Data

the Bottleneck to Trustworthy LLM solutions for Systematic Review Automation

Kavita Kothari & Elias Sandner

09/07/2025

Problem

- Large Language Models have the potential to significantly reduce workload in systematic reviews.[1]
- A large problem remains the consistency and reliability of LLM outputs.[2]
- A growing issue is the availability of LLM tools without (proper) evaluation.
- Developers are using smaller evaluation data sets partly due to lack of availability of large data sets for evaluation.[3]

Systematic Review Presented at Data Science Technology & Application[3]

Authors: Elias Sandner [1; 2]; Luca Fontana [3]; Kavita Kothari [4]; Andre Henriques [5]; Igor Jakovljevic [2]; Alice Simniceanu [3]; Andreas Wagner [2] and Christian Gütl [1]

Affiliations: [1] Cognitive & Digital Science Lab, Technical Unversity Graz, Graz, Austria ; [2] IT Department, CERN, Geneva, Switzerland ; [3] Health Emergencies Programme, WHO, Geneva, Switzerland ; [4] Consultant to Library & Digital Information Networks, WHO, Kobe, Japan ; [5] Occupational Health & Safety and Environmental Protection (HSE) Unit, CERN, Geneva, Switzerland

Keyword(s): Systematic Review, Evidence Synthesis, Large Language Models, Literature Screening Automation, Binary Text Classification.

Abstract: Systematic reviews provide high-quality evidence but require extensive manual screening, making them time-consuming and costly. Recent advancements in general-purpose large language models (LLMs) have shown potential for automating this process. Unlike traditional machine learning, LLMs can classify studies based on natural language instructions without task-specific training data. This systematic review examines existing approaches that apply LLMs to automate the screening phase. Models used, prompting strategies, and evaluation datasets are analyzed, and the reported performance is compared in terms of sensitivity and workload reduction. While several approaches achieve sensitivity above 95%, none consistently reach the 99% threshold required for replacing human screening. The most effective models use ensemble strategies, calibration techniques, or advanced prompting rather than relying solely on the latest LLMs. However, generalizability remains uncertain due to dataset limitatio (More)

Evaluating Large Language Models for Literature Screening: A Systematic Review of Sensitivity and Workload Reduction

Elias Sandner[1,5,a], Luca Fontana[2,b], Kavita Kothari[3,c], Andre Henriques[4,d], Igor Jakovljevic[1,e], Alice Simniceanu[2,f], Andreas Wagner[1,g] and Christian Gütl[5,h]

[1] *IT Department, CERN, Geneva, Switzerland*
[2] *Health Emergencies Programme, WHO, Geneva, Switzerland*
[3] *Consultant to Library & Digital Information Networks, WHO, Kobe, Japan*
[4] *Occupational Health & Safety and Environmental Protection (HSE) Unit, CERN, Geneva, Switzerland*
[5] *Cognitive & Digital Science Lab, Technical Unversity Graz, Graz, Austria*

Keywords: Systematic Review, Evidence Synthesis, Large Language Models, Literature Screening Automation, Binary Text Classification.

Abstract: Systematic reviews provide high-quality evidence but require extensive manual screening, making them time-consuming and costly. Recent advancements in general-purpose large language models (LLMs) have shown potential for automating this process. Unlike traditional machine learning, LLMs can classify studies based on natural language instructions without task-specific training data. This systematic review examines existing approaches that apply LLMs to automate the screening phase. Models used, prompting strategies, and evaluation datasets are analyzed, and the reported performance is compared in terms of sensitivity and workload reduction. While several approaches achieve sensitivity above 95%, none consistently reach the 99% threshold required for replacing human screening. The most effective models use ensemble strategies, calibration techniques, or advanced prompting rather than relying solely on the latest LLMs. However, generalizability remains uncertain due to dataset limitations and the absence of standardized benchmarking. Key challenges in optimizing sensitivity are discussed, and the need for a comprehensive benchmark to enable direct comparison is emphasized. This review provides an overview of LLM-based screening automation, identifying gaps and outlining future directions for improving reliability and applicability in evidence synthesis.

