Can Agents Judge Systematic Reviews Like Humans? Evaluating SLRs with LLM-based Multi-Agent System

Anonymous ACL submission

Abstract

Systematic Literature Reviews (SLRs) are foundational to evidence-based research but remain labor-intensive and prone to inconsistency across disciplines. We present an LLM-based SLR evaluation copilot built on a *Multi-Agent System (MAS)* architecture to assist researchers in assessing the overall quality of the systematic literature reviews. The system automates protocol validation, methodological assessment, and topic relevance checks using a scholarly database. Unlike conventional single-agent methods, our design integrates a specialized agentic approach aligned with PRISMA guidelines to support more structured and interpretable evaluations. We conducted an initial study on five published SLRs from diverse domains, comparing system outputs to expert-annotated PRISMA scores, and observed 84% agreement. While early results are promising, this work represents a first step toward scalable and accurate NLP-driven systems for interdisciplinary workflows and reveals their capacity for rigorous, domain-agnostic knowledge aggregation to streamline the review process.

1 Introduction

Systematic Literature Reviews are foundational to evidence-based research, offering a structured, protocol-driven approach for identifying, analyzing, and synthesizing prior work. In contrast to narrative reviews, SLRs follow a predefined methodology to promote transparency, reproducibility, and reduced bias (Grant and Booth, 2009; Booth et al., 2021). They are widely employed to surface research trends, identify knowledge gaps, and establish a grounded basis for future inquiry across disciplines.

However, the exponential growth of scholarly publications, varying in quality and relevance, has made it increasingly challenging to maintain rigor and comprehensiveness in the SLR process. This overload slows down the review pipeline and introduces risks of redundancy and overlooks literature.

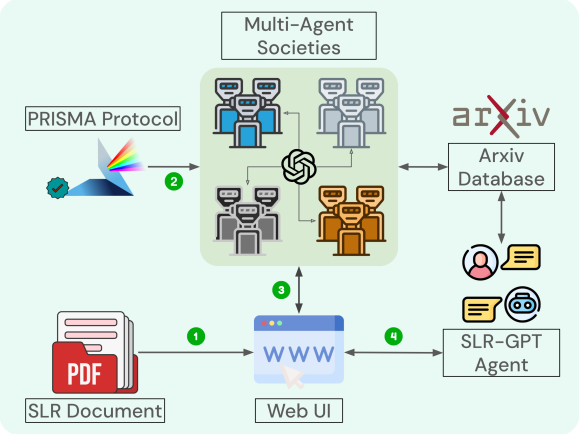

Figure 1: An overview of the proposed multi-agent system LLM-based SLR evaluation framework.

To address these challenges, we propose an AI-augmented SLR system that leverages a multi-agent LLM architecture. Inspired by human-centered co-pilot design principles (Sellen and Horvitz, 2024), our system supports the SLR workflow: from protocol critique and methodological assessment to relevance checking, duplication detection, and collaborative drafting. Figure 1 provides a high-level overview of the proposed system.

The tool is designed with an interdisciplinary lens, recognizing the cross-domain nature of SLRs, from health sciences to software engineering, while leveraging LLM-based agentic architectures to enhance quality and efficiency using the NLP architecture proposed by (Vaswani et al., 2017). By combining automation with human oversight, our approach takes an initial step toward improving accessibility and robustness in evidence synthesis. To guide our study, we focus on the practical capabilities and evaluation of the proposed system within the SLR workflow.

Our research questions (RQs) are as follows:

RQ1: How can a multi-agent LLM system support the protocol validation and compliance steps of the SLR process?

RQ2: How well does the system's output align with PRISMA standards, based on initial expert evaluations, and does the system offer measurable improvements in efficiency or consistency during SLR evaluation?

2 Background and Related Work

SLRs are essential for synthesizing research findings, identifying gaps, and guiding future studies. The PRISMA guidelines (Moher et al., 2009; Page et al., 2021) offer a structured approach to ensure transparency and reproducibility. Tools like Covidence (Covidence Systematic Review Software, 2024) and Rayyan (Ouzzani et al., 2016) assist in screening and data extraction but rely heavily on manual effort, making them prone to fatigue, selection bias, and inconsistency.

