LiRA: A Multi-Agent Framework for Reliable and Readable Literature Review Generation

Gregory Hok Tjoan Go[1, 2], **Khang Ly**[2], **Anders Søgaard**[3] ,
Seyed Amin Tabatabaei[2], **Maarten de Rijke**[1], **Xinyi Chen**[1]

[1]University of Amsterdam
[2]Elsevier B.V.
[3]University of Copenhagen
{g.go, k.ly, s.tabatabaei}@elsevier.com, soegaard@di.ku.dk, {m.derijke, x.chen2}@uva.nl

Abstract

The rapid growth of scientific publications has made it increasingly difficult to keep literature reviews comprehensive and up-to-date. Though prior work has focused on automating retrieval and screening, the writing phase of systematic reviews remains largely under-explored, especially with regard to readability and factual accuracy. To address this, we present **LiRA** (**Li**terature **R**eview **A**gents), a multi-agent collaborative workflow which emulates the human literature review process. LiRA utilizes specialized agents for content outlining, subsection writing, editing, and reviewing, producing cohesive and comprehensive review articles. Evaluated on SciReviewGen and a proprietary ScienceDirect dataset, LiRA outperforms current baselines such as AutoSurvey and MASS-Survey in writing and citation quality, while maintaining competitive similarity to human-written reviews. We further evaluate LiRA in real-world scenarios using document retrieval and assess its robustness to reviewer model variation. Our findings highlight the potential of agentic LLM workflows, even without domain-specific tuning, to improve the reliability and usability of automated scientific writing.

Code —
www.github.com/lira-workflow/auto-review-writing

1 Introduction

Since their inception, literature reviews have been consistently used to streamline the advancement of various scientific fields (Snyder et al. 2016). Of these reviews, one of the most important types is the Systematic Literature Review (SLR), which reproducibly synthesizes a significant portion of existing research relating to a specific research question being addressed (Kitchenham and Charters 2007; Bangdiwala 2024). This role has become increasingly more crucial, evidenced by how quite a few researchers consider them to be original research or potentially even a mandatory step in the scientific process itself (Kraus, Mahto, and Walsh 2023; Palmatier, Houston, and Hulland 2018).

Due to the large amount of new findings and research being disseminated through publications, it has become very difficult to release SLRs in a timely fashion (Qi, Li, and Lyu 2025; Ofori-Boateng et al. 2024; Tian et al. 2025). For example, the estimated time required to complete an SLR has increased significantly in the past few decades in the medical domain (Allen and Olkin 1999; Borah et al. 2017), which is further compounded by the necessity of using expert labor (Atkinson 2025). In tackling this, the vast majority of research relating to SLR automation focuses on the retrieval and screening of scientific papers (Orel et al. 2023; Marshall and Wallace 2019; Chen et al. 2025), as these are the most time-consuming steps (Chai et al. 2021). However, there remains the task of compiling the findings into a comprehensive review paper. Only a small number of works have been published (Kasanishi et al. 2023; Qi, Li, and Lyu 2025; Wang et al. 2024), let alone those which focus on the readability and hallucination mitigation aspects.

In this work, we present **LiRA** (**Li**terature **R**eview **A**gents), an agentic solution aimed at addressing the minimal research related to automated literature review writing. It is a Large Language Model (LLM)-based agentic workflow building upon existing relevant works and integrates some of the most recent SLR-writing guidelines to generate accurate and high-quality reviews, with an additional emphasis on readability. Moreover, it is entirely out-of-the-box, requiring no task-specific pre-training or fine-tuning. We also reduce hallucination in the outputs, which is one of the main barriers in trustworthy automated writing and a key factor preventing the widespread use of similar Artificial Intelligence (AI)-powered systems (Alkaissi and McFarlane 2023; O'Connor et al. 2024; Xu et al. 2023).

To demonstrate LiRA's capabilities, we propose the below research questions:

RQ1 *To what extent are the qualitative Systematic Literature Reviews created by LiRA similar to human-written ones compared to existing literature writing methods when all are given the same set of references?*

RQ2 *How well written are the generated articles compared to existing literature writing methods when all are given the same set of references?*

RQ3 *How well can LiRA properly use citations from the provided sources to generate qualitative Systematic Literature Reviews compared to other methods?*

RQ4 *How well can LiRA be used in real-world settings when using references returned by a scientific document retriever?*

We summarize our main contributions as follows:

- To the best of our knowledge, we propose the first agentic LLM-based automated literature review writing workflow which emulates the human writing process and integrates the findings of other relevant agentic workflows.
- We explore the usage of formally defined guidelines and techniques from relevant similar fields in the agentic workflow, such as the idea of thoroughly analyzing the existing papers before beginning the writing process, establishing a crucial link between theory and application.
- We establish several state-of-the-art baseline results for the automated literature review writing task across multiple settings, comparing between existing systems when using the same LLM type throughout.

2 Related Work

Agentic workflows Comparisons have been made between agentic LLM systems and human cognition, due to how breaking down a task into smaller steps, which these workflows often do, is commonly used to describe how humans solve more intricate problems (Flower and Hayes 1981; Correa et al. 2023). Using this concept, several works show promise in implementing agentic workflows in various fields, such as the medical sciences (Tang et al. 2024) and law (Watson et al. 2025). In some cases, this results in an improvement of more than 90% compared to a simple baseline (Watson et al. 2025).

However, for the task of automated writing specifically, only few works have been published thus far (Qi, Li, and Lyu 2025; Wang et al. 2024; Shao et al. 2024; Tian et al. 2025). Of these works, only two of them address literature reviews or a similar document type and include open-source code (Qi, Li, and Lyu 2025; Wang et al. 2024). Neither paper takes output lengths and how they relate to the readability and evaluation of each proposed system using the listed metrics into account. As a result, no works seem to exist which tackle the issues of readability, and only minimal work exists in addressing the factuality of the resulting articles.

Literature review For centuries (Lind 2014), the process of writing literature reviews has been considered crucial for the development of science (Meerpohl et al. 2012; Higgins et al. 2011; Chalmers, Hedges, and Cooper 2002), as it provides researchers valuable insight on which research questions to answer via an analysis of earlier works (Chalmers and Glasziou 2009; Eagly and Wood 1994). Moreover, improvements have been made to eliminate personal biases (Egger, Smith, and O'Rourke 2001) through the introduction of the SLR, which uses systematic methods of review for the collation and synthesis of findings (Randles and Finnegan 2023; Snyder 2019; Page et al. 2021).

Given how time-consuming (Borah et al. 2017) and costly (Michelson and Reuter 2019) this process is, a need for a viable alternative has emerged. Despite this, there is not much relevant innovation in natural language processing that tackles this issue (Mohammad et al. 2009; Kasanishi et al. 2023; Agarwal et al. 2011), let alone results indicating real-world usability. Therefore, we address all aforementioned problems by introducing an agentic workflow which uses LLMs to generate literature reviews automatically, demonstrating its potential by focusing on both readability and factuality.

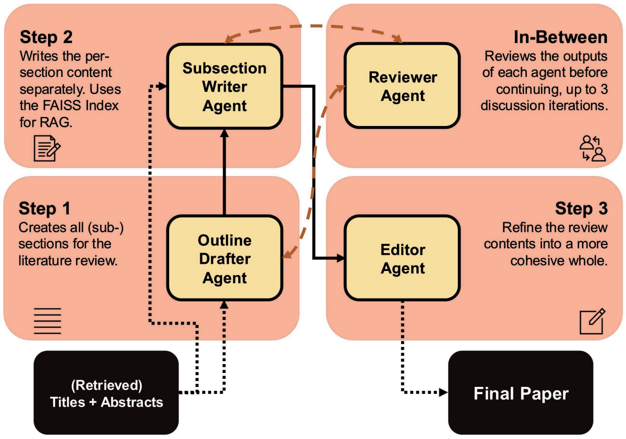

Step 2
Writes the per-section content separately. Uses the FAISS Index for RAG.
Subsection Writer Agent
Reviewer Agent
In-Between
Reviews the outputs of each agent before continuing, up to 3 discussion iterations.
Step 1
Creates all (sub-) sections for the literature review.
Outline Drafter Agent
Editor Agent
Step 3
Refine the review contents into a more cohesive whole.
(Retrieved) Titles + Abstracts
Final Paper

Figure 1: An overview of the LiRA architecture. The narrow dotted arrows represent document input/output, the wide dotted arrows indicate the refinement process, and the filled-in arrows signify the system's main flow. Each agent is explained in the below sections.

