

ALISE: An Automated Literature Screening Engine for Research

Hendrik Roth¹ ^a and Carsten Lanquillon²  ^b

¹Artificial Intelligence (M.Sc.), Johannes Kepler University Linz, Austria

²Business Information Systems, Heilbronn University of Applied Sciences, Heilbronn, Germany

Keywords: Literature Review, Screening, Automation.

Abstract: The screening process needs the most time of a literature review. An automated approach saves a lot of time, making it easier for researchers to review literature. Most current approaches do not consider the full text for screening, which can cause the exclusion of relevant papers. The Automated Literature Screening Engine (ALISE) performs full-text screening based on a research question about the retrieved papers of the literature search. With an average of 61.87% nWSS and a median of 74.38% nWSS, ALISE can save time for reviewers but cannot be used without human screening afterwards. Furthermore, ALISE is sensitive to the given research question(s).


1 INTRODUCTION


A literature review is a widely used research method intended to provide an overview of previous research, identify new research opportunities, or draw new conclusions from previously unrecognised correlations (Rowe, 2014; Okoli, 2015). However, conducting a literature review with an increasing amount of literature is impractical due to the time-consuming search and screening process (van Dinter et al., 2021). This is primarily because extensive screening is required (van Dinter et al., 2021). Researchers initially retrieve many publications, e.g., from a keyword search often numbering in the hundreds or thousands, making thorough review impractical (Kitchenham and Charters, 2007). Hence, researchers typically rely on titles and abstracts for preliminary screening, adapted from established review frameworks (Kitchenham and Charters, 2007; Page et al., 2021). While this title and abstract screening saves time, it comes with limitations. The shortness of titles and abstracts can lead to the omission of relevant publications and thus excluding papers that address the research questions of researchers (Blake, 2010; Penning de Vries et al., 2020; Wang et al., 2020). This problem is reduced by full-text screening (Penning de Vries et al., 2020). However, it has become more difficult as the literature volume increases continuously. In response, re-

searchers have explored automation to aid literature reviews, employing machine learning algorithms for screening and categorization (Noroozi et al., 2023; van Dinter et al., 2021). Many automated methods still hinge on title and abstract screening (van Dinter et al., 2021), perpetuating the risk of overlooking relevant literature. Large language models (LLMs) can effectively comprehend and respond to text-based queries, even rivaling human performance in some tasks (Ouyang et al., 2022; Liu et al., 2023). This makes them suitable for automating the screening of full-text papers. Especially chaining an LLM can achieve higher results on a downstream task rather than using only one standard prompt (Yu et al., 2023; Haji et al., 2023). Despite the possible benefits, there are currently no studies on applying LLM chains to automated literature reviews. For this reason, this paper addresses this gap, aiming to develop an automated full-text literature screening engine based on a given research question while following established literature review protocol guidelines like (Kitchenham and Charters, 2007). To reach this goal, the paper seeks to answer the following research question: *How can the full-text screening process of a literature review be automated using an LLM chain?*

2 RELATED WORK

There are several studies on automated screening processes for literature reviews. (van Dinter et al., 2021)

^a  <https://orcid.org/0009-0007-2602-9679>

^b  <https://orcid.org/0000-0002-9319-1437>

provide an overview of automated literature review approaches, identifying various studies, which focus on title and abstract screening. Yet, only a few approaches focus on full-text screening. This is because several challenges come with screening over full-texts, e.g., PDF files have to be converted to accessible text (Cohen et al., 2010). However, as also stated by (Portenoy and West, 2020), it is questionable if the returned papers by these methods are actually relevant, or only show strong topic similarities. While considering only keywords or topics to identify relevant studies, full-text screening seems to be worse than only screening abstract and title (Dieste and Padua, 2007). Nevertheless, when defining a relevant paper for a literature review as a paper that addresses a research question (Templier and Paré, 2015), this conclusion cannot be made because, logically, a research question of a reviewer may not necessarily be answered directly by the abstract or title, but instead by paragraphs or sentences of a paper (Blake, 2010; Penning de Vries et al., 2020). Hence, (La Quatra et al., 2021) use a text summarizer and correlation calculations to classify if a cited paper contains relevant information in its full text. (Wilson et al., 2023) compare the effectiveness of regular expression matching and a machine learning classifier that was trained particularly for human screening categorization when it came to automated full-text screening. By employing language models as phrase embeddings, (Alchokr et al., 2022) suggested a different method that involved weighting and clustering the literature according to its relevance. Although the authors' approach is conducted on assessing titles and abstracts, they recognized the potential relevance of this method to full-text analysis, highlighting the need for more research in this field. In a different study, (Noroozi et al., 2023) iteratively classified relevant and irrelevant literature during the systematic search process using a random forest classifier based on full-text feature similarity. The goal of this iterative classification strategy was to enhance the accuracy of the screening process and improve the selection of pertinent publications. There is no study yet that uses LLMs for automating the screening process respecting the full text of a paper and a given research question.

