2026 ASEE ANNUAL CONFERENCE June 21–24, 2026 | Charlotte Convention Center, NC

Host Sponsor UNC Charlotte

Paper ID #52646

Assessing the Reliability of Large Language Models for Scientific Information Extraction

Luke Schneider, Pennsylvania State University, Behrend College
Debalina Maitra, Kennesaw State University

Assistant Professor of Teacher Leadership

Jing Zhao, Pennsylvania State University, Behrend College
Dr. Jiawei Gong, Pennsylvania State University, Behrend College

Dr. Jiawei Gong is an Associate Professor of Mechanical Engineering at Penn State Behrend. He earned his PhD (2017) and MS (2014) in Mechanical Engineering from North Dakota State University and a BE in Polymer Materials and Engineering from East China University of Science and Technology (2010). Dr. Gong's research interests include energy materials and devices, particularly dye-sensitized and perovskite solar cells. His recent research focuses on leveraging machine learning for data acquisition and analysis, aimed at advancing solar energy technologies.

Assessing the Reliability of Large Language Models for Scientific Information Extraction

Large language models (LLMs) are increasingly used to extract structured information from scientific publications, yet systematic evaluation of their reliability for information extraction (IE) remains a challenge. In this work, we evaluate the performance of multiple state-of-the-art LLMs on publication-grounded IE tasks, focusing on the extraction of machine learning metadata, including input *features*, *target variables*, *dataset size*, *model types*, and *performance metrics*, from peer-reviewed scientific papers. We adopt a sequential evaluation methodology that distinguishes instruction compliance, information completeness, and factual correctness, enabling clearer attribution of extraction errors. All evaluation metrics are normalized on a continuous scale from 0 to 1, where 1.00 represents perfect performance. Using zero-shot prompts, we benchmark three LLMs under uninformed conditions and observe substantial performance variation, with compliance scores ranging from 0.60 to 1.00, completeness from 0.67 to 0.80, and correctness from 0.87 to 1.00. We further examine an informed setting in which models are exposed to peer-generated extractions, revealing heterogeneous adaptation behaviors: some models improve correctness to a perfect score of 1.00, while others trade instruction adherence for precision. These results demonstrate that high correctness alone does not guarantee reliable IE and that LLMs exhibit distinct extraction strategies. The proposed evaluation approach provides a practical framework for assessing LLM-based IE in educational and research workflows, supporting more trustworthy use of AI tools in scientific literature analysis.

1. Introduction

Large language models (LLMs) are increasingly integrated into scientific workflows, reshaping how researchers interact with published knowledge. Through the multimodal synthesis of text, figures, and tables, LLMs can assist with literature review, hypothesis generation, and automated data curation. Such tasks often involve information extraction, reasoning, summarization, and knowledge generation. Information extraction (IE) focuses on identifying and structuring factual information that is explicitly stated in a publication. In contrast, reasoning requires models to draw logical conclusions or comparisons based on extracted facts, summarization aims to compress these facts while preserving overall meaning, and knowledge generation extends beyond the publication by introducing external knowledge or new hypotheses. Several studies have demonstrated LLMs' capacity to perform generative IE from materials-science literature. For instance, Polak and Morgan [1] used a zero-shot (models are given task instructions without task-specific context) prompt-based approach to guide GPT-4 in extracting structured materials properties from research papers, achieving about 90.8% precision at 87.7% recall without any task-specific fine-tuning. Similarly, Zheng et al. [2] deployed ChatGPT as a "chemistry assistant" to mine over 26,000 metal–organic framework synthesis parameters from 228 publications, reporting precision and recall >90% via carefully engineered instructions.

