Informatics Forum
Edinburgh, UK EH89AB
+44 7516162339
iam@antreas.io
https://antreas.io/
GitHub
Google Scholar

Antreas Antoniou

AI Research Scientist and Engineer

Education

2017–2021 **PhD in Machine Learning**, *The University of Edinburgh*.

2016–2017 **MScR in Data Science with Distinction**, *The University of Edinburgh*.

2014–2015 **MSc in Data Science with Distinction**, *Lancaster University*.

2011–2014 **BEng (Hons) in Computer Systems Engineering**, *Lancaster University*.

Employment

2025-Current **Principal Research Scientist and Head of ML Research**, *Pieces.App*, Led research of on-device models (20-80M params) with sub-100ms inference. Supervised design of RAG systems querying millions of vectors in <20ms, and a novel Evolution+LLM system for deep research report generation. Spearheaded research on massive-context models for single GPUs (under NDA).

2024-2025 **Senior Machine Learning Research Scientist/Engineer**, *Malted AI*, Undertaken fundamental research in LLM training/fine tuning, synthetic data generation and curation, automatic prompt learning, LLM committee-based decision-making and individual and commitee-level distillation of systems into smaller more scalable and cheaper alternatives, which formed the key recipes that the company uses to build their products and services.

2021-2024 **Research Associate on Data-Efficient, Highly Transferable and Robust Generalization Learning**, *University of Edinburgh*, Gained recognition for contributions exceeding normal job expectations, including research support and team leadership. Rewrote codebases for efficiency and adaptability, significantly facilitating research projects. Acted as a mentor, enhancing the research capabilities of team members, Edinburgh, UK.

2020-2021 **Research Intern on Few-Shot Learning**, *Google*, Focused on enhancing the transferability of Google's few-shot learning systems, particularly under conditions of extreme domain shift, Mountain View, USA.

2017-2020 **Machine Learning Practical Lead Teaching Assistant, Group Tutor, Demonstrator and Piazza Instructor**, *University of Edinburgh*, Provided comprehensive support for the Machine Learning Practical course, including tutoring and course material development, Edinburgh, UK.
`https://antreas.io/teaching/`

2016 **Speech-Scientist Intern**, *Amazon*, Worked on the extension and improvement of the capabilities of Amazon Echo, Cambridge, UK.

2015 **Research Associate**, *Lancaster University*, Participated in the Deep Online Cognition project, utilizing a new component-based programming language, Dana, for modular software development, Lancaster, UK.

2014 **Embedded Systems Research Intern**, *Lancaster University*, Designed, built, and programmed new hardware for Blackpool Illuminations, focusing on LED control via PWM and high-voltage frequency control, Lancaster, UK.

2013 **Software Developer Intern**, *Lancaster University*, Developed an Android app to facilitate real-time interaction between presenters and audiences, Lancaster, UK.

Research Projects

2025-Current **On-Device Language and Embedding Models**, *At Pieces, led research for small-scale language models (20-80M parameters) that match the performance of oracle models like Gemini 2.5 Pro while achieving on-CPU inference speeds of 10-100ms.*

2025-Current **High-Performance RAG Systems**, *Supervised the design and development of highly optimized RAG embedding models capable of querying millions of records in under 20ms. All work is under NDA.*

2025-Current **Evolutionary Deep Research Reports**, *Oversaw the system design for "deep study" reports that leverage evolutionary methods to achieve and surpass state-of-the-art performance. All work is under NDA.*

2025-Current **Massive-Context On-Device Models**, *Led blue-sky research into training smaller models capable of handling massive context windows on a single GPU machine. All work is under NDA.*

2024-2025 **Fine-Tuning Large Language Models for Small Datasets**, *At Malted AI, spearheaded experiments to fine-tune large-scale language models (e.g., LLAMA 3.1, Phi 3.5 mini, Qwen2) on small datasets (32 to 2048 samples) to identify optimal fine-tuning strategies. Techniques included LoRA, QLoRA, and full fine-tuning, targeting datasets like DROP and TREC-COVID. A key component was designing and implementing flexible, maintainable, and efficient framework codebases to support a variety of configurations. All results and reports are under NDA.*

2024-2025 **Synthetic Data Generation via Teacher Model Committees**, *At Malted AI, developed a robust system for synthesizing high-quality text data by leveraging committees of teacher models. Researched heuristics linking data quality to agreement patterns among teachers and designed framework codebases to support scalable and modular experimentation pipelines. This work formed the foundation for scalable synthetic data generation systems. Full details are under NDA.*

