

# DEVENDRA VYAS

📍 Germany ✉️ [vyasdevms@gmail.com](mailto:vyasdevms@gmail.com) [in](#) LinkedIn [globe](#) Website

## Summary

---

AI/ML engineer specializing in Agentic AI and Deep Learning: production agentic pipelines, RAG-based knowledge systems, LLM/VLM fine-tuning and perception pipelines. Track record of designing scalable, efficient generative and inference infrastructure, including both deployment in restricted edge environments and on cloud. Expertise spans GenAI, MLOps, and large-scale ML pipelines, with a proven ability to optimize models for performance and efficiency and to communicate complex ML concepts to diverse stakeholders.

## Professional Experience

---

### AI/GenAI Consultant — Dev AI Ventures (Self-Employed)

Jan 2024 – Present

*Dev AI Ventures (Self-Employed)*

*Remote*

- Architected migration from OCR-heavy document processing to RAG-based knowledge systems for Docai.tools (prev. TenderFix) — retrieval architecture, chunking/indexing strategy, embedding-model selection, vector-store design, and private/on-prem deployment for sensitive document workflows.
- Delivered SOTA Vision-Language Model (VLM) research and technical recommendations for Drizz.dev, including preliminary evaluation findings.
- Advised Eventgraphia on deep-learning image-culling architecture and high-level strategy.

### ML Tech Lead

Dec 2023 – Mar 2026

*Verses Global BV (Project: dAIEdge)*

*Remote(Contractor)*

- Led development of an Active Inference-based routing agent leveraging the perception pipeline to perform obstacle avoidance, optimizing the planning stack(Jax based) and reducing planning time from 7 minutes to 21 seconds
- Developed a PyMDP-based saccading agent using a Tapo security camera for active visual exploration, enabling person detection and tracking and later forming the basis of a peer-reviewed conference publication
- Designed and maintained a 3D simulation pipeline in NVIDIA Isaac Omniverse to develop and evaluate routing agents in realistic warehouse environments before real-world deployment
- **Publication:** Towards smart and adaptive agents for active sensing on edge devices, **D. Vyas**, M. De Prado and T. Verbelen, HiPeac 2025 ([Link](#))

### Sr. ML Research Engineer

Dec 2023 – Dec 2025

*Verses AI*

*Remote(Contractor)*

- Designed and deployed an edge-ready 3D perception stack for reliable warehouse object detection and depth-aware navigation on Jetson-class hardware.
- Developed an YOLOv10n perception pipeline for warehouse-specific objects using a BlenderProc2-curated synthetic dataset, improving mAP from 0.15 to 0.62 at 640×640 and reducing the sim-to-real gap by retraining the synthetic-initialized model on cleaned, real warehouse data to reach 0.58 mAP
- Optimized the end-to-end detection stack (pre/post-processing + YOLOv10n inference) with TensorRT on Jetson Orin NX and desktop GPUs, achieving 26 ms latency at 640×640 for real-time edge deployment
- Integrated a DepthAnythingv2-based depth estimation module into the TensorRT perception pipeline, cutting depth latency from 1 s to 50 ms and delivering a combined depth + detection stack at 10 FPS (640×640) and 13 FPS (378×378) on Jetson Orin NX

### Machine Learning Engineer

May 2020 – Jan 2022

*Edge Case Research GmbH (Working Student)*

*Munich, Germany*

- Deployed an end-to-end ADAS perception pipeline with Multiple Object Tracking (MOT), extending 2D object detection to 3D bounding boxes on AWS
- Integrated monitoring and alerting for production ML models using Weights & Biases, with automated performance tracking across model versions to support data-driven iteration

### Machine Learning Engineer

Oct 2019 – Apr 2020

*TerraLoupe GmbH (Working Student)*

*Munich, Germany*

- Designed and automated an end-to-end MLOps pipeline for multi-terabyte aerial imagery segmentation using Docker and CI/CD, integrating Sacred/Omniboard for experiment tracking and reproducibility across 20+ model iterations
- Achieved 40% reduction in experimentation time through toolchain optimization
- Benchmarked GPU vs TPU setups on throughput, memory usage, and cost per inference, and used these results to define the production infrastructure strategy for large-scale segmentation

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |                            |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------|
| <b>ADAS Engineer</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                             | <b>Jan 2019 – Aug 2019</b> |
| <i>KPIT Technologies GmbH</i>                                                                                                                                                                                                                                                                                                                                                                                                                                                    | <i>Munich, Germany</i>     |
| <ul style="list-style-type: none"> <li>Architected Vehicle State Monitor for high-frequency multi-sensor data processing pipeline, handling 100+ Hz sampling rates</li> <li>Developed a GUI dashboard to monitor and visualise real-time vehicle health from streaming telemetry</li> <li>Developed and Automated comprehensive testing framework for vehicle telemetry system</li> </ul>                                                                                        |                            |
| <b>Software Engineer</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                         | <b>Jan 2017 – Aug 2018</b> |
| <i>CNRS (XLIM Lab)</i>                                                                                                                                                                                                                                                                                                                                                                                                                                                           | <i>Poitiers, France</i>    |
| <ul style="list-style-type: none"> <li>Optimized radio propagation simulation engine through algorithm profiling and memory-efficient implementations</li> <li>Achieved 30% runtime reduction and improved numerical stability, enabling larger-scale IoT network simulations</li> <li><b>Publication:</b> CupCarbon: A new platform for the design, simulation and 2D/3D visualization of radio propagation and interferences in IoT networks (<a href="#">Link</a>)</li> </ul> |                            |

