DHEERAJ PARKASH

Data Scientist — Machine Learning Engineer | dheerajparkash.github.io
dheerajparkash36@gmail.com | linkedin.com/in/dheerajparkash/ | +33 7 58 88 06 59 | Paris, Ile-de-France, France

PROFESSIONAL SUMMARY

Data Scientist with strong expertise in machine learning, NLP, and multimodal modeling, supported by solid foundations in statistics and optimization. Experienced in designing scalable data pipelines, improving data quality, and developing predictive models from research to prototype deployment. Proven ability to translate complex analytical outputs into actionable insights for cross-functional engineering and product teams.

SKILLS

Prog/Tools:	Python, SQL, Pandas, NumPy, scikit-learn, PyTorch, TensorFlow, HF Transformers, Docker, Git
ML & AI:	Supervised learning, NLP, Multimodal Modeling, LLM fine-tuning, Model Validation, Explainability
Data Eng:	Data pipelines, ETL design, Feature engineering, Data quality assessment, Azure, MLflow, Databricks
Analytics:	Statistical inference, Time-series modelling, AUC, Log-loss, Regularisation, Experiment Analysis
Prof. Skills:	Data-driven decision-making, stakeholder communication, problem-solving, cross-functional collaboration
Language	Professional English - daily working language, Basic French(learning)

EXPERIENCE

ALTEN ***Paris, France***
Data Scientist Intern **April - September 2025**

- Designed and implemented **scalable multimodal AI pipelines**, validating 11k+ segments for downstream models.
- Built **automated pattern detection** & **data imputation workflows**, improving data quality & consistency metrics.
- Applied **signal processing** and feature transformations to enhance annotation consistency and inter-rater reliability.
- Evaluated AI system performance using **Krippendorff's Alpha, ICC, CCC, and LORO** analyses.
- Trained **multimodal Transformer models** for emotion prediction, improving agreement scores from 0.27 to 0.62.
- Collaborated with **engineering stakeholders** to integrate improved fusion strategies into the deployable prototype.

INRIA/I3S Sophia Antipolis ***Nice, France***
Machine Learning Engineer Intern: **June - August 2024**

- Developed **scalable data pipeline** for multi-party conversational behavior analytics on French-language datasets
- Built **hate speech classification model** combining lexical, graph & contextual features , achieving **78.8% F1-Score**
- Analyzed **interaction graphs** to extract behavioral pattern supporting content moderation and analytics systems.

Data Scientist intern **March - May 2024**

- Built **structured prediction models** for **aggression** and **biases** detection large-scale social media datasets.
- Modeled **discursive and relational interaction patterns** to uncover user behavior relevant to applied NLP systems.

GenAI Research Intern **October - December 2023**

- Fine-tuned and evaluated **LLMs (GPT-2, T5)** using **few-shot learning** for text classification and generation tasks
- Generated **synthetic datasets** to improve coverage of rare and implicit language patterns, enhancing robustness.
- Evaluated and refined model outputs to **reduce false negatives** and improve dataset quality in edge-case scenarios.

EDUCATION

Universite Paris-Saclay & ENS Paris-Saclay **2024-2025**
Masters M2 Data Science, **Grade: 14.7/20** *Paris-France*

Universite Cote d'Azur **2023-2024**
Masters M1 Computer Science, **Grade: 14.8/20 Rank 5th/38** *Nice, France*

Sukkur IBA University **2018-2022**
Bachelors of Science Computer Science, **CGPA: 3.48/4** *Sukkur, Pakistan*

Coursework: *Probability & Statistics, Algorithms for Data Science, Combinatorial Optimization, Operations Research*

PROJECTS

Conditional Graph Generation with Transformers and Diffusion Models GitHub Link **2025**

- Designed **conditional graph generation systems** using Graph Transformers with global attention mechanisms
- Integrated **T5** & **BERT embeddings** to condition graph models, reducing link prediction **MAE from 0.90 to 0.18**.
- Implemented **VAE**, **GAN**, and **diffusion generative** models to improve structural diversity and generation quality.

Explainability and Fairness in NLP Moderation Models GitHub Link **2025**

- Improved **model interpretability** using token-level **necessity and sufficiency** metrics for auditing NLP systems.
- Fine-tuned **BERT** on the **EDOS sexism dataset (14k+ samples)** to analyze bias across protected groups.
- Quantified bias with necessity scores up to **0.96** and sufficiency up to **0.93**, strengthening fairness evaluation.

Scalable Recommendation System Optimization GitHub Link **2024**

- Built an **item-item recommendation engine** using cosine similarity and vectorized Python operations.
- Improved runtime, memory efficiency through optimization techniques relevant to **large-scale**, **production systems**.
- Demonstrated applicability to **streaming and enterprise recommendation environments**.