Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ31033

Can LLMs Truly Understand Relations? An Evaluation of Relation Extraction Capabilities

Yuchen Li

Capital University of Economics and Business, Beijing, China
1243476214@qq.com

Abstract. Large language models (LLMs) are increasingly used for information extraction, and they often perform well when the target facts are stated plainly. In scientific papers, though, many relationships are only implied: the reader has to connect evidence across sentences, track context, and sometimes make several small inferences before a relation becomes clear. How reliably LLMs handle that kind of reading remains an open question. To study this, we propose an evaluation framework that tests relation extraction under progressively heavier reasoning demands. It relies on three prompt templates that grow more structured and more challenging, and we apply them to a curated perovskite solar cell dataset covering three core categories of scientific information—materials, processes, and performance. We then benchmark three representative models. The results show a consistent pattern. When relations are explicit, models extract them with high accuracy. As soon as the task requires deeper inference—especially for implicit or multi-hop relations that hinge on nuanced semantics—performance drops noticeably. The models also fail in different ways: some are comparatively good at maintaining coherent, step-by-step reasoning, while others are more dependable at avoiding hallucinated links or preserving type consistency. Even with these strengths, sizable gaps remain in coverage, reasoning depth, and overall completeness. Taken together, the findings clarify where current LLMs are genuinely useful for scientific information extraction and where they still fall short. They tend to be most reliable in structured extraction settings, while deeper semantic understanding remains the main bottleneck. By quantifying performance across prompt levels, the framework offers a practical yardstick for semantic comprehension and scientific reasoning in relation extraction, and it provides empirical support for downstream applications such as scientific knowledge graph construction.

Keywords: LLMs, relation extraction, semantic understanding, evaluation, information extraction

1. Introduction

Information extraction (IE) sits at the core of modern NLP because it turns raw text into usable structure. In practice, it powers tasks like mining the scientific literature, assembling knowledge graphs, and enabling data-driven discovery in areas such as materials research. Relation extraction (RE) is a key piece of that pipeline: it aims to identify the semantic links between entities so that

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ31033

scientific narratives—often long, dense, and technical—can be converted into representations that support search, analysis, and computation.

Scientific RE is rarely as simple as spotting a clearly stated fact. In fields like materials science and the life sciences, relationships are frequently implied rather than stated outright. Authors spread evidence across sentences, describe causal chains indirectly, or explain mechanisms in a way that assumes domain knowledge. Capturing these links demands more than pattern matching: it requires following the thread of an argument, integrating context across a passage, and drawing inferences that are not explicitly written on the page.

Traditional RE systems have usually relied on large labeled datasets and supervised learning. A common approach frames RE as a classification problem, built on pre-trained encoders such as BERT or RoBERTa. These methods can be strong when relation types are fixed and the test data closely resembles the training distribution. Scientific writing tends to expose their weak points. Transfer across domains is often brittle, indirectly expressed relations are easy to miss, and long-range dependencies—especially those crossing sentence boundaries—remain difficult. As a result, the rich, layered relational structure that scientific texts naturally contain is only partially recovered in many real-world settings.

The rise of large language models (LLMs), including GPT-style systems, Claude, and LLaMA, has reopened the conversation around prompt-based RE in zero-shot or few-shot regimes. The appeal is clear: prompts may reduce the need for costly annotation and, at least in principle, offer a more flexible way to handle open-ended semantics and cross-domain knowledge. Still, it is not obvious how well these models truly grasp scientific relational logic. Prior studies have reported encouraging outcomes on general IE benchmarks, yet scientific RE routinely leans on implicit causality, multi-step reasoning, and mechanistic connections. A further complication is evaluation itself: many existing studies focus on a single dataset slice, a single model, or a single task setup, making it hard to tell whether success comes from shallow extraction or from genuine reasoning.

This work tackles that gap by introducing a structured evaluation framework for LLM-based relation extraction in scientific text. A test set is built from five carefully chosen abstracts in perovskite solar cell (PSC) research, selected to cover a range of scientific problem types and reasoning patterns—materials design, process optimization, transport-layer engineering, physical simulation, and mechanistic chemistry. On top of this corpus, we design hierarchical evaluation templates that probe multiple levels of reasoning, from direct relation identification and cross-sentence integration to multi-hop causal inference and mechanism-level explanation. The evaluation combines quantitative metrics—entity F1, relation F1, relation-type accuracy, hallucination rate, and reasoning consistency—with targeted qualitative analysis, so that the assessment reflects not only structural correctness but also whether the reasoning is plausible and traceable. Controlled experiments on three widely used LLMs then reveal how extraction quality shifts as reasoning demands increase, helping map out where these models are reliable, where they break down, and what that implies for scientific knowledge extraction in high-complexity domains like materials science.

2. Background

Relation extraction (RE) and information extraction (IE) are foundational NLP tasks that convert free-form text into structured representations. This structured output is a prerequisite for downstream systems such as knowledge graphs, question answering, and large-scale information retrieval. Over time, extraction methods have evolved from feature-driven pipelines to neural architectures and, most recently, to prompt-based approaches built on large language models

DOI: 10.54254/2755-2721/2026.TJ31033

(LLMs). Despite this progress, robust extraction in realistic settings is still constrained by three recurring issues: capturing complex semantics expressed implicitly, learning effectively when labeled data are scarce, and transferring reliably across domains with different writing conventions and relation inventories.

