Synapse: Automated Generation of Systematic Review Protocols and AI-Assisted Screening with Large Language Models

Kevin Washington Azevedo da Cruz
Renato Hidaka Torres
Nelson Cruz Sampaio Neto

[1]Universidade Federal do Pará (UFPA)
Belém – PA – Brazil

kevin.cruz@icen.ufpa.br, renatohidaka@ufpa.br, nelsonneto@ufpa.br

Abstract. *The exponential growth of scientific output makes systematic literature reviews (SLRs) increasingly costly to conduct manually. This paper investigates large language models (LLMs) for methodological planning and study screening in a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)-based workflow. Synapse generates review protocols from a researcher's intent and assigns relevance scores to candidate studies. In an exploratory evaluation, protocol generation reached a mean cosine similarity of 0.60 across three protocol cases, while screening reached mean recall 0.967 and WSS@95 0.853 on three low-prevalence SYNERGY corpora. On the same corpora, Synapse matched ASReview's mean recall (0.967) and obtained higher WSS@95 (0.853 versus 0.791).*

1. Introduction

Global scientific output is growing at an accelerating pace. The annual number of Science and Engineering articles rose from about 2 million in 2010 to 3.3 million in 2022 [National Science Board, National Science Foundation 2023], an increase of approximately 65%. This volume imposes operational challenges on researchers who conduct systematic literature reviews (SLRs). The manual stages of searching, screening, and synthesis, whose reporting is standardized by guidelines such as PRISMA [Page et al. 2021], demand many hours and hinder the auditability of methodological decisions.

Large language models (LLMs) can support systematic reviews beyond relevance classification by generating protocols from the researcher's intent, including research questions, search strings, eligibility criteria, extraction schemas, and quality assessment instruments. The joint integration of these functions in a single workflow remains insufficiently characterized.

Recent studies have evaluated GPT-4 and other general-purpose LLMs for screening and extraction, reporting promising but corpus- and imbalance-sensitive results [Khraisha et al. 2024, Wang et al. 2024, Sanghera et al. 2025]. Active-learning systems such as ASReview also prioritize records through iterative reviewer labels [van de Schoot et al. 2021]. These approaches mainly target screening, whereas Synapse combines protocol generation and screening in one workflow.

This scenario motivates systems that integrate protocol generation and screening assistance in a single SLR workflow. We present *Synapse*, which supports selected SLR stages and adopts PRISMA [Page et al. 2021], a widely used reporting guideline for systematic reviews. The researcher remains responsible for final validation of AI-produced decisions (*human-in-the-loop*).

Accordingly, the contributions are: (i) a generative AI *pipeline* for SLR protocol artifacts ; (ii) a title-and-abstract screening mechanism based on relevance scores ; (iii) an exploratory evaluation over three framework–domain cases ; and (iv) a same-corpus comparison with the ASReview active-learning baseline .

2. Theoretical Background

An SLR identifies, appraises, and synthesizes evidence on a research question through an explicit, reproducible protocol designed to reduce bias and support auditability [Kitchenham and Charters 2007, Page et al. 2021]. The standardized flow of stages of an SLR is illustrated in Figure 1. It is conducted according to guidelines such as those of Kitchenham [Kitchenham and Charters 2007], for Software Engineering, and PRISMA [Page et al. 2021], of cross-domain use. SLR automation concentrates efforts on reducing manual work mainly during the Planning and Screening stages, highlighted in Figure 1.

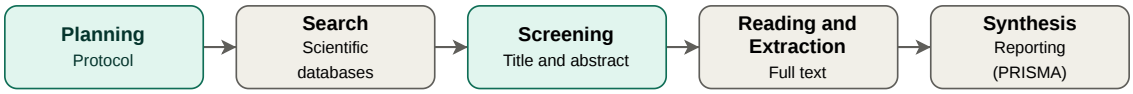


Figure 1. Standard flow of a systematic literature review; the Planning and Screening stages (in green) are those supported by Synapse.

