ROEC: RECORD-ORIENTED EVIDENCE CONSTRUCTION FOR DOCUMENT-LEVEL STRUCTURED INFORMATION EXTRACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Targeted structured information extraction organizes information according to a predefined schema. Extending this setting to the document level is challenging because target information is often sparse and distributed across the document, while non-target content can interfere with extraction. Beyond recovering relevant information, extracted outputs must satisfy predefined requirements and remain supported by the source text. We propose Record-Oriented Evidence Construction (ROEC), which replaces one-step record extraction with Two-Stage Structured Information Extraction. ROEC first extracts core fields and then completes the associated contextual fields, while retaining supporting source sentences throughout extraction. It further incorporates Record-Level Refinement to improve record granularity and conformity with task requirements and reduce redundancy, and Claim Verification and Support Scoring to evaluate extracted claims against retained supporting source sentences and derive a score reflecting the degree of source-text support. Experiments on agricultural research papers show that ROEC improves Relaxed and Strict F1 by 11.83 and 5.86 points over the strongest evaluated baseline under each criterion. The two-stage configuration improves recall and F1 over one-step extraction, while Record-Level Refinement further reduces false positives and increases true positives. Human assessment indicates that the support score reflects differences in source-text support and helps prioritize outputs for manual review of source-support errors.

1 INTRODUCTION

Scientific literature reports findings that support evidence synthesis and comparative analysis, yet much of this information remains embedded in free-form text (Pronesti et al., 2025). In agriculture, such findings describe how interventions or factors affect outcomes across crops, locations, and production conditions. Document-level targeted structured information extraction organizes this information into records according to a predefined schema, as illustrated in Figure 1(a).

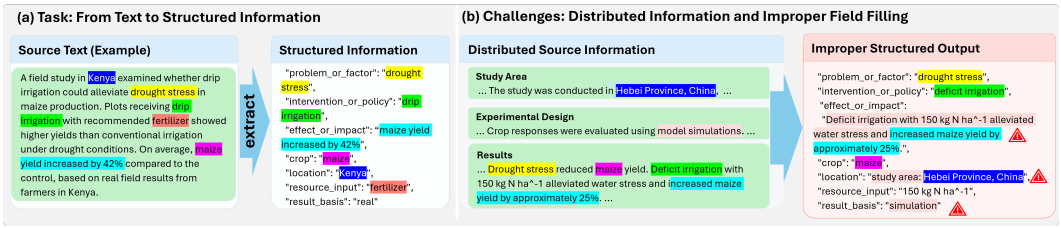


Figure 1: Illustration of the document-level structured information extraction task and its main challenges. (a) Extraction from source text to structured information. (b) Distributed source information and improper field filling.

This setting presents challenges beyond identifying relevant text. Information required for the same structured output may be distributed across distant sections of a document, while different pieces

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

of information must still be correctly associated (Khalighinejad et al., 2024; Zheng et al., 2019). Moreover, extracted content may fail to conform to predefined field requirements and extraction guidelines (Sainz et al., 2024). Locating relevant information therefore does not by itself ensure correct field filling. Figure 1(b) illustrates these two issues through distributed source information and improper field filling. Long-context models remain sensitive to the position of relevant information and to input length (Liu et al., 2024; Du et al., 2025). For extraction using large language models (LLMs), processing the full document, dividing it into chunks, or retrieving relevant passages are alternative ways to organize the input, but these choices alone do not guarantee correct information association or compliance with field requirements. Effective extraction therefore requires both recovering and associating distributed information and ensuring that the resulting structured outputs conform to predefined requirements.

We propose Record-Oriented Evidence Construction (ROEC), which organizes the extraction process around the core content of the target information. Two-Stage Structured Information Extraction facilitates the recovery and association of distributed information by first identifying the key information that defines this core content and then completing the remaining information around it. Record-Level Refinement improves conformity with predefined requirements by examining the extracted structured information for incorrect information combinations, invalid outputs, and redundancy. Claim Verification and Support Scoring addresses source-text support by decomposing the extracted information into verifiable claims and evaluating them against retained supporting source sentences, yielding a support score that reflects the degree of source-text support.

Experiments on agricultural research papers show that ROEC outperforms the evaluated baselines in overall structured information extraction, improving Relaxed and Strict F1 by 11.83 and 5.86 points over the strongest baseline under each criterion. Component analysis shows higher recall and F1 with Two-Stage Structured Information Extraction and further improvements after Record-Level Refinement. Human evaluation further shows a clear association between the support score and human assessments of the degree of source-text support.