2.1. Traditional relation extraction

Traditional relation extraction methods have evolved from feature-based classification and kernel methods to deep neural networks and pre-trained models. Early work, such as Kanimozhi et al. [1], employed string kernels and shallow linguistic features to effectively identify disease–gene relations in the biomedical domain, but struggled to capture complex semantics. Zhang et al. [2] enhanced lexical and syntactic representations by combining multi-scale convolutional networks with graph convolutional networks (GCNs) and metric learning, yet high-level semantic dependencies remained challenging. In the manufacturing domain, Du et al. [3] integrated attention mechanisms with GCNs for knowledge graph construction, improving performance in specific domains but with limited generalizability.

With the development of deep learning, models became better able to capture entity context and structural information. Zhao et al. [4] combined CNNs with GCNs to enhance entity feature extraction, though long-tail relations remained difficult to recognize. Distant supervision approaches, such as Hogan et al. [5], increased coverage using large-scale annotated corpora but were limited in extracting low-frequency relations. Models combining BERT with dependency syntax (e.g., Li et al. [6]) improved recognition of complex relations, albeit at the cost of increased model complexity. Early neural network combinations (Li et al. [7]) improved feature learning but were insufficient for cross-sentence or long-distance relations. Pre-trained models, particularly BERT-based methods (Christou et al. [8]; Su et al. [9]), excelled in extraction accuracy and noise handling but showed limited adaptability to cross-domain and low-resource scenarios. Tian et al. [10] leveraged dependency trees and attention mechanisms to enhance semantic guidance, though noise from automatically generated dependency trees could affect robustness.

In summary, traditional RE methods perform well on standard datasets, but limitations remain in complex semantic understanding, cross-sentence relation extraction, and low-resource scenarios, highlighting areas where LLM-based methods could offer improvements.

2.2. LLM-based information extraction

In recent years, large language models (LLMs) have demonstrated significant potential in information extraction and relation extraction. Dagdelen et al. [11] showed that GPT-3 and LLaMA-2 can jointly perform named entity recognition and relation extraction on scientific texts, flexibly outputting structured information, though still relying on task-specific fine-tuning. Wang et al. [12] proposed a unified instruction-tuning framework, achieving zero-shot and few-shot performance improvements across multi-task IE, demonstrating the generalization ability of instruction-guided LLMs, but limitations remain for highly complex semantic reasoning. For low-resource and few-shot extraction tasks, Haoran et al. [13] combined LLM-based data augmentation with prototypical networks to enhance few-shot Chinese relation recognition, mitigating feature sparsity caused by insufficient data. Xu et al. [14] further explored in-context learning (ICL) and LLM-generated data on top of traditional extraction methods, significantly improving few-shot relation extraction performance, though these methods depend on high-quality task instructions and constraints on generated data.

DOI: 10.54254/2755-2721/2026.TJ31033

In the extraction of pre-defined relation types, Huang et al. [15] proposed the ThreeEERE three-stage framework for complex event relation extraction. By employing improved automatic reasoning chains and gold-standard rules, LLMs achieved performance on multiple tasks that met or partially exceeded supervised methods, demonstrating their effectiveness in complex semantic understanding.

In open information extraction (OpenIE), Ling et al. [16] introduced uncertainty quantification of examples in in-context learning to increase confidence in generated relations, though LLMs still struggle to match the stability of supervised methods. Yu et al. [17] highlighted limitations of LLMs in causal reasoning tasks, noting that the ability to generate reasonable explanations remains constrained. Early scientific text extraction studies, such as Dunn et al. [18], showed that fine-tuning with a small amount of task-specific data can improve the accuracy of complex scientific information extraction, but generalization remains limited.

Overall, LLM-based information extraction methods exhibit strong flexibility and potential in few-shot, multi-task, pre-defined relation, and open-domain extraction scenarios. They complement traditional RE methods by addressing shortcomings in complex semantics and cross-sentence relation handling. However, these approaches still depend on task fine-tuning, example design, or the quality of generated data, and challenges persist in long-tail relations, cross-domain transfer, and complex logical and causal reasoning.

3. Approach

3.1. Workflow

This study aims to establish a general evaluation workflow based on prompt engineering to systematically assess large language models' (LLMs) capabilities in extracting structured information and performing hierarchical consistency reasoning in complex scientific texts. To ensure representativeness of the test scenarios, we selected scientific texts containing multiple entity types, highly specialized terminology, cross-sentence causal chains, and multi-level mechanistic descriptions as the generalized context, requiring models to handle both complex semantic relations and implicit logical structures simultaneously. Based on this, we designed a hierarchical prompt template system that progressively increases task semantic complexity, ranging from basic entity recognition to relation extraction and higher-order mechanistic reasoning, to observe systematic variations in models' prompt sensitivity, information integration ability, and reasoning consistency. The overall technical workflow of this study is illustrated in Figure 1:

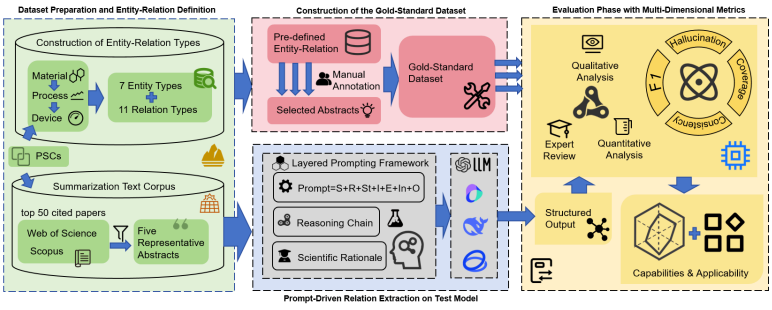

Figure 1. Workflow of entity–relation extraction and evaluation pipeline for typical domain summary texts based on LLM prompt engineering

Our workflow follows four stages.

