# About this map

## Overview

Generative LLM-based tools are artificial intelligence (AI) tools capable of generating natural language text, images, audio, or other media in response to user- or self-inputs known as ‘prompts’. For further background information, please see the published [protocol](https://eppi.ioe.ac.uk/CMS/Portals/0/Generative%20LLM-Based%20Tools%20for%20Health%20and%20Social%20Care%20Applications_Protocol_100624-Final.pdf) that covers this living map (Shemilt 2024).

**This living map of research contains bibliographic records of primary studies and systematic reviews (including protocols for such studies) that evaluate the performance of generative large language model-based (LLM-based) tools for health or social care applications.**

The ['main map'](https://eppi.ioe.ac.uk/eppi-vis/login/open?webdbid=708) contains records of studies that include a substantive focus on clinical health care, public health, or social care applications. Each record included in the 'main map' has been classified (coded) according to the following sets of study characteristics (dimensions)[[1]](#footnote-1):

* Application class(es)
* Article type
* Model(s)
* Mode(s) of model(s) use
* Type(s) of task(s)
* Health and social care categories
* Population(s)
* Version
* Coding status[[2]](#footnote-2)

**The current version of the** [**‘main map’**](https://eppi.ioe.ac.uk/eppi-vis/login/open?webdbid=708) **is ‘Version 5 – 28th March 2024’ (published on 30th June 2025). The current version of the ‘main map’ contains 498 fully coded reports (articles) of which 89 are ‘new’ in this version**.

Alongside the 'main map', we also publish a separate ['sector'](https://eppi.ioe.ac.uk/eppi-vis/login/open?webdbid=714) of the living map containing further eligible records (identified and coded using the same procedures but withheld from publication in the 'main map'[[3]](#endnote-1)) which are ***only*** about specific classes of applications: ‘medical education, health or social care professional development and training’, ‘answering medical (and related) exam questions/ answering medical questions in general’ and/ or ‘answering patient or service user questions/ patient or service user education’[[4]](#endnote-2). Records included in the separate ‘sector’ only have minimal coding: ‘application classes’, ‘article type’ and ‘version’. **The current version of the separate** [**‘sector’**](https://eppi.ioe.ac.uk/eppi-vis/login/open?webdbid=714) **is also ‘Version 5 – 28th March 2024’ (also published on 30th June 2025). The ‘sector’ currently contains 1,241 minimally coded reports of which 355 are ‘new’ in this version**.

## How to cite this map

Shemilt I, Hollands GJ, Khouja C, Raine G, Thomas J. (2024). *Generative large language models for health and social care applications: a living map of research*. London: EPPI Centre.

The published protocol that covers this map is available [here](https://eppi.ioe.ac.uk/CMS/Portals/0/Generative%20LLM-Based%20Tools%20for%20Health%20and%20Social%20Care%20Applications_Protocol_100624-Final.pdf).

## Objectives of the living map

1. To maintain a continual surveillance of the landscape of accumulating research evidence for the use of generative LLM-based tools for health and social care applications;
2. To provide a regularly updated descriptive overview of the landscape of cumulative research evidence, and gaps in the evidence base, for the use of generative LLM-based tools for health and social care applications, classified in terms of its key features and characteristics;
3. To make cumulative research evidence for the use of generative LLM-based tools for health and social care applications more findable, accessible and reusable; and
4. To compile a glossary of key terms and concepts relating to generative LLMs and generative LLM-based tools for health and social care applications (forthcoming).

## Results

### Version 5 – 28th March 2025 (current version)

Between 'Version 4' and ‘Version 5’ of this map (28th March 2025; published 30th June 2025) we have screened a further 288 records and corresponding full-texts of articles (reports) identified from the OpenAlex dataset (up to 6th June 2025) and other sources. None of these records were previously unidentified duplicates. 14 of these records were excluded, while 3 records were referred for a ‘second opinion’, and a further 271 eligible reports were identified. A further 444 records from the OpenAlex dataset (up to 28th March 2025) and other sources were manually or automatically coded and added to the map (‘new’ in ‘Version 5’) and a further 3 records were identified as duplicates during coding. This leaves an updated total of 177 reports awaiting further assessment or coding (-213 between 'Version 4' and ‘Version 5’).

### Of the 444 coded reports first published in ‘Version 5’, 89 have the code(s) ‘clinical health care’, ‘public health’ and/ or ‘social care’ assigned to them for the ‘application classes’ dimension with or without also having the code(s) ‘Medical education, health or social care professional development and training’, ‘Answering medical (and related) exam questions/ answering medical questions in general’ and/ or ‘Answering patient or service user questions / patient or service user education’ assigned to them. Records of these 89 reports are published in ‘Version 5' of the ‘main map’ (web database). The current version of the ‘main map’ therefore includes an updated cumulative total of 498 eligible reports of primary studies or systematic reviews (or protocols for such studies) in ‘All versions’ up to ‘Version 5’.

The other 355 of 444 coded reports first published in ‘Version 5’ have at least one of the code(s) ‘Medical education, health or social care professional development and training’, ‘Answering medical (and related) exam questions/ answering medical questions in general’ and/ or ‘Answering patient or service user questions / patient or service user education’ assigned to them for the ‘application classes’ dimension without also having any of the code(s) ‘clinical health care’, ‘public health’ and/ or ‘social care’ assigned to them. Records of these 355 reports are first published in ‘Version 5’ of the separate ‘sector’ of the living map (web database). The current version of this separate ‘sector’ of the map therefore includes an updated cumulative total of 1,241 eligible reports in ‘All versions’ up to ‘Version 5’.

### Version 4 – 10th December 2024

### Between 'Version 3' and ‘Version 4’, we did not screen any further records and corresponding full-texts of articles (reports) identified from the OpenAlex dataset and other sources[[5]](#footnote-3). As such, ‘Version 4’ of this map (published 30th April 2025) contained records identified from the OpenAlex dataset and other sources up to 10th December 2024. Between 'Version 3' and ‘Version 4’, a further 605 eligible reports of primary studies or systematic reviews (or protocols for such studies) were manually or automatically coded and added to the map (‘new’ in ‘Version 4’); while an updated total of 390 reports were awaiting further assessment or coding (-646 between 'Version 3' and ‘Version 4’).