2.1. Planning and SLR protocols

In the planning stage, the researcher defines the scope of the review, and the use of a *framework* is recommended to structure the research question into components. For example, the PICO *framework* decomposes it into four components: *Population*, *Intervention*, *Comparison*, and *Outcome*. There are also variations that add other elements, such as PICOC, which adds *Context*. From these components filled in by the researcher, the *framework* derives the SLR protocol, which comprises the research questions, the search strings, the eligibility criteria, the extraction schema, and the studies' quality assessment instrument.

2.2. Screening and semantic retrieval

In the screening stage, the researcher evaluates the title and abstract of each record retrieved by the searches, deciding on its inclusion or exclusion according to the eligibility criteria. Once screening is complete, semantic retrieval offers additional support over the set of included studies. Vector representations (or *embeddings*) of title and abstract are compared by cosine similarity, locating related studies by conceptual proximity, regardless of the exact terminology. This is an auxiliary resource, originating from the information retrieval field, that is not part of the formal SLR guidelines and that, in this work, is provided by Synapse.

In this context, during screening, an LLM can estimate the relevance of each record from the title and abstract, supporting the inclusion or exclusion decision. Semantic retrieval, in turn, dispenses with the LLM, relying only on *embeddings* and cosine similarity. In both cases, the *human-in-the-loop* paradigm is adopted, in which the researcher validates the outputs, since eligibility decisions directly affect the validity of the review.

2.3. SLR support tools

Several platforms support conducting SLRs with different degrees of AI automation. The comparison among Parsifal, Rayyan, Covidence, and Synapse is presented in Table 1. Parsifal [Freitas 2015] covers all stages of the flow shown in Figure 1, but keeps screening and protocols manual. Rayyan [Ouzzani et al. 2016] offers screening assisted by a classifier trained from the reviewer's own decisions, which requires the initial labeling of part of the records before the model can provide consistent recommendations. Finally, Covidence [Veritas Health Innovation 2023], integrated into the Cochrane ecosystem, provides automated support for randomized-trial screening tasks.

Table 1. Compact comparison of SLR support features .

Feature	Parsifal	Rayyan	Covidence	Synapse
Review workflow support	✓	✓	✓	✓
AI-based protocol generation	–	–	–	✓
AI-assisted screening	–	✓	✓	✓
Semantic retrieval	Not evaluated	Not evaluated	Not evaluated	Implemented

The selection of the compared tools followed adoption and methodological-coverage criteria [Scott et al. 2021, Harrison et al. 2020]. Platform features were checked in May 2026 and may change over time. Parsifal was included because it supports Software Engineering reviews. Synapse differs by integrating generative protocol artifacts with score-based screening without prior reviewer labels.

3. Synapse development methodology

Synapse supports the planning and screening stages of the SLR process through a *human-in-the-loop* workflow, in which the researcher's manual actions are interleaved with AI modules[1]. The usage flow of the proposed tool is illustrated in Figure 2.

Usage begins when the researcher (1) chooses one of the four supported research-question *frameworks* (see Section 3.1) and fills in its components. The Generation Module (2) produces the protocol artifacts. The researcher (3) runs the database-specific search strings and imports the results ; this step remains manual. The Screening Module (4) assigns relevance scores and suggested decisions, which the researcher (5) validates . Finally, the Semantic Retrieval Module (6) optionally locates related studies by conceptual proximity.

[1]Source code, database *dump*, and execution instructions available at `https://anonymous.4open.science/r/synapse-A68E/`.

