RESEARCH ARTICLE

Extracting Agricultural Experts' Experiential Knowledge from Non-Standardized Text with Quantity and Operation Scope Constraints

Shigang Liu[1,2,3,4,5], **Huarui Wu**[1,2,3,4,5,*] **and Jialong Huang**[1,2,3,4,5]

[1] School of Information Engineering, Dalian Ocean University, Dalian 116023, China
[2] National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China
[3] Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China
[4] Key Laboratory of Digital Village Technology, Ministry of Agriculture and Rural Affairs, Beijing 100097, China
[5] Beijing Research Institute of Agricultural Artificial Intelligence and Robotics, Beijing 100097, China

Abstract

Expert advice on crop management is often written informally, and in such text a number can easily be assigned the wrong role or attached to the wrong operation, material, or measurement basis. We propose QS-IE, a knowledge extraction method constrained by quantity and operation scope. Its unit of extraction is a complete quantitative fact. Given a shared set of numeric candidates and a single base model, QS-IE selects complete candidates from several prompting strategies, verifies them against the source text, and applies scope constraints, so that every extracted quantity stays linked to its source evidence. We built a corpus of 4006 texts on tomato cultivation. Two researchers with agronomy backgrounds jointly annotated and reviewed 1000 of them; the other 3006 were annotated with large language model assistance and then reviewed by hand. The corpus was split into 3206 development texts and 800 independent test texts. On the 1152 reference facts of the test set, QS-IE reached a precision of 84.02%, a recall of 86.72%, and a micro-averaged F1 of 85.35%, which is 3.91 percentage points higher than the strongest baseline, a multi-strategy in-context learning ensemble. Removing the operation scope constraints lowered test F1 by 8.24 percentage points. In a sampled human audit of the test set, 82.00% of the outputs for texts containing target facts were judged usable, and all six engineering checks on storage and evidence traceability were passed. The results indicate that explicit association constraints make quantitative facts extracted from tomato cultivation texts more complete and easier to review, which supports the structured, traceable storage of experiential knowledge that experts express in non-standardized text.

Keywords: agricultural expert knowledge,

Academic Editor:
Muzammil Khan

Submitted: 09 July 2026
Revised: 12 October 2026
Accepted: 17 October 2026
Published: 26 October 2026

Vol. 3, **No.** 4, 2026.
10.62762/TIS.2026.540629

***Corresponding authors:**
✉ Huarui Wu
wuhr@nercita.org.cn

Citation
Liu, S., Wu, H., & Huang, J. (2026). Extracting Agricultural Experts' Experiential Knowledge from Non-Standardized Text with Quantity and Operation Scope Constraints. *ICCK Transactions on Intelligent Systematics*, 3(4), 229–244.

© 2026 by the Authors. Published by Institute of Central Computation and Knowledge. This is an open access article under the CC BY license (https://creativecommons.org/licenses/by/4.0/).

non-standardized advice, knowledge extraction, numeric roles, operation scope, evidence traceability.

1 Introduction

Agricultural experts pass on their experience through discussions, field guidance, informal farm records, and technical documents. Much of this knowledge is fragmentary and depends on context; the reasoning behind a recommendation and the conditions under which it applies are frequently left unstated. Growth stages, materials, and numeric parameters appear side by side, and one number may stand for a formulation content, a dosage, an interval, or a growth condition. Consider the advice "Use 100 g of 50% carbendazim or 80 g of 70% thiophanate-methyl per mu, mixed with 30 kg of water for spray application."[1] Here the percentages are formulation contents, while the gram values are dosages of two alternative materials. Whether the water quantity and the spray operation apply to one or both materials can only be read from the alternative relationship. If one material is combined with the content or dosage of the other prescription, the resulting fact is wrong even though every individual field is correct.

Work on knowledge acquisition and expert elicitation has stressed that the reasoning and applicability of expert knowledge must be preserved [1–4], and participatory agricultural research and horticultural ontologies have dealt with how experience is expressed and how terminology is aligned [5, 6]. In this paper, "non-standardized experience" means recommendations that use varied syntax and conditional associations without following a uniform schema. We study tomato (*Solanum lycopersicum* L.) production experience that has already been written down as text. In such text, association errors make the knowledge hard to audit and hard to organize.

We propose knowledge extraction using quantity and operation scope constraints (QS-IE). The contributions are threefold. First, quantities, numeric roles, recommendation attribution, operations, materials, measurement bases, and source evidence are represented together as one inseparable complete quantitative fact. Second, we constrain numeric roles, recommendation attribution, operation scope, and the alignment between materials and measurement bases, and we verify the outputs of several prompting strategies against the source text; this resolves attribution ambiguities and reduces errors that cross operations or prescriptions. Third, we evaluate complete-fact extraction, usability, and evidence traceability through method comparisons, ablations, human audits, and storage checks on 3206 development texts and 800 independent test texts.

Section 2 reviews related work. Section 3 describes corpus construction, fact representation, QS-IE, and the evaluation settings. Section 4 reports the experiments on complete-fact extraction, constraint ablation, human quality audit, and storage and traceability, and then discusses performance and applicability. Section 5 concludes the paper.

2 Related Work

2.1 Agricultural Knowledge Representation and Contextual Association

Research on agricultural knowledge representation deals with how concepts are organized, how provenance is kept, and how knowledge is adapted to context. Ontology-based models organize agricultural knowledge [7], vocabulary and ontology repositories support terminology alignment [8], and reviews of text-based knowledge extraction in the agricultural domain have surveyed entity recognition, relation extraction, and structured knowledge representation across diverse crop contexts [9]. Pest and disease knowledge graphs support evidence-constrained and traceable retrieval [10], agronomic attribute extraction links parameters to operations [11], and condition-aware retrieval matches knowledge units to queries according to production conditions [12]. For agricultural question answering, Gu et al. [13] fine-tune a model on question-answer pairs derived from a knowledge graph, using structured subgraphs to constrain inference. Wu et al. [14] take a complementary approach, pairing a crop pest and disease knowledge graph with dual retrieval paths—local entity subgraphs for factual queries and community summaries for analytical ones. Our work builds on these studies but focuses on a narrower question: how quantitative parameters and their associated fields should be attributed jointly when several operations and alternative prescriptions occur in the same text.

2.2 Structured Information Extraction and Quantity Association

Structured extraction models entities, relations, and events, for example by generating a target schema or by using contextualized span representations [15, 16]. In agricultural text,

[1] 1 mu $\approx$ 666.7 m^2 (1/15 ha).

domain-specific parsing extracts relationships between terms and operations to support structured knowledge construction [17], and in scientific text, quantity extraction organizes material-property relations through schema supervision with large language models [18]; follow-up questions can further check material-quantity-unit correspondences in passages with many values [19]. Agricultural recommendations add a further difficulty: roles and prescriptions must also be attributed. A single material may have both a formulation content and a dosage, and alternative prescriptions may share an operation or a measurement basis. Correct fields therefore do not guarantee a valid combination, and associations within a prescription need to be verified.

2.3 Operation Scope and Boundaries between Alternative Prescriptions

Two lines of work are relevant here. Semantic graphs of operations capture local dependencies between operations and parameters; for example, graphs of materials synthesis procedures explicitly link operations, materials, and conditions [20]. Flowchart recovery, in contrast, deals with branching control and logical consistency. PAGED represents branches and execution constraints with exclusive, inclusive, and parallel gateways [21], and text2flow refines extracted flowcharts by simulating their execution and feeding back inconsistencies with the source text [22].