3 The LiRA Framework

LiRA emulates the human literature review process by decomposing it into specialized, interacting LLM-based agents. Each agent tackles a key sub-task: either structural planning, fine-grained writing, consistency refinement, or factual verification, which results in a modular and scalable pipeline. This section introduces the core agents and their design motivations.

3.1 Outline Drafter Agent

A key challenge in literature review writing is constructing a coherent structure from a large and unorganized set of references. Instead of relying on the model to implicitly determine the structure during generation, LiRA explicitly drafts an outline to guide the writing process. This agent takes the topic and abstracts (or alternatively the full texts) of the provided references to produce a set of candidate outlines consisting of main sections and subsections. These are then combined into a unified draft structure that includes descriptions for each section and suggested supporting papers.

To manage the context size and focus on only the most relevant content, the outline is constructed from up to 50 references. We use existing heuristics (Wang et al. 2024) that recommend generating approximately 8 sections with around 4 subsections, but the agent can adapt this amount as needed. We also include default sections such as an introduction and conclusion to ensure consistency and completeness.

3.2 Subsection Writer Agent

Generating a full review in a single pass risks superficial coverage and poor section coherence. We tackle this by having a subsection writer agent, which writes each (sub)section individually in parallel, conditioned on the description and a selected subset of relevant references. This design encourages fine-grained generation and makes the outputs more modular and easier to revise.

Relevant references are retrieved using the FAISS index (Douze et al. 2025) based on section-level descriptions, capped at 25% of the total pool per subsection (min 3, max 150). This balances citation diversity with input tractability. Moreover, the reference titles, abstracts, or full texts can be retrieved depending on the availability. The agent outputs ∼1,000 words per subsection, enabling long-form synthesis while staying within model context limits. Also, the article title, abstract, and conclusion are written after the body is generated, mimicking human writing practices and avoiding the premature commitment of top-down generation.

3.3 Editor Agent

Even with structural planning and localized generation, the assembled review may exhibit issues such as redundancy, inconsistency, or stylistic mismatches across sections. These are addressed by an editor agent, which refines the entire draft with a focus on presentation and style. It performs standard editing operations, including improving transitions, enhancing vocabulary, and ensuring overall fluency and cohesion. Importantly, this agent does not alter the factual content, preserving the integrity of the generated information.

If the generated output exceeds the model's context window, the agent detects whether the text ends abruptly—typically when the maximum output length (16,384 tokens) is reached. In such cases, the model is prompted to continue the output using the original input and previously edited content as context. This mechanism effectively doubles the generation capacity, though at the cost of additional input overhead.

3.4 Reviewer Agent

To improve factual accuracy and review quality, we introduce an LLM-based reviewer agent inspired by human editorial workflows. This agent evaluates intermediate outputs (e.g., outlines) based on adapted criteria from systematic review guidelines (Snyder 2019), including content completeness, transparency, clarity, and contribution.

If a component (e.g., an outline or section) fails to meet quality thresholds, the reviewer provides structured feedback and triggers regeneration. The review loop continues up to 3 rounds before fallback progression, balancing refinement with computational efficiency.

3.5 Citation Behavior

Citation hallucination remains a major concern in automated scientific writing, as models may generate plausible-sounding but non-existent references. To address this, LiRA incorporates citation grounding directly into the generation process. Rather than relying on abstract placeholders (e.g., numbered citations), agents cite sources using full article titles, which act as semantic anchors and help the model maintain alignment with the provided references.

After generation, these in-text citations are post-processed into standard numbered references, and any hallucinated titles are redacted during evaluation to ensure fair comparison. This approach improves factual consistency and ensure fair comparison during evaluation.

3.6 Additional Implementation Details

All agents in LiRA were implemented using LangGraph. Moreover, each agent has its own memory, a standard practice for LLM-based agentic workflows to better act upon feedback (Sumers et al. 2023; Qian et al. 2024). This system of storing feedback in memory is comparable to Reflexion (Shinn et al. 2023), where the agent has to adjust its behavior based on the feedback provided. Parallelization is also included to increase the processing speed of certain steps, namely the researcher when analyzing papers and the content writer for generating the article (sub)sections.

Aspects from the design of MetaGPT (Hong et al. 2023) were adapted for more efficient inter-agent communication. Specifically, all agents are required to return their outputs as structured documents to avoid potential inefficiencies relating to information presentation, which are then sent to unique shared message pools for quick information retrieval. Moreover, the input contents are filtered based on the model's maximum context window length (128,000 tokens in our case) to prevent information overload. As a method of improving the model's output quality with minimal intervention, Zero-Shot Chain-of-Thought (Kojima et al. 2022) is included in the prompts for all agents except the researcher and editor, as they do not perform refinement.

4 Experiments

4.1 Baselines

To evaluate LiRA, we compare it against direct prompting and, to the best of our knowledge, the only two publicly available agentic frameworks for survey writing. For fair comparison, all systems (including LiRA) are implemented with `gpt-4o-mini` as the underlying LLM.

Direct prompting (DP) As the simplest baseline, we directly prompt the LLM with task instructions and the full set of reference titles and metadata, asking it to generate a review with the specified sections and length. When the reference list exceeds the context window, it is passed as an attached file, requiring the model's file-reading capability. This baseline tests whether a single prompt can produce a coherent review without decomposition or refinement.

MASS-Survey (MASS) Introduced in an automated survey-writing challenge (Tian et al. 2025), MASS is the only agentic framework from that challenge with publicly available code. Its workflow differs fundamentally from LiRA: Instead of decomposing the writing process into specialized roles with iterative feedback, MASS follows a strictly sequential pipeline. The system first clusters references by topical similarity to construct an outline, then generates section contents and a title directly from these clusters, and finally appends a conclusion. Long reference lists are handled by passing them as attachments.

AutoSurvey (AS) AutoSurvey (Wang et al. 2024) is a multi-stage framework for automatically generating literature surveys in computer science. Given a query (in our case, the original review title), it retrieves relevant publications and uses their titles and abstracts to construct an outline, followed by subsection drafting with refinement steps and partial use of retrieved article content (up to 1,500 tokens). While the paper claims retrieval-grounded drafting, we did not find corresponding functionality in the released code.

For fair comparison, we modified AutoSurvey to restrict retrieval to the references cited in the target human-written review. This required reducing the number of candidate documents per subsection to between 2 and 25% of the references, or falling back to 60 when the fraction exceeded this threshold (as in the original design). We also generalized system prompts from "*You are an expert in Artificial Intelligence*" to "*You are an expert in a relevant field*" to make the system better able to generate articles for multiple domains.

4.2 Metrics

To comprehensively evaluate the generated literature reviews, we consider three complementary dimensions: content similarity to human-written reviews, writing quality, and citation reliability.

Similarity to the human-written review We measure how similar the generated literature reviews are to human-written ones. Metrics include ROUGE-L, heading soft recall (*hsr*), heading entity recall (*her*), and article entity recall (*aer*). Together, these capture lexical overlap, structural alignment, and coverage of key cited works. Full definitions are provided in Appendix A.

Writing quality We evaluate writing quality using both automatic and human assessments. For automatic evaluation, we use the Prometheus 2 LLM (Kim et al. 2024). It is an open-source LLM evaluator that uses the appropriate reference materials (the instruction, reference answer, and score rubric) to provide assessments which mostly align with those of human annotators. Three aspects are evaluated for all generated articles, namely the coverage, structure, and relevance. Here, coverage represents how broad the subject matter of the review is, while structure measures the organization and flow of the review, and relevance indicates how well the review is able to stay on-topic overall. Additional specifications for the model and how it measures the aforementioned aspects can be found in Appendix B.

For human evaluation, we employed subject matter experts (SMEs) who helped evaluate the outputs of the system on the same aspects as mentioned above. For feasibility, the annotation was performed differently for each dataset, though the title, outline, and a section snippet were utilized in both cases. For SciReviewGen, we employed a group of 3 SMEs from Straive to select their preferences between human- and LiRA-written articles for 30 sample snippets, using a rubric as guidance for determining their choice. It must also be noted that the order in which these samples were presented was randomized for each row.

