Wissam **Antoun**

Machine Learning Engineer

Paris, France

(+33) 7 69 81 26 51 | wissam.antoun@gmail.com | WissamAntoun | wissamantoun | Wissam Antoun

wissamantoun.com

Summary

Artificial Intelligence and Machine Learning Engineer specialised in state-of-the-art Natural Language Processing techniques and pretraining LLMs. I'm now a **PhD Candidate in the ALMAnaCH team at Inria-Paris**, continuing my research on LLMs and AI weaponization and safety. I have published and released several state-of-the-art French LLMs such as the **Gaperon French LLM Suite (1.5B, 8B, and 24B trained from scratch), ModernCamemBERT, CamemBERTa (v1 and v2), and CamemBERT v2**. During my masters, I **published AraBERT**, the first and most influential AI models for Arabic text representation, which is now the **most cited Arabic AI paper and GitHub repository**. In 2020, I also **developed AraGPT2 (1.5B), the first Arabic LLM and the first LLM created outside of US and China**. I also ranked in top places in almost all AI competitions that i participated in and was nominated for the Abdul Hadi Debs Endowment Award for Academic Excellence.

In addition, I have experience in teaching as a Graduate Teaching Assistant at the Eng. faculty at the American University of Beirut and as an AI instructor Zaka. After my masters, I joined Siren Analytics as an ML Scientist where I worked on various NLP, IR and Anomaly Detection problems that directly impacts the lebanese public sector. I'm self-motivated, self-taught, and a performer when working on tight deadlines. I'm also a tech geek who loves computer hardware and technology and enjoys self-hosting my own services, DIY projects, and gaming.

Work Experience

ALMAnaCH, Inria

Paris, France

PhD Researcher — *Mar. 2023 -*

- As part of a 2 person team, we trained Gaperon. A French LLM suite with 1.5B, 8B, and 24B trained from scratch for up to 4T tokens. I was responsible for the data curation, part of the pretraining, and SFT.
- Evaluated ModernBERT against DeBERTaV3 to validate architectural claims, resulting in the release of ModernCamemBERT, the first non-English (French) ModernBERT model.
- Studied the effect of scale, model family and instruction tuning on machine-generated text detection. I also have the earliest work on detecting French AI-generated text.

Research Enginner — *Mar. 2022 - Feb. 2023*

- Implemented the pre-training loss and architecture of DeBERTaV3 and trained CamemBERTv2 and CamemBERTa(v2) which the SOTA French encoder model.
- Researching language models for languages displaying high variability, in particular Arabic dialects used on social media.

Siren Analytics

Beirut, Lebanon

Senior Machine Learning Scientist — *Jan. 2021 - Feb. 2022*

- Demoed an early version of a RAG in 2021, before it became mainstream.
- Researched and developed semantic search engines, created embeddings and reranking models.
- Built an OCR API for ID card scanning.
- Designed and developed document AI solutions to extract information from scanned documents.
- Developed a real-time anomaly detection system for covid lockdown mobility permit.
- Developed encryption solutions for model IP protection.
- My work also involved researching and applying latest AI solutions, evaluating software solution vendors, and internship mentoring.

MIND Lab - American University of Beirut

Beirut, Lebanon

Graduate Research Assistant — *Sep. 2018 - Jan. 2021*

- Co-founded AUB's Machine INtelligence Development (MIND) Lab `https://sites.aub.edu.lb/mindlab`
- My research focused on state-of-the-art NLP technologies for natural language processing, generation, and chatbots.
- In early 2020 I developed AraBERT: The first Arabic BERT. The model is now the most cited Arabic AI paper, and most starred Arabic Github Repository. It had reach 10M+ downloads on Hugginface with 500K+ downloads in the last month. I introduced a Arabic specific tokenization technique which adds morphological segmentation prior to WordPiece tokenization.
- In late 2021, I released 13 models including an updated version of AraBERT and a large one, an Arabic Electra mode, and AraGPT2 a suite of Arabic LLMs up to 1.5B params. It was the first Arabic LLM, and the first LLM released outside of US and China. These models were trained and released in the span of 3 months using 128-v3 TPUs provided by Google.
- Researched improving temporal and spacial resolution of remote sensing ML systems for soil moisture prediction.

