

# AKSHAY JOSHI

Applied AI & ML Research | Multimodal GenAI | Tech Leadership

akshay@akjo.tech · +49 155 661 76540 · akjo.tech · Munich, Germany · German Permanent Resident · US B-1 Visa

## Profile Summary

---

Hands-on AI/ML lead specializing in multimodal GenAI - vision-language models, RAG, and agents. Joined Blinkin as its first AI hire and now lead Applied AI, building and deploying AI products for industrial customers with a small cross-functional team. Earlier: AI research & engineering at Google, AMD, German Research Center for AI, and ETHZ.

## Selected Highlights

---

- As Head of Applied AI, led technical execution of a EUR 2.5M EIC deep-tech grant; selected for the EIC Silicon Valley Expansion Programme (incl. TechCrunch Disrupt) and represented Blinkin at CES Las Vegas & Web Summit Lisbon.
- Pretrained and fine-tuned multimodal models on H100 GPUs across ETH's approx. 2000-GPU Euler cluster, using datasets of millions of image-text pairs and over 40M clinical documents, with TPU-based serving on Google Cloud.
- Built compact 34–40B multimodal models, distilled from much larger teacher models, that read pictures, diagrams, infographs, and documents and answer detailed visual questions on par with models several times their size.
- Deployed AI models in inspection systems used by industrial firms incl. Siemens, ABB, Bosch, PERI, and Heinemann, cutting inspection effort by up to 70% through automated checklist completion from pictures, videos & documents.

## Experience

---

### Head of Applied AI

Jan 2026 - Present

Blinkin AI, Germany

Managing Director: Josef Suess

- Supported EUR 80K to EUR 1M-plus soft-booked ARR and enterprise wins with Siemens, EIC (brochure, p. 9), ABB, Bosch Industriekessel, PERI, and Heinemann by leading a 5-person AI/product/backend team across discovery, solution architecture, UI/UX, APIs, and deployment.
- Led AI roadmap and technical execution for a EUR 2.5M non-dilutive EIC deep-tech grant by translating industrial requirements into VLM/RAG milestones, product deliverables, and deployment strategy.
- Strengthened fundraising, product/PMF & U.S. GTM with board advisor Till Kaestner (former LinkedIn DACH Managing Director) and German Accelerator's Silicon Valley AI program, shaping investor materials, enterprise positioning & scale-up strategy.

### Senior Applied AI Scientist

Oct 2023 - Dec 2025

Blinkin AI, Germany

- Reported directly to the CTO; enabled real-time boiler-health analysis for Bosch, verifying critical operational parameters such as pH/temperature/pressure/water-chemistry & flagging risks in under 4 sec, by indexing thousands of procedures & handwritten logs into a multimodal RAG system.
- Built a zero-shot VLM pipeline for ABB medium/high-voltage switchgear by processing up to 30-min long electrical-room videos / over 25 panel images to extract 12 unique component types such as protection relay, circuit breaker at up to 92% accuracy and build asset hierarchy trees with over 87% accuracy.
- Reduced inspection effort by up to 70% and 30-90 min per inspection by deploying multimodal workflows that pre-fill over 100 checklist questions from photos, video, documents, and machine evidence.
- Shipped PERI inventory intelligence by combining segmentation, reasoning & retrieval over 500-plus subcomponents & 30-plus products to identify parts, assemblies, avail. & sourcing needs in under 10 sec.

### Joint Research Fellow

Feb 2022 - May 2023

Google Research, Switzerland

Advisors: Dr. A. Tonioni, Dr. H. Rebecq, Dr. S. Ostermann

- Collaborated with Google Research Machine Perception - AR/VR, in partnership with German Research Center for Artificial Intelligence (DFKI), to design & pretrain DiagramNet, a self-supervised multitask Vision Language Model for diagram representation learning and visual reasoning.
- Built a 1.87M-sample pretraining corpus from web diagrams paired with DBPedia, WikiData & CK12 captions; evaluated SSL objective combinations & proposed an optimal multitask pretraining setup.
- Outperformed DQA-Net and Pix2Struct baselines on AI2D diagram QA with 28.39% and 22.36% relative improvements using DiagramNet pretraining.
- Reduced hallucination and language-prior bias by 14.3% and improved chart-comprehension accuracy by 5.36% rel. through long-caption encoding, multimodal fusion, ViT/CLIP-based entropy minimization & consistency regularization.

