Aton Kamanda

https://atonkamanda.github.io/ | atonkamanda@hotmail.com | https://github.com/atonkamanda | 438 543-4133

Education

University of Namur — Namur
Bachelor of computer science with distinction (minor in mathematics) — *Sept. 2018 – Sept. 2021*

University of Montreal — Montreal
Master of artificial intelligence, 3.95/4.3 GPA — *Sept. 2021 – Aug. 2023*

Experience

Generative AI engineer — October 2024 – Now
Alexa Translations — *Canada, Montreal*

- Implemented a LLM agent workflow leveraging AWS Bedrock & SageMaker for machine translation, processing 10,000+ legal documents monthly while maintaining +90% accuracy on specialized legal terminology.
- Spearheaded the development of a custom OCR platform by fine-tuning image-to-text models for low-resource languages, significantly improving text extraction accuracy on scanned documents.
- Orchestrated comprehensive human feedback collection and training pipeline to develop LLM translation quality evaluators, achieving 90% agreement with human judgment quality through state-of-the-art LLM-as-judge methodology and fine-tuned LLAMA 3 models. The resulting evaluator was subsequently used to clean and validate 10 million translation segments or to correct the translation engine.

Machine learning engineer — March 2024 – September 2024
AwakeAI - Mila incubated startup — *Canada, Montreal*

- Developed a video understanding system based on V-JEPA for real-time activity recognition in nursing homes achieving state-of-the-art performance of 81% on the Toyota Smarthome benchmark.
- Demonstrated model adaptation skills by fine-tuning the V-JEPA architecture on proprietary data, resulting in 80% accuracy on complex action recognition tasks on real-world scenarios. Enhanced system capabilities by seamlessly integrating YOLOv8 and StrongSORT for robust multi-person tracking and action recognition.
- Implemented a highly scalable Kafka-based real-time video streaming pipeline, using inference optimization techniques that enabled processing of tens of simultaneous camera streams on a single A100 at 90 FPS.

Machine Learning Engineer — May 2023 – February 2024
VMware - PhD project with IVADO — *Canada, Montreal*

- Developed a personalized code-completion LLM trained on private codebase by implementing a retrieval-augmented generation system with CodeLlama and ChromaDB, achieving 0.1-10ms query latency on a 50GB vector database.
- Optimized CodeLLAMA 70B model through variational dropout and pruning techniques, resulting in 4x faster inference and 10x memory reduction while maintaining performance by implementing advanced techniques such as variational dropout, pruning, and sparse matrix representations.

Teacher assistant for a graduate deep RL for robots class — January 2023 – May 2023
Mila - Montreal institute for learning algorithms — *Canada, Montreal*

- Course focused on deep RL for robotics and composed mainly of PhD students, I have been in charge of creating entirely new assignments with recent research papers, writing automated tests on Gradescope, grading students, and helping students in their research contributions for the final project. (Course website).

Publications

CodeUltraFeedback: LLM-as-a-Judge for coding preferences alignment | *ACM 2024* — March 2024

- Pioneered CodeUltraFeedback, a 10,000-instruction dataset for LLM alignment, and CODAL-Bench for assessing coding preferences; demonstrated that CodeLlama7B-Instruct, fine-tuned with this data using DPO, outperformed 34B LLMs on CODAL-Bench and improved HumanEval+ functional correctness.The project has reached 60+ stars on github.

Technical strengths

Languages : Python, Julia, C/C++, R, SQL
Data & Developer Tools: Spark, Hadoop, Pandas, AWS, GCP, Azure, Kafka, Docker, Kubernetes
Machine learning: Pytorch, Jax, TensorFlow, MLFlow, Triton, NumPy, Gym, Mujoco, TensorRT, DSPy