ECE Dept. - American University of Beirut

Beirut, Lebanon

Graduate Teaching Assistant

Feb. 2018 - Jan. 2020

- EECE 435L: Software Tools Lab, Co-Instructor (3 semesters). Delivered a range of teaching and assessment activities including tutorials directed towards the delivery of modern Software Tools at the undergraduate level.
- EECE 696: Applied Parallel Programming, Lab instructor (1 semester). Delivered hands-on experience to graduate students on parallel programming and GPU computing using the CUDA C/C++ framework on local computers and on AWS.
- EECE 330: Data Structures and Algorithms, Lab instructor (1 semester). Delivered hands-on experience on fundamental algorithms and data structures used in software applications at the undergraduate level using C++.
- Introduced a plagiarism detection tool for computer code to the Electrical Engineering Department.

Academic Reviewer

Reviewer for the following venues:

2020 - Current

- Major NLP conferences: NACL, EMNLP, COLING, LREC
- Arabic Natural Language Processing Conference
- Journal of King Saud University - Computer and Information Sciences (IF: 13.47)
- Expert Systems with Applications (IF: 6.95)
- Workshop on Open-Source Arabic Corpora and Processing Tools
- ACM Transactions on Asian and Low-Resource Language Information Processing - TALLIP (IF: 1.42)
- Language Resources and Evaluation Journal (IF: 1.36)

Zaka

Online

AI Instructor

May. 2020

- Gave a 4-part online workshop series titled *A Comprehensive NLP series* that was focused on delivering the major NLP concepts through hands-on coding examples along with the NLP fundamentals theory.

Stars of Science

Doha, Qatar

Machine Learning Expert

Feb. 2020 - Apr. 2020

- Provided Machine Learning and Artificial Intelligence solutions as part of the support team that helps contestants throughout their Stars of Science journey

Huawei Technologies (Lebanon) S.A.R.L.

Beirut, Lebanon

Intern

June 2016 - Aug. 2016

- Learned about various LTE principles, base station hardware, and Microwave transmissions systems.

Education

Masters Of Engineering In Electrical And Computer Engineering

Beirut, Lebanon

American University of Beirut

Feb. 2018 - Sep. 2020

- Thesis Title: Transformers for Arabic Natural Language Understanding and Generation
- Major Area: Artificial Intelligence and Machine Learning systems
- Minro Area: Software, Networking and Security
- Awarded Graduate Fellowship with full tuition coverage
- Nominated for the Abdul Hadi Debs Endowment Award for Academic Excellence

Bachelor Of Engineering In Computer And Communication Engineering

Zouk Mosbeh, Lebanon

Notre Dame University - Louaize

Sept. 2013 - Jul. 2017

- GPA: 3.58 Dean's List for 6 semesters

Honors & Awards

International

2021	**First Place**, Arabic Sentiment Analysis 2021 @ KAUST. $10,000 Prize. https://wti.kaust.edu.sa/solve/Arabic-Sentiment-Analysis-Challenge	*Kaggle, Online*
2020	**Second Place**, OSACT4-Shared task on Offensive Language Detection	*Marseille, France*
2019	**Best Domain Detection system** , Qatar International Fake News Detection and Annotation Contest	*Doha, Qatar*

Domestic

2021	**Nominee**, Abdul Hadi Debs Endowment Award for Academic Excellence	*AUB, Lebanon*
2019	**Second Place**, Mindsets-OLX Datathon, Best car price prediction score.(www.kaggle.com/c/mindsets)	*AUB, Lebanon*
2019	**First place**, Big Data, AI, and Media Hackathon Powered by Touch and Anghami	*AUB, Lebanon*

Publications

Scholar Profile: `https://scholar.google.com/citations?user=fEgsyU4AAAAJ&hl=en`

My research metrics (As of November 2025):

- **Citation score: 2459**
- **h-index: 11**
- **i10-index: 12**
- **Total Publications: 22**

