Active Learning and Large Language Models for Systematic Literature Review Screening: A Comparative Analysis

Seminar Applied Predictive Analytics

Summer Term 2025

Supervised by

Prof. Dr. Stefan Lessmann

Humboldt-Universität zu Berlin

School of Business and Economics

Chair of Information Systems

by

Baris Sen (Student no. 628530)

Foudil Chebbour (Student no. 643245)

Jeong Eun Park (Student no. 607172)

Zoë Piemontese-Fischer (Student no. 617563)

Berlin, 31st of August 2025

Abstract

Systematic literature reviews (SLRs) constitute the gold standard for evidence synthesis, yet they remain labor-intensive and prone to human error during screening (Shemilt et al., 2016; Torres-Carrion et al., 2024). Active learning (AL) approaches have demonstrated substantial efficiency gains, typically reducing manual review labour by 50-95% while maintaining high recall (Yu & Menzies, 2019; Miwa et al., 2014). Meanwhile, large language models (LLMs) offer promising capabilities for screening with minimal configuration (Qureshi et al., 2023; Kusa et al., 2023). This study compares AL against LLM-based approaches using four LLMs (Gemini-2.0-Flash, Gemini-2.5-Flash, HU3, Llama-3.3-70B) and three traditional ML models (Random Forest, Naive Bayes, Logistic Regression) on two biomedical datasets from Nelson et al. (2002) and Cohen et al. (2006). Active Learning achieves highest average performance ($F_{0.5}$ = 0.271), outperforming few-shot methods ($F_{0.5}$ = 0.236), though Gemini 2.0 Flash reached the peak individual score ($F_{0.5}$ = 0.256). Traditional models show high recall (>0.95), but poor precision (≈0,21), while LLMs achieve better balance. Results suggest complementary deployment strategies.

KEYWORDS: Systematic Literature Review, Automation, Active Learning, Large Language Models

Table of Contents

1. Introduction

Systematic literature reviews represent the cornerstone of evidence-based practice across multiple scientific disciplines, from medicine and psychology to computer science and environmental studies (Kitchenham & Charters, 2007; Petticrew & Roberts, 2006). However, the title and abstract screening constitutes a significant operational challenge, often requiring weeks to months of expert reviewer time (Borah et al., 2017). Studies indicate that these screening phases can consume 25-50% of total review time, with larger reviews examining thousands of citations (Shemilt et al., 2016; Beller et al., 2013). Despite employing dual-reviewer protocols to ensure comprehensiveness, relevant citations occasionally escape detection, while single-reviewer workflows exhibit miss rates of 5-15% (Waffenschmidt et al., 2019).

Active learning systems address this inefficiency by implementing dynamic re-ranking mechanisms that surface relevant studies substantially earlier than random ordering approaches (Miwa et al., 2014). Empirical studies demonstrate that AL can reduce screening workload by 40-95% while maintaining sensitivity above 95% (Yu & Menzies, 2019; Howard et al., 2016). The approach has shown particular promise in biomedical reviews where class imbalance is severe, with relevant articles comprising 1-10% of retrieved citations (Matwin et al., 2010).

Meanwhile, large language models have emerged as powerful tools throughout the review pipeline, initially for query formulation and data extraction, and increasingly for screening tasks themselves (Qureshi et al., 2023; Syriani et al., 2023). Recent evaluations of LLM screening in English peer-reviewed studies report sensitivity rates of about 0.42, with specificity around 0.92 and overall accuracy of about 0.67 (Syriani et al., 2023). Their advantages include rapid deployment with minimal domain-specific tuning requirements and the ability to process natural language instructions directly (Khraisha et al., 2024).

This study systematically compares Active Learning and LLM-based screening approaches to evaluate their effectiveness under different operational conditions. Using two biomedical datasets from established systematic reviews (Nelson et al., 2002; Cohen et al., 2006), we test multiple models including four LLMs and three traditional ML approaches across varying class imbalance scenarios and prompting configurations. The experimental design employs a unified framework that enables direct performance comparison while maintaining complete methodological transparency through comprehensive tracking of configurations, predictions, and decision processes.

