Enes Uzun

AI Infrastructure Engineer *LLM inference, GPU platforms, agent tooling*

uzunenes.com enes@uzunenes.com linkedin.com/in/uzunenes github.com/uzunenes
Istanbul, Türkiye · Open to relocation worldwide

15,000	250	250+	8
employees use the on-prem LLM chatbot	developers on self-hosted coding models	camera streams analysed in real time	years in production ML

Summary

AI infrastructure engineer at Ford Otosan (Ford Motor Company joint venture) with 8 years of production ML work. I run the on-prem GPU cluster and LLM serving stack used by 15,000 employees and 250 developers. Before that I built real-time computer vision systems in C/C++ that run in four car plants. My recent open-source work is on the reliability of coding agents with self-hosted models.

Experience

Senior Software Engineer
Dec 2022 – Present

Ford Otosan (Ford Motor Company joint venture), Türkiye

- Designed and built the company's LLM gateway. It decides per request whether the data may leave the company: confidential code and documents go to open models on our own GPUs, the rest to Azure OpenAI. Access goes through Entra ID (OAuth2/OIDC).
- Deployed vLLM on on-prem GPUs behind OpenAI-compatible endpoints, so 250 developers use Copilot CLI with internal models (600+ agent requests a day) without code leaving the network or per-token API costs.
- Run the multi-site Kubernetes GPU cluster (5 servers; 2× H200, 4× A100, 4× L40S, 4× V100) with Prometheus and Grafana monitoring. It hosts the internal chatbot (15,000 users), the coding models and the production vision models.
- Turned MobileAI, our vision system in four plants, into a no-code platform where line engineers collect data, label and train models themselves. New models go live only after engineering review and are monitored for drift.
- Lead technical direction for a team of 9. The team's work received 2 global and 2 local Ford innovation awards.

Software Engineer, Network & Systems
May 2019 – Dec 2022

Ford Otosan, Türkiye

- Built Gözcü, a no-code platform for real-time accident detection on CCTV, deployable to a new camera in one click.
- Wrote its inference backend in C/C++ (GStreamer, OpenCV, YOLO): 250+ concurrent RTSP streams, under 200 ms end-to-end latency.
- Gözcü received Ford's President's Health & Safety Award in 2023.

Software Engineer
Jun 2018 – May 2019

Evstek Information Technologies, Istanbul

- Developed the image-analysis software of KE-BOT, a robotic system for FUE hair transplantation (C++, OpenCV). The work was published at IEEE.

Open source

jcode, an agentic coding CLI in Rust
Aug 2026

3 bugs reported and root-caused, all fixed upstream

- Agents on self-hosted models (vLLM, SGLang, Ollama) stalled after context compaction. Traced it to custom providers being routed through the OpenRouter runtime, which dropped orphaned tool outputs (#908).
- Models without a declared `input` list were treated as image-capable, causing an unrecoverable HTTP 400 loop. Isolated it with an A/B matrix; the text-only default I proposed was adopted (#847).
- Shifted symbols were lost on Turkish Q keyboards in the VS Code terminal. Identified the missing Kitty `REPORT_ALTERNATE_KEYS` flag (#870).

Ultralytics YOLO May 2025

PR #20553, merged in release 8.3.131

- Co-authored, with three others, the official YOLO11 example for Triton Inference Server in C++: gRPC client, FP16 preprocessing, NMS, CMake build.

Own projects github.com/uzunenes

- triton-server-hpa: scales Triton on Kubernetes by GPU utilisation (DCGM metrics, Prometheus adapter, HPA), with GPU time-slicing so it runs on one GPU.
- k8s-ai-stack: self-hosted LLM stack on Kubernetes with vLLM, Ollama and OpenWebUI (OAuth2), plus throughput benchmarks.
- puci: small AI coding-agent harness in C17; about 100 kB per session, runs 1,000 agents on one `poll()` loop for evaluation.
- piciem, librtsplink, libmqttlink: image processing in plain C (DFT/FFT, convolution), RTSP and MQTT client libraries.

Skills

- **Inference:** vLLM, SGLang, Triton Inference Server, TensorRT, OpenAI-compatible APIs
- **Platform:** Kubernetes, OpenShift, Docker, ArgoCD, NVIDIA DCGM, Prometheus, Grafana, MLflow, GitHub Actions, Azure DevOps
- **Languages:** Python, C, Bash (expert); C++, Go; Rust (reading and debugging)
- **ML and vision:** PyTorch, YOLO, RT-DETR, OpenCV, GStreamer
- **Systems:** Linux, strace, perf; TCP/UDP, gRPC, MQTT, RTSP
- **Spoken:** Turkish (native), English (professional)

Publications and talks

- *Autoencoder-based video anomaly detection.* 2nd International Conference of Engineering Sciences (ICES), Dec 2023.
- Image analysis for the KE-BOT hair transplantation robot. IEEE Xplore, document 8719213.
- *Multi-site on-prem AI workloads on Kubernetes.* Talk at Kubernetes Istanbul (CNCF community), Jul 2026.

Awards

Global Manufacturing Technical Excellence Award Apr 2025

Ford Motor Company

- For robotic vision in global manufacturing.

President's Health & Safety Award Jun 2023

Ford Motor Company

- For combining AI, computer vision and collaborative robots on the line. Signed by CEO Jim Farley.

Education

M.Sc. Electronics Engineering, coursework completed 2019 – 2023

Gebze Technical University

- Research on unsupervised video anomaly detection with autoencoders (ICES 2023 paper above).

B.Sc. Electronics and Telecommunication Engineering 2013 – 2018

Kocaeli University

- Thesis: real-time lane departure warning on an embedded platform.

Courses

Fast and Efficient LLM Inference with vLLM (Red Hat, 2026); Accelerate Your Model (Google DeepMind on DataCamp, 2026); Executive Leadership (Koç University, 2025); Certified Scrum Team Member (Scrum Inc., 2022).