Anthony Hughes

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

Summary

I bridge the gap between AI research and real-world deployment through a unique combination of 11 years in software engineering and recently research in language models and privacy. My EMNLP (main) paper investigates private summarisation, I recently used circuit interpretability techniques to understand how models leak sensitive personal information, and now researching how interpretability can be extended into a range of attack vectors.

My career trajectory spans from a junior engineer for media companites to building production NLP systems that served major companies, now publishing scientific findings. Currently funded through a competitive UKRI PhD scholarship, I am working to understand mechanisms by which language models process and potentially expose private information, and how we design privacy-preserving techniques around these findings.

I seek to advance AI research by combining deep technical understanding of model internals with practical experience in deploying robust software systems. My goal is to help ensure that as models become more capable, we remain in control of the balance between bias, privacy and personlisation.

Education

- **PhD Computer Science**

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

- **PgDip Speech and Language Technolgoies**

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

Industry

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

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.

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.

- **GFK | Data Engineering Project**

Developed data pipelines integrating with machine learning workflows to drive insights on consumer markets. Key contributor to building informative data science processes for understanding client audiences.

- **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.

Anthony Hughes

Publications

- **How Private are Language Models in Abstractive Summarisation?**
  - EMNLP 2025 | Main Proceedings (Oral)
  - 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/
- **Understanding Inflicted Injuries in Young Children: Toward an Ontology-based Approach**
  - EKAW | 2024
  - https://link.springer.com/chapter/10.1007/978-3-031-77792-9_16
- **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
- **Comparison of Token-and Character-Level Approaches to Restoration of Spaces, Punctuation, and Capitalization in Various Languages**
  - ICNLSP | 2022
  - https://aclanthology.org/2022.icnlsp-1.19/

Preprints

- **PATCH: Mitigating PII Leakage in Language Models with Privacy-Aware Targeted Circuit PatcHing**
  - Under review for EACL 2026
  - https://arxiv.org/abs/2510.07452

Skills

- **Programming Languages:** Python, JavaScript
- **Machine Learning & NLP:** PyTorch, Transformers, Scikit-learn, NLTK, TransformerLens, AutoCircuit
- **Databases & Search:** PostgreSQL/SQL, GraphDB/SPARQL, Elasticsearch, QDrant (vector database)
- **Web Technologies:** HTML/CSS, React.js, Remix
- **High-Performance Computing:** Slurm job scheduling, distributed computing environments

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: *Unintended Interactions between LLM Optimizations and Risks*

Mentoring Experience

- **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*

Invited Talks

- **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*

Awards and Prizes

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

Funding

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

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

Anthony Hughes

Academic Service

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