3 BACKGROUND

3.1 Conducting a Literature Review

There are several common literature review methodologies for various domains. For the information sys-

tems domain, the methodology proposed by (Brocke et al., 2009) is a frequently used methodological framework, whereas the framework of (Kitchenham and Charters, 2007) is often utilized in the software engineering domain. PRIMSA by (Page et al., 2021) is often applied in the biomedical domain, and the methodology by (Snyder, 2019) is common for business research. However, they all basically consist of the same general steps, the differences are mainly references to domain-specific journals and quality assessments or more detailed descriptions of some steps (Templier and Paré, 2015). Therefore, (Templier and Paré, 2015) as well as (Okoli, 2015) modeled general steps for a literature review based on these common methodology frameworks. The only difference between (Okoli, 2015) and (Templier and Paré, 2015) is that they switch the general steps 5 and 6 and, furthermore, split the screening process into two steps (an initial title and abstract screening, which is followed by full-text screening). The last step of (Okoli, 2015) can be ignored because it is about writing the review, not conducting the review. Hence, there are 6 general steps for conducting a literature review based on (Okoli, 2015) and (Templier and Paré, 2015). The steps are iterative and can lead to refinement of the previous steps (Brocke et al., 2009; Templier and Paré, 2015). Figure 1 visualises these six general steps.

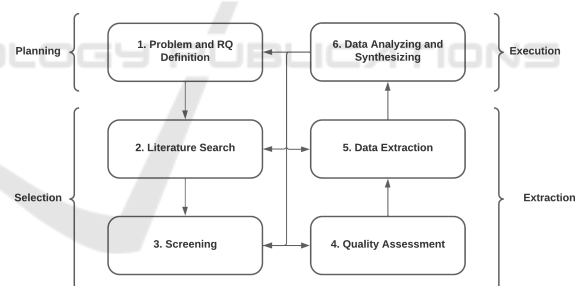


Figure 1: General literature review methodology based on (Templier and Paré, 2015) and (Okoli, 2015).

Step 1 - The first step consists of defining the problem inclusive the research question(s) (Okoli, 2015). (Kitchenham and Charters, 2007) noted, that each literature review must have a research question for guidance of the review. Therefore, this step also includes the definition of the general conditions based on this research question(s) and problem conception, such as the definition of the search terms (Brocke et al., 2009).

Step 2 - After the conceptualisation of the problem and research questions, a search is performed with the defined search terms and filter criteria with the goal to get a literature collection of various litera-

ture databases (Templier and Paré, 2015).

Step 3 - When various papers are retrieved by the search in databases, the literature has to be checked for relevance, also called screening, where the goal is to find papers which helps answer the defined research question (Templier and Paré, 2015). The Screening Process consists typically of initial abstract/title screening to shorten the large volume of retrieved papers, followed by a full-text screening process on the reduced paper corpus (Okoli, 2015; Templier and Paré, 2015).

Step 4 - For each relevant literature, the quality must be assessed (Brocke et al., 2009; Templier and Paré, 2015; Okoli, 2015). Even if a paper is relevant, it can have a low quality and hence must be rejected for inclusion due to quality standards (Okoli, 2015). There are several techniques to assess the quality of a paper (Templier and Paré, 2015).

Step 5 - With the completion of step 4, the reviewers have now a literature corpus for the data extraction related to their research question(s), which represents then the actual findings of the review (Templier and Paré, 2015; Okoli, 2015). The extracted data depends on the study and research question, which then also defines the method which can be used for extracting (Templier and Paré, 2015).

Step 6 - The last step is to analyze and synthesize the extracted data (Okoli, 2015; Templier and Paré, 2015). Typical methods are a concept matrix by (Webster and Watson, 2002) or a table/forest plot as indicated by (Kitchenham and Charters, 2007).

