Process-Guided Agentic Extraction of RCT Evidence Records with Source-Grounded Evaluation

Anonymous ACL submission

Abstract

Extracting randomized controlled trial (RCT) evidence from PDFs is not just schema filling: one result record may depend on tables, notes, captions, model context, and source-specific coding rules. Manual extraction is costly because experts follow staged guidelines: finding evidence units, binding fields to sources, checking support, and applying export policy. We study whether RCT-Flow, a staged agentic workflow, can operationalize these guidelines for automatic record discovery, source binding, field extraction, verification, repair, and frozen export. We benchmark RCT-Flow on a shared ***15-paper*** test set against $\text{Direct}_{\text{Schema}}$, $\text{Table}_{\text{Only}}$, and $\text{SciEx}_{\text{RAG}}$ baselines. The uncalibrated system reaches ***0.740*** full source-supported record F1 versus ***0.404*** for the strongest baseline, while calibration raises accepted field accuracy from ***0.477*** to ***0.689***. A blinded Source-A/B protocol shows that strict mismatches against human references mix extraction errors with annotation conventions, acceptable variants, source silence, and ambiguity. Results support process-guided agentic extraction for this dense RCT evidence-record task. We release workflow, prompt, schema, evaluation, synthetic-fixture, and table-reproduction artifacts.

1 Introduction

Information extraction (IE) increasingly requires models to convert full documents into structured records. Scientific and professional documents often arrive as PDFs whose meaning is distributed across layout, tables, captions, notes, figures, and surrounding prose. Recent work tests this shift in universal long-document IE (Zhang et al., 2026), realistic PDF structure recovery (Li et al., 2025b), multimodal scientific-article extraction (Khaleghinejad et al., 2025), and document-level biomedical evidence extraction (Pronesti et al., 2025). In these settings, the extractor must recover both document structure and task-specific records.

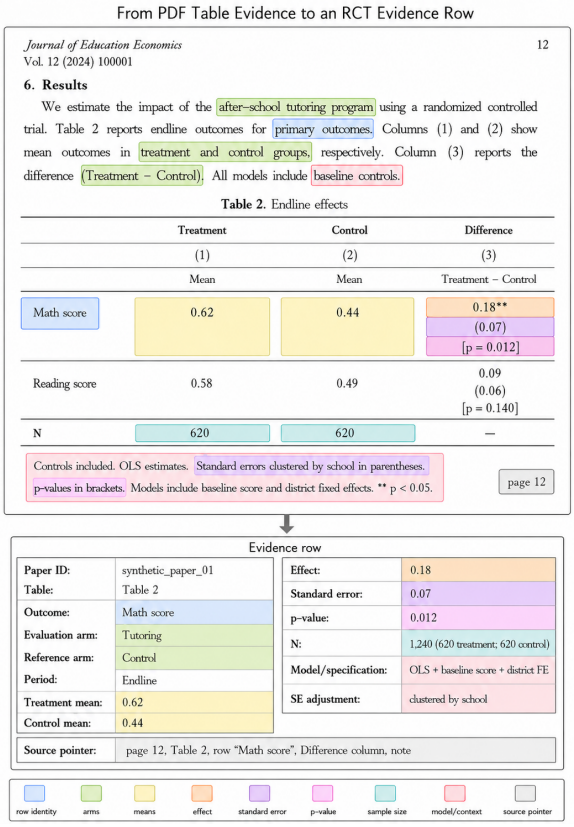

Figure 1: RCT evidence mapped to a structured record.

Randomized controlled trials (RCTs) are a concrete and consequential case. Their numerical evidence supports systematic reviews, meta-analysis, and policy-facing evidence synthesis, but one analyzable result is rarely expressed as an independent sentence or span. A single evidence record may combine a coefficient from one table cell, a standard error from a stacked line, a sample size from a separate table line, intervention and outcome labels from headers or captions, model details from notes, and standardization inputs from baseline or final-value tables (Figure 1). Prior work has studied RCT evidence inference (DeYoung et al., 2020), clinical finding extraction (Naik et al., 2024), and RCT numerical extraction (Yun et al., 2024). Our target

is more specific: PDF-to-RCT-evidence extraction in which the unit is a full-document, table-bound, n-ary evidence record with source pointers.

A straightforward baseline is schema-conditioned extraction: provide a schema, pass the paper or retrieved context to a model, and ask for JSON, as in retrieval-augmented scientific PDF extraction (Li et al., 2025a), PDF-to-JSON extraction (Ferguson et al., 2026), and schema-driven table extraction (Bai et al., 2024). For RCT evidence records, this is incomplete: table lines, columns, panels, notes, captions, and prose jointly define the evidence unit before fields can be filled, and source context determines whether values are supported, missing, ambiguous, outside reference-comparable scope, or derived from reported inputs. Even expert annotators rely on long-staged guidelines to decide which records count, how headers and notes bind to cells, when a field is source-silent, and which source-supported records belong in the export. This makes manual annotation slow and specialized, and motivates our question: can current agentic systems follow *process-level annotation guidelines* rather than only fill schema slots?

Evaluation creates a second difficulty. Strict reference agreement remains necessary for comparability, but an exact mismatch can collapse multiple phenomena into one error label. A mismatch may be a true extraction error, an acceptable numeric or textual variant, an annotation convention, source silence, ambiguity, or a source-supported record outside the reference scope. This is not only a metric problem: human reference annotations are indispensable, but they are themselves coding artifacts whose values may reflect policy decisions, scope choices, or underspecified source evidence. In this setting, evaluation must distinguish whether a disagreement comes from the extractor, the reference coding policy, an acceptable representation variant, or an underdetermined source. Similar concerns arise in structured PDF extraction, where field-specific metrics, array alignment, and missing-value semantics affect conclusions (Ferguson et al., 2026), and in event-argument evaluation, where semantic equivalence and exact-match limitations affect system conclusions (Fane et al., 2025). For RCT evidence extraction, evaluation should therefore report reference agreement while also auditing disagreements against the source.