2 Related Work

Scientific information extraction organizes entities, relations, and experimental findings from scholarly documents into structured representations (Dagdelen et al., 2024). SciREX addresses salient entity identification and document-level n-ary relation extraction (Jain et al., 2020), while SciER provides full-text annotations of entities and relations concerning datasets, methods, and tasks (Zhang et al., 2024). CARE represents clinical findings as structured relations with attributes and numerical outcomes, although its annotations focus on abstracts (Naik et al., 2024). Generative methods produce structured outputs from extraction instructions (Jiao et al., 2023), while GoLLIE emphasizes adherence to detailed annotation guidelines (Sainz et al., 2024). Recent LLM-based work extends scientific extraction to sample records and multimodal structured outputs from full-length articles (Khalighinejad et al., 2024; 2025).

Document-level methods use evidence-guided relation extraction (Xie et al., 2022; Ma et al., 2023) and template filling (Gantt et al., 2024) to recover structured information across sentences. SudokuFill progressively completes scientific records using high-confidence fields to constrain subsequent filling (Li et al., 2026). Beyond field extraction, work on event individuation highlights the difficulty of deciding which information belongs to the same event (Gantt et al., 2023). Factuality evaluation offers complementary tools: FActScore and RefChecker decompose outputs into fine-grained facts or claims for verification (Min et al., 2023; Hu et al., 2024), while AlignScore and MiniCheck assess factual consistency against reference text (Zha et al., 2023; Tang et al., 2024). ROEC brings these concerns together through core-relation-guided record construction, refinement of record granularity and validity, and claim-level support scoring against retained source sentences.

3 Task Definition

We study document-level targeted structured information extraction from agricultural research papers. Given a paper's full text and a predefined schema, the task is to extract zero or more Agricultural Evidence Records (AERs). Each AER represents an individual agricultural finding through a reported intervention–outcome or factor–outcome relation, termed its *core relation*, together with

associated contextual information. Relevant information may span sentences or sections, but all non-empty fields within an AER must describe the same finding. Generic statements that do not establish a specific agricultural finding are excluded, even when some individual fields satisfy the schema.

We consider two AER types: *Intervention–Effect*, describing outcomes associated with an intervention, management practice, technology, or policy; and *Factor–Impact*, describing outcomes associated with an agricultural problem, risk, constraint, or factor. A non-empty *intervention or policy* field determines the former type, and an empty field determines the latter. Both types may occur in the same document, and each AER must satisfy its type-specific required-field constraints. The schema distinguishes *core fields*, which express the core relation, from *contextual fields*, which specify crop, location, resource input, and result basis. Table 1 summarizes the fields and their applicability.

Table 1: AER schema. R: required; O: optional; –: not applicable. I–E: Intervention–Effect; F–I: Factor–Impact. *At least one of *crop* or *location* is required for either AER type.

Field	Information captured	I–E	F–I
Core fields			
problem or factor	Agricultural problem, risk, constraint, or condition	O	R
intervention or policy	Intervention, management practice, technology, or policy	R	–
effect or impact	Reported outcome associated with the intervention or factor	R	R
Contextual fields			
crop	Crop to which the finding applies	O*	O*
location	Study site or region to which the finding applies	O*	O*
resource input	Resource or input explicitly associated with the finding	O	O
result basis	`real` for outcomes that actually occurred; `not_real` for predicted or simulated outcomes	O	O

AER granularity. Each AER should be as atomic as possible. Independently interpretable findings, including separately reported outcomes for distinct interventions, factors, or conditions and independently reported quantitative effects, should be separated when their correspondence can be preserved. Each resulting AER must represent a coherent finding with its applicable contextual information. The same contextual information may be retained across multiple AERs when it is supported by the source text for each corresponding finding. Repeated mentions of the same finding should be consolidated, while findings differing in interventions, factors, conditions, or outcomes should remain distinct.

Field values and source support. Each field should contain the minimal source-text span needed to express its information while preserving necessary qualifiers such as negation, effect direction, quantities, units, and comparison conditions. Source support is required for both field values and their associations: core fields must express a supported relation, and contextual fields must apply to the same finding. Document-level co-occurrence alone is insufficient. Missing optional information remains empty, and AERs lacking required information are invalid; information must not be invented or rewritten to make a record appear complete. The sole exception to preserving source wording is *result basis*, whose `real` or `not_real` label may be assigned from clear evidence in the study design, data source, or methodological description; otherwise, it remains empty. These labels describe the outcome's basis, not whether the extraction is supported by the source text.

4 METHOD

Figure 2 illustrates Record-Oriented Evidence Construction (ROEC). Its central design is to construct AERs around their core relations through Two-Stage Structured Information Extraction while retaining supporting source sentences. Record-Level Refinement addresses AER granularity and validity, while Claim Verification and Support Scoring assesses the degree of source-text support.