1 INTRODUCTION

By synthesizing findings from potentially all relevant studies on a given research question, a Systematic Review (SR) represents the most reliable research methodology for evidence-based conclusions (Shekelle et al., 2013). Therefore, SRs play a crucial role in the medical field, guiding decision-making and shaping clinical practice guidelines (Cook et al., 1997). However, the rigor of systematic reviews makes them highly time- and resource-intensive, often taking months or even years to complete.

Systematic reviews typically begin with a broad database query to ensure comprehensive coverage, followed by human screening—a particularly time-consuming stage of the process (Carver et al., 2013).

Despite following a well-defined procedure, automating the screening phase remains challenging. Existing methods often fall short of human-level sensitivity and lack generalizability across review domains. Traditional ML approaches can support large-scale or living SRs, but their effectiveness is limited by the scarcity of high-quality training data. (Sandner et al., 2024a)

General-purpose LLMs have shown strong performance in classification tasks. Trained on vast text

[a] https://orcid.org/0009-0007-9855-4923
[b] https://orcid.org/0000-0002-8614-4114
[c] https://orcid.org/0000-0002-0759-5225
[d] https://orcid.org/0000-0003-1521-3423
[e] https://orcid.org/0000-0003-1893-9553
[f] https://orcid.org/0000-0003-4068-6177
[g] https://orcid.org/0000-0001-9589-2635
[h] https://orcid.org/0000-0001-9589-1966

Sandner, E., Fontana, L., Kothari, K., Henriques, A., Jakovljevic, I., Simniceanu, A., Wagner, A., Gütl and C.
Evaluating Large Language Models for Literature Screening: A Systematic Review of Sensitivity and Workload Reduction.
In *Proceedings of the 14th International Conference on Data Science, Technology and Applications*

Limitations of Datasets used for Evaluating LLM Based Study Selection[3]

	Num of Reviews	Num of Records	Num of Includes	Blinded screening by 2 reviewers	Domain	Dataset
(Khraisha et al., 2024) - Full Text	1	150	39	partially	Parenting in protected refugee situations	provided by authors
(Khraisha et al., 2024) - TiAb	1	300	103	yes	Parenting in protected refugee situations	provided by authors
(Gargari et al., 2024)	1	330	13	Yes	Light therapy in insomnia disorder	not shared
(Spillias et al., 2024)	1	1098	101	No (1 screener)	Community-Based Fisheries Management (CBFM)	provided by authors
(Li et al., 2024)	3	505	205	Yes	Public Health	Subset of **SYNERGY Dataset**
(Cai et al., 2023)	4	400	40	Yes	Disease	provided by authors
(Tran et al., 2023)	5	22666	1485	Yes	Medical	Available on request
(Issaiy et al., 2024)	6	1180	148	Yes	Radiology	provided by authors
(Guo et al., 2024)	6	24845	538	Yes	Clinical Pharmacology and Therapeutics	provided by authors
(Cao et al., 2024) - TiAb	10	4000	779	yes	Public Health	not shared
(Akinseloyin et al., 2024)	31	39847	885	Yes	Prognosis, Qualitative, Intervention, DTA Studies	Subset of **CLEF-TAR Dataset**
(Wang et al., 2024)	128	657980	10524	yes	DTA and Intervention studies	CLEF-TAR 2017-2019