Operation scope decides which operation a parameter modifies, and prescription boundaries decide which alternative prescription it belongs to. A single operation may carry a dosage, a frequency, and a duration. The fact that two prescriptions share an operation category, such as spray application, is no reason to exchange their materials and dosages, although conditions and measurement bases that are explicitly shared may apply to both. A wide context window may pull in parameters from a neighbouring branch, while splitting at punctuation may cut off shared conditions; in either case, attribution and sharing must be supported by the text. Retrieving similar demonstrations improves their relevance [23], but it cannot replace verification of the associations themselves. QS-IE therefore selects and evaluates complete quantitative facts and uses constraints and source-text verification to test numeric roles, recommendation attribution, and operation-material-measurement basis associations together.

3 Methodology

3.1 Corpus Construction, Reference Annotation, and Data Partitioning

We collected publicly available online materials that record expert knowledge of tomato cultivation. The sources were agricultural service platforms, agricultural and rural affairs departments, agricultural technology companies, and research institutions, and the materials covered extension articles, question answering, and practical guidance. After relevance filtering, cleaning, deduplication, and semantic organization, the 7241 raw records were reduced to 4006 texts; records unrelated to tomato and records without a source or sufficient evidence were removed. Deduplication relied on Unicode NFKC normalization, source associations, question identifiers, and hashes. Texts were segmented along contextual and semantic boundaries so that operation-material-value associations were not broken. Texts that were valid in context but contained no complete quantitative target were kept for false-positive evaluation. Table 1 summarizes the corpus size and partitions.

Table 1. Raw collection size, annotation paths, and experimental partitions.

Item	Count	Unit
Raw records collected	7241	records
Corpus after screening and cleaning	4006	texts
Jointly annotated and reviewed texts	1000	texts
Texts annotated with DeepSeek-V3 assistance and human review	3006	texts
Development set for internal evaluation	3206	texts
Independent test set	800	texts
Question identifiers in the development set	2085	identifiers
Development texts containing complete target facts	1043	texts
Development texts without complete target facts	2163	texts
Numeric candidate occurrences in the development set	6421	occurrences
Complete reference facts in the development set	3891	facts
Complete reference facts in the independent test set	1152	facts

Two researchers with agronomy backgrounds, referred to as A and B, prepared the reference annotations

under a common protocol. Of the 4006 unique texts, they jointly annotated and reviewed 1000 (600 development and 400 test texts). The remaining 3006 texts were pre-annotated with DeepSeek-V3-0324 (hereafter DeepSeek-V3) [24] and then reviewed by hand. For the 2606 development drafts, A and B each reviewed 1303 and then exchanged drafts for cross-review. For the remaining 400 test texts, A reviewed each draft, while B annotated the same texts independently without seeing the drafts. The two results were compared, and disagreements were resolved against the source text to produce the final reference annotations.

Both annotation paths used the same fact schema and checked for omitted targets as well as numeric roles, recommendation attribution, operations, materials, units, measurement bases, and evidence locations. Information that the text did not state was left empty rather than filled in from common knowledge. DeepSeek-V3 was used only to prepare reference annotations, whereas Qwen2.5-72B performed extraction. The development set covers 2085 question identifiers, 3891 reference facts (1043 texts with targets and 2163 without), and 6421 numeric candidate occurrences. The test set contains 1152 reference facts (300 texts with targets and 500 without). Texts, candidates, and facts are counted separately, since one text may contain several facts.

The corpus was divided into a development set D_{dev} for internal evaluation (3206 texts, 80.03%) and an independent test set D_{test} (800 texts, 19.97%). Development texts were used for demonstrations, constraint design, and error analysis; test texts were held out until the configuration was fixed. To limit overlap, partitioning took into account source-linked groups, question identifiers, parent context records, and near-duplicates. Test texts and their references were never used for demonstrations or rule selection. Neither in-context learning nor the deterministic post-processing updates the parameters of the base model, so "independence" here refers to a held-out partition of the same tomato corpus.

The demonstration pool contains 60 annotated development records from 30 source groups. Examples from the source group of the current input are excluded, and at most one example is taken from any group. To examine the influence of demonstrations, we also report results on the 3146 development texts that remain after removing the demonstration-associated groups. Because the development set informed the method design, results on it are not a substitute for test performance.

Numeric candidates are built by rule-based parsing: regular expressions locate values and units in the source text and form a shared candidate list. The base model then assigns numeric roles and links each value to an operation, a material, and a measurement basis, but it does not add numeric candidates of its own. The same procedure is used for both partitions, and reference facts outside candidate coverage are still counted as false negatives (FN).

3.2 Representation of Complete Quantitative Facts

Scientific quantity extraction has used four-way relations among material, property, quantity, and unit [25]. We extend this with numeric role, recommendation attribution, and operation binding, and define a complete quantitative fact as

$$\mathcal{K} = (\text{operation},\ \text{quantity and unit},\ \text{numeric role},\ \text{recommendation attribution},\ \text{material},\ \text{measurement basis},\ \text{evidence},\ \text{context}).$$

The first six fields describe the experiential fact itself, and evidence and context support its validity. Targets are numeric values in expert recommendations that are bound to an operation. A fact with a wrong association cannot be completely correct, even if its value and entities are correct.

The schema defines 17 numeric roles: dosage, dilution parameter, formulation content, repetition count, duration, frequency, environmental target, growth condition, observation, identifier, sequence number, spacing, target count, effect delay, time of day, environmental condition, and other. It also defines 16 operations: root drenching, spray application, irrigation, fertilization, seed soaking, germination promotion, air/sun drying, sowing/transplanting, environmental regulation, light management, pruning, pollination, drainage, field sanitation, crop rotation, and other. Attribution distinguishes expert recommendations, expert descriptions, user questions, and metadata. For instance, "300-fold dilution" is a dilution parameter and is kept apart from a percentage formulation content, and "per mu" is recorded as a measurement basis only when it is stated explicitly.

A value may be bound to several operations supported by the evidence, in which case it yields several facts; conversely, one operation may be linked to values with different roles. In "Spray every 7-10 days for

2-3 applications," both the frequency and the total count are bound to spray application but are stored as separate records. "5-6 leaves" is a seedling-stage condition, not a material dosage. When several agrochemicals are listed, we also check whether they are alternatives. Material and measurement basis may be empty, so that operation parameters supported by the source but unrelated to any material are still kept.

Each record stores identifiers, the positions of the value and the unit, the evidence unit and category of the operation, and the span of the material, which together distinguish facts and locate evidence. Numeric positions are zero-based, half-open intervals. Material quotations are located by the evidence unit, the verbatim text, and the order of occurrence. Identical values at different positions may be different facts, so only records with identical complete canonical keys are merged. Normalization only removes verifiable differences in characters and quotation positions; prescriptions are never merged because they look semantically similar. The original units and text are retained for later verification in context.

3.3 Comparison Methods and Demonstration Retrieval

Retrieval-augmented generation supplies external evidence to a language model [26], and in-context learning conditions a model on a task description and examples without updating its parameters [27]. We retrieve annotated development records to show the model the target schema, and we compare four methods. Schema-guided zero-shot extraction (SG-ZS) receives the schema, the source text, and the numeric candidates. Retrieval-augmented in-context learning (RA-ICL) adds up to five retrieved examples. The multi-strategy in-context learning ensemble (MS-ICL) pools the outputs of retrieval-based, random, diversity-based, and zero-shot prompting and selects complete candidates from them. QS-IE adds numeric-role, recommendation-attribution, and operation-scope constraints, source verification, and material-measurement basis alignment. All four methods share the base model, numeric inputs, schema, and strict matching criterion. Because QS-IE makes additional verification calls, the comparison is between whole pipelines and not between rule sets at equal inference cost.