Meanwhile, a dedicated team from within the company provided scores ranging from 1 to 5 for the AutoSurvey and LiRA articles while using the human-written ones as a baseline. This grading was done using the same rubric as mentioned above. Furthermore, due to this type of annotation being more labor-intensive, only 15 article snippets were used.

Citation quality We evaluate how well generated claims are grounded in appropriate references. Our metric, Citation Quality F1-Score (CQF1), balances precision (penalizing irrelevant or hallucinated citations) and recall (capturing missing but necessary citations), serving as a proxy for hallucination in scientific writing. A full overview of the details is given in Appendix D.

4.3 Datasets

We evaluate the effectiveness of the LiRA framework using two datasets. The primary dataset is SciReviewGen (Kasanishi et al. 2023), a publicly available benchmark built on the Semantic Scholar Open Research Corpus (S2ORC) (Lo et al. 2020). It contains 10,000 review articles in computer science, referencing approximately 690,000 papers. Each review is annotated with structured metadata, including titles, section headers, full texts, and references. Following the setup in Shao et al. (2024), we randomly sample 125 reviews, ensuring each selected article contains a sufficient number of references—averaging around 70 per paper.

To assess the generalizability beyond computer science, we additionally evaluate on an internal dataset of 125 expert-written reviews from ScienceDirect, covering 23 subject areas including business, microbiology, and materials science. The dataset is matched in size to the SciReviewGen subset to support direct cross-domain comparison.

4.4 Results

Across both datasets, LiRA consistently outperforms baseline systems on the majority of evaluation metrics, demonstrating stronger alignment with human-written reviews, higher writing quality, and more reliable citation use.

Similarity to human-written review LiRA achieves the highest ROUGE scores, indicating stronger lexical alignment with human-written reviews. AutoSurvey attains slightly higher heading/entity recall, but largely due to verbosity: on average, AutoSurvey produces 50,000 tokens per article compared to only 22,000 for LiRA. Since recall-based metrics do not normalize for length, longer outputs are naturally favored. Crucially, this shows that LiRA generates concise yet information-dense reviews, rather than inflating scores by producing excessive text.

Writing quality LiRA achieves the best overall writing quality, with a clear advantage in structural coherence. AutoSurvey performs marginally better in coverage, again reflecting its longer outputs, but at the cost of organization and readability. SME evaluations align with these trends: on ScienceDirect, experts strongly preferred LiRA for structure, while AutoSurvey received marginally higher scores for coverage and relevance, suggesting a trade-off between breadth and coherence. On SciReviewGen, SMEs favored LiRA over the human-written reviews, noting its broader and more balanced coverage given the outlines.

Metric	DP	MASS	AS	LiRA
SciReviewGen				
ROUGE	0.06 ± 0.0	0.09 ± 0.0	0.09 ± 0.0	0.13 ± 0.0
hsr	0.69 ± 0.1	0.66 ± 0.1	**0.92 ± 0.4**	0.82 ± 0.1
her	0.06 ± 0.0	0.05 ± 0.0	**0.15 ± 0.0**	0.10 ± 0.0
aer	0.06 ± 0.0	0.09 ± 0.0	**0.34 ± 0.0**	0.27 ± 0.0
ScienceDirect				
ROUGE	0.02 ± 0.0	0.04 ± 0.0	**0.13 ± 0.0**	**0.13 ± 0.0**
hsr	0.24 ± 0.2	0.22 ± 0.2	0.24 ± 0.4	**0.25 ± 0.1**
her	0.03 ± 0.0	0.03 ± 0.0	**0.13 ± 0.0**	0.05 ± 0.0
aer	0.03 ± 0.0	0.05 ± 0.0	**0.25 ± 0.0**	0.17 ± 0.0

Table 1: Results for similarity to the human-written reviews with the baseline settings.

Metric	DP	MASS	AS	LiRA
SciReviewGen				
Coverage	3.53 ± 0.4	4.30 ± 0.3	**4.50 ± 0.1**	4.45 ± 0.1
Structure	3.15 ± 0.9	2.47 ± 1.2	2.30 ± 1.3	**3.38 ± 0.9**
Relevance	4.49 ± 0.2	**4.74 ± 0.2**	4.55 ± 0.2	4.57 ± 0.2
Average	3.72	3.83	3.78	**4.13**
ScienceDirect				
Coverage	3.08 ± 1.0	3.44 ± 1.0	**4.10 ± 0.1**	3.90 ± 0.3
Structure	3.23 ± 1.0	2.59 ± 1.1	2.21 ± 1.3	**3.42 ± 0.9**
Relevance	3.98 ± 0.7	4.15 ± 0.7	4.29 ± 0.3	**4.33 ± 0.3**
Average	3.43	3.39	3.53	**3.88**

Table 2: Writing quality results for the baseline settings.

Dataset	DP	MASS	AS	LiRA
SciReviewGen	0.14	0.13	0.63	**0.76**
ScienceDirect	0.06	0.33	0.55	**0.73**

Table 3: Citation quality results for the baseline settings.

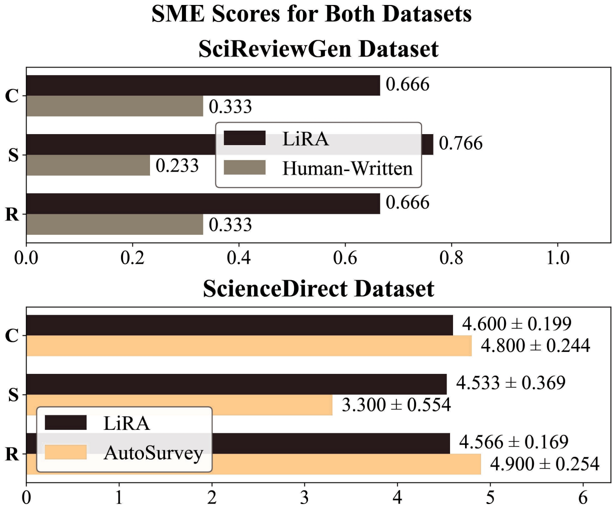


Figure 2: SME evaluation results. Here, **C** indicates Coverage, **S** indicates Structure, and **R** indicates Relevance.

These results highlight LiRA's strength in generating well-structured, concise, and expert-aligned reviews, in some cases even being favored over human-written baselines.

Citation quality LiRA demonstrates the largest gains in citation reliability, achieving the highest Citation Quality F1 (CQF1) scores across both datasets (0.76 on SciReviewGen, 0.73 on ScienceDirect) and substantially outperforming AutoSurvey ($\leq$0.63) and all other baselines. This indicates that LiRA is more effective at grounding claims in appropriate references, avoiding both omissions (recall errors) and hallucinations (precision errors), which stem from LiRA's citation-grounded generation design that explicitly enforces reference anchoring during drafting and refinement.

From this, it can be seen that LiRA overall produces literature reviews that are concise, structurally coherent, and citation-faithful, while maintaining competitive coverage. This balance between quality and reliability highlights LiRA as a more trustworthy and practically useful framework for automated survey writing. achieves the highest ROUGE scores, indicating stronger lexical alignment with human-written reviews. AutoSurvey attains slightly higher heading/entity recall, but largely due to verbosity: on average, AutoSurvey produces 50,000 tokens per article compared to only 22,000 for LiRA. Since recall-based metrics do not normalize for length, longer outputs are naturally favored. Crucially, this shows that LiRA generates concise yet information-dense reviews, rather than inflating scores by producing excessive text.

5 Potential Modifications on LiRA

In this section, we discuss the adjustments tested on LiRA to evaluate its performance more robustly. This involves the usage of a different LLM type for the reviewer agent to potentially mitigate self-bias amplification in the refinement process, and document retriever usage to evaluate if LiRA can be deployed in real-world settings.

5.1 Using a different reviewer model

Method Based on concerns stemming from self-bias amplification (Xu et al. 2024), experimentation was performed using a reviewer model type different from the one used during generation for each component, which is based on existing suggestions. To this end, the `gemma3:4b` model was used through `ollama` (Kamath et al. 2025). This model was chosen because it is open-source, has a context window similar to `gpt-4o-mini`, and has lower hardware requirements compared to most other models. In addition, `ollama` was the selected model provider because of its accessibility and ease of use, with LangGraph already being compatible with it. The parameter values used by `gemma3:4b` were the default `ollama` ones aside from the context window size (128,000) and seed (42).