Despite these promising results, existing work on LLM-based IE in materials science exhibits substantial variation in task formulation, prompt design [3], and evaluation practice. Foppiano et al. [4] benchmarked GPT-3.5 and GPT-4 on materials named-entity recognition and material–property relation extraction, showing that while LLMs can effectively associate materials with reported properties, they remain sensitive to schema ambiguity and often underperform domain-specialized models on exact entity extraction. At a larger scale, Ghosh and Tewari [5] employed GPT-4–based autonomous agents to extract thermoelectric properties from thousands of articles,

achieving F1 scores exceeding 0.9 for numeric property extraction, but also highlighting significant trade-offs between cost, completeness, and verification fidelity in agent-driven pipelines. Gupta et al. [6] combined GPT-based prompting with domain-aware preprocessing to extract over one million polymer property records from literature, reporting strong performance for certain properties but uneven accuracy across different extraction slots. These studies demonstrate that LLMs can automate factual data extraction, yet they also reveal a lack of standardized evaluation frameworks that clearly distinguish missing information, hallucinated content, and instruction-following errors. Moreover, generative IE is frequently entangled with higher-level reasoning or summarization, obscuring whether observed errors arise from factual extraction errors or downstream interpretation.

Rather than relying on a single generative agent, multi-agent architectures distribute extraction, critique, and verification across multiple models, enabling independent assessment of evidence, identification of inconsistencies, and convergence toward a consensus. This paradigm mirrors human scientific practice, in which knowledge claims are refined through peer review. Prior work on LLM debate frameworks and self-consistency mechanisms [7] has demonstrated that cross-examination can reduce hallucinations and improve factual reliability in complex reasoning tasks. Similarly, judge-based and verifier-based architectures [8], where one or more agents explicitly evaluate or arbitrate the outputs of others, have been shown to improve answer faithfulness and calibration. However, despite the growing adoption of these frameworks, it remains unclear whether reported gains arise from improved factual extraction, stricter instruction adherence, or downstream reconciliation of conflicting interpretations.

In this study, we adopt an evaluation framework grounded in the ordered 3C principles [9], i.e., Compliance, Completeness, and Correctness, designed specifically for document-grounded IE. Compliance is evaluated first, as it captures hallucinations [10], non-solicited information, and schema violations, thereby measuring faithfulness to user instructions and source documents. Completeness then assesses the extent to which all required factual elements are extracted, while Correctness evaluates factual accuracy against human expert annotations. Beyond single-model evaluation, we introduce a cross-model adjudication process in which multiple LLMs are exposed to one another's extracted results and allowed to revise their outputs considering the peer evidence, an approach inspired by ensemble reasoning and collaborative verification process. This interaction enables analysis of behavioral properties such as willingness to revise prior answers, responsiveness to stronger evidence, and cooperative error correction, dimensions that are largely invisible to conventional accuracy-based metrics. Because large-scale human labeling for information extraction is costly and often unavailable, especially in scientific domains, the proposed multi-agent framework provides a scalable pathway to improving the reliability and robustness of automated IE systems.

Beyond benchmarking model performance, this framework directly addresses pedagogical needs in teaching and mentoring contexts. In classroom settings, LLMs are increasingly used by students to summarize literature, extract structured information, and assist with technical writing. However, without structured evaluation, students may conflate fluent output with factual reliability. The ordered 3C framework provides a transparent rubric that instructors can use to teach responsible AI-assisted scholarship by distinguishing instruction adherence (Compliance), coverage of required elements (Completeness), and factual alignment with expert ground truth (Correctness).

Similarly, for faculty and research mentors, the framework supports informed decision-making regarding whether and how LLM tools should be integrated into literature review workflows, research training, and data curation pipelines. By decomposing reliability into interpretable dimensions, the proposed methodology enables AI literacy development rather than blind adoption or categorical rejection of LLM tools.

2. Methods

LLMs are prompted to extract structured experimental metadata, defined as explicitly reported publication-level variables such as input features, target variables, dataset size, model types, performance metrics, and problem type from scientific manuscripts (PDFs). Extraction performance is assessed using an ordered 3C sequential evaluation framework, which refers to the ordered application of the 3C criteria (Compliance → Completeness → Correctness), where later dimensions are evaluated only after earlier criteria are satisfied applied sequentially. This ordered design improves interpretability and reduces annotation effort by filtering non-compliant responses early, which is particularly beneficial when using LLM-as-judge methods. The sequential structure also enables modular evaluation: compliance is assessed through schema and instruction checks, completeness through coverage of required information elements, and correctness through direct comparison with human-expert annotations.