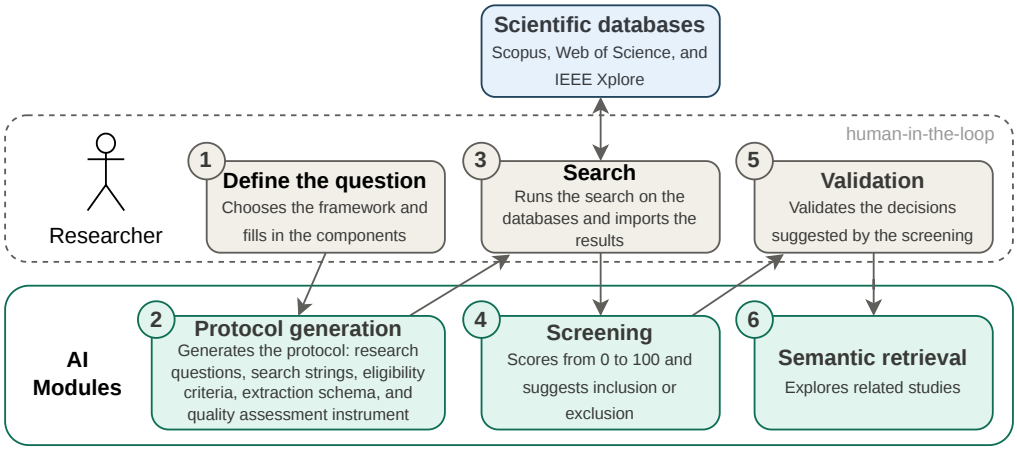


Figure 2. Synapse usage flow: the researcher's manual steps (gray) interleaved with the AI modules (green) and the external scientific databases (blue). The dashed rectangle highlights the *human-in-the-loop* stages.

3.1. Supported Research-Question Frameworks

The choice of the *frameworks* supported by Synapse was grounded in the bibliometric analysis conducted in this work, which collected SLR articles indexed in Scopus between 2020 and 2026[2]. Of the 73 resulting studies, PICO and PICOS were the most frequent, with 28 occurrences each, while PECO and PICOC appeared in 5 each. The counts are framework mentions; seven retrieved studies did not match one of these four categories or were ambiguous.

Therefore, these four *frameworks* were adopted for representativeness and methodological coverage. They span distinct domains, but the present evaluation tests only three of them : PICO (*Population, Intervention, Comparison, Outcome*), aimed at Clinical Medicine; PICOS, which adds *Study Design* and serves Public Health; PECO, which replaces *Intervention* with *Exposure* and is recommended for Environmental Sciences; and PICOC, which adds *Context* and covers Software Engineering. Although Synapse supports PECO, PECO was not included in the present evaluation; therefore, no empirical claim is made about PECO.

3.2. Artificial Intelligence Modules

The AI layer of Synapse is organized into three service modules with delimited responsibilities, all executed on the *backend*. The Protocol Generation and Screening Modules are guided by the DeepSeek LLM [DeepSeek-AI 2024], instructed as described in Section 3.3. The Semantic Retrieval Module, in turn, relies on vector representations (or *embeddings*) produced by SPECTER2 [Singh et al. 2023], a model trained on scientific texts, and on the cosine similarity between these vectors.

Protocol Generation Module. Responsible for step 2 of the usage flow, this module generates structured content from the *framework* components filled in by the researcher. More specifically, it formulates the research questions, creates the search strings,

[2]Search string: `TITLE("systematic literature review" OR "systematic review") AND TITLE-ABS-KEY("PICO" OR "PICOT" OR "PICOS" OR "PICOC" OR "PECO" OR "SPIDER" OR "SPICE" OR "ECLIPSE" OR "PCC") AND PUBYEAR > 2019`.

defines the inclusion/exclusion criteria, dynamically generates the extraction schema, and proposes a quality assessment instrument. The search strings are produced in the format of three repositories: Scopus, Web of Science, and IEEE Xplore.

Screening Module. Responsible for step 4 of the usage flow, this module evaluates the title and abstract of each imported article, assigning a relevance score from 0 to 100 and suggesting an inclusion or exclusion decision. The score is computed by the LLM, instructed to estimate, from the title and abstract, the study's degree of adherence to the research questions and to the inclusion and exclusion criteria. Values close to 100 indicate strong adherence to the inclusion criteria and low adherence to the exclusion criteria. The researcher remains the final validator of the eligibility decisions.

