Reflections on evaluating the utility of an LLM for keywording health research

Claire Stansfield and Ailbhe Finnerty Mutlu
Evidence for Policy and Practice Information and Co-ordinating (EPPI) Centre,
UCL Social Research Institute, University College London (UCL), United Kingdom

Abstract
Generative artificial intelligence shows promise for rapidly annotating collections of research at scale. We reflected on our experiences of evaluating the ability of a Large Language Model (LLM) to apply predefined keywords to records of health research in the context of an evidence repository on vaccine research and research registers of health promotion effectiveness. Five aspects of evaluation helped us articulate key considerations and challenges across the use cases: 1) cyclical prompt development, 2) data availability and quality, 3) performance benchmarks and expectations, 4) task complexity and perspective and 5) workflows and tools.

Key words: *Large Language Models; databases as topic; evaluation studies as topic.*

Introduction

Our context is specialist evidence repositories and research registers that are compiled for purposes of research discovery, such as investigating the research landscape and informing evidence synthesis. They may contain hundreds or thousands of research records, and be annotated with predefined keywords (also called labels, codes) describing broad characteristics of the research, such as study design, topic or thematic focus, or detailed aspects such as who delivered an intervention, country it was delivered, dosage etc. These keywords are often bespoke to the context of the repository and might be assigned from title and abstracts or full-texts. Using an LLM to apply keywording is particularly appealing for achieving rapid keywording at scale in regularly updated (or "living") systems. However, there are challenges in evaluating acceptable performance for accuracy and reliability with a view to implementation. We have separate experiences of using this approach for two use cases and highlight our key reflections here. For a broader perspective, the living documents produced within Responsible Use of AI in Evidence Synthesis (RAISE) is an important reference point. RAISE 2 focuses building and evaluating AI evidence synthesis tools (1).

We begin by outlining the use cases and the approach taken to apply an LLM, followed by reflections across five aspects of evaluation: 1) cyclical prompt development, 2) data availability and quality, 3) performance benchmarks and expectations, 4) task complexity and perspective and 5) workflows and tools.

Context

One author (AFM) is evaluating the utility of an LLM to apply keywords based on full text, to a living evidence repository of Human Papillomavirus (HPV) vaccine research (2, 3). The repository contains research articles on HPV vaccine delivery in low and middle income countries and forms part of an evidence bank that will be updated through regular searches. The bibliographic records have been keyworded according to a taxonomy of HPV-relevant terms. These keywords enable use of the repository to identify relevant research and research gaps to inform research and practice. The first iteration of the HPV taxonomy had around 230 keywords, including 127 individual country names, these keywords ranged from thematic focus, delivery of vaccines to descriptions of the populations of interest. The taxonomy is being further developed by a team of experts in the HPV vaccine community

Address for correspondence: Claire Stansfield, Evidence for Policy and Practice Information and Co-ordinating (EPPI) Centre, UCL Social Research Institute, University College London (UCL), London, United Kingdom.
E-mail: c.stansfield@ucl.ac.uk.

Journal of EAHIL 2026; Vol. 22 (2): 18-23 DOI: 10.32384/jeahil22724

to add more detailed keywords. The initial keywording was done by single human reviewers and pairs of humans independently keyworded a percentage of records (10-20%) that was done by an LLM.

Another author (CS) is evaluating keywords based on titles and abstracts in two registers of health promotion effectiveness, Trials Register of Promoting Health Interventions (TRoPHI) and Database of promoting health effectiveness reviews (DoPHER) (4). The TRoPHI register contains over 25,000 records of research using controlled trials, and descriptive keywords to describe the content related to thematic focus, study design, population groups, and geographical region. Thematic focus forms the largest keyword set comprising 32 keywords on particular areas for health promotion (e.g. injury, mental health, cancer). DoPHER contains over 10,000 reviews of effectiveness and currently has no keywording in the public version. We previously described our planned steps for full automation (5).