We present RCT-Flow, a staged PDF-first workflow for extracting RCT evidence records from development-economics papers under explicit annotation guidelines. RCT-Flow builds semantic source objects from PDFs, inventories candidate result-bearing records, extracts structured records with source pointers, verifies source binding, repairs high-risk fields from source context only, and freezes reference-comparable exports before reference comparison. We pair these frozen outputs with strict reference metrics and blinded Source-A/B review, so evaluation reports both agreement with human references and source-grounded diagnostics for strict mismatches. Development examples expose implicit conventions that can be folded into later workflow variants before a frozen test run; they do not fine-tune weights or update the system during test extraction.

We ask three empirical questions: *i)* whether a process-guided agentic system improves evidence-record construction over schema-, table-, and RAG-style system baselines on the same ***15-paper*** test set; *ii)* how much development annotation is needed for calibration, using ***0***-, ***3***-, ***5***-, and ***10***-paper variants; and *iii)* what strict disagreements look like under a blinded Source-A/B protocol. We also report public checks on ExtractBench and Yun et al., where transfer is harder because the benchmarks do not provide the same process-level annotation workflow. Across these evaluations, we keep extraction source-only, freeze outputs before reference comparison, and use Source-A/B review to interpret strict disagreements against the PDF evidence.

This paper makes four contributions:

1. a process-guided agentic setup for PDF-first RCT evidence extraction that encodes staged human annotation procedures rather than only a schema;

2. a controlled system-level comparison with $\text{Direct}_{\text{Schema}}$, $\text{Table}_{\text{Only}}$, and $\text{SciEx}_{\text{RAG}}$ baselines on the same fixed test set;

3. a calibration-scaling study over ***0***-, ***3***-, ***5***-, and ***10***-paper development settings;

4. a blinded Source-A/B protocol for diagnosing whether strict mismatches reflect extraction error, reference policy, acceptable variants, source silence, or ambiguity.

We release a package with workflow, prompt, schema, synthetic-fixture, table-reproduction, and evaluation-protocol materials.

2 Related Work

RCT evidence extraction. Work on RCT extraction has moved from evidence-bearing text spans toward richer structured objects. Evidence Inference frames trial understanding around intervention-comparator-outcome prompts and evidence labels (DeYoung et al., 2020); later systems extract ICO frames (Wadhwa et al., 2023), clinical findings (Naik et al., 2024), document-level biomedical evidence (Pronesti et al., 2025), and numerical trial results (Yun et al., 2024). Yun et al. (2024) provide the closest public numerical benchmark, but their task is conditioned on intervention, comparator, and outcome information and uses abstract/results text rather than full-PDF discovery of table-bound evidence records. Our setting is therefore narrower in domain but harder in record-level binding: the extractor must recover record identity, estimates, uncertainty, sample sizes, model context, standardization inputs, and source pointers together.

Scientific PDF and schema extraction. Scientific IE increasingly treats papers as structured documents rather than collections of local spans: SciREX extracts document-level entities and relations (Jain et al., 2020), MatViX studies visually rich articles (Khalighinejad et al., 2025), and LitXBench evaluates structured experimental measurements (Chong and Colindres, 2026). PDF-focused work makes the same point from the document side: READoc treats structure recovery as extraction (Li et al., 2025b), schema-driven table IE maps heterogeneous tables to records (Bai et al., 2024), SciEx decomposes scientific PDF extraction into parsing, retrieval, extraction, and aggregation (Li et al., 2025a), and ExtractBench evaluates PDF-to-JSON extraction with field-specific scores, numeric tolerance, array alignment, and missing-value semantics (Ferguson et al., 2026). These works motivate our $\text{Direct}_{\text{Schema}}$, $\text{Table}_{\text{Only}}$, and $\text{SciEx}_{\text{RAG}}$ baselines. Their output contracts, however, usually specify what to extract rather than the staged annotation process for discovering evidence units, binding source context, handling source silence, repairing high-risk fields, and exporting reference-comparable records.

Evaluation beyond exact match. Exact span match can misidentify valid event-argument predictions as errors and can change model rankings (Fane et al., 2025). PDF-to-JSON evaluation must also handle numeric tolerance, alignment, and missing-value semantics (Ferguson et al., 2026). For RCT evidence records, a strict mismatch may be a true extraction error, a numeric variant, an annotation policy difference, source silence, or a source-supported record outside the reference scope. Human references remain the evaluation anchor, but they are not automatically a complete account of source fidelity. We therefore keep reference agreement as the main comparability metric and add a blinded A/B source review: reviewers adjudicate the system–reference pair against the cited PDF evidence.

Agentic document systems. Recent systems use multi-step or agentic decomposition for long and heterogeneous documents, including document agents (Sun et al., 2025), table agents (Cho et al., 2026), and grounded PDF RAG pipelines (Vaghefi et al., 2026). RCT-Flow is therefore not a new general agent architecture. It is a process-guided document-extraction workflow: agents turn the IDEAL-style coding protocol into steps for evidence-unit construction, source binding, field extraction, verification, repair, and frozen export.

3 Task and Method

We study process-guided agentic extraction by pairing a staged annotation workflow with calibration and source-grounded evaluation.