We selected a machine learning research paper as the evaluation domain for two primary reasons. First, ML publications typically contain several structural elements including defined inputs, outputs, datasets, models, and performance metrics, all of which enable objective benchmarking of extraction tasks. Second, this domain aligns with the authors' expertise, allowing reliable construction of expert-generated ground truth. Importantly, the proposed framework is domain-agnostic and can be applied to other scientific disciplines, provided that high-quality expert annotations are available to establish ground truth. The critical requirement is not the specific domain, but the availability of expert-validated reference data against which completeness and correctness can be assessed.

Figure 1 illustrates the evaluation workflow using a representative example, including the user prompt, the model-generated structured response, and expert annotations. Compliance serves as a gatekeeping criterion by determining whether the response adheres to prompt instructions, schema requirements, and document grounding. Completeness then evaluates whether all required information elements are present in the compliant response, independent of factual accuracy. Correctness is assessed last and measures factual alignment with domain-validated ground truth. By conditioning later stages on earlier ones, the framework ensures that errors are attributed to instruction-following, coverage, or factual accuracy, rather than conflating these distinct failure modes.

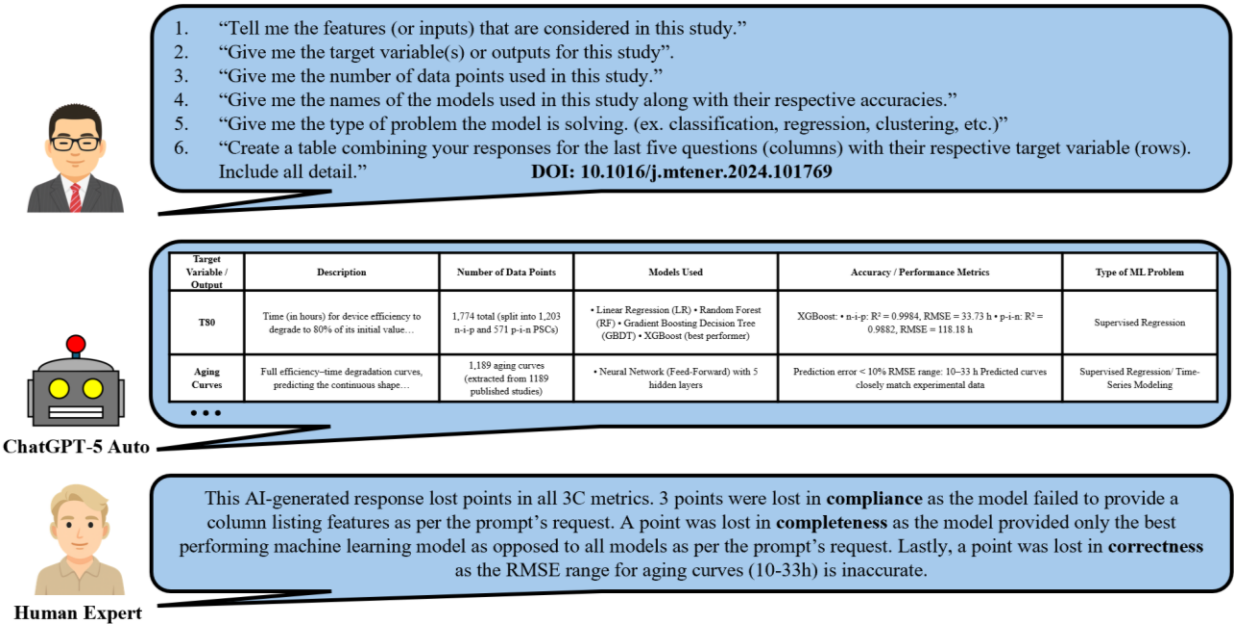

Fig. 1. Example of uninformed user prompt, model response, and expert annotations.