Recent work explores leveraging LLMs for automating SLR stages such as study identification, summarization, and quality assessment (Susnjak, 2023; Ge and Others, 2024; Smith and Others, 2023; Jones and Others, 2022). State-of-the-art LLMs like GPT-4, Claude 3.7, Llama, and Gemini 2.5 (OpenAI, 2025b; DeepMind, 2025; Anthropic-AI, 2025; Meta-AI, 2025) demonstrate impressive few-shot learning capabilities (Brown et al., 2020), making them suitable for structured tasks with minimal supervision.

To enhance performance on complex, multi-step tasks, MAS have emerged as powerful frameworks capable of decomposing problems, enabling cooperative reasoning, and outperforming single-agent models on structured benchmarks (Park et al., 2025; Zhang et al., 2025; Wang et al., 2024). Recent MAS-driven tools illustrate this shift: Google's AI Co-Scientist (Gottweis et al., 2025) leverages Gemini-based agents to generate hypotheses and propose experiments; OpenAI's Deep Research (OpenAI, 2025a) conducts autonomous literature synthesis via web search; and SciSpace (SciSpace, 2025) offers interactive document parsing and drafting.

While existing tools support general scientific exploration and offer basic workflow automation, they often lack the methodological rigor required for systematic reviews. Similarly, recent studies underscore the limitations of core monolithic LLMs in structured tasks: Lieberum et al. reviewed 37 GPT-based SR prototypes and found them largely unvalidated; Penzo et al. demonstrated that LLaMA2-13B exhibits prompt and structure sensitivity; and Wei et al.; Wang et al. reported only modest gains, ranging from 3.9% to 17.9%, from techniques like chain-of-thought and self-consistency across various benchmarks. These findings suggest inherent performance ceilings in end-to-end LLM pipelines for SRs. To address the limitations posed by monolithic LLMs, we introduce a modular, multi-agent LLM system explicitly aligned with PRISMA guidelines, where each checklist item is handled by a specialized agent under expert oversight. Early results indicate improved consistency and reliability. Our open-source implementation[1] promotes transparency, supports scalable, high-quality reviews, and provides a foundation for robust, end-to-end SLR support.

3 Methodology & System Design

3.1 Architecture

Our MAS-LLM SLR Evaluation Framework consists of 27 specialized agents organized into six PRISMA-aligned societies (see Table 1) along with two utility agents (PDF Parsing and Follow-up Conversation). This structure (number of agents and division in societies) mirrors the PRISMA checklist which clearly elicits the checklist items for each part of the systematic reviews (Page et al., 2021; Susnjak, 2023), with one agent per checklist item. Sections like **Methods** have more agents (11) due to their detailed protocol requirements (greater number of checklist items), while sections like **Discussion** require only one agent. Early experiments with multi-item agents resulted in unstable behavior—overloaded agents, degraded performance, and unpredictable agent spawning, so we adopted a one-agent-per-item design with one-shot detailed prompting along with examples for robustness and clarity. All agents use GPT-4.1 (OpenAI, 2025b), a state-of-the-art LLM built for agentic workflows with a 1M token context window. Upon SLR PDF upload, an OCR-enabled Vision–Language Model (Unstructured Technologies, Inc., 2025) converts it into structured text.

A **Coordinator Agent** and **Task Agent** decompose the PRISMA checklist into modular evaluation tasks, dispatching them via few-shot prompts to specialized agents. Each agent uses the arXiv