2024-2025 **Automatic Prompt Optimization for Zero- and Few-Shot Tasks**, *Led the development of a framework at Malted AI to automate prompt discovery using evolutionary algorithms. The system was optimized for flexibility, readability, and maintainability, enabling discovery of optimal prompts for improving task-specific performance under zero-shot and few-shot regimes. This work was foundational to improving task-specific adaptability across various applications. Details remain under NDA.*

2024-2025 **Committee-Based Decision Making for Model Ensembles**, *At Malted AI, investigated ensemble-based decision-making by integrating outputs from teacher models through methods like averaging, threshold-based voting, and small neural networks. Designed the framework code to allow experimentation with different integration strategies while ensuring it was maintainable and adaptable for future research. All results and frameworks are protected under NDA.*

2024-2025 **Model Distillation into Scalable Architectures**, *At Malted AI, conducted research to distill large-scale language models (e.g., 8B parameters) into sub-600M models optimized for efficiency and scalability. Built and optimized codebases to handle experimentation with hyperparameters such as architecture, pretraining dataset, pretraining task, weight decay, learning rate, dataset size, and dropout, focusing on flexibility for team use. Full details of recipes and implementations are under NDA.*

2021-2024 **GATE: Diversifying and Robustifying Representation Learning**, *Led the development of a multi-domain, multi-task, and multi-modal benchmark suite aimed at diversifying empirical evaluations and conclusions [**paper**].*

2021-2023 **TALI: Democratizing Multi-Modal Large Scale Machine Learning**, *Spearheaded an open-source quadra-modal dataset to democratize multi-modal machine learning [**dataset**].*

2017-2021 **PhD Thesis: Meta-Learning for Few-Shot Learning**, *Conducted in-depth research on both supervised and unsupervised few-shot learning. Contributions include [17, 14, 15]. [**thesis link**].*

2017 **MScR Thesis: Data Augmentation Generative Adversarial Networks**, *Developed a Generative Adversarial Network for data augmentation, resulting in improved generalization for machine learning models [20].*

2014 **BEng Dissertation: Fault Tolerant, Self Monitoring Sensors**, *Researched a professional-grade sensing system capable of self-validating its own functionality using signal injection techniques and fault prediction.*

Awards and Nominations

2025 LeRobot Global Hackathon: 7th Place International Finish: Originally conceived and initiated the Edinburgh node, built the organizing team, and managed 60% of organizational work. Led the team (7th/1000+ teams), providing mentorship and all hardware. [website] [video] [github]

2023 Sustained Excellence Contribution Reward as Salary Increment

2020-2021 Staff Award for being the Machine Learning Practical Teaching Assistant – `https://antreas.io/awards`

2019 Received 5 Teaching Award Nominations: Best Practice in Inclusive Learning, Best Support Staff, and two for Best Student Who Tutors – `https://antreas.io/nominations`

2019 Finalist for the Best UK PhD Tutor Award

2019 Top-3 Finalist in UK Open Source Awards for my MAML++ Framework

2018 Nominated for the Best Student Who Tutors Award

2015 Awarded the IBM Prize for Best Data Science Dissertation

2014 Received MSc Data Science Scholarship

2014 Secured 2nd Place in Lancaster University CS Hackathon 2014

Teaching and Mentorship

2022 **Accelerated Deep Learning Fundamentals Course**, *Provided a 6-week accelerated course in deep learning fundamentals for a part-time student lacking the requisite background. Met once a week to ensure skill acquisition.*

2021-2023 **Mentorship and Team Leadership**, *Onboarded new team members and provided weekly teaching sessions on specific university modules (such as MLP) to enrich research capabilities. Conducted regular meetings to discuss research, engineering, and software development problems.*

2021-2023 **One-on-One Research Guidance**, *Conducted one-on-one sessions to debug machine learning and software issues, unblocking stalled research directions.*

Sept. 2017 to Current **Machine Learning Practical Course**, *Teaching Assistant, Group Tutor (Effectively Research Supervisor), Demonstrator and Piazza Instructor*, Full Description at `https://antreas.io/teaching/`.

