

Project Details	
Project Code	MRCPHS24Br Millard
Title	Exploring the value of using large third-party artificial intelligence models in epidemiology, with examples using Twitter data
Research Theme	Population Health Sciences
Summary	Large language models (LLMs) are an artificial intelligence approach that have recently been shown to have extremely promising ability, for example, for conversing with humans or performing tasks such as summarising or extracting information from text. This project will investigate the potential opportunities and challenges of using LLMs in epidemiological research and explore the use of these models for deriving health traits from Twitter data as an example.
Description	<p><b>Aim:</b> This project will investigate the opportunities and challenges of large language models to assist epidemiological research and explore applications of large language models for deriving health phenotypes from Twitter data as an example. <b>Background:</b> Large language models (LLMs) are an artificial intelligence approach that typically have a very large number of parameters (e.g. millions or billions) and have been trained on extremely large datasets. In recent years these models have gained substantial attention, as they have demonstrated extremely promising performance for being used for a variety of tasks, and they are set to disrupt the way tasks are conducted across many areas of life. They are already being adopted to help scientists do their work more efficiently, for example, writing code or helping to write research publications. In epidemiology research there are also opportunities to exploit these pre-trained models, such as for assisting with summarisation, information extraction or prediction using textual data.</p> <p><b>Objective 1:</b> To review the literature and availability of LLMs, to determine the opportunities and challenges for using LLMs to assist epidemiology research. This could include consideration of: (1) The LLMs that are available and the differences between them (e.g. in terms of performance/ capability, environmental impact); (2) limitations of using LLMs for epidemiological research, for example hallucinations, model interpretability; (3) ethical considerations, for example, the use of LLMs with sensitive epidemiological data, and the potential bias in these models; (4) the broad tasks that LLMs have been used for in epidemiology and what performance have LLMs achieved on these tasks.</p> <p><b>Objective 2:</b> Determine the extent that LLMs can be used to summarise or extract information from Twitter data. The student will investigate using LLMs to derive topic information from tweets (e.g. example topics may be politics, or social life), and output can be compared to other topic analysis approaches. The student will also explore using LLMs to extract sentiment or mood from tweets, and this can be compared with existing sentiment analysis approaches. The student will be able to suggest other health-relevant information that may be inferred from tweets and explore approaches to do this. This objective will use open source LLMs that can be downloaded and run on the University's compute services, as they will be applied to sensitive ALSPAC data in Objective 3, that cannot be transferred to external services. <b>Objective 3:</b> Explore the association of the LLM derived data with mental health traits in ALSPAC. The student will set up an analytical pipeline that uses</p>

	a LLM to derive phenotypes from Twitter data available in 750 ALSPAC participants. This will use the approaches developed as part of Objective 2 to derive phenotypes and explore the relationship of these phenotypes with mental health traits, such as anxiety and depression. The specific mental health traits to be explored can be chosen depending on the student's interests.
<b>Supervisory Team</b>	
<b>Lead Supervisor</b>	
Name	Dr Louise Millard
Affiliation	Bristol
College/Faculty	Faculty of Health Sciences
Department/School	Bristol Medical School
Email Address	louise.millard@bristol.ac.uk
<b>Co-Supervisor 1</b>	
Name	Professor Tom Gaunt
Affiliation	Bristol
College/Faculty	Faculty of Health Sciences
Department/School	Bristol Medical School
<b>Co-Supervisor 2</b>	
Name	Dr Oliver Davis
Affiliation	Bristol
College/Faculty	Faculty of Health Sciences
Department/School	Bristol Medical School
<b>Co-Supervisor 3</b>	
Name	Professor Frances Rice
Affiliation	Cardiff
College/Faculty	College of Biomedical and Life Sciences
Department/School	School of Medicine