[1] GitHub link will be added here upon publication

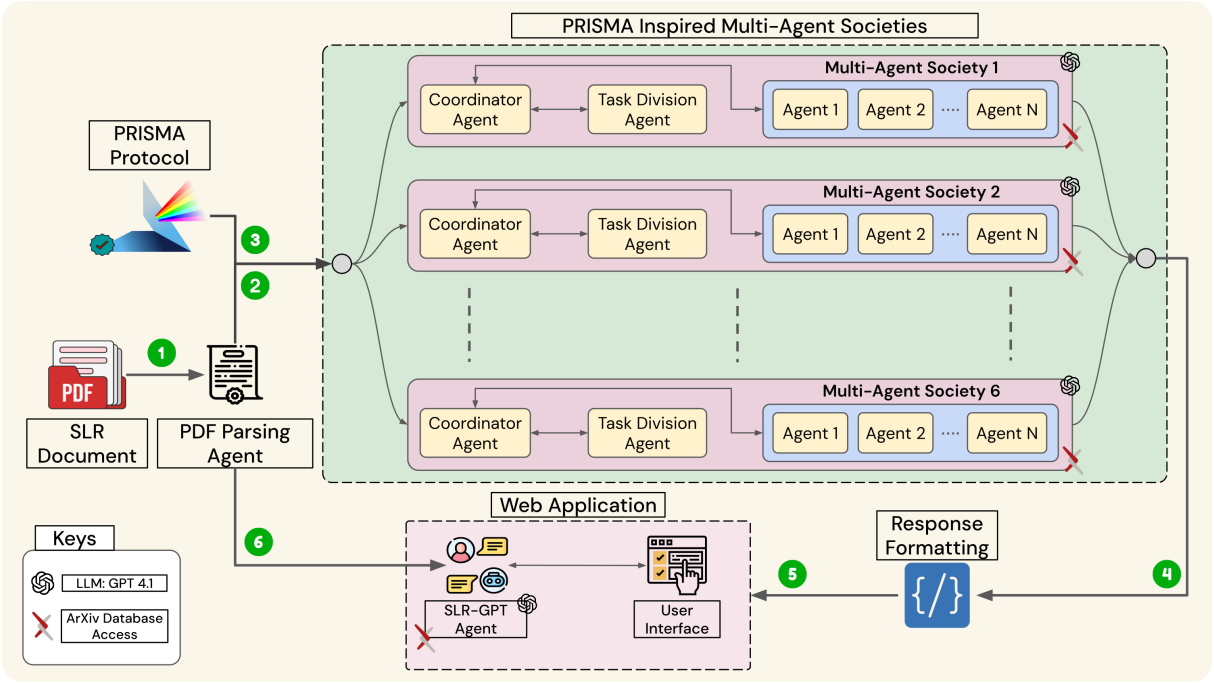

Figure 2: Architecture of the MAS-LLM SLR evaluation framework. The user-uploaded SLR document is processed by multiple agentic societies designed to evaluate it according to PRISMA guidelines. The results are displayed on the web user interface, which also supports follow-up questions and interactive conversations.

Toolkit to retrieve relevant research as needed, assigns a 0–5 score, and provides qualitative feedback. If outputs fall below thresholds, the Coordinator reallocates tasks or spawns new agents. As shown in Figure 2, agent outputs are synthesized into a unified format, accessible via web-interface and provided to the Follow-up Conversation Agent.

Society	Function	Agents
Abstract & Title	Evaluate title clarity and abstract completeness	2
Introduction	Assess rationale, scope, and objectives	2
Methods	Check eligibility criteria, search strategy, and bias assessment	11
Results	Verify result reporting, visualizations, and statistical summaries	7
Discussion	Examine interpretation, limitations, and implications	1
Other Information	Review registration, funding, conflicts of interest, and data policies	4
Standalone	PDF parsing and follow-up dialogue	2

Table 1: Agent societies, their responsibilities, and the number of agents in each society.

The follow-up conversation agent named *SLR-GPT Agent* provides co-pilot style research support through professional interactions. Using the same arXiv Toolkit available to all agents, it suggests new papers, verifies citations, cross-checks literature results, and recommends editorial refinements to maximize PRISMA compliance. By combining structured evaluation outputs from agents from societies with in-context retrieval for both the original manuscript and PRISMA protocol, it transforms assessments into actionable guidance for manuscript improvement. Open source implementation [2] is also shared for reproducibility.

