Anthony Hughes

ajhughes3@sheffield.ac.uk anthonyhughes.github.io @anthonyhughes +447496860312

Summary

I bridge the gap between AI research and real-world deployment through 10 years of software engineering. I now research language models, security and privacy with publications at EMNLP and EACL. My work investigates privacy leaks in language model generated texts. I have recently collaborated with the CrySP group at the University of Waterloo on mechanistic approaches to PII leakage. We developed circuit interpretability techniques to understand the mechanisms behind sensitive information leaks, and designed mitigation strategies to defend against these vulnerabilities.

Currently funded through a fully-funded UKRI PhD scholarship, I'm investigating how machine learning models process and inadvertently expose information. This includes researching training data poisoning and exploring how frameworks like differential privacy can protect both models and individuals' information. These findings can inform the design of more trustworthy AI systems.

I want to do great AI research by combining mathematical and scientific rigour with strong engineering skills. My goal is to help ensure we maintain in control over the trade-offs of security, privacy and utility.

Education

- **PhD Computer Science**

University of Sheffield, 2024-***Current***

Courses: *Time-series Analysis, Machine Learning, Speech Processing and Technologies, Text Processing, and Natural Language Processing.*

- **PgDip Speech and Language Technologies - Leadership**

University of Sheffield, 2023-*2024*

- **MSc Computational Linguistics (Distinction)**

University of Wolverhampton, 2021-2023

- **BSc Computer Science (1st Class, Hons)**

Nottingham Trent University, 2009-2013

Research Experience

- **Visiting Graduate Researcher | July 2025 - September 2025 | Collaborators/Mentors: Vasisht Duddu, N. Asokan**
  Cryptography, Security and Privacy Group (CrySP), University of Waterloo.
  - Submitted to EACL: *Understanding and Mitigating PII Leakage in Language Models: A Mechanistic Approach*
  - Ongoing Project: *How Do Language Models Encode Privacy Norms?*

Selected Publications

- **PATCH: Mitigating PII Leakage in Language Models with Privacy-Aware Targeted Circuit PatcHing**
  - EACL 2026 | Findings
  - https://arxiv.org/abs/2510.07452
- **How Private are Language Models in Abstractive Summarisation?**
  - EMNLP 2025 | Main Proceedings
  - https://arxiv.org/abs/2412.12040
- **Identifying and Aligning Medical Claims Made on Social Media with Medical Evidence**
  - LREC-COLING | 2024
  - https://aclanthology.org/2024.lrec-main.753/
- **Generative Byte-Level Models for Restoring Spaces, Punctuation, and Capitalization in Multiple Languages**
  - Practical Solutions for Diverse Real-World NLP Applications | 2023
  - https://link.springer.com/chapter/10.1007/978-3-031-44260-5_3

Awards and Prizes

- **Best Poster, Insigneo Showcase, July 2025**
- **UKRI PhD Scholarship with the University of Sheffield, 2023-2027**

Selected Projects from Industry

- **Data Language AI | NLP-based Product**

Built an automated text content classification SaaS product, enabling clients easy access text analytics. Led the development of a data visualisation tool allowing customers to view classification quality metrics.

- **Data Graphs | NLP-based Product**

Integrated language models into a SaaS data platform, enabling clients to query their graph data using natural language. Surfaced client graph data for LLM interactions, facilitating more intuitive client engagement with their stored data.

Industry

- **Data Language | Jan 2014 - September 2023**

Junior Engineer → Lead Software Engineer → NLP Engineer

Working as a Lead Software Engineer delivering data and software solutions for a wide range of clients.

- **Ontoba | June 2013 - June 2014**

Software Engineer working on services solving data silo issues with graphs/linked data focussed solutions.

- **Press Assoication | July 2011- September 2012**

Internship (Software Engineer) working on a new digital platform centered around semantic web technologies.

General Tech Skills

- **Programming Languages:** Python, JavaScript, Rust
- **Machine Learning & NLP:** PyTorch, Transformers, Scikit-learn, NLTK, Pre-, Fine-tuning, and Instruction-tuning.
- **Databases & Search:** PostgreSQL/SQL, GraphDB/SPARQL, Elasticsearch, QDrant (vector database)
- **High-Performance Computing:** Slurm job scheduling, distributed computing environments

Mentoring Experience

- **Sajad Rahmanian Ashkezari, Neel Sanjaybhai Faganiya, Lucas Kopp | Sep 2025 → Present**

Project: *How Do Language Models Encode Privacy Norms?*

- **Yangming Cao | Feb 2025 -> Sept 2025**

Project: *Clinical Coding of Medical Texts*

- **Emma Ellwood | Apr 2025 -> Sept 2025**

Project: *Obfuscation of Gender and Familial Information in Medical Summaries*

Talks

- **EurIPS | December 2025 | Foundations of Language Model Security**

Title: *PATCH: Mitigating PII Leakage in Language Models with Privacy-Aware Targeted Circuit PatcHing*

- **EMNLP | November 2025 | Main Conference**

Title: *How Private are Language Models in Abstractive Summarization?*

- **Insigneo Institute | May 2024 | Synthetic Data Workshop**

Title: *Identifying and Aligning Medical Claims Made on Social Media with Medical Evidence*

Funding

- **NIHR/BRC Sheffield | £1650 | APR 2025 → SEPT 2025**

Project: *AI-driven Medical Record Redaction during Adoption and Gender Reassignment in Primary Care*

Academic Service

- **Reviewer, Call For Talks, EurIPS Foundations of Language Model Security | 2025**
- **Reviewer, Student Research Workshop, EACL | 2026**