3.1 RCT Evidence-Record Extraction

We define **RCT evidence-record extraction**: given a scientific PDF, the system must output one structured record for each source-supported quantitative RCT result. Each record is an n-ary relation whose arguments include the paper, table, outcome, treatment arm, reference arm, analytical period, model specification, estimate, uncertainty, sample sizes, means/SDs, standardization inputs, derived effects, and source pointers. This differs from document-level schema filling because the evidence unit must first be discovered from table layout, headers, panels, notes, captions, and surrounding prose. Figure 1 illustrates this object, and Table 1 summarizes the main field families that must be bound to source evidence.

This target follows the IDEAL expert annotation workflow, not only a field list.[1] The protocol has three stages: *i)* establish study structure and candidate exhibits, including interven-

[1] https://ideal-consortium.github.io/codingprotocol.html

Family	Fields	Example value	Source
Record identity	paper, table, panel, outcome, arms, period	tutoring vs control, endline score	stub, headers, caption
Raw effect	coefficient, SE, CI, p-value	coef.=0.184, SE=0.072	numeric cell, parentheses, notes
Model context	model type, controls, preferred specification	OLS with baseline controls	column title, table note
Sample binding	arm N, total N, clusters	N=1,248	line, column, footer, note
Means / SDs	baseline/endline means, SDs, arm-level SEs	control mean=0.51	separate table or column
Derived effects	Hedges' g, g_{se}, standardization inputs	g=0.11	computed from source-bound inputs
Source pointers	page, table, cell, note, caption, text span	p.12, Table 3, Panel A, line 4	PDF layout and reconstructed table

Table 1: Synthetic RCT evidence-record anatomy. Records bind values, sources, and interpretation context.

tions, arms, outcomes, periods, comparisons, and specifications; *ii)* confirm eligible treatment effects by mapping outcome–contrast pairs to estimands, specifications, and data-collection rounds; and *iii)* code intervention, outcome, sampling, randomization, implementation, estimate, precision, and study-metadata fields with review checks. We encode these stages as prompts, skills, routing rules, source-binding checks, repair gates, and export policy, so the system uses them to decide what evidence to search for, which source to bind to each field, when to verify or repair an extraction, and which records to export.

3.2 Study Corpus and Calibration Protocol

Our study uses ***25*** annotated development-economics RCT papers. The main evaluation uses a fixed ***15-paper*** test set. Although small by document count, it is dense by extraction-unit count: it contains ***585*** reference evidence records and, depending on the calibrated variant, more than ***16k*** aligned record-field comparisons.

We evaluate one agentic pipeline under different amounts of annotation feedback. $\text{Agentic}_{\text{Uncal}}$ is the same pipeline before any development-paper calibration; $\text{Agentic}_{\text{3Dev}}$, $\text{Agentic}_{\text{5Dev}}$, and $\text{Agentic}_{\text{10Dev}}$ use ***3***, ***5***, and ***10*** annotated development papers, respectively. Here, calibration means updating prompts, extraction skills, annotation guidelines, schema/export policy, source-binding rules, repair rules, and normalizers; it does not fine-tune model weights or permit access to the ***15-paper*** test references before frozen export. We also report $\text{Agentic}_{\text{25Polished}}$ as an upper-bound diagnostic for a heavily refined pipeline; all ***25*** papers may inform that workflow, but its reported metrics still use the same ***15-paper*** scoring view, so it is not a clean held-out estimate. We score two scopes: **source-scope** extraction, which keeps source-supported candidate records, and **reference-comparable** extraction, which applies a source-only export policy before reference comparison.

3.3 Process-Guided Agentic Pipeline

Figure 2 shows RCT-Flow, an agentic pipeline that operationalizes the human RCT annotation workflow. It turns each PDF into source artifacts, plans evidence units, extracts and source-binds fields, verifies and repairs outputs, and freezes reviewed records for evaluation.

⓪ Workflow setup and guideline operationalization. Before a per-paper extraction run, we convert the annotation workflow and evidence-record schema into an initial extraction setup: prompts, initial extraction skills, routing rules, source-binding checklists, repair gates, normalizers, and reference-comparable export policy. This setup defines what the agents should do, which source evidence they must preserve, and which outputs must exist before validation. It does not inspect test references or fine-tune model weights.

① PDF parsing and source repair. This per-paper stage prepares the PDF as a source object. We render pages, run OCR/parsing, reconstruct layout and tables, and use an agent verifier to check structure consistency, table alignment, and span links. The output is a repaired source packet: page images, extracted text, reconstructed tables, captions, notes, and a Markdown source view. This stage matters because many RCT errors stem not from language generation but from missing table structure, stacked cells, split standard errors, or notes detached from the relevant record.

② Agentic evidence construction. The second per-paper stage is the main source-only extraction