Proposed solution:
Standardised Dataset Collection

1. Prioritisation of the SR stages for LLM evaluation
2. Steps to build a dataset
   a. Define metadata (required / nice-to-have)
   b. Design dataset considering legal limitations (redistribute or provide retrieval script)
   c. Consensus on Format (XML, CSV, SQL)
   d. Hosting Platform (Zenodo, OSF, GitHub, API-Service)
3. Assess existing data sets – integration?
4. Guidance for creating acceptable data sets
5. Community Feedback System: Identify and correct corrupted data

elias.sandner@cern.ch| kavitaukothari@hotmail.com

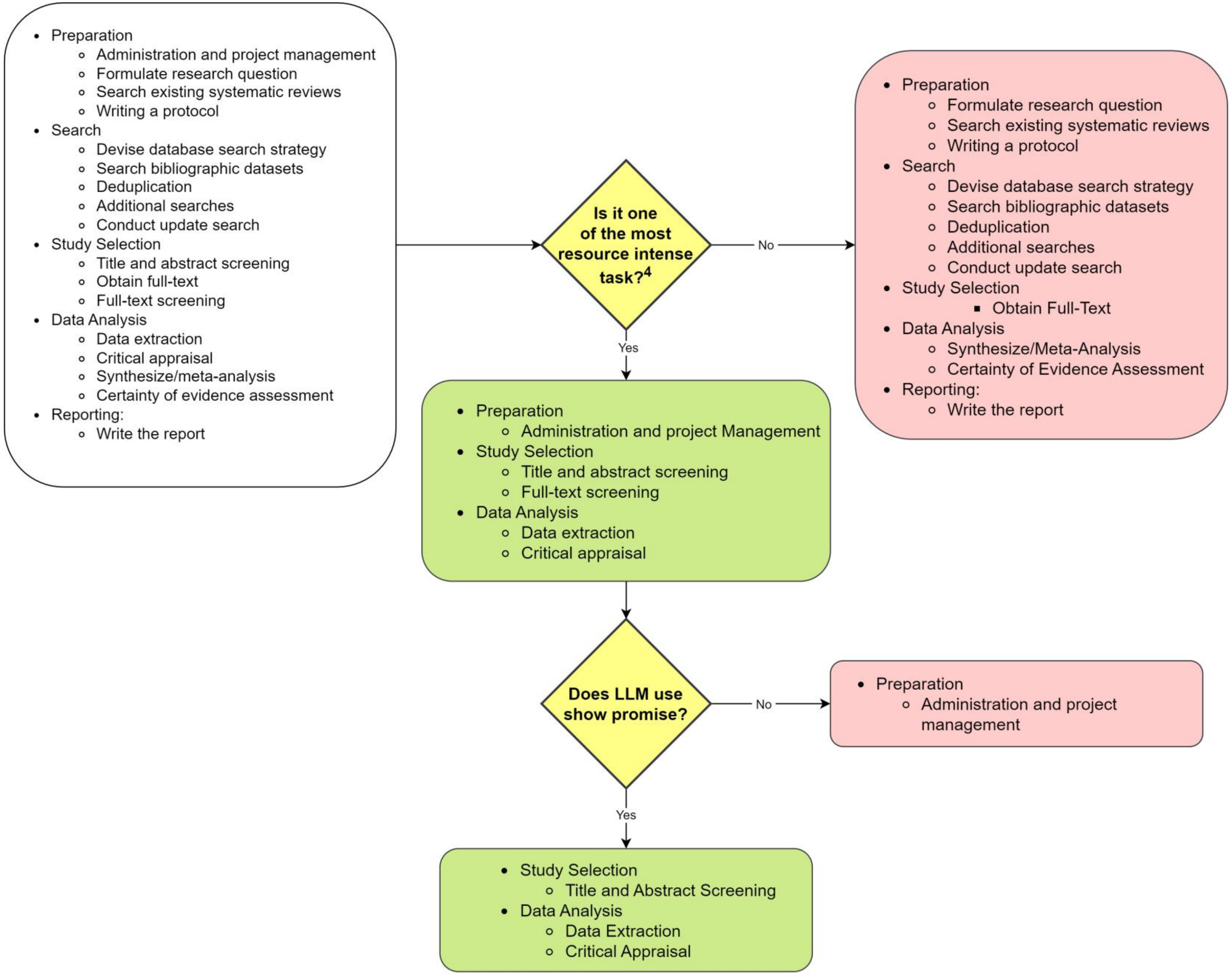
Systematic Review Stages

Preparation	Administration and Project Management	Formulate Research Question	Search existing SRs	Writing a protocol	
Search	Devise database search strategy	Search bibliograhic datasets (n=3)	deduplication	additional searches	Conduct update search
Study Selection	Title and Abstract Screening (n=9)	Obtain full texts	Full Text Screening (n=3)		
Data Analysis	Data Extraction	Critical appraisal (n=3)	Synthesize/meta-analisis	Certainty of evidence assessment	
Reporting	Write the report				

Steps for the systematic review process as defined by Nussbaumer-Streit et al, 2021[4]

Number of LLM evaluations according to Clark et al[5]

Prioritisation of SR Tasks



Systematic Review Stages

Preparation	Administration and Project Management	Formulate Research Question	Search existing SRs	Writing a protocol	
Search	Devise database search strategy	Search bibliograhic datasets	deduplication	additional searches	Conduct update search
Study Selection	Title and Abstract Screening	Obtain full texts	Full Text Screening		
Data Analysis	Data Extraction	Critical appraisal (RoB analysis)	Synthesize/meta-analisis	Certainty of evidence assessment	
Reporting	Write the report				

elias.sandner@cern.ch| kavitaukothari@hotmail.com

STEPS TO BUILD A DATASET:

1. Define required and optional metadata
2. Which metadata could cause copyright issues
3. Define how data should be stored
   a. Define data tables / file structure
   b. XML, CSV, SQL+API-Service
4. Hosting platform (Zenodo, OSF, Github or SQL Database)

Review Metadata
DOI
Title
Abstract
Keywords

Candidate Studies
DOI
Title
Abstract
TiAb Label Screener 1
TiAb Label Screener 2
TiAb consensus
FT Label Screener 1
FT Label Screener 2
FT Label consensus
Authors
Publisher
Publication Date
Full-Text
Publisher-ID
Keywords
Publication Date