3.2 Explainability

As (Kitchenham and Charters, 2007) and PRISMA by (Page et al., 2021) stated in their methodology, the point of the literature review protocol is to record everything in such a way that it is comprehensible and explainable. For this reason, notes should also be made on relevant papers while screening (Kitchenham and Charters, 2007). By taking notes, researchers can keep track of their thought processes, criteria, and justifications for including or excluding specific papers (Okoli, 2015). Most automated methods do not indicate why a paper is relevant, but just return a corpus labelled as relevant without justification (Portenoy and West, 2020). Thus, ALISE must be able to explain why a paper is relevant, as it can also be done when screened manually.

4 APPROACH

4.1 Problem Definition

As ALISE aims to be integrated into commonly used literature review methodology processes, the screening process can be described similar to the general literature review methodology by (Okoli, 2015) or (Templier and Paré, 2015) and thus covers the methodology processes of (Kitchenham and Charters, 2007), (Brocke et al., 2009), (Snyder, 2019), and (Page et al., 2021). Given our study's focus on full-text screening, we omit the initial abstract and title screening step. Furthermore, this task can be seen as a classification task determining as relevant or not relevant paper (Olorisade et al., 2019), which is also respected by the definitions. Typically, the screening process involves the application of inclusion and exclusion criteria (Templier and Paré, 2015). Exclusion criteria, used to apply automatic filters (e.g., language, article type, date), can be applied during the initial literature database search (Brocke et al., 2009). Quality-related exclusion criteria are assessed during the quality assessment step following the screening process (Kitchenham and Charters, 2007). For this reason, the inclusion criteria considered are only from the content perspective as mentioned by (Okoli, 2015), which is to review if the paper addresses the specific research question(s) (Templier and Paré, 2015). With this context, we describe the screening process as follows:

Let RQ be the given research question. Given an initial set $P = \{p_1, p_2, \dots, p_n\}$ of n papers p_i and RQ , retrieved from an initial search (keyword search, snowballing, etc.), the screening process in a literature review involves checking if a paper addresses the research question and documenting the reasons why a paper is considered relevant. Hereby, $\forall p_i \in P$, the relevance labelling function is defined as follows:

$$f(p_i) = \begin{cases} 1, & \text{if } p_i \text{ addresses } RQ \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Thus, the relevance labels $l_i \in \{0, 1\}$ resulting from f are used for each paper $p_i \in P$. Additionally, there is a need for the review protocol to capture the reasons why each paper is considered relevant (Kitchenham and Charters, 2007). This is defined as a set R consisting of paper-specific reasons $r_i \forall p_i \in P$. Hence, for each paper p_i that addresses the research question, t paper specific reasons $r_i = \{reason_1, reason_2, \dots, reason_t\}$ are assigned to explain its relevance.

To conclude, the screening process involves evaluating each paper p_i and documenting the correspond-

ing relevance reasons r_i if it is relevant. Hence, the results of the screening process are two sets $S = \{p_i \mid l_i = 1, p_i \in P\}$ and $R = \{(p_i, r_i) \mid l_i = 1, r_i \text{ explains relevance of } p_i\}$. The set S represents the subset of papers from P that are relevant, hence which address the given research question RQ (label l is equal to 1). The set R contains pairs of papers p_i and their corresponding relevance reasons r_i . Consequently, during the automated screening process, each paper p_i of P is examined, and if it is found to address the research question RQ , l is set to 1, and a relevance reason r_i is documented in R . By selecting papers based on the value of f and documenting the relevance reasons in R , the review protocol ensures transparency and provides a record of the justification behind the inclusion of each relevant paper that addresses the research question in the literature review. If there are multiple research questions, this procedure will be logically performed for all research questions. As output, S can be used for the quality assessment, which is the next step in the general literature review methodology (Templier and Paré, 2015).