The retrieval score is a weighted sum of the Jaccard similarities of character bigrams, unit sets, and operation sets, with weights 0.50, 0.30, and 0.20. Text is NFKC-normalized, and invisible characters and whitespace are removed. Ties in retrieval are broken by ascending identifier, and random selection uses seed 20261002. Diversity selection greedily maximizes 0.80 times the similarity to the target minus 0.20 times the maximum similarity to the examples already chosen, with ties broken by descending identifier. Inputs including demonstrations were limited to 10 000 tokens; as a result, 2920 development records received five retrieved examples and 286 received between zero and four. We do not assume that the same distribution holds for the test data.

For the current text x and a candidate demonstration d, the combined similarity is

$$s(x,d) = 0.5\,J(C_x, C_d) + 0.3\,J(U_x, U_d) + 0.2\,J(O_x, O_d), \tag{1}$$

where C, U, and O denote the sets of character bigrams, units, and operations. The Jaccard similarity is

$$J(A,B) = \frac{|A \cap B|}{\max\left(1, |A \cup B|\right)}, \tag{2}$$

so a component is zero when both sets are empty. These scores only guide lexical example selection and are not probabilities of agronomic correctness. Materials and dosages in a demonstration can never stand in for evidence in the current source text.

3.4 Quantity- and Scope-Constrained Extraction

Figure 1 shows corpus construction and the QS-IE pipeline. Outputs from retrieval-based, random, and diversity-based prompting feed the primary decoding, and zero-shot outputs are used for material consensus repair. For each numeric value, QS-IE keeps the complete role-attribution-operation-material-basis combinations that satisfy the initial source constraints. If any combination occurs at least twice, the most frequent one is chosen; otherwise, the choice is made with the field-agreement score in Eq. (3). Remaining ties are broken in a fixed order: retrieval, random, diversity. Candidates produced by verification are appended before re-decoding, and a strict majority of the candidates in the current decoding decides whether an operation is present. Each value receives one complete combination, which may contain several operation bindings that are emitted as separate facts. Fields from different candidates are never spliced together.

Numeric roles, recommendation attribution, and operation scope are checked both during initial candidate processing and after verification. Values that appear in user questions do not produce

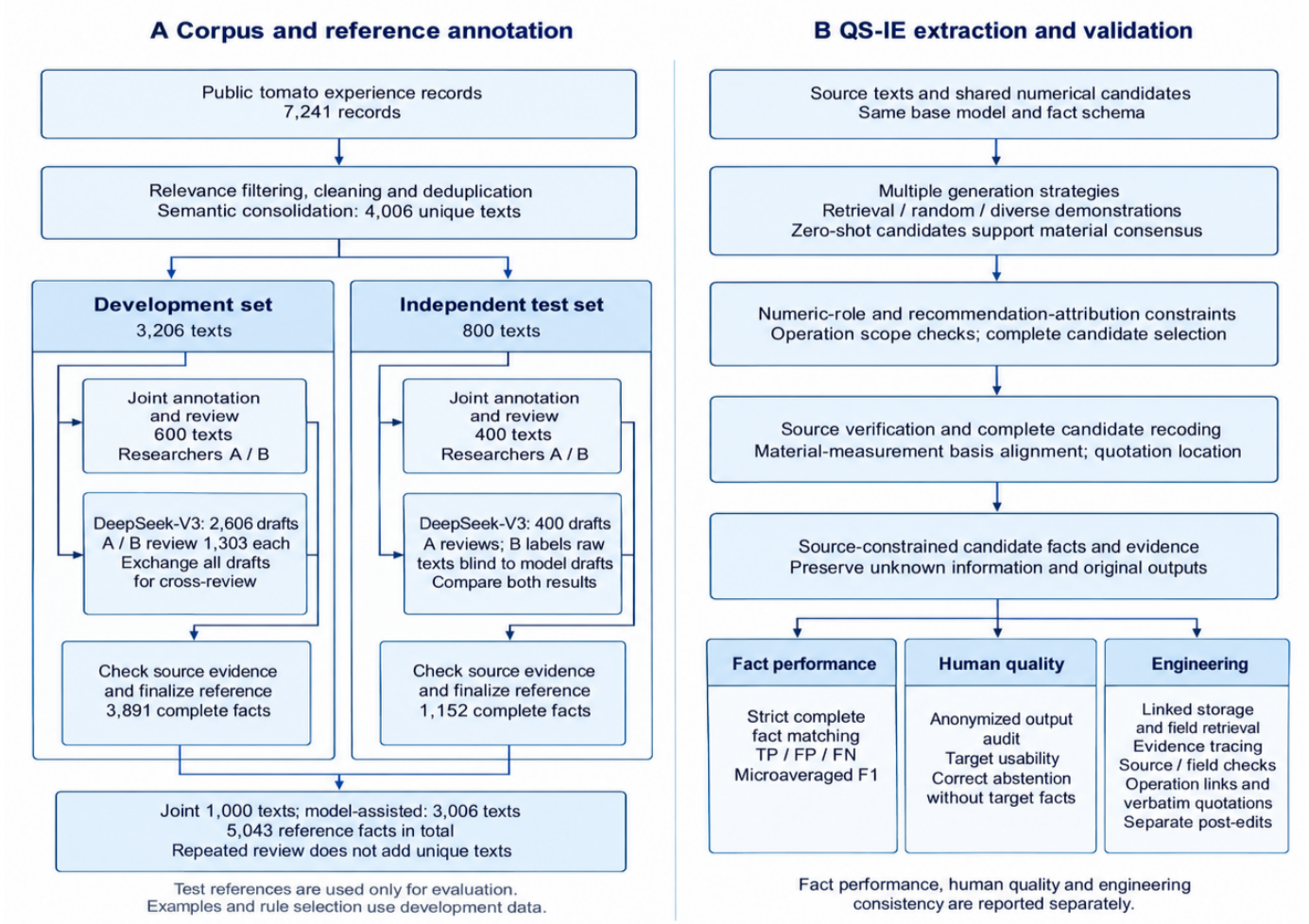


Figure 1. Corpus construction, reference annotation, and QS-IE. (A) The 7241 raw records yield 4006 unique texts, which are split into development and test sets; 1000 texts are jointly annotated and reviewed, and 3006 are annotated with DeepSeek-V3 assistance followed by human review. (B) Complete-candidate selection, source-text verification, and source constraints produce candidate facts with evidence. Test references are used only for evaluation.

recommendation bindings. Non-operation roles, such as identifiers, sequence numbers, observations, and effect delays, do not form operation facts. Percentage contents, dilution factors, intervals, counts, and durations are handled separately. A material is attributed according to the evidence unit of the operation and is never carried across an explicit boundary between alternative prescriptions. Conditions, ratios, or growth conditions that are explicitly shared may be bound to several operations through context, but application methods are not inferred from knowledge about agrochemicals. Missing materials, measurement bases, threshold directions, and implementation conditions are left as unknown.

Every non-metadata value in an expert answer is verified once against the source text; one additional verification is allowed when a quotation cannot be located. The original and verified candidates are decoded together and then filtered by the predetermined source constraints. Material consensus repair requires that at least two baseline candidates agree with the primary candidate on all other associated fields. In addition, the material must be the only option in the same comma-delimited local clause, occur exactly once, lie within 24 characters of the value, be at least two characters long, and contain no numeric or formulation markers. Valid local quotations that already exist are kept.

The field-agreement score of a complete candidate f is

$$\mathrm{score}(f) = \sum_{k \in \mathcal{F}} w_k \sum_{g \in \mathcal{G}_v} \mathbb{I}(f_k = g_k), \tag{3}$$

where $\mathcal{G}_v$ is the candidate set for the current numeric value v, and $\mathcal{F}$ contains the fields role, recommendation attribution, operation, material, and measurement basis, with weights w_k of 3, 2, 4, 3, and

2, respectively. $\mathbb{I}(\cdot)$ equals 1 when the two fields are equal and 0 otherwise.

