Completing A Systematic Review in Hours instead of Months with Interactive AI Agents

Rui Qiu[1]* **Shijie Chen**[1]* **Yu Su**[1] **Po-Yin Yen**[2] **Han-Wei Shen**[1]
[1]The Ohio State University [2]Washington University School of Medicine
{qiu.580, chen.10216, su.809, shen.94}@osu.edu, yenp@wustl.edu

Abstract

Systematic reviews (SRs) are vital for evidence-based practice in high stakes disciplines, such as healthcare, but are often impeded by labor-intensive and lengthy processes that can span months. Due to the high demand for domain expertise, existing automatic summarization methods fail to accurately identify relevant studies and generate high-quality summaries. To that end, we introduce *Insight*Agent, a human-centered interactive AI agent powered by large language models that revolutionizes the systematic review workflow. *Insight*Agent partitions a large literature corpus based on semantics and employs a multi-agent design for more focused processing of literature, leading to significant improvement in the quality of generated SRs. *Insight*Agent also provides intuitive visualizations of the corpus and agent trajectories, allowing users to effortlessly monitor the actions of the agent and provide real-time feedback based on their expertise. Our user studies with 9 medical professionals demonstrate that the visualization and interaction mechanisms can effectively improve the quality of synthesized SRs by 27.2%, reaching 79.7% of human-written quality. At the same time, user satisfaction is improved by 34.4%. With *Insight*Agent, it only takes a clinician about 1.5 hours, rather than months, to complete a high-quality systematic review. *Insight*Agent demonstrates great potential in facilitating more timely and informed decision-making in high stake application scenarios[1].

1 Introduction

Systematic reviews (SRs) are the cornerstone of evidence-based practice (EBP) across high stakes disciplines, such as healthcare, providing comprehensive synthesis of research evidence to inform clinical decision-making (Majid et al., 2011). Notably, the number of published SRs indexed in PubMed per year has increased from less than 50 in the 1990s to almost 36,000 in 2022 (Brignardello-Petersen et al., 2025), indicating huge amount of human effort have been dedicated to conducting SRs. Nevertheless, conducting systematic reviews remains a labor-intensive and time-consuming process that can take months to complete (Chandler et al., 2019; Hanney et al., 2015).

The systematic review process comprises several key steps: formulating a research question, collecting a corpus of literature, defining inclusion and exclusion criteria, screening relevant records, summarizing these studies, synthesizing the findings, and generating a final report. While the initial steps are well-supported by information retrieval tools, the latter stages — specifically, record screening, literature summarization, and finding synthesis — remain significant bottlenecks (Hanney et al., 2015). These steps demand intensive effort in reading, comprehending, and integrating information from a large volume of studies, posing challenges in efficiency and consistency.

So far, automation techniques for SRs mainly focus on record screening and shows varying sensitivity (recall), resulting in only limited adoption and modest saving of manual labor (Tóth et al., 2024). For the evidence synthesis stage, more recent literature review systems based on large language models, such as ChatCite (Li et al., 2024) and AutoSurvey (Wang et al., 2024), still overlook subtle but important details, resulting in generic summaries with untraceable sources, making them unsuitable for rigorous systematic reviews.

To overcome these limitations, we introduce ***Insight*Agent**, the first automated system capable of generating high-quality systematic reviews. We propose a human-centered agent framework that equips language agents (Su et al., 2024) with a user-friendly graphical interface which enables real-time

*Equal contribution.
[1]Code and data are available at: https://github.com/OSU-NLP-Group/InsightAgent.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 31559–31593
July 27 - August 1, 2025

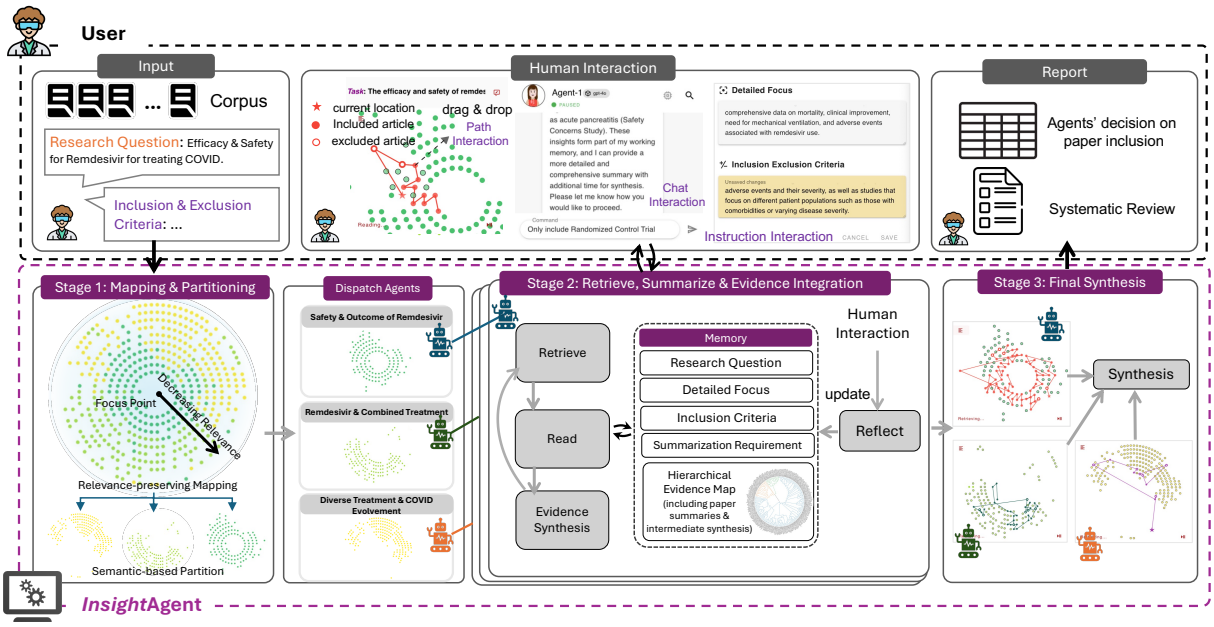

Figure 1: Overview of the *Insight*Agent workflow. In Stage 1, the corpus is mapped into semantic clusters. In Stage 2, multiple agents concurrently read and synthesize evidence under real-time user guidance for each cluster. Finally, in Stage 3, findings of all agents are integrated into a complete systematic review.

user oversight using visualization and the incorporation of user expertise via interactions. As illustrated in Figure 1, *Insight*Agent projects a large corpus into a circular *relevance-preserving map* (Qiu et al., 2024), where more relevant articles appear toward the center and semantically similar articles are clustered together, and dispatches multiple agents to read and summarize these clusters in parallel. This multi-agent design is inspired by the recommended multi-reviewer strategies in systematic reviews (Chandler et al., 2019), where assigning different subsets of studies to independent reviewers can reduce individual bias and accelerate the initial screening. Each agent then explores the cluster from the center, identifies and reads relevant articles, and integrates findings into an interim SR. In this process, *Insight*Agent simultaneously generates a provenance tree that clearly tracks supporting articles for each interim finding.

With these visualization techniques, users can intuitively monitor the reading trajectory of each agent and intervene via various types of interactions to adjust the focus of the agent when needed. At last, users can inspect the provenance tree to ensure the summaries in the output SR are properly supported by evidences.

We comprehensively evaluate the effectiveness of *Insight*Agent using 15 existing systematic reviews in the biomedical domain with 9 domain experts and medical students. With GPT-4o (Hurst et al., 2024) as the backbone, the multi-agent design of *Insight*Agent allows it to produce systematic reviews with 15.6% higher quality than AutoSurvey. Interactions with users further improve article identification accuracy by 47% (F1 points), generated review quality by 27.2%, and overall user satisfaction by 34.4%. On average, a clinician can complete a systematic review in 1.5 hours using *Insight*Agent and reach 79.7% of human-written quality. *Insight*Agent drastically cuts the time needed for crafting a high-quality systematic review from months to hours, demonstrating the great potential of human-centered AI agents in accelerating evidence-based discoveries.

2 Related Work

2.1 Language Agents for Literature Survey

Recent advancements in large language models (LLMs) have opened new possibilities for automate the literature review process. Several works have applied LLMs to automate literature reviews. AutoSurvey (Wang et al., 2024) generates literature summaries by constructing an outline and progressively refining it, while ChatCite (Li et al., 2024) extracts key elements from research papers and incrementally generates task-specific summaries. LitLLM (Agarwal et al., 2024) retrieves papers through keyword-based queries and produces sum-

maries using zero-shot generation methods. Following a similar paradigm, Lai et al. (2024) take a step-by-step approach and generate sections of a literature survey in sequence; Iyer et al. (2024) facilitate semantic exploration of astronomical literature using LLMs to improve context-based retrieval.

So far, existing LLM agents for literature review mostly operate in a fully autonomous fashion. The lack of user interaction and transparency in these systems presents significant limitations. Autonomous agents without human involvement often struggle to maintain coherence and transparency in their decision-making processes. Our proposed system, *Insight*Agent, addresses these gaps by enabling real-time user monitoring and intervention for the agents' decision making through an intuitive graphical user interface. Through a human-centered interface, users can visually monitor agents' tasks, guide their progress, and interact with them to ensure coherence and relevance.

2.2 Visual Analytics for Information-seeking and Decision-making

Visual analytics (VA) methods embed visualization into the data analysis processes and can effectively facilitate decision-making and information-seeking (Isenberg et al., 2016; Lee and Uppal, 2020; Qiu et al., 2022). In the context of information-seeking, VA has been applied primarily in two ways: (1) sense making and interpretability, and (2) retrieval, classification, and decision-making.

Sensemaking and Interpretability. VA systems assist researchers in comprehending thematic and relational structures within extensive document collections. For instance, HINTs (Lee and Ma, 2024) employ hypergraph representations to highlight complex entity-topic relationships, whereas Qiu et al. (2024) utilize adaptive 2D layouts to map documents according to user queries.

Retrieval, Classification, and Decision-Making. VA methodologies also focus on targeted tasks like document retrieval and classification, which are crucial in systematic reviews. Docflow (Qiu et al., 2022) categorizes documents in response to user-specified queries through answer embedding similarity to streamline the record screening process. Studies also suggest that coupling machine learning–based retrieval with interactive visualization can significantly improve precision and recall in document retrieval and information-seeking(da Silva et al., 2023). Beyond retrieval, research has shown that thoughtful interface design reduces cognitive biases (Cho et al., 2017; Oral et al., 2023) and facilitates strategic planning (Nazemi et al., 2022).

Building on these insights, our approach leverages *LLM-driven agents* with a spatial document layout to facilitate systematic reviews, from where users can observe agent actions, refine corpus exploration, and achieve more effective evidence synthesis through a transparent, VA-based interface.

3 *Insight*Agent

In this section, we present *Insight*Agent, a human-centered approach that equips autonomous language agents with an interactive graphical user interface for improved summary quality and user satisfaction in high stakes domains. *Insight*Agent operates in three stages: (1) **corpus mapping and partitioning**, (2) **record screening and evidence synthesis**, and (3) **final synthesis**. In each stage, *Insight*Agent harnesses the capabilities of LLMs and leverage domain knwoledge from expert users, ensuring that the systematic review process is efficient and accurate.

3.1 Stage 1: Corpus Mapping & Partitioning

The first stage of *Insight*Agent aims to project a large biomedical corpus into an intuitive layout and partition it for parallel processing by multiple agents. This mapping not only prepares a more focused action space for each agent, but also provides a visual interface for the user to monitor agent trajectories and intervene when needed (Figure 1).

Corpus Mapping. To visualize the overall structure of the corpus, we use the radial-based relevance and similarity map (RSS map) (Qiu et al., 2024). Each article in the corpus is presented as a dot, whose positions is decided by two factors: (1) relevance to the *research question* - more relevant articles appear closer to the center, and (2) semantic similarity to other articles - semantically similar articles are placed in a nearby region.

Corpus Partitioning. Once the documents are positioned in the radial layout, we apply K-means clustering to partition the corpus into semantically distinct clusters. The optimal number of clusters k is automatically determined by the Elbow method (Onumanyi et al., 2022), which evaluates cluster compactness through within-cluster and inter-cluster distances, resulting in an average of

nine clusters across our evaluations. Subsequently, multiple agents are instantiated and operate in parallel, with each assigned to a separate cluster and an optional reading focus. Users may flexibly refine these clusters and adjust agents' reading based on their domain knowledge and research objectives. This corpus partitioning reduces noise and individual agent workload, significantly enhancing retrieval accuracy and summarization quality. A quantitative comparison demonstrating the effectiveness of our multi-agent design over a single-agent approach is presented in Appendix B.

3.2 Stage 2: Reading and Evidence Synthesis

In this stage, each agent explores its assigned cluster, identifies relevant documents, generates incremental summaries, and integrates newly acquired evidence into their knowledge base. Throughout these processes, *Insight*Agent offers rich visualization and interactive controls so that users can easily monitor and refine the agents' reading strategies to enhance the quality of the final systematic review (Figure 1-Step 2).

Agent Setup and Record Screening. When an agent is created for a given corpus, it is initialized with a *research question* Q_i, *inclusion and exclusion criteria* (e.g., study type), and a *summary requirement* (e.g., desired level of detail or a specific analytic focus). The agent then begins screening articles from the center of the RSS map—corresponding to the most relevant documents—and progressively moves outward. At each step, the agent selects its next article from a defined *receptive field*, consisting of the immediate neighbors of the current document within the relevance-preserving map. To maintain consistency across all agent operations, we standardize this receptive field to always include eight nearest neighboring articles. This design ensures agents systematically navigate through the corpus, reducing randomness and promoting coherent exploration. As each article is reviewed, the agent determines whether it satisfies the predefined inclusion criteria, dynamically updating its short-term reading strategy accordingly.

Summary Generation & Memory Mechanism. For each relevant article, the agent summarizes the key findings and their relevance to Q_i. These summaries are stored in a *local memory*, along with metadata such as timestamps and article embeddings. Whenever the agent encounters overlapping or contradictory information, it merges existing summaries with the new one:

$$M_{k+1} = f(M_k, S_j), \tag{1}$$

where S_j is the freshly generated summary for the current document, and M_k is a previously stored interim evidence synthesis. This incremental evidence synthesis process avoids redundancy and gradually constructs a coherent subdomain knowledge base. Importantly, each agent's memory remains isolated from others until the final synthesis stage, allowing it to develop a specialized perspective on its assigned topic.