Eligibility Criteria
Raw Eligibility Criteria
Inclusion Criteria <List>
Exclusion Criteria <List>

Step 3: Existing Datasets

- Collect existing datasets
- Evaluate quality/format of existing datasets
- Evaluate feasibility to transition to defined format
- Decide if they are outdated (contamination)

Status In progress

Data sets for systematic review automation evaluation

Introduction

Purpose

To develop a list of existing data sets available for automation evaluation (benchmarking) in systematic reviews (SRs)

Problem statement

There are several data sets already existing for use of automation evaluation yet these are scattered on different platforms with different formats. As the first step to streamline what is already available, we propose to list all the sources in one document.

Please add any datasets that can be used for evaluating systematic review automation. Please share any datasets, regardless of whether or not you have checked the reliability of the dataset.

Name of Dataset	SR stage	Link to Dataset	Shared by

Datasets shared by community

- Search: 1
- Data Extraction: 2
- Quality Assessment: 2
- Literature Screening: 13
- Collections of Reviews: 3

Google Sheet

Name of Dataset	SR stage	Link to Dataset	Shared by
Synergy	Literature Screening	https://github.com/asreview/synergy-dataset	Elias Sandner, Kavita Kothari
CSMeD	Literature Screening	https://github.com/WojciechKusa/CSMeD-baselines	Elias Sandner, Kavita Kothari
Health Attribution Database	Topic Database	https://www.healthattribution.org/database	Tim Repke
CLEF TAR	Search	https://github.com/CLEF-TAR/tar	Pawel Jemiolo
PIK-HIC (climate & health, impacts of climate change)	Literature Screening & Abstract Coding	TBD Partially available on https://climateliterature.org/#/project/climatehealth and as part of lancet countdown 5.3.1 / 5.3.2	Max Callaghan / Tim Repke
Carbon pricing effectiveness review	Literature screening & full-text coding	TBD	Klaas Miersch
NIEHS	TBD	Not available online anymore; DESTinY has a copy	James Thomas (?)
Map of carbon dioxide removal	Literature Screening & Abstract Coding	Published soon: https://www.researchsquare.com/article/rs-4109712/v1 Some data on: https://climateliterature.org/#/project/cdrmap	Tim Repke / Sarah Lück
Sustainable Development Goals (UNEP SDG Synthesis coalition)	TBD	https://www.unevaluation.org/repository/member-publications?tab=2	Diana Danilenko / UNEG & SDG Synthesis Coalition