	SciReviewGen		ScienceDirect	
Metric	**LiRA**	**LiRA +gemma3**	**LiRA**	**LiRA +gemma3**
ROUGE	**0.13 ± 0.0**	**0.13 ± 0.0**	**0.13 ± 0.0**	**0.13 ± 0.0**
hsr	0.82 ± 0.1	**0.83 ± 0.1**	0.25 ± 0.1	**0.25 ± 0.1**
her	0.10 ± 0.0	**0.10 ± 0.0**	**0.05 ± 0.0**	0.05 ± 0.0
aer	**0.27 ± 0.0**	0.26 ± 0.0	**0.17 ± 0.0**	0.17 ± 0.0
CQF1	0.76	**0.77**	**0.73**	0.73
Coverage	**4.45 ± 0.1**	4.45 ± 0.1	**3.90 ± 0.3**	3.89 ± 0.4
Structure	3.38 ± 0.9	**3.47 ± 0.9**	**3.42 ± 0.9**	3.26 ± 1.0
Relevance	**4.57 ± 0.2**	4.64 ± 0.2	**4.33 ± 0.3**	4.29 ± 0.3

Table 4: The gemma3 results for the ScienceDirect dataset. Note that if both numbers in a row are bolded, it means they returned the exact same value.

Results The results for this setting can be found in Table 4. Though the extent of bias mitigation itself cannot be measured properly with the current metrics, we can remark how using `gemma3:4b` has little impact on the scores overall across all metrics. This suggests that alternative model configurations may not significantly alter the article output quality, therefore indicating potentially that the pipeline can work well even when using different LLMs in the process.

5.2 Retrieval Usage

Method All prior experiments assumed the availability of gold references from a reference article to generate a review. This is not the case for real-world settings, however, as novel literature reviews are required to keep up with current developments. Therefore, we evaluate if the system can generate reviews similar enough to the human-written ones when provided with retrieved references instead, hence the inclusion of **RQ4**. Specifically, an internal API for embedding-similarity search was used, which can be called to retrieve as many references as listed in the human-written review.

Results We examined if the results when using retrieval differed significantly compared to the baseline researcher setting, which was tested using the appropriate statistical tests. From the results (shown in Table 5), we note that only two results were significantly different compared to the baseline, indicating that LiRA is capable of performing similarly despite the different references used. More details on the statistical tests can be found in Appendix F.

6 Deployment

The deployment of LiRA will use the following steps. First, it will be developed using the Python version of LangGraph, which is an open-source and production-ready agentic framework. Furthermore, the `gpt-4o-mini` LLM from AzureOpenAI will be used, with the possibility of using other models given LangGraph's extensive support for various other endpoints such as form `ollama`.

As the use case of LiRA requires it to generate novel literature reviews not based on existing reviews, a document retrieval system will be added on to the system. It would function by asking the user for a review topic as input, which would then be enriched using an LLM (i.e., `gpt-4o-mini`) and afterwards used for embedding-based retrieval using an internal API. This API by default has access to a large collection of scientific articles, and can be replaced depending on the specific circumstances.

Metric	**LiRA**	**LiRA +retriever**
ROUGE	**0.130 ± 0.021**	0.128 ± 0.019
hsr	**0.257 ± 0.130**	0.251 ± 0.116
her	**0.056 ± 0.043**	0.054 ± 0.044
aer	**0.170 ± 0.057**	0.152 ± 0.054*
Coverage	**3.892 ± 0.407**	3.839 ± 0.415*
Structure	3.264 ± 1.049	**3.411 ± 1.023**
Relevance	**4.296 ± 0.389**	4.270 ± 0.372

Table 5: The ScienceDirect retrieval results. The stars indicate results significantly lower than the baseline.

7 Conclusion, Limitations, and Future Work

This work introduces **LiRA**, an agentic workflow designed for the automatic writing of literature reviews, which integrates the concepts of research before writing and refinement in its core pipeline. The results obtained show that LiRA is capable of performing the task of automated literature review quite well, outperforming all tested open-source methods when accounting for the varying output lengths, indicating a positive result for essentially every research question proposed. Moreover, it reduces hallucination through improved citation behavior and can demonstrably be used in real-world settings.

Several improvements could be made, mainly regarding the irreproducibility of results due to the usage of `gpt-4o-mini` for all experiments. This can be solved by using seedable models instead, which should be feasible given current LLM availability. In addition, there is a lack of open-source datasets for this task specifically, which hinders the generalizability of all results to other scientific fields. Therefore, we encourage authors to create additional datasets, ideally in a similar format to SciReviewGen, to facilitate the evaluation of similar systems in the future.

Furthermore, opportunities exist to create more end-to-end pipelines, as the current project does not take into account factors such as primary studies and risk of bias in randomized trials (i.e., the implementation of automated tools based on Higgins et al. (2024)). Doing this would allow for the integration of more steps within the literature review writing process, namely the screening and search criteria definition steps, which would allow for better paper reproducibility.

Acknowledgments

We would like to thank Marcela Haldan and Alexandra Noti for their help in providing annotations for the ScienceDirect articles. In addition, this research was (partially) supported by the Dutch Research Council (NWO), under project numbers 024.004.022, NWA.1389.20.183, and KICH3.LTP.20.-006, and the European Union under grant agreements No. 101070212 (FINDHR) and No. 101201510 (UNITE). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of their respective employers, funders and/or granting authorities.

References

Agarwal, N.; Reddy, R. S.; Gvr, K.; and Rosé, C. P. 2011. Towards Multi-Document Summarization of Scientific Articles:Making Interesting Comparisons with SciSumm. In Nenkova, A.; Hirschberg, J.; and Liu, Y., eds., *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages*, 8–15. Portland, Oregon: Association for Computational Linguistics.

Alkaissi, H.; and McFarlane, S. I. 2023. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*.

Allen, I. E.; and Olkin, I. 1999. Estimating Time to Conduct a Meta-analysis From Number of Citations Retrieved. *JAMA*, 282(7): 634–635.

Atkinson, C. F. 2025. AI-pocalypse now: Automating the Systematic Literature Review with SPARK (Systematic Processing and Automated Review Kit) – Gathering, Organising, Filtering, and Scaffolding. *MethodsX*, 14: 103129.

Bangdiwala, S. I. 2024. The Importance of Systematic Reviews. *International Journal of Injury Control and Safety Promotion*, 31(3): 347–349. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/17457300.2024.2388484.

Borah, R.; Brown, A. W.; Capers, P. L.; and Kaiser, K. A. 2017. Analysis of the Time and Workers Needed to Conduct Systematic Reviews of Medical Interventions Using Data from the PROSPERO Registry. *BMJ Open*, 7(2): e012545. Publisher: British Medical Journal Publishing Group Section: Health informatics.

Chai, K. E. K.; Lines, R. L. J.; Gucciardi, D. F.; and Ng, L. 2021. Research Screener: A Machine Learning Tool to Semi-automate Abstract Screening for Systematic Reviews. *Systematic Reviews*, 10(1): 93.

Chalmers, I.; and Glasziou, P. 2009. Avoidable Waste in the Production and Reporting of Research Evidence. *The Lancet*, 374(9683): 86–89. Publisher: Elsevier.

Chalmers, I.; Hedges, L. V.; and Cooper, H. 2002. A Brief History of Research Synthesis. *Evaluation & the Health Professions*, 25(1): 12–37. Publisher: SAGE Publications Inc.

Chen, H.; Jiang, Z.; Liu, X.; Xue, C. C.; Yew, S. M. E.; Sheng, B.; Zheng, Y.-F.; Wang, X.; Wu, Y.; Sivaprasad, S.; Wong, T. Y.; Chaudhary, V.; and Tham, Y. C. 2025. Can Large Language Models Fully Automate or Partially Assist Paper Selection in Systematic Reviews? *British Journal of Ophthalmology*. Publisher: BMJ Publishing Group Ltd Section: Epidemiology.

Correa, C. G.; Ho, M. K.; Callaway, F.; Daw, N. D.; and Griffiths, T. L. 2023. Humans Decompose Tasks by Trading Off Utility and Computational Cost. *PLOS Computational Biology*, 19(6): e1011087. Publisher: Public Library of Science.