4.2 Technical Details

To assess whether a paper addresses a given RQ , we utilize an LLM chain as described by (Wu et al., 2022), since it has the potential to outperform various classical retriever-reader architectures (Yu et al., 2023). Our chain is inspired by the generate-read chain of (Yu et al., 2023) and the multi-hop QA chain of (Haji et al., 2023) using Flan-t5-XL due to hardware limitations. Thus, the LLM chain with manual prompt templates first generates the evidence E based on the chunks C which serve as an answer to RQ and, then, generates the final answer A using E as context. Here, the chunks C were created by a straight-forward approach. The template length was subtracted from the maximum input token length of 512 to determine the chunk size $l_{chunk} = 512 - l_{template}$. For each sentence, it was checked whether adding the sentence to the current chunk would exceed the token limit in order to avoid truncated sentences. Chunking by logical sections of the paper also seemed intuitive, but handling long sections was challenging and kind of arbitrary in some cases, so we chose the simpler and more straight-forward approach of cutting right before reaching the token limit. The usage of the evidence-answer chain also enables simultaneously getting the reason r_i for a paper when conducting MRC because it generates the evidence for the answers and the answer itself as a reason. This also implicates the labelling function because if the RQ is not answerable by a paper p_i , the LLM chain

returns *unanswerable*. If the answer is not unanswerable, $l = f(p_i) = 1$, otherwise $l = f(p_i) = 0$. When $l = 1$, the reason r_i can be returned by referring to the extracted pieces of evidence related to that question. However, most retrieved papers P from the search are in PDF format (van Dinter et al., 2021), necessitating a PDF-to-text conversion before being used as textual input for the evidence-answer chain. This is a challenge due to the diverse layouts of scientific texts, including multiple columns, different headers and footers, variable abstract positions, and figures and tables affecting text flow (Bast and Korzen, 2017). Addressing these issues, (Tauchert et al., 2020) employed optical character recognition (OCR) to convert scientific PDFs into plain text format. While they used OCR-tesseract, better libraries have emerged, with Grobid being a notable choice as evaluated by (Miah et al., 2022). Grobid is also utilized by the Semantic Scholar Open Research Corpus (Lo et al., 2020), offering both effectiveness and scalability for handling large volumes of scientific papers. For this reason, we chose the s2orc json converter of the Semantic Scholar Open Research Corpus (Lo et al., 2020). Figure 2 visualises the flow of ALISE.

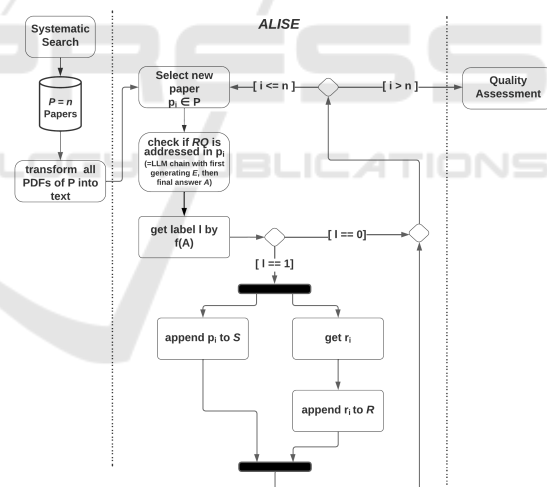


Figure 2: Implementation of ALISE.

5 EVALUATION

5.1 Metrics

ALISEs goal is to assist scientists in the screening process and reduce the time and effort of reviewers. To evaluate its performance, we follow the precedent set by other automated screening approaches against human performance, like (Cohen et al., 2006; Kusa

et al., 2023). We use standard NLP metrics based on the confusion matrix: true positives (TP) for correctly classified papers, false positives (FP) for papers ALISE incorrectly labels as relevant, true negatives (TN) for correctly classified irrelevant papers, and false negatives (FN) for papers ALISE misses. The evaluation metrics equations are provided below:

$$Accuracy(Acc) = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$Precision(Pr) = \frac{TP}{TP + FP} \quad (3)$$

$$Recall(Re) = \frac{TP}{TP + FN} \quad (4)$$

$$F1 = 2 \times \frac{Pr \times Re}{Pr + Re} \quad (5)$$

$$WSS = \frac{TN + FN}{N} - (1.0 - Re) \quad (6)$$

The WSS metric measures the work saved over sampling (Cohen et al., 2006). It represents the ratio of articles initially identified through a literature search that researchers can skip reading because they have already been screened out by ALISE.

$$nWSS = \frac{TN}{TN + FP} \quad (7)$$

The nWSS metric by (Kusa et al., 2023) is the normalized WSS metric to enable better comparisons between different literature reviews, hence not the same reviews must be evaluated as a baseline. Furthermore, the nWSS is equal to the true negative rate (Kusa et al., 2023).

