, TruffleHog, Microsoft Presidio, LLM-as-judge) against GDPR/DSGVO and EU AI Act requirements.

Gujarat Technological University

Jul 2020 – May 2024

Bachelor of Engineering in Computer Science & Engineering · Grade 9.10/10 (German equivalent: 1.4) Gujarat, India

SKILLS

Core Technical Expertise

AI & Machine Learning: Artificial Intelligence (AI), Machine Learning, Deep Learning, Natural Language Processing (NLP), Computer Vision, Generative AI, Reinforcement Learning, Recommender Systems, clustering, feature engineering, model evaluation

LLM & Generative AI: Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), Prompt Engineering, Fine-tuning (LoRA, PEFT), Quantization (4-bit NF4, INT8), Embeddings, Semantic Search, Hybrid Search, Re-Ranking, Chunking, Guardrails, Structured Outputs, Function Calling, Tool Calling, Schema Validation, Pydantic, JSON Schema, hallucination mitigation

Agents, Orchestration & Automation: LangGraph, LangChain, AI agents, agentic workflows, multi-agent systems, state handling, self-correcting critic agents; n8n, API integrations, automated data flows, multi-step AI workflows, backend-service orchestration

AI-Assisted Development: Claude Code, OpenAI Codex, Cursor, GitHub Copilot; parallel task decomposition, code review, debugging and test-based validation

Speech & Vision: OpenAI Whisper (Large), Speech-to-Text, Automatic Speech Recognition (ASR), Qwen-VL, PyTesseract (OCR), YOLOv8, ResNet, MobileNetV2, Grad-CAM, LIME, OpenCV

Languages & Libraries: Python, SQL, JavaScript, TypeScript, PyTorch, scikit-learn, XGBoost, Gradient Boosting, pandas, NumPy, Hugging Face Transformers, bitsandbytes, Streamlit, Node.js, React

Backend & MLOps: FastAPI, Flask, REST APIs, microservices, Docker, containerization, Kubernetes, Hetzner Cloud, Linux, NGINX, Git; model deployment, GPU inference, batch processing, error handling, logging, data quality validation

Data & Databases: MongoDB (NoSQL), Qdrant (vector database), PostgreSQL, SQL, Supabase, data pipelines, embedding pipelines, scalable data processing

Trustworthy AI: PII detection, secret scanning, guardrails, Microsoft Presidio, gitleaks, TruffleHog, LLM-as-judge, GDPR / DSGVO, EU AI Act, prompt injection

Ways of working: End-to-end ownership, cross-functional collaboration with stakeholders, structured and solution-oriented working style, analytical thinking, prototyping and proof of concept, Agile teamwork

Spoken languages: German (B2), English (C1), Hindi (native), Gujarati (native)

Additional Working Knowledge

AWS, Microsoft Azure, GitHub Actions, TensorFlow, MLflow – practical working knowledge

Current Professional Development

LLM evaluation and benchmarking with Ragas and Langfuse; LLM tracing and observability; GitLab CI and CI/CD workflows; Model Context Protocol (MCP); LLMOps, model monitoring, drift detection, human-in-the-loop review gates, escalation design and AI governance