Semantic Retrieval Module. It acts in step 6 of the usage flow, being optionally triggered by the researcher over the imported corpus. It generates vector representations of the included articles and, through cosine similarity, locates related studies by conceptual proximity regardless of exact terminology. This module is implemented but was not empirically evaluated in this study.

3.3. Prompt Engineering Strategy

The experiments used the `deepseek-v4-flash` model through the DeepSeek API in July 2026. The interaction relies on a dual-*prompt* architecture and uses *temperature* $= 0$ to reduce sampling variability; this setting does not guarantee identical responses from a hosted API. Prompts follow distinct patterns according to the stage: planning or screening.

Planning *prompts*. The system message starts with a domain persona (e.g., "expert in bibliographic search strategies") and dynamically injects a semantic instruction according to the selected *framework*, making each component explicit and preventing confusion between close schemas. For PECO, for example, it clarifies that E corresponds to *Exposure*, and not to an intervention. The filled-in components are rendered as structured text in the user message, with the syntax of the search strings adapted to each scientific database (`TITLE-ABS-KEY()` in Scopus, `TS=` in Web of Science, `"All Metadata":` in IEEE Xplore). Finally, the *prompt* instructs the model to respond only with the numbered list, without introductory or explanatory text, so that the response can be read and processed automatically by the system.

Screening *prompts*. Screening uses structured instructions that guide the analysis of title and abstract, the verification of the eligibility criteria, and the assignment of the relevance score. The results are returned in JSON with a pre-specified structure and categorical fields of predefined values, restricted to the available metadata. A validation mechanism applies default values in case of non-conformity.

3.4. Technological Architecture

Synapse uses a React/Vite *frontend*, an asynchronous FastAPI *backend*, PostgreSQL 16 for transactional data, and Qdrant for vector storage. The backend orchestrates DeepSeek through LangChain [DeepSeek-AI 2024, Chase 2022]; SPECTER2 produces 768-dimensional embeddings for semantic retrieval [Singh et al. 2023]. The architectural description is intentionally concise because the evaluation, rather than interface or implementation detail, is the focus of this revision.

4. Experimental Evaluation

The evaluation was conducted to answer two evaluation questions: **(EQ1)** how closely do the automatically generated protocols correspond to manually authored protocols? and **(EQ2)** what is the quality of AI-assisted screening when evaluated as a classifier against known human decisions?

4.1. Adherence of Protocol Generation (EQ1)

Experimental design and metric. To evaluate protocol generation, three published systematic reviews were replicated. The cases were selected to vary the research-question *framework* and knowledge area. Because each framework–domain combination appears once, this is an exploratory analysis and framework, domain, and case-specific effects cannot be separated. SYNERGY was reserved for EQ2 because it provides labeled screening corpora, whereas EQ1 retained published reviews with reconstructable protocol artifacts. No same-model LLM ablation or alternative protocol-generation baseline was run for EQ1. They are: **(P1)** ecosystem services in mountains [Mengist et al. 2020] (PICOC, Environmental Sciences); **(P2)** risk factors of the post-COVID-19 condition [Tsampasian et al. 2023] (PICOS, Medicine); and **(P3)** associations among interoception, vagal tone, and emotion regulation [Pinna and Edwards 2020] (PICO, Psychology).

For each of the three reviews, two projects were created. **Project A** (reference) manually reproduces the published protocol, and **Project B** is generated by Synapse from only the *framework* components filled in by the researcher, automatically producing the remaining artifacts: research questions, search strings, inclusion/exclusion criteria, extraction schema, and quality assessment instrument. To allow inspection of the artifacts and reproduction of the analysis, Projects A and B of each protocol are available in a public repository[3].

The adherence between Projects A and B is measured by element-wise cosine similarity. Each item is converted into a 768-dimensional *embedding* by the multilingual Sentence-BERT model[4] [Reimers and Gurevych 2019], and the pairs are formed by optimal matching with the Hungarian Algorithm [Kuhn 1955]. Project A is treated as a manually reconstructed reference rather than an independently blinded or adjudicated gold standard.

