Abdalrahman Ibrahim

Head of AI Products Development

Linz, Austria • Open to remote / relocation

abdalrahman.m5959@gmail.com • +43 660 622 8105 • github.com/geekgineer • linkedin.com/in/abdalrahman-m-amer • geekgineer.com

SUMMARY

AI platform and products leader with production experience in private LLM infrastructure, agentic systems, and edge AI for autonomous robots. Architected and now lead the GenAI platform at AGILOX — a private microservices stack serving 250+ employees with an internal AI gateway handling around 27M tokens per week. Author of **YOLOs-CPP** (around 900 stars), a widely used C++ inference engine for YOLO models. PhD candidate at the University of Klagenfurt (Agentic AI for Robotics & Transportation).

EXPERIENCE

Head of AI Products Development Jan 2026 – Present
AGILOX (AMR Robotics) Linz, Austria

- Lead AI products and platform across the company. Own the Orbit private LLM platform (12 microservices), the Keysmith AI gateway, and the cross-org AI products roadmap.
- Set technical direction for private GenAI infrastructure, agentic systems, on-robot AI, and the underlying observability and cost-attribution layer.

AI Lead Engineer May 2023 – Jan 2026
AGILOX (AMR Robotics) Linz, Austria

- Architected **Orbit**, a private LLM microservices platform serving **250+ employees**. Migrated from commercial APIs to self-hosted, quantized open-weight models — cut monthly LLM spend from **$8,000 to €1,500 (around 80%)** while securing full data sovereignty.
- Built **Keysmith**, an internal AI gateway issuing virtual API keys for developer and agentic-coding workflows. Routing, fallbacks, per-team cost attribution, observability dashboards. Representative 7-day window (Nov 2025): **12,970 requests, 27.3M tokens, 93.5% success**, $0.0002 average cost per request.
- Designed a custom agentic orchestration layer for intent-aware routing and **sub-200ms token streaming**, sidestepping the overhead of heavy frameworks.
- Shipped a **Hybrid Graph+Vector RAG** over complex technical documentation, with evaluation-driven tuning loops that measurably reduced hallucinations in internal engineering support.
- Optimized private model serving with **vLLM** and **AWQ quantization** for high-throughput local execution of 70B+ open-weight models on private GPU clusters (Runpod).
- **VAgent**: integrated LLMs with ROS 2 telemetry over the Model Context Protocol (MCP), enabling natural-language fleet diagnostics and real-time fleet health monitoring for AMR operations.

Senior Research Engineer Oct 2022 – May 2023
AGILOX (AMR Robotics) Linz, Austria

- Delivered production LiDAR + Camera perception stacks for autonomous mobile robots; optimized edge-hardware deployment and built simulation-driven data pipelines that cut physical sensor testing time by around 40%.

Researcher Oct 2021 – Oct 2022
Infineon Technologies Villach, Austria

- Graph Neural Networks for automated analog circuit recognition: **+15% accuracy** via small-data augmentation, **-30% compute time**. Published in IEEE Access (2024).

Robotics & AI Engineer 2018 – 2021
Zewail City / iRobotX Egypt

- ROS-based social robots and deep-learning pipelines for medical CT imaging; evaluated Visual SLAM and reinforcement-learning

control for humanoid robotics.

SELECTED OPEN-SOURCE PROJECTS

YOLOs-CPP (around 900 stars) — Cross-platform, production-ready C++ inference engine for YOLO v5–v12 and YOLO26 on ONNX Runtime. Unified API for detection, segmentation, pose estimation, OBB, and classification.

YOLOs-CPP-RT — CUDA-accelerated single-header variant of YOLOs-CPP for real-time inference.

ros2_yolos_cpp — ROS 2 adapters for YOLOs-CPP, deployed in mobile-robot perception stacks.

SmolVLM2-ROS 2 — Modular ROS 2 package for SmolVLM2 vision-language inference, strategy-pattern architecture.

Depths-CPP — Depth-Anything-V2 via ONNX Runtime, single-header C++.

CloudPeek — Single-header C++ point cloud viewer with zero PCL or Open3D dependency.

PUBLICATIONS

FlexiNet: Real-Time Ego Vehicle Speed Estimation. IEEE Access, 2025. doi.org/10.1109/ACCESS.2025.3562229.
A GNN System for Analog Circuits' Structure Recognition. IEEE Access, 2024. doi.org/10.1109/ACCESS.2024.3367598.

EDUCATION

Ph.D. Candidate, Smart Systems — In progress
Alpen-Adria-Universität Klagenfurt, Austria — Part-time
Focus: Agentic AI for Robotics & Transportation.

M.Sc. Autonomous Systems & Robotics — 2022
Alpen-Adria-Universität Klagenfurt, Austria

B.Sc. Mechatronics Engineering — 2018
Misr University for Science and Technology, Egypt
First of class, CGPA 3.96/4.0, Honor Degree.

SPEAKING & EXTERNAL

NVIDIA GTC Paris 2025 — AGILOX exhibit, on-robot AI for AMRs.
Interlogistics Applications Fair — Live SAM2 segmentation demo.

TECHNICAL STACK

AI Infrastructure: Private LLM hosting, vLLM, AWQ/GPTQ quantization, AI gateways, cost attribution, observability
Agentic & RAG: Custom orchestration (no LangChain), MCP, intent routing, token streaming, Hybrid Graph+Vector retrieval, eval-driven tuning
Cloud & DevOps: Azure, Docker, GitOps, CI/CD, microservices, Terraform
Robotics: ROS 2, LiDAR/Camera perception, sensor fusion, edge deployment (Jetson-class)
Languages: C++ (modern), Python, CUDA, Bash
Libraries: ONNX Runtime, PyTorch, TensorRT