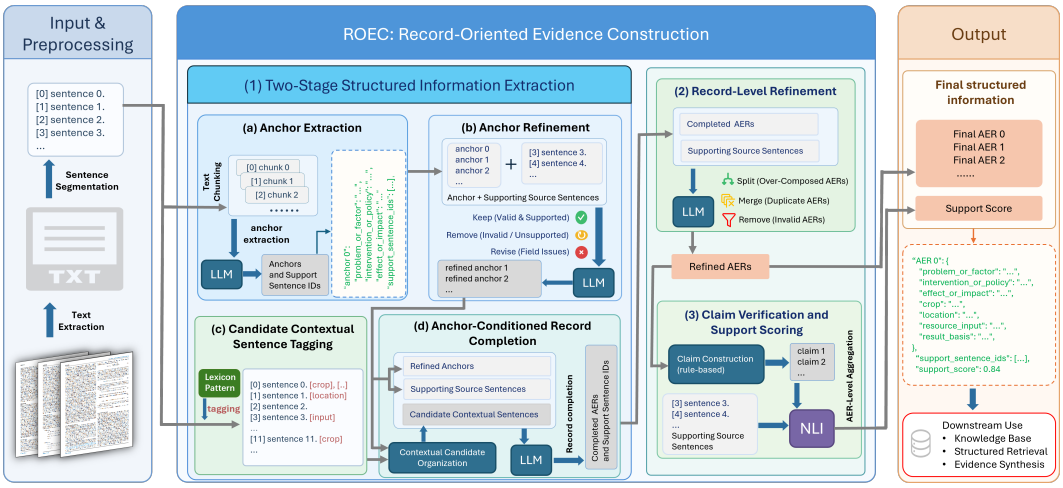


Figure 2: Overview of ROEC, showing AER construction, refinement, and support scoring with supporting source sentences retained throughout the framework.

4.1 Two-Stage Structured Information Extraction

Drawing on staged field filling (Li et al., 2026), ROEC first extracts core relations and then completes their contextual fields. An *anchor* comprises core-field values expressing a candidate AER's core relation, together with supporting source sentence identifiers. Figure 2(a–d) illustrates the extraction, refinement, context organization, and completion workflow.

In the first stage, the LLM extracts candidate anchors from source-text chunks (Figure 2(a)). Each anchor is then inspected with its cited supporting sentences to check required core-field presence, relation coherence, and source support. As shown in Figure 2(b), the LLM retains valid anchors, revises field values when supported corrections are available, and removes invalid or unsupported candidates, producing *refined anchors*.

The second stage organizes context around refined anchors, drawing on target-conditioned context organization (Zhang et al., 2025). Field-specific lexicons and sentence patterns tag candidate contextual sentences (Figure 2(c)). Anchors are grouped under a fixed input-size threshold, prioritizing those sharing supporting sentence identifiers without merging them. For each supporting sentence, ROEC selects tagged candidates within a three-sentence window on either side and retains intervening sentences. It also selects at most one candidate per contextual tag outside the local window, choosing the nearest to the group's supporting sentences regardless of section boundaries. Selected sentences are deduplicated and restored to document order.

As illustrated in Figure 2(d), the completion input combines refined anchors, their supporting sentences, and the selected context. The LLM determines which contextual information belongs to each anchor's finding and fills the corresponding fields while keeping core-field values unchanged. Field values follow Section 3, and optional fields remain empty when suitable values cannot be established. Sentences cited for filled contextual fields are retained alongside the anchor's supporting sentences. The resulting *completed AERs* and their supporting sentence identifiers are passed to Record-Level Refinement.

4.2 Record-Level Refinement

Record-Level Refinement uses an LLM to split combined findings, merge duplicates, and remove invalid AERs. Splitting and merging are guided by finding identity, drawing on work on event individuation (Gantt et al., 2023). ROEC selects potentially over-composed AERs using predefined core-field length thresholds and identifies potential duplicates within each document using lexical and semantic field similarity (Section 5.2). For each operation, the selected AERs are grouped with their retained supporting source sentences into input blocks, as illustrated in module (2) of Figure 2. ROEC controls block size to avoid overly large or small inputs. The LLM splits independently

reportable findings into separate AERs, merges records describing the same finding, and removes records judged to violate the required-field or source-support requirements in Section 3. Any revisions must preserve source-supported information and follow the task's granularity and minimal-text requirements. The resulting outputs are termed *refined AERs*.

4.3 Claim Verification and Support Scoring

As illustrated in module (3) of Figure 2, Claim Verification and Support Scoring takes refined AERs and their retained supporting source sentences as input and estimates source-text support through claim construction, NLI-based verification, and AER-level aggregation. Claim construction provides smaller units for verification without changing AER granularity.

Claim construction. Drawing on fine-grained factual verification through claim decomposition (Min et al., 2023; Hu et al., 2024), ROEC uses predefined rule-based templates to construct two groups of claims from each refined AER. Core claims express applicable relations among its core fields, whereas contextual claims associate each non-empty contextual field with the corresponding core information. For example, a crop-related contextual claim expresses that the AER's core relation applies to the specified crop. A claim is generated only when the fields required by its template are present.