3.5 Inference Settings and Evaluation Metrics

Extraction uses Qwen2.5-72B [28] (Ollama identifier `qwen2.5:72b`) with temperature 0, seed 17, a context of 16 384 tokens, an input budget of 10 000 tokens, and at most 6144 output tokens; no parameters are updated. Strict one-to-one matching uses the canonical key of a complete fact defined in Section 3.2: matched extractions are true positives (TP), unmatched extractions are false positives (FP), and unmatched references are FN. Micro-averaged precision (P), recall (R), and F1 are computed with Eq. (4) [29]. A fact with a mismatched field can therefore count as both an FP and an FN. The development candidate list misses 15 numeric occurrences that correspond to 16 reference facts, and these remain FN. Test recall is likewise computed over all references, including those outside candidate coverage.

$$P = \frac{TP}{TP + FP}, \qquad R = \frac{TP}{TP + FN}, \qquad F1 = \frac{2TP}{2TP + FP + FN} \tag{4}$$

For the development set, we use paired bootstrap resampling of source groups with 2000 repetitions (seed 17); in each repetition the same groups are drawn for all methods, and TP, FP, and FN are recomputed. The 2.5th and 97.5th percentiles give 95% confidence intervals (CIs). Note that the 1971 development source groups and the 2085 question identifiers are different levels of aggregation. The intervals describe sampling uncertainty within this corpus, not the full uncertainty of method selection. Test results are point estimates under the same strict criteria, and neither the development intervals nor the significance conclusions are carried over to the test set.

Abbreviations: SG-ZS, schema-guided zero-shot extraction; RA-ICL, retrieval-augmented in-context learning; MS-ICL, multi-strategy in-context learning ensemble; QS-IE, knowledge extraction using quantity and operation scope constraints; TP, strictly matched facts; FP, unmatched extracted facts; FN, unmatched reference facts; P, precision; R, recall; CI, confidence interval; pp, percentage points.

3.6 Structured Storage and Evidence Traceability Verification

Experiment 4 evaluates structured storage, field retrieval, and evidence traceability. The 3206 development texts with 3915 QS-IE candidates and the 800 test texts with 1189 candidates are stored as linked records that keep the verbatim values, units, roles, operations, materials, measurement bases, complete evidence, and source hashes. Single values and ranges that can be parsed receive lower and upper bounds, and other expressions are kept verbatim. No new causal relationships, threshold directions, or prescription branches are inferred. An SQLite database links source, evidence-unit, and candidate-fact tables, so that operations, quantities, and their evidence remain available for review [18, 30].

Engineering checks examine the consistency of source hashes, field round trips, operation-evidence links, and quotation offsets in the case studies. On the development set, the checks cover all texts and candidates plus 20 quotations in six cases. Six queries cover material dosage, dilution parameters, transplanting count, seedling leaf stage, spray frequency, and records without quantitative output. These checks address whether records can be stored and traced; the case queries are qualitative, and no separate question-answering accuracy or F1 is computed. Human post-edits are stored apart from the original outputs and do not enter the automatic extraction statistics.

4 Experiments

4.1 Experimental Design and Human Quality Audit

4.1.1 Four Experiments and Development-Set Audit

Experiment 1 compares complete-fact extraction by SG-ZS, RA-ICL, MS-ICL, and QS-IE. Experiment 2 removes, one at a time, the numeric-role and recommendation-attribution constraints, the operation-scope constraints, and the material-measurement basis alignment. Two further analyses remove model-based source verification (on the independent test set) and material consensus repair (on the development set). Full QS-IE reuses its outputs from Experiment 1, whereas the ablations rerun the affected stages, with input budgets and failures handled in the same way for all variants. On the development set, challenge labels are D1 (multiple operations or alternative prescriptions), D2 (composite materials or multiple measurement bases), and D4 (mixed numeric roles); a text may carry more than one label. D3 (omitted context) contains only 4 records and 22 reference facts and is therefore left out of the main quantitative conclusions.

Experiment 3 audits output quality. The

development-set audit covers 200 source texts and the 800 original outputs of the four methods. Sampling follows quotas for challenge coverage, excludes source groups linked to demonstrations, and takes at most one text per source group and question; the sampling and anonymization seeds are 20261004 and 20261006. Of the 200 texts, 151 contain target facts (120 challenge texts and 31 other target texts) and 49 contain no complete target fact. Method identities are hidden as O1-O4, with a separate random mapping for each record. Two reviewers with agronomy backgrounds independently rate all 800 outputs, giving 1600 ratings in total. Faithfulness measures whether an output is supported by the source; completeness measures omissions of target quantitative facts and required units; binding accuracy measures whether numeric role, operation, material, and measurement basis are consistent; and overall usability measures whether the output could be used for knowledge organization without substantive revision. These 800 outputs come from 200 development texts and should not be confused with the 800 independent test texts.

For each method, the two reviewers give 302 ratings on the 151 target-bearing texts. Empty and failed outputs are counted as failures. An output is usable overall only if it passes faithfulness, completeness, and binding accuracy. The correct abstention rate on the 49 non-target texts is based on 98 ratings, and the acceptable decision rate uses all 400 ratings. Both reviewers have equal weight, and the number of ratings is not the number of independent texts. Observed agreement and Cohen's κ [31] are computed from the paired binary ratings of the 604 anonymized target-bearing outputs of the four methods. Because of its challenge quotas, this audit is meant to diagnose differences between methods rather than to estimate usability in the population.

4.1.2 *Independent Test Procedure*

The independent test uses the 800 predefined tomato cultivation texts together with the fact schema, role definitions, evidence-location rules, and method configurations fixed on the development set. The base model, example pool, retrieval weights, operation-scope constraints, and input/output budgets were all fixed in advance, and test references were kept separate from the inference inputs. All four methods produce original outputs on the same test set, which are used for complete-fact evaluation and for the ablations. Evaluation uses the final human-confirmed reference annotations described in Section 3.1, not the model drafts. Negative texts are kept according to the definition of a target fact; a text that contains only descriptive numbers, a diagnosis, or no complete quantitative recommendation is not a low-quality text, and target facts are always determined from source evidence.

Operations are distinguished from implementation conditions. For example, "fertilize when the soil is dry" states a condition for fertilization and does not imply a drying operation. In "fertilization during post-transplant recovery," transplanting may simply mark the timing of fertilization, and it is annotated as a separate operation only when it is explicitly recommended. Chemicals, application methods, measurement bases, and comparison directions that are not stated remain unknown, and material quotations are located verbatim.

All four methods are evaluated with the same complete-fact matching criteria against 1152 reference facts (TP + FN = 1152). Incorrect facts extracted from negative texts count as FP, and empty outputs, failed records, and reference facts not covered by numeric candidates all remain in the evaluation.

The test-set audit samples 200 source texts, 150 with target facts and 50 without complete target facts. Two reviewers independently assess the anonymized outputs of all four methods. For each method, faithfulness, completeness, binding accuracy, and overall usability are based on 300 target-bearing ratings, and overall usability again requires the same output to pass the other three criteria. Correct abstention uses 100 non-target ratings, and acceptable decisions use all 400 ratings. Reviewer agreement is calculated from the paired binary ratings of the 600 target-bearing outputs across all methods. Reference construction and output auditing are carried out separately.

The engineering checks use the linked structure of source text, evidence unit, and candidate fact to verify source hashes, field storage round trips, operation-evidence traceability, verbatim quotation locations, retrieval by source record, and retrieval by operation and role. Query results are compared with the fact sets defined by the corresponding source records or by the operation and role conditions. Source texts and records are also kept for non-target texts. These checks concern storage, retrieval, and the consistency of the evidence chain; they complement, but do not replace, complete-fact matching and human audits.