DOI: 10.54254/2755-2721/2026.TJ31033

1) Data collection and curation. We first assembled a domain-specific corpus of high-quality papers, then shortlisted abstracts that are information-dense and exhibit complete, well-formed scientific reasoning (e.g., clear problem–method–result–interpretation structure). To reflect the way scientific relations are expressed in practice, we defined an entity–relation schema that covers common multi-entity settings, hierarchical relation patterns, and cross-sentence causal links. Using this schema, we created a manually annotated gold-standard set and enforced high annotation consistency through explicit guidelines and iterative adjudication.

2) Hierarchical prompt template design. We then designed a suite of prompts with increasing semantic demands. The sequence starts with constrained, structured entity extraction, and gradually extends to relation identification, implicit causality, and mechanism-oriented explanation. This progression allows us to probe how model behavior changes as the required reasoning depth increases.

3) Model application. The same prompt suite was applied to multiple candidate LLMs to obtain structured outputs under comparable conditions, enabling controlled cross-model comparison.

4) Comprehensive evaluation. Finally, we combined quantitative metrics with expert review to evaluate performance across the prompt hierarchy. By aligning evaluation difficulty with task complexity, we characterize where models remain reliable, where they degrade, and what this implies for practical deployment boundaries and suitable application scenarios.

3.2. Hierarchical prompt template design logic

In this study, a layered template system was employed in prompt engineering to systematically evaluate different LLMs' structured generation capabilities and reasoning differences in scientific relation extraction tasks. Unlike previous approaches that relied on a single prompt format, we adopted a unified experimental design combining multiple templates (M_j , T_i) to analyze the role of prompt structure in semantic control, causal explicitness, and knowledge explanation. To ensure controlled and comparable evaluation, all models were executed on the same corpus DDD, with the relation extraction mapping function defined as:

$$\mathrm{f} : (\mathrm{M_j}, \mathrm{T_i}, \mathrm{D}) \rightarrow \mathrm{R_{ij}} = \ \{\mathrm{r_{ij}(text)} \ , \ \mathrm{r_{ij}(json)}\} \tag{1}$$

where $r_{ij}(text)$ represents the model-generated natural language reasoning or scientific explanation, and $r_{ij}(json)$ represents the standardized entity–relation structured output.

(1) Template 1: Structured Instruction-based Template

Template 1 constrains model behavior with a strict instruction structure—"Role–Task–Example–Format"—which has been shown to provide good task transferability and output stability in multiple information extraction studies (e.g., REBEL, InstructIE, SciIE; Wang et al. [19], Gui et al. [20], Zhang et al. [21]). Formally, the prompt structure is expressed as:

$$\mathrm{Prompt} = \mathrm{S} + \mathrm{R} + \mathrm{St} + \mathrm{I} + \mathrm{E} + \mathrm{In} + \mathrm{O} \tag{2}$$

The first three components ensure model semantic self-localization and context understanding, the middle two reinforce task boundaries and example alignment, and the last two control execution and output consistency. This template explicitly defines task boundaries and output standards, strengthens execution consistency, reduces hallucination tendencies, and is suitable for standard entity recognition and basic relation extraction (Standard IE) as well as fact-level tasks with "low reasoning load."

DOI: 10.54254/2755-2721/2026.TJ31033

(2) Template 2: Reasoning-enhanced Template

To address Template 1's limitations in cross-sentence, multi-hop, or implicit relation recognition, Template 2 introduces a systematic reasoning chain, requiring models to explicitly output reasoning steps to capture deeper causal structures. This design draws on the Chain-of-Thought (CoT) reasoning paradigm, guiding the extraction process via step-by-step reasoning:

$$\mathrm{E}_1 \xrightarrow{\text{process}} \mathrm{E}_2 \xrightarrow{\text{mechanism}} \mathrm{E}_3 \tag{3}$$

This template requires the model to distinguish between explicit relations (surface-level), implicit relations (implicit-level), and multi-step causal relations (deep causal-level), generating both reasoning text and standardized JSON objects. Compared to Template 1, Template 2 primarily enhances semantic integration and causal explicitness, suitable for "medium reasoning load" complex relation extraction tasks.

(3) Template 3: Domain-grounded Causal Template

In scientific relation extraction tasks, many relations depend not only on linguistic causal reasoning but also on domain knowledge explanation and argumentation. Template 3 combines the reasoning chain with domain knowledge, guiding the model to provide scientific rationale for each extracted relation, thereby improving consistency and validity in complex scientific texts and enabling causal interpretability and knowledge consistency:

$$\mathrm{r_{ij}} \;=\; \langle (\mathrm{E_a}\,,\ \mathrm{E_b})\ ,\ \text{relation type}\ ,\ \text{rationale} \rangle \tag{4}$$

This template effectively enhances the model's mechanistic judgment and understanding of specialized terminology, emphasizing the transition from relation identification to knowledge explanation, and is suitable for mechanistic, theoretical, or multi-variable interaction tasks.

