Philippe Bergna

Philippe-Bergna | pbergna753@gmail.com | philippe753 | +44 7858362489 | British Citizen

AI Research Engineer with a strong track record in deep learning, adversarial robustness, and large-scale model evaluation. Led research and production efforts across computer vision, LLMs, and active learning pipelines. Experienced in training foundation models from scratch, optimizing data efficiency, and deploying scalable ML systems—including a Retrieval-Augmented Generation (RAG) chatbot for querying AI literature.

EDUCATION

University of Bristol — Bristol, Uk
BEng Engineering Mathematics with Study Abroad — 09/2018 - 06/2022
MEng Engineering Mathematics — 09/2021 - 06/2022

Katholieke Universiteit Leuven — Leuven, Belgium
MS Artificial Intelligence (year abroad) — 09/2020 - 07/2021
Took an AI master's degree in my third year at University.

PROFESSIONAL EXPERIENCE

Advai — London, UK

Research Engineer – Machine Learning & Robust AI Systems — Nov 2023 – present
Machine Learning Researcher — Sep 2022 – Nov 2023

Outline
Conducted cutting-edge AI safety research in both LLMs and Computer Vision, focusing on adversarial attacks, model robustness, jail breaking.

Key Responsibilities
- **Adversarial CV Attacks:** Designed both digital and 3D physical adversarial examples to stress-test facial verification and object detection systems (e.g., YOLO, Faster-RCNN).
- **LLM Evaluation:** Designed behavioral stress tests and prompt variations to assess robustness and alignment failures in GPT-3 and LLaMA 3 models.
- **Model Training:** Trained foundational CV and LLM models from scratch for robustness evaluation and benchmarking.
- **Active Learning Pipelines:** Built adversarial and Bayesian sampling strategies to optimize data efficiency in object detection tasks.
- **Consultant Role:** Lead advisor for model evaluation, adversarial robustness, and OOD detection strategy across client projects.

Key Achievements
- **Selected 1 of ~500** applicants and one of few research scientists hired without a PhD.
- Developed 3D adversarial camouflages with a **95% evasion rate**, securing multimillion-pound UK MoD contracts.
- Facial spoofing attacks achieved **55% success** on iProov systems.
- OOD active learning reduced dataset size by **50%** while improving mAP by **15%**.
- Jailbreak tests revealed **7 novel bypass strategies** with ~10% success against GPT-3/LLaMA-3 safety filters.
- Extended open-source `PyRelationAL` to CV active learning tasks.

University of Bristol — Bristol, UK

Teaching Assistant — 09/2021 – 06/2022

Outline
Guided undergraduates through core modules in Artificial Intelligence, Data Science, and others by delivering interactive tutorials, mentoring projects, and elucidating complex machine learning concepts.

Key Responsibilities

- **Lecturing & Tutoring:** Delivered tutorials and interactive coding sessions on key machine learning topics—such as convolutional and recurrent neural networks, attention mechanisms, and dimensionality reduction—ensuring students understood both theoretical underpinnings and practical implementation.
- **Project Supervision:** Supervised final-year projects in Computer Vision; guided students on dataset design, model development (e.g., CNNs, ResNets), metric analysis, and reproducibility.

Key Achievements

- The only TA without a PhD teaching the Artificial Intelligence course module.

Front-End Developer — Remote

Alphard Technology — 06/2020 – 09/2020

Outline

Built a cross-platform React Native app, designing UI components and ensuring responsiveness with designers and backend engineers.

Key Responsibilities

- Built and styled front-end components using **React Native** and **JavaScript**, ensuring smooth cross-platform functionality.
- Integrated UI with back-end APIs, implemented navigation flows, and handled edge-case UI logic for dynamic user input.
- Collaborated using **Git**, **GitKraken**, and participated in agile sprint cycles including stand-ups and code reviews.
- Supported deployment and CI/CD processes using **AWS** tools and services.

PUBLICATIONS

- **Philippe Bergna**, Jake Thomas. "Multi-Dimensional Adversarial Input Preprocessing for Out-of-Distribution Detection" submitted to the International Conference on Machine Learning (ICML).
- **Philippe Bergna**, Richard Bergna. "Trust but Verify: Quantifying Model Uncertainty in Adversarial Prompt Attacks on LLMs". In preparation for submission to the Neural Information Processing Systems Foundation (NeurIPS).

PROJECTS

For more information about my projects, checkout my website and GitHub
https://philippe-753.github.io/philippebergna.github.io/project/
https://github.com/

Out-of-Distribution Research Paper
Proposed and first-authored a novel method for OOD detection using small, input-specific adversarial perturbations (*adversarial probes*) to test the stability of model confidence under targeted shifts. These probes act as diagnostic tools, revealing inconsistencies in model behaviour that correlate with OOD data. Demonstrated strong performance across standard OOD benchmarks, particularly in near-OOD settings such as CIFAR-10 vs CIFAR-100 and TinyImageNet.

AI Safety RAG Chatbot
Built and deployed a Retrieval-Augmented Generation (RAG) chatbot for querying AI safety topics. Curated a custom corpus of 50 AI safety papers, indexed with FAISS and LangChain, and integrated with ChatGPT-3.5 (easily switchable to any GPT-family model). Deployed publicly on personal website:
`https://philippe-753.github.io/philippebergna.github.io/project/rag-chatbot/`.

Extracting Training Data from LLMs
I'm currently replicating and analysing the results of a recent paper demonstrating how training data can be extracted from aligned large language models (LLMs) via next-token sampling and finetuning (https://openreview.net/pdf?id=vjel3nWP2a), and exploring the implications for memorization and data leakage in modern LLMs.

TECHNICAL SKILLS

Programming Languages

- Python (Expert), MATLAB, JavaScript (Advance), C++, C# (intermediate).

Python Libraries & Software

- Expert with Pytorch, TensorFlow, Hugging Face, Keras, Numpy, Pandas, Scikit-Learn, MLflow, wandb, Docker, Kubernetes, Langchain, FAISS and more.
- Work with Jupyter Notebook, Pycharm, and Visual Studios (preferred).