Our work advances the field through three primary contributions: (1) Development of a unified computational framework enabling direct comparison between AL and LLM screening approaches with comprehensive provenance tracking of prompts, predictions, and experimental configurations; (2) Systematic evaluation of zero-shot, few-shot, and hybrid LLM-AL strategies against established classical AL baseline, including analysis of sensitivity to seed composition and textual feature variants; (3) Implementation of a fully auditable experimental methodology that captures complete decision trails for reproducibility and error analysis, addressing growing concerns about reproducibility in ML-assisted systematic reviews.

2. Related Work

The application of active learning (AL) to systematic review screening has advanced considerably over the past two decades. Early deployments established that AL-based tools can cut screening effort while maintaining high recall; for example, Wallace et al. (2012) reported ≈ 40% workload reduction with no missed relevant studies in prospective case studies using Abstrackr (100% sensitivity). Miwa et al. (2014) further compared certainty- and uncertainty-based AL across clinical and public-health reviews, finding that certainty selection and weighting positive instances improve early identification of relevant studies and that topic features can bolster performance on complex, imbalanced corpora. Comparative evaluations by Gates et al. (2019) likewise showed semi-automated AL workflows achieving ≥95% sensitivity with ~40% workload reduction in practice, underscoring AL's real-world utility. The CLEF eHealth evaluation campaigns (2017–2018) extended this line of work by creating standardized benchmark tasks for Technology-Assisted Reviews in Diagnostic Test Accuracy domains, enabling consistent comparative evaluation of AL-based and other automated screening methods (Kanoulas et al., 2017; Suominen et al., 2018).

The emergence of large-scale pre-trained models has transformed natural language processing capabilities across domains (Min et al. 2021). Early applications in systematic reviews focused on query expansion and study categorization (Chen & Zhang, 2025). Recent work has explored direct screening applications using models like GPT-3/4, BERT, and domain-specific variants (Syriani et al., 2023).

Studies report varying effectiveness depending on domain characteristics and prompt design. Khraisha et al. (2024) found GPT-4's sensitivity ranged from 38% to 100% with generally high specificity (>80%) when applied to systematic review screening, with performance strongly influenced by dataset imbalance and prompt reliability, while Syriani et al. (2023) observed high sensitivity (75–90%) but low precision (15–20%) for software engineering reviews. The consensus emphasizes that progress in automating systematic reviews depends on validated datasets, interoperability across tools, and clear quality criteria (O'Connor et al., 2017).

Recent work has proposed integrating LLM predictions with AL to enhance Technology-Assisted Review. Bron et al. (2024) introduced a combined method that leverages LLM classifications alongside AL's iterative retraining to optimize document screening. In their case study, this approach achieved 96.7% recall, outperforming LLM-only screening and saving more work than leading AL systems like AutoTAR (ibid). The findings suggest that LLM integration can accelerate relevant document retrieval while maintaining AL efficiency, though evidence is still limited to single-domain evaluations and broader comparisons remain needed. This shows potential for further testing a hybrid approach combining AL with LLMs for Systematic Literature Review.

Reproducibility challenges in ML-assisted systematic reviews have received increasing attention (Bannach-Brown et al., 2019; Thomas et al., 2013). The CLEF eHealth evaluation campaigns established standardized benchmark tasks and metrics, enabling comparative evaluation across diverse biomedical topics but also highlighting that many studies rely on custom datasets with limited generalizability (Kanoulas et al., 2017; Suominen et al., 2018). Recent work emphasizes the need for transparent reporting of experimental configurations, dataset characteristics, and complete audit trails to ensure reproducibility and facilitate trustworthy deployment of automation tools (Marshall & Wallace, 2019; Shemilt et al., 2016).

3. Methodology

3.1. Experimental Datasets

Our empirical testing employs two established benchmark corpora from the Synergy dataset collection, an open-source resource of labeled records from systematic reviews that is widely

used in screening automation research and integrated into tools such as ASReview (De Bruin et al., 2023).

This Nelson (2002) dataset contains 368 biomedical titles and abstracts from a comprehensive systematic review examining postmenopausal hormone replacement therapy and health outcomes (Nelson et al., 2002). The dataset exhibits significant class imbalance with 80 relevant articles (21.74% prevalence) and 288 irrelevant articles, representing the challenging conditions typical of biomedical systematic reviews. The original review applied rigorous inclusion criteria focusing on randomized controlled trials and observational studies with specific outcome measures, providing high-quality ground truth labels for evaluation.