(4) Overall Advantages of the Hierarchical Design

The three templates form a layered system from structured constraints to scientific reasoning, corresponding to different levels of model cognitive demand. Formally, they can be abstracted as:

$$\mathrm{T}_1 \subset\ \mathrm{T}_2\ \subset\ \mathrm{T}_3 \tag{5}$$

where T_1 emphasizes format consistency and task boundaries, T_2 emphasizes logical coherence and causal explicitness, and T_3 emphasizes scientific interpretability and knowledge reasoning ability.

This hierarchical prompt system provides a unified, extensible experimental framework for comparing models' performance across basic extraction, reasoning-enhanced, and knowledge-explanation capabilities.

4. Evaluation

4.1. Requirements

To ensure that different large language models (LLMs) can be systematically, comparably, and reproducibly evaluated in structured scientific information extraction, this study defines overall requirements along three dimensions: task complexity, methodological consistency, and evaluation fairness. First, scientific relation extraction inherently exhibits hierarchical characteristics. It encompasses basic tasks such as entity recognition and explicit relation extraction, as well as higher-

DOI: 10.54254/2755-2721/2026.TJ31033

order tasks like cross-sentence reasoning, multi-hop causal chains, and mechanistic explanations. Therefore, the evaluation framework must differentiate models' performance across three capabilities, i.e., basic recognition, semantic reasoning and scientific explanation, to fully reveal their adaptability at different cognitive levels. Second, to avoid performance biases caused by corpus variations, prompt expressions, or output formats, this study maintains a unified corpus, consistent entity and relation definitions, and standardized output specifications across all model–template combinations. This ensures that observed performance differences primarily reflect the models' intrinsic capabilities and the guiding effects of prompt structures on reasoning behavior, rather than experimental artifacts. Finally, the three-layer prompt templates naturally form a capability gradient. The structured-instruction template assesses basic extraction ability, the reasoning-enhanced template identifies differences in complex semantic modeling, and the domain-grounded explanation template distinguishes mechanistic understanding and causal reasoning. Because this gradient derives from the semantic hierarchy of the tasks themselves, rather than artificially imposed difficulty, it enables fair, interpretable, and methodologically meaningful comparisons under consistent conditions.

4.2. Experimental setup

4.2.1. Selection of paper abstracts

To examine how well LLMs handle relation extraction in scientific writing, we use scientific literature as our evaluation setting. Scientific abstracts are typically organized around a recognizable rhetorical structure (problem, method, results, interpretation) and express information at multiple semantic levels. This makes them suitable for testing not only entity recognition, but also relation inference and, when applicable, mechanism-oriented explanation.

We focus on materials science because many papers in this field follow a familiar storyline: what a material is made of shapes how it can be processed, and those processing choices ultimately show up in the measured performance. This cause-and-effect flow aligns well with hierarchical relation extraction. Some links are spelled out plainly in the text, while others sit one or two steps away and have to be inferred from context and scientific logic.

Within materials science, perovskite solar cells (PSCs) offer a particularly useful testbed. Abstracts in PSC research often combine straightforward, single-step statements—such as a specific material choice paired with a reported efficiency—with more layered chains like "material modification → process control → device performance." They also tend to use standardized experimental phrasing while weaving in mechanistic claims about why an improvement happens, not just that it happens. That combination makes PSC abstracts well suited for probing whether a model can accurately extract what is stated directly, connect supporting evidence that is scattered across sentences, and produce a credible explanation when the text hints at an underlying mechanism rather than spelling it out.

Overall, this selection strategy yields a compact but representative abstract set that supports evaluation across multiple relation types and reasoning depths.

4.2.2. Definition of entity and relation types

After fixing the evaluation domain, we define an entity–relation schema tailored to PSC abstracts so that relation extraction can be formulated as a structured task. We specify seven top-level entity categories: Material, Performance, Method, Property, Device Structure, Scaling Issue, and

DOI: 10.54254/2755-2721/2026.TJ31033

Environmental Impact. To capture the experimental pipeline with sufficient granularity, Material is further partitioned into perovskite absorber, ETL/HTL, electrode, additive, interface modifier, and substrate, covering the main components and interventions that appear from input design to device assembly. For Method, we distinguish fabrication, processing, engineering, characterization, and simulation, which correspond to the procedures most frequently reported in PSC studies.

This schema is designed to follow the information flow typically conveyed in scientific writing: from material composition and experimental operations to mechanisms and measured outcomes. Concretely, entities such as materials, additives, and solvents represent the objects and input conditions under study; fabrication steps and instrumentation describe how experiments are carried out; and mechanism-related properties and performance metrics connect directly to the causal and explanatory relations we aim to extract.

Based on these entity definitions, we introduce eleven relation labels: ACHIEVES, IMPROVES, USED_IN, EXHIBITS, ENABLES, REDUCES, INCREASES, AFFECTS, CAUSES, COMPARED_WITH, and REQUIRES. Together, these labels cover both single-hop relations stated explicitly and multi-step chains that link materials, processes, intermediate mechanisms, and device performance. For instance, a material modification may ENABLE a processing adjustment; process control can IMPROVE a performance metric; and an additive may AFFECT an interfacial mechanism, which in turn influences device behavior..