Claim verification. ROEC formulates source-text support assessment as evidence-based natural language inference (NLI) (Wadden et al., 2020; Jullien et al., 2023). For a claim q and evidence text e, let $p_E(q, e)$ denote the entailment probability returned by the NLI model, with e as the premise and q as the hypothesis. Let $\mathcal{E}_q$ denote the relevant supporting source sentences retained for q. ROEC evaluates each sentence individually, then selects the two sentences with the highest individual entailment probabilities and evaluates their concatenation in document order. Let these two sentences be $e_{(1)}$ and $e_{(2)}$. The claim-level support score is

$$\mathrm{Sup}(q) = \max\left\{ \max_{e \in \mathcal{E}_q} p_E(q, e),\ p_E\big(q, e_{(1)} \oplus e_{(2)}\big) \right\}, \tag{1}$$

where $\oplus$ denotes concatenation in document order. When only one supporting source sentence is available, its individual entailment probability is used.

AER-level aggregation. Claim-level support scores are averaged separately for the core claims and contextual claims of each refined AER. Let $\overline{\mathrm{Sup}}_{\mathrm{core}}$ and $\overline{\mathrm{Sup}}_{\mathrm{contextual}}$ denote these two means. The AER-level support score is

$$\mathrm{Sup}_{\mathrm{AER}} = 0.7\,\overline{\mathrm{Sup}}_{\mathrm{core}} + 0.3\,\overline{\mathrm{Sup}}_{\mathrm{contextual}}. \tag{2}$$

The weighting places greater emphasis on the core claims. This score estimates source-text support against the retained supporting sentences; it does not directly measure compliance with every task requirement or agreement with a gold AER under the extraction evaluation protocol.

5 Experiments and Results

5.1 Experimental Setup

Data and annotation. The development set consists of 15 articles from SCI-indexed agricultural journals. For the test set, we applied rule-based retrieval to research articles in PubMed Central (PMC) to identify papers likely to contain target information, then randomly selected 30 articles from the retrieved candidates for manual annotation.

Table 2 summarizes both sets. Each document was independently annotated by two domain experts. A third expert consolidated their annotations and resolved discrepancies, followed by a joint review by all three experts to finalize the gold AERs. The development set was used for prompt design, field-specific lexicon construction, and configuration and threshold selection; the test set was reserved for final evaluation.

Table 2: Statistics of the development and test sets.

Statistic	Dev	Test
Documents	15	30
Gold AERs	257	421
Intervention–Effect AERs	105	316
Factor–Impact AERs	152	105
Mean AERs / doc.	17.1	14.0
Median AERs / doc.	12	13
AER range / doc.	2–62	0–43
Mean document tokens	14,971.8	8,603.3
Median document tokens	14,287	8,122
Document token range	6,559–25,163	3,847–18,152

Comparison methods. We compare ROEC with four methods using the same AER schema. **Full** extracts all fields jointly from the complete document. **Chunk** performs extraction independently on overlapping text chunks and aggregates the resulting AERs. **Retrieval** selects sentences using field-specific lexicons, restores their document order, and groups them into chunks for extraction. **LangExtract** (Goel, 2026) is an open-source structured information extraction framework adapted to the task definition and AER schema while retaining its original extraction procedure.

Model configuration. All reported results involving LLM-based methods, including ROEC and the baselines, are obtained using `qwen/qwen3-32b` as the LLM. NLI computations use `dleemiller/ModernCE-base-nli`, and embedding-based similarity uses `sentence-transformers/all-MiniLM-L6-v2`.

5.2 Evaluation Protocol

Drawing on matching-based evaluation for structured prediction (Chen et al., 2023), we evaluate AER extraction using partial matching to accommodate missing fields and surface-form variation in free-text values. For each predicted–gold AER pair, field-level similarity combines token-set F1 and embedding cosine similarity for core fields, while contextual fields use token-set F1:

$$\mathrm{Sim}_f = \begin{cases} 0.6F1_{\mathrm{tok}} + 0.4\,\mathrm{Sim}_{\mathrm{emb}}, & f \in \mathcal{F}_{\mathrm{core}}, \\ F1_{\mathrm{tok}}, & f \in \mathcal{F}_{\mathrm{contextual}}, \end{cases} \tag{3}$$

where the similarities are computed between the corresponding field values, and $\mathcal{F}_{\mathrm{core}}$ and $\mathcal{F}_{\mathrm{contextual}}$ are the field groups in Table 1. The AER-level similarity is

$$\mathrm{Sim}_{\mathrm{pred,gold}} = 0.7\,\overline{\mathrm{Sim}}_{\mathrm{core}} + 0.3\,\overline{\mathrm{Sim}}_{\mathrm{contextual}}, \tag{4}$$

where each mean excludes fields empty in both AERs; fields empty in only one AER receive a similarity of zero and remain in the mean. The same similarity calculation is used to select potentially redundant AERs in Record-Level Refinement.

