William Gitta Lugoloobi

KabakaWilliam | wlugoloobi | williamgitta@gmail.com | +447887301898

I'm a Rhodes Scholar and PhD student at Oxford working on security and reliability in post-trained LLMs for reasoning and agentic tasks. My recent work uses steering vectors to modify how LLMs perceive problem difficulty in code and math settings. Steering towards "easy" reduces hallucinations and chain-of-thought length while improving accuracy on math benchmarks. My forthcoming paper develops practical defences in agents against prompt-injection attacks during GRPO on poisoned data.

Education

University of Oxford
DPhil Social Data Science — *Oct. 2024 – Present*

- **Thesis:** Rewarding Reasoning: Post-Training Choices and Security Trade-offs in Agentic LLMs.
- Supervised by Prof. Joss Wright and Prof. Chris Russell .
- Fully funded by the Rhodes Scholarship.

University of Oxford
MSc Social Data Science — *Oct. 2023 – Aug. 2024*

- Courses included Applied Machine Learning, Natural Language Processing, Data Analytics at Scale, Frontiers of Data Science, and Applied Analytical Statistics.
- Fully funded by the Rhodes Scholarship.
- **Thesis:** *From Headlines to Stories: Utilising Transformers for Story Driven OCR of Newspapers.*
- Thesis was well commended for demonstrating that traditional content analysis could be operationalised as a Document Layout Analysis Task.

Northwestern University
BS Journalism and Strategic Communications — *Aug. 2019 – May. 2023*

- Graduated Summa Cum Laude.
- Founded the University's first data science club and lobbied for uptake in CS and AI courses in the journalism department.

Publications

Lugoloobi, William and Chris Russell (Oct. 2025). *LLMs Encode How Difficult Problems Are.* (under review (2025)). DOI: 10.48550/arXiv.2510.18147. URL: http://arxiv.org/abs/2510.18147.

Projects

Berserk MCP — Link to Repo

Created an MCP Server and Client with FastMCP and VLLM to give LLM Agents the ability to search the web for information on when a new chapter of Berserk (my favourite manga) is released

SOTA Verl GRPO Implementation — Link to PR

Contributed a GRPO training script to Verl, exceeding SOTA-level MATH500 for a 1.5B model with a single A100 in under an hour.

Work Experience

Software Engineering Intern, Qatar Computing Research Institute Apr 2022 — Jul 2022

- Prototyped and deployed a Python/Docker image-classification model to detect natural disasters from social-media images.
- Placed 3rd out of 50 interns in QCRI's Summer Intern Research Competition.
- Guided researchers in porting existing code into a production-ready application.

Data Scientist/Analyst Intern, VICE Media Group Feb 2022 — Apr 2022

- Built a data pipeline for VICE Arabia's digital content using Python (Beautiful Soup) and external APIs.
- Designed a dashboard for campaign performance tracking and benchmarking across teams.

Data Science Intern, Al Jazeera Media Group Apr 2021 — Aug 2021

- Implemented a deep-learning model (TensorFlow, Azure Databricks) to detect manipulated images and fake news from social media.
- Created mockups and front-end designs for a newsroom "breaking news" dashboard to streamline information from multiple sources.
- Evaluated Al Jazeera's newswires and identified undelivered articles to the company's Microsoft SQL Server database.

Teaching

University of Oxford, Oxford Internet Institute Jan 2025 — Present

- Teaching Assistant, *Introduction to Natural Language Processing* (Oxford's core NLP course).
- Led weekly tutorials on Naive Bayes, RNNs, and Transformers with over 40+ MSc/PhD students.
- Designed and delivered mini-lectures on Weights & Biases for experiment tracking, re-implementing GPT-2 locally, and using shared GPU clusters (SLURM/vLLM).
- Received commendations from the course convenor and students for clarity in lectures, plus for introducing practical tooling.

Northwestern University Aug 2022 — December 2022

- Teaching Assistant, *Introduction to Web Design* and *Statistics*
- Hosted weekly office hours to aid students struggling to complete problem sheets
- Convened seminars on how to use both R and Python for statistical analysis

Skills

Programming Languages	R, Python, TypeScript, Javascript
Most used Libraries	Pytorch, VLLM, Numpy, Verl, HuggingFace, Tidyverse, React.js, Tailwind
Leadership	Keble MCR President, Northwestern Data Science Club President, Rhodes Scholarship