5.2 Dataset

The common dataset of (Cohen et al., 2006) for evaluating automated screening was not used due to its limitation on titles and abstracts, whereas ALISE uses full-texts. In the evaluation of automated screening approaches, researchers often contend with the challenge of manual annotation. Some studies evaluate these approaches based on a single literature review (Noroozi et al., 2023), while others consider two different reviews (Alchokr et al., 2022). Our selection of three literature reviews from random searches on ACM, IEEE, and SpringerLink due to the labor-intensive nature of manual annotation, followed specific criteria. We considered reviews that were peer-reviewed, reproducible (yielding consistent search results with the provided searches), accessible (in literature databases to which we had access), well-documented (with relevant papers clearly listed, such

as in a table), and comprehensible (with well-defined inclusion and exclusion criteria to minimize FP during manual annotation). The following reviews met our specified criteria, while many others were unsuitable due to factors such as irreproducibility, inaccessible databases, or the impracticality of manually downloading thousands of papers. Consequently, our evaluation baseline comprises three literature reviews: literature review 1 (LR1) (Jakob, 2022), literature review 2 (LR2) (da Silva Junior et al., 2022), and literature review 3 (LR3) (Omran and Treude, 2017). Table 1 provides an overview of the evaluated literature reviews regarding the number of papers screened (n) and how many papers are actually relevant.

Table 1: Overview of LRs used for evaluation.

Literature Review	n	relevant
LR1 (Jakob, 2022)	101	60
LR2 (da Silva Junior et al., 2022)	262	6
LR3 (Omran and Treude, 2017)	232	33

5.3 Setup

Manually downloading all papers from the three selected literature reviews was necessary since there are no open API accesses available for obtaining full-text content from SpringerLink, IEEE, and ACM Digital Library or the automation of this task was longer than manually downloading. To save time, and considering that automated downloading was solely for evaluation purposes and not part of the screening process, we opted for manual downloads. The automated screening process ran on an NVIDIA Tesla T4, with no modifications to the quantization of Flan-t5-XL. Where the literature reviews had multiple research questions, one search was performed for each research question, and duplicate results were removed. Each iteration took approximately 45 minutes to two hours, resulting in a total evaluation time ranging from 2.5 to 6 hours, depending on the number of papers evaluated.

5.4 Results

This section presents the evaluation results for each literature review. In the evaluation of the literature

Table 2: Confusion matrix of all literature reviews.

LR	TP	FP	FN	TN
LR1	57	17	1	23
LR2	6	61	0	192
LR2*	10	57	0	192
LR3	33	197	0	2
LR3*	21	54	12	145
LR3**	53	22	12	145

review by (Jakob, 2022) (LR1), out of an initial population of 101 papers, 98 were evaluated due to limited full-text access. ALISE achieved 57 TPs, 17 FPs, 1 FNs, and 23 TNs. See table 2 for the confusion matrix values. For the literature review by (da Silva Junior et al., 2022) (LR2), which initially screened on title and abstracts, 67 papers were classified as relevant by ALISE. Subsequently, two independent reviewers of the related domain manually reviewed the FPs of the first evaluation LR2, leading to 10 TPs and 57 FPs. The final confusion matrix values are listed in table 2 as LR2*. The third literature review by (Omran and Treude, 2017) (LR3), initially evaluated with the same research questions, has a gold standard of 33 relevant papers. However, ALISE classified 230 papers as relevant out of 232. An error analysis revealed several issues causing this misclassification, including sensitivity to certain keywords. Due to the search by keyword with "natural language" in several major high-ranked software engineering conferences, this results in all papers mentioning natural language and additionally in 226 out of 232 also "process", causing ALISE to classify nearly all papers as relevant to the first RQ of (Omran and Treude, 2017). Whereas RQ two and three of LR3 results in lower relevant papers, we identified that Flan-t5-XL also classified NLP algorithms like latent Dirichlet allocation as NLP library, which is not the library used for implementation, but an algorithm. The fourth question "If so, how was the choice justified?" makes no sense when iterating over each question because it is related to the third question as a follow-up. However, this completely failed evaluation shows two valuable conclusions: Follow-up research questions currently cannot be handled when iterating over the questions and Flan-t5-XL requires some more input rather than just buzzwords, e.g. an example what an NLP library is or what is covered under *natural language processing*. The evaluation of LR3 was repeated with a new research question, yielding 32 TPs and 22 FPs. The confusion matrix for this evaluation is in table 2 as LR3*. After reevaluation regarding the new research question, we encountered the same issue of FPs as in the evaluation of LR2. Two independent reviewers, both research engineers in NLP, followed the same procedure as in LR2: conducting individual assessments followed by a final comparison and discussion. Of the initial 54 FPs, the reviewers identified 32 as genuinely relevant due to their mention of NLP libraries used in research implementation. This significant disparity in the number of relevant papers not identified by (Omran and Treude, 2017) can be explained by their screening strategy. This evaluation is referred to as LR3** based on the indepen-

