JUDE KHOUJA +1 240 200 4022 jude@latynt.com

RESEARCH

AI social cognition modeling and evaluation. Computational models for social simulation and inference.

SKILLS

NLP: LLM IFT, DPO, LoRA, Representation learning (contrastive and triplet loss), Semantic Clustering, Summarization
DL/ML: Transformers, MoE, EncDec, RNNs, DPO, LoRA **Tools:** Python, Pytorch, Deepspeed, hydra
Management: Starting, recruiting and managing ML teams and functions. Planning and executing high level AI strategies and initiatives. Establishing academic and data partnerships.

HONORS

Fulbright Scholarship, U.S. Department of State 2010.

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)

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

HIGHLIGHTED EXPERIENCE

PRINCIPAL APPLIED SCIENTIST, FORETHOUGHT.AI Dec 22 - Oct 23
Lead work in Large Language Modeling instruction-finetuning and evaluation. Trained a FLAN2-style LLM outperforming specialized internal models in all 12 tasks using production, synthetic and human annotations.

SR. STAFF APPLIED SCIENTIST, FORETHOUGHT.AI Feb 21 - Nov 22
Designed, built and deployed a new AI product (discover) for semantic clustering of millions of support tickets which included training new T5-based abstractive summarizer and a custom encoder model trained with triplet loss from synthetic and human labels. New models improved offline clustering coverage by 46% relative and 8% in online metrics for beta customers while reducing duplicates.

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

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

SENIOR APPLIED SCIENTIST, MICROSOFT May 18 – Sep 19
Contributed to scaling up Language Model distributed training to tens of billions of words and explored the use of sub-word representations (BPE, Wordpiece) in speech services's language modeling team.

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.