Transparent Evidence Integration. To maintain accountability for how each conclusion is formed, *Insight*Agent logs every summary merge and evidence update in a *dependency graph*, which functions as an attribution or provenance structure. Leaf nodes in this tree represent article-level summaries and other nodes represent interim syntheses. Different color-coding or labeling denotes contributions from distinct agents, allowing domain experts to verify sources and scrutinize any disputed findings.

User Interventions. *Insight*Agent offers three types of real-time user interaction interfaces to effectively collaborate with users in the systematic review process:

- **Path Navigation.** On the RSS map, users can *drag* an agent's pointer to a missed relevant article and the agent will read the article next and update its local memory accordingly.
- **Chat Navigation.** Users can issue natural-language directives (e.g., "Focus on randomized controlled trials"), causing the agent to reflect on changes to Q_i, inclusion criteria, or summary requirements. It then revises its retrieval strategy and merges new findings or discards outdated ones.
- **Instruct Navigation.** For more fine-grained control, experts can directly edit the agent's parameters, such as specifying stricter inclusion criteria or change to a different summarization format. Upon receiving these instructions, the agent double-checks relevant memory entries to ensure previously stored summaries align with the updated requirements.

Whenever user interventions alter the agent's behavior, the agent enters a *reflection phase*, during which it reconciles any conflicts in local memory,

and adjusts its reading strategy to be consistent with the latest directives. By combining iterative retrieval, localized summarization, and user-driven oversight, *Insight*Agent builds a flexible, transparent evidence base that will later feed into the final synthesis stage. We report the details of our interface design in appendix 6.

3.3 Stage 3: Final Synthesis

Once the dispatched agents have finished building localized evidence bases, *Insight*Agent integrates these subdomain findings into a coherent final summary following the template specifies by the user. The template usually has several sections, including *Introduction*, *Study Design*, *Key Findings*, *Discussion*, and *Conclusion*. Citations in the resulting extended abstract follow a structured format: *Insight*Agent uses `[citation_number]` to refer back to original documents or interim summaries, which ensures the supporting sources for each argument are clearly traceable. The dependency tree is updated accordingly. We provide an example final synthesis template in Appendix A.

4 Experiments

4.1 Experiment Setup

We perform a comprehensive human evaluation to assess *Insight*Agent's effectiveness in accelerating systematic reviews. Our study involves 15 published systematic reviews. Among them, 13 were published in 2024 and 2 were from 2022 and 2023. We reconstruct the literature corpus following their published search strategy, with corpus sizes ranging from 72 to 7,356 articles (the average inclusion rate is 5.7%). Details of the chosen systematic reviews are provided in Appendix D. Following common practices in record screening and considering the cost and LLM context length constraints, we only use the title and abstract of each article in this study. We invite 9 medical experts with prior experience in publishing systematic reviews to participate in the evaluation.

Our study primarily focuses on four axes:

(1) **Record Screening.** We measure how accurately can *Insight*Agent identify relevant studies in the corpus during record screening. We report precision, recall and F1 scores.

(2) **Systematic Review Quality.** We use a rubric-based method in which each generated report is independently scored by two experts on multiple dimensions, such as comprehensiveness and writing quality, totaling 100 points. Each systematic review is scored by two different experts and we report the average rating. The rubric is collaboratively defined by domain experts, with peer-reviewed human systematic reviews as the ground truth (100 points). Appendix E provides the detailed rubric.

(3) **User Experience** After using each system, participants complete a questionnaire evaluating the usability of the interface, perceived precision, and overall satisfaction. We also conduct brief follow-up interviews for qualitative feedback on transparency, user control, and confidence in the output systematic review. The questionnaire is available in Appendix G.

Evaluation Procedure. Each of the nine experts is randomly assigned two to four systematic reviews, with each systematic review independently evaluated by at least two experts familiar with its topic. To ensure fairness and consistency in comparing different methods (*Insight*Agent, AutoSurvey, and ChatCite), we assign reviews generated by different methods for the same systematic review to the same expert.

While we do not set a time limit, we observe that on average it only takes participants about 1.5 hours to finish a session and produce a systematic review. This marks a substantial speedup compared to manual systematic reviews, which typically takes months.

Implementation Details. We experiment with two popular large language models as the system backbone: the proprietary GPT-4o (Hurst et al., 2024) and the open-weight Llama 3.3 70B (Grattafiori et al., 2024). We use default hyperparameters for text generation.

To evaluate the effectiveness of user interactions, we test two variants: *Insight*Agent, the default configuration that supports real-time user interactions, and *Insight*Agent$_{auto}$, which disables interactions and autonomously completes a systematic review in a few minutes.

Baseline Systems. We compare *Insight*Agent with two recent fully autonomous LLM-based literature review systems:

- **ChatCite** (Li et al., 2024), an incremental reflective summarization system for literature review.

- **AutoSurvey** (Wang et al., 2024), an LLM-based system combining embedding-based retrieval with automated survey generation.

For ChatCite, we follow Li et al. (2024) and implement the key element extractor and comparative generator in biomedical settings. For AutoSurvey, we use the released implementation and retrieve the top 100 articles for each SR for summarization.

4.2 Experiment Results

4.2.1 Record Screening

System	Recall (%)	Precision (%)	F1%
BM25 (Top-100)	54.30	16.50	25.3
ChatCite	–	–	–
AutoSurvey (Top-100)	70.30	20.43	31.6
*Insight*Agent$_{auto}$ (Llama 3.3)	66.40	62.40	64.3
*Insight*Agent (Llama 3.3)	<u>87.90</u>	**80.10**	<u>83.8</u>
*Insight*Agent$_{auto}$ (GPT-4o)	71.10	51.90	60.0
*Insight*Agent (GPT-4o)	**98.50**	<u>79.80</u>	**88.2**

Table 1: Performance in record screening. We bold the best performance and underline the second best.

Table 1 presents the performance of each system in identifying relevant articles. We include BM25 (Robertson and Walker, 1994) for reference. We restrict both BM25 and AutoSurvey's retrieval to top-100, which is a practical cutoff given that all of the systematic reviews in our evaluation include fewer than 100 articles. ChatCite does not perform retrieval, instead it summarizes all user-provided articles. Hence, we omit it from this comparison.

The results indicates that *Insight*Agent$_{auto}$ with both GPT-4o and Llama 3.3 outperform BM25 and AutoSurvey (GPT-4o) by a large margin in record screening, with a substantial advantage in precision (62.4%/41.9% vs 20.4%), demonstrating the effectiveness of our multi-agent design and corpus partitioning strategy in more accurately identify articles relevant to the research question.

Impressively *Insight*Agent further substantially improves both recall and precision, with *Insight*Agent(GPT-4o) reaching a near-perfect 98.5% recall. These results show that our user-centered design with an interactive interface can effectively help users monitor the agent's reading progress and correct agent mistakes based on their domain knowledge. The comprehensive and accurate record screening results lays a solid foundation for *Insight*Agent to generate high-quality systematic reviews in the final synthesis stage.

4.2.2 Quality of the Generated Summaries

System	Score
ChatCite (GPT-4)	47.1
AutoSurvey (GPT-4o)	54.0
*Insight*Agent$_{auto}$ (Llama 3.3)	60.9
*Insight*Agent (Llama 3.3)	<u>70.2</u>
*Insight*Agent$_{auto}$ (GPT-4o)	62.4
*Insight*Agent (GPT-4o)	**79.7**

Table 2: Quality of generated systematic reviews rated by human experts. We bold the best performance and underline the second best.

Table 2 presents the quality of generated systematic reviews evaluated by experts. We observe that *Insight*Agent$_{auto}$ using the weaker Llama 3.3 model already outperform both baselines using more powerful GPT-4 and GPT-4o models. Compared to AutoSurvey, *Insight*Agent$_{auto}$(GPT-4o) improves generated review quality by 8.4 points, marking a significant step forward for fully autonomous agents for systematic reviews.

When human oversight and guidance is available, *Insight*Agent with both base LLMs further show substantial improvement in review quality by 9.3 points (Llama 3.3) and 17.3 points (GPT-4o). *Insight*Agent(GPT-4o) reaches a quality rating of 79.7 points, making it the first system to be practically useful for domain experts to accelerate the systematic review process.

We break down the quality improvement by evaluation dimensions in Figure 2. While *Insight*Agent shows improvement in all aspects, the advantage more prominent in the comprehensiveness and accuracy of reviews and the derived insights and conclusions. Qualitatively, evaluators find that *Insight*Agent consistently produces more comprehensive and relevant summaries than ChatCite and AutoSurvey, which frequently include irrelevant information partly due to their low precision in record screening. Furthermore, *Insight*Agent is able to deliver more reliable findings, sometimes even suggesting novel insights absent in the original human-written SR. At last, participants report that *Insight*Agent's graphical interface and well-designed interaction features make it easier to trace evidence, significantly boosting their confidence in the generated reviews. We present a detailed qualitative comparison for one SR in Appendix H.

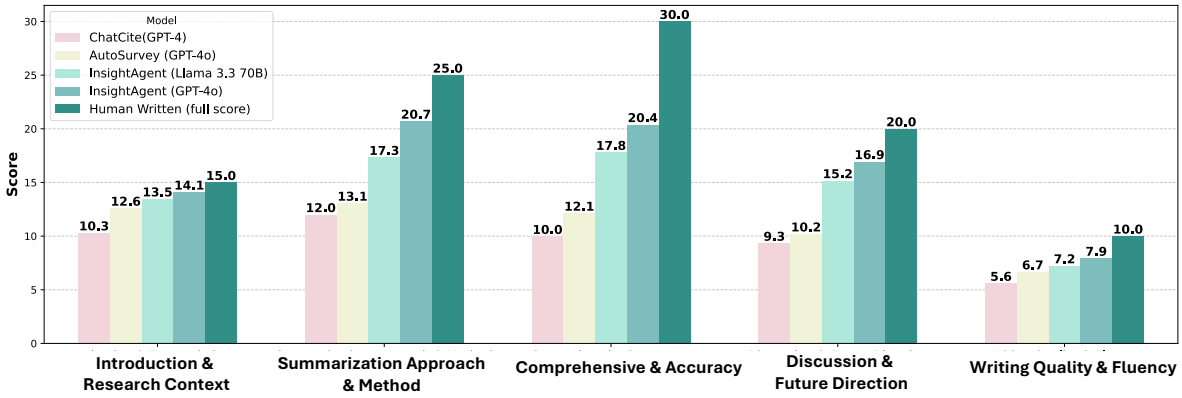

Figure 2: Detailed comparison of summarization quality of *Insight*Agent against ChatCite and AutoSurvey across five evaluation dimensions: Introduction & Research Context, Summarization Approach & Method, Comprehensiveness & Accuracy, Discussion & Future Directions, and Writing Quality & Fluency.

Question	***Insight*Agent$_{auto}$ (Llama 3.3)**	***Insight*Agent (Llama 3.3)**	***Insight*Agent$_{auto}$ (GPT-4o)**	***Insight*Agent (GPT-4o)**
System was easy to use.	2.0/5	3.9/5	2.1/5	**4.0**/5
Confidence in recommendations.	2.6/5	4.2/5	2.8/5	**4.5**/5
Visualizations aided understanding.	2.9/5	**4.2**/5	2.9/5	**4.2**/5
Ability to guide or correct agents.	2.7/5	3.9/5	2.9/5	**4.6**/5
Overall satisfaction.	3.0/5	4.0/5	3.2/5	**4.3**/5

Table 3: Usability & Trustworthiness Likert Scores. We bold the best performance and underline the second best.

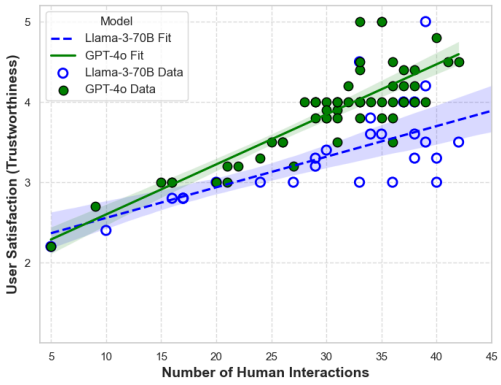

Figure 3: Relationship between the number of human interactions and perceived system trustworthiness for GPT-4o and LLamA 3.3 70B. Scatter points represent individual data samples, while regression lines with confidence bands illustrate the overall trend.

4.2.3 User Experience

After each user study session, we administer a questionnaire to collect feedback on the usability and trustworthiness of *Insight*Agent, shown in Table 3. Participants evaluated *Insight*Agent in both *autonomous* and *interactive* modes. While *Insight*Agent with GPT-4o generally achieves higher scores (e.g., 4.4/5 for "System was easy to use" compared to 3.9/5 for Llama 3.3 70B), both backbone models can benefit from user guidance. For instance, the *confidence in recommendations* metric increases from 2.8/5 to 4.2/5 for GPT-4o and from 2.6/5 to 4.0/5 for Llama 3.3 70B, indicating that real-time oversight and domain-expert input significantly boost trust in the system's output.

To examine how interactions influence the user experience of *Insight*Agent, we plot the user satisfaction scores and the number of interventions for each SR in Figure 3. Overall, both GPT-4o and Llama 3.3 70B exhibit a clear trend: as participants engage in more interventions, their confidence in the system consistently grows. Among the two backbone LLMs, GPT-4o demonstrates stronger abilities in collaborating with users, echoing our results in record screening and final synthesis stages.

These quantitative results clearly demonstrate the practical benefits of incorporating human interaction into *Insight*Agent. A paired t-test confirms that the observed improvements are statistically significant: for *Insight*Agent (GPT-4o), interaction increases summary quality by 27.2% ($p = 3.43 \times 10^{-7}$) and user satisfaction by 34.4% ($p = 1.89 \times 10^{-6}$). These statistically robust improvements underline the substantial impact of human guidance within the review process

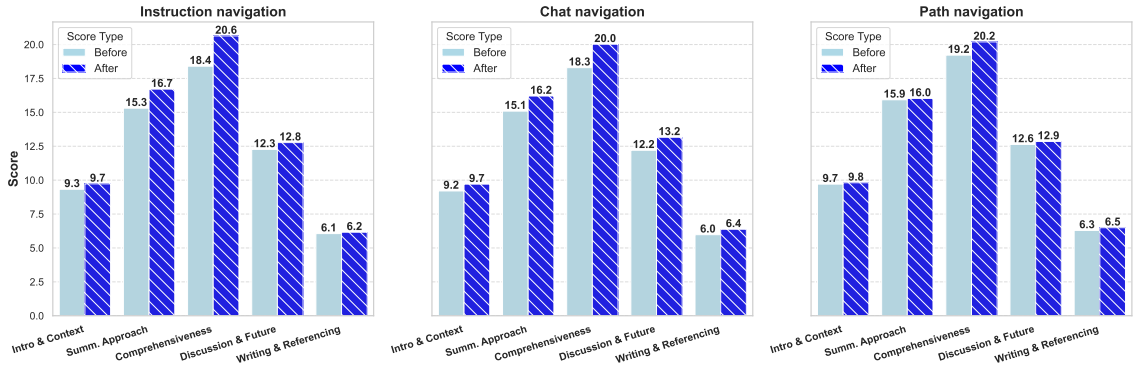

Figure 4: Impact of interaction types on summary quality improvement across different evaluation perspectives. Each subplot compares scores before and after interaction for Instruction, Chat, and Path navigation.