Each use case had a bespoke set of keywords that were defined (prior to automation) to support consistent and reliable annotation of multiple keywords to a research record by humans. Each keyword was converted into an LLM prompt in the form of a command or question to guide the LLM to produce a desired action or answer (6). The cases here focus on prompts requiring the LLM to provide a true or false response, so that if an answer is true, the keyword is annotated to the record about that research. The prompts generally followed a similar structure, detailing the role of the LLM and the criteria to be met. One prompt was used per

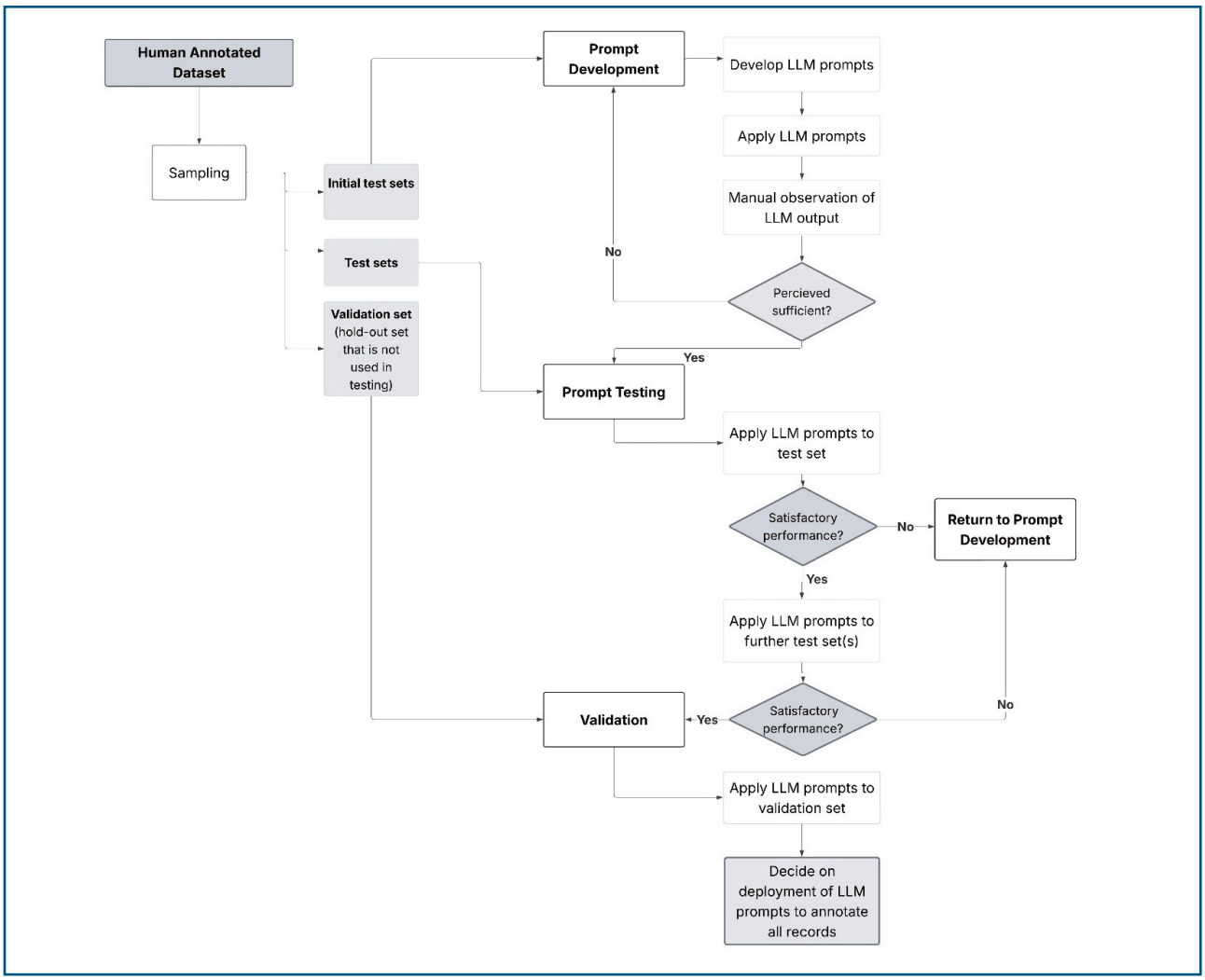

Fig. 1. *The cyclical process of prompt development and evaluation.*

keyword. For example, in TRoPHI, over 90 question prompts are applied to one title and abstract record.
Mostly, our evaluations focused on the reliability of an individual keyword being annotated to a single research record (title/abstract or full-text article), rather than the collective set of keywords annotated to a single record. We used the GPT 4.1 LLM within EPPI-Reviewer, using parameters fixed within this interface (7, 8), as this interface is developed within our research centre. Our reflections here relate to evaluations generally rather than on the performance of a specific LLM model.

Cyclical prompt development and evaluation

In both cases, we began by iteratively writing prompts and comparing performance of the LLM with prior human annotations on relatively small samples of records. Most records had multiple keywords assigned to them, it was only possible to manually check reasons for discrepancies over relatively small samples, as it involved considerable time checking the text that the keywords were applied to (either full text or title/abstract). Once the prompts were perceived as sufficient from this first step, they were evaluated on further samples to obtain measures of reliability compared with humans, followed by further investigations, and cycles of revision and testing of prompts. Strategies to improve prompts included: sense checking with another human, asking an LLM to edit prompts, and trial and error observations of what seemed to work on small samples. An overview of the process is shown in *Figure 1.*
Repeated cycles of prompt development and testing proved challenging as it required having a representative sample of the keywords. It was important to evaluate new prompts on new, previously unseen, sample records that had not informed prompt refinements and these samples needed to represent all the keywords being tested. Ideally these samples should be sufficiently large but depended on the data or the resources available to manually annotate new data.
