LLM Utility For Selection Of Health Policy Papers In Systematic Reviews

Terms of Reference

Samuel Bickel-Barlow, Krisha Chandnani, Cyril Estier, Juan Gomez-Illingworth, Sofie Wiedemeyer

London School of Economics & UK Health Security Agency
October 2025 - March 2026

Goal

The objective of this project is to analyse the effectiveness of automating the selection process for systematic reviews in the field of health policy by using LLMs. The analysis will optimise, measure, and report the accuracy of LLMs in accomplishing this task within an automated systematic review pipeline.

Background

The UK Health Security Agency (UKHSA) (the client) is interested in evaluating the accuracy of large language models (LLMs) in screening paper abstracts for inclusion in systematic reviews of health policy topics. Systematic reviews attempt to identify, appraise and synthesise evidence to answer specific research questions. This makes them a valuable tool for UKHSA policy analysts seeking a complete and unbiased review of relevant evidence to help answer health policy questions. These reviews are normally created and updated by volunteer researchers, who follow detailed protocols to increase consistency and reduce bias. Volunteers must identify and evaluate hundreds of potentially relevant papers for inclusion before even commencing a systematic review. This labour-intensive process could potentially be automated by LLMs, whose ability to intelligently handle text data could make them ideally suited to this task, if they can make decisions about paper inclusion at least as accurately as human reviewers. This application of LLMs would be hugely valuable as it greatly reduces the time and effort required to create systematic reviews, a process that currently takes a year or more. Partial automation using LLMs would allow volunteer researchers to produce more reviews covering a greater number of research questions. UKHSA experts would therefore have a larger corpus of questions with systematically reviewed research to draw from when formulating health policy.

Project Stages

The LSE team will evaluate the accuracy of several open source LLMs in screening abstracts of papers for inclusion in systematic reviews of health policy topics (the project). In order to achieve this, it will:

i) Research systematic reviews and review the application of LLMs to similar questions.
ii) Collect and prepare data from Cochrane reviews and relevant papers.
iii) Develop a data processing pipeline in Python, which calls open-source LLMs to select papers for inclusion in systematic reviews providing all necessary context.
iv) Explore prompt optimisation techniques and test extensions like agents, fine-tuning, and automated refinement.
v) Assess the accuracy of various models and techniques using Cochrane reviews as a benchmark, employing metrics including precision, recall, F1 score, agreement rates, and Cohen's Kappa.
vi) Conduct error analysis to identify patterns in model failures and provide actionable recommendations.

vii) Produce a detailed report of our findings, a comprehensive GitHub repository with documentation, and an annotated validation dataset for future UKHSA evaluations.

Roles & Responsibilities

LSE Team

The team of MPA in Data Science for Public Policy (DSPP) candidates at the London School of Economics and Political Science (LSE) is as follows:

Name	Email Address
Sam Bickel-Barlow	s.c.bickel-barlow@lse.ac.uk
Krisha Chandnani	k.chandnani@lse.ac.uk
Cyril Estier	c.m.estier@lse.ac.uk
Juan Gomez-Illingworth	j.gomez-illingworth@lse.ac.uk
Sofie Wiedemeyer	s.wiedemeyer@lse.ac.uk

The MPA-DSPP candidates listed above will plan and execute all tasks and produce all deliverables outlined in the ToR by the due date.

The LSE Supervisor for this project is Dr Ryan Hübert (r.hubert@lse.ac.uk). Dr Hübert will guide and advise the team, but will not be directly involved in project activities.

UKHSA Team

The UKHSA team is as follows:

Name	Email Address
Becca Dikuyi	becca.dikuyi@ukhsa.gov.uk
Ollie Higgins	ollie.higgins@ukhsa.gov.uk
Toby Nonnenmacher	toby.nonnenmacher@ukhsa.gov.uk

The UKHSA team will set project objectives and expectations. The LSE team will schedule meetings with the UKHSA team to update on progress and ask for clarification or feedback when needed.