After LLMs independently generate responses, their outputs are shared with other models through a human-generated meta-prompt that requests revision, as illustrated in Fig. 2. In the first stage (AI-determined answer), multiple LLMs respond to the same questions, producing parallel extractions without cross-model influence. These initial responses are evaluated using the ordered 3C framework to establish a single-agent baseline. In the second stage (AI-determined agreeance), each model is exposed to peer responses under identical constraints and invited to revise its output. Evaluation in this stage reflects behavioral dynamics rather than output quality alone. In particular, willingness to change measures whether and how an agent updates its response after being exposed to peer extractions, while additional behaviors such as cooperative verification and cross correction assess whether revisions are evidence-driven. Comparing responses before and after cross-agent exposure isolates epistemic adaptability and collaborative robustness, enabling systematic analysis of how multi-agent interaction influences information extraction performance. The prompts used for both uninformed and informed tests can be seen in Table 1.

Table 1. Specific prompts for both uninformed and informed tests.

Specific Test	**Prompts**
Uninformed Test	1. "Tell me the features (or inputs) that are considered in this study." 2. "Give me the target variable(s) or outputs for this study". 3. "Give me the number of data points used in this study." 4. "Give me the names of the models used in this study along with their respective accuracies." 5. "Give me the type of problem the model is solving. (ex. classification, regression, clustering, etc.)" 6. "Create a table combining your responses for the last five questions (columns) with their respective target variable (rows). Include all detail."
Informed Test	"Here are the responses of two other large language models to the exact same set of prompts and exact same paper and supplementary information **Original prompts, paper, supplementary info, and other responses attached*. If you would like, you may update your response based off of responses provided but the other LLMs. Please create the same table structure as before, simply (if applicable) updated."

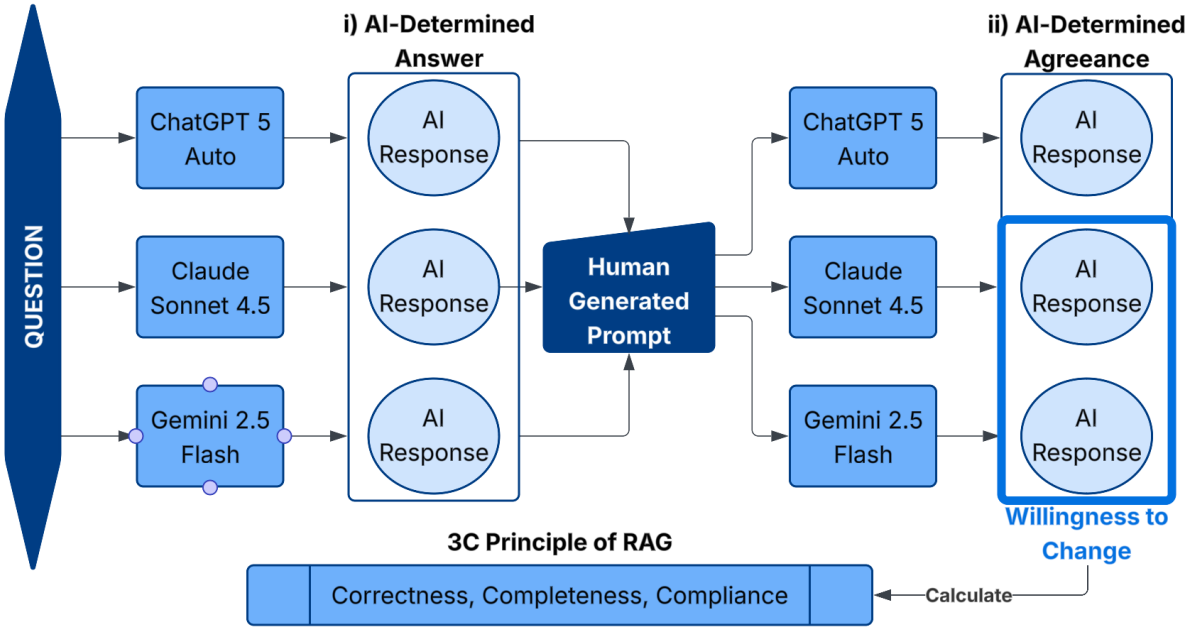

Fig. 2. Schematic of the two-stage multi-agent evaluation pipeline.

