Mahrokh Hassani

Netherlands

☎ +31 612837312 • ✉ mahrokhhassani99@gmail.com

mahrokh-hassani-1a0002353

Profile

AI Developer – LLM & NLP Systems with experience designing and evaluating real-world language systems for content moderation, safety, and multilingual NLP.

Strong background in building end-to-end LLM pipelines, including data processing, model selection, fine-tuning, prompt-based generation, retrieval-augmented generation (RAG), and evaluation under practical constraints. Experienced with transformer-based models and attribution techniques to improve reliability, explainability, and robustness of LLM outputs.

Comfortable working close to product and research teams, iterating on system behavior using offline metrics and human-in-the-loop feedback, and translating research insights into deployable NLP solutions.

Technical Skills

Languages: Python, C++, HTML, SQL
ML Frameworks: PyTorch, Hugging Face Transformers, scikit-learn, XGBoost, TensorFlow (working knowledge)
LLMs & NLP Transformer-based models, LLM fine-tuning, prompt engineering, Retrieval-Augmented Generation (RAG), semantic similarity, Multilingual NLP, content moderation, hallucination detection, Token- and step-level attribution, LLM evaluation & benchmarking
Systems & Tooling: FastAPI, Docker, Git, Linux, Jupyter Experiment tracking & evaluation pipelines, Microsoft Azure (LLM services, experimentation, deployment basics)
Tools: Docker, Git, Linux, Jupyter, Google Colab, FastAPI, VS Code, LaTeX, Azure
Visualization & Analysis: Python visualization (matplotlib, seaborn), Power BI (basic)

Education

University of Groningen — Groningen, Netherlands
MSc Information Science — Sep 2024 – July 2025
Specialization: Artificial Intelligence, NLP, Data Science.
Thesis (Grade: 8.2): *Hate Identification and Detoxification in Social Media* — designed a dual-stage pipeline combining a ModernBERT-based toxicity classifier with an mT0-XL detoxifier; evaluated on PAN-2024/2025 and Jigsaw fine-grained toxicity splits using STA, SIM, and ChrF1; reproducible PyTorch/HF codebase with ablations on prompts and decoding.
Selected Coursework: Advanced Topics in NLP; Computational Semantics; Language Technology Project; Shared Task in Information Science; Research Seminar in Information Science; Learning from Data; Semantic Web Technology; User Interface Evaluation.

Malayer University — Iran
BSc Computer Engineering — 2018 – 2022
Thesis: Crossroad Service Level Detection using YOLOv5 (C++, OpenCV).
Selected Coursework: Algorithms and Data Structures; Operating Systems; Database Systems; Computer Networks; Software Engineering; Artificial Intelligence; Probability and Statistics; Digital Logic Design.

Experience

AI Developer Intern — BrainBite — Nov 2025 – Present
Remote, Netherlands

- Designed and implemented an agent localization evaluation tool to assess multilingual and region-specific behavior of LLM-based tutoring agents, supporting quality control and consistency across localized content.

- Built LLM-based pipelines for personalized educational content, with emphasis on controllability, safety, and response quality across different user contexts.
- Implemented and evaluated prompt-based generation, retrieval-augmented generation (RAG), and structured reasoning loops to improve factuality and alignment.
- Developed evaluation workflows to compare agent behavior across languages, prompts, retrieval strategies, and decoding settings, enabling data-driven iteration and model selection.
- Collaborated with AI engineers to integrate evaluation and generation pipelines into product-facing prototypes, balancing quality, latency, and reliability constraints.

University of Groningen Jan 2025 – July 2025
AI/NLP Research Assistant

- Designed, fine-tuned, and evaluated transformer-based NLP systems for hate speech, irony, and hallucination detection, focusing on robustness across domains and languages.
- Built multilingual LLM-based detoxification pipelines using mT0-XL and prompt-based generation, emphasizing controllability, safety, and output consistency.
- Developed LLM attribution and explainability tools (ContextCite, Inseq) to analyze token- and step-level behavior, supporting debugging and trust assessment of model outputs.
- Implemented reusable evaluation pipelines combining automatic metrics (ChrF, BERTScore, uncertainty estimation) with structured error analysis to guide model iteration.
- Contributed to a SemEval-2025 shared task on hallucination detection, achieving results within the top-performing tier and validating system performance in a competitive benchmark setting.

MAPNA Group Tehran, Iran
IT Intern May 2021 – Jul 2021

- Supported cybersecurity and system maintenance, reducing downtime incidents and contributing to ISMS research that informed company policy updates.
- Assisted in network monitoring and troubleshooting, helping to decrease recurring system errors and improving operational reliability.
- Documented IT procedures and security protocols, which streamlined onboarding for new staff and increased compliance with internal audits.
- Collaborated with senior engineers to analyze system vulnerabilities, leading to actionable recommendations that strengthened infrastructure security.

Projects

Hate Identification & Detoxification System: Python, PyTorch, Hugging Face. Built a dual-stage NLP system combining a transformer-based toxicity classifier with LLM-based text rewriting to reduce harmful content while preserving semantic meaning. Designed prompt-controlled detoxification workflows and evaluated outputs using automatic metrics (STA, SIM, ChrF) and qualitative error analysis. Structured the pipeline to support multilingual inputs and reproducible experimentation, enabling controlled comparison of model and decoding choices.

Hallucination Detection & Evaluation System (SemEval-2025): Python, LLaMA-3, Wikipedia API, RAG. Developed an LLM evaluation system to detect hallucinations using retrieval-augmented generation and uncertainty-aware scoring. Implemented retrieval pipelines and decision logic to distinguish grounded from hallucinated outputs across knowledge-intensive prompts. Validated system performance in the SemEval-2025 shared task, ranking within the top-performing tier among participating teams.

Fanfiction Popularity Prediction: Python, XGBoost, LLMs. Created ML pipeline to predict story popularity using metadata and semantic features.

Multilingual LLM Attribution & Explainability Tool: Python, ContextCite, Inseq. Built attribution pipelines to analyze token- and step-level reasoning behavior in multilingual LLM outputs. Used attribution signals to diagnose failure modes, spurious reasoning patterns, and cross-lingual inconsistencies. Framed explainability outputs as debugging aids for improving LLM reliability rather than post-hoc analysis artifacts.

F1 Race Strategy Simulation & Decision Support System: Python. Designed a time-series simulation framework to evaluate race strategy decisions such as undercut/overcut timing under tire degradation and pit-stop constraints. Modeled lap-time degradation and scenario-based outcomes to compare strategic trade-offs between track position and tire performance. Built visualizations to surface optimal decision windows, framing the system as a decision-support tool rather than a prediction model.

Languages

- **English** – Fluent (C1/C2)
- **Dutch** – Learning
- **Persian (Farsi)** – Native
- **Korean** – Intermediate (B2 conversational)
- **Turkish** – Upper Intermediate (B2)
- **French** – Upper Intermediate (B2)

Interests

Music production, reading, sports, creative writing.