Douze, M.; Guzhva, A.; Deng, C.; Johnson, J.; Szilvasy, G.; Mazaré, P.-E.; Lomeli, M.; Hosseini, L.; and Jégou, H. 2025. The Faiss Library. ArXiv:2401.08281 [cs].

Eagly, A. H.; and Wood, W. 1994. Using Research Syntheses to Plan Future Research. In *The Handbook of Research Synthesis*, 485–500. New York, NY, US: Russell Sage Foundation. ISBN 978-0-87154-226-7.

Egger, M.; Smith, G. D.; and O'Rourke, K. 2001. Introduction: Rationale, Potentials, and Promise of Systematic Reviews. In *Systematic Reviews in Health Care*, 1–19. John Wiley & Sons, Ltd. ISBN 978-0-470-69392-6.

Flower, L.; and Hayes, J. R. 1981. A Cognitive Process Theory of Writing. *College Composition and Communication*, 32(4): 365.

Fränti, P.; and Mariescu-Istodor, R. 2023. Soft Precision and Recall. *Pattern Recognition Letters*, 167: 115–121.

Higgins, J.; Thomas, J.; Chandler, J.; Cumpston, M.; Li, T.; Page, M.; et al., eds. 2024. *Cochrane Handbook for Systematic Reviews of Interventions version 6.5 (updated August 2024)*. https://training.cochrane.org/handbook/current.

Higgins, J. P. T.; Altman, D. G.; Gøtzsche, P. C.; Jüni, P.; Moher, D.; Oxman, A. D.; Savović, J.; Schulz, K. F.; Weeks, L.; and Sterne, J. A. C. 2011. The Cochrane Collaboration's Tool for Assessing Risk of Bias in Randomised Trials. *BMJ*, 343: d5928. Publisher: British Medical Journal Publishing Group Section: Research Methods &amp; Reporting.

Hong, S.; Zhuge, M.; Chen, J.; Zheng, X.; Cheng, Y.; Wang, J.; Zhang, C.; Wang, Z.; Yau, S. K. S.; Lin, Z.; Zhou, L.; Ran, C.; Xiao, L.; Wu, C.; and Schmidhuber, J. 2023. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. In *Proceedings of the Twelfth International Conference on Learning Representations*.

Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; Rouillard, L.; Mesnard, T.; Cideron, G.; Grill, J.-b.; Ramos, S.; Yvinec, E.; Casbon, M.; Pot, E.; Penchev, I.; Liu, G.; Visin, F.; Kenealy, K.; Beyer, L.; Zhai, X.; Tsitsulin, A.; Busa-Fekete, R.; Feng, A.; Sachdeva, N.; Coleman, B.; Gao, Y.; Mustafa, B.; Barr, I.; Parisotto, E.; Tian, D.; Eyal, M.; Cherry, C.; Peter, J.-T.; Sinopalnikov, D.; Bhupatiraju, S.; Agarwal, R.; Kazemi, M.; Malkin, D.; Kumar, R.; Vilar, D.; Brusilovsky, I.; Luo, J.; Steiner, A.; Friesen, A.; Sharma, A.; Sharma, A.; Gilady, A. M.; Goedeckemeyer, A.; Saade, A.; Feng, A.; Kolesnikov, A.; Bendebury, A.; Abdagic, A.; Vadi, A.; György, A.; Pinto, A. S.; Das, A.; Bapna, A.; Miech, A.; Yang, A.; Paterson, A.; Shenoy, A.; Chakrabarti, A.; Piot, B.; Wu, B.; Shahriari, B.; Petrini, B.; Chen, C.; Lan, C. L.; Choquette-Choo, C. A.; Carey, C. J.; Brick, C.; Deutsch, D.; Eisenbud, D.; Cattle, D.; Cheng, D.; Paparas, D.; Sreepathihalli, D. S.; Reid, D.; Tran, D.; Zelle, D.; Noland, E.; Huizenga, E.; Kharitonov, E.; Liu, F.; Amirkhanyan, G.; Cameron, G.; Hashemi, H.; Klimczak-Plucińska, H.; Singh, H.; Mehta, H.; Lehri, H. T.; Hazimeh, H.; Ballantyne, I.;

Szpektor, I.; Nardini, I.; Pouget-Abadie, J.; Chan, J.; Stanton, J.; Wieting, J.; Lai, J.; Orbay, J.; Fernandez, J.; Newlan, J.; Ji, J.-y.; Singh, J.; Black, K.; Yu, K.; Hui, K.; Vodrahalli, K.; Greff, K.; Qiu, L.; Valentine, M.; Coelho, M.; Ritter, M.; Hoffman, M.; Watson, M.; Chaturvedi, M.; Moynihan, M.; Ma, M.; Babar, N.; Noy, N.; Byrd, N.; Roy, N.; Momchev, N.; Chauhan, N.; Sachdeva, N.; Bunyan, O.; Botarda, P.; Caron, P.; Rubenstein, P. K.; Culliton, P.; Schmid, P.; Sessa, P. G.; Xu, P.; Stanczyk, P.; Tafti, P.; Shivanna, R.; Wu, R.; Pan, R.; Rokni, R.; Willoughby, R.; Vallu, R.; Mullins, R.; Jerome, S.; Smoot, S.; Girgin, S.; Iqbal, S.; Reddy, S.; Sheth, S.; Põder, S.; Bhatnagar, S.; Panyam, S. R.; Eiger, S.; Zhang, S.; Liu, T.; Yacovone, T.; Liechty, T.; Kalra, U.; Evci, U.; Misra, V.; Roseberry, V.; Feinberg, V.; Kolesnikov, V.; Han, W.; Kwon, W.; Chen, X.; Chow, Y.; Zhu, Y.; Wei, Z.; Egyed, Z.; Cotruta, V.; Giang, M.; Kirk, P.; Rao, A.; Black, K.; Babar, N.; Lo, J.; Moreira, E.; Martins, L. G.; Sanseviero, O.; Gonzalez, L.; Gleicher, Z.; Warkentin, T.; Mirrokni, V.; Senter, E.; Collins, E.; Barral, J.; Ghahramani, Z.; Hadsell, R.; Matias, Y.; Sculley, D.; Petrov, S.; Fiedel, N.; Shazeer, N.; Vinyals, O.; Dean, J.; Hassabis, D.; Kavukcuoglu, K.; Farabet, C.; Buchatskaya, E.; Alayrac, J.-B.; Anil, R.; Dmitry; Lepikhin; Borgeaud, S.; Bachem, O.; Joulin, A.; Andreev, A.; Hardin, C.; Dadashi, R.; and Hussenot, L. 2025. Gemma 3 Technical Report. ArXiv:2503.19786 [cs].

Kasanishi, T.; Isonuma, M.; Mori, J.; and Sakata, I. 2023. SciReviewGen: A Large-scale Dataset for Automatic Literature Review Generation. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 6695–6715. Toronto, Canada: Association for Computational Linguistics.

Kim, S.; Suk, J.; Longpre, S.; Lin, B. Y.; Shin, J.; Welleck, S.; Neubig, G.; Lee, M.; Lee, K.; and Seo, M. 2024. Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 4334–4353. Miami, Florida, USA: Association for Computational Linguistics.

Kitchenham, B.; and Charters, S. 2007. Guidelines for Performing Systematic Literature Reviews in Software Engineering. *Proceedings of the 2nd international workshop on Evidential assessment of software technologies*, 2.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large Language Models are Zero-Shot Reasoners. *Advances in Neural Information Processing Systems*, 35: 22199–22213.

Kraus, S.; Mahto, R. V.; and Walsh, S. T. 2023. The Importance of Literature Reviews in Small Business and Entrepreneurship Research. *Journal of Small Business Management*, 61(3): 1095–1106. Publisher: Routledge _eprint: https://doi.org/10.1080/00472778.2021.1955128.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Lind, J. 2014. *A Treatise of the Scurvy, in Three Parts: Containing an Inquiry into the Nature, Causes, and Cure, of that Disease*. Cambridge Library Collection - History of Medicine. Cambridge: Cambridge University Press. ISBN 978-1-108-06998-4.