3. Results and Discussion

Responses were evaluated by domain experts who calibrated outputs against ground truth to assess compliance, completeness, and correctness. For benchmark construction, authors independently read the full manuscript used for evaluation, generated structured ground-truth annotations, and reconciled discrepancies through discussion until consensus was reached. This consensus-based expert annotation served as the reference standard for compliance, completeness, and correctness scoring. Figure 3 summarizes single-agent IE performance under uninformed conditions using the ordered 3C evaluation framework. Here, the uninformed condition denotes single-agent extraction in which models generate responses independently, without exposure to peer outputs. Clear performance differences emerge across compliance, completeness, and correctness. Gemini 2.5 Flash achieves perfect compliance, producing structurally valid outputs that strictly adhere to prompt instructions and schema constraints. In contrast, Claude 4.5 Sonnet and ChatGPT-5 Auto both omit the features column, resulting in identical compliance scores. However, the underlying causes differ: Claude lists all machine learning (ML) models evaluated in the study, whereas ChatGPT reports only the best-performing model. This distinction affects downstream completeness. As a result, ChatGPT exhibits the lowest completeness score (0.67), while Claude and Gemini both achieve higher completeness (0.80). Notably, Claude's completeness deficit is primarily driven by its compliance violation (missing features), whereas Gemini's lower completeness arises despite full structural adherence, reflecting limitations in coverage rather than formatting. Correctness is evaluated independently of compliance and completeness. Under this criterion, Claude 4.5 Sonnet achieves perfect correctness, indicating that all extracted information provided in the response aligns with the ground truth. ChatGPT-5 Auto incurs a single error in model-performance extraction, while Gemini 2.5 Flash exhibits multiple factual inaccuracies despite its high compliance. Consequently, correctness follows the ordering Claude 4.5 Sonnet > ChatGPT-5 Auto > Gemini 2.5 Flash, independent of schema adherence or coverage. These model-specific tradeoffs are consistent with prior studies of LLM-based information extraction in materials science. Shi et al [9], compared GPT-4 Turbo, Claude 3 Opus, and Gemini 1.5 Pro on

synthesis-condition extraction for metal–organic frameworks and observed that Claude and Gemini achieved higher completeness, while Gemini exhibited stronger instruction adherence and structured responses. Despite lower quantitative extraction performance, GPT-4 demonstrated strong logical reasoning and contextual inference capabilities.

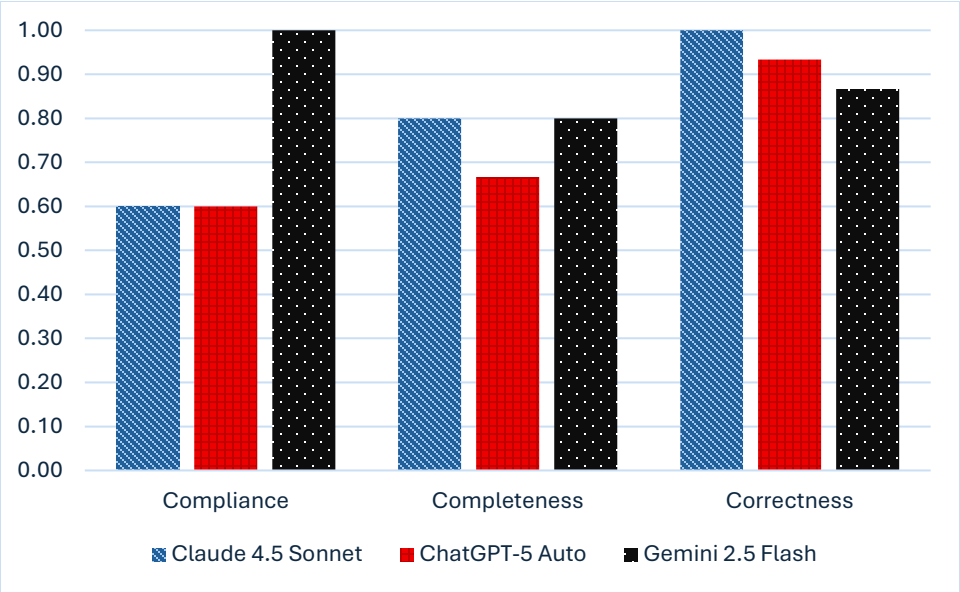

Fig. 3. Uninformed, zero-shot prompt IE performance.