Contamination

Problem:

- Content of evaluation datasets (the underlying systematic reviews) may have been part of the training data for LLMs
- It can not be guaranteed that the LLM is actually **reasoning** - it may just "**remember**" the correct answer

Solution 1:

Retrospective evaluation using systematic reviews published after training cut-off

- Difficult to build large dataset and use the latest LLM models

Solution 2:

Prospective evaluation (expensive) & time lag

Reasoning or memorisation?

Company	Model	Training cut-off date
Anthropic	Claude 4 Opus	January 2025
Meta	LLAMA 4	August 2024
OpenAI	GPT 4.1	June 2024
OpenAI	GPT 4	September 2021

LLM Training Cut-Offs[6]

OpenAI Models

Model Name	Company	Cut-off Date	Source
GPT-1	OpenAI	2018.10	Source
GPT-2	OpenAI	2019.11	Source
GPT-3	OpenAI	2020.10	Source
GPT-3.5*	OpenAI	2021.09	Source
GPT-4*	OpenAI	2021.09	Source
GPT-4-turbo (2024-04-09)	OpenAI	2023.12	Source
GPT-4o (2024-05-13)	OpenAI	2023.10	Source
GPT-4o mini (2024-07-18)	OpenAI	2023.10	Source
GPT-4o-realtime-preview (2024-10-01-preview)	OpenAI	2023.10	Source
GPT-4.1	OpenAI	2024.06.01	Source
GPT-4.1-mini	OpenAI	2024.06.01	Source
OpenAI o1-preview (2024-09-12)	OpenAI	2023.10	Source
OpenAI o1-mini (2024-09-12)	OpenAI	2023.10	Source
o1	OpenAI	2023.10.01	Source
o1-pro	OpenAI	2023.10.01	Source
o3	OpenAI	2024.06.01	Source
o3-mini	OpenAI	2023.10.01	Source
o3-pro	OpenAI	2024.06.01	Source
o4-mini	OpenAI	2024.06.01	Source

Anthropic Models

Model Name	Company	Cut-off Date	Source
Claude Instant 1.2	Anthropic	2023.01	Source
Claude 2	Anthropic	early 2023	Source
Claude 2.1	Anthropic	2023.01	Source
Claude 3 Opus	Anthropic	2023.08	Source
Claude 3 Sonnet	Anthropic	2023.08	Source
Claude 3 Haiku	Anthropic	2023.08	Source
Claude 3.5 Sonnet	Anthropic	2024.04	Source
Claude 3.5 Haiku	Anthropic	2024.07	Source
Claude 3.7 Sonnet	Anthropic	2024.11	Source
Claude 4 Opus	Anthropic	2025.03	Source
Claude 4 Sonnet	Anthropic	2025.03	Source