Lo, K.; Wang, L. L.; Neumann, M.; Kinney, R.; and Weld, D. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4969–4983. Online: Association for Computational Linguistics.

Marshall, I. J.; and Wallace, B. C. 2019. Toward Systematic Review Automation: A Practical Guide to Using Machine Learning Tools in Research Synthesis. *Systematic Reviews*, 8(1): 163.

Meerpohl, J. J.; Herrle, F.; Antes, G.; and Elm, E. v. 2012. Scientific Value of Systematic Reviews: Survey of Editors of Core Clinical Journals. *PLOS ONE*, 7(5): e35732. Publisher: Public Library of Science.

Michelson, M.; and Reuter, K. 2019. The Significant Cost of Systematic Reviews and Meta-analyses: A Call for Greater Involvement of Machine Learning to Assess the Promise of Clinical Trials. *Contemporary Clinical Trials Communications*, 16: 100443.

Mohammad, S.; Dorr, B.; Egan, M.; Hassan, A.; Muthukrishan, P.; Qazvinian, V.; Radev, D.; and Zajic, D. 2009. Using Citations to Generate Surveys of Scientific Paradigms. In Ostendorf, M.; Collins, M.; Narayanan, S.; Oard, D. W.; and Vanderwende, L., eds., *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 584–592. Boulder, Colorado: Association for Computational Linguistics.

Ofori-Boateng, R.; Aceves-Martins, M.; Wiratunga, N.; and Moreno-Garcia, C. F. 2024. Towards the Automation of Systematic Reviews Using Natural Language Processing, Machine Learning, and Deep Learning: A Comprehensive Review. *Artificial Intelligence Review*, 57(8): 200.

Orel, E.; Ciglenecki, I.; Thiabaud, A.; Temerev, A.; Calmy, A.; Keiser, O.; and Merzouki, A. 2023. An Automated Literature Review Tool (LiteRev) for Streamlining and Accelerating Research Using Natural Language Processing and Machine Learning: Descriptive Performance Evaluation Study. *Journal of Medical Internet Research*, 25: e39736.

O'Connor, A. M.; Clark, J.; Thomas, J.; Spijker, R.; Kusa, W.; Walker, V. R.; and Bond, M. 2024. Large Language Models, Updates, and Evaluation of Automation Tools for Systematic Reviews: A Summary of Significant Discussions at the Eighth Meeting of the International Collaboration for the Automation of Systematic Reviews (ICASR). *Systematic Reviews*, 13(1): 290.

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. *BMJ*, 372: n71.

Palmatier, R. W.; Houston, M. B.; and Hulland, J. 2018. Review Articles: Purpose, Process, and Structure. *Journal of the Academy of Marketing Science*, 46(1): 1–5.

Qi, R.; Li, W.; and Lyu, H. 2025. Generation of Scientific Literature Surveys Based on Large Language Models (LLM) and Multi-Agent Systems (MAS). In Wong, D. F.; Wei, Z.; and Yang, M., eds., *Natural Language Processing and Chinese Computing*, 169–180. Singapore: Springer Nature. ISBN 978-981-97-9443-0.

Qian, C.; Liu, W.; Liu, H.; Chen, N.; Dang, Y.; Li, J.; Yang, C.; Chen, W.; Su, Y.; Cong, X.; Xu, J.; Li, D.; Liu, Z.; and Sun, M. 2024. ChatDev: Communicative Agents for Software Development. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15174–15186. Bangkok, Thailand: Association for Computational Linguistics.

Randles, R.; and Finnegan, A. 2023. Guidelines for Writing a Systematic Review. *Nurse Education Today*, 125: 105803.

Shao, Y.; Jiang, Y.; Kanell, T.; Xu, P.; Khattab, O.; and Lam, M. 2024. Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 6252–6278. Mexico City, Mexico: Association for Computational Linguistics.

Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; and Yao, S. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, 8634–8652. Red Hook, NY, USA: Curran Associates Inc.

Snyder, H. 2019. Literature Review as a Research Methodology: An Overview and Guidelines. *Journal of Business Research*, 104: 333–339.

Snyder, H.; Witell, L.; Gustafsson, A.; Fombelle, P.; and Kristensson, P. 2016. Identifying Categories of Service Innovation: A Review and Synthesis of the Literature. *Journal of Business Research*, 69(7): 2401–2408.

Sumers, T.; Yao, S.; Narasimhan, K.; and Griffiths, T. 2023. Cognitive Architectures for Language Agents. *Transactions on Machine Learning Research*.

Tang, X.; Zou, A.; Zhang, Z.; Li, Z.; Zhao, Y.; Zhang, X.; Cohan, A.; and Gerstein, M. 2024. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 599–621. Bangkok, Thailand: Association for Computational Linguistics.

Tian, Y.; Gu, X.; Li, A.; Zhang, H.; Xu, R.; Li, Y.; and Liu, M. 2025. Overview of the NLPCC2024 Shared Task 6: Scientific Literature Survey Generation. In Wong, D. F.; Wei, Z.; and Yang, M., eds., *Natural Language Processing and Chinese Computing*, 400–408. Singapore: Springer Nature. ISBN 978-981-97-9443-0.

Wang, Y.; Guo, Q.; Yao, W.; Zhang, H.; Zhang, X.; Wu, Z.; Zhang, M.; Dai, X.; Zhang, M.; Wen, Q.; Ye, W.; Zhang, S.; and Zhang, Y. 2024. AutoSurvey: Large Language Models Can Automatically Write Surveys. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems*.

Watson, W.; Cho, N.; Srishankar, N.; Zeng, Z.; Cecchi, L.; Scott, D.; Siddagangappa, S.; Kaur, R.; Balch, T.; and Veloso, M. 2025. LAW: Legal Agentic Workflows for Custody and Fund Services Contracts. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; Schockaert, S.; Darwish, K.; and Agarwal, A., eds., *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, 583–594. Abu Dhabi, UAE: Association for Computational Linguistics.

Xu, F.; Song, Y.; Iyyer, M.; and Choi, E. 2023. A Critical Evaluation of Evaluations for Long-form Question Answering. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3225–3245. Toronto, Canada: Association for Computational Linguistics.

Xu, W.; Zhu, G.; Zhao, X.; Pan, L.; Li, L.; and Wang, W. 2024. Pride and Prejudice: LLM Amplifies Self-Bias in Self-Refinement. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15474–15492. Bangkok, Thailand: Association for Computational Linguistics.

A Metrics Measuring the Similarity with Human-Written Articles

This section includes an in-depth explanation of how the textual similarity metrics were implemented and how they function.

- **ROUGE-L**: This metric is frequently used for text evaluation and functions by counting the number of recalled units in a text from the reference (Lin 2004). For this project, we used the ROUGE-L metric from the HuggingFace `evaluate` package.
- **Heading Soft Recall:** Inspired by the idea of soft recall (Fränti and Mariescu-Istodor 2023), this metric quantitatively measures the heading coverage of a review without relying on exact matches, where a higher score indicates a more comprehensive outline by virtue of being similar to the human-written one. Here, we use the original implementation provided by Shao et al. (2024).

  Let S be a set, and let both P and G be sets of predicted and gold headers, respectively. For each item in S, the soft count is defined as the inverse sum of its similarity to all other items within S, denoted as follows:

$$\text{count}(S_i) = \frac{1}{\sum_{j=1}^{K} \text{Sim}(S_i, S_j)},$$

$$\text{Sim}(S_i, S_j) = \cos\left(\text{embed}(S_i), \text{embed}(S_j)\right),$$

where embed$(\cdot)$ is parametrized by the `parahprase-MiniLM-L6-v2` model provided by Sentence-Transformers. This is then used to measure the cardinality of the set, defined as the sum of its individual items:

$$\text{card}(S) = \sum_{i=1}^{K} \text{count}(S_i).$$

The heading soft recall (hsr) is then defined as:

$$\text{hsr} = \frac{\text{card}(G \cap P)}{\text{card}(G)},$$
$$\text{card}(G \cap P) = \text{card}(G) + \text{card}(P) - \text{card}(G \cup P).$$

- **Heading and Article Entity Recall:** These metrics indicate the proportion of named entities in the baseline gold review mentioned in the generated text, between headings or the full review texts, respectively. The implementation for this is also taken from Shao et al. (2024). However, due to the presence of more scientific terms, we instead use the SciSpacy named entity tagger for all settings, specifically using the `en_core_sci_lg` model, as it has the largest available vocabulary.