Selected Works (ordered by relevance):

Gaperon: A Peppered English-French Generative Language Model Suite
Godey, Nathan, **Wissam Antoun**, Rian Touchent, Rachel Bawden, Éric de la Clergerie, Benoît Sagot, and Djamé Seddah
arXiv:2510.25771, arXiv preprint — 2025 — **Cited by: -**

AraBERT: Transformer-based Model for Arabic Language Understanding
Antoun, **Wissam**, Fady Baly, and Hazem Hajj
Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools co-located with LREC 2020, Marseille, France — 2020 — **Cited by: 1603**

AraGPT2: Pre-Trained Transformer for Arabic Language Generation
Antoun, **Wissam**, Fady Baly, and Hazem Hajj
Proceedings of the Sixth Arabic Natural Language Processing Workshop co-located with EACL, Kyiv, Ukraine — 2021 — **Cited by: 183**

AraELECTRA: Pre-Training Text Discriminators for Arabic Language Understanding
Antoun, **Wissam**, Fady Baly, and Hazem Hajj
Proceedings of the Sixth Arabic Natural Language Processing Workshop co-located with EACL, Kyiv, Ukraine — 2021 — **Cited by: 198**

Towards a robust detection of language model generated text: is ChatGPT that easy to detect?
Antoun, **Wissam**, Virginie Mouilleron, Benoît Sagot, and Djamé Seddah
Proceedings of CORIA-TALN 2023, Paris, France — 2023 — **Cited by: 54**

From text to source: Results in detecting large language model-generated content
Antoun, **Wissam**, Benoît Sagot, and Djamé Seddah
Proceedings of LREC-COLING 2024, Torino, Italia — 2023 — **Cited by: 17**

ModernBERT or DeBERTaV3? Examining Architecture and Data Influence on Transformer Encoder Models Performance
Antoun, **Wissam**, Benoît Sagot, and Djamé Seddah
arXiv:2504.08716, arXiv preprint — 2025 — **Cited by: -**

CamemBERT 2.0: A Smarter French Language Model Aged to Perfection
Antoun, **Wissam**, Francis Kulumba, Rian Touchent, Éric de la Clergerie, Benoît Sagot, and Djamé Seddah
arXiv:2411.08868, arXiv preprint — 2024 — **Cited by: 10**

Data-Efficient French Language Modeling with CamemBERTa
Antoun, **Wissam**, Benoît Sagot, and Djamé Seddah
Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canda — 2023 — **Cited by: 9**

Empathetic BERT2BERT Conversational Model: Learning Arabic Language Generation with little data
Naous, Tarek, **Wissam Antoun**, Reem Mahmoud, and Hazem Hajj
Proceedings of the Sixth Arabic Natural Language Processing Workshop co-located with EACL, Kyiv, Ukraine — 2021 — **Cited by: 32**

hULMonA: The Universal Language Model in Arabic
ElJundi, Obeida, **Wissam Antoun**, Nour El Droubi, Hazem Hajj, Wassim El-Hajj, and Khaled Shaban
Proceedings of the Fourth Arabic Natural Language Processing Workshop co-located with ACL, Florence, Italy — 2019 — **Cited by: 95**

Multi-Task Learning using AraBert for Offensive Language Detection
Djandji, Marc, Fady Baly, **Wissam Antoun**, and Hazem Hajj
Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools co-located with LREC, Marseille, France — 2020 — **Cited by: 66**

State of the Art Models for Fake News Detection Tasks
Antoun, **Wissam**, Fady Baly, Rim Achour, Amir Hussein, and Hazem Hajj
2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT'2020), Doha, Qatar — 2020 — **Cited by: 80**

Skills

Python | PyTorch | Transformers | vLLM | SGLang | Docker | Slurm | TensorFlow | Tirton Inference Server | CUDA | Parallel Programming | Keras | Ansible | nginx | FastAPI | Elastic Stack | Kafka | Detectron2 | Bash | Git | Horovod | C++ | Matlab | Java | SQL | LaTex | VSCode | Copilot