4.2 Experiment 1: Complete-Fact Extraction Performance

4.2.1 Independent Test Performance

Table 2. Strict complete-fact extraction performance on the independent test set (800 texts).

Method	TP	FP	FN	P (%)	R (%)	F1 (%)
SG-ZS	765	238	387	76.27	66.41	71.00
RA-ICL	924	242	228	79.25	80.21	79.72
MS-ICL	939	215	213	81.37	81.51	81.44
QS-IE	**999**	**190**	**153**	**84.02**	**86.72**	**85.35**

Note: TP, FP, and FN are fact counts. P, R, and F1 are computed from the integer counts and rounded to two decimal places. TP + FN = 1152 for all methods, and the numbers of extracted facts (TP + FP) are 1003, 1166, 1154, and 1189, respectively. Differences between methods and between partitions are computed from unrounded F1 values.

Table 2 reports strict extraction results on the 800 test texts and 1152 reference facts. QS-IE produces 1189 candidate facts and reaches a precision of 84.02%, a recall of 86.72%, and an F1 of 85.35%, which is 3.91 percentage points above the strongest baseline, MS-ICL. The ranking of the methods is the same on both partitions, and the test F1 of QS-IE is 0.67 percentage points higher than its development F1. As point estimates, these differences do not establish statistical significance, nor do they show generalization to other crops.

4.2.2 Development and Internal Evaluation Comparison

Table 3 gives the strict complete-fact results on the development set. QS-IE reaches an F1 of 84.68%, which is 13.43, 2.35, and 1.59 percentage points higher than SG-ZS, RA-ICL, and MS-ICL, respectively (computed from unrounded values). The 95% CI of its paired F1 difference from MS-ICL is [0.50, 2.74] percentage points.

On the 600 jointly annotated and reviewed development texts, the F1 values of the four methods are 71.26%, 80.19%, 81.67%, and 85.73%; on the 2606 texts annotated with DeepSeek-V3 assistance and human review, they are 71.25%, 82.82%, 83.42%, and 84.43%. After excluding the groups linked to demonstrations, QS-IE obtains an F1 of 84.70%. Because the two subsets differ in fact density and annotation pathway, their differences should not be read as a causal effect of model-assisted annotation.

4.3 Experiment 2: Constraint Ablation and Mechanism Analysis

4.3.1 Independent Test Ablations

Table 4 lists the ablation results on the test set. Removing the numeric-role and recommendation-attribution constraints, the operation-scope constraints, and the material-measurement basis alignment lowers F1 by 5.42, 8.24, and 6.72 percentage points, respectively, and removing model-based source verification lowers it by 0.97 percentage points. The largest drop comes from removing operation scope, as on the development set, which suggests that operation binding matters most for complete-fact correctness in this task.

4.3.2 Development-Set Mechanisms and Challenge Scenarios

Table 5 lists the development-set ablations. Removing the numeric-role and recommendation-attribution constraints, the operation-scope constraints, and the material-measurement basis alignment lowers F1 by 4.27, 9.45, and 4.94 percentage points, respectively, and removing material consensus repair lowers it by 0.15 percentage points (computed from unrounded values). The first three rerun variants contain 2, 3, and 2 over-budget records, which are kept in the evaluation. Removing operation scope again causes the largest drop. Since removing one module also changes later inference steps, these results do not isolate the causal effect of any single rule.

Figure 2 compares D1, D2, D4, and target-bearing controls without challenge labels. QS-IE reaches F1 values of 84.52%, 87.50%, 91.29%, and 77.74% on these groups, respectively. D1 merges multiple operations and alternative prescriptions, which often occur together; because the two are not reported separately, the D1 result cannot be attributed to either challenge alone. The groups also differ in fact density and semantic composition, so the ranking of F1 values is not a ranking of difficulty. The 685 records with at least one D1-D4 label consist of 587 target-bearing and 98 non-target records and contain 2841 reference facts. Another 452 target-bearing controls contain 1030 facts, and 4 target-bearing records with uncertain labels contain 20 facts, giving 3891 reference facts across the three groups. On the 685 challenge records, QS-IE reaches an F1 of 88.07%, and its record-level false-positive rate on the 2163 non-target development records is 2.17%.

An automated analysis of field mismatches finds complete-fact errors or omissions in 374 QS-IE

Table 3. Strict complete-fact extraction performance on the development set (3206 texts).

Method	TP	FP	FN	P (%)	R (%)	F1 (%)	[95% CI]
SG-ZS	2651	899	1240	74.68	68.13	71.25	[68.90, 73.53]
RA-ICL	3186	663	705	82.77	81.88	82.33	[79.95, 84.68]
MS-ICL	3208	623	683	83.74	82.45	83.09	[80.94, 85.35]
QS-IE	**3305**	**610**	**586**	**84.42**	**84.94**	**84.68**	**[82.63, 86.67]**

Note: All methods use the same 3206 texts and 3891 reference facts. The 95% CIs are based on 2000 paired bootstrap resamples of source-linked groups.

Table 4. Ablation results on the independent test set (800 texts).

Setting	TP	FP	FN	P (%)	R (%)	F1 (%)	Drop (pp)
QS-IE	999	190	153	84.02	86.72	85.35	0.00
Without numeric role and recommendation attribution constraints	946	269	206	77.86	82.12	79.93	5.42
Without operation scope constraints	881	252	271	77.76	76.48	77.11	8.24
Without material-measurement basis alignment	920	268	232	77.44	79.86	78.63	6.72
Without model-based source-text verification	991	206	161	82.79	86.02	84.38	0.97

Table 5. Ablation results on the development set (3206 texts).

Variant	P (%)	R (%)	F1 (%)	Drop (pp)
QS-IE	84.42	84.94	84.68	0.00
Without numeric role and recommendation attribution constraints	79.65	81.19	80.41	4.27
Without operation scope constraints	77.53	73.07	75.23	9.45
Without material-measurement basis alignment	79.97	79.52	79.74	4.94
Without material consensus repair	84.27	84.79	84.52	0.15

Note: Full QS-IE reuses its outputs from Experiment 1. Drops are computed from unrounded F1 values; pp denotes percentage points.

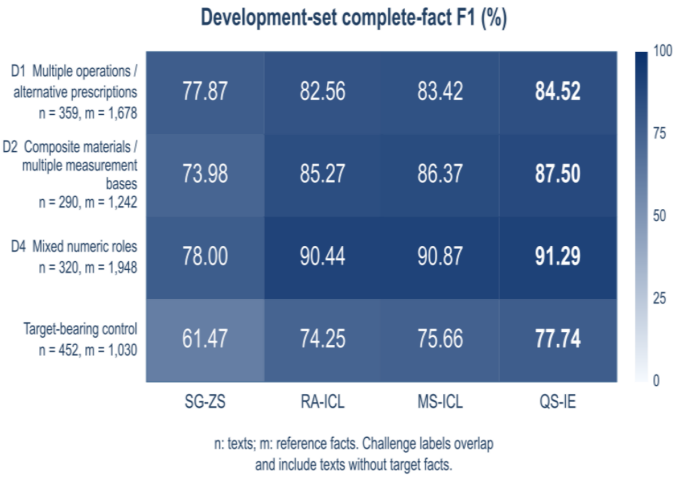


Figure 2. Complete-fact F1 (%) on the development-set challenge scenarios and target-bearing controls. n is the number of texts with each label, and m is the number of reference facts. Challenge labels may overlap and include non-target texts, whereas the controls contain only target-bearing texts. D3 (4 texts, 22 facts) and the 4 target-bearing records with uncertain labels are not plotted. The color scale ranges from 0 to 100%.

development records. Operation binding, numeric role, material, and measurement basis are involved in 244, 148, 106, and 9 of these records, respectively. Because a record can carry several labels, the counts do not add up to the number of erroneous records, and the automated breakdown does not assign a single, human-adjudicated cause to each error. Together with the ablations, it points to operation association as the main target for improvement. Local material repair should also check operation and seasonal scope, so that correct local fields do not hide a wrong overall binding.