The Cohen (2006) Antihistamines Dataset comprises 310 citations focusing on antihistamine-related studies, originally compiled for citation classification method evaluation. The dataset presents extreme class imbalance with only 16 relevant articles (5.16% prevalence) and 294 irrelevant citations, creating substantial methodological challenges due to sparse positive examples and ambiguous inclusion boundaries. This corpus represents one of the most challenging scenarios in screening automation due to its severe imbalance and limited discriminative features.

Both corpora form part of the Synergy collection's standardized benchmarks, containing PubMed IDs (PMIDs), DOIs, OpenAlex identifiers, and binary inclusion labels (label_included: 0/1) as assigned by domain experts during the original systematic reviews (De Bruin et al., 2023). The datasets maintain linkages enabling retrieval of titles and abstracts through external databases when needed for experimental processing. The datasets reflect realistic class imbalance characteristics encountered in biomedical systematic reviews, where relevant articles typically comprise about 5% of retrieved citations, with the Cohen dataset representing an extreme case of this imbalance (Lanera et al. 2019).

3.2. Screening Methodologies

We compare two screening approaches within a unified modular framework. Active Learning employs an iterative strategy with simulated feedback, while Low-Shot Learning uses static single-pass prompting to classify the entire inference set.

The Active Learning approach begins with a certain number of examples of the positive class and/or the negative class and iteratively expands the labeled set through strategic candidate selection. At each iteration, models retrain on all accumulated labels and identify the next

candidate using predict_next(), which returns the highest predicted positive probability for classical models or the first LLM-identified positive item under the configured schema. Feedback simulation queries dataset gold labels and immediately appends them to the training set. The process continues until a conservative stopping criterion is met: when 5% of the inference pool receives negative labels, remaining items in the current batch are auto-assigned negative labels.

In contrast, Low-Shot Learning performs a single conditioning step, either few-shot with provided examples of both classes or zero-shot with examples of only one class (the positive class in this case), which is followed by prediction across the entire inference set without further iteration or feedback incorporation.

Both approaches utilize a common technical architecture where all models implement APABaseModel with standardized train(), predict(), and predict_next() methods. Classical baselines (Naive Bayes, Logistic Regression, Random Forest) employ joint-corpus TF-IDF vectorization with optional text preprocessing, while LLMs extend BaseLLM to support structured prompting with JSON output validated against Pydantic schemas. We evaluate three prompt formatting variants: TOKEN, ID, and ID_TOKEN.

3.3. Evaluation Metrics and Experimental Design

Our evaluation centers on $F_{0.5}$ (precision-weighted) as the primary metric, supplemented by precision, recall, F1, and accuracy. Statistical significance testing uses paired t-tests on $F_{0.5}$ scores across matched experimental configurations, without bootstrap confidence intervals.

The experimental design manipulates five key factors: screening approach (Active vs. Low-Shot), initial seed composition, textual feature sets (title+abstract with optional keywords), prompt formats (TOKEN, ID, ID_TOKEN), and model families (LLMs vs. classical baselines). Each experimental run persists complete configuration details, prompt templates, and per-record predictions to SQLite databases for full reproducibility. Classical models use random_state=42 for consistency, though no global framework seed is imposed. All feedback remains simulated through ground-truth label queries, with configurable batch processing and delay parameters defaulting to full batching with zero delays.

4. Results

Our systematic evaluation across two benchmark datasets reveals distinct performance patterns about LLM superiority in screening tasks. While Gemini 2.5 Flash achieved the highest individual $F_{0.5}$ score (0.28) under ZeroShot conditions, overall approach comparisons show Active Learning maintaining the strongest average performance ($F_{0.5}$ = 0.271) compared to ZeroShot (0.250) and FewShot (0.236) methods.

Model-specific results demonstrate more nuanced patterns than expected. Gemini-2.0-Flash emerged as the most consistent performer with an average $F_{0.5}$ score of 0.256, followed closely by Random Forest (0.252) and HU3 (0.252). Traditional machine learning approaches proved surprisingly competitive, with Random Forest matching several LLM variants and significantly outperforming Gemini-2.5-Flash (0.242) and Llama (0.226). Statistical testing confirmed highly significant differences between Active Learning and ZeroShot approaches ($p < 0.001$), though Active vs. FewShot ($p = 0.109$) and FewShot vs. ZeroShot ($p = 0.301$) comparisons revealed no meaningful performance distinctions.