3.2 Evaluation Methodology

We assess our framework on five published SLRs from diverse fields (Medical, E-commerce, AI, Metaverse, IoT), comparing agent outputs against ratings by three expert SLR reviewers. Both agents and human experts score each PRISMA item on a 0–5 scale (0 = Not Addressed, ..., 5= Thoroughly Addressed), enabling a standardized, ordinal evaluation across sections. Agent prompts incorporate PRISMA guidelines and one-shot exemplars to standardize evaluation; human reviewers assess the original manuscripts along with PRISMA guidelines. We quantify alignment using Mean Absolute Error (MAE), agreement level using MAE and additional statistical analysis to pinpoint areas where multi-agent collaboration aligns best with human experts. This benchmark demonstrates the tech-

[2] GitHub link will be added here upon publication

nical viability and interdisciplinary applicability of agentic LLMs in streamlining SLR workflows across diverse scientific domains.

4 Results

4.1 MAE & Agreement Level

We evaluated our system on five published SLRs from diverse domains (Medical, AI, Ecommerce, IoT, Metaverse), comparing its PRISMA-based scores against expert human reviewers (Section 3.2). Agreement percentages, computed as $100\% - (\text{MAE}/5 \times 100)$, are shown in Figure 3.

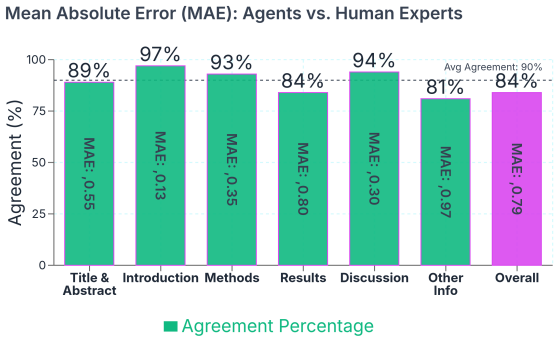

Figure 3: Agreement between agents and humans. Avg. agreement is 84%, with the highest alignment in Introduction (97%) and lower in Other Information (81%).

According to our results, the overall agreement is **84%**, with strongest alignment in *Introduction* (97%), *Discussion* (94%), and *Methods* (93%). Alignment dips only slightly to *Results* (84%) and *Other Information* (81%). In absolute terms, the highest agreement exceeds the overall agreement by 13 points, while the lowest is just 3 points below, indicating that all section-level agreements fall within a narrow window around the mean (close alignment overall). These agreement levels demonstrate that our system faithfully reproduces expert judgments on core review components by aligning itself with the PRISMA protocol while highlighting opportunities for future improvement. This alignment with experts also shows that our proposed MAS supports the protocol validation and compliance mainly through its system design and architectural choices mentioned in Section 3.

To address the significant delays in traditional peer review, averaging 15 weeks for the first round, with 10 weeks in medical sciences, 14 in natural sciences, and 17 in social sciences (SciRev, 2014), and up to 25 weeks in fields like Economics (Huisman and Smits, 2017), we present a automated MAS review system. It analyzes papers in just 15–20 minutes based on length and complexity, offering early-stage insights that can guide and accelerate subsequent human reviews, effectively reducing overall turnaround time.

4.2 Paper-wise & Inter-expert Analysis

Per-paper analysis (Appendix Fig.4) shows agent scores consistently track human evaluations across all five SLRs. Expert comparison (Appendix Fig.5) indicates the highest inter-rater variability. High inter-expert agreement—Intraclass Correlation Coefficient = 0.924, Krippendorff's $\alpha = 0.889$, Pearson $\rho = 0.898$ (Appendix Table 2), validates our human benchmark and suggests remaining divergences highlight targets for system refinement.

5 Future Directions

To ensure practical utility, we plan to deploy an interactive, browser-based interface that allows users to ask questions, revise summaries, and re-score sections, enabling both quantitative and qualitative evaluation of UI design and agent responsiveness. We will test the system with real-world users, systematic reviewers, and authors to assess its collaborative effectiveness. Structured feedback will be collected using Likert scales to evaluate clarity, usefulness, and trust, following best practices in human–AI interaction (Zhao et al., 2024; Rong et al., 2022). This feedback will be integrated into a preference-tuning loop to better align agent behavior with user preferences (Shao et al., 2025).