Communication & Meetings

The LSE team will maintain regular communication with the UKHSA team throughout the project duration. A preliminary meeting schedule includes:

- **Kick-off meeting** (04 November 2025): Project initiation, clarification of objectives and expectations.
- **Progress update meetings** (approximately monthly, November 2025 - February 2026): Updates on data collection, model development, and preliminary findings.
- **Mid-project check-in** (January 2026): Review of initial results and adjust approach if needed.

- **Pre-final review** (early March 2026): Present draft findings for feedback.
- **Final presentation** (late March 2026): Present completed deliverables.

Additional ad-hoc meetings may be scheduled as needed for clarification or urgent matters. The LSE team will provide written progress updates via email between scheduled meetings.

Data Sources

Cochrane Database of Systematic Reviews

Cochrane Library is one of the primary organisations engaged in the creation of systematic reviews. Their database includes thousands of high-quality systematic reviews on many health-related topics. The large number of high-quality systematic reviews makes the Cochrane Database of Systematic Reviews a good candidate for a source of "ground truth" in our evaluation of LLMs' application to the creation of these reviews.

PubMed Database

PubMed is a database maintained by the U.S. National Library of Medicine. It constitutes an index of biomedical literature, compiling citations and abstracts from journals and databases, e.g. the Cochrane Database of Systematic Reviews (CDSR). Even though PubMed does not provide the full reviews (which are available in the Cochrane Library published by Wiley), it appears to include the information we need for this project, and there are fewer restrictions concerning how the data can be processed.

Scope & Limitations

Project Scope

The scope of the team's analysis will focus on the evaluation of LLMs for screening paper abstracts for inclusion of papers in systematic reviews of health policy topics. The team will not train, test, and evaluate LLMs' capacity to generate systematic reviews in their entirety.

The team will only use Cochrane systematic reviews with a clear set of inclusion and exclusion criteria to create an annotated validation dataset unless otherwise directed by the client. The team anticipates using approximately X systematic reviews to create a robust validation dataset.

Technical Constraints

In conducting data collection and analysis, the team will only use financial resources provided by the LSE, including for data subscriptions, tokens for LLM calls, and processing power. If this limitation creates significant barriers to the analysis, the team will flag them as early as possible with the client.

While the team will evaluate multiple open-source models, rapidly evolving LLM capabilities mean that newer models may emerge during the project timeline. The team will prioritise models that are stable, well-documented, and likely to remain accessible to UKHSA post-project.

Analysis Limitations

The team will attempt to establish human baseline performance where possible, but acknowledges that perfect ground truth may not exist given variability in human reviewer decisions. Results from this analysis will be based on Cochrane systematic reviews in health policy. Performance may vary when applied to other domains or review methodologies.

Research Ethics & Data Governance

The team will adhere to the following ethical and data governance principles:

Data Access & Permissions

The team has access to the Cochrane systematic reviews' abstracts and citations through PubMed's API. The information will be obtained and processed in accordance with the corresponding terms and conditions.

Reproducibility & Transparency

All code and methodologies will be documented in the GitHub repository to ensure transparency and enable independent verification of results.