These quantitative findings are further supported by qualitative feedback gathered from post-session interviews. Participants repeatedly highlight the importance of being able to *direct* the agent's reading paths and *monitor* its exploration dynamics through visualizations:

- *"I felt more in control when I could redirect the agent to areas I knew were important."*
- *"Visualizing exploration paths helped me trust that critical topics weren't being missed."*

Overall, these results show that *Insight*Agent's interactive features successfully enhance user confidence through engagement. In the following section, we further show how different forms of user intervention contribute to these improvement.

4.2.4 Effectiveness of Interactions

*Insight*Agent supports diverse types of interactions, offering users flexible control in the systematic review process. To rigorously evaluate the impact of a single user interaction, we use *Insight*Agent$_{auto}$ to complete two interim synthesis before and after an interaction. For each of the three interaction types, we sample 50 such interim synthesis pairs in our user studies and ask participants to score the auto-completed systematic reviews. The results are reported in Figure 4.

Instruction Navigation. By directly revising the research question, inclusion & exclusion criteria, or summarization requirements, users effectively realign the agent's entire reading and synthesis strategy. As illustrated in Figure 4 (left), instruction navigation noticeably improves the coverage of relevant literature and findings in the synthesized SRs. Although participants less frequently use this type of interaction (see Appendix F for usage statistics), they consistently cite it as a powerful way to "reset" or "refine" the agent's focus, thus yielding pronounced benefits in final reporting quality.

Chat Navigation. As shown in Figure 4 (center), chat navigation brings modest improvement across all dimensions. We find chat is used in a flexible way. Some participants use chat primarily for asking clarifying questions without intervening the agent's decision-making, leading to minimal changes in the final summary. In other cases, users leverage chat to propose new angles of investigation or update the summarization format, significantly improving the completeness of the review.

Path Navigation. Figure 4 (right) shows that path navigation also exhibits moderate but consistent improvements, especially in the comprehensiveness of SRs. By pinpointing overlooked articles in the corpus, participants can ensure that relevant studies were included, thereby enhancing the coverage of the final review.

4.3 Error Analysis

While *Insight*Agent has made a significant stride toward automating systematic reviews, we still identify two key limitations upon comparing agent generated SRs with human-written ones (See Appendix H for detailed SRs, evaluations, and analysis). A more detailed error analysis is in Appendix I.

Insufficient Statistical Analysis. Human experts often perform statistical analysis over data from multiple relevant studies to derive numeric evidence that supports rigorous conclusions. While *Insight*Agent shows promise in evidence synthesis, it is not yet capable of performing such analysis.

Limited Planning and Evidence Weighting. Human-generated systematic reviews often control the proportion of articles drawn from different sources, such as high-quality randomized controlled trials (RCTs) versus observational studies (e.g., "60% of data is from RCTs and 40% from observational cohorts") and weight these sources accordingly. While *Insight*Agent can accurately identify relevant studies, it generally treats the articles equally and has limited capacity in considering such global constraints and adjust review plans.

These limitations call for future research in augmenting AI agents systematic reviews with advanced analysis and planning modules.

5 Conclusion

We introduced *Insight*Agent, a human-centered language agent for accelerating systematic reviews. *Insight*Agent adopts a novel multi-agent design and is equipped with an intuitive graphical interface that supports both agent decision monitoring and user interactions, resulting in improved quality and user experience in high stakes domains like healthcare. Through comprehensive human study, we show that a single domain expert can finish a high-quality systematic review in only 1.5 hours, rather than months, and reach 79.7% of human-written quality. Our work demonstrates the great potential of interactive AI agents in accelerating systematic reviews and further facilitate scientific research.

Limitations

Despite promising results, this work faces three primary limitations. First, we conducted a relatively small-scale user study, owing to both the high human effort required for evaluation and the limited availability of fully completed systematic reviews. Second, due to cost and LLM context length limits, this work's setup is restricted to reading only the abstract of each study rather than full texts, potentially omitting critical details that can shape final conclusions. Third, the system lacks the ability to extract and synthesize numerical statistics or weight evidence based on study design; consequently, effect sizes, incidence rates, and other quantitative measures are not rigorously reported, and stronger studies are not afforded greater influence in the synthesized results. We deem developing more advanced systematic review agents capable of quantitative analysis and evidence weighting as promising and important future work directions.

References

Shubham Agarwal, Issam H Laradji, Laurent Charlin, and Christopher Pal. 2024. Litllm: A toolkit for scientific literature review. *arXiv preprint arXiv:2402.01788*.

Pinar Avsar, Declan Patton, Janet Cuddigan, and Zena Moore. 2024. A systematic review on the impact of sub-epidermal moisture assessments on pressure ulcer/injury care delivery pathways. *International Wound Journal*, 21(6):e14928.

Sahar Borna, Michael J Maniaci, Clifton R Haider, Cesar A Gomez-Cabello, Sophia M Pressman, Syed Ali Haider, Bart M Demaerschalk, Jennifer B Cowart, and Antonio Jorge Forte. 2024. Artificial intelligence support for informal patient caregivers: A systematic review. *Bioengineering*, 11(5):483.

Romina Brignardello-Petersen, Nancy Santesso, and Gordon H Guyatt. 2025. Systematic reviews of the literature: an introduction to current methods. *American Journal of Epidemiology*, 194(2):536–542.

Jacqueline Chandler, Miranda Cumpston, Tianjing Li, Matthew J Page, and VJHW Welch. 2019. Cochrane handbook for systematic reviews of interventions. *Hoboken: Wiley*.

Isaac Cho, Ryan Wesslen, Alireza Karduni, Sashank Santhanam, Samira Shaikh, and Wenwen Dou. 2017. The anchoring effect in decision-making with visual analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 116–126. IEEE.

Sherlon Almeida da Silva, Evangelos E Milios, and Maria Cristina F de Oliveira. 2023. Evaluating visual analytics for relevant information retrieval in document collections. *Interacting with Computers*, 35(2):247–261.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Stephen R Hanney, Sophie Castle-Clarke, Jonathan Grant, Susan Guthrie, Chris Henshall, Jorge Mestre-Ferrandiz, Michele Pistollato, Alexandra Pollitt, Jon Sussex, and Steven Wooding. 2015. How long does biomedical research take? studying the time taken between biomedical and health research and its translation into products, policy, and practice. *Health research policy and systems*, 13:1–18.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Karice Hyun, Matthew A Hollings, Nashid Hafiz, Clara Zwack, Caroline Free, Pablo Perel, Clara K Chow,

et al. 2024. Mobile phone text messaging for medication adherence in secondary prevention of cardiovascular disease. *Cochrane Database of Systematic Reviews*, (3).

Kieko Iida, Mina Ishimaru, Mayuko Tsujimura, and Ayumi Wakasugi. 2025. Community-dwelling older people's experiences of advance care planning with health care professionals: a qualitative systematic review. *JBI Evidence Synthesis*, 23(1):69–107.

Petra Isenberg, Tobias Isenberg, Michael Sedlmair, Jian Chen, and Torsten Möller. 2016. Visualization as seen through its research paper keywords. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):771–780.

Kartheik G Iyer, Mikaeel Yunus, Charles O'Neill, Christine Ye, Alina Hyk, Kiera McCormick, Ioana Ciuca, John F Wu, Alberto Accomazzi, Simone Astarita, et al. 2024. pathfinder: A semantic framework for literature review and knowledge discovery in astronomy. *arXiv preprint arXiv:2408.01556*.

Yuxuan Lai, Yupeng Wu, Yidan Wang, Wenpeng Hu, and Chen Zheng. 2024. Instruct large language models to generate scientific literature survey step by step. *arXiv preprint arXiv:2408.07884*.

Gero Langer, Ching Shan Wan, Astrid Fink, Lukas Schwingshackl, and Daniela Schoberer. 2024. Nutritional interventions for preventing and treating pressure ulcers. *Cochrane Database of Systematic Reviews*, (2).

Eva K Lee and Karan Uppal. 2020. Cerc: an interactive content extraction, recognition, and construction tool for clinical and biomedical text. *BMC medical informatics and decision making*, 20(14):1–14.

Sam Yu-Te Lee and Kwan-Liu Ma. 2024. Hints: Sensemaking on large collections of documents with hypergraph visualization and intelligent agents. *arXiv preprint arXiv:2403.02752*.

Yutong Li, Lu Chen, Aiwei Liu, Kai Yu, and Lijie Wen. 2024. Chatcite: Llm agent with human workflow guidance for comparative literature summary. *arXiv preprint arXiv:2403.02574*.

Shaheen Majid, Schubert Foo, Brendan Luyt, Xue Zhang, Yin-Leng Theng, Yun-Ke Chang, and Intan A Mokhtar. 2011. Adopting evidence-based practice in clinical decision making: nurses' perceptions, knowledge, and barriers. *Journal of the Medical Library Association: JMLA*, 99(3):229.

IOM ROUNDTABLE ON EVIDENCE-BASED MEDICINE. 2011. Learning what works best: The nation's need for evidence on comparative effectiveness in health care: An issue overview. In *Learning What Works: Infrastructure Required for Comparative Effectiveness Research: Workshop Summary*. National Academies Press (US).

Moses Mukosha, Abigail Hatcher, Wilbroad Mutale, Mwansa Ketty Lubeya, Jamie L Conklin, and Benjamin H Chi. 2024. Prevalence of persistent hypertension following pregnancy complicated by hypertensive disorders in low-and middle-income countries: a systematic review. *Frontiers in Global Women's Health*, 5:1315763.

Kawa Nazemi, Dirk Burkhardt, and Alexander Kock. 2022. Visual analytics for technology and innovation management: An interaction approach for strategic decision making. *Multimedia Tools and Applications*, 81(11):14803–14830.

Adeiza James Onumanyi, Daisy Nkele Molokomme, Sherrin John Isaac, and Adnan M Abu-Mahfouz. 2022. Autoelbow: An automatic elbow detection method for estimating the number of clusters in a dataset. *Applied Sciences*, 12(15):7515.

Emre Oral, Ria Chawla, Michel Wijkstra, Narges Mahyar, and Evanthia Dimara. 2023. From information to choice: A critical inquiry into visualization tools for decision making. *IEEE Transactions on Visualization and Computer Graphics*.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. 2021. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, 372.

Raffaella Panza, Valentina Cattivera, Jacopo Colella, Maria Elisabetta Baldassarre, Manuela Capozza, Luca Zagaroli, Maria Laura Iezzi, Nicola Laforgia, and Maurizio Delvecchio. 2024. Insulin delivery technology for treatment of infants with neonatal diabetes mellitus: A systematic review. *Diabetes Therapy*, 15(11):2293–2308.

Suhyun Park, Jenna Marquard, Robin R Austin, David Pieczkiewicz, Ratchada Jantraporn, and Connie White Delaney. 2024. A systematic review of nurses' perceptions of electronic health record usability based on the human factor goals of satisfaction, performance, and safety. *CIN: Computers, Informatics, Nursing*, 42(3):168–175.

Rui Qiu, Yamei Tu, , Po-Yin Yen, and Han-Wei Shen. 2024. Vadis: A visual analytics pipeline for dynamic document representation and information seeking. *IEEE Transactions on Visualization and Computer Graphics*, 30(2):1533–1548.

Rui Qiu, Yamei Tu, Yu-Shuen Wang, Po-Yin Yen, and Han-Wei Shen. 2022. Docflow: A visual analytics system for question-based document retrieval and categorization. *IEEE Transactions on Visualization and Computer Graphics*.

Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and*

Development in Information Retrieval, organised by Dublin City University, pages 232–241. Springer.

Kelsey J Sharrad, Olatokunbo Sanwo, Sofia Cuevas-Asturias, Kayleigh M Kew, Kristin V Carson-Chahhoud, and Katharine C Pike. 2024. Psychological interventions for asthma in children and adolescents. *Cochrane Database of Systematic Reviews*, (1).

Yu Su, Diyi Yang, Shunyu Yao, and Tao Yu. 2024. Language agents: Foundations, prospects, and risks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 17–24, Miami, Florida, USA. Association for Computational Linguistics.

Anu Surendran, Lisa Beccaria, Sharon Rees, and Peter Mcilveen. 2024. Cognitive mental workload of emergency nursing: A scoping review. *Nursing Open*, 11(2):e2111.

Carol Stephanie C Tan-Lim and Natasha Ann R Esteban-Ipac. 2024. Among patients with covid-19, should remdesivir be used for treatment? a systematic review and meta-analysis. *Acta Medica Philippina*, 58(14):50.

Barbara Tóth, László Berek, László Gulácsi, Márta Péntek, and Zsombor Zrubka. 2024. Automation of systematic reviews of biomedical literature: a scoping review of studies indexed in pubmed. *Systematic reviews*, 13(1):174.

Jennifer K van Heerden, Elizabeth H Louw, Friedrich Thienemann, Mark E Engel, and Brian W Allwood. 2024. The prevalence of pulmonary hypertension in post-tuberculosis and active tuberculosis populations: a systematic review and meta-analysis. *European Respiratory Review*, 33(171).

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, et al. 2024. Autosurvey: Large language models can automatically write surveys. *arXiv preprint arXiv:2406.10252*.

Ziyu Wang, Bopeng Qiu, Jie Gao, and Juan Del Coso. 2022. Effects of caffeine intake on endurance running performance and time to exhaustion: A systematic review and meta-analysis. *Nutrients*, 15(1):148.

Chuncheng Wu, Ping Zhao, Ping Xu, Chaomin Wan, Surjit Singh, Shoban Babu Varthya, and Shuang-Hong Luo. 2024. Evening versus morning dosing regimen drug therapy for hypertension. *Cochrane Database of Systematic Reviews*, (2).

Mengxin Xie, Tianjiao Tang, and Hongsheng Liang. 2023. Efficacy of single-pill combination in uncontrolled essential hypertension: A systematic review and network meta-analysis. *Clinical Cardiology*, 46(8):886–898.

