Anthony Hughes

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

I am a PhD student focusing on privacy in language models, with publications at EMNLP and EACL. Prior to my PhD, I spent ten years in industry as a software engineer.

My PhD develops algorithms for protecting private information in language models, combining mid- and post-training techniques, formal differential privacy guarantees, and circuit-based interpretability to understand how models memorise and surface sensitive information.

I am now putting my efforts toward safety research. As a SPAR fellow, I have been working on benchmarking and developing methods for detecting whether models have been backdoored through poisoned training data. This includes concerning scenarios like secret loyalties. Additionally, my independent work examines vulnerabilities in safety-critical components such as classifiers and probes.

I'm interested in work that deeply understands both the adversary and the defender, with the aim of building robust monitoring and science of models. Right now a fellowship feels a natural next step toward full-time technical safety work.

Education

- **Technical AI Safety | May 2026 - June 2026**
  - Blue Dot Impact
- **PhD Computer Science**
  - University of Sheffield, 2024-***Current (estimated completion April 2027)***
- **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

- **Research Internship | June 2026 - August 2026 |**
  - Collaborators/Mentors: **Jacob Imola, Gautam Kamath (University of Waterloo)**
  - Project: ***Certified Unlearning***
- **Safety Fellowship | March 2026 - May 2026 |**
  - Mentors**: Andrew Draganov (Arcadia/LASR Labs)**
  - Project: ***Detecting Whether an LLM has Been Poisoned (under review at ICML mech interp workshop)***
- **Visiting Graduate Researcher | July 2025 - September 2025 |**
  - Collaborators/Mentors**: Vasisht Duddu, N. Asokan (University of Waterloo)**
  - Accepted EACL: ***PATCH: Mitigating PII Leakage in Language Models with Privacy-Aware Targeted Circuit PatcHing***

Safety Projects

- **Detecting whether an LLM Has Been Poisoned**
  - https://sparai.org/projects/sp26/rec4IhwDkqFTaz45j
- **Boundary-targeted Membership Inference Attacks on Safety Classifiers**
  - https://arxiv.org/abs/2605.22373

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://aclanthology.org/2025.emnlp-main.1531/

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*

Academic Service

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

Mentoring Experience

- **Fatemeh Honarvar Nazari | Jan 2026 → Present**

Project: *Understanding how models reason about private information.*

- **Sajad Rahmanian Ashkezari, Neel Sanjaybhai Faganiya, Lucas Kopp | Sep 2025 → March 2026**

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*

Industry

- **JetBrains | March 2026 - July 2026 | Research Scientist Intern**
  - Privacy preservation of code and proprietary information.
- **Data Language | Jan 2014 - September 2023 | Junior Engineer → Lead Software Engineer → NLP Engineer**
  - Built a classification SaaS product, enabling clients easy access to text analytics. Led the development of a data viz tool allowing customers to view classification quality metrics.
  - 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.
- **Ontoba | June 2013 - June 2014 | SWE**

Working on solving data silo issues with graphs/linked data focussed solutions.

- **Press Association | July 2011- September 2012 | SWE (Intern)**

Working on a new digital platform centered around semantic web technologies.

Awards and Prizes

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