Matching is restricted to AERs of the same type within the same document. Candidate pairs are scored by AER-level similarity and screened for explicit negation, direction, and numerical conflicts. For NLI-based conflict checking, claims are constructed from both AERs using the rule-based templates described in Section 4.3. Claims expressing corresponding relations are compared bidirectionally, with the predicted and gold claims alternately serving as the premise and hypothesis. The NLI outputs are used to screen for clear contradictions. At each evaluation threshold, pairs are eligible for matching only if they pass the conflict checks and meet the similarity threshold. Drawing on one-to-one template alignment (Gantt et al., 2024), eligible pairs are aligned using a globally optimized one-to-one bipartite matching. Matched pairs count as true positives (TP); unmatched predictions count as false positives (FP), and unmatched gold AERs count as false negatives (FN).

We report Relaxed and Strict results using similarity thresholds τ_{relaxed} and τ_{strict}, respectively, with $\tau_{\mathrm{strict}} > \tau_{\mathrm{relaxed}}$. Strict evaluation therefore uses a higher partial-matching threshold rather than requiring exact matching. TP, FP, and FN are pooled across the test set to compute micro-averaged precision, recall, and F1. These metrics measure agreement with gold AERs, whereas the AER-level support score estimates source-text support against the retained supporting sentences.

5.3 Main Results

Table 3 shows that ROEC achieves the highest Relaxed and Strict F1 scores among the evaluated methods, improving over the strongest baseline under each criterion by 11.83 and 5.86 points, respectively. ROEC also achieves the highest Relaxed recall, exceeding Retrieval by 11.64 points, while outperforming Chunk, Retrieval, and LangExtract in Relaxed precision. Compared with these three baselines, ROEC produces fewer AERs yet recovers more gold AERs, indicating that its recall gain is not simply due to generating more predictions. Full retains the highest precision but has substantially lower recall and F1. Together, these results demonstrate improved overall AER extraction performance under both matching criteria on the test set.

Table 3: Overall AER extraction performance on the test set. Pred denotes the number of predicted AERs. P, R, and F1 are reported on a 0–100 scale; the best score in each metric is bold.

Method	Pred	Relaxed P	Relaxed R	Relaxed F1	Strict F1
Full	201	**43.78**	20.90	28.30	12.86
Chunk	988	23.68	55.58	33.22	13.48
Retrieval	1055	22.94	57.48	32.79	16.12
LangExtract	1000	23.00	54.63	32.37	13.37
ROEC	871	33.41	**69.12**	**45.05**	**21.98**

5.4 Ablation Analysis of Two-Stage Extraction and Record-Level Refinement

We use Chunk as the *One-step* baseline because it shares the chunk-based extraction setting with ROEC. *Two-stage* replaces joint field extraction with Two-Stage Structured Information Extraction, including anchor refinement; *Full ROEC* additionally applies Record-Level Refinement.

Table 4 presents the progressive comparison from One-step extraction to Two-stage extraction and Full ROEC. One-step corresponds to the Chunk baseline. With nearly unchanged prediction counts, Two-stage recovers 44 additional gold AERs and raises Relaxed recall from 55.58 to 66.03, while also improving precision. Relaxed and Strict F1 increase by 6.16 and 4.37 points, respectively, supporting the effectiveness of the two-stage extraction configuration.

Adding Record-Level Refinement reduces false positives by 133 while increasing true positives by 13, yielding further gains of 5.67 points in Relaxed F1 and 4.13 points in Strict F1. These results indicate that refinement improves the balance between recovering gold AERs and limiting unmatched predictions under the evaluated configuration.

Table 4: Progressive ablation of Two-Stage Structured Information Extraction and Record-Level Refinement. Pred denotes the number of predicted AERs. TP, FP, FN, P, R, and F1 use Relaxed matching unless otherwise specified. P, R, and F1 are reported on a 0–100 scale.

Method	Pred	TP	FP	FN	P	R	F1	Strict F1
One-step	988	234	754	187	23.68	55.58	33.22	13.48
Two-stage	991	278	713	143	28.05	66.03	39.38	17.85
Full ROEC	871	291	580	130	33.41	69.12	45.05	21.98

Figure 3 illustrates the two successive comparisons. In (a), Two-stage recovers the crop and location omitted by One-step. In particular, the location is recovered from a different section of the paper and matches the gold annotation, illustrating contextual-field completion across section boundaries. In (b), Two-stage combines treatments M1–M3 and their effects into one AER, whereas Full ROEC produces three separate AERs preserving the correspondence between each treatment and its quantitative effect. This example illustrates how Record-Level Refinement can improve AER granularity.