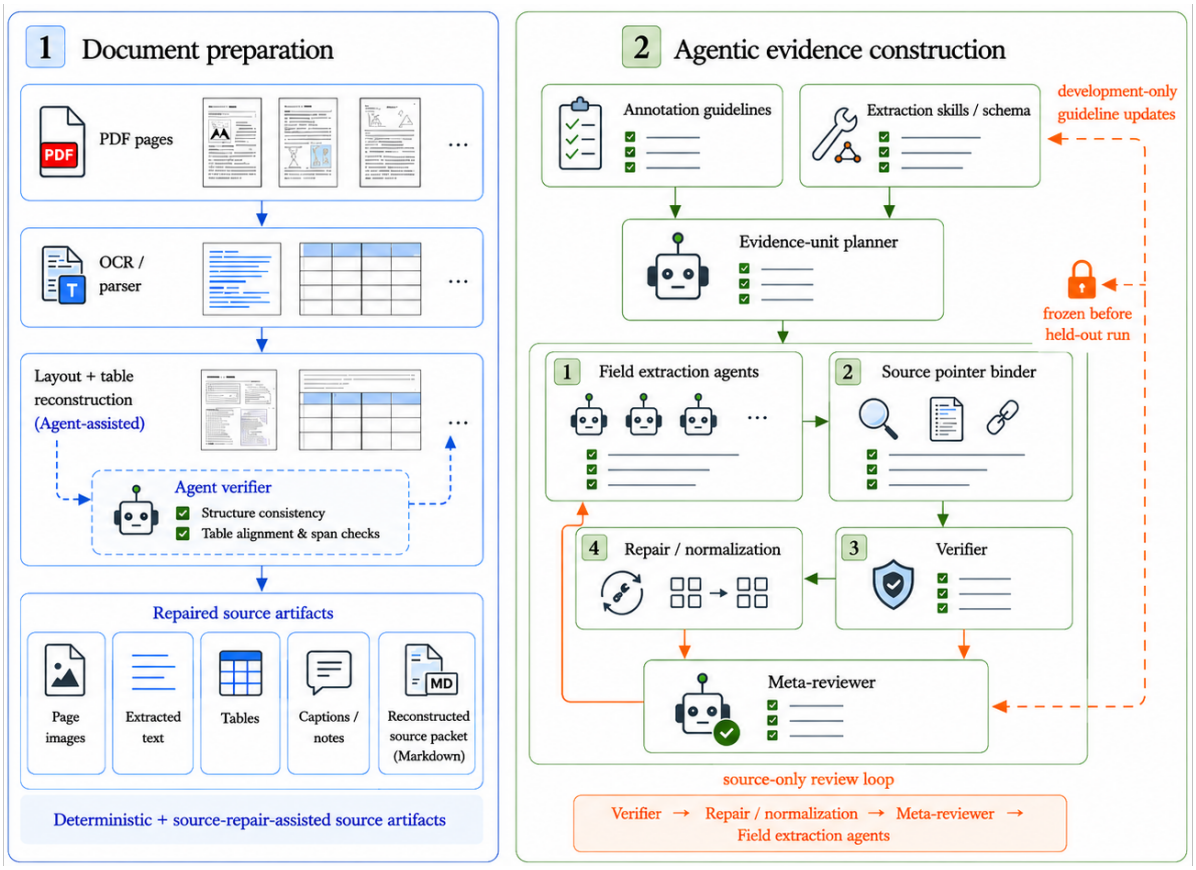

Figure 2: RCT-Flow pipeline. PDFs are repaired into source artifacts, then a process-guided agentic workflow plans evidence units, extracts fields, binds source pointers, verifies support, repairs high-risk outputs, and meta-reviews records before frozen export. Development-set diagnostics may update later frozen variants, but not the current frozen output.

loop. Starting from the repaired source packet, workflow guidelines, schema, and extraction skills, *i) an evidence-unit planner* inventories candidate result-bearing tables, panels, evidence units, and contrasts before any fields are filled; *ii) parallel field-extraction agents* recover record identity, estimates, uncertainty, sample sizes, model context, standardization inputs, and optional reporting fields; *iii) a source pointer binder and verifier* attach values to source handles and check whether each record and value is supported by the cited PDF evidence; and *iv) repair/normalization agents and a meta-reviewer* revisit weak fields, reject unsupported records, accept supported records, or route cases back to field extraction or repair. During development runs, meta-review and post-freeze mismatch analysis can also record reusable learning events, such as missing source-binding checks, ambiguous normalization conventions, weak-field routines, or export-policy gaps. These events may update skills, schemas, prompts, or guidelines for a later frozen variant. During a test run, the loop uses the PDF and source artifacts, but not the reference annotations, and it cannot alter the frozen export.

③ **Frozen export and evaluation.** The final stage separates extraction from evaluation. *i) Frozen export* writes source-scope evidence records, source pointers, a reference-comparable export, and audit packets with sources, logs, intermediates, and metadata. *ii) Reference comparison* then runs record alignment and field scoring against human reference annotations only after the export is locked. *iii) Source-fidelity review* audits strict disagreements against the PDF source, as described next. Development papers may use these post-freeze diagnostics to update later prompts, skills, schemas, or guidelines; held-out/test papers use them only for reporting and error analysis, never to change the scored output.

3.4 Source-A/B Review Protocol

Strict field disagreement is not always an extraction error. After frozen export and deterministic field comparison, sampled strict mismatches are

Label	Meaning
Source favors extraction	The PDF supports the extracted value over the reference value.
Source favors reference	The PDF supports the reference value over the extracted value.
Acceptable variant	Both values are defensible under the source and coding policy.
Policy / scope difference	The mismatch follows an annotation or export convention.
Source silent / optional	The field is absent, optional, or not recoverable from the source.
Ambiguous	The source is insufficient for a decisive preference.

Table 2: Source-A/B diagnostic labels. Labels explain strict mismatches; they do not replace agreement.

converted into blinded Source-A/B packets. Each packet contains the focal field, record context, indexed PDF evidence, and two candidate values, with extraction/reference provenance hidden where feasible. Automatic labels may be used for triage and packet construction, but the reported Source-A/B audit relies on human review. Table 2 lists the diagnostic labels. On development papers, recurring labels can become learning events for later workflow versions. On test papers, labels are diagnostic only and cannot change the frozen output.