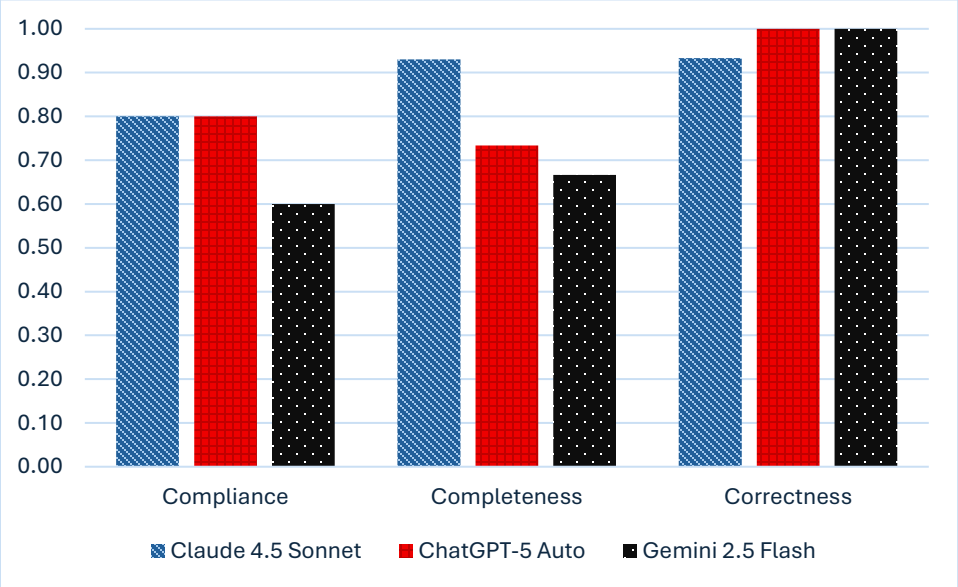

Fig. 4. Informed, zero-shot prompt IE performance.

Following this, an informed test was conducted, meaning models were shown peer-generated extractions and invited to revise their responses under identical prompts. Under the informed condition (Fig. 4), the three models exhibit distinct adaptation patterns in response to improved evidence. Claude 4.5 Sonnet shows substantial gains in both compliance (from 0.60 to 0.80) and completeness (from 0.80 to 0.93), while maintaining high correctness (0.93). This behavior

indicates strong adaptability: when reliable evidence is available, the model more effectively constrains its output to the source document while expanding coverage. ChatGPT-5 Auto benefits consistently across all three dimensions, reaching perfect correctness (1.00) from an initial 0.93, improving compliance from 0.60 to 0.80, and increasing completeness modestly from 0.67 to 0.73. These gains suggest a verification-oriented extraction strategy that leverages peer evidence to improve factual accuracy without aggressive expansion of output breadth. In contrast, Gemini 2.5 Flash exhibits a tradeoff: correctness improves to 1.00, but compliance decreases from 1.00 to 0.60, while completeness remains comparatively low at 0.67. This pattern indicates a precision-dominant response strategy in which factual confidence increases at the expense of strict instruction adherence. The net changes across metrics are summarized in Figure 5, highlighting that peer evidence does not uniformly improve all evaluation dimensions and instead amplifies model-specific extraction strategies. A summary of model-specific behavioral tendencies observed across evaluation dimensions can be seen in Table 2.

