Tito Nicola Drugman

TitoNicolaDrugman | titonicoladrugman | drugmantito@gmail.com | Milan, Italy

Summary

M.Sc. student in Computer Science & Engineering (AI track) at Politecnico di Milano, with a research interest in information retrieval and in the deployment of language models under resource and confidentiality constraints. My recent work sits at the intersection of the two: a fully on-premises RAG pipeline for code generation over a 7,600-snippet knowledge base, comparing lexical, dense and hybrid retrieval and two multi-hop strategies; an automated labelling and page-state classification pipeline for a vision-language web agent, built to remove manual annotation from the evaluation loop; and a 4.56M-parameter encoder-only Transformer quantised to INT8 and deployed on an STM32N6 NPU, which is the same compression problem that governs local LLM inference at a larger scale. I run open-weight models locally on a dual-GPU Linux workstation and work daily on Ubuntu. From October 2026 to March 2027 I will be an AI trainee at the European Court of Auditors in Luxembourg, working on on-premises LLM deployment, RAG, OCR and document classification.

Education

2024 - present	M.Sc. in Computer Science & Engineering (Artificial Intelligence) at **Politecnico di Milano**
Fall 2025	Exchange semester at **The Chinese University of Hong Kong, Shenzhen**
2021 - 2024	B.Sc. in Artificial Intelligence at **Università di Pavia**, **Università Milano Statale**, **Università Milano Bicocca**
2019 - 2020	Academic year abroad at **LaSalle High School** (Ohio, USA)

Projects

RAG for LLM-Based Code Generation [GitHub] Sep 2025 - Feb 2026

- Designed a reproducible ingestion, chunking and indexing pipeline over a knowledge base of 7,600+ code snippets, corpus-agnostic by design and transferable to technical-documentation and audit domains.
- Benchmarked three single-hop retrievers (lexical BM25, dense CodeBERT 768-d embeddings, hybrid via Reciprocal Rank Fusion) and two LLM-driven multi-hop pipelines (query decomposition, iterative refinement).
- Deployed the full system on-premises on local Linux hardware (2× RTX 5060 Ti) with zero external API calls, using Hugging Face Transformers and 4-bit NF4 quantisation (bitsandbytes) to fit CodeLlama-7B-Instruct in available VRAM.
- Semantic retrieval reached 0.2567 CodeBLEU against a 0.1859 no-retrieval baseline (+38% relative); multi-hop decomposition reached 0.2505.
- Engineering: deterministic generation, resume-safe execution, OOM handling, version-controlled configurations.

Autonomous Web Agents (WebVoyager) [GitHub] Sep 2025 - Jan 2026

- Replaced proprietary GPT-4o with locally hosted open-weight models: Qwen3-VL:8b and Qwen3-VL:30b served via Ollama, with DeepSeek-v3 (685B MoE) accessed through the OpenRouter API as a third comparison point.
- Built a fully automated labelling pipeline using locally hosted Qwen3-VL:30b to classify webpage screenshots into 12 semantic categories, producing a 1,101-image dataset with zero manual annotation.
- Fine-tuned a ResNet-18 page-state classifier and integrated it as a visual prior in the agent loop, cutting reasoning latency by 17% (34.5% overall accuracy, 56.0% excluding the dominant HOMEPAGE class).
- Co-authored the grayscale + colored Set-of-Mark grounding study (52.4% vs. 51.1%) and the GPT-4o reference replication (62.7% task success vs. 59.1% in the original paper).
- Stack: Python, Selenium, Ollama, PyTorch, Hugging Face, OpenRouter.

On-Device Transformer on an STM32N6 NPU [GitHub] Mar 2025 - Sep 2025

- Designed, trained and INT8-quantised a custom 4.56M-parameter encoder-only Transformer from scratch for autoregressive text generation, under strict flash and memory budgets.
- Applied per-channel symmetric INT8 quantisation of both weights and activations using a held-out calibration set, a methodology directly transferable to compressing larger LLMs for on-premises inference.
- Engineered the end-to-end pipeline: model design in PyTorch, training on $\sim$1.4M words from Project Gutenberg, quantisation calibration, C-code generation via STM32CubeIDE / X-Cube-AI, NPU configuration and on-device validation.
- On-board results: 73.14 ms/token latency, 13.67 tokens/s throughput, 257/407 operators accelerated on the NPU; produced a quantitative CNN-vs-Transformer comparison under identical hardware constraints.