dent reviewers' annotations. Based on these confu-

Table 3: Evaluation results of ALISE.

	Acc	Re	Pr	F1	WSS	nWSS
LR1	81.63	98.28	77.03	86.36	22.77	57.50
LR2	76.45	100.00	08.96	16.44	74.13	75.89
LR2*	78.00	100.00	14.93	25.97	74.13	77.11
LR3	15.09	100.00	14.35	25.10	00.87	01.05
LR3*	71.55	63.64	28.00	38.89	31.31	72.86
LR3**	85.34	81.54	70.67	75.71	49.21	86.83

sion matrices, table 3 provides a summary of evaluation metrics for all literature reviews. The accuracy ranges from 15.09% to 85.34%, with perfect recalls for LR2, LR2*, and LR3. Precision varies between 8.96% (LR2) and 77.03% (LR1). The F1 score ranges from 16.44% to 86.36%. WSS metrics vary widely, with some in the intermediate range, while the nWSS has only one outlier LR3.

5.5 Result Analysis

(Cohen et al., 2006) noted that the goal of automated screening tools should be at least having a 95.00% recall compared to the human baseline and a WSS as high as possible. (Kusa et al., 2023) adapted this goal also with the nWSS. The evaluation metrics show (table 3), that ALISE can surpass this goal by reaching 98.28% and 100.00% for LR1 and LR2. ALISE also reached 100.00% recall for the first evaluation of LR3. Yet, this must be taken with caution because nearly every paper was classified as relevant for LR3 (see table 3). This is also then represented by the low WSS and nWSS of 00.87% and 1.05% indicating that nearly no work was saved by manually screening the literature. In contrast, exceeding 72.86% nWSS for the majority of literature reviews, this is a strong indication that, in general, ALISE is capable of saving a lot of time for human reviewers. Nevertheless, ALISE cannot be used without manual human evaluation after classification due to the classification of some FPs in each evaluated literature review. Otherwise, the nWSS would have also been perfect 100.00%. Furthermore, the outlier of LR3 and the following evaluation LR3** shows the sensitivity for the research question used as input since there is an improvement of 85,78% nWSS score. Consequently, the results mark the validation of ALISE being used for automated screening over full-texts, but having some limitations. The nWSS also makes it possible to compare these results with other automated tools. In this study, ALISE has an average nWSS score of 61.87%, which is better than the best method in average evaluated by (Kusa et al., 2023) with 57.21%. Without evaluation LR3 being an outlier due to the

wrong research questions needed for the LLM, the average is even 74.04% nWSS which clearly outperforms the best method stated in (Kusa et al., 2023). Except for model E evaluated by (Kusa et al., 2023) with an average nWSS of 55.50%, all other five evaluated models are below an average of 41.41%. However, a user of ALISE may also not initially define the research question(s) as well as required for the model and would then have to iteratively adjust that, so the average without the outlier should be viewed with caution. For this reason, the average with outlier and the median of 74.38% is more meaningful. The median of the best model D evaluated by (Kusa et al., 2023) is 60.9%. This also indicates a strong proof that ALISE can be used as an automated screening method when considering its limitations.

6 CONCLUSION

ALISE can be used as an automated screening tool over full-texts for literature reviews. Not only is relevant literature classified as relevant, ALISE also provides reasons for the review protocol why a paper is relevant. An evaluation of three different literature reviews was conducted to measure the performance of ALISE. The highest nWSS score is 86.83%, indicating a large time saving for the reviewers after the literature search. With an average of 61.87% nWSS considering all evaluated literature reviews, and a median of 74.38% nWSS, ALISE can save a lot of time but cannot be used without a followed human screening iteration over the literature classified as relevant by ALISE. However, there are some limitations when using ALISE regarding RQ sensitivity and hardware.