Dataset characteristics fundamentally shaped automation effectiveness, with the Nelson corpus (21.7% positive rate) enabling substantially higher performance across all approaches (mean $F_{0.5}$ = 0.247) than the Cohen antihistamines dataset (5.16% positive rate, mean $F_{0.5}$ = 0.072). This 3.4-fold performance difference underscores how class imbalance and feature discriminability constrain screening automation potential, regardless of methodological sophistication.

Prompt engineering emerged as a critical optimization factor, with structured ID_TOKEN formats achieving superior performance ($F_{0.5}$ = 0.271) over single-mode ID-only or TOKEN-only approaches (both 0.245). Similarly, focused title and abstract combinations (0.261) consistently outperformed keyword-enhanced prompts (0.222), suggesting that LLMs benefit from concentrated, high-quality textual input rather than expanded feature sets that may introduce noise.

Class imbalance sensitivity patterns varied markedly between approaches. ZeroShot methods performed optimally with minimal positive examples (1-3 instances) but deteriorated when additional positive seeds were introduced, suggesting that these approaches rely on distributional assumptions aligned with naturally sparse positive rates. Active Learning maintained stable performance across varying imbalance conditions, indicating greater robustness to initial seed composition and supporting its utility in diverse screening contexts.

5. Discussion

Our comparative evaluation demonstrates that screening methodology selection depends critically on dataset characteristics, implementation constraints, and workflow requirements. The results contradict assumptions about universal LLM advantages, showing that iterative active learning often outperforms static prompting despite sophisticated pre-trained models. Active learning's adaptive selection mechanism creates more efficient learning curves than approaches relying on fixed example sets.

However, the findings reveal distinct operational profiles across method types. LLMs achieve higher precision suitable for time-constrained scenarios where false positives are costly, while traditional ML methods deliver extremely high recall (>0.95) but poor precision, increasing reviewer workload through false positives. This trade-off suggests complementary rather than competitive deployment: LLMs for initial precision-oriented filtering, traditional ML for comprehensive high-recall sweeps.

Dataset characteristics fundamentally constrain automation effectiveness regardless of methodological sophistication. The Cohen dataset's severe class imbalance (5.16% positive rate) and sparse discriminative features created substantially greater challenges than the more balanced Nelson dataset (21.7% positive rate), producing a 3.4-fold performance difference. This variation underscores that automation effectiveness cannot be assumed transferable across review types without considering underlying data characteristics.

Prompt engineering emerges as a critical optimization factor extending beyond technical implementation. ID_TOKEN formats outperformed single-mode approaches, while title+abstract combinations exceeded keyword-enhanced prompts, indicating that LLMs benefit from focused, structured input rather than expanded feature sets. This sensitivity suggests that LLM screening systems require domain-specific optimization rather than universal implementations.

Several factors constrain result interpretation. Ground truth labels remain subject to 5-15% inter-rater variability, experimental design choices may influence comparative outcomes, and both datasets represent biomedical reviews with limited domain generalizability. The evaluation timeframe (2025) captures a specific point in rapidly evolving LLM development.

6. Conclusions

Our systematic evaluation of active learning versus LLM-based approaches for literature screening demonstrates complex performance patterns that resist simple generalizations about methodological superiority. While Active Learning achieved the highest average $F_{0.5}$ score (0.271) across approaches, individual model performance varied considerably, with Gemini 2.5 Flash under ZeroShot conditions reaching the peak individual score (0.28).

Model rankings reveal competitive performance across different architectures. Gemini-2.0-Flash achieved the highest average $F_{0.5}$ score (0.256), followed closely by Random Forest (0.252) and HU3 (0.252). This finding challenges assumptions about LLM dominance, demonstrating that traditional machine learning approaches remain viable alternatives. Statistical testing confirmed significant differences between Active Learning and ZeroShot methods ($p < 0.001$), though Active versus FewShot comparisons showed no significant differences ($p = 0.109$).

The performance characteristics reveal distinct operational profiles suited to different screening phases. LLMs demonstrate balanced precision-recall trade-offs appropriate for efficiency-focused screening, while traditional ML approaches exhibit extremely high recall (>0.95) but poor precision, making them suitable for comprehensive coverage scenarios where missing relevant studies carries high costs.