6 Conclusion

We introduced a multi-agent system for evaluating Systematic Literature Reviews aligned with the PRISMA protocol. By leveraging specialized agents for protocol validation, topic relevance, structural assessment, and ArXiv integration, the system aims to reduce the manual burden of interdisciplinary SLR evaluation. The built-in SLR-GPT Assistant supports citation checks and editorial feedback. An initial small-scale empirical evaluation on five SLRs from diverse domains showed 84% agreement with expert SLR judgments. While promising, these results are preliminary. This work offers a first step toward scalable, accurate MAS for SLRs, with future efforts focused on broader evaluations and system refinement with a focus on human-AI collaboration.

Limitations

Our MAS-LLM framework shows early promise, but several limitations should be acknowledged. The current evaluation spans five SLRs from distinct domains. In subsequent studies, we are looking towards increasing the number of papers and hence the credibility of our results. Agent performance is bounded by the capabilities of current LLMs, which may overlook fine-grained domain knowledge in technical contexts. The system's integration with arXiv enhances open-access coverage but excludes other key databases like PubMed or Scopus, creating potential gaps. Moreover, the system currently supports only evaluation, not real-time drafting or collaboration. Nonetheless, this study demonstrates the feasibility of structured, agentic LLM support for SLRs and lays the groundwork for more scalable and interactive systems in future work.

References

Anthropic-AI. 2025. Claude 3.7 Sonnet and Claude Code. Technical report, Anthropic AI. Retrieved from https://www.anthropic.com/news/claude-3-7-sonnet.

Andrew Booth, Martyn-St James, Mark Clowes, Anthea Sutton, and 1 others. 2021. Systematic approaches to a successful literature review.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Covidence Systematic Review Software. 2024. Covidence: Streamlining systematic review workflow. Online resource. Available at https://www.covidence.org/.

DeepMind. 2025. Gemini: Revolutionizing AI with multimodal capabilities. Technical report, DeepMind. Retrieved from https://deepmind.google/technologies/gemini/.

X Ge and Others. 2024. AI-driven systematic reviews in health research. *Medical Informatics Journal*.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, and 1 others. 2025. Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864*.

Maria J Grant and Andrew Booth. 2009. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health information & libraries journal*, 26(2):91–108.

Jeroen Huisman and Jeroen Smits. 2017. Duration and quality of the peer review process: the author's perspective. *Scientometrics*, 113(1):633–650.

L Jones and Others. 2022. Automating sentiment analysis in systematic reviews. *AI Review Journal*.

Judith-Lisa Lieberum, Markus Töws, Maria-Inti Metzendorf, Felix Heilmeyer, Waldemar Siemens, Christian Haverkamp, Daniel Böhringer, Joerg J. Meerpohl, and Angelika Eisele-Metzger. 2025. Large language models for conducting systematic reviews: on the rise, but not yet ready for use—a scoping review. *Journal of Clinical Epidemiology*, 181:111746.

Meta-AI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal AI innovation. Technical report, Meta AI. Retrieved from https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G Altman. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *BMJ*, 339:b2535.

OpenAI. 2025a. Introducing deep research. https://openai.com/index/introducing-deep-research/. Accessed: 2025-05-06.

OpenAI. 2025b. Introducing GPT-4.1 in the API.

Mourad Ouzzani, Hossam Hammady, Zbys Fedorowicz, and Ahmed Elmagarmid. 2016. Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews*, 5:210.

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, and 1 others. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372:n71.

Sooyoung Park, Franziska Roesner, Tadayoshi Kohno, and Daniel Weld. 2025. Why do multi-agent LLM systems fail? *arXiv preprint arXiv:2503.13657*.

Nicolò Penzo, Maryam Sajedinia, Bruno Lepri, Sara Tonelli, and Marco Guerini. 2024. Do LLMs suffer from multi-party hangover? a diagnostic approach to addressee recognition and response selection in conversations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11210–11233. Association for Computational Linguistics.

Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel,

Gjergji Kasneci, and Enkelejda Kasneci. 2022. Towards human-centered explainable ai: A survey of user studies for model explanations. *arXiv preprint arXiv:2210.11584*.

