Mohamed Eldagla

meldagla3@gmail.com · mohamed-eldagla.github.io · Google Scholar · GitHub · LinkedIn

Education

German International University (GIU) | *Master's in Data Science* October 2025 – July 2027 (expected)
Advisor: Caroline Sabty; thesis mentor: Alfonso Amayuelas (UCSB & Stanford HAI)

Egypt-Japan University of Science and Technology (E-JUST) | *BSc in Computer Science* October 2021 – June 2025
Artificial Intelligence and Data Science
Coursework following CMU 11-711 (NLP), Stanford CS231n (Deep Learning), Berkeley CS188 (AI), MIT 6.006 (Algorithms)

Research Experience

Fatima Institute for Global AI Research ↗ | *Pre-doctoral Fellow* April 2026 – Present
Curiosity-Driven Learning in LLM Agents (master's thesis)

- Studying curiosity-driven LLM exploration across inference, SFT, and GRPO. [P6]
- Adapting Schmidhuber's curiosity signals (surprisal, learning progress, information gain) to modern LLM agents.
- Built CuriosityLab, an environment for clean signal ablations; adding learning progress raises exploration 1.5× (pilot).

SpecDrift: Multilingual AI Safety with T. T. Mayeesha (UNT) & M. K. Morol (Anthropic)

- Auditing whether Anthropic and OpenAI models follow their own published constitutions and specs. [P5]
- Red-teaming with Petri using persona-based, multi-turn attacks across Bangla, Egyptian, and Sudanese contexts.
- Designed tiered risk taxonomies for Egyptian dialects, including region-specific risks in Upper Egypt.

PRISM AI Safety Fellowship ↗ | *Research Fellow* June 2026 – November 2026
Reward Model Interpretability; mentored by Rachel Freedman (CHAI, Berkeley)

- Reproduced Meta's LoRe ↗ with co-authors and found a data-formatting shortcut; fixing it cut accuracy (98% → 59%).
- Developed RBD to separate shared from individual preferences; per-user terms added no accuracy on real data. [P1]
- Showed RBD recovers synthetic per-user preferences and beats the base reward model (0.64 → 0.67 accuracy).

Medical Informatics & Neuro Dynamics (MIND) School, E-JUST | *Research Assistant* May 2023 – July 2026
Advisor: Ahmed Anter

- Developed an LSTM-CNN model for EEG seizure detection: 99.1% test accuracy (2-class) and 90.7% (5-class). [P3]
- Probed and intervened on medical VLM activations; sycophancy appears to reflect uncertainty, not concealment. [P2]

Université Paris-Est Créteil (UPEC) | *Data Science Intern, NLP* June 2025 – August 2025

- Built a 105K-record multilingual event dataset (real and synthetic) across English, French, and Tunisian Arabic.
- Prototyped ensemble knowledge distillation from Jais 13B and TunBERT into TinyLlama 1.1B for on-device use.

Publications

NeurIPS 2026 InterpScience Workshop **[P1] Reward Basis Decomposition for Interpreting Preference Heterogeneity ↗**
Mohamed I. Eldagla*, Ifesi Onubogu*, Prerana Rane*, Harsh Sharma, Rachel Freedman *(*equal contribution)*
🏆 **(Oral, Top 1.5%)**

NeurIPS 2026 Med-Reasoner Workshop **[P2] Uncertainty, Not Concealment: A Mechanistic Reappraisal of Sycophancy in Vision-Language Models ↗**
Mohamed I. Eldagla*, Zeyad H. ElShafei*, Ahmed M. Anter *(*equal contribution)*

Knowledge-Based Systems 2025 **[P3] Blockchain-Integrated Multi-Modal LSTM-CNN Fusion for High-Precision Epileptic Seizure Detection from EEG Signals ↗**
Mahmoud O. Mokhiamar, Abdelrahman Mahmoud, **Mohamed I. Eldagla**, Yassine Aribi, Ahmed M. Anter. *Vol. 323, 113703.*

IEEE Japan-Africa Conf. 2025 **[P4] EEG Spectrogram Augmentation Using Transformer and Quantum-Inspired Generative Models for Epilepsy Classification ↗**
Ahmed A. Elshamy, Heba A. Metwaly, **Mohamed I. Eldagla**, Zyad AlKayem, Ahmed B. Zaky

Pre-print 2026 **[P5] Specification Drift in Multilingual, Multi-Turn LLM Interactions** *For ML conference submission*
Mohamed I. Eldagla*, Nusrat Jahan Lia*, Azza Osman, Tasmiah Tahsin Mayeesha, Md Kishor Morol *(*equal contribution)*