### Of the 605 coded reports first published in ‘Version 4’, 189 have the code(s) ‘clinical health care’, ‘public health’ and/ or ‘social care’ assigned to them for the ‘application classes’ dimension with or without also having the code(s) ‘Medical education, health or social care professional development and training’, ‘Answering medical (and related) exam questions/ answering medical questions in general’ and/ or ‘Answering patient or service user questions / patient or service user education’ assigned to them. Records of these 189 reports were first published in ‘Version 4' of the ‘main map’ (web database). From ‘Version 4’ onwards, we had also removed 57 records of ‘commentary / review-type articles’ from the ‘main map’[[6]](#footnote-4). ‘Version 4’ of the ‘main map’ therefore included an updated cumulative total of 423 eligible reports of primary studies or systematic reviews (or protocols for such studies) in ‘All versions’ (up to ‘Version 4’).

### The other 416 of 605 coded reports first published in ‘Version 4’ have at least one of the code(s) ‘Medical education, health or social care professional development and training’, ‘Answering medical (and related) exam questions/ answering medical questions in general’ and/ or ‘Answering patient or service user questions / patient or service user education’ assigned to them for the ‘application classes’ dimension without also having any of the code(s) ‘clinical health care’, ‘public health’ and/ or ‘social care’ assigned to them. Records of these 416 reports were first published in ‘Version 4’ of the separate ‘sector’ of the living map (web database). From ‘Version 4’ onwards, we had also removed 10 records of ‘commentary / review-type articles’ from this ‘sector’ of the map3. ‘Version 4’ of the separate ‘sector’ of the map therefore included an updated cumulative total of 886 eligible reports in ‘All versions’ (up to ‘Version 4’).

### Version 3 – 26th November 2024

### Between 'Version 2' and ‘Version 3’ of this map (26th November 2024; published 28th February 2025) we had screened a further 335 records and corresponding full-texts of articles (reports) identified from the OpenAlex dataset (up to 26th November 2024) and other sources. None of these records were previously unidentified duplicates. 76 of these records were excluded while a further 259 eligible reports were identified. A further 95 records were manually coded and added to the map (‘new’ in ‘Version 3’). This left an updated total of 1,036 reports awaiting further assessment or coding (+158 between 'Version 2' and ‘Version 3’).

### Of the 95 coded reports first published in ‘Version 3’, 35 had the code(s) ‘clinical health care’, ‘public health’ and/ or ‘social care’ assigned to them for the ‘application classes’ dimension with or without also having had the code(s) ‘Medical education, health or social care professional development and training’, ‘Answering medical (and related) exam questions/ answering medical questions in general’ and/ or ‘Answering patient or service user questions / patient or service user education’ assigned to them. These 35 records were published in ‘Version 3' of the ‘main map’ (web database), which included a cumulative total of 291 eligible reports in ‘All versions’ up to ‘Version 3’.

### The other 60 of 95 coded reports first published in ‘Version 3’ had at least one of the code(s) ‘Medical education, health or social care professional development and training’, ‘Answering medical (and related) exam questions/ answering medical questions in general’ and/ or ‘Answering patient or service user questions / patient or service user education’ assigned to them for the ‘application classes’ dimension without also having had any of the code(s) ‘clinical health care’, ‘public health’ and/ or ‘social care’ assigned to them. Records of these 60 reports were first published in ‘Version 3’ of the separate ‘sector’ of the living map (web database), which included a cumulative total of 482 eligible reports in ‘All versions’ up to ‘Version 3’.

### Version 2 – 14th October 2024

Between 'Version 1' and ‘Version 2’ of this map (14th October 2024; published 20th December 2024) we had screened a further 413 records and corresponding full-texts of articles (reports) identified from the OpenAlex dataset (up to 14th October 2024) and other sources. None of these records were previously unidentified duplicates. 103 of these records were excluded while a further 310 eligible reports were identified. A further 276 records were manually coded and added to the map, with 2 other records previously awaiting further assessment reviewed, included and coded: a total of 278 ‘new’ reports in ‘Version 2’. This left an updated total of 878 reports awaiting further assessment or coding (-13 between 'Version 1' and ‘Version 2’).

Of the 278 coded reports first published in ‘Version 2’, 110 had the code(s) ‘clinical health care’, ‘public health’ and/ or ‘social care’ assigned to them for the ‘application classes’ dimension with or without also having had the code(s) ‘Medical education, health or social care professional development and training’, ‘Answering medical (and related) exam questions/ answering medical questions in general’ and/ or ‘Answering patient or service user questions / patient or service user education’ assigned to them. These records were published in ‘Version 2' of the ‘main map’ (web database), which included a cumulative total of 256 eligible reports in ‘All versions’ up to ‘Version 2’.

The other 168 of 278 coded reports first published in ‘Version 2’ had at least one of the code(s) ‘Medical education, health or social care professional development and training’, ‘Answering medical (and related) exam questions/ answering medical questions in general’ and/ or ‘Answering patient or service user questions / patient or service user education’ assigned to them for the ‘application classes’ dimension without also having had any of the code(s) ‘clinical health care’, ‘public health’ and/ or ‘social care’ assigned to them. Records of these 168 reports were first published in ‘Version 2’ of the separate ‘sector’ of the living map (web database), which included a cumulative total of 422 eligible reports in ‘All versions’ up to ‘Version 2’.

### Version 1 – 2nd September 2024

Up to 'Version 1' of this map (2nd September 2024; published 18th October 2024) we had screened 2,610 records and corresponding full-texts of articles (reports) identified from the OpenAlex dataset (up to 2nd September 2024) and other sources. 69 records were previously unidentified duplicates, which we discarded. 1,250 records were excluded. 400 eligible reports were manually coded. A further 891 reports were awaiting further assessment or coding.

Of the 400 coded records up to ‘Version 1’, 146 records had the code(s) ‘clinical health care’, ‘public health’ and/ or ‘social care’ assigned to them for the ‘application classes’ dimension with or without also having the code(s) ‘Medical education, health or social care professional development and training’, ‘Answering medical (and related) exam questions/ answering medical questions in general’ and/ or ‘Answering patient or service user questions / patient or service user education’ assigned to them. These records were published in ‘Version 1' of the ‘main map’ (web database), which included a total of 146 eligible reports in ‘All versions’ up to ‘Version 1’.