A Prompt Template

we present the complete prompt in *Insight*Agent for retrieve (Prompt 1), read (Prompt 2), synthesis (Prompt 3), and reflection (Prompt 4) actions.

B Quantitative Evaluation of Multi-Agent Design

To quantitatively demonstrate the effectiveness of the multi-agent design, we compare the retrieval performance of InsightAgent under two different conditions: an automatically determined number of clusters (multi-agent scenario) and a single-cluster scenario (single-agent). The evaluation metrics considered include recall, precision, and the F1 score. Results are summarized in Table 4, which indicates that the multi-agent configuration, using an automatically determined number of clusters (averaging nine), consistently outperforms the single-agent scenario across both evaluated LLM backbones (Llama 3.3 and GPT-4o). Specifically, the multi-agent setup yields notably higher recall, precision, and F1 scores, underscoring its superior capability to identify relevant articles accurately. In contrast, the single-agent design results in a substantial performance decline, highlighting the importance of corpus partitioning and parallel processing for ensuring the quality of review output.

C *Insight*Agent Interface

*Insight*Agent is designed to give users maximum flexibility in monitoring systematic reviews (Figure 6). The main *Canvas* (**top center**) hosts an *infinite scrollable area* where researchers can place one or more *Environment* components (*e.g.*, E1, E2, E3). Each environment projects a corpus of documents onto an interactive 2D map, displaying article distributions and agent "line-of-action" paths for reading. By hovering over any dot in an environment, users can view that article's metadata (title, authors, abstract, etc.), then decide which agent should read or skip it. This canvas-based design affords more fluid iteration than a static layout, letting users freely reposition cards, open new environments, or consolidate findings as needed.

The *Corpus View* (C1) displays metadata for each document. It tracks which agent has read a given paper, whether it was included or excluded, and can export those decisions in CSV form for reference or reporting. The *Chat Window* (A1) offers a direct interface for conversations with a chosen agent, enabling chat-based instructions or updates

Table 4: Retrieval Performance Comparison Between Single-Agent and Multi-Agent Designs

System Configuration	Recall (%)	Precision (%)	F1 (%)
InsightAgent$_{auto}$ (Llama 3.3) - K$_{auto}$ (9 on avg.)	66.4	62.4	64.3
InsightAgent$_{auto}$ (Llama 3.3) - K = 1	56.6	52.0	54.2
InsightAgent$_{auto}$ (GPT-4o) - K$_{auto}$ (9 on avg.)	71.1	51.9	60.0
InsightAgent$_{auto}$ (GPT-4o) - K = 1	60.9	44.7	51.6

(such as clarifying a misunderstanding or adjusting the focus of the agent's reading). Alongside the chat, each agent's *Memory Hierarchy* (A1–M) visualizes how individual evidence items are synthesized into higher-level summaries. Users can expand or hover over nodes to review summarized findings and cross-check citations, further strengthening traceability.

The menu on the left (A) lists the system's primary components: Environments, Agents, and a Collaboration Panel. Researchers can add multiple agents to a single environment—enabling parallel reading of different subtopics—or move a single agent across multiple environments by simply dragging its icon from one environment to another. The agent's ongoing tasks and progress appear in the *Agent Status* panel on the right, which also displays the research question, detailed focus, inclusion/exclusion criteria, and summarization requirements. These can be edited in real time, representing an *instruction-based interaction* that updates the agent's parameters mid-stream. Meanwhile, "path-based" interactions occur within the environment cards (E2, E3), where the user drags the agent pointer around the projected document space to direct its reading order. Finally, the Collaboration Panel allows multiple agents (Agent 0 and Agent 1, for example) to be grouped, letting them exchange findings or produce a unified synthesis through a shared chat or memory structure.

This flexible interface facilitates a more iterative, user-driven approach than is typical in static review tools. By combining path navigation, direct textual instructions, and real-time agent collaboration, *Insight*Agent broadens researchers' capacity to organize, track, and refine systematic reviews in a manner best suited to their investigative goals.

D Completed Systematic Reviews for Evaluation

We evaluated *Insight*Agent using 15 complete systematic reviews spanning diverse biomedical topics (Table 6). All these reviews are publicly available and we replicate the corpus from PubMed, which are intended for research use.

Each review's search strategy was replicated according to recommended guidelines (Page et al., 2021; Chandler et al., 2019; MEDICINE, 2011) to retrieve its original corpus of candidate articles, and the final set of included studies was recorded from the published review. The table summarizes each review's publication year, title, total retrieved corpus size, and number of ultimately included articles. By systematically re-creating these 15 datasets, we ensure that our framework's performance is assessed on authentic, rigorously vetted reviews, thereby enhancing the validity and comparability of our evaluation.

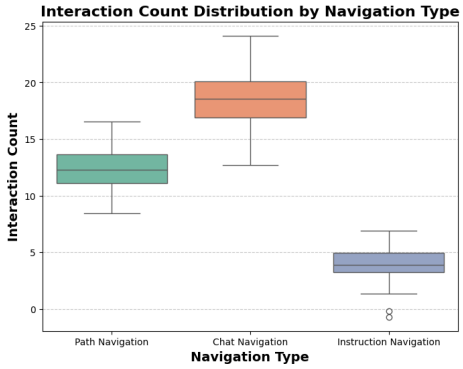

Figure 5: The number of interactions users conducted on average to complete systematic reviews across different navigation types.

E Evaluation Rubric

Our evaluation rubric (Table 7, 8, 9, 10, 11) is developed by domain experts in biomedical research and guided by established systematic review frameworks (Page et al., 2021). To make is suitable for

Table 5: Prevalence of Major Error Types in Extended Abstracts (number of reports flagged / total reports (%)).

Error Type	**ChatCite** ($R = 15$)	**AutoSurvey** ($R = 15$)	**InsightAgent** (Llama-3.3, $R = 30$)	**InsightAgent** (GPT-4o, $R = 59$)
Insufficient Quantitative Evidence	12/15 (80%)	10/15 (67%)	8/30 (27%)	13/59 (22%)
Limited Evidence Weighting	10/15 (67%)	8/15 (53%)	11/30 (37%)	18/59 (31%)
Omitted Heterogeneity Discussion	13/15 (87%)	10/15 (67%)	5/30 (17%)	7/59 (11%)
Hallucinations / Faithfulness Errors	8/15 (53%)	6/15 (40%)	6/30 (20%)	7/59 (12%)

our evaluation setting, we tailor it to agent-based summarization of a biomedical systematic review in extended-abstract format. It evaluates five main dimensions: clarity of the research question and context, methods used (including data extraction and corpus coverage), completeness and accuracy of key biomedical findings, depth of discussion and practical takeaways, and overall writing quality. Each sub-criterion outlines concise "full points" expectations and corresponding deductions, offering a detailed yet focused way to gauge how effectively an extended abstract captures core biomedical evidence and implications within the constraints of a shorter-form document.

F Distribution of Interactions

We report the count of each type of interactions that users performed during the evaluation in Figure 5

G Usability Questionnaire

We conduct a user study inviting domain experts to perform a systematic review using *Insight*Agent, which integrates both a Radial Map and a Hierarchical Map for visualization and interaction. To assess the effectiveness of these designs, we formulate a set of tasks spanning cluster identification, path adjustment, and evidence synthesis navigation. Based on these tasks, and user's overall experience while using the system, we develop a structured questionnaire that covers five core usability categories: ease of use, confidence in recommendations, visualization-aided understanding, ability to guide or correct the agent, and overall satisfaction. **Table 12, 13** list all the questionnaire items, which allowed us to evaluate how effectively our visualizations and interactive features supported systematic review workflows.

H Reports Generated by human, *Insight*Agent, AutoSurvey and ChatCite

We presents systematic review summaries generated by *Insight*Agent (**Table 15**), AutoSurvey (**Table 16**) and ChatCite (**Table 17**), along with a human-conducted abstract reported in the orignal paper (Tan-Lim and Esteban-Ipac, 2024) (**Table 14**). During the user study, participants leveraged 15 path navigation, 24 chat-based interactions, and 1 instruction navigation to refine the generated summary.

Comparison with Human-Curated Summary. **Table 15, 16, 17** present the detailed evaluations based on our evaluation rubric (Appendix E). The evaluation highlights that *Insight*Agent-produced reports captured essential findings with greater numerical specificity, while also offering structured synthesis across multiple perspectives.

Specifically, table 15 indicates that *Insight*Agent generated more extensive numeric details than the manually written review, an observation echoed by multiple participants. For instance, *Insight*Agent highlighted differential efficacy based on patient severity and time to treatment initiation, a nuanced perspective the human-curated report had touched upon only briefly. One expert remarked that *"The system's summary pinpoints how remdesivir's impact may vary depending not just on disease severity but also treatment delay and patient demographics—a useful angle we hadn't fully explored."* *Insight*Agent's structured approach also aligns closely with the human reference in terms of core conclusions, suggesting its capacity to synthesize abstract-level data effectively.

AutoSurvey and ChatCite. Tables 16 and 17 show that while AutoSurvey retrieves multiple potentially relevant abstracts, it often introduces tangential discussions or unverified claims due to broad embedding-based retrieval. Clinical partners noted an "excess of irrelevant evidence" that diluted the final summary's coherence. ChatCite,

by contrast, provided shorter, more direct statements, yet lacked a cohesive structure suitable for systematic reviews, creating an "overwhelming" presentation of scattered facts. Participants found it challenging to assemble ChatCite's unstructured bullet points into a narrative aligning with standard review guidelines.

Strengths and Limitations. Overall, *Insight*Agent delivers a focused, multi-perspective synthesis with greater numeric specificity—enabling rapid exploration of heterogeneous studies that traditionally demand extensive manual labor. In some cases, InsightAgent even surfaced more detailed effect estimates than the human-written review, driven by domain experts interactively prompting for specific statistics to support their evolving queries. However, two key limitations remain. First, our reliance on abstract-level inputs can omit critical numeric details—e.g., the WHO Solidarity trial's mortality comparison (14.5% vs. 15.6%, $p = 0.12$) and subgroup counts across nine RCTs totalling 13,085 patients were not fully captured (Table 15-E_1, E_2). Second, without a dedicated statistical module, the system cannot compute confidence intervals or meta-analytic effect sizes, limiting its ability to perform quantitative synthesis beyond reported values. Additional error categories, such as omitted heterogeneity discussion and hallucinations—are examined in detail in Appendix I. Future work should integrate structured data extraction and lightweight statistical analyses to enrich automatically generated summaries and further close the gap with human expert reviews.

I Detailed Error Analysis

Below we define four principal error categories observed in system-generated SR extended abstracts, describe how each is detected (via rubric subscores or expert feedback), and formalize its prevalence rate:

$$\text{Rate}_e = \frac{N_e}{R},$$

where for each error type e, N_e denotes the number of reports flagged under that category, and R denotes the total number of reports (see Table 5).

1. Insufficient Statistical Analysis & Quantitative Evidence This error occurs when a summary omits concrete numerical synthesis (e.g., pooled effect sizes or incidence rates) that human reviewers would compute across studies. Detection is triggered if either the rubric subscore on item 3.3 ("Use of Quantitative or Specific Evidence") is ≤ 2, or an expert explicitly notes missing statistics. As shown in Table 5, this error affects 80% of ChatCite and 67% of AutoSurvey reports, but is reduced to 27% for InsightAgent (Llama-3.3) and 22% for InsightAgent (GPT-4o). *Example feedback: "The summary notes improved symptom scores but fails to report the 25% mean improvement and its confidence interval."*

2. Limited Evidence Weighting This error arises when the summary treats all studies equally, without accounting for design or quality differences (e.g., RCT vs. observational). We flag a report when rubric subscores on item 2.3 ("Coverage & Representativeness") or 2.5 ("Risk of Bias") are ≤ 2, or an expert comment highlights uniform weighting. InsightAgent variants reduce this error from 67% (ChatCite) and 53% (AutoSurvey) to 37% (Llama-3.3) and 31% (GPT-4o). *Example feedback: "All trials are summarized identically, but the large multi-center RCT should carry greater emphasis than small cohort studies."*

3. Omitted Heterogeneity Discussion A report is flagged if it fails to acknowledge conflicting or subgroup findings (rubric item 3.4 "Variability / Heterogeneity" ≤ 2 or expert notes missing subgroup analysis). Baseline systems omit heterogeneity in 87% (ChatCite) and 67% (AutoSurvey) of cases, whereas InsightAgent lowers this to 17% (Llama-3.3) and 11% (GPT-4o). This reduction reflects how our interactive pipeline allows domain experts to prompt the agent for subgroup or variability checks during review, ensuring conflicting findings are surfaced before final synthesis. *Example feedback: "There is no mention that remdesivir's benefit differs by treatment delay or patient age, which emerged in several trials."*

4. Hallucinations or Faithfulness Errors This error covers any invented or misrepresented facts. We flag a report if rubric item 3.5 ("Faithfulness to Source Material") is ≤ 2 or an expert identifies a discrepancy. Hallucinations occur in 53% of ChatCite and 40% of AutoSurvey summaries, but fall to 20% (Llama-3.3) and 12% (GPT-4o) with InsightAgent. The integrated evidence-tracking graph and real-time review controls enable users to verify each claim's provenance on the fly, substantially limiting faithfulness errors. *Example feedback: "The summary claims remdesivir reduced*

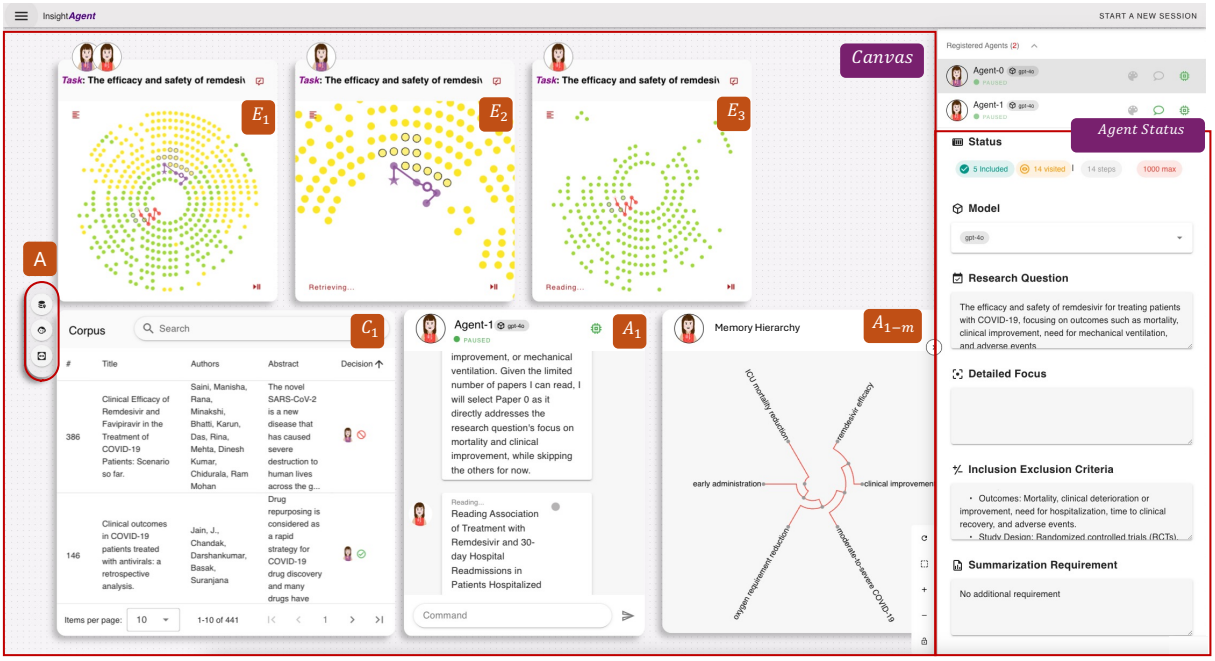

Figure 6: A screenshot of the *Insight*Agent interface while conducting a systematic review. The central *Canvas* is an infinite- scrollable space for creating multiple *Environments* (E1, E2, E3) and attaching any number of agents or collaboration panels. This design allows users to freely drag and drop agents, documents, or synthesis outputs as they refine the review process.