An ongoing challenge is deciding how much to iterate prompts or when to stop. Identifying content within a sample can be useful to improve a prompt (e.g. setting out certain ways authors describe an intervention or population) but care is needed to avoid being too specific in order to anticipate other relevant scenarios that may apply to that keyword. Stopping requires accepting limitations of the possibility of the prompt at the current time or abandoning prompts that perform less well. Another aspect is that performance differs between LLM models and if tested, can add further cycles within the prompt development and evaluation process.

Data quality and availability

Ensuring there is sufficient volume of good quality data of records with keyword annotations with which to compare to the LLM generated data is important. However, this is not always available for various reasons. For example, having all data annotated independently by multiple humans is ideal, though requires resources to generate. If this data is not available there should be at least a proportion of records that are annotated by more than one human, and the reliability of these annotations between humans should be calculated. There also needs to be sufficient representation of all the keywords in the data including those that are rarely used. Producing truly random samples of records is not always possible due to factors such as small sample sizes or uneven distributions of keywords across samples.
In the case of TRoPHI, random and stratified samples of dual keywording by humans were generated (using approximately 700 records), and pairwise comparisons of humans and LLM data were made. However, when 200 records from this set became the only validation samples (that had not been used to inform prompt refinement), some keywords were underrepresented owing the smaller size of set. These sets were compared with larger samples of single human and LLM data (approximately 3,000) to observe performance at a larger scale. However, we recognise that these volumes of data are not routinely available, including our other use cases.
Furthermore, it is recognised that there are often inconsistencies between humans when applying multiple keywords for this type of task (9, 10). The quality and structure of the text itself can vary. The interpretation of text and interpretation of the keywords can vary between humans. Humans may also have varying level of expertise with the keywording task. There are studies which have identified that the human is influenced by an AI decision if they are presented with AI judgements for the same task (9). Sometimes the human an-

notators may also have used text beyond that being used by the LLM, such as a journal name or author keywords. There may also be inconsistencies applied by the same human over time, or changes in interpretation of how keywords are applied over time, or human fatigue may be another factor in identifying all keywords. For example, in the HPV repository, when there were many relevant keywords relating to a full-text document, humans missed one or two which were identified by the LLM.