The other 254 of 400 reports coded up to ‘Version 1’ had at least one of the code(s) ‘Medical education, health or social care professional development and training’, ‘Answering medical (and related) exam questions/ answering medical questions in general’ and/ or ‘Answering patient or service user questions / patient or service user education’ assigned to them for the ‘application classes’ dimension without also having any of the code(s) ‘clinical health care’, ‘public health’ and/ or ‘social care’ assigned to them. Records of these 254 reports were published in ‘Version 1’ of a separate ‘sector’ of the living map (web database), which included a total of 254 eligible reports in ‘All versions’ up to ‘Version 1’.

## Identifying the evidence

To identify eligible articles for this living map, we first conducted a preliminary scoping search of the internet using keyword searches to identify an initial corpus of bibliographic records of 51 potentially eligible, or borderline eligible, articles. These records were imported into [EPPI Reviewer (ER)](https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=2914) software (Thomas 2024), where we used [OpenAlex tools](https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=3754) ‘match records’ features to semi-automatically match 50 of 51 of these records to a corresponding record in the OpenAlex dataset[[7]](#endnote-3).

Second, the matched ‘initial corpus’ OpenAlex records were used to ‘seed’ an initial set of three OpenAlex network graph searches. These OpenAlex searches retrieved records that were connected (on the date of search, 9th January 2004) to ‘seed’ records in the network graph, via either a one-hop (i) forwards (‘cited by’) or (ii) backwards (‘bibliography’) citation relationship, or a one-hop (iii) forwards ‘related publications’ (‘recommended by’) relationship (and published since 1st January 2018). The retrieved records from these searches were also imported into EPPI Reviewer. Third, we also imported into ER selected subsets of bibliographic records already identified for inclusion in a preceding map of research on artificial intelligence and equity (EPPI Centre, 2023).

Following semi-automated de-duplication between and within these sources using ER ‘manage duplicates’ tools, we allocated a random sample of 94 retained, unique records for a first phase of pilot eligibility screening. Pilot eligibility screening was undertaken by 4 researchers working independently, who each applied an earlier version of the screening codes pictured below, with reference to ‘living’ screening guidance notes (see below). Screening was based on title-abstract when possible, or full-text when needed. As such, corresponding full-texts are retrieved for those records whose eligibility cannot be determined based on information in the title-abstract alone, as well as for all eligible records meeting eligibility criteria (if available). The researchers could refer a record for a ‘second opinion’ if they were unsure of its eligibility and referred records (and corresponding full-text articles) were discussed between the four researchers before a final, consensus decision.

Inter-rater agreement between the 4 researchers over 94 records allocated for the first phase of pilot screening was 0.82 (95% CI: 0.74 to 0.90) at the included / excluded level, and 0.63 (95% CI: 0.58 to 0.68) at the specific code level. We therefore undertook a second phase of pilot eligibility screening using a further random sample of 95 records, using the same procedures described above. Inter-rater agreement between the 4 researchers over 95 records allocated for the second phase pilot screening was 0.97 (95% CI: 0.85 to 1.00) at the included / excluded level, and 0.78 (95% CI: 0.70 to 0.85) at the specific code level. Based on these results, we transitioned to a workflow in which one of researchers screened each record but also retained the option of referring records for a ‘second opinion’ discussion if needed.

Next, we developed, tested, refined and executed conventional electronic searches of Embase (Ovid SP) and MEDLINE (Ovid SP) databases. These conventional searches were intentionally configured to prioritise precision over recall, using the search strategies reproduced in Appendix 1. The searches were conducted during January 2024 and the retrieved records were imported into ER and de-duplicated.

Also during January 2024, we implemented auto-update searches of the OpenAlex dataset. These auto-update searches are, like network graph searches (see above), ‘seeded’ by the accumulating set of included records identified by eligibility screening. The ‘seed’ records are subscribed to a novel machine learning ‘recommender’ model (the ‘auto-update model’) (Tenti 2023). This model automatically scores all ‘new’ records that have been added to the latest version of the OpenAlex dataset (updated every ~1 month) and recommends those most likely to be eligible for inclusion in this living EGM. Starting in January 2024, we have imported the top 500 records by auto-update model score from each auto-update search, with import filters applied (see Appendix 2), from each consecutive update of the OpenAlex dataset (i.e. 13 auto-update searches up to Version 4 of this living EGM, the latest from the OpenAlex dataset ‘snapshot’ on 14th March 2025).

From March 2024 onwards, we have also deployed a series of further OpenAlex network graph searches (‘seeded’ by the accumulating set of included records identified by eligibility screening), as well as concurrent OpenAlex custom searches (see below), to coincide with our monthly OpenAlex auto-update searches. We have imported records from each of these three kinds of automated OpenAlex searches into our continual updating workflows every ~2 months. The most recent network graph and custom searches were run over the live OpenAlex dataset on 28th March 2025.

An OpenAlex custom search is similar to a conventional electronic database search, as it combines target sets of ‘title-abstract’ keywords, and OpenAlex ‘concepts’, using Boolean operators. OpenAlex ‘concepts’ are comparable to index terms in conventional electronic research databases (e.g. MeSH terms in MEDLINE), except that they are automatically assigned to each record in the dataset by an OpenAlex machine learning algorithm, and they are not organised into a hierarchical ‘tree’ structure. The OpenAlex custom search strategy we deploy for this living map is reproduced in Appendix 1.

Beyond the pilot phase of eligibility screening (see above), once records from OpenAlex searches have been imported with filters (see Appendix 2) and de-duplicated using the same procedure described above, the retained records are prioritised for manual screening using active learning (Miwa 2014, O’Mara-Eves 2015) (‘priority screening mode’) in EPPI Reviewer. Unique records from each round of OpenAlex continual updating searches are added to the pool of records not yet screened, to be re-prioritised using active learning. In addition, we have trained, calibrated, evaluated two versions of a custom binary machine learning classifier using machine learning features in ER. Since the inaugural version, the latest version of this classifier is deployed to discard records scoring below a ‘safe’ threshold score *before* adding the retained records scoring above the same threshold score to our priority screening (active learning) workflow.