4.2.3. Dataset source and abstract selection

To ensure that the evaluation framework covers the three core scientific information categories—Material, Process, and Performance—and tests model capabilities across different reasoning depths, this study systematically selected high-quality abstracts from highly cited papers in the perovskite solar cell (PSC) domain. The primary sources included Web of Science, Scopus, and frequently cited key works identified in domain review articles. During the initial screening, 50 paper abstracts were collected. After removing duplicates, overly brief abstracts, or those lacking structured information, five representative abstracts were ultimately selected based on task complexity, relation type coverage, and diversity of technical approaches, forming the final refined dataset.

Taken together, the five abstracts span several common research lines in PSC studies, including materials design, interfacial chemistry, transport-layer engineering, device-level simulation, and process-oriented comparison. We selected them to meet four criteria that are directly tied to our evaluation goals.

First, entity coverage is broad enough to support structured extraction: each abstract contains the core categories needed for our schema (e.g., MATERIAL, METHOD, PROPERTY, PERFORMANCE, and DEVICE_STRUCTURE). Second, the texts exhibit relation diversity, with frequent instances of relations such as IMPROVES, REDUCES, CAUSES, AFFECTS, and ENABLES, along with related mechanistic links. Third, they provide a range of reasoning depth, from relations stated explicitly in a single sentence to cross-sentence dependencies and multi-hop causal chains. Finally, each abstract presents a traceable scientific narrative, typically following a material–process–performance pathway, which allows us to evaluate extraction outputs against interpretable domain semantics rather than purely surface matching.

The five abstracts complement each other and present progressive complexity: basic abstracts emphasize clear entity–relation structures; medium-complexity abstracts introduce cross-sentence inferences and multi-factor interactions; high-complexity abstracts incorporate mechanistic explanations and implicit causal chains. This rigorously structured, progressively challenging, and

DOI: 10.54254/2755-2721/2026.TJ31033

semantically dense corpus allows for comprehensive assessment of model basic extraction, logical reasoning, and scientific explanation abilities using a minimal dataset.

4.2.4. Large language models and versions

To systematically evaluate the adaptability of LLMs in relation extraction and mechanistic explanation tasks within materials science, three representative domestic LLMs were selected: Demobao, deepseek-V2.3, and chatGLM-4.6. These models represent different development directions of domestic AI vendors and differ in English text comprehension and structured JSON output capabilities. Using consistent prompt engineering across models, their performance in entity–relation extraction was compared to reveal the strengths and limitations of domestic LLMs in scientific text processing. Default model parameters were used during experiments, with no fine-tuning applied.

Table 1. Representative domestic large language models and technical features

Model & Version	Technical Architecture	Development Direction	Focus	Expected Performance Advantages
Demobao (DEMOBAO)	Dense Transformer	Enterprise Ecosystem-Oriented	Strong social media interaction, content moderation, and user engagement; highly adaptive	Excels in JSON format compliance, field stability, and consistency of structured output
deepseek-V2.3	Mixture of Experts (MoE)	Open-Source-Oriented	Focuses on open-source ecosystem, offers strong customization, adapts to different domain needs	Strong in professional English terminology understanding, cross-sentence integration, and standardized scientific expression
chatGLM-4.6	Cross-Domain Alignment Strategy	Core Technology-Oriented	Expert at structured tasks such as document parsing, dialogue modeling, and reasoning	Excels in causal modeling depth, chain-of-thought reasoning, and mechanistic explanation generation

4.2.5. Data processing

To ensure reproducibility and comparability of model performance, the five selected PSC abstracts were uniformly processed as follows: Core content extraction: redundant background and repeated information were removed to retain key statements regarding materials, processes, performance, and their causal relationships. Manual annotation: abstracts were annotated based on the pre-defined seven entity types and eleven relation types to create a Gold Standard dataset. Annotation strictly followed uniform guidelines, covering single-hop causality, multi-hop complex causal chains, and mechanistic explanations. JSON conversion: annotated data were converted into a standardized JSON format, recording each entity's category, text position, and corresponding relation type, while preserving original text indices for consistent model input and output comparison. Cross-checking: all JSON data were reviewed by two domain experts to minimize subjective bias, ensure accurate entity–relation alignment, maintain logical consistency, and eliminate annotation omissions or formatting errors.

4.3. Metrics

To comprehensively assess the relation extraction and mechanistic explanation capabilities of large language models in the materials science domain, this study designed task-specific evaluation metrics for each template level, ensuring that the core abilities of each task type are systematically and quantitatively measured. Template 1 focuses primarily on basic entity and relation extraction tasks. The following technical metrics were selected: Entity F1, Loose Entity F1, Relation F1

DOI: 10.54254/2755-2721/2026.TJ31033

(Strict), Relation F1 (Relaxed), Entity Type-level Accuracy, Relation Type-level Accuracy, Relation Count Bias, and Redundancy Ratio.

For Template 2, which extends to multi-hop causal chains and reasoning processes, additional metrics were introduced beyond the basic extraction measures: hallucination rate, coverage, reasoning consistency, and expert qualitative scoring. Expert scoring evaluates the completeness of generated causal chains, the logical rigor, and alignment with domain knowledge, using a 0–5 scale (0 indicates completely incorrect or missing reasoning, 5 indicates fully logical and highly consistent with domain mechanisms).