LIMITATIONS

ALISE shares LLM limitations, making it sensitive to the RQ and the chain prompts. In addition, fast inference requires a GPU, which makes it costly. Furthermore, the PDF conversion may introduce errors, potentially affecting the results. Two of the three LRs evaluated initially screened titles and abstracts before full-text, introducing the possibility of FNs not identified by either reviewers or ALISE. Utilizing LRs with full-text screening from the start could have mitigated this issue. An inadequately defined research question can lead to suboptimal results, negating the time savings and potentially requiring significant refinement. Even if subsequent iterations were error-free, the cumulative computation time may exceed the manual screening.

REFERENCES

- Alchokr, R., Borkar, M., Thotadarya, S., Saake, G., and Leich, T. (2022). Supporting systematic literature reviews using deep-learning-based language models. In *Proceedings of the 1st International Workshop on Natural Language-Based Software Engineering, NLBSE '22*, page 67–74, New York, NY, USA. Association for Computing Machinery.
- Bast, H. and Korzen, C. (2017). A benchmark and evaluation for text extraction from pdf. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–10.
- Blake, C. (2010). Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *Journal of Biomedical Informatics*, 43(2):173–189.
- Brocke, J. v., Simons, A., Niehaves, B., Riemer, K., Platfaut, R., and Clevén, A. (2009). Reconstructing the giant: On the importance of rigour in documenting the literature search process. In *European Conference on Information Systems*.
- 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.
- Cohen, K. B., Johnson, H. L., Verspoor, K., Roeder, C., and Hunter, L. E. (2010). The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC bioinformatics*, 11(1):492.
- da Silva Junior, B. A., Silva, J., Cavalheiro, S., and Foss, L. (2022). Pattern recognition in computing education: A systematic review. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, pages 232–243, Porto Alegre, RS, Brasil. SBC.
- Dieste, O. and Padua, A. G. (2007). Developing search strategies for detecting relevant experiments for systematic reviews. In *First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007)*, pages 215–224.
- Haji, S., Suekane, K., Sano, H., and Takagi, T. (2023). Exploratory inference chain: Exploratorily chaining multi-hop inferences with large language models for question-answering. In *2023 IEEE 17th International Conference on Semantic Computing (ICSC)*, pages 175–182.
- Jakob, D. (2022). Voice controlled devices and older adults – a systematic literature review. In Gao, Q. and Zhou, J., editors, *Human Aspects of IT for the Aged Population. Design, Interaction and Technology Acceptance*, pages 175–200, Cham. Springer International Publishing.
- Kitchenham, B. A. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE-2007-01, School of Computer Science and Mathematics, Keele University.
- Kusa, W., Lipani, A., Knoth, P., and Hanbury, A. (2023). An analysis of work saved over sampling in the evaluation of automated citation screening in systematic