B LLM Evaluation

Here, we outline the details for LLM evaluation. For the specific LLM model, we used `prometheus-7b-v2.0` using (Kim et al. 2024) with the same configurations outlined in Shao et al. (2024), namely a temperature of 0.01, cumulative probability of top tokens of 0.95, maximum number of new tokens of 512, and repetition penalty of 1.03. Also, because of hardware limitations, the model had to be run with fp8-quantization and a memory usage value of 0.8. Note that no seeding was used here, with the justification being that it would encourage more diverse outputs from the model, thus better mimicking reviews made by separate experts. Moreover, to simulate the process of evaluation by Subject Matter Experts (SMEs) and account for randomness, the evaluation was performed three times for all articles, with the final score being the average between all runs. Furthermore, due to the model's context size, the evaluation for the criteria had to be done in the following ways:

- For coverage and relevance, each generated review and its original article were split into their sections, with each section being evaluated separately using the same criteria. If any section is still too long for any of the two, then said section is split further into two parts each. The criteria for determining this is if the combined length of the generated and original section exceeds the context window (32,768 tokens) minus roughly three thousand tokens to account for the rubric and other components.
- For structure, evaluation was performed using only the outlines of both articles. This is because they are reasonably short and serve as a proper approximation for how the respective contents are arranged.

The rubric provided (shown in table 7) to the LLM contains scores ranging from 1 to 5, with the coverage, structure, and relevance evaluated for all generated literature reviews. Furthermore, the description for each score bracket is specified to accurately reflect the proper gradings. These were adapted from the rubric used by Wang et al. (2024), with minor adjustments to accommodate for the output structure. Also, because three runs were performed per setting, the results reported are the means of the average scores and the averaged resulting standard deviations.

C Sample SME Evaluation

We include a sample annotation for the SciReviewGen dataset in figure 3. It can be seen that the SMEs were provided with the title, article snippet, and outline, which were then used to determine which article was qualitatively better.

D Citation Quality

This section provides an overview of how the citation quality F1-score functions. It uses an LLM (`gpt-4o-mini`) to evaluate how well the article cites the given reference abstracts. More precisely, for a given article containing a set of claims $|C|$, an LLM performing Natural Language Inference denoted as m determines the entailment of a claim c_i given the references listed $Ref_i = \{r_{i_1}, r_{i_2}, \cdots\}$, returning 1 if the references support the claim and 0 otherwise.

This is then used to calculate the citation recall (R_c) and precision (P_c), which are defined as the following:

$$R_c = \frac{\sum_{i=1}^{|C|} m(c_i, Ref_i)}{|C|},$$
$$P_c = \frac{\sum_{i=1}^{|C|} \sum_{j=1}^{|Ref_i|} m(c_i, Ref_i) \cap n(c_i, r_{i_j})}{\sum_{j=1}^{|C|} |Ref_i|},$$

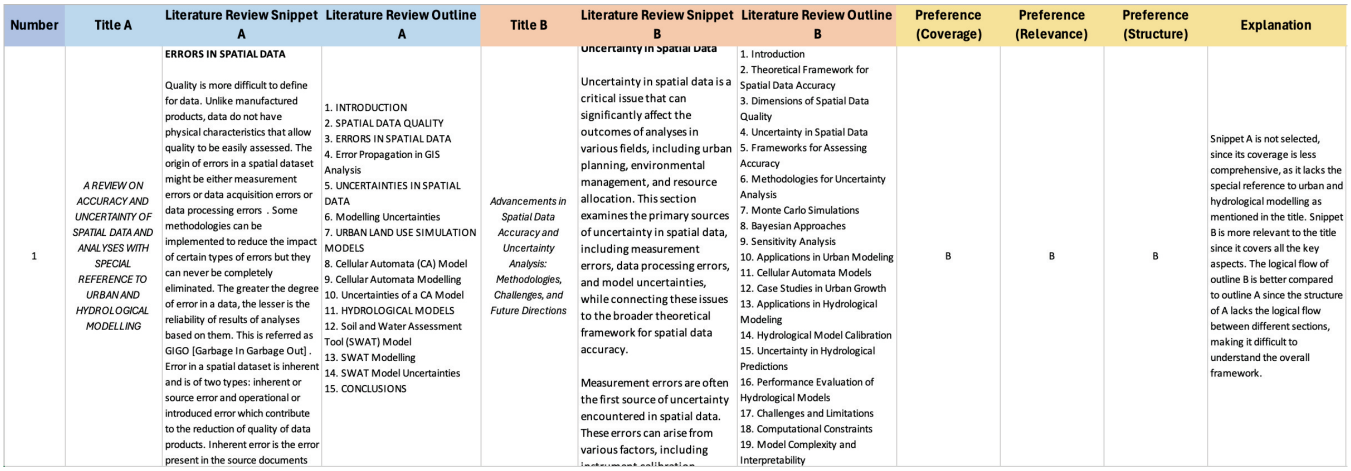
where $n(c_i, r_{i_j}) = (m(c_i, \{r_{i_j}\}) = 1) \cup (m(c_i, Ref_i \setminus \{r_{i_j}\}) = 0)$, which indicates if the paper r_{i_j} is related to the claim c_i.

Given the consideration that longer articles are more difficult to cite properly for, a scaling factor was included for the recall, as it is reliant on the number of claims present. We argue that for a given article, the amount of effort required to properly cite references for the claims is roughly exponential due to the amount of information which has to be remembered from each paper and the relations between all papers. For example, if given references $Ref = \{r_1, r_2, r_3\}$, not only must one consider the relations between individual pairs $((r_1, r_2), (r_2, r_3), (r_1, r_3))$, but also the potential relations between sets of papers. Thus, if a generated article is too short, its performance is penalized accordingly.

The scaling factor is defined as the following:

$$\text{Scaling} = 1 - e^{(-k \times n_{\text{claims}})},$$

where k is a scaling factor (set as 0.01 based on the assumption that a paper has roughly 100 citations on average) and n_{claims} is the average number of claims per article across the generated articles. By definition, this value asymptotically approaches 1 due to the exponential term when n is large enough. This is applied to the aggregate recall score to smooth out the weighting, which is then used to calculate a final F1-score.



Number	Title A	Literature Review Snippet A	Literature Review Outline A	Title B	Literature Review Snippet B	Literature Review Outline B	Preference (Coverage)	Preference (Relevance)	Preference (Structure)	Explanation
1	*A REVIEW ON ACCURACY AND UNCERTAINTY OF SPATIAL DATA AND ANALYSES WITH SPECIAL REFERENCE TO URBAN AND HYDROLOGICAL MODELLING*	**ERRORS IN SPATIAL DATA** Quality is more difficult to define for data. Unlike manufactured products, data do not have physical characteristics that allow quality to be easily assessed. The origin of errors in a spatial dataset might be either measurement errors or data acquisition errors or data processing errors . Some methodologies can be implemented to reduce the impact of certain types of errors but they can never be completely eliminated. The greater the degree of error in a data, the lesser is the reliability of results of analyses based on them. This is referred as GIGO [Garbage In Garbage Out] . Error in a spatial dataset is inherent and is of two types: inherent or source error and operational or introduced error which contribute to the reduction of quality of data products. Inherent error is the error present in the source documents	1. INTRODUCTION 2. SPATIAL DATA QUALITY 3. ERRORS IN SPATIAL DATA 4. Error Propagation in GIS Analysis 5. UNCERTAINTIES IN SPATIAL DATA 6. Modelling Uncertainties 7. URBAN LAND USE SIMULATION MODELS 8. Cellular Automata (CA) Model 9. Cellular Automata Modelling 10. Uncertainties of a CA Model 11. HYDROLOGICAL MODELS 12. Soil and Water Assessment Tool (SWAT) Model 13. SWAT Modelling 14. SWAT Model Uncertainties 15. CONCLUSIONS	*Advancements in Spatial Data Accuracy and Uncertainty Analysis: Methodologies, Challenges, and Future Directions*	**Uncertainty in Spatial Data** Uncertainty in spatial data is a critical issue that can significantly affect the outcomes of analyses in various fields, including urban planning, environmental management, and resource allocation. This section examines the primary sources of uncertainty in spatial data, including measurement errors, data processing errors, and model uncertainties, while connecting these issues to the broader theoretical framework for spatial data accuracy. Measurement errors are often the first source of uncertainty encountered in spatial data. These errors can arise from various factors, including instrument calibration	1. Introduction 2. Theoretical Framework for Spatial Data Accuracy 3. Dimensions of Spatial Data Quality 4. Uncertainty in Spatial Data 5. Frameworks for Assessing Accuracy 6. Methodologies for Uncertainty Analysis 7. Monte Carlo Simulations 8. Bayesian Approaches 9. Sensitivity Analysis 10. Applications in Urban Modeling 11. Cellular Automata Models 12. Case Studies in Urban Growth 13. Applications in Hydrological Modeling 14. Hydrological Model Calibration 15. Uncertainty in Hydrological Predictions 16. Performance Evaluation of Hydrological Models 17. Challenges and Limitations 18. Computational Constraints 19. Model Complexity and Interpretability	B	B	B	Snippet A is not selected, since its coverage is less comprehensive, as it lacks the special reference to urban and hydrological modelling as mentioned in the title. Snippet B is more relevant to the title since it covers all the key aspects. The logical flow of outline B is better compared to outline A since the structure of A lacks the logical flow between different sections, making it difficult to understand the overall framework.