Dataset characteristics fundamentally constrain automation effectiveness regardless of methodological sophistication. The Cohen (2006) dataset's severe class imbalance (5.16% positive rate) and sparse discriminative features produced substantially lower performance (mean $F_{0.5} = 0.072$) compared to the more balanced Nelson (2002) dataset (21.7% positive rate, mean $F_{0.5} = 0.247$). This 3.4-fold performance difference emphasizes that screening automation effectiveness cannot be assumed transferable across review types without careful consideration of corpus characteristics.

Prompt engineering emerges as a critical optimization factor, with ID_TOKEN methods outperforming single-mode approaches (0.271 vs. 0.245) and title+abstract combinations proving more effective than keyword-enhanced inputs (0.261 vs. 0.222). These results indicate that successful LLM deployment requires structured, domain-specific optimization rather than generic implementations.

Future systematic review methodology should embrace strategic deployment of complementary approaches rather than seeking universal solutions. The evidence supports multi-stage workflows employing ZeroShot LLMs for rapid initial filtering, Active Learning for precision-recall optimization, and traditional ML methods for comprehensive final sweeps. This integrated approach leverages the distinct strengths of each methodology while mitigating their respective limitations, representing a promising direction for evidence synthesis acceleration while maintaining methodological rigor essential for reliable systematic reviews.

Reproducibility Appendix

All supporting materials are provided in our GitHub repository. The experimental_env folder contains the testing environment, including a Jupyter notebook that demonstrates the experimental setup and an SQL database with the data from all experiments. The prod_env folder provides a production-ready tool derived from the implementation used in the experimental phase. Finally, the analysis folder includes a Jupyter notebook that explores the results, conducts the analysis, and draws insights from the experiments.

References

Bannach-Brown, A., Przybyla, P., Thomas, J., Rice, A. S., Ananiadou, S., Liao, J., & Macleod, M. R. (2019). Machine learning algorithms for systematic review: Reducing workload in a preclinical review of animal studies and reducing human screening error. *Systematic Reviews, 8*(1), 23. https://doi.org/10.1186/s13643-019-0942-7

Beller, E. M., Glasziou, P. P., Altman, D. G., Hopewell, S., Bastian, H., Chalmers, I., & PRISMA for Abstracts Group. (2013). PRISMA for abstracts: Reporting systematic reviews in journal and conference abstracts. *PLOS Medicine, 10*(4), e1001419. https://doi.org/10.1371/journal.pmed.1001419

Borah, R., Brown, A. W., Capers, P. L., & Kaiser, K. A. (2017). Analysis of the time and workers needed to conduct systematic reviews and meta-analyses. *JAMA Internal Medicine, 177*(7), 1085–1087. https://doi.org/10.1136/bmjopen-2016-012545

Bron, M. P., Greijn, B., Coimbra, B. M., van de Schoot, R., & Bagheri, A. (2024). Combining large language model classifications and active learning for improved technology-assisted review. *CEUR Workshop Proceedings, 3770,* 77–95.

Chen, X., & Zhang, X. (2025). *Large language models streamline automated systematic review: A preliminary study*. arXiv. https://doi.org/10.48550/arXiv.2502.15702

Cohen, A. M., Hersh, W. R., Peterson, K., & 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. https://doi.org/10.1197/jamia.M1929

De Bruin, J., van de Schoot, R., van Aert, R., O'Mara-Eves, A., & Tsafnat, G. (2023). The Synergy dataset: A benchmark collection for research on systematic review automation. *Data in Brief, 49,* 109457. https://doi.org/10.34894/HE6NAQ

Gates, A., Guitard, S., Pillay, J., Elliott, S. A., Dyson, M. P., Newton, A. S., & Hartling, L. (2019). Performance and usability of machine learning for screening in systematic reviews: A comparative evaluation of three tools. *Systematic Reviews, 9*(1), 187. https://doi.org/10.1186/s13643-019-1222-2

Howard, B. E., Phillips, J., Miller, K., Tandon, A., Mav, D., Shah, M. R., & Thomas, R. S. (2016). SWIFT-Review: A text-mining workbench for systematic review. *Systematic Reviews, 5*(1), 87. https://doi.org/10.1186/s13643-016-0263-z

Kanoulas, E., Li, D., Azzopardi, L., & Spijker, R. (2017). CLEF 2017 technologically assisted reviews in empirical medicine overview. *CEUR Workshop Proceedings, 1866,* 1–30. http://ceur-ws.org/Vol-1866/

Khraisha, Q., Put, S., Kappenberg, J., Warraitch, A., & 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, 15*(4), 616–626. https://doi.org/10.1002/jrsm.1715

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering* (EBSE Technical Report EBSE-2007-01). Keele University & University of Durham.