## Machine Learning Researcher - HiWi

Nov 2020 - Jul 2023

German Research Center for Artificial Intelligence, Germany

Advisors: Prof. P. Loos, Dr. P. Pfeiffer

- Developed novel SSL training methods for discriminative LLMs & Subword Tokenizers for Biomedical NLU. Models were pre-trained on 40 million clinical reports from PubMed.
- Achieved 2% avg. abs. improvement over state-of-the-art transformer models of similar scale in PubMedQA (Que. Ans.), EBM NLP (PICO), BC5-disease (Medical NER) & BIOSSES (Sentence Sim.) downstream tasks in the Microsoft Biomedical Language Understanding and Reasoning Benchmark.
- Built highly parallel & computationally efficient Semantic Search (retrieves & ranks 2.5 million documents in under 3 sec) & Recommendation System for Smart Vigilance in Medical Product Research & Dev.
- Set up Google Language Interpretability Tool workflows to debug models with Integrated/Smooth Grad saliency maps, attention, prediction visualizations, embedding analysis & counterfactual checks.

## Software Engineer

Feb 2018 - Sep 2019

AMD Research & Development, India

Manager: Satish Kumar Narayanaswamy

- Secure platform firmware for AMD Ryzen processors, including PSP initialization and PlayReady DRM support across desktop/mobile x86 families.
- Built and validated Ryzen Master SDK features for CPU/memory overclocking, core parking, and SMT controls; delivered 70 Windows SDK APIs.
- Created over 100 unit tests and CLI validation coverage, working across firmware, SDK & product teams.

## Education

---

ETH Zurich, Switzerland

Feb 2022 - May 2023

Master of Science Research Thesis

GPA: 1.2/5.0 (Best: 1.0)

Research Area: Vision & Language, Visual Reasoning

Advisor: Prof. Mrinmaya Sachan

University of Saarland, Germany

Oct 2019 - Jul 2023

Master of Science in Data Science & Artificial Intelligence

GPA: 1.7/5.0 (Best: 1.0)

Research Area: NLU, Computational Semantics

Advisor: Prof. Josef van Genabith

Visvesvaraya Technological University, India

Aug 2013 - Jun 2017

Bachelor of Engineering in Computer Science & Engineering

GPA: 8.03/10 (Best: 10.0)

## Publications & Preprints

---

- **Akshay Joshi**, “*Self-Supervised Multimodal Representation Learning for Diagram Understanding*”, Master’s Thesis, Saarland University/German Research Center for AI, 2023.
- **Akshay Joshi**, Peter Pfeiffer, Annalena Kohnert, Philip Hake, Peter Fettke, Peter Loos, “*PubMedSMBERT: A Pretrained Biomedical Language Model for Medical Smart Vigilance*”, Internal, 2023. Provided on request.
- Annalena Kohnert, **Akshay Joshi**, Peter Pfeiffer, “*Evaluating Medical LM Pretraining*”, EU Project Manuscript, 2023. Provided on request.
- **Akshay Joshi**, Ankit Agrawal, Sushmita Nair, “*Art Style Classification with Self-Trained Ensemble of AutoEncoding Transformations*”, arXiv:2012.03377, 2020.

## Skills

---

- **Engineering Stack**: Python, C/C++, SQL, Docker, Kubernetes, Git, CI/CD, AWS, GCP (TPU), Azure
- **GenAI Stack**: Multimodal RAG, Multi-agent Orchestration, Agent Handoffs/Tracing, Tool Calling, Model Context Protocol (MCP), LLM Evaluation, Guardrails
- **Training & Optimization**: PyTorch, Hugging Face Transformers/Tokenizers/Accelerate, PEFT, LoRA, SFT, DPO/RLHF, DeepSpeed, Quantization, Curriculum Learning
- **Serving & Retrieval**: vLLM, TensorRT-LLM, LlamaIndex, LangChain/LangGraph, Qdrant, ClickHouse, Elasticsearch, FAISS, Redis, MLflow, FastAPI, W&B
- **Spoken Languages**: English (CEFR C1), German (CEFR B1)

## Projects

---

- Low-rank Subspace Representation Finetuning & Multi-Teacher Quantized Distillation of Llama 3/3.1.
- MICCAI Radiogenomic Classification of fMRI to detect latent MGMT Methylation via 3D CNN.

## Achievements

---

- Graduated in the **top 5%** of a **106**-student Bachelors class in Computer Science & Engineering.