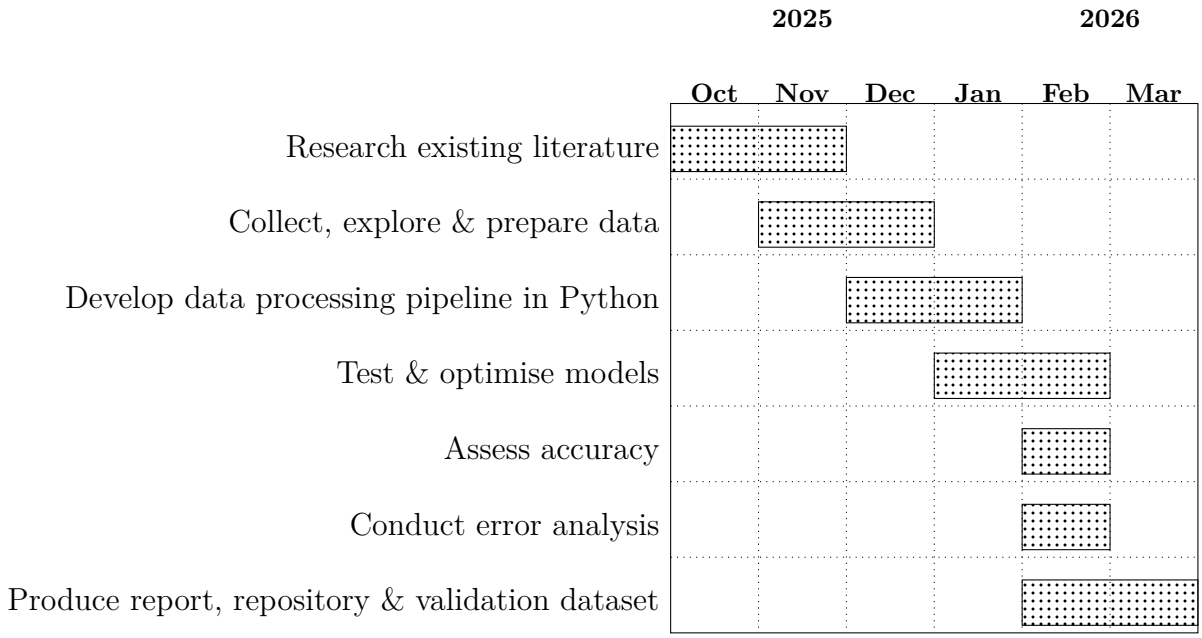
Deliverables

At the end of the project, the LSE team will deliver:

1. A comprehensive GitHub repository containing all project code, documentation, and outputs, providing UKHSA with a reproducible and transparent record of the methodology. The repository will include:
   (a) All analysis scripts and model implementation code;
   (b) Data processing pipelines;
   (c) Documentation for setup and usage; and,
   (d) The detailed capstone project PDF report.
2. A detailed capstone project PDF report of findings by 26/03/2026, which will compare the accuracy of various LLM models and techniques in the inclusion of articles for systematic reviews and final recommendations.
3. An annotated validation data set, which will include systematic reviews and corresponding true and predicted included and excluded papers.
4. A slide deck for presenting key findings and conclusions to the client.

Timeline

The project stages will be executed as follows:



The timeline includes buffer periods to accommodate unforeseen technical challenges, data access issues, or the need for additional model iterations based on client feedback.

Key Terms

Agents – autonomous or semi-autonomous systems built on top of LLMs that can plan, reason, and take actions (such as calling tools, browsing data, or writing code) to accomplish user-defined goals without constant human input.

Agreement Rates – The percentage of times that two models, raters, or processes produce the same results or outcomes.

$$\text{Agreement Rate} = \frac{\text{Number of Agreements}}{\text{Total Cases}}$$

Annotated Validation Dataset – a curated set of examples labelled or commented on (often by humans) that is used to test and evaluate how well a model performs on known, verified data without influencing the model's training process.

Automated Refinement – a feedback-driven process where model outputs are iteratively improved using evaluation metrics, human feedback, or other automated systems to enhance performance and reliability over time.

Cohen's Kappa - a statistic that gauges agreement among raters for categorical items; it adjusts for agreement by chance.

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where:

- p_o is the observed agreement proportion among raters.
- p_e is the expected chance agreement proportion.

Fine-tuning – the process of further training a pre-trained LLM on a smaller, specialised dataset to adapt it for a specific task, domain, or tone, while retaining the broader knowledge learned during initial training.

F1 Score – The harmonic mean of the precision and recall metrics, i.e., it penalises considerable imbalances.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Large Language Model (LLM) - a type of artificial intelligence that uses deep learning techniques to understand and generate human language.

Precision – The percentage of correctly predicted positives.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Prompt Engineering/Optimisation – the practice of designing, refining, and testing inputs (prompts) to an LLM to achieve more accurate, relevant, or efficient responses. It involves understanding how model behaviour changes with different instructions or structures.

Recall – The percentage of real positives that were identified correctly.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Systematic Review – an academic procedure that attempts to create a comprehensive and systematic overview by using clearly defined methods of research to answer a particular

question.

Training – the process by which a machine learning model learns patterns from data. During training, the model adjusts its internal parameters to minimise errors when predicting or generating outputs based on input data.