4 Experiments

4.1 Experimental Setup

Task and data. The main task is PDF-to-RCT evidence-record extraction on the ***25-paper*** study corpus described above. The study-corpus calibration comparison evaluates all systems on the same fixed ***15-paper*** test set. Development annotation is used only for calibrating source-only extraction behavior: prompts, agent skills, guidelines, schema/export policy, source-binding rules, repair rules, and normalization. Development references can inform later workflow versions only after their exports are frozen; test references are never available before frozen export.[2] We also run two public stress tests: *i)* ExtractBench (Ferguson et al., 2026), a general PDF-to-JSON benchmark, and Yun's (Yun et al., 2024) RCT numerical extraction task , which is closer in domain but uses prespecified intervention, comparator, and outcome inputs rather than full record discovery.

Systems. For current-method context, we evaluate three baseline harnesses: $\text{Direct}_{\text{Schema}}$, a one-pass schema-prompting extractor following text-to-structure IE work (Lu et al., 2022); $\text{Table}_{\text{Only}}$, which follows schema-driven table extraction by using table cells and captions without full document context (Bai et al., 2024); and $\text{SciEx}_{\text{RAG}}$, a local SciEx-style RAG schema extractor implemented with the same target schema and a matched source-only extraction harness (Li et al., 2025a). This is not an execution of the official SciEx system. We score these frozen outputs on the same ***15-paper*** test set. For calibration scaling, we use the remaining ***10*** papers as development data and compare four frozen agentic variants: no development calibration ($\text{Agentic}_{\text{Uncal}}$), calibration on 3 papers ($\text{Agentic}_{\text{3Dev}}$), calibration on 5 papers ($\text{Agentic}_{\text{5Dev}}$), and calibration on all 10 development papers ($\text{Agentic}_{\text{10Dev}}$). We also report an all-paper polished diagnostic ($\text{Agentic}_{\text{25Polished}}$), where all ***25*** annotated papers may inform the workflow. Its reported metrics are still computed on the same ***15-paper*** scoring view, so this is not a clean held-out estimate.

Models and evaluation. Release artifacts list provider/model identifiers, decoding settings, source-only boundaries, and frozen aggregate outputs where available. We report frozen system outputs rather than a model-family ablation: baselines, agentic variants, and historical diagnostic outputs differ in workflow structure and are therefore not interpreted as evidence about one provider or model alone. Reference metrics are record precision, recall, and F1, accepted field accuracy on aligned records, and paper-clustered uncertainty estimates. Field accuracy counts an extracted field as accepted when it matches the human reference annotation exactly, after normalization, or within numeric tolerance. We report record alignment in two views: **reference-comparable**, where source-only export policy filters predictions to records comparable with the human reference annotations, and **source-supported inventory**, where source-supported records outside the reference-comparable scope are retained. Record alignment is used only for evaluation. Source-fidelity review is reported as diagnostic evidence, not as an adjusted score. All systems use the same test set, target schema, normalizers, record-alignment procedure, and gold-comparable filter. Because the baselines and agentic pipeline differ in inputs and runtime structure, the comparison is at the system level, not an isolated test of a single component such as agent orchestration.

[2] Appendix B defines the calibration and leakage boundary.

System	Type	Calibration	Source-supported			Reference-comparable			Field acc.
			P	R	F1	P	R	F1	
Current-method baselines									
$\text{Direct}_{\text{Schema}}$	Prompt	0 papers	0.3756	0.4359	0.4035	0.9813	0.4479	0.6150	0.3044
$\text{Table}_{\text{Only}}$	Table-only	0 papers	0.4981	0.2274	0.3122	0.9784	0.2325	0.3757	0.3071
$\text{SciEx}_{\text{RAG}}$	RAG	0 papers	0.3519	0.2598	0.2989	0.9345	0.2684	0.4170	0.3098
Agentic calibration ladder									
$\text{Agentic}_{\text{Uncal}}$	Agentic	0 papers	***0.6128***	***0.9333***	***0.7398***	***0.9894***	***0.9538***	***0.9713***	0.4765
$\text{Agentic}_{\text{3Dev}}$	Agentic	3 papers	0.5878	0.8923	0.7088	0.9835	0.9162	0.9487	0.6887
$\text{Agentic}_{\text{5Dev}}$	Agentic	5 papers	0.5991	0.9094	0.7223	0.9801	0.9282	0.9535	0.6875
$\text{Agentic}_{\text{10Dev}}$	Agentic	10 papers	0.5906	0.8974	0.7123	0.9572	0.9179	0.9372	***0.6895***
$\text{Agentic}_{\text{25Polished}}$	Agentic	All 25 diag.	*0.6510*	*0.9915*	*0.7859*	*0.9983*	*0.9949*	*0.9966*	*0.6669*

Table 3: Main results on the fixed ***15-paper*** test set, including $\text{Agentic}_{\text{25Polished}}$. Record P/R/F1 are shown first for the source-supported inventory and then for reference-comparable export. Field acc. is micro accepted field accuracy on aligned records. ***Bold italic*** marks the best non-diagnostic value. $\text{SciEx}_{\text{RAG}}$ is local SciEx-style, not official SciEx. $\text{Agentic}_{\text{25Polished}}$ may use all ***25*** papers for workflow polishing and is not selected as best. This is a system-level comparison, not an isolated estimate of agent orchestration alone.