5.5 Analysis of Source-Text Support Scores

We assess whether the support score reflects the degree of source-text support and helps prioritize AERs requiring manual review. Two annotators independently evaluate all 871 refined AERs using only the AERs and their corresponding supporting source sentences, without access to support

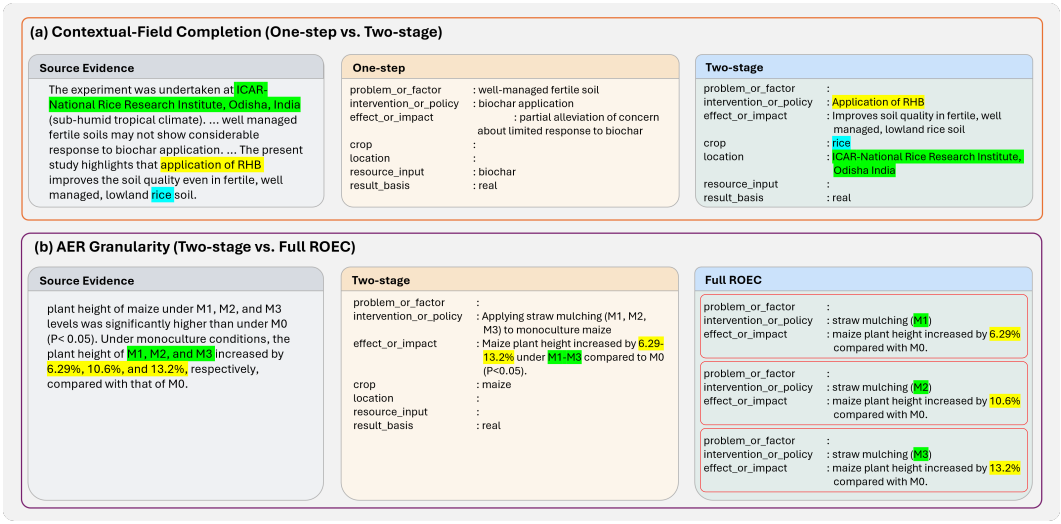


Figure 3: Illustrative examples of (a) contextual-field completion and (b) AER granularity. Highlighting links source text to corresponding AER values. Only the core fields are shown for Full ROEC in (b); the remaining fields are unchanged from the Two-Stage output.

scores. A third expert independently reviews the assessments, consolidates the final labels, and adjudicates disagreements or ambiguities. Each AER is labeled *Fully Supported* if its core relation and all non-empty contextual fields are supported, *Partially Supported* if its core relation is supported but at least one non-empty contextual field is not, or *Unsupported* if its core relation is not supported. The final assessments comprise 459 Fully Supported, 258 Partially Supported, and 154 Unsupported AERs.

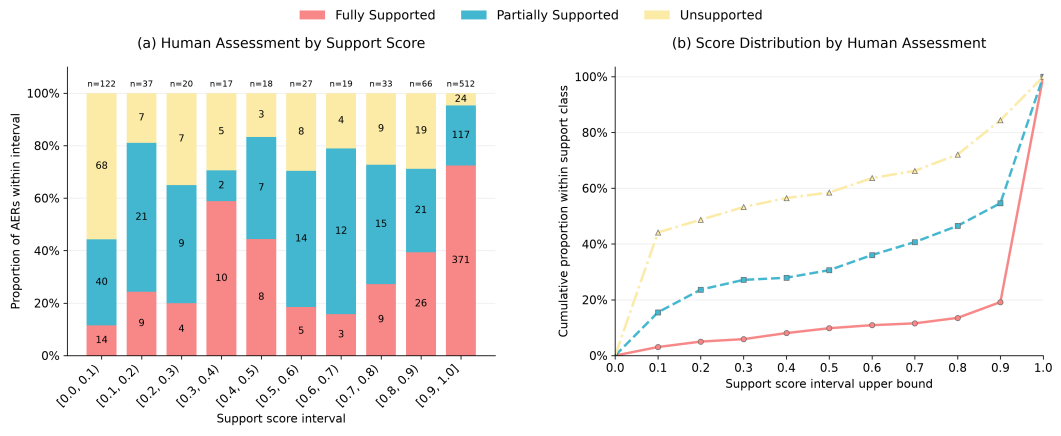


Figure 4: Human assessment of source-text support for 871 AERs. (a) Support-level proportions within score intervals; segment labels indicate counts and n denotes the interval total. (b) Cumulative score distributions normalized within each support level. Markers show proportions below each interval upper bound, with 1.0 included at the final boundary; connecting lines are visual guides.