GenHack 2025: Urban Heat Island Bias Correction [GitHub] Oct 2025 - Nov 2025

- Three-person hackathon organised by École Polytechnique, BNP Paribas and Kayrros: built an ML pipeline correcting systematic Urban Heat Island bias in 9 km ERA5-Land reanalysis data.
- Fused four heterogeneous sources (ERA5-Land, Sentinel-2 NDVI, ECA&D ground stations, JRC Global Surface Water) into a unified 3M+ row training set across IT/ES/FR/PT.
- Compared Random Forest and HistGradientBoosting regressors with a station-based split to prevent geographic leakage; engineered a land-cover segmentation step to resolve urban/water spectral ambiguity.
- Reduced raw ERA5 error from 3.01°C to 1.94°C RMSE (-35.5%), MAE by -45% and systematic urban bias by -93% on hold-out stations; feature importance was dominated by elevation and coordinates rather than urban fraction. Stack: Python, scikit-learn, GeoPandas, Xarray, Dask.

Wildfire Segmentation (Sen2Fire) [GitHub] 2025

- Benchmarked nnU-Net, SegFormer, Dual-Stream SegFormer and hybrid Transformer-CNN architectures on Sen2Fire (Sentinel-2 multispectral and Sentinel-5P aerosol imagery).
- Reached an F1-score of 0.3675 with nnU-Net against a 0.2810 published baseline and challenged the original study's conclusion by showing that robust architectures exploit the full 13-band input rather than hand-selected band subsets.
- Implemented Test Time Augmentation and analysed its accuracy-latency trade-off.

Work Experience

AI Engineer Intern, STMicroelectronics Mar 2024 - Jul 2024

- Designed and implemented an automated hyperparameter optimisation framework in Python, extending the Keras Tuner library, for neural networks deployed on microcontroller-class hardware, targeting an industrial audio anomaly-detection use case for predictive maintenance.
- Engineered the search procedure to evaluate candidate architectures jointly against accuracy and strict embedded constraints (ROM footprint, RAM usage, Multiply-Accumulate operations), so that only deployable models are returned.
- Validated the framework's ROM/RAM/MACC estimators against STM32Cube.AI ground-truth measurements on three MLPerf Tiny reference models (anomaly detection, image classification, keyword spotting).
- Demonstrated the end-to-end constraint-aware search on two MLPerf Tiny tasks, producing smaller constraint-compliant models with performance comparable to the larger MLPerf baselines.
- Diagnosed and engineered a workaround for an upstream Keras Tuner bug that broke long unattended remote tuning runs under hardware-constraint filtering.
- Worked on internal Linux workstations over corporate VPN in a confidential environment; maintained reproducibility through version-controlled Conda environments and systematic experiment logs.

Robotics Trainer & Certification Assessor, Fanuc-Sanoma / Comau-Pearson Oct 2022 - present

- Deliver technical training and administer official industry certifications on industrial communication protocols, robot programming and Industry 4.0 to mixed audiences of engineers, educators and students across Northern Italy.
- Program and configure industrial robots and validate motion programs in RoboGuide; run hands-on sessions on Comau e.DO cells.
- Produce structured course materials and assessment criteria and lead "Train the Trainer" sessions.

Science Museum Guide, Leonardo da Vinci National Museum Mar 2023 - present

- Lead tours in the Leonardo da Vinci Galleries and the Air and Transport Pavilion, including the official narration inside the Enrico Toti submarine.
- Communicate complex technical and historical concepts to a diverse public audience.

Languages, Skills & Additionals

Languages	Italian (Native), English (C1).
LLMs, RAG & IR	RAG (BM25, dense CodeBERT embeddings, hybrid via Reciprocal Rank Fusion), multi-hop retrieval (query decomposition, iterative refinement), document ingestion/chunking/indexing, prompt engineering, Transformer architectures, CodeLlama, Qwen3-VL, DeepSeek-v3, Llama-3.
Local & On-Prem AI	Ollama, llama.cpp, Hugging Face Transformers (local inference), bitsandbytes (4-bit NF4), local vision-language model deployment, OpenRouter, API-free pipelines.
ML & Computer Vision	PyTorch, TensorFlow, Keras, scikit-learn, deep learning, quantisation-aware deployment, image classification, semantic segmentation.
Embedded AI	ST X-Cube-AI, STM32CubeIDE, STM32N6570-DK, TFLite INT8 quantisation, MLPerf Tiny.
Systems & Tooling	Ubuntu Linux (daily driver), SSH, Bash, Docker, Podman (rootless containers), Git, GitHub, Selenium, Anaconda, Jupyter, LaTeX.
Programming & Data	Python, C, Java, MATLAB, SQL, NumPy, Pandas, Matplotlib, GeoPandas, Xarray, Dask, MongoDB, Neo4j, Redis, Cassandra, YAML/JSON.
Robotics & Training	FANUC, COMAU, RoboGuide; certified technical instructor & certification examiner.
Interests	On-device & local AI, geopolitics, public transport, travelling.