

Igor Shilov

📍 London, UK ✉ hello@igorshilov.com ☎ +447523506334 🔗 igorshilov.com 🌐 igorshh

Summary

AI security and privacy researcher with over a decade of software engineering experience. Currently in the final year of a PhD at Imperial College London. Research interests include adversarial robustness, memorization, and model internals.

Experience

Goodfire AI , Research Fellow	Jun - Oct, 2026 London, UK
• Research project on mechanistic interpretability. Developed the neuron community discovery method to identify neuron-sparse representations in LLMs.	
Anthropic , AI Safety Fellow	Mar - Sep, 2025 London, UK
• Research project on capability removal. Developed Selective Gradient Masking, a pretraining intervention to localize target knowledge in LLMs for later removal.	
Meta , Staff Software Engineer	2017 - 2023 London, UK
• PyTorch framework for differential privacy. Lead developer and maintainer of Opacus, a PyTorch library for differentially private model training, used by Meta teams and external researchers (2,000+ GitHub stars). • Privacy-preserving ML. As a research engineer, applied differential privacy and federated learning to products and supported research on differential privacy and privacy attacks. • StopNCII.org. Led development of the privacy-preserving platform with the UK-based NGO Revenge Porn Helpline, using on-device perceptual hashing to help combat non-consensual intimate image sharing. • Safety ML. Led a team of four engineers building models to detect and remove harmful content.	
Zvooq , Senior Software Engineer	2014 - 2017 Moscow, Russia
• Recommender system. Led a team of four engineers building a music recommendation engine. • Analytics engine. Built an internal analytics system for copyright-holder reporting.	
Mail.ru , Software Engineer	2013 - 2014 Moscow, Russia
• Search ranking. Improved the ranking algorithm for Mail.ru's search engine, which held roughly 10% of the Russian market at the time.	
Côc Côc , Software Engineer	2011 - 2012 Moscow, Russia Hanoi, Vietnam
• Search infrastructure. Built a distributed web crawler and improved query intent classification.	

Education

PhD student	Imperial College London Computing Research	2023 - Present London, UK
Specialist degree (5 years)	Lomonosov Moscow State University Applied Mathematics and Computer Science	2008 - 2013 Moscow, Russia

Selected Publications

- Claudini: Autoresearch Discovers State-of-the-Art Adversarial Attack Algorithms for LLMs** 2026
Alexander Panfilov*, Peter Romov*, Igor Shilov*, Yves-Alexandre de Montjoye,
Jonas Geiping, Maksym Andriushchenko
Advances in Neural Information Processing Systems (NeurIPS), 2026
- The mosaic memory of large language models** 2026
Igor Shilov*, Matthieu Meeus*, Yves-Alexandre de Montjoye
Nature Communications, 2026
- Beyond Data Filtering: Knowledge Localization for Capability Removal in LLMs** 2025
Igor Shilov, Alex Cloud, Aryo Pradipta Gema, Jacob Goldman-Wetzler,
Nina Panickssery, Henry Sleight, Erik Jones, Cem Anil
arXiv preprint 2512.05648
- Free Record-Level Privacy Risk Evaluation Through Artifact-Based Methods** 2025
Joseph Pollock*, Igor Shilov*, Euodia Dodd, Yves-Alexandre de Montjoye
USENIX Security Symposium, 2025
- SoK: Membership Inference Attacks on LLMs Are Rushing Nowhere (and How to Fix It)** 2025
Matthieu Meeus, Igor Shilov, Shubham Jain, Manuel Faysse, Marek Rei, Yves-Alexandre de Montjoye
IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), 2025
- Copyright Traps for Large Language Models** 2024
Matthieu Meeus*, Igor Shilov*, Manuel Faysse, Yves-Alexandre de Montjoye
International Conference on Machine Learning (ICML), 2024
- Opacus: User-Friendly Differential Privacy Library in PyTorch** 2021
Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek,
John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, Ilya Mironov
NeurIPS Workshop on Privacy in Machine Learning, 2021

Complete Publication List

- Estimating Model-Level Membership Inference Vulnerability Without Reference Models** 2026
Euodia Dodd, Nataša Krčo, Igor Shilov, Matthew Wicker, Yves-Alexandre de Montjoye
Advances in Neural Information Processing Systems (NeurIPS), 2026
- Exploring the limits of strong membership inference attacks on large language models** 2025
Jamie Hayes, Ilia Shumailov, Christopher A. Choquette-Choo, Matthew Jagielski,
Georgios Kaissis, Milad Nasr, Meenatchi Sundaram Muthu Selva Annamalai,
Niloofar Mireshghallah, Igor Shilov, Matthieu Meeus, Yves-Alexandre de Montjoye,
Katherine Lee, Franziska Boenisch, Adam Dziedzic, A. Feder Cooper
Advances in Neural Information Processing Systems (NeurIPS), 2025
- RippleBench: Capturing Ripple Effects Using Existing Knowledge Repositories** 2025
Roy Rinberg, Usha Bhalla, Igor Shilov, Flavio P Calmon, Rohit Gandikota
Mechanistic Interpretability Workshop at NeurIPS, 2025

Certification for Differentially Private Prediction in Gradient-Based Training Matthew Wicker, Philip Sosnin, <u>Igor Shilov</u> , Adrianna Janik, Mark Niklas Mueller, Yves-Alexandre de Montjoye, Adrian Weller, Calvin Tsay <i>International Conference on Machine Learning (ICML), 2025</i>	2025
Counterfactual Influence as a Distributional Quantity Matthieu Meeus, <u>Igor Shilov</u> , Georgios Kaissis, Yves-Alexandre de Montjoye <i>The Impact of Memorization on Trustworthy Foundation Models Workshop at ICML, 2025</i>	2025
Sub-optimal Learning in Meta-Classfier Attacks: A Study of Membership Inference on Differentially Private Location Aggregates Yuhan Liu, Florent Guepin, <u>Igor Shilov</u> , Yves-Alexandre de Montjoye <i>arXiv preprint 2412.20456</i>	2024
Watermarking Training Data of Music Generation Models Pascal Epple, <u>Igor Shilov</u> , Bozhidar Stevanoski, Yves-Alexandre de Montjoye <i>arXiv preprint 2412.08549</i>	2024
Defending against Reconstruction Attacks with Rényi Differential Privacy Pierre Stock, <u>Igor Shilov</u> , Ilya Mironov, Alexandre Sablayrolles <i>arXiv preprint 2202.07623</i>	2022
Antipodes of label differential privacy: PATE and ALIBI Mani Malek Esmaeili, Ilya Mironov, Karthik Prasad, <u>Igor Shilov</u> , Florian Tramèr <i>Advances in Neural Information Processing Systems (NeurIPS), 2021</i>	2021