Template 3 requires the model to provide scientifically valid mechanistic explanations and accurately handle domain-specific terminology. Additional metrics for this template include scientific plausibility, mechanistic understanding accuracy, causal judgment capability, domain terminology handling, reasoning depth score, and expert qualitative scoring. Expert scoring assesses whether the mechanistic explanations align with accepted domain theories, the accuracy of terminology usage, and the clarity and rationality of reasoning logic, using the same 0–5 scale (0 indicates erroneous or incoherent explanations, 5 indicates scientifically valid and logically complete explanations).

4.4. Results

4.4.1. Template 1: basic entity and relation extraction

Table 2. Average metrics of the three models under template 1

Model	Entity F1	LooseEntity F1	RelationF1-Strict	RelationF1-Relaxed	EntityType-levelAccuracy	RelationType-levelAccuracy	RelationCount Bias	RedundancyRatio
Demobao	0.81	0.91	0.73	0.86	0.92	0.88	0.62	0.15
DeepSeek	0.81	0.92	0.76	0.88	0.91	0.87	0.53	0.13
chatGLM-4.6	0.82	0.93	0.72	0.85	0.90	0.85	0.56	0.16

Table 2 presents the average performance metrics of the three models under Template 1. In entity recognition, all three models perform comparably. Entity F1 scores range from 0.81–0.82, with chatGLM-4.6 slightly higher (0.82), while Demobao and DeepSeek are both 0.81. The ranking for Loose Entity F1 is consistent, with chatGLM-4.6 (0.93) exceeding DeepSeek (0.92) and Demobao (0.91), indicating that under relaxed matching conditions, chatGLM-4.6's entity recognition aligns more closely with the reference standard. Entity type-level accuracy differences are small, ranging from 0.90 to 0.92.

The differences between models become much easier to see once the task shifts from extracting entities to extracting relations. DeepSeek stands out as the most reliable overall, with the best scores on both strict F1 (0.76) and relaxed F1 (0.88). Demobao comes next, and chatGLM-4.6 trails on these two metrics. Relaxed F1 is higher than strict F1 across the board, which is what you would expect, but the ranking stays the same. That consistency points to a broader distinction: the models are separated less by any single "preferred" relation pattern and more by how well they cope with the many ways scientific text can phrase the same underlying link. In comparison, relation-type accuracy is fairly steady across the three models (0.85–0.88), with only small gaps.

The clearest divergence shows up when looking at quantity-related diagnostics. Demobao has the strongest tendency to generate extra relations, reflected in the highest relation-count bias (0.62) and a relatively high redundancy ratio (0.15). DeepSeek is more restrained and better calibrated on both

DOI: 10.54254/2755-2721/2026.TJ31033

measures (bias 0.53; redundancy 0.13). chatGLM-4.6 sits in the middle on bias (0.56) but shows the highest redundancy ratio (0.16). Put together, the picture is less about raw capability and more about calibration: each model strikes a different balance between generating aggressively to maximize recall and keeping outputs tight enough to avoid repetitive or unnecessary relations.

Overall, differences among the three models in entity recognition are minor, whereas variations in relation extraction quality, relation count bias, and redundancy are more pronounced. The strict and relaxed relation F1 rankings are consistent, but quantity-related metrics provide greater differentiation.

4.4.2. Template 2: multi-hop causal chain reasoning

Table 3. Average metrics of the three models under template 2

Model	Hallucination Rate	Coverage	Reasoning Consistency	Qualitative Score	Entity F1	Relation F1	Type-level Accuracy	Relation Count Bias
Demobao	0.11	0.85	0.86	4.6	0.85	0.82	0.90	0.52
DeepSeek	0.10	0.86	0.85	4.4	0.87	0.80	0.91	0.45
chatGLM-4.6	0.09	0.82	0.84	4.4	0.84	0.79	0.88	0.50

Table 3 shows that in Template 2's multi-hop causal chain reasoning task, all three models maintain relatively high levels of entity recognition, relation generation, and logical consistency, yet distinct patterns emerge across metrics. For Entity F1, DeepSeek achieves the highest score (0.87), with Demobao and chatGLM-4.6 slightly lower (0.85 and 0.84), indicating that DeepSeek maintains higher entity recognition accuracy in complex texts. Regarding Relation F1, Demobao (0.82) slightly surpasses DeepSeek and chatGLM-4.6 (0.80 and 0.79), reflecting stronger coverage of causal relation entries. Type-level accuracy differences are minimal (0.90, 0.91, 0.88), demonstrating stable performance in entity and relation type identification.

For reasoning chain quality, reasoning consistency is highest for Demobao (0.86), followed by DeepSeek (0.85) and chatGLM-4.6 (0.84). Coverage is similar for DeepSeek (0.86) and Demobao (0.85), while chatGLM-4.6 is slightly lower (0.82). All three models maintain low hallucination rates, with chatGLM-4.6 being the lowest (0.09), DeepSeek at 0.10, and Demobao at 0.11. Qualitative scores (4.4–4.6) indicate that all models produce structurally clear reasoning texts, with minor differences mainly in local detail completeness and logical coherence. For type-level accuracy, the models are close (0.88–0.91), while relation count bias shows moderate differences: DeepSeek is the lowest (0.45), with chatGLM-4.6 and Demobao at 0.50 and 0.52, respectively. In summary, all three models can effectively perform multi-hop causal chain reasoning under Template 2, but distinct patterns are observed: DeepSeek excels in entity recognition. Demobao shows advantages in causal relation coverage and reasoning consistency but has a slightly higher hallucination rate. chatGLM-4.6 demonstrates the best hallucination control, though coverage and relation F1 are slightly conservative.