Figure 3: Annotation sample for the SciReviewGen dataset.

For the references, only the abstracts were used, as doing so would replicate the process of how SMEs generally perform the same task. Furthermore, the implementation was taken from Wang et al. (2024) with some adjustments to better handle longer inputs. Specifically, assuming a claim-reference pair is too long, the reference is split into a few sections, and then the metric checks if any split supports the claim. This assumes the rest of the references are also supported by the sources as a result.

E Ablation and Modification

Figures 4 and 5 show the results for additional studies relating to system design modification. More precisely, this involves two experiments, with the first being an ablation study of the editor model. This was performed because this component can be considered less impactful towards the output compared to most other agents in the system design.

The other experiment performed was the addition of a researcher agent. As a method of better imitating the process of literature review writing by humans, this agent takes each reference abstract and analyzes the contents based on several pre-determined questions based on existing literature (Snyder 2019). This results in a list of findings (i.e., key points, limitations) which are then utilized by the other agents in place of the original abstract.

The results indicate that the editor model has a generally small but negative impact towards the final results, though it does seem to improve the article coverage and relevance slightly in both cases. Meanwhile, the addition of a researcher agent also tends to decrease overall performance, again with marginal differences except in some metrics. As such, these results suggest that the editor agent in **LiRA** provides a trade-off in terms of article quality and similarity, while the researcher agent as it currently stands tends to reduce performance, likely because the abstracts are already full of information and do not require further processing.

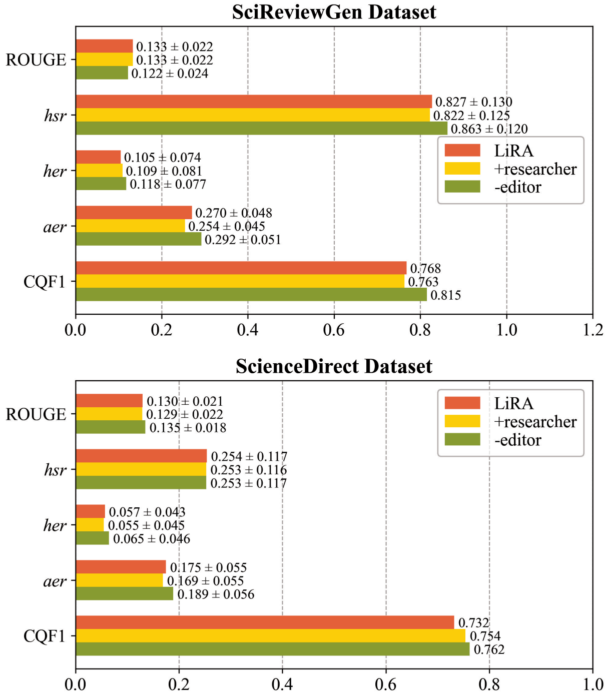


Figure 4: Textual similarity and citation quality results for the different reviewer model setting.

F Retrieval Setting Statistical Tests

We include the statistical significance results for all results related to the retriever setting comparison in table 6. For all ROUGE- and recall-based metrics, the two-tailed Mann-Whitney U-test was used, as almost all of the results were not normally distributed (determined using the Shapiro-Wilk test on each set of results). Additionally, the results for the Prometheus gradings use the same test, though the results across all runs were averaged by sample first to combine the

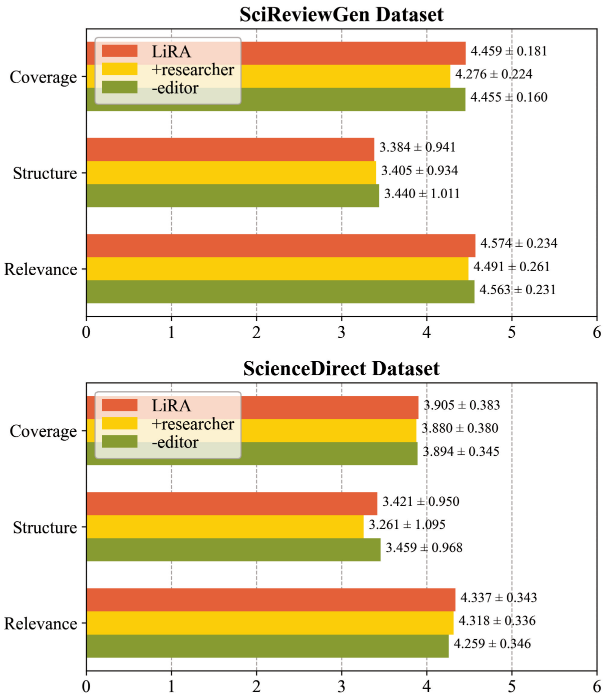


Figure 5: Prometheus evaluation results for the different reviewer model setting.

Metric	**U Statistic value**	**p-value**
ROUGE-L	8503.5	.226
hsr	8041.0	.690
her	8414.5	.292
aer	10097.5	**<.001**
Coverage	8714.5	.114
Structure	7909.5	.864
Relevance	8970.5	**.042**

Table 6: The statistical significance test results for the retriever experiment setting.

results first by group.

All tests and normality checks were performed using the corresponding functions available in the `scipy.stats` module. Moreover, the significance level was chosen to be $\alpha = 0.05$.

Criteria	**Score**	**Description**
Coverage	5	The literature review section comprehensively covers all key and peripheral topics, providing detailed discussions and extensive information.
	4	The literature review section covers most key areas of the topic comprehensively, with only very minor topics left out.
	3	The literature review section is generally comprehensive in coverage but still misses a few key points that are not fully discussed.
	2	The literature review section covers some parts of the topic but has noticeable omissions, with significant areas either underrepresented or missing.
	1	The literature review section has very limited coverage, only touching on a small portion of the topic and lacking discussion on key areas.
Structure	5	The literature review outline is tightly structured and logically clear, with all sections and content arranged most reasonably, and transitions between adjacent sections smooth without redundancy.
	4	The literature review outline has good logical consistency, with content well arranged and natural transitions, only slightly rigid in a few parts.
	3	The literature review outline has a generally reasonable logical structure, with most content arranged orderly, though some links and transitions could be improved such as repeated subsections.
	2	The literature review outline has weak logical flow with some content arranged in a disordered or unreasonable manner.
	1	The literature review outline lacks logic, with no clear connections between sections, making it difficult to understand the overall framework.
Relevance	5	The literature review section is exceptionally focused and entirely on topic; the article is tightly centered on the subject, with every piece of information contributing to a comprehensive understanding of the topic.
	4	The literature review section is mostly on topic and focused; the narrative has a consistent relevance to the core subject with infrequent digressions.
	3	The literature review section is generally on topic, despite a few unrelated details.
	2	The literature review section is somewhat on topic but with several digressions; the core subject is evident but not consistently adhered to.
	1	The literature review section is outdated or unrelated to the field it purports to review, offering no alignment with the topic.

Table 7: The rubric criteria alongside their descriptions, adapted from (Wang et al. 2024).