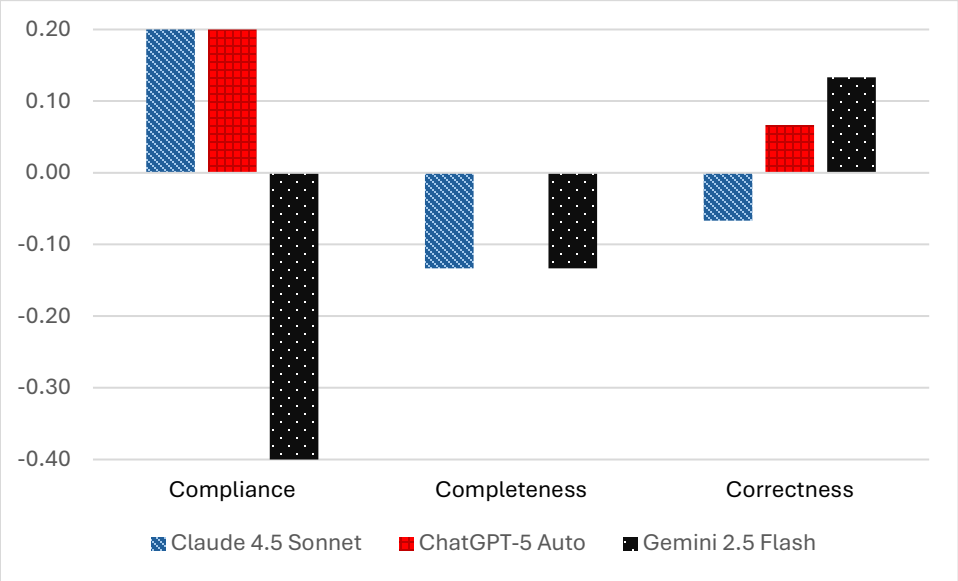

Fig. 5. Changes of informed LLM IE with respect to the uninformed conditions.

Table 2. Observed advantages and disadvantages of LLMs through uninformed and informed testing

LLM	Observed Advantages	Observed Disadvantages
Claude 4.5 Sonnet	High factual precision; strong correctness under both conditions; adaptable to peer evidence	Occasional schema violations; compliance inconsistencies
ChatGPT-5 Auto	Balanced performance; verification-oriented strategy; improved correctness under informed setting	Conservative coverage; moderate completeness limitations
Gemini 2.5 Flash	Strong schema adherence in uninformed conditions; structured output reliability	Tradeoff between compliance and correctness under informed setting; coverage variability

4. Conclusions and Future Work

This work systematically evaluated large language models for document-grounded information extraction using an ordered 3C framework and quantified how cross-model adjudication alters

extraction behavior. Under uninformed, zero-shot conditions, the three evaluated models exhibited distinct performance profiles: Gemini 2.5 Flash achieved perfect compliance (1.00) but lower correctness (0.87), Claude 4.5 Sonnet demonstrated perfect correctness (1.00) despite lower compliance (0.60), and ChatGPT-5 Auto showed more balanced but moderate performance (compliance 0.60, completeness 0.67, correctness 0.93). When peer evidence was introduced, correctness improved across all models, reaching 1.00 for both ChatGPT-5 Auto and Gemini 2.5 Flash, yet compliance and completeness did not uniformly improve. Notably, Gemini 2.5 Flash's compliance decreased from 1.00 to 0.60 under informed conditions, while Claude 4.5 Sonnet's completeness increased from 0.80 to 1.00, indicating that peer exposure amplifies model-specific extraction strategies rather than enforcing convergence across evaluation dimensions. These results demonstrate that compliance, completeness, and correctness are non-redundant dimensions of IE performance and that gains in factual accuracy do not guarantee improved instruction adherence or information coverage. Consequently, single-metric accuracy evaluations are insufficient for assessing LLM reliability in scientific extraction tasks.

This study represents an initial exploration, but broader benchmarking across multiple papers, LLM models, and model versions is needed to assess the generalizability of the proposed framework. The present evaluation was conducted on a single publication using three contemporary models; future work should examine whether similar compliance–completeness–correctness tradeoffs persist across different scientific domains, document structures, and model architectures. Establishing cross-domain robustness will be essential for validating the ordered 3C framework as a general evaluation standard for document-grounded information extraction.