4.2 Main Results

Annotation workflow guidance. Table 3 shows that agents following the annotator workflow produce far more source-supported evidence records than $\text{Direct}_{\text{Schema}}$, $\text{Table}_{\text{Only}}$, and $\text{SciEx}_{\text{RAG}}$. In the source-supported view, the uncalibrated agentic system reaches ***0.7398*** record F1, vs. ***0.4035*** for $\text{Direct}_{\text{Schema}}$, the strongest current-method baseline. This supports the claim that a staged agentic workflow benefits from explicit process-level guidelines when the workflow, source artifacts, and export policy are specified. The reference-comparable view is higher for all systems because the source-only export policy filters the record universe before reference comparison, but the relative gap remains large.

Fast calibration from few papers. Calibration affects a different aspect of the task: using three development papers raises field accuracy from ***0.4765*** to ***0.6887***, while five or ten papers yield only minor further gains. Field-family analysis supports an annotation-convention explanation: the written workflow already enables broad record discovery, while a few development papers surface convention-heavy field, normalization, source-binding, and export-policy choices implicit in annotator practice. This also explains why field accuracy jumps while source-supported record F1 does not monotonically improve: calibration can make the system more conservative about record inclusion. Process-guided extraction thus benefits from annotated examples, but most field-accuracy gains appear after only a few development papers.[3]

Why errors remain. $\text{Agentic}_{\text{25Polished}}$ estimates the upper bound of a heavily refined workflow on the same ***15-paper*** scoring view: allowing all ***25*** papers to inform the process raises source-supported record F1 to ***0.7859***, but field accuracy remains imperfect. Remaining strict mismatches may reflect true extraction failures, underspecified conventions, ambiguous evidence, source-silent fields, or human reference errors. The source-fidelity audit below distinguishes these cases instead of treating all strict mismatches as the same error type.

Why field accuracy lags record F1. The gap between record F1 and field accuracy is expected in this task and should not be interpreted as pure source-fidelity failure. Record matching asks whether the system found the correct table-bound evidence unit. Field accuracy is much stricter: it scores many convention-heavy and optional cells against a human reference annotation set after deterministic normalization. Disagreements often involve model/specification labels, adjusted-SE conventions, p-values encoded through table notes or stars, arm-level means and sample sizes bound through multi-level headers, standardization inputs, and derived standardized effects. Some are true extraction or binding errors; others reflect acceptable label variants, reference/export policy, source-silent optional fields, or ambiguous source evidence. The Source-A/B diagnostics are therefore not a replacement for strict field accuracy, but

[3] Appendix C reports per-field-family coverage.

Benchmark	System	Main score	Field/err.
ExtractBench	$\text{Direct}_{\text{Schema}}$	***0.3320***	***0.4040***
ExtractBench	$\text{Agentic}_{\text{Generic}}$	0.2269	0.2874
ExtractBench	$\text{Agentic}_{\text{Adapted}}$	0.2057	0.2818
Yun binary	$\text{Direct}_{\text{Schema}}$	0.5263	0.1127
Yun binary	$\text{Agentic}_{\text{Adapted}}$	***0.5965***	***0.1069***
Yun continuous	$\text{Direct}_{\text{Schema}}$	0.3629	0.3835
Yun continuous	$\text{Agentic}_{\text{Adapted}}$	***0.4041***	***0.2904***

Table 4: Public stress tests. ExtractBench reports mean overall / field score; Yun reports exact extraction / effect error, where lower error is better.

Item	Value
Field	`Coefficientvalue`
Extraction	0.000227
Reference	0.00027
Strict score	Mismatch
Source-A/B	Source favors extraction
Evidence	Indexed coefficient-cell evidence supports 0.000227; no indexed evidence in the packet supports 0.00027.

Table 5: Redacted Source-A/B example. Strict scoring marks the field wrong; source review favors extraction.

an interpretation layer explaining what strict mismatches mean.

4.3 Public Stress Tests

The public stress tests qualify the private-corpus result. On ExtractBench, agentic adapters underperform direct schema prompting, showing that agentic orchestration alone does not guarantee better PDF-to-JSON extraction. On Yun, the guideline-aware adapter improves binary and continuous extraction relative to our direct-schema adapter. We read this pattern as task-specific transfer: process-guided agents work best when the task provides a staged annotation workflow, not just a schema.[4]

4.4 Source-Fidelity Diagnostics

The Source-A/B and repair diagnostics use the ***25-paper*** diagnostic set, not the ***15-paper*** main scoring denominator. This diagnostic pass yields ***15,267*** field-level Source-A/B packets, from which we human-audit ***711***; it also audits ***577*** dropped source-scope records and logs ***90,412*** source-context repair decisions. Tables 5 and 6 show why strict reference agreement needs source-grounded interpretation. The example in Table 5 is scored as wrong by strict field comparison because the extraction gives ***0.000227*** and the human reference annotation gives ***0.00027***; blinded source review

Diagnostic outcome	Count	Share
Source favors extraction value	119	16.7%
Source favors reference value	194	27.3%
Acceptable variant / policy-equivalent	155	21.8%
Source silent / optional	116	16.3%
Ambiguous / insufficient evidence	127	17.9%
Dropped source-scope records		
Export-policy exclusion	369	64.0%
Source-supported outside scope	195	33.8%
Duplicate / near-duplicate	12	2.1%
Insufficient support	1	0.2%

Table 6: Source-fidelity diagnostics from the ***25-paper*** audit.

favors the frozen extraction because the indexed coefficient cell supports ***0.000227***. Across the human-audited packets, ***16.7%*** of strict mismatches favor the extraction, ***21.8%*** are acceptable variants or policy-equivalent, and ***34.2%*** are source-silent or ambiguous. The agent is still wrong in many cases: ***27.3%*** favor the reference. But the audit shows that strict field mismatch is not synonymous with extraction failure. Records removed before reference-comparable export show the same pattern: most are policy-excluded or source-supported outside scope, not unsupported hallucinations. These diagnostics do not correct the main scores; they explain what the strict scores compress.[5]