ICU stay by 50%, yet no cited trial reports this figure."

By combining rubric-based thresholds with expert annotations, we achieve both quantitative rigor and real-world validity in pinpointing key failure modes of our agentic synthesis pipeline. The substantial reduction in omitted heterogeneity and faithfulness errors underscores the efficacy of human–in–the–loop oversight, while remaining error rates highlight avenues for enhancing statistical integration and bias mitigation.

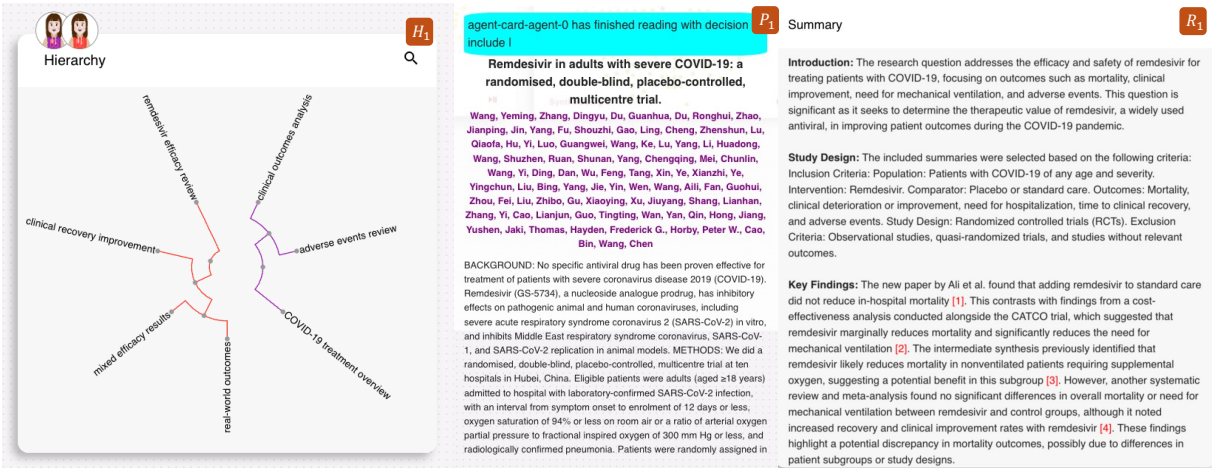

Figure 7: An integrated *Synthesis View* in *Insight*Agent in an early stage of review, featuring two agents collaborating on the same systematic review. The *Hierarchy* panel (left) merges each agent's memory structure into a single visualization, color-coded to indicate which concepts originate from which agent. Hovering over a node reveals the underlying article or synthesized findings (e.g., `P1` for paper details, `R1` for an intermediate summary). This arrangement allows users to examine and reconcile multiple evidence streams in real time, comparing granular article content against higher-level synthesized conclusions.

Retrieve Prompt

You are a biomedical research agent assisting with a systematic review on the question: {query}. You have access to a corpus of documents and must decide which papers to read *during this iteration* based on their titles and initial abstracts. In doing so, you should consider whether each paper can contribute new evidence toward the research question.
Before this iteration, you have already read or processed the following papers:
{paper_already_read}
Below is a summary of any key findings you have discovered thus far:
{findings_so_far}
Here are the newly available papers for you to evaluate:
{available_papers}
Please consult the following inclusion criteria to determine relevance:
{inclusion_criteria}
Select one or more papers (by their indexes) if they provide additional, relevant information for the question {query} with detailed focus {detailed_focus}.
Otherwise, return "skip" if none appear pertinent or if they merely repeat information you have already processed.
Before making your selection, explain your reasoning and the rationale behind your choices.
Inspiration from prior conversation history is shown below: {inspiration_conversation_history}
Your output must be in the JSON format below (and *nothing else*):

```json
  {
    "thought": "<string>",
    "selected_papers": [
      // e.g., "1", "2", or "skip" if none are relevant
    ]
  }
```

Read Prompt

You are a biomedical research agent assisting with a systematic review on the question: {query} with detailed focus {detailed_focus}. Below, you will read a newly assigned paper, extracting any information that may relate to the review's question.
Before this iteration, you have processed:
{paper_already_read}
Key findings so far:
{findings_so_far}
The paper you need read is:
{paper_to_read}
Be aware that this paper might not be relevant to {query} with the detailed focus {detailed_focus}, or should be exclude considering the inclusion exclusion criteria: {inclusion_criteria}. If it is relevant, produce an overall thought reflecting your updated understanding of the review, integrating what you have already discovered. If it is not relevant, explain briefly why you are excluding it.
If the content conflicts with previously read materials, do not resolve the conflict; simply include it in your overall thought for future exploration.
Consider the inspiration from prior conversation history with human before decision making:
{inspiration_conversation_history}
You must respond with a single JSON object, wrapped in an array, following the schema below.

```json
  {
    "analysis": "<string>",
    "response_preparation_analysis": "<string>",
    "related_to_query": true/false,
    "reason_of_exclusion": "<string>",
    "summary_of_the_paper": "<string>",
    "summary_phrase": "<string>",
    "thought": "<string>"
  }
```

Notes:
- analysis: includes all relevant details from the paper pertaining to the review, plus any interesting extra information.
- response_preparation_analysis: how you intend to fulfill the user's needs, given any prior instructions.
- related_to_query: a boolean indicating if the paper addresses "query".
- reason_of_exclusion: if false above, explain why.
- summary_of_the_paper: a concise overview of the content; return "not included" if no relevant details are found.
- summary_phrase: a short, up-to-three-word phrase describing how the paper connects to "query".
- thought: your overall updated understanding of the review, without extraneous details.

Synthesize Prompt

You are a research agent focusing on a biomedical systematic review for the question {query}. You have just received a new paper summary and must integrate it with existing paper summaries or intermediate syntheses, if relevant. Your output should merge any overlapping or complementary information into a final synthesized summary as described below. Cite sources using <citation>citation_number</citation> at the end of relevant sentences.
Here is the newly provided paper summary:
{current_summary_index}:{paper_summary}
Below are previously generated summaries or intermediate syntheses, each with an ID:
{previous_summaries}
You should:
[1] Identify the most relevant existing summary or synthesis to merge with this new paper. If none are relevant, clearly state why.
[2] If a relevant item is another paper summary, combine them into a new intermediate synthesis. If it is an existing synthesis, add the new paper's insights to that synthesis. If no match is found, explicitly explain that outcome.
[3] Incorporate the user's summarization requirement {summarization_requirement} into the final integrated summary if it applies to the content.
[4] Structure the final synthesized summary (plain text inside each section) with the following sections, and cite sources using <citation>citation_number</citation>:
==========
• Introduction: Introduce the research question {query} and its broader context, explaining its significance.
• Study Design: Describe how the included summaries were chosen. Mention that the following inclusion-exclusion criteria were applied: {inclusion_exclusion_criteria}. Do not mention any search strategy.
• Key Findings: Present core insights from all integrated evidence, highlighting any similarities, differences, patterns, or contradictions. Cite each source via <citation>id</citation> or <citation>{current_summary_index}</citation> for the new paper.
• Conclusion: Summarize the overall outcomes, discussing whether they address the original question and any overarching patterns or implications.
• Discussion: Reflect on the strength of the evidence, potential gaps, limitations, and suggest directions for future work. If certain aspects remain unclear, note that consulting the full text of some articles may be necessary.
==========
[5] Only use information provided in these summaries or syntheses; do not introduce details beyond the given text. Retain any citations from the identified synthesis if you are updating it.
After deciding which summaries or syntheses to include, return a single JSON object following the schema below, and nothing else:

```
{
  "identified_relevant_summaries": ["<id1>", "<id2>", ...],
  "reasoning": "<string>",
  "synthesized_summary": "<string>",
  "thought": "<string>"
}
```

Notes:
- "identified_relevant_summaries" is a list of IDs for any relevant summaries or empty if none match.
- "reasoning" explains why these items were selected (or not) in the context of {query}.
- "synthesized_summary" is the new or updated synthesis, containing the five sections (Introduction, Study Design, Key Findings, Conclusion, Discussion) formatted in HTML with citations.
- "thought" is your overall perspective on {query} thus far, mindful of potential conflicts but without arbitrating them.

Reflect Prompt

You are a research agent assisting a human expert in conducting a systematic review for the question: {query}. Below are the criteria you have been using to include or exclude studies:
{include_exclude_criteria}
You have already read several papers and obtained certain findings (summaries or insights). The human expert has now provided further input or questions. They may also have changed their guidance on which paper to read ({paper_reading_instruction_if_any}), or revised the summarization requirement. If no such specific instruction exists, this variable will be empty.
{findings_so_far}
(Note: these findings represent your current overall insights.)
{conversation_history}
You must reflect on this new feedback or instruction to determine how your process should evolve. For instance:
- If the human suggests focusing on a different paper than you previously chose, you should note how your plan will adapt in the next iteration.
- If they introduce or change the summarization requirement, you should note the updates you will make.
- If they provide further critiques or clarifications, incorporate them into your plan.
- If they only express general approval or clarifications about your existing findings, you may continue without major changes.
At the end of your reflection, you must produce a single JSON object, formatted following this schema:

```json
{
  "reflection": "<string>",
  "updates_on_additional_requirement": "<string>",
  "updates_on_criteria": "<string>",
  "updates_on_summarization_requirement": "<string>"
}
```

Note:
-"reflection" is your overall reasoning on how to modify your process or maintain it, based on the new human feedback.
-"updates_on_additional_requirement" describes any further research directions or sub-questions you plan to pursue at the user's request; if there is no update, leave this empty.
-"updates_on_criteria" indicates any changes to your inclusion/exclusion criteria after reflecting on human critiques; if none, leave this empty.
-"updates_on_summarization_requirement" details any shifts in how you will summarize information in subsequent steps (e.g., more concise, focusing on specific methodology or outcome), or remains empty if no change is needed.

Year	Title	Corpus/Inclusion
2024	A systematic review on the impact of sub-epidermal (Avsar et al., 2024)	712 Total / 10 Included
2024	Artificial Intelligence Support for Informal Patient Caregivers (Borna et al., 2024)	381 Total / 10 Included
2024	Nutritional interventions for preventing (Langer et al., 2024)	72 Total / 23 Included
2024	Community dwelling older people (Iida et al., 2025)	455 Total / 5 Included
2024	Prevalence of persistent hypertension following pregnancy (Mukosha et al., 2024)	752 Total / 22 Included
2024	Insulin Delivery Technology for Treatment in Infants (Panza et al., 2024)	204 Total / 22 Included
2024	A Systematic Review of Nurses' Perceptions (Park et al., 2024)	449 Total / 21 Included
2024	Text Messaging for Medication Adherence in CVD Prevention (Hyun et al., 2024)	4,162 Total / 18 Included
2024	Psychological interventions for asthma in children and adolescents (Sharrad et al., 2024)	6,665 Total / 34 Included
2024	Cognitive mental workload of emergency nursing (Surendran et al., 2024)	353 Total / 57 Included
2024	Remdesivir for Patients with COVID-19 (Tan-Lim and Esteban-Ipac, 2024)	491 Total / 9 Included
2024	The prevalence of pulmonary hypertension in post-tuberculosis (van Heerden et al., 2024)	542 Total / 34 Included
2022	Effects of Caffeine Intake on endurance (Wang et al., 2022)	1,107 Total / 21 Included
2024	Evening versus morning dosing regimen (Wu et al., 2024)	7,356 Total / 27 Included
2023	Efficacy of single-pill combination (Xie et al., 2023)	1,697 Total / 32 Included

Table 6: 15 completed systematical reviews used in evaluation, including publication year, review title, total retrieved corpus, and final number of included articles.

Detailed Evaluation Rubric – Category 1: Introduction & Research Context (15 points)

These items evaluate how well the extended abstract states the research question, provides justification, and outlines clear objectives for practice or research.

- 1.1 Clarity of Research Question (5 points)
  - *Full Points (5):*
    - The extended abstract explicitly states the research question (e.g., comparing interventions X and Y).
    - The scope is clear (population, condition, outcomes).
  - *Deductions:*
    - (-1 ~2 points) if the question is vaguely stated or missing key details.
    - (-3 ~4 points) if the research question is unclear or scattered.
    - (-5 points) if the question is not stated at all or is entirely irrelevant.
- 1.2 Justification & Relevance (5 points)
  - *Full Points (5):*
    - Provides a compelling reason for why this extended abstract is necessary (e.g., gaps in existing knowledge).
    - Shows clear alignment with user-provided context or corpus.
  - *Deductions:*
    - (-1 ~2 points) if the justification is too brief or lacks evidence.
    - (-3 ~4 points) if little rationale is given and significance is unclear.
    - (-5 points) if no justification is provided at all.
- 1.3 Objectives & Significance (5 points)
  - *Full Points (5):*
    - Clearly identifies intended impact (practice, policy, or future research).
    - States how the agent-based summary could help clinicians, researchers, or stakeholders.
  - *Deductions:*
    - (-1 ~2 points) if objectives are too general or only partially linked.
    - (-3 ~4 points) if objectives are not mentioned or lack practical implications.
    - (-5 points) if it is entirely missing or irrelevant.

