Project Details	
Project Code	MRCPHS25Br Liu
Title	Automating knowledge synthesis in biomedical literature using AI and Large Language Models
Research Theme	Population Health Sciences
Summary	There are millions of new scientific findings published every year, and the evidence they contain vary by their study designs and analysis methods (e.g. randomised trials, Mendelian randomisation studies, observational studies, etc). The student will use artificial intelligence (AI) methods such as Large Language Models (LLMs, e.g. LLaMA and GPT-4) for knowledge synthesis, i.e. to extract and summarise from large amounts of heterogenous information to generate broader perspectives and new insights, to be able to answer questions such as the effect of obesity on breast cancer as identified by the vast volumes of literature.
Description	Background: Knowledge synthesis can be a slow and cumbersome process but is an essential tool for medical and public health policy-makers. Formal systematic reviews require rigid protocols and extensive human effort from trained professionals, whereas a massive volume of research evidence emerges every year. Recent advances in computational approaches in Artificial Intelligence (AI) and Natural Language Processing (NLP), such as Named Entity Recognition and text summarization powered by Large Language Models (LLMs, such as BERT, LLaMA, or the GPT model series), have substantially improved the efficiency and accuracy of information extraction at a massive scale. These AI and NLP methods could potentially offer the opportunity to complement manual systematic reviews with automated knowledge synthesis reports that are updated in real-time as new literature emerges, and therefore improve the efficiency and responsiveness of decision-making of systematic reviewers, public health professionals and clinicians. This project aims to substantially improve the rapid review of literature by developing methods to automate knowledge synthesis from published research articles. The student will explore use of Large Language Models and NLP methods to automate: 1) the identification and extraction of key research information, and 2) the assessment of research quality and risk of bias from a published or pre-print article. The student will evaluate the developed automation pipeline for case studies such as on the effect of dietary and lifestyle factors on the likelihood of cancers and other diseases. Objective 1: To investigate the performance of various automated approaches for text summarization. The student will evaluate and compare several NLP methods including rule-based approaches, small-scale LLMs (i.e. models such as Llama-3-8B with small parameter size and can be used in resource constrained compute environments) as well as flagship LLMs (e.g. GPT-4o or Llama-3-30b) for the summarization of literature text as available from research article databases/online sources (e.g. PubMed, and preprint servers such as medRxiv) regarding the overall research objectives, key findings etc. The student will also evaluate the respective functionalities and performance, bias, as well as their potential environmental impact from these approaches. This objective aims to train the student regarding the current widely-used

	methods and underlying biomedical literature data, as well as to further lead to the adequate combination and balance between multiple methods for subsequent objectives. The student will be able to define their own evaluation criteria after consultation with stakeholders and experts. Objective 2: To apply or develop robust methods for the extraction and harmonization of evidence from biomedical literature. To extent from the research work for Obj. 1, the student will apply the various natural language processing methods to identify and extract key information from scientific text, such as the involved risk factors and disease outcomes (e.g. lifestyle and dietary factors, disease names), research methods and study designs (e.g. is it an observational study or randomised trials), as well as quantitative information such as the effect sizes, standard errors from the statistical analysis. In addition, the student develop approaches to harmonize the evidence across heterogenous study types (such as randomised trials, non-randomised interventions, observational studies) to enable comprehensive triangulation of evidence, between these different studies and also with biomedical databases and platforms. The student will have the opportunity to decide on the details of the methods used in information extraction and harmonization. Objective 3: To develop a novel framework to automate the assessment of risk of bias from research articles. Risk of bias refers to the risk that the research findings reported by a study are inaccurate due to limitations of the study, such as selection bias, measurement error or missing information. The student will build on methods developed in Obj. 2 to robustly assess the risk of bias of a study, and automate such assessment in published and pre-print research articles as well as contrast and validate the results from their developed methods with established frameworks (e.g. RobotReviewer). The student will steer the choice of methods, which could involve developing new components or repurposing existing ones. This research project has the potential to create significant impact in the research of systematic reviews and evidence synthesis via the innovation in Artificial Intelligence, Natural Language Processing and Large Language Model technologies. The student will be supervised by an interdisciplinary team from data mining, computer science and public health backgrounds in the Data Mining Epidemiological Relationships Programme at MRC Integrative Epidemiology Unit, University of Bristol and the CardiffNLP Lab at Department of Computer Science in Cardiff University. The student will also have the opportunity to have industry placement at AMPLIFY to translate their skills in machine learning and data science to address business problems.
Supervisory Team	
Lead Supervisor	
Name	Dr Yi Liu
Affiliation	Bristol
College/Faculty	Faculty of Health and Life Sciences
Department/School	Population Health Sciences, Bristol Medical School
Email Address	yi6240.liu@bristol.ac.uk

Co-Supervisor 1	
Name	Dr Louise Millard
Affiliation	Bristol
College/Faculty	Faculty of Health and Life Sciences
Department/School	Population Health Sciences, Bristol Medical School
Co-Supervisor 2	
Name	Dr Luis Espinosa-Anke
Affiliation	Cardiff
College/Faculty	College of Physical Sciences and Engineering
Department/School	School of Computer Science and Informatics
Co-Supervisor 3	
Name	Professor Tom Gaunt
Affiliation	Bristol
College/Faculty	Faculty of Health and Life Sciences
Department/School	Population Health Sciences, Bristol Medical School