Figure 4(a) shows that Fully Supported AERs account for 72.5% of predictions in $[0.9, 1.0]$, compared with 11.5% in $[0, 0.1)$. The class-wise cumulative distributions in Figure 4(b) further show that 80.8% of Fully Supported AERs have scores of at least 0.9, compared with 45.3% of Partially Supported and 15.6% of Unsupported AERs. These results indicate an association between support scores and human-assessed source-text support, although the proportions fluctuate across intermediate score intervals rather than changing monotonically.

To assess review prioritization, we treat Partially Supported and Unsupported AERs as requiring review for source-support errors. Reviewing the 179 AERs with scores below 0.3, representing 20.6% of all outputs, covers 152 of the 412 AERs requiring review (36.9%), approximately 1.79 times the expected coverage of random review at the same budget. However, 141 AERs requiring review remain in the highest-score interval. The support score therefore helps prioritize manual review but does not establish that high-score AERs can be accepted without inspection.

6 Discussion

Refinement improves extraction through both validity filtering and correction of record granularity, the latter being related to the problem of event individuation (Gantt et al., 2023). Our qualitative inspection suggests that splitting AERs containing multiple findings into independently interpretable records is a main contributor to the increase in TP after refinement. Nevertheless, substantial FP remain across all evaluated methods. We observed two recurring patterns: records in which some core fields satisfy the requirements while another is filled inappropriately, and broad conclusions that do not establish a specific target finding, such as general statements that climate change severely affects global agriculture. Such outputs may contain relevant, source-supported information without qualifying for inclusion in the gold annotations. Refinement can retain them when local plausibility and partial field conformity are mistaken for validity of the record as a whole. This exposes a difficulty beyond determining record granularity: the boundary between a qualifying finding and a broadly relevant statement is challenging to express and enforce consistently through prompts. During development, we revised prompts to clarify general task requirements rather than tailoring them to individual examples, but alternative formulations did not eliminate these errors. Refinement therefore makes meaningful progress toward task-conforming outputs while leaving record-level inclusion as an important challenge. Information extraction methods need to address this boundary alongside locating relevant content and assigning it to fields.

Support scoring provides a complementary contribution by assessing whether extracted information is supported by the retained source sentences. Our human assessment indicates that the scores capture meaningful differences in source support and help prioritize records for manual review. Their imperfect agreement with human judgments, however, limits their use as definitive assessments. Possible sources of disagreement include rule-based claims that do not fully capture a record's meaning and limitations of the NLI model in evaluating the resulting claim–evidence pairs. Consequently, a high entailment score for the constructed claims need not imply that a human reviewer would regard the entire AER as fully supported. This reinforces the distinction between evidence attribution and overall output quality (Buchmann et al., 2024). The contribution of support scoring is thus an additional, evidence-oriented signal for review, rather than a replacement for human judgment or task-validity checks. Future work should investigate claim construction that more faithfully preserves record meaning and verification methods that better assess domain-specific relations and evidence distributed across sentences.

7 Conclusion

We presented ROEC, a framework for document-level targeted structured information extraction that uses core relations to guide contextual-field completion, refines record granularity and validity, and assesses extracted claims against retained supporting source sentences. Experiments on agricultural research papers show that ROEC improves Relaxed and Strict F1 by 11.83 and 5.86 points over the strongest evaluated baseline under each criterion. Progressive ablation shows that the two-stage configuration improves recall and F1 with nearly unchanged prediction counts, while record-level refinement further reduces false positives and increases true positives. These results support the effectiveness of separating core-relation extraction from contextual completion and subsequently refining the resulting records. Human assessment further indicates that support scores reflect differences in source-text support and help prioritize manual review, although high scores do not establish that records can be accepted without inspection. Together, these findings support an approach that addresses information recovery, conformity with record-level requirements, and source-text support as complementary aspects of structured extraction. Future work will strengthen record-validity checks and contextual attribution, improve claim verification, and evaluate transfer across domains, schemas, and model configurations.

AI USE STATEMENT

Generative AI tools were used in this work to assist with writing and language polishing, draft portions of the manuscript, support literature retrieval and discovery, and assist with research ideation and execution, including code development and debugging. All AI-assisted content was reviewed and revised by the authors. Suggested literature and related-work information were checked against the original sources, and AI-assisted code and experimental procedures were inspected and tested before use. Research decisions, experimental design, interpretation of results, and the final claims presented in the paper were determined and verified by the authors. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We do not believe there are significant ethical issues associated with this research.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we describe the task formulation, main methodological components, key model and experimental settings, and evaluation protocol. Upon acceptance, we intend to release the implementation and relevant configurations, along with supporting materials that can be shared under applicable data-use restrictions.

REFERENCES

Jan Buchmann, Xiao Liu, and Iryna Gurevych. Attribute or abstain: Large language models as long document assistants. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8113–8140, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.463. URL https://aclanthology.org/2024.emnlp-main.463/.