SciRev. 2014. Average duration first review round 15 weeks. Accessed: 2025-05-18.

SciSpace. 2025. Scispace: AI chat for scientific pdfs. https://typeset.io/. Accessed: 2025-05-06.

Abigail Sellen and Eric Horvitz. 2024. The rise of the AI co-pilot: Lessons for design from aviation and beyond. *Communications of the ACM*, 67(7):18–23.

Zekai Shao, Yi Shan, Yixuan He, Yuxuan Yao, Junhong Wang, Xiaolong Zhang, Yu Zhang, and Siming Chen. 2025. Do language model agents align with humans in rating visualizations? an empirical study. *arXiv preprint arXiv:2505.06702*.

J Smith and Others. 2023. Sentiment analysis in systematic literature reviews. *Computational Linguistics*.

Teo Susnjak. 2023. PRISMA-DFLLM: An extension of PRISMA for systematic literature reviews using domain-specific finetuned large language models. *arXiv preprint arXiv:2306.14905*.

Unstructured Technologies, Inc. 2025. Unstructured - your GenAI has a data problem. Accessed: February 27, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xuezhi Wang, Yixin Zhou, Dale Schuurmans, and 1 others. 2023. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the 2023 EMNLP*, pages 4116–4128. Association for Computational Linguistics.

Zihan Wang, Xiang Ren, and 1 others. 2024. Llm multi-agent systems: Challenges and open problems. *arXiv preprint arXiv:2402.03578*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Harm Le, and 1 others. 2022. Chain of thought prompting elicits reasoning in large language models. In *Proceedings of the 2022 EMNLP*, pages 2480–2493. Association for Computational Linguistics.

Ke Zhang, Yali Zhang, Jinlong Li, and 1 others. 2025. A comprehensive survey on multi-agent cooperative decision-making. *Information Fusion*. ArXiv:2503.13415.

Lingjun Zhao, Nguyen X. Khanh, and Hal Daumé III. 2024. Successfully guiding humans with imperfect instructions by highlighting potential errors and suggesting corrections. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 719–736.

A Appendix

A.1 Papers (SLR)-Wise Analysis

Figure 4 presents a paper-wise comparison of average PRISMA scores assigned by our MAS-LLM system versus human experts across five SLRs. The mean absolute error (MAE) by paper ranges from 0.05 (Paper 3) to 0.44 (Paper 2), demonstrating consistently strong alignment and identifying specific instances for targeted model refinement.

Figure 4: Paper-wise comparison of average scores (Agents vs. Humans).

A.2 Human Experts' Scores

Figure 5 illustrates inter-expert variability across PRISMA checklist societies. Methods sections exhibit the highest reviewer disagreement—reflecting the inherent complexity of methodological assessments—whereas Title & Abstract sections achieve the greatest consensus.

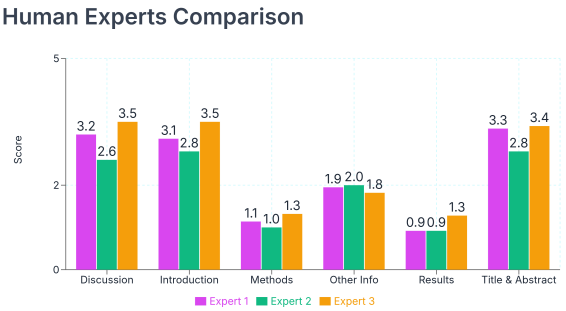

Figure 5: Variation in human expert scores by society.

A.3 Inter-Expert Reliability Analysis

Table 2 summarizes inter-expert reliability metrics, including Intraclass Correlation Coefficient (ICC), Krippendorff's Alpha, and average Pearson ρ, all exceeding 0.88. These high values confirm the robustness of our expert benchmark and substantiate the validity of comparing agent outputs against human ratings.

Metric	**Value**	**Interpretation**
Intraclass Correlation Coefficient (ICC)	0.924	Excellent reliability
Krippendorff's Alpha	0.889	Strong agreement
Avg. Inter-Human Pearson ρ	0.898	Strong correlation

Table 2: Inter-expert agreement metrics.