Quantitative analysis. The mean similarity by category and protocol is presented in Table 2. A consistent gradient emerges across the three protocols: the artifacts that express research intent are the most preserved, with research questions (0.74) and inclusion/exclusion criteria (0.65), while the structural artifacts are the most reformulated, with the extraction schema (0.48) and the quality assessment instrument (0.49) in the lowest similarity ranges.

The mean similarity is computed only over matched pairs and must be interpreted jointly with the generated/reference cardinality ratio (G/R). This ratio is not semantic coverage: values below one indicate condensation and values above one indicate expansion . One example is the P1 search strings, which present the highest mean similarity

[3] `https://anonymous.4open.science/r/pa1_results-7305/`

[4] `paraphrase-multilingual-mpnet-base-v2` from the `sentence-transformers` library.

Table 2. Cosine similarity (mean of pairs) and generated/reference cardinality ratio by category and protocol (P1: PICOC/Environmental; P2: PICOS/Medicine; P3: PICO/Psychology).

	P1		P2		P3	
Category	Sim.	G/R	Sim.	G/R	Sim.	G/R
Research questions	0.74	5/6	0.77	4/4	0.69	5/3
Inclusion criteria	0.68	6/8	0.55	5/5	0.70	5/6
Exclusion criteria	0.73	6/9	0.61	6/5	0.60	6/5
Search strings	0.82	3/12	0.53	3/1	0.48	3/5
Extraction schema	0.60	5/14	0.35	4/10	0.47	5/7
Quality assessment	0.54	6/4	0.56	6/3	0.37	7/8

of the evaluation (0.82), but in which the system condensed the twelve reference strings into only three, semantically close ones. In this case, a high similarity coexists with the reduction in the number of items.

A pattern in the opposite direction occurs in the extraction schema. For example, in P2, the mean similarity is low (0.35) and none of the generated fields adequately corresponds to the ten reference fields. While the reference defines bibliographic-methodological fields to extract from each study (for example, the effect measure or the type of diagnostic test), the generation produces fields aimed at answering the research questions.

The two cases show that similarity and correspondence are independent dimensions. A high similarity may coexist with strong condensation of the items, as in the P1 search strings, and a low similarity, with the absence of adequate correspondence, as in the P2 extraction schema. Textual similarity, on its own, does not reveal whether the generated protocol covers the scope of the reference nor whether it adequately fulfills its function. It remains to be seen whether a high similarity corresponds to an adequate protocol.

Illustrative analysis. Three descriptive pairs were selected to show that cosine similarity and substantive adequacy can diverge; they were not independently rated by multiple evaluators. In P2, a research-question pair with similarity 0.87 preserved the same hospitalization/ICU intent. In P1, search strings reached 0.81 similarity but condensed 12 reference strings into three, showing that high similarity does not imply scope preservation. In P3, a quality item with similarity 0.20 replaced a generic criterion with a topic-specific validated-scale criterion, illustrating that low similarity can reflect legitimate reformulation. These examples are illustrative rather than a representative qualitative sample.

4.2. Quality of AI-Assisted Screening (EQ2)

Dataset. To evaluate screening as a binary classifier of records between relevant and irrelevant, the SYNERGY dataset [De Bruin et al. 2023] was used, which gathers 26 published systematic reviews. For each review, a set of records is made available, the articles collected by the search, with title, abstract, and a binary inclusion or exclusion label assigned by the authors. This label is the reference (or *ground truth*) of this evaluation. A

record is considered relevant when it was included by the authors in the published review.

Metrics. In an SLR, only a small fraction of the collected records is relevant. This proportion is denoted by r, that is, the ratio between the number of relevant records and the total number of records in the set. This strong imbalance between the two classes, relevant and irrelevant, makes accuracy misleading, since classifying everything as irrelevant would get most of the records right.