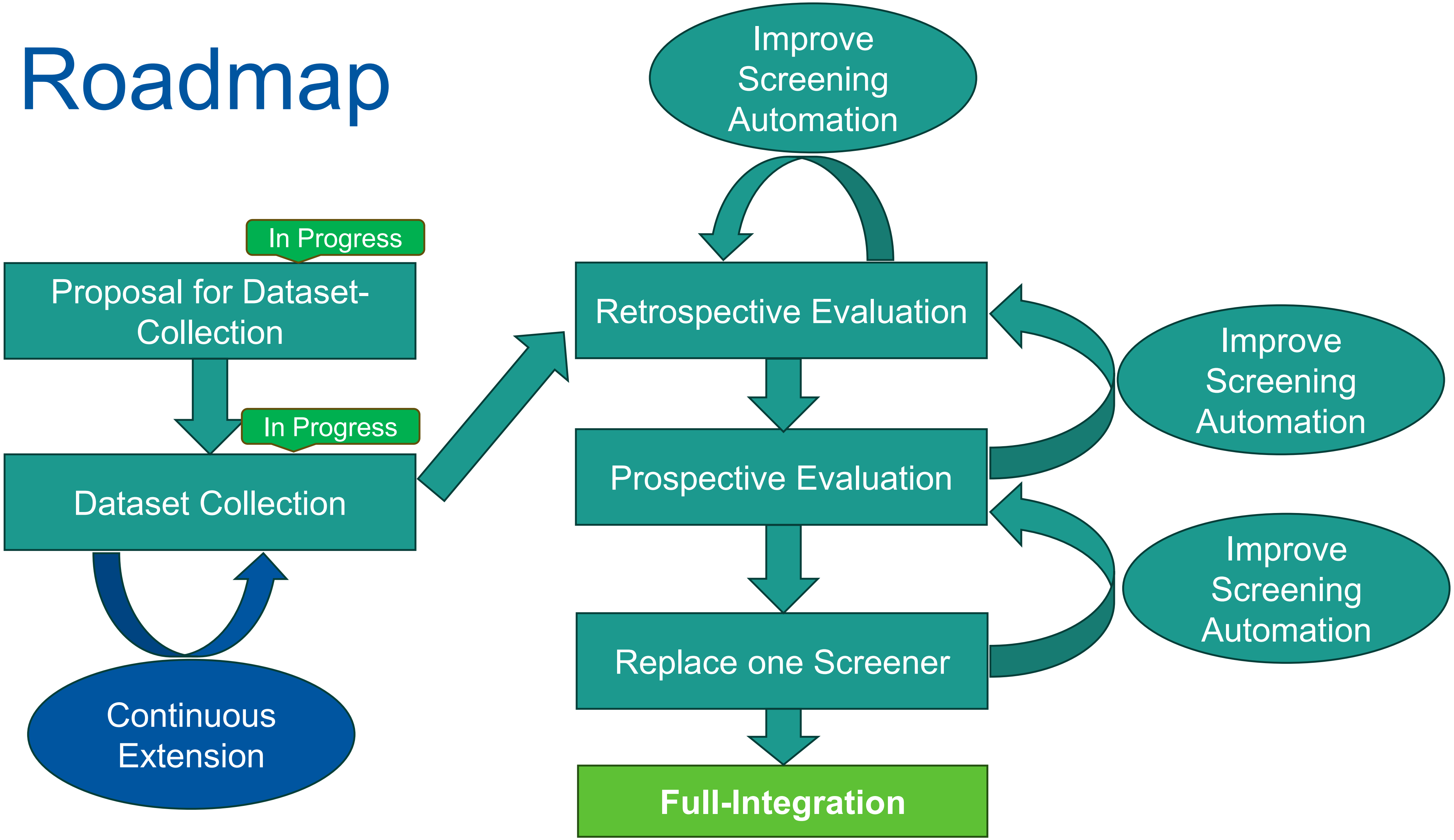
Meta Models

Model Name	Company	Cut-off Date	Source
LLama-2-7B,13B,70B	Meta	Pretraining 2022.09, Finetuning 2023.07	Source
LLama-3-7B	Meta	2023.03	Source
LLama-3-70B	Meta	2023.12	Source
Llama-3.1-8B	Meta	2023.12	Source
Llama-3.1-70B	Meta	2023.12	Source
Llama-3.2-1B	Meta	2023.12	Source
Llama-3.2-3B	Meta	2023.12	Source
Llama-3.3-70B	Meta	2023.12	Source
Llama-4-Scout (17Bx16E)	Meta	2024.08	Source
Llama-4-Maverick (17Bx128E)	Meta	2024.08	Source