5 Conclusion

We framed quantitative RCT extraction from PDFs as document-level scientific IE over table-bound evidence records, rather than flat schema filling. RCT-Flow operationalizes a human annotation workflow through evidence-unit planning, source-bound field extraction, verification, repair, export policy, and frozen evaluation. On a shared ***15-paper*** test set, this process-guided agentic pipeline improves record construction over direct schema prompting, table-only extraction, and a local SciEx-style RAG baseline; a few development papers mainly improve convention-heavy field completion. Source-A/B diagnostics show that strict mismatches mix extraction errors with annotation policy, acceptable variants, source silence, and ambiguity. The contribution is therefore not a new general architecture for scientific IE, but evidence that explicit annotation workflows, source binding, and frozen source-aware evaluation can make agentic extraction effective for a dense RCT evidence-record contract.

[4] Appendix E explains the public benchmark scope.

[5] Appendix D expands Source-A/B details.

6 Limitations

The strongest empirical claim is internal to one development-economics RCT evidence-record corpus. The fixed ***15-paper*** test set supports a controlled within-corpus comparison, but it does not establish broad state of the art for scientific IE. The baseline systems are matched source-only harnesses rather than independent production systems, and the experiment does not isolate which pipeline stage accounts for the observed gains.

The approach is domain-specialized. It assumes an RCT evidence-record schema, field-specific coding rules, process-level annotation guidelines, and source packets that preserve table layout, notes, captions, and page provenance. Errors in OCR, table recovery, note binding, or guideline coverage can therefore propagate through the pipeline. The ***25-paper*** polished condition is an upper-bound diagnostic, not a clean held-out estimate; the ***3***-, ***5***-, and ***10***-paper calibration variants are the appropriate calibration-scaling comparisons.

The public benchmark evidence is limited. ExtractBench is a general PDF-to-JSON benchmark and is negative for our current agentic adapters. Yun et al. is closer to RCT extraction, but it uses intervention-comparator-outcome conditioning and abstract/results text rather than full-PDF record discovery. These results support a narrow task-specific claim rather than broad transfer.

The implementation depends on closed frontier models, incomplete operational telemetry, and private source documents. The release package records provider/model identifiers, run-configuration notes, aggregate outputs, prompts, schemas, rubrics, redacted examples, synthetic fixtures, and reproduction scripts where available, but raw PDFs and human reference annotations cannot be released. Cost comparisons are therefore incomplete, and reproducibility is limited to released aggregate artifacts and table-generation scripts.

Finally, the evaluation views and audits should not be conflated. The source-supported inventory and reference-comparable export answer different questions, and record alignment can be ambiguous. Source-A/B and source-scope audits diagnose disagreement types but do not replace reference agreement; automatic labels are triage, and human review was performed by project-internal reviewers rather than an external annotator population. Appendix F gives artifact, human-review, compute, and AI-assistance details.

References

Fan Bai, Junmo Kang, Gabriel Stanovsky, Dayne Freitag, Mark Dredze, and Alan Ritter. 2024. Schema-driven information extraction from heterogeneous tables. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10252–10273. Association for Computational Linguistics.

Nicole Cho, Kirsty Fielding, William Watson, Sumitra Ganesh, and Manuela Veloso. 2026. TASER: Table agents for schema-guided extraction and recommendation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 226–252. Association for Computational Linguistics.

Curtis Chong and Jorge Colindres. 2026. LitXBench: A benchmark for extracting experiments from scientific literature. *Preprint*, arXiv:2604.07649.

Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C. Wallace. 2020. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132. Association for Computational Linguistics.

Enfa Fane, Md Nayem Uddin, Oghenevovwe Ikumariegbe, Daniyal Kashif, Eduardo Blanco, and Steven Corman. 2025. BEMEAE: Moving beyond exact span match for event argument extraction. 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)*, pages 5734–5749, Albuquerque, New Mexico. Association for Computational Linguistics.

Nick Ferguson, Josh Pennington, Narek Beghian, Aravind Mohan, Douwe Kiela, Sheshansh Agrawal, and Thien Hang Nguyen. 2026. ExtractBench: A benchmark and evaluation methodology for complex structured extraction. *Preprint*, arXiv:2602.12247.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7506–7516. Association for Computational Linguistics.

Ghazal Khalighinejad, Sharon Scott, Ollie Liu, Kelly L. Anderson, Rickard Stureborg, Aman Tyagi, and Bhuwan Dhingra. 2025. 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)*, pages 3636–3655. Association for Computational Linguistics.

Sha Li, Ayush Sadekar, Nathan Self, Yiqi Su, Lars Andersland, Mira Chaplin, Annabel Zhang, Hyoju Yang, James B. Henderson, Krista Wigginton, Linsey Marr, T. M. Murali, and Naren Ramakrishnan.

2025a. Exploring LLMs for scientific information extraction using the SciEx framework. *Preprint*, arXiv:2512.10004. Accepted to the KGML Bridge at AAAI 2026 (non-archival).

Zichao Li, Aizier Abulaiti, Yaojie Lu, Xuanang Chen, Jia Zheng, Hongyu Lin, Xianpei Han, Shanshan Jiang, Bin Dong, and Le Sun. 2025b. READoc: A unified benchmark for realistic document structured extraction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21889–21905. Association for Computational Linguistics.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772. Association for Computational Linguistics.