Table 7: Detailed Evaluation Rubric – Category 1

Detailed Evaluation Rubric – Category 2: Summarization Approach & Method (25 points)

These items focus on how the agent or system generated the extended abstract, inclusion/exclusion criteria, data extraction, and potential biases. Also adapted from traditional methodology to reflect the user-provided corpus.

- 2.1 Agent-Based Summarization Description (5 points)
  - *Full Points (5):*
    - Describes how the system generated the abstract (e.g., from user corpus, rules, filters).
    - Mentions the basic workflow (e.g., "The agent filtered 50 articles…").
  - *Deductions:*
    - (-1 ~2 points) if the system's role is only vaguely mentioned.
    - (-3 ~4 points) if the method is unclear or incomplete.
    - (-5 points) if no method is described at all.
- 2.2 Inclusion/Exclusion Rationale & Data Sources (5 points)
  - *Full Points (5):*
    - Briefly states what types of articles or data were included (e.g., RCTs, observational) and any user-driven filters.
    - Mentions the source of the corpus (e.g., "User uploaded 50 full-text articles from PubMed…").
  - *Deductions:*
    - (-1 ~2 points) if criteria or corpus sources are vague or incomplete.
    - (-3 ~4 points) if it's unclear what data was included/excluded.
    - (-5 points) if no mention of data sources or selection criteria is made.
- 2.3 Coverage & Representativeness (5 points)
  - *Full Points (5):*
    - Demonstrates that the abstract captures the key range of the user-provided corpus.
    - Incorporates relevant findings from diverse studies (if applicable).
  - *Deductions:*
    - (-1 ~2 points) if only a narrow subset is summarized without justification.
    - (-3 ~4 points) if major, important studies/themes are missing.
    - (-5 points) if it's unclear the summary reflects the corpus at all.
- 2.4 Data Extraction & Reliability (5 points)
  - *Full Points (5):*
    - Explains how key info (outcomes, participants, etc.) was extracted or verified.
    - Mentions how errors/biases in extraction might be handled.
  - *Deductions:*
    - (-1 ~2 points) if extraction process is vaguely described.
    - (-3 ~4 points) if there is no clarity on relevant details.
    - (-5 points) if steps are absent or contradictory.
- 2.5 Risk of Bias / Limitations (5 points)
  - *Full Points (5):*
    - Acknowledges potential weaknesses (e.g., model may overlook nuances).
    - Identifies any known biases in the agent's approach or corpus.
  - *Deductions:*
    - (-1 ~2 points) if risk of bias is only superficially noted.
    - (-3 ~4 points) if major limitations are ignored.
    - (-5 points) if no mention of limitations or bias at all.

Table 8: Detailed Evaluation Rubric – Category 2

Detailed Evaluation Rubric – Category 3: Comprehensiveness & Accuracy of Summaries (30 points)

These items focus on how thoroughly and accurately the extended abstract covers all major findings, uses quantitative evidence, and addresses heterogeneity or conflicting data.

- 3.1 Relevance & Completeness (10 points)
  - *Full Points (10):*
    - Accurately reflects all major findings relevant to the research question.
    - Notes key outcomes, interventions, or theories from the corpus.
  - *Deductions:*
    - (-1 ~3 points) if some significant findings are missing or incomplete.
    - (-4 ~7 points) if large sections are omitted, misleading the summary.
    - (-8 ~10 points) if it fails to capture the essence entirely.
- 3.2 Clarity & Cohesion of Findings (5 points)
  - *Full Points (5):*
    - Information flows logically, with clear transitions; no abrupt contradictions.
  - *Deductions:*
    - (-1 ~2 points) if transitions are unclear or haphazard.
    - (-3 ~5 points) if the text is difficult to follow or contradictory.
- 3.3 Use of Quantitative or Specific Evidence (5 points)
  - *Full Points (5):*
    - Includes specific data points or effect sizes where possible.
    - Avoids vague statements like "Most studies are positive" without data.
  - *Deductions:*
    - (-1 ~2 points) if only partial specificity or missing effect sizes.
    - (-3 ~4 points) if mostly qualitative, omitting key quantitative data.
    - (-5 points) if no concrete data despite availability.
- 3.4 Discussion of Variability / Heterogeneity (5 points)
  - *Full Points (5):*
    - Addresses differences among studies and acknowledges inconsistent findings.
  - *Deductions:*
    - (-1 ~2 points) if variability is briefly mentioned but not explained.
    - (-3 ~4 points) if contradictory findings lack context.
    - (-5 points) if no mention of heterogeneity at all.
- 3.5 Faithfulness to Source Material (5 points)
  - *Full Points (5):*
    - Conclusions and claims clearly stem from the user-provided corpus.
    - No major distortions or misinterpretations of original study conclusions.
  - *Deductions:*
    - (-1 ~2 points) if minor discrepancies or paraphrasing errors occur.
    - (-3 ~4 points) if some interpretations conflict with source data.
    - (-5 points) if the summary significantly misrepresents the corpus or invents data.

Table 9: Detailed Evaluation Rubric – Category 3

Detailed Evaluation Rubric – Category 4: Discussion, Implications & Future Directions (20 points)

These items evaluate how well the summary interprets findings, gives practical recommendations, and proposes future research.

- 4.1 Interpretation of Findings (10 points)
  - *Full Points (10):*
    - Interprets the summarized data in context (e.g., "These findings suggest X is more effective…").
    - Identifies overarching trends, limitations, and potential biases clearly.
  - *Deductions:*
    - (-1 ~3 points) if interpretation is present but lacks clarity or depth.
    - (-4 ~7 points) if major outcomes are not linked to broader discussion.
    - (-8 ~10 points) if there is little or no interpretation, or contradictions.
- 4.2 Practical Implications or Recommendations (5 points)
  - *Full Points (5):*
    - Offers thoughtful suggestions for clinicians, policymakers, or researchers.
    - Ties the summarized evidence back to real-world application.
  - *Deductions:*
    - (-1 ~2 points) if implications are only briefly stated.
    - (-3 ~4 points) if recommendations are vague or lack linkage to data.
    - (-5 points) if no discussion of implications is provided.
- 4.3 Suggestions for Future Research (5 points)
  - *Full Points (5):*
    - Identifies gaps or unresolved questions (e.g., limited data on long-term effects of X).
    - Proposes clear avenues for deeper investigation.
  - *Deductions:*
    - (-1 ~2 points) if future directions are superficial.
    - (-3 ~4 points) if they are barely mentioned or unconnected to corpus.
    - (-5 points) if no future research direction is provided at all.

Table 10: Detailed Evaluation Rubric – Category 4

Detailed Evaluation Rubric – Category 5: Writing Quality, Fluency & Referencing (10 points)

These items assess clarity, coherence, and proper referencing or attribution in the extended abstract.

- 5.1 Readability & Structure (5 points)
  - *Full Points (5):*
    - The text is well-organized, concise, and easy to understand.
    - Logical headings or paragraph breaks guide the reader.
  - *Deductions:*
    - (-1 ~2 points) if minor clarity issues occur sporadically.
    - (-3 ~4 points) if the writing is frequently unclear or structurally confusing.
    - (-5 points) if the writing is so poor that the extended abstract is barely comprehensible.
- 5.2 Referencing & Supporting Evidence (5 points)
  - *Full Points (5):*
    - References (or placeholders) are used where the agent draws on specific studies or data.
    - Citations are relevant; any direct quotations are properly attributed.
  - *Deductions:*
    - (-1 ~2 points) if references are incomplete or inconsistent.
    - (-3 ~4 points) if missing references for major claims.
    - (-5 points) if there are no references at all in contexts that need them.

Table 11: Detailed Evaluation Rubric – Category 5

Detailed Questionnaires after User Study

Category 1: The System Was Easy to Use

These questions evaluate how easily participants can navigate and operate different components of *Insight*Agent.

1. Radial Chart – Search & Navigation
   - How easy was it to locate the target article using the search function?
     Scale: (1 = Very difficult, 5 = Very easy)
2. Hierarchical Map – Exploration
   - How easy was it to explore and understand the synthesis process?
     Scale: (1 = Very difficult, 5 = Very easy)
3. Hierarchical Map – Tracing Citations
   - How easy was it to trace the citation path for the summary?
     Scale: (1 = Very difficult, 5 = Very easy)
4. Hierarchical Map – Locating Specific Evidence
   - How easy was it to locate specific evidence within the Hierarchical Map?
     Scale: (1 = Very difficult, 5 = Very easy)

Category 2: Confidence in Recommendations

These questions measure the participant's trust in the system's outputs, their own confidence in interpreting the agent's actions, and overall trust in the final systematic review.

1. Radial Chart – Agent's Progress
   - How confident are you in your understanding of the agent's progress?
     Scale: (1 = Not confident at all, 5 = Very confident)
2. Hierarchical Map – Source Identification
   - How confident are you that you identified all the original sources correctly?
     Scale: (1 = Not confident at all, 5 = Very confident)
3. Hierarchical Map – Agent Incorporating Feedback
   - How confident are you that the agent will incorporate your feedback into future summaries?
     Scale: (1 = Not confident at all, 5 = Very confident)
4. Hierarchical Map – Synthesis Accuracy
   - How confident are you that the agent is synthesizing the correct information?
     Scale: Scale: (1 = Not confident at all, 5 = Very confident)

Category 3: Visualization-Aided Understanding

These items assess how well the visual interfaces (Radial Chart and Hierarchical Map) help users understand cluster boundaries, topic coverage, citation structures, and the synthesis process.

1. Radial Chart – Cluster Clarity
   - How clear were the cluster boundaries on the Radial Chart?
     Scale: (1 = Not clear at all, 5 = Very clear)
2. Radial Chart – Topic Identification
   - How easy was it to understand the main topics of each cluster?
     Scale: (1 = Very difficult, 5 = Very easy)
3. Radial Chart – Node Information
   - How clear was the information provided when hovering over the article node?
     Scale: (1 = Not clear at all, 5 = Very clear)
4. Radial Chart – Topic Coverage
   - How comprehensive was the Radial Chart in displaying topic coverage?
     Scale: (1 = Not comprehensive at all, 5 = Very comprehensive)
5. Radial Chart – Missing Areas
   - How easy was it to identify missing areas in the agent's coverage?
     Scale: (1 = Very difficult, 5 = Very easy)
6. Hierarchical Map – Synthesis Structure
   - How clear was the hierarchical structure of the evidence synthesis?
     Scale: (1 = Not clear at all, 5 = Very clear)
7. Hierarchical Map – Agent's Synthesis Process
   - How clear was the agent's synthesis process as displayed in the Hierarchical Map?
     Scale: (1 = Not clear at all, 5 = Very clear)

Table 12: Detailed Questionnaire

Detailed Questionnaires after User Study

Category 4: Ability to Guide or Correct Agents

These questions measure how effectively participants can detect and fix issues, give feedback, or redirect the agent using the system's interfaces.

1. Radial Chart – Path Correction
   - How easy was it to identify incorrect paths on the Radial Chart?
     Scale: (1 = Very difficult, 5 = Very easy)
2. Radial Chart – Drag-and-Drop
   - How intuitive was the drag-and-drop functionality for adjusting paths?
     Scale: (1 = Not intuitive at all, 5 = Very intuitive)
3. Hierarchical Map – Missing Citations
   - How easy was it to identify missing citations or incorrect paths?
     Scale: (1 = Very difficult, 5 = Very easy)
4. Hierarchical Map – Feedback on Summaries
   - How intuitive was the feedback process within the Hierarchical Map?
     Scale: (1 = Not intuitive at all, 5 = Very intuitive)

Category 5: Overall Satisfaction

These questions capture general satisfaction and user impressions of how well the system performs and meets user needs.

1. Overall System Trustworthiness
   - To what extent do you trust the generated systematic review?
     Scale: (1 = Not confident at all, 5 = Very confident)

Table 13: Detailed Questionnaire CONT.

Table 14: Report generated by **Human (original report)** on Systematic review for Remdesivir be Used for Treatment (Tan-Lim and Esteban-Ipac, 2024) [preferred statements are highlighted in green, less preferred statements are highlighted in orange, not preferred statements are highlighted in red]

Human

Background. Remdesivir is an intravenously administered antiviral drug that inhibits RNA-dependent RNA polymerase. In vitro studies have shown that remdesivir can inhibit the growth of the COVID-19 virus in infected Vero cells and can inhibit infection in human cell lines.
Objective. To determine the efficacy and safety of remdesivir in treating patients with COVID-19 infection.
Methods. A systematic search of electronic medical literature databases was done from inception until September 4, 2022. Search for ongoing studies and preprints was also done. Risk of bias assessment was done using Cochrane risk of bias tool version 2.0. Measures of effect used were relative risk (RR) and 95% confidence interval (CI). Subgroup analysis by disease severity was preplanned. The estimates for efficacy and safety of remdesivir was calculated using Review Manager 5.4 software.
Results. Nine randomized controlled trials with 13,085 participants were identified(**Report a detailed number of participants / patients in these studies**). Eight of the included studies recruited confirmed COVID-19 patients needing hospitalization, while one study limited recruitment to nonhospitalized patients. Remdesivir showed significant benefit for outpatients with mild to moderate disease with at least one risk factor for disease progression in terms of COVID 19-related hospitalization (RR 0.13 95% CI 0.03 to 0.59), all-cause hospitalization (RR 0.28, 95% CI 0.10 to 0.75), and need for medically-attended visits (RR 0.19, 95% CI 0.07 to 0.56).(**detailed synthesized statistics**)

Discussion. Remdesivir demonstrates benefit in patients with mild, moderate, and severe COVID-19 but not in critical cases requiring mechanical ventilation. Its safety profile aligns with international recommendations.

Conclusion. Current evidence supports the use of remdesivir in selected COVID-19 patients.