Therefore, *recall* is adopted as the priority metric, which is the fraction of the relevant records that the system retrieves, since the omission of relevant studies constitutes a critical failure. Precision, which is the fraction of the retrieved records that are relevant, is reported as context. We also report WSS@95 (*Work Saved over Sampling* at 95% *recall*) [Cohen et al. 2006], using the standard definition

$$WSS@r = \frac{TN + FN}{N} - (1 - r).$$

At $r = 0.95$, the 0.05 random-sampling reference is subtracted. Consequently, WSS@95 is an offline workload-reduction measure and should not be interpreted as guaranteed safe exclusion .

Selected sets. Three sets from SYNERGY were chosen based on predefined criteria: (i) distinct domains, to mirror the generalization logic of EQ1 ; (ii) low proportion of relevant records ($r \leq 10\%$), a range representative of the imbalance found in systematic reviews and in which WSS@95 is informative; and (iii) eligibility criteria sufficiently detailed in the published text for a faithful reconstruction of the protocol. They are: **Donners_2021** [Donners et al. 2021] (Medicine, 248 records, $r = 5.6\%$), on the pharmacokinetics of the drug emicizumab; **Oud_2018** [Oud et al. 2018] (Psychology, 901 records, $r = 2.2\%$), on psychotherapies for *borderline* personality disorder; and **Hall_2012** [Hall et al. 2012] (Software Engineering, 8,681 evaluated records, $r = 1.2\%$), on fault prediction models.

Evaluation procedure. The system assigns each record a relevance score from 0 to 100, and a cutoff threshold separates the included (score equal to or above) from the excluded (score below). To compute the conventional offline WSS@95 measure, the threshold is selected retrospectively using the test labels: it starts from the highest score and is lowered until recall reaches at least 95%. Since the scores are discrete, the adopted threshold is the first point that satisfies this criterion. This threshold is not available to a reviewer during a new review and is not an online stopping rule.

Results. The test verifies, in each review, how much the system retrieves of the relevant studies (*recall*) and how much reading the reviewer saves with the determined threshold (WSS@95). The results are presented in Table 3 and Figure 3.

Across the three sets, the system missed six relevant articles in total , one in Oud_2018 and five in Hall_ 2012, but also produced 192 false positives. The predicted-inclusion lists contain 44, 84, and 196 records, respectively, for a total of 324 records requiring manual assessment; mean precision is 0.350. The mean WSS@95 of 0.853 corresponds to approximately 0.903 for $(TN + FN)/N$ at the selected retrospective operating points. High recall therefore coexists with a non-negligible false-positive burden, and human validation remains necessary.

Table 3. Performance of AI-assisted screening on the selected sets , with 95% Wilson score confidence intervals for recall and precision .

Set	r (%)	Recall [95% CI]	Precision [95% CI]	WSS@95
Donners_2021 (Medicine)	5.6	1.000 [0.785, 1.000]	0.318 [0.200, 0.466]	0.773
Oud_2018 (Psychology)	2.2	0.950 [0.764, 0.991]	0.226 [0.150, 0.326]	0.857
Hall_2012 (Software Eng.)	1.2	0.952 [0.892, 0.979]	0.505 [0.436, 0.574]	0.928
Mean	–	**0.967**	**0.350**	**0.853**

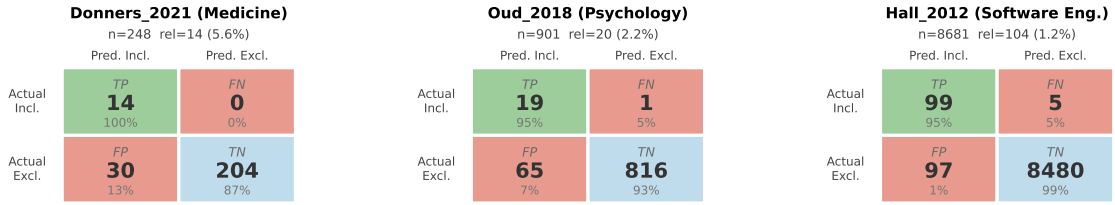


Figure 3. Confusion matrices of screening on the three selected sets. The six false negatives coexist with 192 false positives across the three selected sets .