Performance benchmarks and expectations

RAISE 2 (1) includes a taxonomy of common performance metrics in evidence synthesis. In our use cases, the benchmarks used to evaluate performance of LLM needed to reflect variability in the accuracy and consistency in the comparison data that was available. Inter-rater reliability (IRR) was compared to evaluate the reliability of keywording instructions, rather than directly comparing the accuracy of the LLM with human data. This is consistent with the domains of subject classification (10), content analysis (12), where pre-existing human annotations of multiple keywords are not a sufficient gold standard. We used IRR to assess the performance of records keyworded by human and the LLM, or where available, between two humans and two LLMs. The IRR is calculated from the rates of agreement and disagreement of the application of each keyword and evaluates whether different observers of the same records yield the same data within accepted levels of error (12). An advantage is that this allows comparison of consistency, and comparison over large sets of records. Disadvantages are that it might not detect consistent errors unless checking individual instances of disagreement or it does not work well for 'rarely' used keywords that are not sufficiently represented in evaluation. IRR metrics vary on the level to which they adjust for randomness in their agreements and disagreements, and the number of people or entities the IRR is being measured between (11).

Across our evaluations Krippendorff's alpha and Fleiss Kappa were used based on pair-wise comparisons (e.g. human-human, human-LLM, LLM-LLM). Evaluating IRR provided an overview of performance at scale, and comparison across different pairs of annotators. However, we are aware that sample size of some (under) represented keywords could affect some results. Furthermore, this metric alone does not provide a complete picture, as it does not highlight consistent omissions or show nuances of how keywords are applied. We found it useful to review how keywording was applied in individual cases, though such investigations were targeted owing to the time involved. A further challenge is the level of performance to accept. Acceptable performance of annotation tasks is ill-defined, and expectations may vary regarding use cases. The type of task and the real-world implications inform the level of acceptable reliability and tasks which can have a more serious consequence require a higher level of reliability (11). For example, if the evidence that is aligned to a set of keywords is intended to be relied upon to inform policy and decision making (without using further approaches for discovery), then a high standard is required as the consequences are a hindrance to discovery and there is potential for altering the findings of a systematic review (this is highlighted in RAISE 2 (1)). On the other hand, if the use case is more for exploratory discovery, rapid mapping, or if the limitations of keywording are communicated then performance standards are more arbitrary. For example, if keywording of a record acts as a signpost it is less critical if it is liberally applied (as researchers can filter out records manually following their own checks), and also having other options available (such as free-text searches) to discover useful research not keyworded.

Task complexity and perspective

The complexity of the task is another consideration. Identifying records which can be considered edge cases of records, within the register or repository and where the keywording encompasses a complex area, can be helpful to incorporate as examples of edge cases into training material given to the human annotators and the LLM prompts in order to improve reliability across a wide variety of records. However, there are some concepts which are complicated for an LLM to infer from a piece of text and refinement of prompts might not sufficiently improve performance. We found this in some cases where human agreement was low (such as assignment to certain topic areas), which indicates a high level of task difficulty to infer from text. Furthermore, humans and LLMs address tasks in different ways and types of difficulties differ. For some keywords, we found the LLM prompt needed to have clearer instructions, and contain more context about a

keyword, compared with human. Having a keyword category for 'unclear' was useful to isolate records that the LLM had not been able to annotate.

We also needed to be mindful of being influenced in favour of wanting an LLM to perform well in order to improve efficiencies, and mitigating this with appropriate testing. We observed when using information from the title/abstract that some keywords were more suited to describe primary research than reviews, as the abstract of the primary research contained specific information on a population or context compared with a systematic review covering multiple populations or contexts. This highlighted the importance of testing prompts that we wanted to re-use within their context of use.