Table 15: Report generated by ***Insight*Agent** on Systematic review for Remdesivir be Used for Treatment (Tan-Lim and Esteban-Ipac, 2024) [preferred statements are highlighted in green, less preferred statements are highlighted in orange, not preferred statements are highlighted in red]

***Insight*Agent(GPT-4o)**
Introduction: Remdesivir has been widely used as an antiviral treatment for COVID-19, but its effectiveness remains debated. This review evaluates the efficacy and safety of remdesivir in treating COVID-19 patients, focusing on outcomes such as mortality, clinical improvement, mechanical ventilation, and adverse events. The review synthesizes findings from multiple randomized controlled trials (RCTs) to provide a comprehensive assessment of its clinical impact. **Study Design:** ^{E_2}This systematic review analyzed 491 studies, of which 11 met the inclusion criteria for final synthesis. These studies were RCTs comparing remdesivir with placebo or standard care in patients with COVID-19 of varying severity (**More detailed number of participants in these studies would make it more convincing**). The inclusion criteria required studies to assess mortality, clinical improvement, hospitalization, time to recovery, or adverse events. Observational studies, quasi-randomized trials, and those without relevant outcomes were excluded. The included trials span different populations and care settings, allowing for a broad assessment of remdesivir's effectiveness [1-11]. **Key Findings:** **Mortality Impact: (preferred discussion topic for findings)** The WHO Solidarity trial found no significant reduction in mortality among hospitalized COVID-19 patients treated with ^{E_1} remdesivir (14.5% vs. 15.6%, RR 0.91, p=0.12) [2]. A meta-analysis combining multiple trials showed a modest mortality reduction in non-ventilated patients (RR 0.86, p=0.02) (**More synthesized stats result is preferred**) but no benefit for ventilated patients [2]. The NOR-Solidarity trial reported similar mortality rates between remdesivir and standard care groups [4]. **Clinical Improvement & Recovery Time: (preferred discussion topic for findings)** The DisCoVeRy trial found no substantial improvement in clinical outcomes with remdesivir treatment [1]. The PINETREE study reported that early remdesivir use in high-risk outpatients significantly reduced hospitalization or death (HR 0.13, p=0.008) [3]. The Spinner et al. trial observed that a 5-day remdesivir course resulted in a statistically significant improvement in clinical status, whereas the 10-day regimen showed no clear benefit [6]. **Need for Mechanical Ventilation:** The WHO Solidarity trial found that remdesivir reduced the progression to mechanical ventilation in non-ventilated patients (RR 0.88, p=0.04) [2]. In ventilated patients, remdesivir did not improve survival or clinical outcomes (RR 1.13, p=0.32) [2]. **Adverse Events & Safety Profile: (preferred discussion topic for findings)** The NOR-Solidarity trial found no significant difference in adverse events between remdesivir and control groups [4]. Common adverse effects associated with remdesivir included nausea, hypokalemia, and headache, though these were generally mild [6]. Remdesivir was linked to a slight delay in hospital discharge due to its intravenous administration requirement [2]. **Conclusion:** Remdesivir provides modest clinical benefits in non-ventilated COVID-19 patients and may prevent disease progression when administered early in high-risk outpatients.(**aligns with human conclusion**) However, its impact on overall mortality remains limited, and it offers no advantage for patients already on mechanical ventilation. Given the mixed results, remdesivir's role in COVID-19 treatment should be considered alongside other therapeutic options. **Discussion:** The evidence suggests that remdesivir's effectiveness depends on factors such as disease severity and timing of administration. The benefits appear more significant in early-stage disease, while hospitalized patients see limited improvement.(**Preferred, more nuanced conclusion**) Study heterogeneity, differences in patient populations, and variations in standard care practices contribute to the mixed findings. Future research should explore combination therapies, optimize treatment timing, and evaluate long-term effects. Additionally, given the practical challenges of intravenous administration, alternative formulations or delivery methods may improve its accessibility. (**Preferred, insightful suggestions based on the synthesized findings**)

Table 16: Report generated by **AutoSurvey** on Systematic review for Remdesivir be Used for Treatment (Tan-Lim and Esteban-Ipac, 2024) [preferred statements are highlighted in green, less preferred statements are highlighted in orange, not preferred statements are highlighted in red]

AutoSurvey

Abstract
Remdesivir has been widely used as an antiviral treatment for COVID-19. This review evaluates its efficacy and safety by examining mortality rates, clinical improvement, mechanical ventilation needs, and adverse events. The findings indicate that remdesivir reduces hospitalization duration but has limited impact on mortality. It is generally well-tolerated, though concerns about nephrotoxicity exist. The drug is more effective in early-stage treatment and in combination with other therapies. However, racial and socioeconomic disparities affect its accessibility. Future research should explore optimized administration strategies and combination treatments.

1. Introduction
The COVID-19 pandemic has driven an urgent search for effective treatment options to mitigate the severe health outcomes associated with SARS-CoV-2 infection. Among the antiviral treatments explored, remdesivir emerged as a promising candidate due to its mechanism of action as an RNA-dependent RNA polymerase (RdRp) inhibitor. It was the first antiviral approved by the U.S. Food and Drug Administration (FDA) for COVID-19 treatment. Despite its initial promise, clinical trials and real-world studies have provided mixed results regarding its efficacy. While some studies suggest that remdesivir can shorten hospitalization duration, others indicate that it does not significantly reduce mortality rates. This review aims to comprehensively assess the impact of remdesivir on mortality, disease progression, the need for mechanical ventilation, and its safety profile.

2. Mechanism of Action of Remdesivir (Irrelevant section)
Remdesivir functions by inhibiting the viral RNA polymerase enzyme, which is essential for SARS-CoV-2 replication **...**

3. Clinical Efficacy of Remdesivir
Several studies have evaluated remdesivir's role in reducing COVID-19 mortality. The WHO Solidarity Trial, one of the largest global studies, found no significant mortality benefit associated with remdesivir use. However, other randomized controlled trials have suggested that remdesivir may offer minor improvements in clinical outcomes for patients with moderate disease, particularly when administered early in the course of infection. Despite these mixed findings, remdesivir has been shown to shorten hospitalization duration in many patient populations. Studies have reported faster viral clearance in younger patients and those with better baseline health conditions, though the effect on overall recovery remains uncertain. Some trials also indicate that remdesivir may reduce the need for invasive mechanical ventilation, but these results are inconsistent across different clinical settings.(**lack quantitative evidence and details**)

4. Safety and Adverse Events
The safety profile of remdesivir has been a topic of ongoing investigation, with nephrotoxicity emerging as a significant concern. Causal discovery models analyzing electronic health records suggest that remdesivir may contribute to acute kidney injury (AKI), though its risk appears lower in patients with pre-existing chronic kidney disease. Additionally, concerns about liver toxicity have been raised, but long-term studies indicate that remdesivir does not significantly affect mitochondrial DNA integrity. Clinical trials have reported transient elevations in liver enzyme levels, which require monitoring in high-risk patients. Other commonly reported adverse events include nausea, hypokalemia, and headaches. Despite these risks, remdesivir remains a relatively well-tolerated antiviral when compared to alternative treatments such as molnupiravir and nirmatrelvir, which may exhibit different side effect profiles.(**lack quantitative evidence and details**)

5. Comparative Effectiveness and Combination Therapies
When compared to other antiviral options, remdesivir has demonstrated mixed efficacy. While molnupiravir and Paxlovid offer alternative mechanisms for viral inhibition, their clinical effectiveness varies based on patient demographics and disease severity. Some studies suggest that oral GS-441524 derivatives could surpass remdesivir in terms of convenience and effectiveness due to improved bioavailability. (**non-irrelevant evidence and discussion**) Combination therapy approaches have also been explored to enhance remdesivir's therapeutic potential. Studies have shown that combining remdesivir with corticosteroids or monoclonal antibodies can yield synergistic benefits. Additionally, remdesivir combined with azithromycin or ivermectin has been investigated for improved antiviral activity. In kidney transplant patients, remdesivir has been associated with protective effects against donor-specific antibody formation, further demonstrating its potential utility in specialized populations.

6. Disparities in Treatment and Access (Irrelevant section)
Access to remdesivir treatment has not been uniform across different racial and socioeconomic groups **...**

7. Economic and Cost-effectiveness Analysis (Irrelevant section)
The cost-effectiveness of remdesivir has been a subject of debate among healthcare policymakers. **...**

8. Limitations of Current Evidence
Despite extensive research, there are several limitations in the current evidence base regarding remdesivir's efficacy. Conflicting findings across clinical trials can be attributed to variability in study designs, differences in patient selection criteria, and inconsistencies in outcome measures. Large-scale trials such as the WHO Solidarity study provide valuable insights, but their results may not fully capture remdesivir's benefits in specific patient subgroups. Furthermore, the long-term impact of remdesivir on post-acute COVID-19 conditions remains unclear. Biases in hospital protocols, co-administration of other treatments, and differences in the effectiveness of remdesivir against emerging variants also complicate the interpretation of findings.

9. Conclusion and Future Directions
In summary, remdesivir has demonstrated some benefits in reducing hospitalization duration, but its impact on mortality remains limited. (**more conclusive result should be made**) The drug has a generally acceptable safety profile, though concerns about nephrotoxicity necessitate careful monitoring. Combination therapies show promise in enhancing remdesivir's effectiveness, but more research is needed to identify optimal treatment strategies. Clinically, remdesivir is best suited for moderate COVID-19 cases when administered early in the disease course. (**Result aligned with human report**) Moving forward, future studies should focus on exploring remdesivir's efficacy in combination with other antivirals, improving oral formulations for easier administration, and addressing healthcare disparities to ensure equitable access.

Table 17: Report generated by ChatCite on Systematic review for Remdesivir be Used for Treatment (Tan-Lim and Esteban-Ipac, 2024) [preferred statements are highlighted in green, less preferred statements are highlighted in orange, not preferred statements are highlighted in red]

ChatCite
Remdesivir has demonstrated variable efficacy in treating COVID-19, with its benefits largely dependent on disease severity, timing of administration, and patient characteristics. The WHO Solidarity Trial, one of the largest studies, found that remdesivir did not significantly reduce mortality among hospitalized COVID-19 patients compared to standard care (RR 0.91, 95% CI 0.82-1.02, p=0.12) (**more synthesized stats as evidence would be stronger**), with no observed survival benefit for patients already on mechanical ventilation [1]. However, among hospitalized patients requiring oxygen but not yet ventilated, remdesivir showed a modest reduction in mortality (RR 0.87, 95% CI 0.76-0.99, p=0.03) and in disease progression to ventilation (RR 0.88, 95% CI 0.77-1.00, p=0.04), suggesting potential benefits for less severe cases [1]. The NOR-Solidarity trial similarly found no significant difference in mortality, viral clearance, or reduction in respiratory failure between remdesivir, hydroxychloroquine, and standard care, highlighting its limited antiviral effect in hospitalized patients [3]. Additionally, a study on moderate-to-severe COVID-19 cases found no significant clinical improvement or difference in recovery time between patients receiving remdesivir and those who did not, further questioning its efficacy in already hospitalized patients [4]. Conversely, evidence from outpatient settings suggests that remdesivir may have a role in preventing disease progression if administered early (**Needs specific evidence or more precise wording**). A randomized, placebo-controlled trial in high-risk, nonhospitalized patients found that a 3-day course of remdesivir reduced hospitalization or death risk by 87% (HR 0.13, 95% CI 0.03-0.59, p=0.008) compared to placebo, with a lower incidence of COVID-19-related medically attended visits [2]. These findings indicate that early intervention with remdesivir in at-risk populations may be effective in preventing severe disease progression. Among patients with moderate COVID-19, a 5-day course of remdesivir led to a statistically significant improvement in clinical status at day 11 compared to standard care (OR 1.65, 95% CI 1.09-2.48, p=0.02), though the clinical relevance of this improvement remains uncertain [5]. In contrast, a 10-day regimen showed no significant difference in clinical outcomes, raising questions about the optimal duration of treatment [5]. Another study found that while remdesivir-treated patients showed a numerically faster recovery trend, this was not statistically significant, and its effect was more pronounced in patients treated earlier in the disease course [6]. Regarding safety, remdesivir was generally well tolerated, with most trials reporting similar or slightly higher rates of adverse events compared to standard care or placebo. In a trial focusing on severe COVID-19, adverse events occurred in 66% of remdesivir-treated patients compared to 64% in the placebo group, but a higher proportion of remdesivir recipients discontinued treatment due to adverse events (12% vs. 5%) [6]. Another study reported comparable rates of adverse events between remdesivir and placebo groups (42.3% vs. 46.3%), indicating an acceptable safety profile [2]. While nausea, hypokalemia, and headaches were slightly more frequent among remdesivir recipients, these side effects were generally manageable [5]. However, the requirement for intravenous administration may limit its accessibility in outpatient settings. Collectively, these findings suggest that remdesivir provides minimal benefit for critically ill hospitalized patients but may help prevent hospitalization and severe disease when used early in high-risk outpatients. Further research is needed to determine the optimal treatment duration and patient selection criteria for maximizing its efficacy. **No Explicit Conclusion**

Evaluation Details of the Review Report Done by *Insight*Agent (Table 15)

1. **Introduction & Research Context (15 Points)**
   (a) **Clarity of Research Question (5/5)**
      - The extended abstract explicitly states the research question: evaluating the efficacy and safety of remdesivir in treating COVID-19 patients.
      - The scope is clear, defining key outcomes: mortality, clinical improvement, mechanical ventilation, and adverse events.
   (b) **Justification & Relevance (4/5)**
      - The abstract mentions ongoing debates about remdesivir's effectiveness, justifying the review.
      - However, it does not explicitly highlight gaps in prior systematic reviews that this study aims to address.
      - A stronger justification would include a brief mention of inconsistencies in guidelines or conflicting clinical trial results.
   (c) **Objectives & Significance for Practice/Research (5/5)**
      - The objectives are well-articulated: providing a comprehensive synthesis of randomized controlled trials (RCTs).
      - The conclusion explicitly discusses clinical decision-making, reinforcing its significance for practice and research.

2. **Summarization Approach & Method (25 Points)**
   (a) **Agent-Based/Automated Summarization Description (5/5)**
      - The abstract effectively demonstrates *Insight*Agent's capabilities in selecting and synthesizing relevant studies. It also highlights human involvement in modifying agent actions, ensuring higher accuracy and contextual relevance.
   (b) **Inclusion/Exclusion Rationale & Data Sources (4/5)**
      - The review clearly states that it analyzed 491 studies and selected 11 as relevant, which is a strong methodological detail. However, it does not explicitly clarify why certain studies were excluded beyond generic criteria (RCTs, relevant outcomes). A small addition about why 3 human-selected studies were excluded would enhance transparency.
   (c) **Coverage & Representativeness (5/5)**
      - The summary reflects a broad range of study populations and care settings, showing high representativeness.
      - Compared to human-generated reviews, users noted more details and a better overview of multiple perspectives.
   (d) **Data Extraction & Reliability (4/5)**
      - The process of data extraction (e.g., identifying mortality rates, hazard ratios) is evident, but there is no explicit mention of how errors were mitigated.
      - An improvement would be a brief statement about verification mechanisms InsightAgent used to prevent misinterpretation of numerical results.
   (e) **Risk of Bias / Limitations of Automated Summarization (3/5)**
      - The abstract does not discuss potential biases in how InsightAgent selects or synthesizes studies.
      - Given that automated agents might miss nuanced contextual details or overweigh certain study designs, this should have been acknowledged.