Statistical analysis. We report 95% Wilson intervals [Wilson 1927]; recall intervals are [0.785, 1.000], [0.764, 0.991], and [0.892, 0.979] for Donners_2021, Oud_2018, and Hall_2012, respectively. With only three datasets, cross-domain significance claims are avoided; paired same-corpus tests are reported below.

4.3. Comparison with Related Approaches

To provide a same-corpus baseline, the three sets were also processed with ASReview [van de Schoot et al. 2021], version 3.0.7, using its default Naive Bayes/TF–IDF configuration, one relevant and one irrelevant prior, and a fixed random seed. Synapse and ASReview reached the same mean recall (0.967), while Synapse obtained a higher mean WSS@95 (0.853 versus 0.791).

Table 4. Screening comparison with ASReview on the same corpora.

	Recall		WSS@95	
Set	Synapse	ASReview	Synapse	ASReview
Donners_2021	1.000	1.000	0.773	0.736
Oud_2018	0.950	0.950	0.857	0.721
Hall_2012	0.952	0.952	0.928	0.915
Mean	**0.967**	**0.967**	**0.853**	**0.791**

Exact paired tests were significant for Oud_2018 and Hall_2012 ($p < 0.001$) but not for Donners_2021 ($p = 0.233$). No separate BM25 baseline was executed, and EQ1 has no same-model LLM ablation; therefore, claims are limited to the same-corpus ASReview comparison and do not establish superiority over all ranking, active-learning, or generation approaches.

Regarding the broader literature, published LLM-assisted screening results vary with corpus, prevalence, prompting strategy, and threshold [Khraisha et al. 2024, Sanghera et al. 2025]. The same-corpus ASReview experiment is therefore the strongest comparison reported here, while external literature figures provide context rather than a head-to-head benchmark.

5. Conclusion

This work presented Synapse, a *human-in-the-loop* system that combines LLM-based SLR protocol generation with score-based title-and-abstract screening. In exploratory EQ1, three framework–domain cases reached mean cosine similarity 0.60, with stronger preservation of research-intent artifacts than structural artifacts. Because each framework and domain is represented by one case, framework, domain, and review-specific effects cannot be separated; the illustrative pairs were not independently rated, and Project A is a reconstruction rather than an independently blinded or adjudicated gold standard.

In EQ2, Synapse reached mean recall 0.967 and WSS@95 0.853 on three low-prevalence SYNERGY corpora, matching ASReview's mean recall and obtaining higher mean WSS@95 (0.853 versus 0.791). The comparison is retrospective and offline because the threshold is selected from labels. The 192 false positives, 324 predicted inclusions, and mean precision 0.350 show that high recall does not remove manual validation.

Other limitations are that PECO and semantic retrieval were implemented but not empirically evaluated; EQ1 lacks a same-model LLM ablation and the study did not execute a separate BM25 baseline; a single hosted LLM model was used.

Future work should enlarge EQ1, evaluate all eligible SYNERGY reviews, assess semantic retrieval, add prospective stopping rules that do not inspect test labels, and replicate the experiments with additional models and independent human assessment.

References

Chase, H. (2022). LangChain. `https://github.com/langchain-ai/langchain`. Acesso em: maio 2026.

Cohen, A. M., Hersh, W. R., Peterson, K., and Yen, P.-Y. (2006). Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*, 13(2):206–219.

De Bruin, J., Ma, Y., Ferdinands, G., Teijema, J., and Van de Schoot, R. (2023). SYNERGY — open machine learning dataset on study selection in systematic reviews.

DeepSeek-AI (2024). DeepSeek-V3 technical report.