4.4 Experiment 3: Human Quality Audit

4.4.1 Independent Test Audit

Table 6 summarizes the test-set audit. QS-IE achieves an overall usability of 82.00%, a correct abstention

rate of 96.00%, and an acceptable decision rate of 85.50%; the first and last are 7.00 and 4.25 percentage points higher than those of MS-ICL. Pooled over the four methods, Cohen's κ is 0.62, 0.85, 0.83, and 0.81 for faithfulness, completeness, binding accuracy, and overall usability, respectively. These figures describe the audited sample only and are not estimates for all 800 test texts.

4.4.2 *Development-Set Audit*

Figure 3 shows the audit of 200 development texts. Out of 302 target-bearing ratings, QS-IE passes 277 for faithfulness, 283 for completeness, 269 for binding accuracy, and 253 for overall usability. The two reviewers judge 127 and 126 of the 151 outputs usable, giving a pooled overall usability of 83.77%, 10.60 percentage points above MS-ICL. This difference applies only to the challenge-quota audit sample.

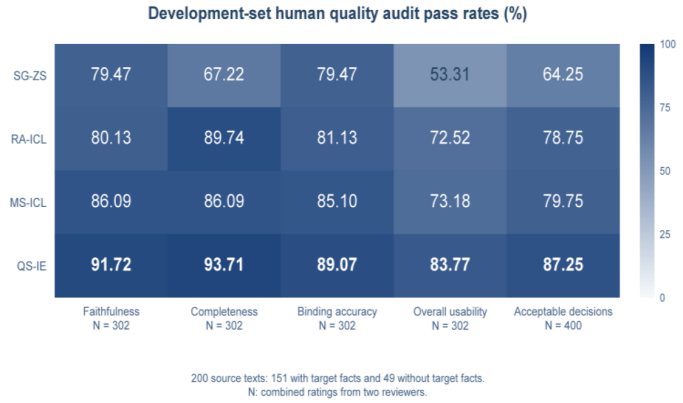


Figure 3. Pooled human audit pass rates (%) on 200 development texts. The first four columns use 302 ratings of 151 target-bearing texts; the last column (acceptable decisions) uses all 400 ratings of the 200 texts and includes correct abstention on 49 non-target texts. The two reviewers rate each output independently, so the number of ratings is not the number of independent texts. The color scale ranges from 0 to 100%.

For the 49 non-target audit records, SG-ZS, RA-ICL, and QS-IE each abstain correctly in 96 of 98 ratings (97.96%), and MS-ICL in all 98 (100.00%). These quota-sample rates are not extrapolated to all non-target development texts.

The 253 usable target-bearing judgments and 96 correct abstentions of QS-IE give 349 acceptable decisions, that is, 349/400 = 87.25%; the corresponding rates of the three baselines are 64.25%, 78.75%, and 79.75%. Acceptable decisions cover both extraction and abstention, whereas overall usability concerns target-bearing records only.

For the 604 anonymized target-bearing outputs, the paired reviewer ratings yield Cohen's κ values of 0.62, 0.85, 0.80, and 0.81 and observed agreement of 89.90%, 95.86%, 94.54%, and 92.05% for faithfulness, completeness, binding accuracy, and overall usability, respectively. These values measure agreement in auditing outputs, not agreement in building the reference annotations. Faithfulness attracts the most disagreement, and the binding accuracy of QS-IE is lower than its faithfulness and completeness, indicating that field associations remain the main thing to revise.

4.5 Experiment 4: Structured Storage and Evidence Traceability

4.5.1 *Independent Test Storage and Traceability Checks*

All 800 test texts and 1189 QS-IE candidate facts are stored as linked SQLite records. All six engineering checks pass, and Table 7 lists their scope and results. All 9 predefined types of injected errors are detected.

4.5.2 *Development-Set Engineering Checks and Cases*

All 3206 development texts and 3915 original candidates are stored as linked records. The source-hash, field round-trip, and operation-evidence link checks pass, and all 20 quotation locations in the six cases are consistent (Table 8). Of the 3915 candidates, 2554 are single values, 1240 are ranges, and 121 are expressions kept verbatim. In all, 1062 records have candidate outputs and 2144 have none; the latter may include omissions and cannot all be counted as correct abstentions. Range bounds represent only intervals stated in the source. Expressions that cannot be parsed remain as text, and no threshold directions or judgments such as "optimal" are added. Figure 4 shows the chain from storage to retrieval and evidence tracing.

Six field queries retrieve material dosage, dilution parameters, transplanting counts, seedling leaf stages, and spraying frequency, and show that source texts without quantitative output are retained. Table 9 presents three cases: mismatched fertilization timing and materials, mismatched seasons and measurement bases, and abstention for an image-dependent text without an explicit quantitative recommendation. The six cases contain 23 original fact cards.

The three cases in Table 9 are taken from the six cases and show how source evidence supports the manual correction of associations. The post-edited results are excluded from F1 and from the human-audit pass rates, and these cases do not demonstrate overall revision success or any automatic error-correction ability.

Table 6. Human quality audit of the four methods on the independent test set (pooled pass rates, %).

Method	Faithfulness	Completeness	Binding accuracy	Overall usability	Correct abstention	Acceptable decisions
SG-ZS	78.00	64.33	80.00	52.33	98.00	63.75
RA-ICL	84.33	86.00	83.00	72.67	98.00	79.00
MS-ICL	86.33	87.33	84.33	75.00	100.00	81.25
QS-IE	**87.67**	**94.67**	**85.33**	**82.00**	96.00	**85.50**

Note: Values are pooled pass rates (%) from two reviewers. For each method, the first four criteria use 300 target-bearing ratings, correct abstention uses 100 non-target ratings, and acceptable decisions use all 400 ratings.

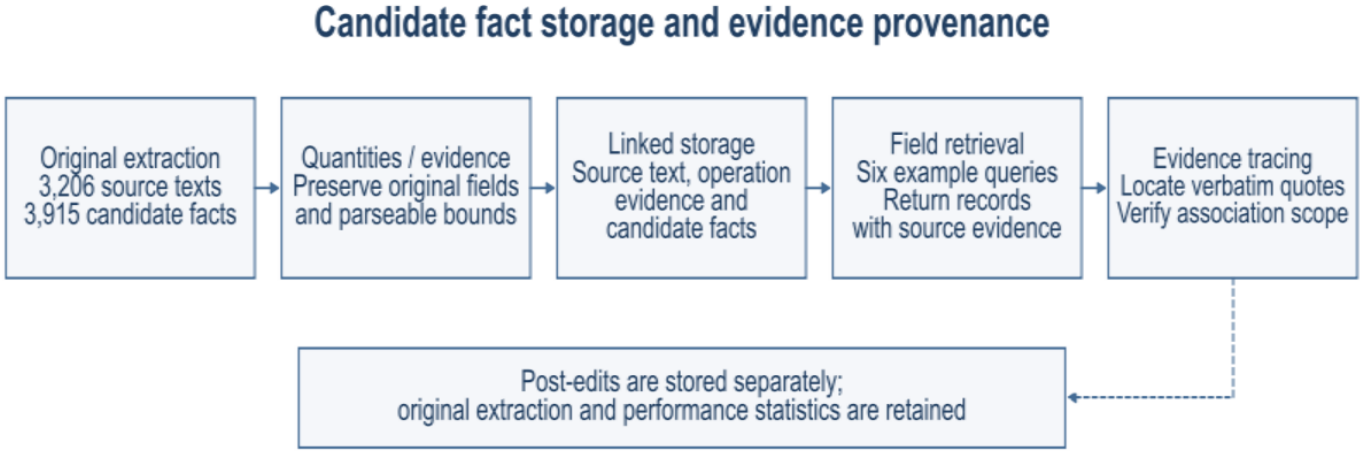


Figure 4. Linked storage, field retrieval, and evidence traceability of development-set candidate facts. The 3915 candidates include both TP and FP. Post-edits are stored separately and do not replace the original extraction outputs or their performance statistics.