DOI: 10.54254/2755-2721/2026.TJ31033

4.4.3. Template 3: mechanism explanation and technical terminology handling

Table 4. Average metrics of the three models under template 3

Model	Hallucination Rate	Coverage	Reasoning Consistency	Qualitative Score	Entity F1	Relation F1	Type-level Accuracy	Relation Count Bias	Scientific Plausibility	Mechanism Understanding Accuracy	Causal Judgment	Terminology Handling	Reasoning Depth Score
Demobao	0.26	0.40	0.74	3.2	0.48	0.28	0.72	-0.16	High	High	Medium-High	High	1.12
DeepSeek	0.22	0.45	0.68	3.8	0.54	0.39	0.74	-0.11	High	High	Medium-High	High	1.14
chatGLM-4.6	0.27	0.56	0.67	4.0	0.62	0.47	0.71	0.24	High	High	High	High	1.16

Table 4 shows that in Template 3's scientific mechanism explanation task, the three models exhibit clear differences in coverage, relation F1, reasoning depth, and technical terminology handling. Coverage: chatGLM-4.6 achieves the highest (0.56), followed by DeepSeek (0.45) and Demobao (0.40), indicating that chatGLM-4.6 retains more information points when generating mechanism chains. Relation F1: chatGLM-4.6 leads (0.47), with DeepSeek (0.39) and Demobao (0.28) lower, showing its superior completeness and fine-grained handling in mechanistic relation extraction.For entity and relation type identification, type-level accuracy is similar across models (0.71–0.74), but relation count bias indicates directional differences: chatGLM-4.6 tends to generate more relations (0.24), whereas Demobao and DeepSeek are relatively conservative (negative bias). Regarding hallucination, DeepSeek is lowest (0.22), while Demobao and chatGLM-4.6 are slightly higher (0.26 and 0.27), reflecting differences in stability under high-complexity tasks. Reasoning capabilities show another pattern: Demobao has the highest reasoning consistency (0.74) but the lowest reasoning depth score (1.12), whereas chatGLM-4.6 has slightly lower consistency (0.67) but the highest reasoning depth (1.16). DeepSeek is intermediate in both metrics (0.68 and 1.14). This indicates that the models emphasize different aspects—maintaining logical coherence versus expanding deep reasoning. In qualitative dimensions, all three models are rated "High" for scientific plausibility and terminology handling, but slight differences exist in mechanism understanding accuracy and causal judgment, with chatGLM-4.6 slightly outperforming the others overall. In summary, for Template 3's complex scientific mechanism explanation task, all three models can generate basic mechanism outputs but show distinct observable strengths: chatGLM-4.6 excels in coverage, relation completeness, and reasoning depth. DeepSeek scores higher in hallucination control and type-level stability. Demobao has an advantage in reasoning consistency but is relatively weaker in relation extraction.

Overall, as templates progress from structured constraints to causal reasoning and finally to scientific mechanism explanation, models maintain basic recognition accuracy while facing increasing challenges in relation completeness, logical consistency, and information coverage, reflecting the hierarchical impact of template complexity on task difficulty.

4.5. Overall performance analysis

Across the three prompt templates, we observe a consistent pattern: performance separates as the required reasoning depth increases and the output format becomes less constrained. Under Template 1, the tasks are relatively direct and tightly specified. Models mainly need to recover entities and relations that are stated explicitly, and the structured format limits variation in wording. Not

DOI: 10.54254/2755-2721/2026.TJ31033

surprisingly, this setting yields the strongest overall scores, with outputs that are comparatively standardized and internally consistent.

Template 2 adds chain-of-thought (CoT) prompting and asks models to derive intermediate conclusions from textual evidence. In practice, most intermediate steps still depend on local context within or near the evidence spans used during manual annotation, so Template 2 aligns reasonably well with intra- and inter-sentence entity–relation extraction. This alignment often increases the coverage of annotated types and improves F1 for explicit extraction. At the same time, the multi-step format exposes weaknesses in reasoning control: models may skip necessary links, repeat steps, or produce causal sequences that do not close coherently. In other words, a model can arrive at a correct extracted relation while the stated reasoning path is incomplete or internally inconsistent—an effect we summarize as correct outputs with unreliable reasoning traces.

Under Template 3, the task shifts further toward mechanism-level scientific explanation. Outputs move from concise structured extraction to longer, semi-open narratives, which increases both generation flexibility and the space for interpretation. This change encourages models to supply missing links by drawing on background knowledge rather than staying strictly within the source text. A common consequence is that multiple relations are merged into a single broad statement, reducing relation granularity and weakening direct alignment with the gold annotations. Accordingly, Template 3 does not reliably increase annotated-type coverage; instead, narrative answers are more likely to omit entities, blur relation semantics, drift in terminology mapping, and accumulate small errors across the explanation. These issues also lead to formatting deviations and, overall, lower quantitative scores.