### Eligibility criteria

Prioritised records and corresponding full-texts are screened by one researcher who applies our eligibility criteria[[8]](#footnote-5) (reproduced below) based on the latest version of our living screening guidance notes (also reproduced below) using the same procedure described for the pilot phase (including the option to refer a record for a ‘second opinion’). A single reason for exclusion (if applicable) is recorded as the first applicable exclusion criterion in this hierarchical list.

**Excluded - Published before 2018**

This criterion is applied to records published before 1st January 2018 (i.e. with publication year of 2017 or earlier), which have not already been identified and removed prior to assigning records for eligibility screening.

**Excluded - Duplicate record**

This criterion is applied to any records, that are duplicates of one or more other (eligible or ineligible) records (i.e. further duplicates not already identified and removed prior to assigning records for eligibility screening).

**Excluded - Reported in a language other than English**

This criterion is applied when the corresponding full-text report is published in any language other than English. This code is also applied when the title and abstract (if available) is reported in a language other than English and the corresponding full text is not available.

**No major focus on generative LLM-based tool(s) OR no major focus on health or social care application(s)**

This criterion is applied to records published in full journal articles, that *EITHER*:

Do not include a major focus on one or more generative large language model-based tool(s). Major focus is defined as a sole, predominant or substantive focus (judgement required). Generative LLM-based tools is a class of artificial intelligence tools capable of generating text, images, audio and/or other media (e.g. computer code) in response to user- (or self-) prompts. To be eligible, tool(s) must (explicitly or implicitly/ probably) be underpinned by LLMs with both encoder *and* decoder (encoder + decode = generative) architecture (i.e. LLM must perform one or more generative decoder operation(s)). If not explicitly generative, but explicitly an LLM-based tool *and* all issues/ topics covered are relevant to generative, then eligible on this criterion (but may still be ineligible on others, below). If focus is on BERT (or closely related) models, exclude.

*OR*

Do not explicitly involve a major focus on the use of generative LLM-based tools for tasks directly related to one or more health or social care applications. Major focus is defined as a sole, predominant or substantive focus (judgement required). Social care is defined as the provision of social work, personal care, protection or social support services to children or adults in need or at risk, or adults with needs arising from illness, disability, old age or poverty. Health and social care applications encompass medical education applications (including answering medical exam questions) and other applications related to workforce development or the training of health or social care professionals, including public health professionals. Health and social care applications also encompass patient or service user education applications.

This criterion is not applied to records published in full journal articles that explicitly include a major focus on applying generative LLM-based tools for tasks related to the prioritisation, planning, production, of empirical research studies, encompassing systematic reviews, other types of evidence synthesis, primary research and/or modelling studies – whether or not the application is to health or social care evidence synthesis/ research (see also the next criterion (below) instead.

**Excluded - Evidence synthesis or research application**

This criterion is applied to records published in pre-print or full journal articles with a major focus on applying generative LLM-based tools for tasks related to the prioritisation, planning, production, of empirical research studies, encompassing primary research, evidence synthesis (including systematic reviews), modelling, technology appraisals, and clinical or public health guideline development processes and tasks; and without a major focus on the use of generative LLM-based tools for tasks directly related to health and/or social care applications. This criterion is not applied to records that include a major focus on applications which explicitly have a health and/or social care professional educational/ training/ clinical development role as well as a potential research/ publication function (e.g. a generative LLM being used for helping to write case reports in healthcare).

**Excluded - Letters/ correspondence, ‘research highlights’ (or similar) articles, or corrections / errata explicitly linked to a probably eligible single primary research study**

This criterion is applied to records that explicitly refer to, summarise and/or ‘highlight’ a single eligible primary research report only - published in the same issue, or elsewhere (see for example: <https://doi.org/10.1038/s44222-023-00097-7>). This criterion is not applied to letters/ correspondence that incorporate an original/ de novo formal evaluation (empirical studies) or substantive discussion (non-empirical) of performance of generative LLM-based tools for tasks related to one or more health or social care applications.

**Excluded - No formal evaluation (empirical studies) or substantive discussion (non-empirical) of performance for one or more specific tasks**

This criterion is applied to records that explicitly involve applying generative LLM-based tools for tasks directly related to one or more health or social care applications but which do not incorporate a formal evaluation (empirical studies) or major focus on discussing (commentary/ review-type articles) the performance of such tools. To qualify as a (record/ report of an) empirical study with formal evaluation, the report must include one or more (implicit or explicit) research aim (or research question), a description of the methods used, and at least some results. We are primarily interested in articles that make claims about the performance of eligible tools for specific tasks.

**Excluded - Not a full journal article**

This criterion is applied to records that are not full journal articles. For example, pre-print articles, dissertation and theses, and grey literature publications, are excluded on this criterion. Conference proceedings articles are eligible if the report is a full-text article; but excluded if the report is an abstract only. Letters to editors and/or other correspondence articles are included if otherwise eligible. This criterion is also applied to otherwise eligible retracted articles.

**Included - Primary research**

This criterion is applied to records of eligible primary research studies that formally evaluate the performance (i.e. beneficial, adverse and/or differential impacts) of generative LLM-based tools for any specific health care (including public health) and/or social care application. Include reports of studies that investigate the acceptability of, or people's experiences / views concerning, (use of) generative LLM-based tools for specific health and/or social care applications. Include reports of studies that formally evaluate the cost-effectiveness and/or costs of using generative LLM-based tools for health and/or social care applications. To qualify as a record of an empirical primary study with formal evaluation, the report must include one or more (implicit or explicit) research aim (or research question), a description of the methods used, and at least some results. This criterion is also applied to reports of protocols for eligible primary studies. If the article reports both an eligible primary research component and an eligible systematic review/ other evidence synthesis / modelling component, or both an eligible primary research component and an eligible commentary or review component, then the ‘Included – Primary research’ criterion is applied.

**Included - Systematic review / other research synthesis / modelling**

This criterion is applied to records of eligible systematic reviews, other research synthesis, or modelling studies that evaluate the performance (i.e. beneficial, adverse and/or differential impacts) of generative LLM-based tools for any specific health care (including public health) and/or social care application.