Apr. 2015 to May 2015 **Digital Innovation**, *Teaching Assistant.*

Research Collaboration and Support

2025 **Edinburgh LeRobot Hackathon: Key Organizer and Lead**, *Originally conceived and initiated bringing the global LeRobot Hackathon to the University of Edinburgh, built the organizing team, and managed 60% of the organizational work. Sourced and provided all specialized hardware and high-performance GPU infrastructure. Served as an expert mentor for all teams throughout the 30-hour competition. [**event website**].*

2023 **Community Support for EIDF A100 GPU Cluster**, *One of the early adopters and a key community support member. Created the Slack server, answered hundreds of questions, scheduled community meetings for key issues, and served as a bridge between users and developers.*

2023 **Tooling and Infrastructure**, *Developed a Python package called kubejobs, simplifying Kubernetes job specifications. Authored early documentation for the university's EIDF A100 GPU cluster.*

2022 **POEM Project**, *Directly supported the project by offering bi-weekly engineering and research support. Orchestrated 40 experiments in just two days to meet the paper submission deadline.*

2021 **HDR UK Medical Deep Learning Project**, *Provided significant contributions to the project by rewriting the codebase to improve quality and efficiency. Achieved a 35x speed-up in the training/evaluation pipelines. Provided weekly perspectives, suggestions, and direct support.*

Teaching

2025 **Guest Lecture on Model Compression**, *University of Edinburgh*, Machine Learning Systems Course, Delivered a guest lecture on model compression techniques and industry applications [Slides].

2025 **Invited Talks and Workshops**, *University of Edinburgh*, Community Engagement, Delivered a series of invited talks and workshops, including "GenAI Superpowers" at the Teach-A-Thon [Slides], "LLMs for Teaching" [Slides], and "What I learned after 3000 hours using ChatGPT" [Slides].

2021-2023 **Supervisory Role for MSc Students**, *University of Edinburgh*, Main Supervisory Contact, Acted as the main supervisory contact for 4 MSc students, two of whom successfully completed their thesis with marks between 60-70%, and one of which was of quality and significance suitable for conference submission.

2021-2022 **Accelerated Deep Learning Fundamentals Course**, *University of Edinburgh*, Course Instructor, Designed and delivered a 6-week accelerated deep learning fundamentals course to aid a part-time student lacking the required background, meeting with them once a week.

Sept. 2017 to Current **Machine Learning Practical Course**, *University of Edinburgh*, Teaching Assistant, Group Tutor (Effectively Research Supervisor), Demonstrator and Piazza Instructor, Directly supervised over 60 student projects across 17 groups, alongside developing course materials and managing cloud infrastructure. Full Description at https://antreas.io/teaching/.

Apr. 2015 to May 2015 **Digital Innovation**, *University of Edinburgh*, Teaching Assistant.

Proactive Technical Leadership

Onboarding and Mentorship: Proactively **onboarded new team members**, ensured **smooth and low-friction remote development** on GPU machines, and provided **weekly teaching** on university modules to enrich research capabilities. Conducted **regular meetings** to discuss research, engineering, and software development challenges

Research Quality Enhancement:	Conducted **one-on-one sessions** to debug machine learning model training and generalization behaviors and software issues, thereby **unblocking stalled research directions**
HDR UK Medical Project:	Contributed significantly by **rewriting the project codebase** to increase code quality, adaptability, and efficiency. Achieved a **35x speed up** in the training/evaluation pipelines
Research Collaborations:	Directly supported multiple projects like **POEM** by providing **bi-weekly engineering and research support**. **Orchestrated 40 experiments** in two days to meet project deadlines
Infrastructure and Tooling	**Kubernetes Cluster**: Built and deployed a Kubernetes cluster for the research group. Developed Python tooling for cluster management and gave a tutorial to the group on Kubernetes usage. **Research Framework**: Authored a minimal machine learning research framework for the group, following best practices. **Deep Learning Server**: Successfully procured a £50K deep learning research server through comprehensive market analysis, vendor negotiations, and overseeing the setup process
Community Engagement and Support:	**Early adopter** of the university's EIDF A100 GPU cluster. Authored **early documentation** and a Python package named **kubejobs** to facilitate user engagement. Founded the cluster's **Slack server**, answered hundreds of user questions, and organized **community meetings** for key issue discussions
Research Orchestration:	Improved research orchestration by **learning how to deploy Kubernetes clusters**, subsequently setting up a cluster for the group. Conducted a **tutorial** on Kubernetes usage and provided **Python tooling** for easier management
Knowledge Sharing:	**Built a comprehensive wiki page** for the research group to document best practices, tools, and other valuable resources

Open Source Philosophy and Contributions

Philosophy: My conviction in the *democratization of Machine Learning research* stems from two core beliefs: the irreplaceable value of *individual expertise* and the *power of collective collaboration*. For me, open source is not merely a development model; it is a **fundamental necessity** for driving innovation and maintaining ethical standards in the field.
Contributions: I have been an active participant in the open-source ecosystem, contributing to various projects, educational resources, and frameworks. My contributions are publicly available on my GitHub profile and include:

- **How to Train Your MAML**: A user-friendly framework that simplifies the complexities of MAML meta-learning.
- **Minimal-ML-Template**: A scalable and adaptable machine learning template designed to minimize the initial setup overhead for researchers.
- **Kubejobs**: A Python package that simplifies working with Kubernetes, facilitating an easier and more efficient experience for users.