5. Conclusion

In this work, we evaluate large language models (LLMs) for relation extraction in scientific texts and introduce a hierarchical framework that spans three increasingly demanding settings: entity recognition, cross-sentence relation reasoning, and mechanism-oriented explanation. The framework is designed to support unified and comparable testing, with procedures that can be reproduced across models and prompt variants.

Overall, the tested models perform reliably on explicit entity identification and straightforward relations stated directly in the text. The gap widens, however, as tasks require deeper integration of evidence—especially when relations are expressed across sentences, when causal links are implicit, or when the text calls for mechanistic reasoning rather than surface extraction. These failures are not uniform: models differ in how well they manage domain terminology, how coherent their explanations remain, and how consistently they adhere to a structured output format. Still, one pattern is stable across models: pushing for broader coverage tends to increase redundant relations and raises the risk of hallucinated links, and multi-hop causal chains remain difficult to reconstruct completely in some abstracts.

Future work will proceed in two directions. First, we will strengthen context and incorporate external retrieval to supply background knowledge that is often required for multi-step scientific reasoning. The goal is to improve multi-hop linkage, mechanism identification, and scientific consistency while reducing hallucination. Second, we will move toward automated prompt optimization—through template search and prompt rewriting—to increase the stability of structured outputs and improve generalization across tasks and domains.

DOI: 10.54254/2755-2721/2026.TJ31033

References

[1] Kanimozhi U, Manjula D. A kernel-based SVM for semantic relations extraction from biomedical literature [J]. International Journal of Advanced Intelligence Paradigms, 2023, 24(3-4): 341-354.

[2] Zhang M, Qian T, Liu B. Exploit feature and relation hierarchy for relation extraction [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 917-930.

[3] Du K, Yang B, Wang S, et al. Relation extraction for manufacturing knowledge graphs based on feature fusion of attention mechanism and graph convolution network [J]. Knowledge-Based Systems, 2022, 255: 109703.

[4] Zhao Y, Yuan X, Yuan Y, et al. Relation extraction: advancements through deep learning and entity-related features [J]. Social Network Analysis and Mining, 2023, 13(1): 92.

[5] Hogan W. An overview of distant supervision for relation extraction with a focus on denoising and pre-training methods [J]. arXiv preprint arXiv: 2207.08286, 2022.

[6] Li Y, Wang H, Zhang D. An Entity-Relation Extraction Method Based on the Mixture-of-Experts Model and Dependency Parsing [J]. Applied Sciences, 2025, 15(4): 2119.

[7] Li Y. The combination of CNN, RNN, and DNN for relation extraction [C]//2021 2nd International Conference on Computing and Data Science (CDS). IEEE, 2021: 585-590.

[8] Christou D, Tsoumakas G. Improving distantly-supervised relation extraction through bert-based label and instance embeddings [J]. IEEE Access, 2021, 9: 62574-62582.

[9] Su P, Vijay-Shanker K. Investigation of improving the pre-training and fine-tuning of BERT model for biomedical relation extraction [J]. BMC bioinformatics, 2022, 23(1): 120.

[10] Tian Y, Chen G, Song Y, et al. Dependency-driven relation extraction with attentive graph convolutional networks [C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 4458-4471.

[11] Dagdelen J, Dunn A, Lee S, et al. Structured information extraction from scientific text with large language models [J]. Nature communications, 2024, 15(1): 1418.

[12] Wang X, Zhou W, Zu C, et al. Instructuie: Multi-task instruction tuning for unified information extraction [J]. arXiv preprint arXiv: 2304.08085, 2023.

[13] Haoran X, Lou K, Chan S. Improved Chinese Few-Shot relation extraction using Large Language Model for Data Augmentation and Prototypical Network [C]//2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2024: 3787-3794.

[14] Xu X, Zhu Y, Wang X, et al. How to unleash the power of large language models for few-shot relation extraction? [J]. arXiv preprint arXiv: 2305.01555, 2023.

[15] Huang F, Huang Q, Zhao Y T, et al. A three-stage framework for event-event relation extraction with large language model [C]//International Conference on Neural Information Processing. Singapore: Springer Nature Singapore, 2023: 434-446.

[16] Ling C, Zhao X, Zhang X, et al. Improving open information extraction with large language models: A study on demonstration uncertainty [J]. arXiv preprint arXiv: 2309.03433, 2023.

[17] Yu L, Chen D, Xiong S, et al. Causaleval: Towards better causal reasoning in language models [C]//Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2025: 12512-12540.

[18] Dunn A, Dagdelen J, Walker N, et al. Structured information extraction from complex scientific text with fine-tuned large language models [J]. arXiv preprint arXiv: 2212.05238, 2022.

[19] Wang Q, Li C, Liu Y, et al. An Adaptive Framework Embedded with LLM for Knowledge Graph Construction [J]. IEEE Transactions on Multimedia, 2025.

[20] Gui H, Qiao S, Zhang J, et al. Instructie: A bilingual instruction-based information extraction dataset [C]//International Semantic Web Conference. Cham: Springer Nature Switzerland, 2024: 59-79.

[21] Zhang Q, Chen Z, Pan H, et al. SciER: An entity and relation extraction dataset for datasets, methods, and tasks in scientific documents [J]. arXiv preprint arXiv: 2410.21155, 2024.

[22] Salvatore N, Wang H, Zhang Q. Lost in the Middle: An Emergent Property from Information Retrieval Demands in LLMs [J]. arXiv preprint arXiv: 2510.10276, 2025.