We define ‘systematic review or other research synthesis’ as any record reporting secondary data that reports: some search terms; clearly defined inclusion criteria; and some information on the selection process (at least N of references located by searches and N of studies included). If the report is an otherwise eligible review article that does not meet the latter criteria for being considered a systematic review, the criterion ‘Included - Commentary or review-type article’ is applied instead (see the next inclusion criterion, below). We include any systematic review which aimed to include eligible studies, whether or not any were located. We include updates to eligible systematic reviews, living systematic reviews, or other research synthesis, if the record presents new data and the original review or research synthesis meets the criterion described above.

We also include any modelling studies that are at least partly based on relevant empirical data (e.g. data used as inputs to the model, or data against which the model is being calibrated or tested).

If the article reports both an eligible primary research component and an eligible systematic review/ other evidence synthesis / modelling component, the ‘Included – Primary research’ criterion is applied.

**Included – Protocol**

Apply this code only to protocols of eligible primary research studies, systematic reviews, other research synthesis or modelling studies.

## Classifying the evidence

### Overview

This is a living map, and our classification (coding) scheme and guidance notes continually evolve as we encounter new studies during manual coding and discuss and resolve ‘boundary cases’. This coding scheme also went through multiple cycles of piloting and revision before we settled on the inaugural version.

We apply the current coding scheme (reproduced below) to each record included in the ‘main map’ based on its ‘applications class(es)’ code(s) (which is applied by human researchers during eligibility screening). First, we classify all records (reports) for which the corresponding full-text article can be retrieved and uploaded with the assistance of generative-AI tools (currently GPT4-o) in EPPI Reviewer[[9]](#footnote-6) using the full-text PDF markdown text. Second, any ‘main map’ records (reports) for which the corresponding full-text article cannot be retrieved, or for which generative AI-assisted coding fails, are manually classified by humans using the following procedure: Two of three researchers provisionally code each record and a fourth researcher checks the other two’s coding, arbitrates when needed and creates a final version of the coding. There is also the option for any of the researchers to refer records for a ‘second opinion’ in relation to either coding or eligibility. These ‘second opinions’ are sometimes easily resolved by the fourth researcher, or else are escalated for discussion among all four researchers to reach a consensus decision.

Our full coding scheme for ‘main map’ records (reports) comprises 9 dimensions:

* Application class(es)
* Article type
* Model(s)
* Mode(s) of model(s) use
* Task(s) Type(s)
* Health / social care topic(s)
* Population(s)
* Map version
* Coding status

Coding of systematic reviews (and other forms of evidence synthesis) is based on characteristics of the included studies (rather than on the ES eligibility criteria).

If an article focuses on (i) medical education applications and/or (ii) answering medical (and related) exam questions or answering medical questions in general (no specific application), and/or (iii) answering patient or service user questions, *and* also on eligible clinical health, public health and/or social care applications, then further ‘main map’ coding is intended to be applied based solely on the clinical health, public health and/or social care applications (and no further coding is intentionally applied that is only applicable to these other applications (i), (ii) and/or (iii)).

Further information on each dimension of the current classification (coding) scheme is provided below.

### Application class(es)

Included records are first partitioned into broad, top-level classes of health or social care applications, as follows (we select all codes that apply):

* Clinical health care. Excludes medical education / medical exams, health professional development and training, patient or service user education, and public health. Includes reports focused on patient question answering tools that explicitly culminate in, or otherwise involve, one or more clinical decision (code the latter reports as ‘Clinical health care’).
* Public health. Excludes education for public health professionals / related exams, public health professional development and training, and public health-related patient or service user education).
* Social care. Excludes education for social care professionals / related exams, social care professional development or training, and social care-related service user education.
* Medical education, health or social care professional development and training. Also use this code for applications that involve generating (but not answering) medical (or related) exam questions.
* Answering medical (and related) exam questions or answering medical questions in general (no specific application). We also use this code for applications that involve both generating and answering medical (or related) exam questions.
* Answering patient or service user questions / patient or service user education. Answering patient questions about medical conditions, treatments, or health services, unless use of the tool explicitly culminates in, or otherwise involves, one or more clinical decisions (code the latter reports as ‘clinical health care’). We also use this code for applications that involve generating patient education materials.

Records of articles focused exclusively on (i) medical education applications and/or (ii) answering medical (and related) exam questions or answering medical questions in general (no specific application), and/or (iii) answering patient or service user questions, or patient or service user education, are partitioned into their own separate ‘sector’ of the map and minimal further map coding (‘article type’, ‘version’ and ‘coding status’ – see below) is applied to these records by a single human researcher.

### Article type

Included records are segmented into three broad types. We select one code:

* Primary research
* Systematic reviews (and other research syntheses)
* Protocol (for a primary research study or systematic review).

### Model(s)

Each record is classified according to the specific LLM(s) under investigation.

We select all codes that apply:

* Not about specific model(s)
* BioGPT
* Claude / Claude 2
* DeepSeek
* Falcon 40-B
* Gemini / Bard
* GPT-2 / GPT-3 / GPT-4 / ChatGPT (OpenAI family)
* LLaMA / LLaMA-2 / 3 / 4
* Med-PaLM
* Mistral / Mixtral
* PaLM / PaLM 2
* PanGu-α / PanGu-Σ
* PubMed GPT
* Other(s) – specified
* Other(s) – not specified

### Mode(s) of model(s) use

We select all codes that apply and at least one code:

* Standard LLM. Select this code if it can reasonably be inferred that one or more eligible tool(s) is underpinned by a standard generative LLM (i.e. an LLM that has not been fine-tuned, and that is not being used in a Retrieval Augmented Generation (RAG) workflow).
* Fine-tuned LLM. Select this code if it is explicitly stated that one or more eligible tool(s) is underpinned by an LLM that has been fine-tuned.
* Retrieval Augmented Generation (RAG). Select this code if it is explicitly stated that one or more eligible tool(s) is underpinned by an LLM that is being used in a Retrieval Augmented Generation (RAG) workflow.
* Unclear. We select this code if none of the other codes apply.

### Task(s) type(s)

This code set comprises a series of concise, output-focussed[[10]](#endnote-4) task descriptors. We prospectively formulated these codes during pilot phases of coding based on the article authors’ verbatim description(s) of task(s) and their output(s).

