Kenza Amara

0041772655438 | 22amara.kenza@gmail.com | Zurich, Switzerland

RESEARCH INTEREST

Alignment, AI Interpretability, Multimodal AI, working with Large Language Models, Large Vision-Language Models, and Graph Neural Networks.

EXPERIENCE

Stealth AI & Robotics Startup Sep-Nov 2025
Research Engineer *Zurich, Switzerland*

- Designed and implemented the delta wing for drone structures and integrated it into the main development pipeline.
- Collected and processed data on aircraft components using web scraping, 2D segmentation, and 3D modeling.

IBM Mar-Oct 2024
Multimodal XAI - Research Collaboration *Zurich, Switzerland*

- Analyze the impact of the text and image modalities on the LVLM answer and rationale (LLaVa-Vicuna-7B,...)
- Produce a synthetic text-image dataset for VQA and develop a user interface for testing LVLMs.
- Identify the effect of input perturbations (image annotation, complementary/contradictory text descriptions,...)

Microsoft Jun-Aug 2022
AI Research PhD Internship *Cambridge, UK*

- Optimized GNN regression objective with a substructure-aware loss to account for common core structures in molecule pairs.
- Improved feature attribution methods on a recently proposed explainability benchmark.

Meta May-Jul 2021
AI Research Internship *Paris, France*

- Reinterpreted binary-hashing and product quantizers as auto-encoders
- Designed backward-compatible decoders that improve the reconstruction of the vectors from the same codes, significantly outperforming in nearest-neighbor search.

Daikin Jun-Sept 2018
ML Engineer Internship *Osaka, Japan*

- Developed an optimized ML model regularized by thermodynamic laws
- Improved the predicted power consumption of air-conditioning systems.

EDUCATION

ETH AI Center, Zurich *Ph.D. in Computer Science*, 2021-2025
PhD Advisors: Mennatallah El-Assady, Andreas Krause, Ce Zhang.

ETH University, Zurich *MSc in Environmental Systems and Policy* 2019-2021

Ecole Polytechnique, Paris *MSc in Computer Science* 2016-2019

Lycée Henri 4, Paris *Scientific Preparatory Program* 2014-2016

SKILLS

Code: Python, R, C/C++, Java, TensorFlow, PyTorch, Scikit-Learn, Keras

Communication: LateX, React/D3.js, HTML/CSS/PHP, Linux, Microsoft Windows

Languages:

- *Fluent* - English, French, German
- *Intermediate* - Spanish, Japanese

LEADERSHIP

Talks

XAI Conference 2024 - *Oral*, “Challenges & Opportunities in Text Generation Explainability”

Tech for Sustainability 2023 (Microsoft & EY) - *Talk*, “Cyclone Tracking durch Machine Learning und Remote Sensing zur Identifizierung von Biodiversität”

PML4DC ICLR 2022 - *Panel Discussion*, Practical ML for Developing Countries

AMLD EPFL 2022 - *Lightning talk*, Advances of ML Approaches for Financial Decision-Making

Business - Entrepreneurship

University of St. Gallen - *Corporate Finance* 2023

EuroMUN Maastricht 2018

Teaching

ETH University - *Teaching Assistant*, WebDev 2024

ETH University - *Teaching Assistant*, Interactive Machine Learning 2024

Sino-French Nuclear Engineering Institute, China - *Teaching Assistant*, Mathematics and Physics 2016-2017

Project Supervision

Which language do LLMs think in?, Arundhati Balasubramaniam, MSc

Investigating Hierarchical Feature Types in Multimodal Models and the Impact on Polysemanticity, Sinie van der Ben, PhD

Exploring GNN-LLM’s Ability to Model Complex Relationships for Recommendation Systems, Sebastian Hönig, MSc

Path-based Explainability of Open-ended LLM Responses using Knowledge Graphs, Lukas Mautner, MSc

Beyond Binary Selection: Exploring Soft Masking Techniques for Language Model Explainability, Steffen Backmann, MSc

Other

Hack-Nation - Global AI Hackathon & Acceleration Program 8-9 Nov 2025

PUBLICATIONS

Preprint: https://www.arxiv.org/abs/2505.07610

Concept-Level Explainability for Auditing & Steering LLM Responses

Kenza Amara, Rita Sevastjanova, Mennatallah El-Assady

Preprint: https://arxiv.org/abs/2410.01690

Why Context Matters in VQA & Reasoning: Semantic Interventions for VLM Input Modalities

Kenza Amara, Lukas Klein, Carsten T. Lüth, Paul F Jaeger, Hendrik Strobelt, Mennatallah El-Assady

ICLR 2025 - **Bidirectional AI-Human Alignment** Workshop

Processing, Priming, Probing: Human Interventions for Explainability Alignment

Kenza Amara

Neurips 2024 - Workshops: **ATTRIB, Interpretable AI, SafeGenAI, RedTeaming, CALM, Statistical Frontiers LLMs**

Interactive Semantic Interventions for VLMs: A Human-in-the-Loop Approach to Interpretability

Kenza Amara, Lukas Klein, Carsten T. Lüth, Hendrik Strobelt, Mennatallah El-Assady, Paul F Jaeger

Neurips 2024 - Datasets and Benchmarks

PowerGraph: A power grid benchmark dataset for graph neural networks

Kenza Amara, Anna Varbella, Blazhe Gjorgiev, Giovanni Sansavini

ACL 2024
SyntaxShap: A Syntax-aware Explainability Method for Text Generation
Kenza Amara, Rita Sevastjanova, Mennatallah El-Assady

XAI 2024
Challenges and Opportunities in Text Generation Explainability
Kenza Amara, Rita Sevastjanova, Mennatallah El-Assady

VIS 2024 - NLVIS Workshop
iToT: An Interactive System for Customized Tree-of-Thought Generation
Alan Boyle, Isha Gupta, Sebastian Hönig, Luks Mautner, Kenza Amara, Furui Cheng, Mennatallah El-Assady

Neurips 2023 - XAI Workshop
GInX-Eval: Towards In-Distribution Evaluation of Graph Neural Network Explanations
Kenza Amara, Rex Ying, Mennatallah El-Assady

IEEE Computer Society
Generative Explanation for Graph Neural Network: Methods and Evaluation
Jialin Chen, Kenza Amara, Junchi Yu, Rex Ying

Journal of Cheminformatics
Explaining compound activity predictions with a substructure-aware loss for graph neural networks
Kenza Amara, Jose Jimenez Luna, Raquel Rodriguez Perez

LoG 2022
GraphFramEx: Towards Systematic Evaluation of Explainability Methods for Graph Neural Networks
Kenza Amara, Rex Ying, Zitao Zhang, Zhihao Han, Yinan Shan, Ulrik Brandes, Sebastian Schemm, Ce Zhang

ICMR 2022
Nearest neighbor search with compact codes: A decoder perspective
Kenza Amara, Matthijs Douze, Alexandre Sablayrolles, Hervé Jégou

AAAI 2022
ReforesTree: A Dataset for Estimating Tropical Forest Carbon Stock with Deep Learning and Aerial Imagery
Gyri Reiersen, David Dao, Bjorn Lütjens, Konstantin Klemmer, Kenza Amara, Attila Steinegger, Ce Zhang, Xiaoxiang Zhu

KDD 2020 - Fragile Earth Workshop
OneForest: Towards a Global Species Dataset by Fusing Remote Sensing and Citizen Science Data with Graph Neural Networks
Kenza Amara, David Dao, Charlotte Bunne, Bjorn Lütjens, Dava Newman, Ce Zhang, and Tom Crowther