|                                                                                                                                                                                                                                                                                                                                                                                                                                                          |                                                                                                                   |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|
| <b>Research Experience</b>                                                                                                                                                                                                                                                                                                                                                                                                                               | <b>Google Scholar Profile</b>  |
| <b>Synthetic Adversarial Generation with Diffusion Models [Master's Thesis]</b>                                                                                                                                                                                                                                                                                                                                                                          | <b>Feb 2023 – Aug 2023</b>                                                                                        |
| <i>Fraunhofer AISEC — Grade: 1.0/5.0</i>                                                                                                                                                                                                                                                                                                                                                                                                                 | <i>Munich, Germany</i>                                                                                            |
| <ul style="list-style-type: none"> <li>Generated DDPM-based synthetic adversarial datasets for CIFAR-10 robustness re-training, improving ResNet18 accuracy by 8-10%</li> </ul>                                                                                                                                                                                                                                                                          |                                                                                                                   |
| <b>Fraunhofer AISEC Research (Robustness/Privacy)</b>                                                                                                                                                                                                                                                                                                                                                                                                    | <b>Feb 2022 – Sep 2023</b>                                                                                        |
| <i>Fraunhofer AISEC</i>                                                                                                                                                                                                                                                                                                                                                                                                                                  | <i>Munich, Germany</i>                                                                                            |
| <ul style="list-style-type: none"> <li>Led robustness/privacy research on CNNs, Lipschitz constants, and differential privacy threats; co-authored pre-print with automated containerized CI/CD pipelines for reproducible evaluation.</li> <li><b>Pre-Print:</b> Gradient Masking and the Underestimated Robustness Threats of Differential Privacy in Deep Learning, F Boenisch, P Sperl, <b>D Vyas</b>, K Böttinger (<a href="#">Link</a>)</li> </ul> |                                                                                                                   |
| <b>Pedestrian Segmentation &amp; LiDAR PoC</b>                                                                                                                                                                                                                                                                                                                                                                                                           | <b>Nov 2020 – Mar 2021</b>                                                                                        |
| <i>Warp Automotive, UnternehmerTUM</i>                                                                                                                                                                                                                                                                                                                                                                                                                   | <i>Munich, Germany</i>                                                                                            |
| <ul style="list-style-type: none"> <li>Developed semantic segmentation for urban pedestrians and LiDAR point clouds; refactored pipeline with MLOps (multi-stage Docker, experiment tracking, Go-Task automation)</li> </ul>                                                                                                                                                                                                                             |                                                                                                                   |
| <b>3D Medical Image Reconstruction</b>                                                                                                                                                                                                                                                                                                                                                                                                                   | <b>Apr 2020 – Jul 2020</b>                                                                                        |
| <i>CAMP Chair, TUM</i>                                                                                                                                                                                                                                                                                                                                                | <i>Munich, Germany</i>                                                                                            |
| <ul style="list-style-type: none"> <li>Built U-Net for multi-coil MRI 3D reconstruction (3rd place globally); researched NeRFs for CT-from-XRay reconstruction</li> <li><b>Publication:</b> Multi-Coil MRI Reconstruction Challenge, <i>Frontiers in Neuroscience</i> (<a href="#">Link</a>)</li> </ul>                                                                                                                                                  |                                                                                                                   |

Education

|                                                    |                            |
|----------------------------------------------------|----------------------------|
| <b>Technical University of Munich</b>              | <b>Oct 2019 – Aug 2023</b> |
| <i>M.Sc. Informatics   Majors: ML and CV</i>       | <i>Munich, Germany</i>     |
| <b>The LNM Institute of Information Technology</b> | <b>Jul 2012 – Jun 2016</b> |
| <i>B.Tech Computer Science</i>                     | <i>Jaipur, India</i>       |

Technical Skills

|                                |                                                                                     |
|--------------------------------|-------------------------------------------------------------------------------------|
| <b>Gen AI:</b>                 | vLLM, Ollama, n8n, LoRA, RAG pipelines, agent workflows                             |
| <b>Enterprise Integration:</b> | REST APIs, knowledge systems, documentation workflows, Jira/Confluence integrations |
| <b>MLOps and Infra:</b>        | FastAPI, WandB/MLFlow, Docker, Git, CI/CD(GitHub Actions, Jenkins), Kubernetes      |
| <b>Cloud:</b>                  | Google Cloud Engine, AWS                                                            |
| <b>Databases:</b>              | PostGres(pgvector), MongoDB                                                         |
| <b>Libraries:</b>              | PyTorch, Jax, Tensorflow, Pandas, OpenCV ROS/ROS2                                   |
| <b>Languages:</b>              | Python, C++, MATLAB, Java, bash                                                     |

|                                                                                                                                                                                                             |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Misc.</b>                                                                                                                                                                                                |
| <ul style="list-style-type: none"> <li><b>Languages:</b> English(Bilingual/Native), Hindi(Native), German(Intermediate), French(Basics)</li> <li>Stand-up comedian, (ex)Podcaster, Mentor TUM.ai</li> </ul> |