We select all codes that apply, up to 3 specific types (or we select ‘various tasks’ if 4 or more / more than 3 specific types):

* Various types of tasks. We apply this code to reports focused on **four or more (i.e. more than three)** different types of specific tasks.
* Diagnosing / recognising / detecting (e.g. symptoms)
* Extracting information (where information is the output)
* Predicting (e.g. disease risk(s)).
* Recommending / suggesting (e.g. treatments). If this code is applicable, then ‘supporting / assisting clinical decisions may or may not also be applicable.
* Summarising / simplifying (e.g. discharge summaries, simplified radiology reports)
* Supporting / assisting clinical decisions (e.g. clinical decision aids). If this code is applicable, then ‘recommending / suggesting’ is invariably also likely to be applicable. For this code to be used in addition to ‘diagnosing…’ there must also explicitly be a subsequent (in the pathway) management / treatment decision that is being informed by model output(s).
* Supporting / assisting patients (e.g. speech assistant for people with aphasia)
* Translating - language
* Triaging (e.g. prioritising patients)
* Writing / transcribing (e.g. medical notes, letters or e-mails)
* Other(s) – specify
* Unclear / not reported

### Health / social care topic(s)

Included articles (records) are also partitioned into broad areas of health and disease, and/or social care, based on the following topic categories (adapted from UK Clinical Research Collaboration 2024), or ‘Not applicable’.

We select all codes that apply (up to 5 specific areas for health or select ‘generic health relevance if 6 or more/ more than 5 areas):

* Generic health relevance. Apply this code when the report is judged relevant to **6 or more (i.e. more than 5)** [health categories](https://hrcsonline.net/getting-started/general-approach-to-coding/assigning-health-categories/). Also apply this code when the report is judged of relevance to all diseases and conditions or to health and well-being in general.
* Blood
* Cancer and neoplasms. Use this code only for cancers/ neoplasms regardless of type(s) and location(s).
* Cardiovascular
* Congenital disorders
* Diagnostic services
* Ear
* Eye
* Infection
* Inflammatory and immune system
* Injuries and accidents
* Mental health
* Metabolic and endocrine
* Musculoskeletal
* Neurological
* Oral and gastrointestinal
* Plastic surgery
* Renal and urogenital
* Reproductive health and childbirth
* Respiratory
* Skin
* Stroke
* Public health
* Disputed aetiology
* Adult social care
* Children’s social care
* Other(s) - specify

### Population(s)

In addition to health and social care categories that may reflect the specific health conditions experienced by study participants, we also apply codes for the following participant characteristics for studies included in the main map, if applicable. These codes are only applied to reports of primary research studies with human participants and SRs of studies that consider human participants with specific characteristics[[11]](#endnote-5), excluding ‘simulated cases’/ ‘simulated patients’.

**Age**

We select all codes that apply, or ‘Not applicable’:

* Infants
* Children
* Adults
* Older adults
* Unclear / not reported
* Not applicable

Infants = 0-24 months; Children = 2-17 years; Adults = 18 to 64 years; Older adults = 65+ years.

If the article does not report any information about the age of the participants = Select: ‘Unclear / not reported’.

If the article does not report in detail the age of the participants, but it describes the participants as “adults” = Select: ‘Adults’ AND ‘Older adults’.

If the article states that “children” were in the study population = Select: ‘Children’ BUT NOT ‘Infants’.

If the age is reported as a mean, select the most appropriate check box. Example: “mean age 43 years” = Select: ‘Adults’.

If the age is reported as a mean and its standard deviation, standard error, or other measure of spread or variability, select the most appropriate check box or boxes. Example: “mean age 18.0 ±10.0 years” = Select: ‘Children’ AND ‘Adults’

If the age is reported as "older than 18 years" and no upper age limit (nor any other clear indication that the upper age limit of participants is likely to be <65 years) is reported = Select: 'Adults' AND 'Older adults'

**Sex**

We select all codes that apply, or ‘Not applicable’:

* Male
* Female
* Unclear / not reported
* Not applicable

**Countries / Region**

We select all codes that apply, or ‘Not applicable’. For systematic reviews and other research syntheses, we select ‘Not applicable’ unless eligibility criteria for considering studies include specific geographical restrictions:

* United Kingdom
* Other Europe
* Canada
* United States of America
* Other North / Central America
* South America
* Africa
* Asia
* Oceania
* Unclear / not reported
* Not applicable

### Map version

These codes are assigned to each included record based on the version number of the living map in which the record is first published. We select ‘All versions’ and the (one) other applicable version code.

* All versions
* Version 1 – 2nd September 2024
* Version 2 – 14th October 2024
* Version 3 – 26th November 2024
* Version 4 – 10th December 2024

### Coding status

One of these codes is assigned to each included record based on whether the report has been classified (coded) either manually by human researchers (‘humans’) or with the assistance of generative AI tools (’generative AI’).

* Humans
* Generative AI

## Acknowledgements

We would like to thank members of our project advisory group for their helpful suggestions on the presentation of this living map, including its evidence and gap maps: Caroline Green (University of Oxford), Chloe Park (Department of Health and Social Care), Ellen Coughlan (The Health Foundation), Iain Marshall (King's College London), Nell Thornton (The Health Foundation), Pritesh Mistry (The King's Fund) and Sam Taylor (Department of Health and Social Care). We would also like to thank our EPPI Centre colleagues who were co-authors of the [protocol](https://eppi.ioe.ac.uk/CMS/Portals/0/Generative%20LLM-Based%20Tools%20for%20Health%20and%20Social%20Care%20Applications_Protocol_100624-Final.pdf) for this living map, and who also provided feedback on its presentation: Alison O'Mara-Eves, Dylan Kneale and Katy Sutcliffe. Finally, we would like to thank our other EPPI Centre colleagues for their feedback on our searches for eligible studies: Claire Stansfield and Hossein Dehdarirad.

## Funding

This work has been commissioned by the National Institute for Health Research Policy Research Programme (NIHR PRP) for the UK Department of Health and Social Care (DHSC). It is being funded through the NIHR PRP contract with the EPPI Centre at University College London (UCL): Reviews facility to support national policy development and implementation, NIHR200701. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the NIHR or the DHSC.

## Funder involvement

This work is part of an ongoing programme of research funded by the DHSC (see also ‘Contributions’, below). DHSC policy team members will be consulted at various stages to help us understand the policy context of the issue under study and identify the focus and scope of the analysis.

