

Lounis Hamroun

Software / ML Engineer | Paris, France | Open to relocation | 6-month internship from Jan. 2027

lounis.hamroun@gmail.com | lounishamroun.com | [GitHub](#) | [LinkedIn](#)

EDUCATION

Télécom Paris

Mastère Spécialisé IA Expert Data & MLOps

Palaiseau, France

2025 – 2027

IMT Nord Europe

Engineering Degree in Computer Science

Lille, France

Sept. 2022 – Sept. 2025

SELECTED TECHNICAL PROJECTS

GPT-2 From Scratch | *PyTorch, Transformers* | *Article*

Summer 2026

- Reimplemented GPT-2 inference from first principles in PyTorch, including causal self-attention, transformer blocks, weight loading, and autoregressive generation.
- Validated the implementation layer by layer against Hugging Face GPT-2 by numerically comparing intermediate activations with tolerance-based tests.
- Built component-level tests and debugged issues involving tensor shapes, causal masking, parameter loading, and weight sharing.

GPU Kernel Optimization with Triton | *PyTorch, Triton, GPU Benchmarking* | *Article*

2026 – Present

- Implemented a row-wise softmax kernel in Triton and benchmarked it against `torch.softmax` and a naive PyTorch baseline on an RTX 5090 across varying row widths using CUDA-event timing, warmups, and repeated measurements.
- Characterized a shape-dependent performance crossover: PyTorch led for small and medium row widths, while Triton became faster after a sharp performance drop near $N \approx 8700$; examined the coinciding `BLOCK_SIZE` jump from 8192 to 16384.

Renault AutoVI – Visual Anomaly Detection | *PyTorch, PatchCore*

Dec. 2025 – Jun. 2026

- Built an unsupervised PatchCore pipeline for a Télécom Paris industry project with Renault, covering six automotive-part inspection scenarios.
- Implemented frozen ResNet50 feature extraction, patch embeddings, memory-bank retrieval, and L2 nearest-neighbor anomaly scoring.
- Evaluated defect detection using PR-AUC and ROC/Youden thresholding.

EXPERIENCE

Air Liquide

Apprentice MLOps Engineer

Paris La Défense, France

Jan. 2025 – Aug. 2025

- Refactored an AWS Bedrock-based industrial AI assistant into a modular Model Context Protocol (MCP) client-server architecture, separating conversation logic from business-tool execution.
- Implemented MCP tools connecting the assistant to internal IoT business services, creating reusable interfaces for tool integrations.
- Contributed to the infrastructure design of a GCP-based RAG system using Vertex AI, Cloud Storage, Terraform, and Azure DevOps CI/CD.

TECHNICAL SKILLS

Programming: Python, SQL, C

ML / AI: PyTorch, Triton, Hugging Face Transformers

Systems & Performance: Linux, Docker, Git, GPU profiling, benchmarking

Cloud / MLOps: AWS, GCP, Vertex AI, Terraform, Azure DevOps CI/CD