Google Models

Model Name	Company	Cut-off Date	Source
Gemini 1.0 Pro	Google	2023.02	Source
Gemini 1.5 Pro	Google	2024.05	Source
Gemini 1.5 Flash	Google	2024.05	Source
Gemini 2.0 Flash	Google	2024.06	Source
Gemini 2.0 Flash Thinking	Google	2024.05	Source
Gemini 2.0 Flash-Lite	Google	2025.01	Source
Gemini 2.0 Pro Experimental	Google	2025.01	Source
Gemini 2.5 Flash	Google	2025.01	Source
Gemini 2.5 Pro	Google	2025.01	Source

Roadmap



elias.sandner@cern.ch| kavitaukothari@hotmail.com

Step 4: Strategy for adding new Data

- Create guidance on how to transfer, format data from current tools into an acceptable format for LLM evaluation.
- Allow researchers to submit their data
- Quality control of submitted data sets
- Mechanism to remove outdated sets or archive outdated sets. (when is a data set too old and cause contamination if used for evaluation)
- Add to dataset collection

DEST Hackathon: Data Sharing Circle

Gain consensus on:

- Priority SR tasks
- Metadata
- File Structure
- ...

Discuss:

- How to build on existing datasets
- Legal limitations
- Data contamination
- Dataset Maintenance

Deliverable: Proposal for addressing data scarcity

- Roadmap
- Resource estimation

elias.sandner@cern.ch| kavitaukothari@hotmail.com

References

1. Sandner, E., Hu, B., Simiceanu, A., Fontana, L., Jakovljevic, I., Henriques, A., ... & Gütl, C. (2024, November). Screening Automation for Systematic Reviews: A 5-Tier Prompting Approach Meeting Cochrane's Sensitivity Requirement.
2. Thomas J, Flemyng E, Noel-Storr, A. et al. Responsible AI in Evidence Synthesis (RAISE): guidance and recommendations (version 2; updated 3 June 2025). In: Open Science Framework [https://osf.io/], Washington DC: Center for Open Science. DOI 10.17605/OSF.IO/FWAUD (accessed 08/07/2025).
3. Sandner, E., Fontana, L., Kothari, K., Henriques, A., Jakovljevic, I., Simniceanu, A., Wagner, A., Gütl and C. (2025). Evaluating Large Language Models for Literature Screening: A Systematic Review of Sensitivity and Workload Reduction. In Proceedings of the 14th International Conference on Data Science, Technology and Applications - Volume 1: DATA; ISBN 978-989-758-758-0; ISSN 2184-285X, SciTePress, pages 508-517. DOI: 10.5220/0013562900003967
4. Nussbaumer-Streit B, Ellen M, Klerings I, Sfetcu R, Riva N, Mahmić-Kaknjo M, Poulentzas G, Martinez P, Baladia E, Ziganshina LE, Marqués ME, Aguilar L, Kassianos AP, Frampton G, Silva AG, Affengruber L, Spjker R, Thomas J, Berg RC, Kontogiani M, Sousa M, Kontogiorgis C, Gartlehner G; working group 3 in the EVBRES COST Action (https://evbres.eu). Resource use during systematic review production varies widely: a scoping review. J Clin Epidemiol. 2021 Nov;139:287-296. doi: 10.1016/j.jclinepi.2021.05.019. Epub 2021 Jun 4. PMID: 34091021.
5. Clark J, Barton B, Albarqouni L, et al. Generative artificial intelligence use in evidence synthesis: A systematic review. *Research Synthesis Methods*. 2025;16(4):601-619. doi:10.1017/rsm.2025.16
6. Wang, H. (2023). *LLM Knowledge Cutoff Dates*. GitHub repository. Retrieved July 8, 2025, from https://github.com/HaoooWang/llm-knowledge-cutoff-dates

CERN | World Health Organization

collaboration