Impact: These projects reflect my ongoing commitment to *open science*, *efficiency*, and *community-driven research and development*.

References

[1] Antreas Antoniou et al. *EEVEE and GATE: Finding the right benchmarks and how to run them seamlessly*. 2024. URL: https://openreview.net/forum?id=0SMhqvgHST.

[2] Antreas Antoniou et al. *TALI: A Dataset of Temporal Audio Language and Images*. 2024. URL: https://huggingface.co/datasets/Antreas/TALI.

[3] Linus Ericsson et al. "einspace: Searching for Neural Architectures from Fundamental Operations". In: *NeurIPS* (2024).

[4] Alessandro Fontanella et al. "Development of a deep learning method to identify acute ischaemic stroke lesions on brain CT". In: *Stroke and Vascular Neurology* (2024).

[5] Kiyoon Kim et al. "Adversarial Augmentation Training Makes Action Recognition Models More Robust to Realistic Video Distribution Shifts". In: *arXiv preprint arXiv:2401.11406* (2024).

[6] Fady Rezk et al. "Liouna: Biologically Plausible Learning for Efficient Pre-Training of Transferrable Deep Models". In: *2nd Workshop on Advancing Neural Network Training: Computational Efficiency, Scalability, and Resource Optimization (WANT@ ICML 2024)*. 2024.

[7] Alessandro Fontanella et al. "ACAT: Adversarial Counterfactual Attention for Classification and Detection in Medical Imaging". In: *ICML 2023* (2023).

[8] Alessandro Fontanella et al. "Challenges of building medical image datasets for development of deep learning software in stroke". In: *arXiv preprint arXiv:2309.15081* (2023).

[9] Adam Jelley et al. "Contrastive Meta-Learning for Partially Observable Few-Shot Learning". In: *ICLR 2023* (2023).

[10] Fady Rezk et al. "Is Scaling Learned Optimizers Worth It? Evaluating The Value of VeLO's 4000 TPU Months". In: *Proceedings on*. PMLR. 2023, pp. 65–83.

[11] Antreas Antoniou. "Meta learning for supervised and unsupervised few-shot learning". In: (2021).

[12] Timothy Hospedales et al. "Meta-learning in neural networks: A survey". In: *IEEE transactions on pattern analysis and machine intelligence* 44.9 (2021), pp. 5149–5169.

[13] Antreas Antoniou et al. "Defining benchmarks for continual few-shot learning". In: *Meta-Learrning Workshop, NeurIPS 2020*. 2020.

[14] Antreas Antoniou and Amos Storkey. "Assume, Augment and Learn: Unsupervised Few-Shot Meta-Learning via Random Labels and Data Augmentation". In: *arXiv preprint arXiv:1902.09884* (2019).

[15] Antreas Antoniou and Amos J Storkey. "Learning to Learn by Self-Critique". In: *NeurIPS 2019*. 2019, pp. 9936–9946.

[16] Antreas Antoniou et al. *Meta-meta-learning for Neural Architecture Search through arXiv Descent*. 2019.

[17] Antreas Antoniou, Harrison Edwards, and Amos Storkey. "How to train your MAML". In: *ICLR 2019*. 2018.

[18] Antreas Antoniou et al. "Dilated Densenets for Relational Reasoning". In: *arXiv preprint arXiv:1811.00410* (2018).

[19] Luke N Darlow et al. "CINIC-10 is not ImageNet or CIFAR-10". In: *arXiv preprint arXiv:1810.03505* (2018).

[20] Antreas Antoniou, Amos Storkey, and Harrison Edwards. "Data Augmentation Generative Adversarial Networks". In: *ICANN 2018* (2017).

[21] Antreas Antoniou and Plamen Angelov. "A general purpose intelligent surveillance system for mobile devices using deep learning". In: *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2016, pp. 2879–2886.