Table 7. Engineering checks on the independent test set.

Check	Passed / total	Pass rate (%)
Source-text hash consistency	800 / 800	100.00
Round-trip consistency of fact fields	1189 / 1189	100.00
Traceability of operation evidence	1189 / 1189	100.00
Verbatim localization of linked quotations	3092 / 3092	100.00
Retrieval consistency by source record	800 / 800	100.00
Retrieval consistency by operation and role	65 / 65	100.00

Table 8. Engineering checks for development-set storage and evidence traceability.

Check	Passed / total	Pass rate (%)
Source-text hash checks	3206 / 3206	100.00
Round-trip comparisons of candidate fact fields	3915 / 3915	100.00
Operation evidence link checks	3915 / 3915	100.00
Quotation location checks in the cases	20 / 20	100.00

Note: The quality of candidate content is evaluated in Experiments 1 and 3. The 20 quotations come from the six case checks.

4.6 Discussion

The comparisons and ablations on both partitions support the use of field-association constraints. Removing operation scope lowers development and test F1 by 9.45 and 8.24 percentage points, respectively, which indicates that operation binding has a strong effect on complete-fact extraction. This agrees with the observation that, in agricultural recommendations, several operations and alternative prescriptions often share nearby numbers. Removing material consensus repair lowers development F1 by only 0.15 percentage points, and removing source verification lowers test F1 by 0.97 percentage points. Research on multi-prompt aggregation estimates the accuracy of prompts and the dependencies among them through weak supervision [32]; our method, by contrast, selects candidates by frequency and field agreement. Since dependencies among candidates are not modeled, different strategies may share systematic errors, which is why the candidates still have to be checked against

Table 9. Representative cases of source-constrained knowledge structuring.

Source record	Review focus	Original issue and treatment
652	Fertilization timing, material, and operation	The original candidates misidentified the timing of fertilization after transplanting and treated root drenching as a material. Post-editing clarified the timing and separated the two alternative application methods (application with irrigation water or root drenching); unspecified materials were left for further review.
1304	Transplanting quantity, seasonal context, and measurement basis	The original candidates omitted the per-mu measurement basis and assigned a wrong role to a calendar period. Post-editing kept the measurement basis, seasonal context, and environmental threshold conditions stated in the source.
3106	Image dependence and missing quantities	The source gave no explicit quantitative recommendation. The candidate set remained empty, the source and its image dependence were retained, and no dose was invented.

source evidence.

Compared with agricultural knowledge graphs and context-aware question answering [10–14], semantic graphs of operations, and flowchart extraction [20–22], this study asks a more specific question: whether numeric role, recommendation attribution, and the operation-material-measurement basis association jointly support a complete quantitative fact in the source. A correct operation category does not turn a cross-prescription combination into a TP. QS-IE needs additional verification calls, and whether it keeps its advantage at equal inference cost remains to be tested.

Micro-averaged F1 is computed over facts, whereas overall usability is judged on record-level outputs, so the two measures capture different aspects of quality. In addition, the development audit uses challenge quotas, and the audit distributions differ between partitions. Similar numbers therefore do not imply equal quality, and the audit rates do not estimate how much of the corpus could be stored as knowledge.

Once experiential knowledge is structured, its records can be traced to their sources and revised [2, 34]; expert-knowledge parameterization has also been shown to extend the applicability of crop management models across different environments and techniques [33]. Storage and traceability keep the original fields and evidence, so missing materials, operation mismatches, and unknown conditions become visible to human inspectors; they do not, however, correct semantic errors automatically. All engineering checks pass on both partitions, which supports the consistency of storage, retrieval, and the evidence chain within the checked scope, while the correctness of facts still has to be evaluated separately. The six cases show revisions that are supported by evidence and add no prescription parameters absent from the source. Any gains in production or in agricultural services would need dedicated field trials.

This study has several limitations. All texts concern tomato cultivation, and the method has not been validated across crops, regions, or independent institutions. The development set guided method selection and constraint design; the bootstrap intervals reflect sampling uncertainty within this corpus only, and the test point estimates do not establish statistical significance. The audit sampling limits extrapolation to the full corpus, and paired reviewer ratings are not independent samples. Because evaluation relies on shared numeric candidates, it mainly measures role identification and field binding; end-to-end numeric discovery without predefined candidates has not been tested. The schema covers non-quantitative techniques, implicit exceptions, and complex conditional logic only partially, and with only 4 D3 texts the robustness to omitted context cannot be assessed. Future work should examine transfer to other crops and end-to-end extraction, and should quantify inference cost and human revision effort.

5 Conclusion

QS-IE extracts complete quantitative facts from agricultural experts' experiential knowledge by modeling numeric role, recommendation attribution, operation scope, and material-measurement basis associations together. On 800 independent test texts with 1152 reference facts, it reaches a precision of 84.02%, a recall of 86.72%, and a micro-averaged

F1 of 85.35%, 3.91 percentage points higher than MS-ICL. Removing operation scope causes the largest ablation drop (8.24 percentage points), and the development set shows the same trend. In the test audit, the overall usability and acceptable decision rates are 82.00% and 85.50%, and all six storage and evidence-traceability checks pass. These results support the use of QS-IE for structuring experiential knowledge of tomato cultivation under source constraints, which can provide traceable knowledge for intelligent consultation and science communication. Generalization to other crops, end-to-end extraction without predefined numeric candidates, and the agricultural benefits of the approach still need to be validated.

Data Availability Statement

The corpus, annotation files, and extraction code produced in this study are not publicly available, as the source materials are subject to copyright and the terms of the platforms from which they were collected. The method configurations, random seeds, and model identifiers are reported in full in the Methodology section to support reproducibility of the extraction pipeline. Readers seeking further information may contact the corresponding author.

Funding

This work was supported by the National Major Agricultural Science and Technology Project under Grant 0304.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that GPT-6.1 Sol was used for language editing and translation of the manuscript. The authors have carefully reviewed, revised, and verified the AI-assisted output and take full responsibility for the content of the manuscript.

Ethical Approval and Consent to Participate

This study involved no human or animal experiments. The source texts used for corpus construction are publicly available and contain no private or sensitive information. The researchers who participated in annotation and quality auditing provided informed consent; their contributions were used solely for the analysis and publication described in this study, and no personal data were collected or retained beyond what was necessary for annotation review. No clinical intervention, deception, or collection of personal data from third parties was involved. This study was therefore exempt from formal ethical review in accordance with applicable institutional guidelines.