Yunmo Chen, William Gantt, Tongfei Chen, Aaron Steven White, and Benjamin Van Durme. A unified view of evaluation metrics for structured prediction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12868–12882, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.795. URL https://aclanthology.org/2023.emnlp-main.795/.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. Structured information extraction from scientific text with large language models. *Nature Communications*, 15:1418, 2024. doi: 10.1038/s41467-024-45563-x.

Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Babu Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A Huerta, and Hao Peng. Context length alone hurts LLM performance despite perfect retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23281–23298. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1264.

William Gantt, Reno Kriz, Yunmo Chen, Siddharth Vashishtha, and Aaron Steven White. On event individuation for document-level information extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12938–12958. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.862. URL https://aclanthology.org/2023.findings-emnlp.862/.

William Gantt, Shabnam Behzad, Hannah An, Yunmo Chen, Aaron Steven White, Benjamin Van Durme, and Mahsa Yarmohammadi. MultiMUC: Multilingual template filling on MUC-4. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 349–368. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.eacl-long.21. URL https://aclanthology.org/2024.eacl-long.21/.

Akshay Goel. LangExtract. Open-source software, 2026. URL https://github.com/google/langextract.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. Knowledge-centric hallucination detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6953–6975. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.395.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. SciREX: A challenge dataset for document-level information extraction. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7506–7516, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.670. URL https://aclanthology.org/2020.acl-main.670/.

Yizhu Jiao, Ming Zhong, Sha Li, Ruining Zhao, Siru Ouyang, Heng Ji, and Jiawei Han. Instruct and extract: Instruction tuning for on-demand information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10030–10051. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.620.

Mael Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Dónal Landers, and Andre Freitas. NLI4CT: Multi-evidence natural language inference for clinical trial reports. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 16745–16764. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.1041. URL https://aclanthology.org/2023.emnlp-main.1041/.

Ghazal Khalighinejad, Defne Circi, L. Brinson, and Bhuwan Dhingra. Extracting polymer nanocomposite samples from full-length documents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13163–13175. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.779.

Ghazal Khalighinejad, Sharon Scott, Ollie Liu, Kelly L. Anderson, Rickard Stureborg, Aman Tyagi, and Bhuwan Dhingra. MatViX: Multimodal information extraction from visually rich articles. In *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)*, pp. 3636–3655. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.185.

Yang Li, Yajiao Wang, Yu Zhang, Yuanzhe Zhang, Maodi Hu, Mengting Zhang, Xi Sun, Hua Yue, and Zhixiong Zhang. SudokuFill: A multi-agent progressive filling framework for document-level scientific information extraction. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 33112–33138. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1657. URL https://aclanthology.org/2026.findings-acl.1657/.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638.

Youmi Ma, An Wang, and Naoaki Okazaki. DREEAM: Guiding attention with evidence for improving document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1971–1983. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.eacl-main.145. URL https://aclanthology.org/2023.eacl-main.145/.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Aakanksha Naik, Bailey Kuehl, Erin Bransom, Doug Downey, and Tom Hope. CARE: Extracting experimental findings from clinical literature. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 4580–4596. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-naacl.285. URL https://aclanthology.org/2024.findings-naacl.285/.

Massimiliano Pronesti, Joao H Bettencourt-Silva, Paul Flanagan, Alessandra Pascale, Oisín Redmond, Anya Belz, and Yufang Hou. Query-driven document-level scientific evidence extraction from biomedical studies. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28034–28051. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1359. URL https://aclanthology.org/2025.acl-long.1359/.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. GoLLIE: Annotation guidelines improve zero-shot information-extraction. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Y3wpuxd7u9.

Liyan Tang, Philippe Laban, and Greg Durrett. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8818–8847. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.499. URL https://aclanthology.org/2024.emnlp-main.499/.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, and Jiawei Han. Eider: Empowering document-level relation extraction with efficient evidence extraction and inference-stage fusion. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 257–268, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.23. URL https://aclanthology.org/2022.findings-acl.23/.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11328–11348. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.634. URL https://aclanthology.org/2023.acl-long.634/.

Fu Zhang, Hongsen Yu, Jingwei Cheng, and Huangming Xu. Entity pair-guided relation summarization and retrieval in LLMs for document-level relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4022–4037. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.224. URL https://aclanthology.org/2025.findings-naacl.224/.

Qi Zhang, Zhijia Chen, Huitong Pan, Cornelia Caragea, Longin Jan Latecki, and Eduard Dragut. SciER: An entity and relation extraction dataset for datasets, methods, and tasks in scientific documents. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13083–13100, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.726. URL https://aclanthology.org/2024.emnlp-main.726/.

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. Doc2EDAG: An end-to-end document-level framework for chinese financial event extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 337–346, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1032. URL https://aclanthology.org/D19-1032/.