Table 18: Detailed Evaluation Result.

Evaluation Details of the Review Report Done by *Insight*Agent (Table 15)

3. **Comprehensiveness & Accuracy of Summaries (30 Points)**
   (a) **Relevance & Completeness (10/10)**
   - All major findings are well-captured: mortality impact, clinical improvement, need for mechanical ventilation, and adverse events.
   - Users reported that InsightAgent-generated summaries contained more details than the original human reports while maintaining essential findings.

   (b) **Clarity & Cohesion of Findings (5/5)**
   - The flow is logical, moving from mortality impact to adverse events, which makes the findings easy to follow. The use of structured headings improves readability.

   (c) **Use of Quantitative or Specific Evidence (5/5)**
   - Effect sizes, risk ratios, and p-values are included, making the summary data-driven and robust.
   - Compared to human-written summaries, this version includes more numerical detail.

   (d) **Discussion of Variability / Heterogeneity (3/5)**
   - The abstract mentions study heterogeneity but lacks specific examples of variability (e.g., differences in patient age, severity, treatment duration).
   - A stronger discussion would include how the studies differed in methodology and how that impacts interpretation.

   (e) **Faithfulness to Source Material (5/5)**
   - No misinterpretations or misleading paraphrasing were detected.
   - InsightAgent appears to have accurately preserved key conclusions from each study.

4. **Discussion, Implications & Future Directions (20 Points)**
   (a) **Interpretation of Findings (9/10)**
   - InsightAgent provides a good interpretation of when remdesivir works best (non-ventilated patients, early administration).
   - One missing point: it could have elaborated more on why remdesivir fails in ventilated patients.

   (b) **Practical Implications or Recommendations (5/5)**
   - The conclusion ties findings back to clinical application, helping doctors and policymakers decide how to use remdesivir.

   (c) **Suggestions for Future Research (4/5)**
   - The abstract mentions the need for combination therapies and treatment timing research, which is excellent.
   - However, it does not discuss long-term effects or cost-effectiveness, which could be valuable additions.

5. **Writing Quality, Fluency & Referencing (10 Points)**
   (a) **Readability & Structure (5/5)**
   - Clear and structured writing enhances readability.

   (b) **Referencing & Supporting Evidence (5/5)**
   - References are well-integrated, ensuring credibility.

Final Score: 90/100

User Feedback & Discussion After the Study:

- "Compared to our manually written systematic review, the *Insight*Agent-generated version had more detailed effect sizes and numerical results. It also synthesized the studies in a more structured way."
- "One thing that surprised us was how *Insight*Agent made it easy to explore multiple perspectives—something that usually takes weeks of manual reading and synthesizing"
- "We noticed that while original human review focused on a narrower set of 9 studies, These 11 studies picked by system still managed to capture the key essence of the human-written summary."
- "An interesting perspective we hadn't considered emerged, remdesivir's impact varies not just by disease severity but also by treatment delay and patient demographics. This kind of insight can be hard to extract manually."

Table 19: Detailed Evaluation Result CONT.

Table 20: Detailed Evaluation Result.

Evaluation Details of the Review Report Done by AutoSurvey (Table 16)

1. Introduction & Research Context (15 Points)
   (a) Clarity of Research Question (3/5)
      - The report mentions remdesivir's efficacy and safety, but the research question is not explicitly stated.
      - The discussion covers too many broad topics (e.g., healthcare disparities, economic costs) that dilute the focus.
      - **Deduction (-2 points)**: The research question is vague, and the reader must infer its scope.
   (b) Justification & Relevance (3/5)
      - The report provides background on remdesivir's approval and initial studies, but it does not justify why this review is necessary.
      - Missing discussion of research gaps—instead, it assumes all aspects of remdesivir's effects need evaluation.
      - **Deduction (-2 points)**: Justification lacks depth, failing to define why this review matters in the context of prior literature.
   (c) Objectives & Significance for Practice/Research (2/5)
      - The report does not clearly define its objectives beyond general coverage of remdesivir.
      - It touches on clinical applications but does not connect findings to real-world decision-making for clinicians or policymakers.
      - **Deduction (-3 points)**: The significance is weak, with no clear intent to influence practice or guide research.
2. Summarization Approach & Method (25 Points)
   (a) Agent-Based/Automated Summarization Description (2/5)
      - The methodology is not described, leaving the source of included studies unclear.
      - Unlike InsightAgent, there is no evidence that human users had control over the summarization process.
      - **Deduction (-3 points)**: Without transparency in how studies were selected and summarized, the report's reliability is questionable.
   (b) Inclusion/Exclusion Rationale & Data Sources (1/5)
      - No mention of inclusion or exclusion criteria, making it impossible to verify study quality.
      - The report includes many unrelated sections (e.g., healthcare disparities, cost-effectiveness), suggesting irrelevant studies were used.
      - **Deduction (-4 points)**: Without proper inclusion criteria, findings cannot be trusted as evidence-based.
   (c) Coverage & Representativeness (5/5)
      - The report includes a wide range of topics, covering mechanisms, efficacy, safety, socioeconomic factors, and economic analysis.
      - While comprehensive, it suffers from information overload, lacking focus on answering the research question.
      - No deduction since the breadth is a strength, but coverage does not guarantee quality.
   (d) Data Extraction & Reliability (2/5)
      - Some numerical results are reported, but there is no clarity on how data was extracted.
      - Effect sizes, statistical significance, and study population details are inconsistently mentioned.
      - **Deduction (-3 points)**: The lack of clear data extraction methodology reduces trustworthiness.
   (e) Risk of Bias / Limitations of Automated Summarization (1/5)
      - No discussion of bias or limitations.
      - The report presents all findings as equally valid without distinguishing high- vs. low-quality studies.
      - **Deduction (-4 points)**: Failure to acknowledge limitations severely weakens reliability.

Evaluation Details of the Review Report Done by AutoSurvey (Table 16)

3. Comprehensiveness & Accuracy of Summaries (30 Points)
   (a) Relevance & Completeness (5/10)
      - The report is comprehensive but includes many irrelevant details, such as healthcare equity, economic analysis, and alternative antivirals.
      - **Deduction (-5 points)**: While complete, much of the information does not directly contribute to answering the research question.
   (b) Clarity & Cohesion of Findings (3/5)
      - The flow is inconsistent, jumping between clinical findings, economic considerations, and disparities.
      - Some sections contradict each other, such as saying remdesivir "reduces hospitalization" but also claiming "its impact on overall recovery is uncertain".
      - **Deduction (-2 points)**: Inconsistent messaging reduces clarity.
4. Discussion, Implications & Future Directions (20 Points)
   (a) Interpretation of Findings (6/10)
      - The interpretation lacks depth and fails to highlight key takeaways for different patient populations.
      - While the report mentions remdesivir's role in different treatment settings, it does not explain why outcomes vary.
      - **Deduction (-4 points)**: No clear synthesis of what the findings mean in a clinical context.
   (b) Practical Implications or Recommendations (3/5)
      - The economic and disparity discussion is present, but practical clinical recommendations are missing.
      - **Deduction (-2 points)**: The report does not offer concrete takeaways for treatment decisions.
5. Writing Quality, Fluency & Referencing (10 Points)
   (a) Readability & Structure (3/5)
      - The writing is clear but overly verbose, with some sections appearing redundant.
      - **Deduction (-2 points)**: Information overload reduces readability.
   (b) Referencing & Supporting Evidence (3/5)
      - Some claims are supported by numerical values, but no clear sourcing for study selection.
      - **Deduction (-2 points)**: Lack of transparent referencing makes it hard to verify findings.

Final Score: 56/100

User Feedback & Discussion After the Study:

- "There's no way to verify that the studies included were high quality. The report assumes all findings are equally valid, which is a red flag"
- "The mention of healthcare disparities and economics is interesting but seems misplaced—it doesn't add much to evaluating remdesivir's efficacy"
- "This report felt overwhelming. It covers a lot of ground but lacks focus on what really matters."
- "InsightAgent's report was structured in a way that was easier to digest and more trustworthy, whereas this one felt like a long Wikipedia article without clear filtering of relevant studies."

Table 21: Detailed Evaluation Result of AutoSurvey CONT.

Evaluation Details of the Review Report Done by ChatCite (Table 17)

1. Introduction & Research Context (15 Points)
   (a) Clarity of Research Question (2/5)
      - The report does not explicitly state a research question but rather presents findings in an unstructured format.
      - The focus on remdesivir's effectiveness and safety is evident, but the scope is unclear—it does not specify key outcomes or the target patient population.
      - **Deduction (-3 points)**: The lack of an explicitly stated research objective reduces clarity.
   (b) Justification & Relevance (3/5)
      - The report briefly mentions the WHO Solidarity Trial and other studies but fails to justify why a review of remdesivir is necessary.
      - It does not discuss existing controversies, gaps in knowledge, or why this analysis is valuable in the current research landscape.
      - **Deduction (-2 points)**: Justification is weak, with no discussion of why this review is needed.
   (c) Objectives & Significance for Practice/Research (2/5)
      - There is no explicit statement about the intended impact of the report.
      - The report presents clinical findings but does not connect them to practice or research applications.
      - **Deduction (-3 points)**: The absence of clear objectives reduces its practical value.
2. Summarization Approach & Method (25 Points)
   (a) Agent-Based/Automated Summarization Description (1/5)
      - There is no mention of how studies were selected, summarized, or synthesized.
      - Unlike InsightAgent, which allowed for human refinement and structured summarization, this system appears to have blindly aggregated text.
      - **Deduction (-4 points)**: The lack of any methodological transparency reduces reliability.
   (b) Inclusion/Exclusion Rationale & Data Sources (1/5)
      - The report does not mention inclusion or exclusion criteria, making it impossible to assess the quality of the included studies.
      - Some included results appear cherry-picked without providing a systematic justification.
      - **Deduction (-4 points)**: Without clear inclusion criteria, the review lacks credibility.
   (c) Coverage & Representativeness (4/5)
      - The report covers multiple studies, different patient populations, and treatment settings, showing breadth.
      - However, it presents findings in a disorganized way, making it hard to follow or extract meaningful insights.
      - **Deduction (-1 point)**: The report is broad but lacks focus.
   (d) Data Extraction & Reliability (2/5)
      - Although numerical findings are presented, the report does not clarify how data was extracted or verified.
      - Some statistical values lack context, making it unclear how they were derived.
      - **Deduction (-3 points)**: Data extraction lacks transparency, making verification difficult.
   (e) Risk of Bias / Limitations of Automated Summarization (1/5)
      - No mention of bias or limitations in the data, methodology, or findings.
      - Unlike InsightAgent, which allowed for user refinement and multiple perspectives, this report blindly presents findings without addressing quality concerns.
      - **Deduction (-4 points)**: Failure to address bias makes the findings unreliable.

Table 22: Detailed Evaluation Result.

Evaluation Details of the Review Report Done by ChatCite (Table 17)

3. Summarization Approach & Method (25 Points)
   (a) Agent-Based/Automated Summarization Description (1/5)
      i. There is no mention of how studies were selected, summarized, or synthesized. Unlike InsightAgent, which allowed for human refinement and structured summarization, this system appears to have blindly aggregated text.
      ii. Deduction (-4 points): The lack of any methodological transparency reduces reliability.
   (b) Inclusion/Exclusion Rationale & Data Sources (1/5)
      i. The report does not mention inclusion or exclusion criteria, making it impossible to assess the quality of the included studies. Some included results appear cherry-picked without providing a systematic justification.
      ii. Deduction (-4 points): Without clear inclusion criteria, the review lacks credibility.
   (c) Coverage & Representativeness (4/5)
      i. The report covers multiple studies, different patient populations, and treatment settings, showing breadth. However, it presents findings in a disorganized way, making it hard to follow or extract meaningful insights.
      ii. Deduction (-1 point): The report is broad but lacks focus.
   (d) Data Extraction & Reliability (2/5)
      i. Although numerical findings are presented, the report does not clarify how data was extracted or verified. Some statistical values lack context, making it unclear how they were derived.
      ii. Deduction (-3 points): Data extraction lacks transparency, making verification difficult.
   (e) Risk of Bias / Limitations of Automated Summarization (1/5)
      i. No mention of bias or limitations in the data, methodology, or findings.
      ii. Deduction (-4 points): Failure to address bias makes the findings unreliable.

4. Comprehensiveness & Accuracy of Summaries (30 Points)
   (a) Relevance & Completeness (5/10)
      i. The report includes key study findings but does not structure them properly, making it difficult to determine which results are most relevant. Critical aspects like study variability and key takeaways, are missing or buried within dense paragraphs.
      ii. Deduction (-5 points): The information is present but poorly structured, making it hard to extract key insights.
   (b) Clarity & Cohesion of Findings (2/5)
      i. The lack of structured headings makes the report very difficult to read. Findings jump between mortality, disease progression, and adverse events without clear transitions.
      ii. Deduction (-3 points): Poor organization reduces readability.
   (c) Use of Quantitative or Specific Evidence (3/5)
      i. Some statistical values are included, which improves credibility.
      ii. However, there is no discussion on how these numbers were chosen or whether they are the most critical findings.
      iii. Deduction (-2 points): Lack of transparency on study selection weakens trust in the quantitative evidence.

5. Discussion, Implications & Future Directions (20 Points)
   (a) Interpretation of Findings (4/10)
      i. The report does not provide clear interpretations of the results. Findings are listed, but there is little attempt to synthesize the information into meaningful conclusions.
      ii. Deduction (-6 points): No clear synthesis of what the findings mean for clinical practice.

6. Writing Quality, Fluency & Referencing (10 Points)
   (a) Readability & Structure (1/5)
      i. The entire report is written as one long paragraph, making it difficult to extract key findings. Deduction (-4 points): The poor structure severely reduces readability.

Final Score: 42/100
User Feedback & Discussion After the Study:
- “It’s hard to extract key takeaways because everything is crammed into a single block of text"
- ‘The result is not explicit and makes me hard to trust, as I don’t know how these studies were chosen.”
- “Compared to InsightAgent’s output, this feels more like a collection of findings without clear organization.”

Table 23: Detailed Evaluation Result of ChatCite CONT.