- literature reviews. *Intelligent Systems with Applications*, 18:200193.
- La Quatra, M., Cagliero, L., and Baralis, E. (2021). Leveraging full-text article exploration for citation analysis. *Scientometrics*, 126(10):8275–8293.
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).
- Lo, K., Wang, L. L., Neumann, M., Kinney, R., and Weld, D. (2020). S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.
- Miah, M. S. U., Sulaiman, J., Sarwar, T. B., Naseer, A., Ashraf, F., Zamli, K. Z., and Jose, R. (2022). Sentence boundary extraction from scientific literature of electric double layer capacitor domain: Tools and techniques. *Applied Sciences*, 12(3).
- Noroozi, M., Moghaddam, H. R., Shah, A., Charkhgard, H., Sarkar, S., Das, T. K., and Pohland, T. (2023). An ai-assisted systematic literature review of the impact of vehicle automation on energy consumption. *IEEE Transactions on Intelligent Vehicles*, pages 1–22.
- Okoli, C. (2015). A guide to conducting a standalone systematic literature review. *Commun. Assoc. Inf. Syst.*, 37:43.
- Olorisade, B. K., Brereton, P., and Andras, P. (2019). The use of bibliography enriched features for automatic citation screening. *Journal of Biomedical Informatics*, 94:103202.
- Omrán, F. N. A. A. and Treude, C. (2017). Choosing an nlp library for analyzing software documentation: A systematic literature review and a series of experiments. In *Proceedings of the 14th International Conference on Mining Software Repositories, MSR '17*, page 187–197. IEEE Press.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.
- Page, M. J., McKenzie, J. E., 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 Moher, D. (2021). The prisma 2020 statement: An updated guideline for reporting systematic reviews. *Journal of Clinical Epidemiology*, 134:178–189.
- Penning de Vries, B. B., van Smeden, M., Rosendaal, F. R., and Groenwold, R. H. (2020). Title, abstract, and keyword searching resulted in poor recovery of articles in systematic reviews of epidemiologic practice. *Journal of Clinical Epidemiology*, 121:55–61.
- Portenoy, J. and West, J. D. (2020). Constructing and evaluating automated literature review systems. *Scientometrics*, 125(3):3233–3251.
- Rowe, F. (2014). What literature review is not: diversity, boundaries and recommendations. *European Journal of Information Systems*, 23(3):241–255.
- Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104:333–339.
- Tauchert, C., Bender, M., Mesbah, N., and Buxmann, P. (2020). Towards an integrative approach for automated literature reviews using machine learning. In *Hawaii International Conference on System Sciences*.
- Templier, M. and Paré, G. (2015). A framework for guiding and evaluating literature reviews. *Commun. Assoc. Inf. Syst.*, 37:6.
- van Dinter, R., Tekinerdogan, B., and Catal, C. (2021). Automation of systematic literature reviews: A systematic literature review. *Information and Software Technology*, 136:106589.
- Wang, Z., Nayfeh, T., Tetzlaff, J., O’Blenis, P., and Murad, M. H. (2020). Error rates of human reviewers during abstract screening in systematic reviews. *PLOS ONE*, 15(1):1–8.
- Webster, J. and Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, 26(2):xiii–xxiii.
- Wilson, E., Cruz, F., Maclean, D., Ghanawi, J., McCann, S. K., Brennan, P. M., Liao, J., Sena, E. S., and Macleod, M. (2023). Screening for in vitro systematic reviews: a comparison of screening methods and training of a machine learning classifier. *Clinical Science*, 137(2):181–193.
- Wu, T., Terry, M., and Cai, C. J. (2022). Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22*, New York, NY, USA. Association for Computing Machinery.
- Yu, W., Iyer, D., Wang, S., Xu, Y., Ju, M., Sanyal, S., Zhu, C., Zeng, M., and Jiang, M. (2023). Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations*.

APPENDIX

In this appendix, we list some implementation details of ALISE.

Libraries

- Langchain (<https://python.langchain.com>)

- s2orc-doc2json (<https://github.com/allenai/s2orc-doc2json>)
- Transformers (<https://github.com/huggingface/transformers>)

Langchain was used for the evidence-answer LLM chain with Flan-t5-XL coming with transformers and hugging face hub. To transform the pdf papers into strings, we used the s2orc-doc2json converter.

Prompt Templates

Figure 3 contains the evidence and answer prompt templates used. These templates performed best in the evidence-response chain evaluation with the QASPER dataset.

```
[ ]: evidence_prompt_template = """Return only the full sentences from the given text that provide,
    .-an answer to the question. Avoid answers that are incorrect or provides incomplete,
    .-justification for the question.
    If you find a relevant span, return the full sentence of this span, not only the span.
    If you can't find a sentence that provides an answer, return \"None\".

Text:
{context}

Question: {question}
Relevant Sentences: """

EVIDENCE_PROMPT = PromptTemplate(
    template=evidence_prompt_template, input_variables=["context", "question"]
)

answer_prompt_template = """Answer the question based on the given extracted sentences.
    Read the sentences carefully and consider any relevant information before answering.

    If you are unable to generate an answer based on the provided sentences, respond with,
    .-\\"unanswerable\".
    If the question is boolean, respond with \"yes\" or \"no\" or if it is unanswerable with,
    .-\\"unanswerable\".
    If there are more bulletpoints in the answer, list them all and do not explain each further.

Provided Sentences:
{context}

Question:
{question}

Answer: """

ANSWER_PROMPT = PromptTemplate(
    template=answer_prompt_template, input_variables=["context", "question"]
)
```

Figure 3: Prompt templates.