References

[1] Rafea, A., El-Dessouki, A., Hassan, H., & Mohamed, S. (1993). Development and implementation of a knowledge acquisition methodology for crop management expert systems. *Computers and electronics in agriculture, 8*(2), 129-146. [CrossRef]

[2] Girard, N., & Hubert, B. (1999). Modelling expert knowledge with knowledge-based systems to design decision aids: The example of a knowledge-based model on grazing management. *Agricultural Systems, 59*(2), 123-144. [CrossRef]

[3] Cornelissen, A. M. G., van den Berg, J., Koops, W. J., & Kaymak, U. (2003). Elicitation of expert knowledge for fuzzy evaluation of agricultural production systems. *Agriculture, ecosystems & environment, 95*(1), 1-18. [CrossRef]

[4] Robledo, J., Sulá, A. I. P., Jaworski, L. G., Mouafo-Tchinda, R. A., Andersen Onofre, K. F., Thomas-Sharma, S., & Garrett, K. A. (2025). Expert knowledge elicitation: Accessing the big data in experts' brains. *Phytopathology®, 115*(10), 1245-1259. [CrossRef]

[5] Semeon, G., & Garfield, M. J. (2019). Framework for externalization of tacit knowledge in participatory agricultural research in Ethiopia: The case of Farmers Research Group (FRG). In *Proceedings of the 52nd Hawaii International Conference on System Sciences* (pp. 5339-5347). https://aisel.aisnet.org/hicss-52/ks/culture_ks/4/

[6] Koenderink, N. J., Top, J. L., & van Vliet, L. J. (2005, August). Expert-based ontology construction: A case-study in horticulture. In *16th International Workshop on Database and Expert Systems Applications (DEXA'05)* (pp. 383-387). IEEE. [CrossRef]

[7] Ngo, Q. H., Kechadi, T., & Le-Khac, N. A. (2022). Knowledge representation in digital agriculture: A step towards standardised model. *Computers and Electronics in Agriculture, 199*, 107127. [CrossRef]

[8] Jonquet, C., Toulet, A., Arnaud, E., Aubin, S., Yeumo, E. D., Emonet, V., ... & Larmande, P. (2018). AgroPortal: A vocabulary and ontology repository for agronomy. *Computers and Electronics in Agriculture, 144*, 126-143. [CrossRef]

[9] Nismi Mol, E. A., & Santosh Kumar, M. B. (2023). Review on knowledge extraction from text and scope in agriculture domain. *Artificial intelligence review, 56*(5), 4403-4445. [CrossRef]

[10] Wu, H., Xie, N., Fan, J., Wang, X., Liu, H., Zheng, H., ... & Lv, Q. (2026). Autonomous construction of crop pest and disease knowledge graphs at scale via LLM-driven expert modeling. *Computers and Electronics in Agriculture, 250*, 111970. [CrossRef]

[11] Guo, W., & Zhao, C. (2026). Wheat agronomic knowledge extraction and spatio-temporal knowledge graph construction. *Applied Sciences, 16*(10), 4776. [CrossRef]

[12] Wang, Y., Zhao, K., Li, H., Wang, R., & Chen, W. (2026). A condition-aware retrieval-augmented decision support framework for tomato cultivation management. *Frontiers in Plant Science, 17*, 1802748. [CrossRef]

[13] Gu, W., Wang, X., Wang, Z., Gao, Y., Zhang, X., Zheng, J., & Liu, K. (2026). PlantExpert: A knowledge graph-augmented large language model for agricultural question answering. *Smart Agricultural Technology, 15*, 102467. [CrossRef]

[14] Wu, H., Xie, N., Wang, X., Fan, J., Li, Y., & Meng, Z. (2025). Crop GraphRAG: pest and disease knowledge base Q&A system for sustainable crop protection. *Frontiers in Plant Science, 16*, 1696872. [CrossRef]

[15] Lu, Y., Liu, Q., Dai, D., Xiao, X., Lin, H., Han, X., ... & Wu, H. (2022, May). Unified structure generation for universal information extraction. In *Proceedings of the 60th annual meeting of the association for computational linguistics* (*volume 1: Long papers*) (pp. 5755-5772). [CrossRef]

[16] Wadden, D., Wennberg, U., Luan, Y., & Hajishirzi, H. (2019, November). Entity, relation, and event extraction with contextualized span representations. 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. 5784-5789). [CrossRef]

[17] Kaushik, N., & Chatterjee, N. (2018). Automatic relationship extraction from agricultural text for ontology construction. *Information processing in agriculture, 5*(1), 60-73. [CrossRef]

[18] Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., ... & Jain, A. (2024). Structured information extraction from scientific text with large language models. *Nature communications, 15*(1), 1418. [CrossRef]

[19] Polak, M. P., & Morgan, D. (2024). Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications, 15*(1), 1569. [CrossRef]

[20] Mysore, S., Jensen, Z., Kim, E., Huang, K., Chang, H. S., Strubell, E., ... & Olivetti, E. (2019, August). The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures. In *Proceedings of the 13th linguistic annotation workshop* (pp. 56-64). [CrossRef]

[21] Du, W., Liao, W., Liang, H., & Lei, W. (2024, August). PAGED: A benchmark for procedural graphs extraction from documents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (*Volume 1: Long Papers*) (pp. 10829-10846). [CrossRef]

[22] Ying, W., Liu, Y., Zhao, X., Cheng, W., Chen, Z., Yu, W., Fu, Y., & Chen, H. (2026). Multi-agent procedural graph extraction with structural and logical refinement. In *Findings of the Association for Computational Linguistics: EACL 2026* (pp. 3021-3034). [CrossRef]

[23] Liu, J., Shen, D., Zhang, Y., Dolan, W. B., Carin, L., & Chen, W. (2022, May). What makes good in-context examples for GPT-3?. In *Proceedings of Deep Learning Inside Out* (*DeeLIO 2022*): *The 3rd workshop on knowledge extraction and integration for deep learning architectures* (pp. 100-114). [CrossRef]

[24] Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., ... & Piao, Y. (2024). Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*. [CrossRef]

[25] Court, C. J., & Cole, J. M. (2018). Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. *Scientific data, 5*(1), 180111. [CrossRef]

[26] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems, 33*, 9459-9474. https://papers.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[27] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems, 33*, 1877-1901. https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[28] Hui, B., Yang, J., Cui, Z., Yang, J., Liu, D., Zhang, L., ... & Lin, J. (2024). Qwen2.5-coder technical report. *arXiv preprint arXiv:2409.12186*. [CrossRef]

[29] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management, 45*(4), 427-437. [CrossRef]

[30] Buneman, P., Khanna, S., & Wang-Chiew, T. (2001, January). Why and where: A characterization of data provenance. In *International conference on database theory* (pp. 316-330). Berlin, Heidelberg: Springer Berlin Heidelberg. [CrossRef]

[31] Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational linguistics, 34*(4), 555-596. [CrossRef]

[32] Arora, S., Narayan, A., Chen, M. F., Orr, L., Guha, N., Bhatia, K., ... & Ré, C. (2022). Ask me anything: A

simple strategy for prompting language models. *arXiv preprint arXiv:2210.02441*. [CrossRef]

[33] Ballot, R., Loyce, C., Jeuffroy, M.-H., Ronceux, A., Gombert, J., Lesur-Dumoulin, C., & Guichard, L. (2018). First cropping system model based on expert-knowledge parameterization. *Agronomy for Sustainable Development, 38*, 33. [CrossRef]

[34] Buche, P., Cuq, B., Fortin, J., & Sipieter, C. (2019). Expertise-based decision support for managing food quality in agri-food companies. *Computers and Electronics in Agriculture, 163*, 104843. [CrossRef]

Shigang Liu is currently pursuing a master's degree at the School of Information Engineering, Dalian Ocean University, Dalian, China. His research interests include agricultural knowledge graphs, natural language processing, and intelligent agricultural applications. (Email: Nercita_ZSCQ_02@126.com)

Huarui Wu received the Ph.D. degree in engineering and is a Research Professor and doctoral supervisor at the Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing, China. His research focuses on intelligent services based on agricultural big data and on agricultural information technology. (Email: wuhr@nercita.org.cn)

Jialong Huang is currently pursuing a master's degree at the School of Information Engineering, Dalian Ocean University, Dalian, China. His research focuses on agricultural AI agent technologies and their applications. (Email: nercita_glwj_01@163.com)