Pre-print 2026 **[P6] The Curious Agent: Can Intrinsic Curiosity Guide LLM Exploration?** *For ML conference submission*
Mohamed I. Eldagla, Ahmed Hammad, Alfonso Amayuelas

Honors & Awards

2026 **Fatima Fellowship** Fatima Institute for Global AI Research; 1 of 67 fellows from 700+ applicants; includes compute support
2026 **PRISM AI Safety Fellowship** 1 of 50 fellows from 800+ applicants
2026 **Black in AI Emerging Leaders in AI (ELAI)** Top-ranked in 2026–27 cohort; mentored by **Alexander Rush**
2026 **MIT AI Alignment, AI Safety Fundamentals Fellow** Reading group on alignment and interpretability
2025 **UPEC AI Research Internship** 1 of 30 from 1,300+ applicants for a master's-level internship, as an undergraduate
2025 **GIU MSc Data Science Fellowship** Full tuition for academic excellence
2025 **E-JUST Academic Excellence** Ranked 1st of 120 in the final semester (GPA 4.00); graduated in the top 10% of the class
2024 **SAS Curiosity Cup** Global data science competition; top 15 of 107 participants
2024 **Data Science Intern, DataScience Middle East** 2nd place across the two-month hybrid cohort; SAS partner
2023 **NTA Presidential Leadership Program** AI, CS, and leadership
2022 **E-JUST Misr El Kheir Merit Scholarship** 75% undergraduate tuition for academic excellence

Talks

Dec. 2026 **Reward Basis Decomposition for Interpreting Preference Heterogeneity↗** Oral, *NeurIPS 2026 Workshop*
Oct. 2026 **Do People Really Want Different Things? Reward Basis Decomposition Explained↗** PRISM Fellowship
Aug. 2026 **Babel's Backdoor: When Multilingual Diversity Becomes an Agent's Vulnerability** GIU
Jan. 2026 **Limitations of MediaPipe Pose in Real Videos and Mitigation Strategies: A Critical Analysis** GIU
May 2025 **AI for a Sustainable Future** Ministry of Planning and Economic Development, Egypt
Feb. 2025 **Prerequisites Are All You Need: Learning Machine Learning in the Era of AI** E-JUST

Professional Experience

Egyptian Cabinet Information and Decision Support Center | *AI Engineer Intern* July 2025 – September 2025

- As sole developer, built "Chat Egypt," a Prime Minister-sponsored RAG chatbot for checking circulating rumors.
- Built pipelines that verify news content against trusted government and media sources; system later deployed.

Teaching & Volunteering

German International University (GIU) | *Teaching Assistant, Faculty of Informatics* October 2025 – Present

- Teaching Data Engineering & Visualization, Big Data & NoSQL, Programming I, and Databases (50–154 students).
- Covering data pipelines, MongoDB, Cassandra, Spark MLlib, Kafka, Hadoop, and SQL through hands-on labs.

Code in Place, Stanford University | *Volunteer Section Leader Mentor* April 2026 – July 2026

- Mentored students in a CS106A-based introductory programming course with 16,000+ learners.
- Held one-on-one help sessions with students several times a week.

Digital Egypt Cubs Initiative (DECI), MCIT | *Instructor, via Udacity* July 2023 – December 2024

- Taught Python, data structures, and algorithms to 40 students aged 12–17.
- Ran weekly four-hour, activity-based lessons.

Google Developer Student Clubs, E-JUST | *Core Team Member & Instructor* October 2021 – May 2023

- Taught C++, data structures, and algorithms to 50+ undergraduates.
- Led study sessions following Harvard's CS50x online curriculum for 100+ students.

UN Climate Change Conference (COP27), UNFCCC | *ICT Volunteer* October 2022 – November 2022

- Co-developed an indoor navigation system for the COP27 mobile app with MCIT and UN engineers.
- Managed the UN Climate Change YouTube channel from the conference's Master Control Room.

Academic Service

2026 – Present **Lead, AI Safety in Medicine Group** MIND School, E-JUST
2026 **Reviewer** *International Conference on Advancements in Smart, Secure and Intelligent Computing* (ASSIC 2026)

Skills & Interests

Programming Python, C/C++, JavaScript, SQL
ML & LLMs PyTorch, Hugging Face, PEFT/LoRA, vLLM, TensorFlow, OpenCV
Data & Tools Spark, Kafka, MongoDB, Pandas, Git, Docker, LaTeX
Interests Soccer, long-distance running, cycling (top 5 of 80 in a 25 km race), memorizing phone numbers