

**RESEARCH AREA** Theory of Mind enabled LLMs for improved AI social reasoning. Applications in social simulation and inference.

**TECHNICAL SKILLS** **NLP:** LLM IFT, Representation learning (contrastive and triplet loss), Semantic Clustering  
**DL/RL:** Transformers, MoE, GRPO, PPO, RNNs, LoRA **Frameworks:** Pytorch, Deepspeed, hydra, Python

**MANAGEMENT SKILLS** Establishing ML teams and functions. Cross organizational planning and execution of AI strategies. Building and running partnerships with academic and private institutions.

**HONORS** Fulbright Scholarship 2010

### EDUCATION & Training

**University of Oxford** - 2023 - present (part-time)  
Ph.D. in Information, Communication and the Social Sciences

**Stanford University** - 2017  
Graduate coursework in NLP with Deep Learning (CS224n)

**University of Maryland College Park** - 2012  
Masters in Information Management

**Damascus University** - 2007  
B.E. Computer Science, Focus: Artificial Intelligence

### SELECTED PUBLICATIONS

(In review) LINGOLY-TOO: Disentangling Memorisation from Reasoning with Linguistic Templatisation and Orthographic Obfuscation (2025)

TopoX: A Suite of Python Packages for Machine Learning on Topological Domains. JMLR (2024)

Stance Prediction and Claim Verification: An Arabic Perspective. Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER) workshop at ACL (2020)

Mr. LDA: A Flexible Large Scale Topic Modeling Package using Variational Inference in MapReduce. ACM International Conference on WWW (2012)

### HIGHLIGHTED EXPERIENCE

**PRINCIPAL APPLIED SCIENTIST, FORETHOUGHT.AI** Dec 22 - Oct 23  
Lead work in conversational LLM instruction-finetuning and evaluation. Finetuned up to 30B LLMs outperforming specialized internal models in 12 tasks using production, synthetic and human annotations.

**SR. STAFF APPLIED SCIENTIST, FORETHOUGHT.AI** Feb 21 - Nov 22  
Lead a new AI product for themes categorization of millions of support tickets that outperformed SotA public models in clustering coverage by 46% relative and 8% in online metrics for beta customers. Custom models included a T5-based abstractive summarizer and a custom encoder model trained with triplet loss from synthetic and human labels.

Grew and managed the ML team and established ML infra functions. Setup hiring and interviewing processes and advised the leadership team on ML initiatives, OKRs and objectives.

Designed and implemented company's first ML experimentation repo and infrastructure using pytorch, hydra, mlflow and sagemaker.

**SR. PRINCIPAL ML SCIENTIST, SAGE INTACCT** Apr 20 – Jan 21  
Lead the company's first AI product by revamping all unsupervised models for time tracking which improved relative F1-score 60-80% over previous experiments. Helped build the ML engineering team and hire ML and Data engineers.

**SENIOR APPLIED SCIENTIST, MICROSOFT** May 18 – Sep 19  
Trained Language Models for speech recognition with distributed training of LSTMs on tens of billions of words. Implemented several sub-word representations (BPE, Wordpiece) to improve training efficiency.

**SR. /PRINCIPAL DATA SCIENTIST, SALESFORCE** Sep 15 – Feb 18  
Drove the team's NLP DL foundation and automated business processes by building text classification models for the customer support processes.  
Developed reusable machine learning pipeline and feature engineering libraries.

**DATA SCIENCE LEAD, CRITTERCISM** Jul 14 – Nov 14  
Applied clustering techniques and text matching for cardinality reduction of mobile device types.  
Built an internet facing analytics portal for tracking live mobile performance metrics worldwide (Patented).  
Built AWS on demand analytics infrastructure using S3 and EMR.

**DATA ANALYTICS LEAD, IREX** Nov 13 – Jun 14  
Lead the organizational technical projects, processes and team for documenting human rights violations from crowdsourced content. Oversaw the design and implementation of video and image based annotation systems.

**BIG DATA SPECIALIST, ORACLE** Mar 13 – Oct 13  
Prototyped machine learning proof of concepts in the public sector using technologies including Hadoop, Hive, Pig, R Enterprise, Mahout and other proprietary and open source tools.

**DATA SCIENTIST, ORACLE** Jun 12 – Aug 12  
Lead Data Scientist in the Oracle/NCI partnership project that won the "2012 Best Government Big Data Solution" Award.  
Developed MapReduce programs in Java and Python for generating synthetic Gene data of 900 million patients.

**GRADUATE RESEARCH ASSISTANT, UMIACS** Oct 11 – May 12  
Evaluated distributed Topic Modeling (LDA) algorithms and applied them for unsupervised lexicon expansion.  
Developed an Arabic version of the Word Count tool (LIWC) for sentiment analysis and honor dictionary validation.