

Parsa Mazaheri

San Francisco Bay Area | 650-390-5422 | p.mazaheri98@gmail.com | [Linkedin: parsa-mazaheri](#) | [GitHub](#) | [Google-Scholar](#)

Machine Learning Engineer with 5+ years of experience in machine learning, GenAI, software engineering, and large-scale systems. Focused on foundation models, Reasoning, Multimodal Learning, Synthetic data, and Agentic AI.

EXPERIENCE

Machine Learning Engineer | *Apple Inc* - (*GenAI*)

Sep 2024 – Present

- Part of **Foundation Models** team in DMLI doing research & engineering on frontier on-device and server foundation models, with focus on **reasoning, agentic, and STEM/scientific reasoning** (text & multimodal).
- Pre-training:** Expanded expert-level STEM/math pre-training mixture at web scale and pioneered hard-problem curricula via targeted generation; boosted MATH/Science benchmarks by **12–23%** (eval-gated).
- Synthetic data at scale:** Led end-to-end synthetic QA & derivation-style generation at **70B+ tokens** (text & multimodal), spanning target selection → generation → QC; improved scientific reasoning benchmarks by **8–19%**.
- Multimodal reasoning:** Improved visual understanding and multimodal problem-solving by generating large-scale **500M+ image-caption pairs** and **+10M image-QA** for image-grounded reasoning tasks.
- Reasoning & CoT:** Engineered reasoning improvements via structured multi-step training data, CoT-oriented instruction design, and synthesis across math, science, and general reasoning tasks. (post-training)

Machine Learning Engineer | *University of California Org* - (*GenAI*)

Jul 2023 – Sep 2024

- Designed and deployed UC virtual assistant (Chatbot) powered by **Llama3 & GPT4** LLMs with RAG
- Automation & RAG:** Developed an **Action mechanism** with **intent classifier** routing to application-specific actions a **Vector-DB** using **FAISS** for precise querying and knowledge retrieval.
- Designed **Purchase Analysis system** for UCOP's with **\$1.118 Billion** annual transactions, using **predictive analytics**, and compliance checks for funds and transaction behaviors, reducing manual reviews.
- Owned end-to-end system design (data ingestion → retrieval → orchestration → deployment), improving response consistency and reliability for production usage improving the experience for procurement and buyers.

Tech Lead | *Tapsell Inc.* - (*Adtech/E-Commerce*)

Jun 2021 - Aug 2022

- Led** a team of 5 to build an end-to-end sponsored-content (reportage) sales platform (payments + editor), automating fulfillment and increasing reportage revenue by **80%** at one of Iran's largest adtech companies.
- Built a **backlink marketplace** integrated with the sponsored-content platform, providing campaign analytics and optimization recommendations that improved advertiser performance by **30%**.
- Improved the **Ads API platform** (delivery and targeting), driving **+12% CTR** across **45M** unique users and **28B** monthly impressions on the web.
- Streamlined user-insights collection with **sentiment analysis** (BERT, ClickHouse), improving audience targeting and increasing banner-ad revenue by **24%**.

Software Engineer | *Asre Danesh Inc* - (*Fintech*)

Jul 2019 - Jan 2021

- ML:** Designed a Credit Card scanner OCR (C++ and PyTorch) – worked with attention, transformers, and model optimization, cutting data entry time by 2x
- Data Science:** Architected advanced **data processing pipeline** with Numpy, Spark to efficiently handle and analyze large volumes of trading data, enabling faster **trade decisions**.
- Developed an **automated DSaaS** to provide forecast insights for automating a manual multi-component process

PUBLICATIONS

- REPOT: Recoverable Program-of-Thought via Checkpoint Repair** [arXiv]. Parsa Mazaheri. arXiv 2026.
- AgentAtlas: Beyond Outcome Leaderboards for LLM Agents** [arXiv]. Parsa Mazaheri, Kasra Mazaheri. arXiv 2026.
- Apple Intelligence Foundation Language Models: Tech Report 2025** [arXiv]. Apple Foundation Models Team. arXiv preprint, 2025.
- Enhancing Reliability in Large Language Models through Automated Hallucination Detection** [IJCHML]. Parsa Mazaheri, Selin Ugur, Mariam Gonzaliam. International Journal of Computer and Human-Machine Learning, 2026.
- Bridging Accuracy and Interpretability in Large Language Models: A Hybrid AI Approach** [IJCHML]. Zhang Hao, Parsa Mazaheri, Ece Arslan. International Journal of Computer and Human-Machine Learning, 2026.
- Don't Believe Everything You Read: Enhancing Summarization Interpretability through Automatic Identification of Hallucinations in Large Language Models** [arXiv]. Priyesh Vakharia, Devavrat Joshi, Meenal Chavan, Dhananjay Sonawane, Bhriugu Garg, Parsa Mazaheri. arXiv preprint, 2023.

PROJECTS

- **Patent Law Copilot:** Designed a Copilot using GenAI (**GPT3.5**), for writing patents claims with references & a Patent finder (finding relevant patents **by giving a description**) and giving differences for the two.
- **Semantic Parsing:** Worked with Prof. Ziyu Yao on a **Semantic Parsing** project on cross-domain database with **Google’s T5 and Bert**. Pioneered a confidence metric, which increased agents’ accuracy in answers by **8%** on uncertain domains.
- **IOB Slot Tagging:** Developed a Deep learning model using **GloVe and LSTM** (PyTorch) for natural language utterances tagging. Implemented an **Bi-LSTM** model with **F-1 score of 0.87** on test data, accurately classifying tokens in text.
- **Aspect-Based Sentiment Analysis:** Pioneered a novel **Bert**-Based ABSA model for identify the aspects of given target entities on Semeval 2014 dataset achieving **0.82** F-1 score.

ACCOMPLISHMENTS

- Ranked within the **Top 0.5%** among **165,000+** participants in the nationwide university entrance exam (Aug 2019)
- Admitted to the **National Organization for Development of Exceptional Talents (NODET)** through a highly competitive nationwide exam (Aug 2013)
- **International Collegiate Programming Contest (ICPC)** 2017 Finalist – 2nd Place Regional

EDUCATION

- University of California Santa Cruz** | *MS Artificial Intelligence - NLP* **Sep. 2022 – Dec 2023**
- M.S. Thesis: Paper on Interpreting Hallucination in LLMs ([arxiv](#)) - Advisor: Prof. Ian Lane
 - Coursework: Deep Learning, Conversational Agents, Machine Translation, Generative Transformers
- K.N. Toosi University of Technology** | *BS Computer Engineering* **Sep 2017 – Jun 2021**
- B.S. Thesis: Compact Pedestrian Detection using YOLO
 - **International Collegiate Programming Contest (ICPC)** 2017 Finalist – 2nd place Regional

TECHNICAL SKILLS

Languages: Python, Java, C++, GoLang, SQL
Libraries: PyTorch, TensorFlow, Jax, HuggingFace, Scikit-Learn, Transformers, Pandas, Numpy, Spark, OpenCV, Gensim, spaCy, Matplotlib, NLTK, Airflow, vllm, Hive, XGBoost, Onnx, LLM, CUDA, TensorRT, Triton, VertexAI, LLaMa
Tools: AWS, GCP, Kafka, RestAPI, PostgreSQL, MongoDB, Docker, Kubernetes, Git, Prometheus, CUDA, MLOps, Natural Language Processing, Computer Vision, Auto-regressive Models, Parallel Programming, FastAPI, Django, React, Splunk, Grafana