Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Central Bank of Ireland PhD Programme in AI & Data Science (4-years full-time)

Project title: Explaining and Mitigating Bias in Advice from Large Language Models.

Project supervisor: Dr. Eoin Delaney (Trinity College Dublin) and Prof. Mark Keane (University College Dublin).

Project locations: School of Computer Science and Statistics, Trinity College Dublin.

Application deadline: 02 June 2025. Applications will be reviewed on a rolling-basis and early application is strongly recommended.

Start date: September 2025.

PhD structure: The funding for the project includes a tax-free stipend of €25,000 per annum. In addition to stipend, fees will be covered for four years and there is a budget for equipment and conference travel. This project is supported by the Central Bank of Ireland and the Insight Centre for Data Analytics which is funded by Research Ireland.

PhD topic: This project aims to address how language models can *nudge* people towards irresponsible behavior by amplifying existing cognitive biases. We aim to: (1) develop novel human-centered explanation techniques that help people to critically evaluate generative AI models with a focus on financial advice; (2) provide empirical evidence on how advice from Large Language Models (LLMs) can be overconfident and biased; and (3) create practical guidelines for responsible AI system design. By combining insights from cognitive psychology with explainable AI (XAI) techniques, our work directly supports the Central Bank's regulatory objectives under the EU AI Act by addressing systems that could deploy subliminal, manipulative, or deceptive techniques, impairing informed financial decision-making. The resulting frameworks from this project will equip practitioners with evidence-based approaches for evaluating and regulating AI advisory systems while protecting people as these technologies become increasingly

prevalent.

The Institution: The School of Computer Science and Statistics at Trinity College Dublin is a collegiate, friendly, and research-intensive centre for academic study and research excellence. The School has been ranked 1st in Ireland, top 25 in Europe, and top 100 Worldwide (QS Subject Rankings 2018, 2019, 2020, 2021, 2023).

Requirements: Applicants should have (or expect to attain prior to project start) at least a 2.1 honours degree or equivalent in the areas of computer science, mathematics, applied mathematics, statistics, human computer interaction or related disciplines. Applicants must demonstrate proficiency in machine learning or statistical modelling and have experience with computing through Python. Demonstration of open source project work (e.g., GitHub repositories), and familiarity with machine learning and deep learning frameworks (scikit-learn, PyTorch, LLM API's) is a plus. Experience working with large language models or other generative models is an added bonus. Applicants should demonstrate an interest in human-centered machine learning, interpretability and applied machine learning. Applicants for whom English is a second language will be required to demonstrate their competence in the English language in line with Trinity College Dublin requirements as appropriate.

Application: Applicants should email Dr. Eoin Delaney (eoin.delaney@tcd.ie) to apply. The application should include a comprehensive CV (2-pages max), academic transcripts of the degree/ degrees, and a short cover letter/statement of purpose (2-pages max) indicating how their skills align with the project and their motivation for applying. Please submit these documents as a single pdf. Please include "CBI PhD Application" followed by your name in the subject line. The application CV should, at minimum, include the applicant's name, educational institution, qualification stating overall grade/percentage (predicted grades are acceptable for those still studying) and contact details of two academic referees. Informal queries can be made to: eoin.delaney@tcd.ie.