Donners, A. A. M. T., Rademaker, C. M. A., Bevers, L. A. M., Schutgens, R. E. G., Fischer, K., Leebeek, F. W. G., and Huitema, A. D. R. (2021). Pharmacokinetics and associated efficacy of emicizumab in humans: A systematic review. *Clinical Pharmacokinetics*, 60(11):1395–1406.

Freitas, F. (2015). Parsifal: A tool to support researchers to perform systematic literature reviews. `https://parsif.al`. Acesso em: maio 2026.

Hall, T., Beecham, S., Bowes, D., Gray, D., and Counsell, S. (2012). A systematic literature review on fault prediction performance in software engineering. *IEEE Transactions on Software Engineering*, 38(6):1276–1304.

Harrison, H., Griffin, S. J., Kuhn, I., and Usher-Smith, J. A. (2020). Software tools to support title and abstract screening for systematic reviews in healthcare: an evaluation. *BMC Medical Research Methodology*, 20(1):7.

Khraisha, Q., Put, S., Kappenberg, J., Warraitch, A., and Hadfield, K. (2024). Can large language models replace humans in systematic reviews? evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Research Synthesis Methods*.

Kitchenham, B. A. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE 2007-001, Keele University and Durham University.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97.

Mengist, W., Soromessa, T., and Legese, G. (2020). Ecosystem services research in mountainous regions: A systematic literature review on current knowledge and research gaps. *Science of the Total Environment*, 702:134581.

National Science Board, National Science Foundation (2023). Publications output: U.S. trends and international comparisons. Science and engineering indicators 2024. nsb-2023-33, National Center for Science and Engineering Statistics (NCSES), Alexandria, VA. Acesso em: jun. 2026.

Oud, M., Arntz, A., Hermens, M. L. M., Verhoef, R., and Kendall, T. (2018). Specialized psychotherapies for adults with borderline personality disorder: A systematic review and meta-analysis. *Australian & New Zealand Journal of Psychiatry*, 52(10):949–961.

Ouzzani, M., Hammady, H., Fedorowicz, Z., and Elmagarmid, A. (2016). Rayyan — a web and mobile app for systematic reviews. *Systematic Reviews*, 5(1):210.

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., and McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*, 372:n160.

Pinna, T. and Edwards, D. J. (2020). A systematic review of associations between interoception, vagal tone, and emotional regulation: Potential applications for mental health, wellbeing, psychological flexibility, and chronic conditions. *Frontiers in Psychology*, 11:1792.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Sanghera, R., Thirunavukarasu, A. J., El Khoury, M., O'Logbon, J., Chen, Y., Watt, A., Mahmood, M., Butt, H., Nishimura, G., and Soltan, A. (2025). High-performance automated abstract screening with large language model ensembles. *Journal of the American Medical Informatics Association*, 32(5):893–905.

Scott, A. M., Forbes, C., Clark, J., Carter, M., Glasziou, P., and Munn, Z. (2021). Systematic review automation tools improve efficiency but lack of knowledge impedes their adoption: a survey. *Journal of Clinical Epidemiology*, 138:80–94.

Singh, A., D'Arcy, M., Cohan, A., Downey, D., and Feldman, S. (2023). SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5548–5566, Singapore. Association for Computational Linguistics.

Tsampasian, V., Elghazaly, H., Chattopadhyay, R., Debski, M., Naing, T. K. P., Garg, P., Clark, A., Ntatsaki, E., and Vassiliou, V. S. (2023). Risk factors associated with post-COVID-19 condition: A systematic review and meta-analysis. *JAMA Internal Medicine*, 183(6):566–580.

van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., and Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, 3(2):125–133.

Veritas Health Innovation (2023). Covidence systematic review software. `https://www.covidence.org`. Acesso em: maio 2026.

Wang, S., Scells, H., Zhuang, S., Potthast, M., Koopman, B., and Zuccon, G. (2024). Zero-shot generative large language models for systematic review screening automation. In *Advances in Information Retrieval (ECIR 2024)*.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.