## Conflicts of interest

None.

## Contributions

This work is being undertaken by a team of researchers based at the [EPPI Centre](https://eppi.ioe.ac.uk/cms/) and the [Centre for Reviews and Dissemination (CRD)](https://www.york.ac.uk/crd/)under the auspices of the [UK NIHR Policy Research Programme Reviews Facility](https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=73) (NIHR PRP Reviews Facility). The NIHR PRP Reviews Facility is a collaboration between the EPPI Centre, University College London, the Centre for Reviews and Dissemination (CRD), University of York, and the [Department of Public Health, Environments and Society (PHES)](https://www.lshtm.ac.uk/research/faculties/php/public-health-environments-and-society), London School of Hygiene and Tropical Medicine.

## References

Care Quality Commission. (2024). *Care Quality Commission (CQC) - Guidance for Providers > Service Types*. [Webpage]. Available from: https://www.cqc.org.uk/guidance-providers/regulations-enforcement/service-types

EPPI Centre. (2024). *AI and equity: what are the benefits and harms associated with this new generation of decision-making tools?* [Web page]. London: EPPI Centre. Available from: https://eppi.ioe.ac.uk/cms/Projects/Technologyresearchdevelopment/AIandequity/tabid/3923/Default.aspx

Miwa M, Thomas J, O'Mara-Eves A, Ananiadou S. (2014). Reducing systematic review workload through certainty-based screening. *Journal of Biomedical Informatics*; 51: 242-253. Available from: https://doi.org/10.1016/j.jbi.2014.06.005

O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. (2015). Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic Reviews*; 4:5. Available from: https://doi.org/10.1186/2046-4053-4-5

Shemilt I, Hollands GJ, Khouja C, Raine G, Kneale D, O’Mara-Eves A, Sutcliffe K, Thomas J. (2024). *Generative Large Language Model-Based Tools for Health and Social Care Applications: A Living Map and Critical Review (Protocol)*. London: EPPI Centre. ISBN: 978-1-911605-57-7 Available from: https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=3926

Tenti P, Pasi G, Thomas J, Peñaloza R. (2023). A Content-Based Recommendation Model for Living Evidence. [Pre-print]. *Social Sciences Research Network*. Available from: http://dx.doi.org/10.2139/ssrn.4406117

Thomas J, Graziosi S, Brunton J, Ghouze Z, O'Driscoll P, Bond M. (2024). *EPPI-Reviewer: advanced software for systematic reviews, maps and other evidence synthesis* [Software]. London: EPPI Centre. Available from: https://eppi.ioe.ac.uk/cms/er

UK Clinical Research Collaboration. (2024). *UKCRC Health Research Classification System - Health Categories*. [Web page]. Available from: https://hrcsonline.net/health-categories/

## Appendix 1. Conventional electronic searches

**MEDLINE (Ovid SP) -- 1st January 2018 to 25th January 2024.**

1. large language model$.ti.

2. LLM$.ti.

3. generative artificial intelligence.ti.

4. generative ai.ti.

5. generative pre-trained transformer$.ti.

6. gpt$.ti.

7. chatgpt.ti.

8. or/1-7

9. limit 8 to yr="2018 -Current"

10. limit 9 to english language

**MEDLINE (Ovid SP) -- 1st January 2018 to 1st March 2024.**

1. large language model$.ti.

2. LLM$.ti.

3. generative artificial intelligence.ti.

4. generative ai.ti.

5. generative pre-trained transformer$.ti.

6. gpt$.ti.

7. chatgpt.ti.

8. bard.ti.

9. llama.ti.

10. claude.ti.

11. med-palm.ti.

12. mistral.ti.

13. mixtral.ti.

14. or/1-13

15. limit 14 to yr="2018 -Current"

16. limit 15 to english language

**Embase (Ovid SP) -- 1st January 2018 to 25th January 2024.**

1. large language model$.ti.

2. LLM$.ti.

3. generative artificial intelligence.ti.

4. generative ai.ti.

5. generative pre-trained transformer$.ti.

6. gpt$.ti.

7. chatgpt.ti.

8. or/1-7

9. limit 8 to yr="2018 -Current"

10. limit 9 to english language

11. limit 10 to conference abstract status

12. 10 not 11

**Embase (Ovid SP) -- 1st January 2018 to 1st March 2024.**

1. large language model$.ti.

2. LLM$.ti.

3. generative artificial intelligence.ti.

4. generative ai.ti.

5. generative pre-trained transformer$.ti.

6. gpt$.ti.

7. chatgpt.ti.

8. bard.ti.

9. llama.ti.

10. claude.ti.

11. med-palm.ti.

12. mistral.ti.

13. mixtral.ti.

14. or/1-13

15. limit 14 to yr="2018 -Current"

16. limit 15 to english language

17. limit 16 to conference abstract status

18. 16 not 17

**OpenAlex Custom / Boolean Search --** **1st January 2018 to present (first run on 11th March 2024 and updated once every ~2 months from date of preceding search)**

1. primary\_location.source.id:S4389157821,from\_publication\_date:YYYY-MM-DD
2. display\_name.search:large language model,from\_publication\_date:YYYY-MM-DD
3. display\_name.search:LLM,from\_publication\_date:YYYY-MM-DD
4. display\_name.search:generative artificial intelligence,from\_publication\_date:YYYY-MM-DD
5. display\_name.search:generative ai,from\_publication\_date:YYYY-MM-DD
6. display\_name.search:generative pre-trained transformer,from\_publication\_date:YYYY-MM-DD
7. display\_name.search:gpt,from\_publication\_date:YYYY-MM-DD
8. display\_name.search:chatgpt,from\_publication\_date:YYYY-MM-DD
9. display\_name.search:bard,from\_publication\_date:YYYY-MM-DD
10. display\_name.search:llama,from\_publication\_date:YYYY-MM-DD
11. display\_name.search:claude,from\_publication\_date:YYYY-MM-DD
12. display\_name.search:med-palm,from\_publication\_date:YYYY-MM-DD
13. display\_name.search:mistral,from\_publication\_date:YYYY-MM-DD
14. display\_name.search:mixtral,from\_publication\_date:YYYY-MM-DD
15. 1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 OR 11 OR 12 OR 13 OR 14

## Appendix 2. OpenAlex import filters

The first set of import filters below (‘Set 1’) were applied to the results of all three kinds of OpenAlex searches (auto-update, network graph, custom) up to and including those executed on 19th July 2024. The second set of import filters below (‘Set 2’) were applied to the results of all three kinds of OpenAlex searches from those executed on 30th August 2024 onwards. The change from ‘Set 1’ to ‘Set 2’ reflects a decision – representing a deviation from our protocol (Shemilt 2024) – which we made during August 2024, to exclude pre-print articles from this living map (we had previously been including and coding eligible pre-prints).