Workflows and tools

A key facilitator to evaluation is having efficient workflows and tools to conduct the evaluations. It was important to have a process to manage different versions of the test LLM prompts, collate decisions of the human and LLM model about each keyword for each research record, and calculate performance metrics for each sample. It took time to develop suitable templates and processes; for example, converting outputs of LLM and human decisions, and calculating inter-rater reliability were done through developing bespoke Excel templates for each project. However, we are aware of a shiny app being developed that would improve this process. As we progressed more tools have become available to support evaluations within the interface we used, such as a tool to compare the performance of different LLMs, and multiple iterations of LLMs, across small samples (13). Other development underway includes an opensource data extraction and evaluation toolkit that will allow easy iteration of instructions for keywording and data extraction as well as their evaluation through comparison of human annotated data and a human-in-the-loop evaluation process.

Conclusion

Implementing LLM predefined keywording currently requires considerable development and testing time and may be an efficient method for automating some types of annotations for evidence repositories and research registers at scale. The five areas reflected upon here helped us articulate some considerations and challenges of evaluation and implementation. It also provides a structure for further conversations beyond individual use cases.

Funding

AFM has funding from Wellcome Trust (313586/Z/24/Z) and CS has funding from National Institute for Health and Care Research Policy Review Facility (NIHR200701).

Acknowledgements

A number of colleagues from EPPI Centre, UCL and from the London-York PRP Evidence Review Facility have participated in discussions that have informed this work.

Submitted on invitation.
Accepted on 11 May 2026.

REFERENCES

1. Thomas J, Hair K, Noel-Storr, A. et al. Responsible use of AI in evidence SynthEsis (RAISE 2026): building and evaluating AI evidence synthesis tools (version 3; updated 13 March 2026). In: Open Science Framework [https://osf.io/fwaud/files/wbf4y], Washington DC: Center for Open Science. DOI 10.17605/OSF.IO/FWAUD
2. Eni TE, Bond M, Finnerty Mutlu A, Tifuh OB, Wehtuogenyi TR, Leopold L, Lontum A, Okwen TT, Mbinkar BN, Cholong BJ, Okwen PM, Yong NB, Jeppesen BT. Living evidence repository and evidence maps on human papillomavirus vaccines delivery. 2025. EPPI Visualiser database. https://eppi.ioe.ac.uk/eppi-vis/login/open?webdbid=923
3. HPV living evidence partnership. Available from: https://aliveevidence.org/hpv-living-evidence/
4. EPPI Centre. Databases. Available from: https://eppi.ioe.ac.uk/cms/databases
5. Stansfield C, Thomas J. Applying automation to maintain research registers in health promotion. J Eur Assoc Health Info Libr. 2024;20(2):26-9.

6. Homiar A, Thomas J, Ostinelli EG, Kennett J, Friedrich C, Cuijpers P, et al. Development and evaluation of prompts for a large language model to screen titles and abstracts in a living systematic review. BMJ Mental Health. 2025;28:e301762.
7. Thomas J, Graziosi S, Brunton J, Ghouze Z, O'-Driscoll P, Bond M, Koryakina A. EPPI-Reviewer: advanced software for systematic reviews, maps and evidence synthesis. EPPI Centre, UCL Social Research Institute, University College London; 2023.
8. EPPI-Reviewer. Automated screening and data extraction using Large Language Models. Available from: https://eppi.ioe.ac.uk/cms/er4/Help/Automated-coding-using-LLMs
9. Gartlehner G, Kahwati L, Nussbaumer-Streit B, et al. From promise to practice: challenges and pitfalls in the evaluation of large language model for data extraction in evidence synthesis. BMJ Evidence-Based Medicine. 2025;30:385-9.
10. Golub K, Soergel D, Buchanan G, Tudhope D, Lykke M, Hiom D. A framework for evaluating automatic indexing or classification in the context of retrieval. Journal of the Association for Information Science and Technology. 2016;67(1):3-16.
11. Krippendorff K. Reliability. In: Content analysis: An introduction to its methodology. (4. ed.). SAGE Publications, Inc; 2019. 277-360.
12. Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. Communication Methods and Measures. 2007;1(1):77-89.
doi: 10.1080/19312450709336664
13. EPPI-Reviewer. Latest changes (16/04/2026 - V6.18.0.0).
Available from:
https://eppi.ioe.ac.uk/cms/er4/Help/Version-History-Announcements/Latest-Changes-16-04-2026-V61800