

ATRI VIVEK SHARMA

[linkedin.com/in/atri-sharma](https://www.linkedin.com/in/atri-sharma) ◇ London, UK ◇ atriviveksharma@gmail.com

EDUCATION

PhD in Computer Science

Imperial College London

2023 - 2027 (expected)

London, UK

- Conducting research on AI safety, adversarial robustness and model evaluation for machine learning systems under the supervision of **Prof. Alessio Lomuscio**, including conversational LLM agents, RAG systems, and decision-tree ensembles
- Supported by the **UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence**.

MEng in Aeronautical Engineering

Imperial College London

2017 - 2021

London, UK

- First Class Honours (78%), Dean's List

EXPERIENCE

Research Scientist (Part-time)

Safe Intelligence Ltd.

Nov 2023 - Present

London, UK

- Core contributor at Imperial College London spintout for the product design and R&D for a SaaS product for testing and validation of AI agents
- Designed iterative, optimisation-based methods for multi-turn jailbreaking of AI agents, achieving an attack success rate of up to 50% against Sonnet 5, GPT-5.6, and Gemini 3.6-powered agents evaluated on HarmBench behaviours, exposing alignment failures that emerge over extended interactions
- Developed optimisation-based black-box methods that generate intent-preserving adversarial queries to stress-test RAG agents, eliciting evidence-contradicting intrinsic hallucinations in up to 54% of responses from GPT-5 and Gemini 3 models
- Led the synthetic test-data generation workstream, designing methods to achieve broad domain coverage for systematic agent evaluation
- Proposed and implemented a robust gradient boosting algorithm for tabular data models, improving robust accuracy by 44% over the state of the art and 113% over conventional XGBoost

Product Engineer

Pangaea Data Ltd.

Jul 2021 - Sep 2023

London, UK

- Directed a four-person cross-functional NLP and product team to identify patients with rare diseases from electronic health records, collaborating with clinicians and stakeholders to refine research objectives and validate product design
- Designed a clinical NLP pipeline that used BERT-based models and regex keyword detection to mine unstructured text, extract ontology concepts, and identify patients through a decision-tree model, distributed on Microsoft Azure Marketplace
- Led a proof of concept for a major pharmaceutical company, applying the pipeline to data from approximately 65,000 patients to identify cohorts with cachexia, COPD, and lupus nephritis; the approach identified 2-5x as many patients compared to diagnostic code baseline.
- Authored technical white papers, research grant submissions and project proposals leading to 2 POCs

SELECTED PUBLICATIONS

- [Intent-Preserving Adversarial User Simulation for Multi-Turn Agent Evaluation](#)
[Atri Vivek Sharma](#), Brian Formento, Alessio Lomuscio
UserSim Workshop at NeurIPS, 2026 (to appear)
- [Eliciting Intrinsic Hallucinations in LLMs via Semantically Equivalent Adversarial Attacks](#)
[Atri Vivek Sharma](#), Brian Formento, Alessio Lomuscio
Conference on Language Modeling (COLM), 2026
- [Learning Robust XGBoost Ensembles for Regression Tasks](#)
[Atri Vivek Sharma](#), Panagiotis Kouvaros, Alessio Lomuscio
Uncertainty in Artificial Intelligence (UAI), 2025

SKILLS

- **Programming and Development:** Python, C++, SQL, Git, Linux
- **Machine Learning and Distributed Computing:** PyTorch, TensorFlow, Hugging Face Transformers, Ray
- **LLM and Agent Tools:** LangChain, LlamaIndex, OpenAI Agents SDK, Amazon Bedrock, Ollama, vLLM
- **Cloud and Infrastructure:** Microsoft Azure, Google Cloud, AWS, Docker, PBS HPC software