# Ali Aghababaei

**NLP & Generative AI Engineer**

*Orientation Year Visa | authorized to work in the Netherlands without sponsorship*


| | | | |
|---|---|---|---|
| Email: | ali.aghababaei96@gmail.com | LinkedIn: | linkedin.com/in/aliaghababaeii |
| Phone: | +31 6 83 35 97 62 | Portfolio: | aghababaeiali.github.io |
| Location: | Rotterdam, Netherlands | GitHub: | github.com/aghababaeiali |


## Professional Summary


Applied AI Engineer specializing in LLMs, NLP, and production-ready AI systems. Hands-on experience training, fine-tuning, and evaluating transformer models (BERT, RoBERTa, FLAN-T5) and building end-to-end RAG pipelines with vector-based retrieval, MLflow experiment tracking, CI/CD automation, and Docker-based deployment. Built an 8-node agentic LLM system with LangGraph, evaluated with RAGAS (Faithfulness: 0.921). First co-author on a 2025 arXiv paper on explainability in NLP, combining applied research with practical model validation. MSc ICT, University of Padova, 110/110. Based in Rotterdam, authorized to work in the Netherlands without sponsorship.


## Technical Skills

**Languages:** Python, SQL

**ML & AI:** PyTorch, Scikit-learn, Hugging Face Transformers, NumPy, Pandas, Matplotlib

**NLP & GenAI:** BERT, RoBERTa, T5/FLAN-T5, LangChain, LangGraph, CrewAI, Agentic AI, Retrieval-Augmented Generation (RAG), Vector Search, sentence-transformers, SpaCy, LoRA/PEFT, HuggingFace PEFT, Prompt Engineering, Groq API, OpenAI API, Large Language Models (LLMs)

**Evaluation:** RAGAS, ROUGE, F1, BERTScore, A/B Testing, Explainable AI (XAI), Integrated Gradients, SHAP, Captum, LLM Evaluation

**Deployment & Tools:** FastAPI, Docker, Gradio, Hugging Face Hub/Spaces, ChromaDB, Vector Databases, Git, pytest, Render, CI/CD, GitHub Actions, MLOps, MLflow, Experiment Tracking, REST APIs, GCP, AWS

## Projects

**Agentic RAG System (LangGraph, ChromaDB, RAGAS)** Live | GitHub
*Apr 2026*

- Built an 8-node agentic AI system with LangGraph: arXiv search with query reformulation loop, session-isolated ChromaDB retrieval with per-paper re-ranking, contradiction detection, and grounded answer synthesis with citations
- Evaluated with RAGAS across 20 queries: **Faithfulness 0.921**, Answer Relevancy 0.885; deployed live on Hugging Face Spaces

**Dutch Constitution Q&A: RAG Pipeline with Evaluation** Live | GitHub
*Apr 2026*

- Built end-to-end RAG pipeline (PDF ingestion, MiniLM embeddings, ChromaDB, Llama 3.1 via Groq) served via FastAPI and Gradio; **GitHub Actions CI/CD pipeline** runs 13 pytest tests on every push and auto-deploys to Hugging Face Spaces on success
- Tracked all evaluations in **MLflow**; identified retrieval failures through tracking, replaced character-based chunking with structure-aware Article-level splitting, improving **Faithfulness from 0.55 to 0.86 (+0.31)** via documented A/B comparison

**Explainable NLP with Integrated Gradients (BERT, RoBERTa, Captum)** Paper | GitHub
*Jan 2025*

- Deliberately fine-tuned BERT and RoBERTa with controlled overfitting to sharpen class-distinctive token attribution; applied Integrated Gradients via Captum for token-level explainability; benchmarked against DeepLIFT and GradientSHAP; built custom SpaCy rendering pipeline for subword tokenization and negation alignment
- Validated findings with social psychologists; contributed to a published arXiv preprint (2025) as first co-author

**NLP Sentiment Analysis Pipeline (DistilBERT, FastAPI, Docker)** Live | GitHub
*Nov 2025*

- Fine-tuned DistilBERT on 50K IMDB reviews achieving **88% accuracy and F1**; deployed via FastAPI

and Docker on Render with auto-redeploy; model published to Hugging Face Hub

**Text Summarization Pipeline (FLAN-T5, LoRA, FastAPI)** Live | GitHub
*Dec 2025*

- Fine-tuned FLAN-T5 with LoRA on Samsum; **ROUGE-1: 45.09** vs baseline 43.03; deployed via FastAPI and Gradio on Hugging Face Spaces

## Experience

**Research Assistant** University of Padova, Italy
*Jun 2024 – Dec 2024 · Part-time · Hybrid*

- Fine-tuned BERT and RoBERTa on sociopsychological datasets to extract class-distinctive keywords for dictionary expansion and recursive dataset construction; applied Integrated Gradients (Captum) for token-level attribution, benchmarked against DeepLIFT and GradientSHAP
- Built a custom SpaCy rendering pipeline for subword tokenization and negation alignment; findings validated through expert human evaluation by social psychologists
- Contributed to a published arXiv preprint (2025) as first co-author alongside researchers from University of Padova and SWPS University Warsaw

**Technical Expert** Mazand Phase Industrial Power Company, Iran
*Nov 2019 – Mar 2022 · Full-time*

- Designed electrical panels and test methodologies, producing mechanical and electrical drafts in AutoCAD and ePLAN
- Supervised project execution, commissioning, and repair of power systems and electrical inverters
- Delivered product training to clients, translating technical specifications into practical operating procedures

**Electrical Engineering Intern** Sorak Taban Electricity Distribution Company, Iran
*Jun 2018 – Aug 2018*

Assisted in infrastructure planning and technical documentation during undergraduate studies.

## Education

**M.Sc. in ICT for Internet and Multimedia** University of Padova, Italy
*Sep 2022 – Jul 2025*
Grade: 110/110 Thesis: *Application of Integrated Gradients Explainability to Sociopsychological Semantic Markers*

**Erasmus Exchange Program** Warsaw University of Technology, Poland
*Oct 2023 – Feb 2024*
Grade: 4.75/5.0

**B.Sc. in Electrical Engineering** Ferdowsi University of Mashhad, Iran
*Sep 2014 – Sep 2019*

## Publications

**Application of Integrated Gradients Explainability to Sociopsychological Semantic Markers**
arXiv preprint (2025) · arxiv.org/abs/2503.04989

**Artificial Intelligence in Agro-Food Systems: From Farm to Fork**
*Foods* Journal, MDPI (2025) · doi.org/10.3390/foods14030411

## Languages

English (C2, IELTS 8.5) Persian (Native) Italian (A2) Dutch (beginner, learning)