Kusa, W., Mendez, O., Knoth, P., Hanbury, A., & Goeuriot, L. (2023). CLEF-2023 systematic review automation track overview. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023).* CEUR-WS. https://doi.org/10.1145/3578337.3605135

Lanera, C., Berchialla, P., Sharma, A., Minto, C., Gregori, D., & Baldi, I. (2019). Screening PubMed abstracts: Is class imbalance always a challenge to machine learning? *Systematic Reviews, 8*(1), 317. https://doi.org/10.1186/s13643-019-1245-8

Marshall, I. J., & Wallace, B. C. (2019). Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. *Systematic Reviews, 8*(1), 163. https://doi.org/10.1186/s13643-019-1074-9

Matwin, S., Kouznetsov, A., Inkpen, D., Frunza, O., & O'Blenis, P. (2010). A new algorithm for reducing the workload of experts in performing systematic reviews. *Journal of the American Medical Informatics Association, 17*(4), 446–453. https://doi.org/10.1136/jamia.2010.004325

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heinz, I., & Roth, D. (2021). Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A survey. *ACM Computing Surveys, Volume 56, Issue 2.* https://arxiv.org/abs/2111.01243

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. https://doi.org/10.1016/j.jbi.2014.06.005

Nelson, H. D., Humphrey, L. L., Nygren, P., Teutsch, S. M., & Allan, J. D. (2002). Postmenopausal hormone replacement therapy: Scientific review. *JAMA, 288*(7), 872–881. https://doi.org/10.1001/jama.288.7.872

O'Connor, A. M., Tsafnat, G., Gilbert, S. B., Thayer, K. A., & Wolfe, M. S. (2017). Moving toward the automation of the systematic review process: A summary of discussions at the second meeting of ICASR. *Systematic Reviews, 6*(1), 174. https://doi.org/10.1186/s13643-017-0667-4

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide.* Blackwell Publishing. https://doi.org/10.1002/9780470754887

Qureshi, R., Shaughnessy, D., Gill, K. A. R., Robinson, K. A., Li, T., & Agai, E. (2023). Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? *Systematic Reviews, 12*(1), 72. https://doi.org/10.1186/s13643-023-02243-z

Shemilt, I., Khan, N., Park, S., & Thomas, J. (2016). Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. *Systematic Reviews, 5*(1), 140. https://doi.org/10.1186/s13643-016-0315-4

Suominen, H., Kelly, L., Goeuriot, L., Neves, M., Kanoulas, E., Li, D., & Zuccon, G. (2018). Overview of the CLEF eHealth evaluation lab 2018. In *Proceedings of CLEF 2018* (pp. 286–301). Springer. https://doi.org/10.1007/978-3-319-98932-7_26

Syriani, E., David, I., & Kumar, G. (2023). Assessing the ability of ChatGPT to screen articles for systematic reviews. *arXiv Preprint* arXiv:2307.06464.

Thomas, J., Newman, M., & Oliver, S. (2013). Rapid evidence assessments of research to inform social policy: Taking stock and moving forward. *Evidence & Policy, 13*(1), 5–27. https://doi.org/10.1332/174426413X662572

Torres-Carrion, P., González González, C., Aciar, S., & Rodriguez, G. (2018, April 18). *Methodology for systematic literature review applied to engineering and education*. In *2018 IEEE Global Engineering Education Conference (EDUCON)*. IEEE. https://doi.org/10.1109/EDUCON.2018.8363388

Waffenschmidt, S., Knelangen, M., Sieben, W., Bühn, S., & Pieper, D. (2019). Single screening versus conventional double screening for study selection in systematic reviews: A methodological systematic review. *BMC Medical Research Methodology, 19*(1), 132. https://doi.org/10.1186/s12874-019-0782-0

Wallace, B. C., Small, K., Brodley, C. E., & Trikalinos, T. A. (2012). Deploying an interactive machine learning system in an evidence-based practice center: Abstrackr. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium* (pp. 819–824). ACM. https://doi.org/10.1145/2110363.2110464

Yu, Z., & Menzies, T. (2019). FAST2: An intelligent assistant for finding relevant papers. *Expert Systems with Applications, 120,* 57–71. https://doi.org/10.1016/j.eswa.2018.11.021