In addition, prompt engineering warrants further investigation as a potential source of compliance and completeness variation. Although all models in this study were evaluated using identical zero-shot prompts, it remains unclear how prompt specificity, schema rigidity, or explicit structural constraints influence extraction stability. Future experiments should examine whether refined or more constrained prompts reduce schema violations, improve coverage of figure-based or table-based information, and mitigate instruction-following errors without sacrificing factual accuracy. Alternative system designs, including few-shot prompting or retrieval-augmented workflows, may also provide improvements in reliability relative to zero-shot IE.

The performance differences observed between uninformed and informed conditions further motivate systematic study of multi-agent interaction dynamics. While peer exposure improved correctness in some cases, tradeoffs emerged in compliance and completeness, suggesting that cross-model influence does not uniformly enhance all reliability dimensions. Future research should investigate whether additional rounds of interaction lead to convergence, divergence, or oscillatory behavior, and whether specific orchestration strategies can amplify complementary model strengths. Exploring structured coordination mechanisms such as assigning distinct verification or schema-enforcement roles to different agents may clarify whether reliable information extraction is best achieved through consensus-based negotiation or through modular specialization within multi-agent systems.

Acknowledgement

This work was supported, in part, by a seed grant from the Penn State Office of the Vice President for Commonwealth Campuses.

References

1. Polak MP, Morgan D (2024) Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, 15(1):1569. https://doi.org/10.1038/s41467-024-45914-8

2. Zheng Z, Zhang O, Borgs C, Chayes JT, Yaghi OM (2023) ChatGPT Chemistry Assistant for Text Mining and the Prediction of MOF Synthesis. *Journal of the American Chemical Society*, 145(32):18048–18062. https://doi.org/10.1021/jacs.3c05819

3. Zhou Y, Muresanu AI, Han Z, Paster K, Pitis S, Chan H, Ba J (2023) Large Language Models Are Human-Level Prompt Engineers. https://doi.org/10.48550/arXiv.2211.01910

4. Foppiano L, Lambard G, Amagasa T, Ishii M (2024) Mining experimental data from materials science literature with large language models: an evaluation study. *Science and Technology of Advanced Materials: Methods*, 4(1):2356506. https://doi.org/10.1080/27660400.2024.2356506

5. Ghosh S, Tewari A (2025) Automated Extraction of Material Properties using LLM-based AI Agents. https://doi.org/10.48550/arXiv.2510.01235

6. Gupta S, Mahmood A, Shetty P, Adeboye A, Ramprasad R (2024) Data extraction from polymer literature using large language models. *Communications Materials*, 5(1):269. https://doi.org/10.1038/s43246-024-00708-9

7. Du Y, Li S, Torralba A, Tenenbaum JB, Mordatch I (2024) Improving factuality and reasoning in language models through multiagent debate. *Proceedings of the 41st International Conference on Machine Learning*, 235:11733–11763.

8. Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, Chen A, Goldie A, Mirhoseini A, McKinnon C, Chen C, Olsson C, Olah C, Hernandez D, Drain D, Ganguli D, Li D, Tran-Johnson E, Perez E, Kerr J, Mueller J, Ladish J, Landau J, Ndousse K, Lukosuite K, Lovitt L, Sellitto M, Elhage N, Schiefer N, Mercado N, DasSarma N, Lasenby R, Larson R, Ringer S, Johnston S, Kravec S, Showk SE, Fort S, Lanham T, Telleen-Lawton T, Conerly T, Henighan T, Hume T, Bowman SR, Hatfield-Dodds Z, Mann B, Amodei D, Joseph N, McCandlish S, Brown T, Kaplan J (2022) Constitutional AI: Harmlessness from AI Feedback. https://doi.org/10.48550/arXiv.2212.08073

9. Shi Y, Rampal N, Zhao C, Fu DJ, Borgs C, Chayes JT, Yaghi OM (2025) Comparison of LLMs in extracting synthesis conditions and generating Q&A datasets for metal–organic frameworks. *Digital Discovery*, :10.1039.D5DD00081E. https://doi.org/10.1039/D5DD00081E

10. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P (2023) Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, 55(12):248:1-248:38. https://doi.org/10.1145/3571730