### Set 1

**Title**

supplemental material, supplementary material, decision letter, faculty opinions, author comment

### Set 2

**Journal**

medrxiv,biorxiv,chemrxiv,vixra,arxiv,techrxiv,crimrxiv,new scientist,the conversation,research square,nejm journal watch,veterinary record,chemical & engineering news,physics today,reactions weekly,authorea,carolina digital repository,psycextra dataset,repec,zenodo,social science research network,doaj,preprints,ebooks,book series

**URL**

arxiv,vixra,preprint,preprints,tathapi.com,osf.io

**DOI**

preprint,10.22541,10.21203,10.1101,10.2139,10.36227,10.3410/f.,10.21428,zenodo,protocols.io,chemrxiv,techrxiv,figshare,10.2196/preprints.,/REVIEWS/,10.2210/PDB,/REVIEW1,/REVIEW2,/REVIEW3,/REVIEW4,osf.io,protocols.io

**Title**

supplemental material,supplementary material,decision letter,faculty opinions,author comment

## Endnotes

1. In early versions of the main map, records were also classified according to ‘service type(s)’. Since production of the main map transitioned to a generative AI-assisted coding workflow (see ‘Classifying the evidence’, below in this document) – i.e. from Version 4 onwards – we have discontinued use of the ‘service type(s)’ code set. We took this decision due to the unreliability of automated classification of records for ‘service type(s)’ using generative AI tools. Although records in the main map had previously classified according to ‘service type(s)’ by human coders up to Version 3, we have also discontinued publishing ‘service type(s)’ classifications for the latter records from Version 4 onwards; however, these metadata are available on request to [i.shemilt@ucl.ac.uk](mailto:i.shemilt@ucl.ac.uk). [↑](#footnote-ref-1)
2. ‘Coding status’ classifies records according to whether they have been coded by ‘humans’ only or using our generative AI-assisted (‘gen AI’) workflow. See ‘Classifying the evidence’, below in this document, for further details. [↑](#footnote-ref-2)
3. From the early stages of scoping and identifying records for inclusion in this living map, it became evident that we would be likely to encounter large numbers of studies about the use of generative LLM-based tools for ‘medical education, health or social care professional development and training’, ‘answering medical (and related) exam questions/ answering medical questions in general’ and/ or ‘answering patient or service user questions/ patient or service user education’ applications.   
     
   We also observed that many, but not all, studies of the latter classes of applications published to date comprise small studies, with relatively narrow scope, and/or with similar content. From a map design perspective, including all of those records in the same map as the much smaller sets of records of studies about ‘clinical health care’, ‘public health’ and/ or ‘social care’ applications would make it difficult or impossible to visually discern any differences between cells or segments in user-configured or pre-configured evidence and gap maps (EGMs) incorporating the ‘Application class(es)’ dimension; rendering those EGMs essentially unusable/ uninterpretable.   
     
   The sheer volume of the former, larger sets of studies would also mean that fully classifying (coding) them would limit our ability to fully classify (code) studies of ‘clinical health care’, ‘public health’ and/ or ‘social care’ applications. Based on early consultations with key stakeholders, we judge the latter classes of applications are likely to be higher priority for decision making (compared with the former classes), potentially warranting more coverage in our planned critical review. [↑](#endnote-ref-1)
4. Records included in the ‘main map’ may or may not *also* be about ‘medical education, health or social care professional development and training’, ‘answering medical (and related) exam questions/ answering medical questions in general’ and/ or ‘answering patient or service user questions/ patient or service user education’ applications of one or more generative LLM-based tools. However, map coding is applied based *solely* on the clinical health, public health and/or social care applications (and no further coding is applied that is only applicable to these other applications). [↑](#endnote-ref-2)
5. Instead, during that period, we had invested our manual effort in transitioning to a generative AI-assisted classification workflow. [↑](#footnote-ref-3)
6. We took this decision due to the unreliability of automated classification of records of ‘commentary/ review-type’ articles using generative AI tools. The removed records are available on request to [i.shemilt@ucl.ac.uk](mailto:i.shemilt@ucl.ac.uk). [↑](#footnote-ref-4)
7. [OpenAlex](https://openalex.org/) is an open access dataset and knowledge graph comprising ~250 million bibliographic records of research articles (reports) from across science, connected in a very large network graph of conceptual, citation and other (e.g. author) relationships. The OpenAlex dataset is automatically, and continuously, updated with new records as new articles (reports) are published online. We have developed OpenAlex tools (in EPPI Reviewer) to enable regular automated searches of this dataset - primarily to support (continual) semi-automated updating of (living) systematic reviews, (living) evidence maps, and related evidence synthesis use scenarios. [↑](#endnote-ref-3)
8. From Version 4 onwards, we have removed all records of ‘commentary/ review-type’ articles, previously published up to Version 3, from the ‘main map’ and the separate ‘sector’. We took this decision due to the unreliability of automated classification of records of ‘commentary/ review-type’ articles using generative AI tools. These removed records are available on request to [i.shemilt@ucl.ac.uk](mailto:i.shemilt@ucl.ac.uk). [↑](#footnote-ref-5)
9. Full details of current generative AI prompts and pre-specified recoding procedures are available on request to [i.shemilt@ucl.ac.uk](mailto:i.shemilt@ucl.ac.uk). [↑](#footnote-ref-6)
10. ‘Output-focussed’ means ‘focussed on the output(s) that are generated by the LLM-based tool’. [↑](#endnote-ref-4)
11. Coding of systematic reviews and/or other evidence syntheses is based on characteristics of included studies (rather than on eligibility criteria, e.g. PICO). [↑](#endnote-ref-5)