Aakanksha Naik, Bailey Kuehl, Erin Bransom, Doug Downey, and Tom Hope. 2024. CARE: Extracting experimental findings from clinical literature. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4580–4596. Association for Computational Linguistics.

Massimiliano Pronesti, Joao H. Bettencourt-Silva, Paul Flanagan, Alessandra Pascale, Oisín Redmond, Anya Belz, and Yufang Hou. 2025. 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)*, pages 28034–28051. Association for Computational Linguistics.

Li Sun, Liu He, Shuyue Jia, Yangfan He, and Chenyu You. 2025. DocAgent: An agentic framework for multi-modal long-context document understanding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17701–17716. Association for Computational Linguistics.

Saeid Vaghefi, Aymane Hachcham, Veronica Grasso, Nakiete Msemo, Chiara Colesanti Senni, and Markus Leippold. 2026. AI for climate finance: Agentic retrieval and multi-step reasoning for early warning system investments. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 492–511. Association for Computational Linguistics.

Somin Wadhwa, Jay DeYoung, Benjamin Nye, Silvio Amir, and Byron C. Wallace. 2023. Jointly extracting interventions, outcomes, and findings from RCT reports with LLMs. In *Proceedings of Machine Learning for Health*.

Hye Sun Yun, David Pogrebitskiy, Iain James Marshall, and Byron C. Wallace. 2024. Automatically extracting numerical results from randomized controlled trials with large language models. In *Proceedings of the 9th Machine Learning for Healthcare Conference*, volume 252 of *Proceedings of Machine Learning Research*. PMLR.

Shuyi Zhang, Zhenbin Chen, Shuting Li, Kewei Tu, Li Jing, Zixia Jia, and Zilong Zheng. 2026. MMUIE: Massive multi-domain universal information extraction for long documents. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 6338–6370. Association for Computational Linguistics.

A Task, Schema, and Human Workflow

RCT evidence-record extraction is not flat schema filling. The system must first identify treatment-effect evidence units, then bind record fields to source context and annotation policy. The release package includes field definitions and source-binding rules; here we summarize the process-level workflow that makes the task difficult.

The agentic workflow follows the public IDEAL coding protocol for impact-evaluation RCTs, which describes staged coding rather than direct field entry. Stage 1 establishes study and experiment structure: intervention arms, comparison structure, outcomes, time periods, exhibits, and design features. Stage 2 locates treatment effects by matching outcomes, arm comparisons, and specifications to reported results in tables, figures, appendices, and prose. Stage 3 fills detailed experiment, intervention, outcome, sample, and treatment-effect fields, with quality checkpoints between stages. We encode this process as prompts, skills, routing rules, source-binding checks, repair gates, and export policy; development papers can reveal missing conventions for later frozen variants, but test references remain unavailable until after frozen export.

B Calibration and Evaluation Protocol

The protocol below defines what information can shape a calibrated variant, how records are aligned, what is scored, and what comparison boundaries apply. Development references can only influence later workflow versions after the development export is frozen. Test references are unavailable during test extraction and cannot change frozen test outputs.

B.1 Calibration Boundary

During development calibration, source PDFs and source packets are available, and human reference annotations may shape later workflow versions only after the corresponding development export is frozen. Allowed changes are prompts, skills, guidelines, schema/export policy, source-binding, repair, and normalization; model weights are never fine-tuned.

During frozen test extraction, only source PDFs, source packets, and fixed workflow artifacts are available. Test references are unavailable and cannot change frozen test outputs. Post-freeze evaluation uses the frozen outputs for metrics, record alignment, Source-A/B packet construction, and error analysis. $\text{Agentic}_{\text{25Polished}}$ remains diagnostic because all ***25*** papers may inform workflow polishing, although it is reported on the same ***15-paper*** scoring view.

Scope	System	Match	Unm. ref.	Unm. pred.	Ambig.
Gold	$\text{Agentic}_{\text{Uncal}}$	558	21	6	6
Gold	$\text{Agentic}_{\text{3Dev}}$	536	31	9	18
Gold	$\text{Agentic}_{\text{5Dev}}$	543	16	11	26
Gold	$\text{Agentic}_{\text{10Dev}}$	537	13	24	35
Gold	$\text{Agentic}_{\text{25Polished}}$	582	0	1	3
Full	$\text{Agentic}_{\text{Uncal}}$	546	3	345	36
Full	$\text{Agentic}_{\text{3Dev}}$	522	4	366	59
Full	$\text{Agentic}_{\text{5Dev}}$	532	6	356	47
Full	$\text{Agentic}_{\text{10Dev}}$	525	4	364	56
Full	$\text{Agentic}_{\text{25Polished}}$	580	0	311	5

Table 7: Record-alignment counts for the finalized ***15-paper*** calibration-scaling comparison. *Unm.* abbreviates unmatched.

B.2 Metrics and Record Alignment

We report record precision, recall, and F1 in two scopes. The *reference-comparable* view scores records after source-only export policy filters predictions to the reference-comparable record universe. The *source-supported inventory* view scores a broader source-supported inventory; extra records in this view are not automatically hallucinations. Accepted field coverage is computed after evaluation-stage record alignment and type-aware field normalization. Source-A/B labels are diagnostic and never correct the main scores. Table 7 reports record-alignment denominators for the finalized comparison.

B.3 Comparison Boundary

All main comparisons use the same fixed ***15-paper*** test set, target schema, record-alignment procedure, normalizers, and gold-comparable filter. We interpret differences between baselines, clean agentic variants, and $\text{Agentic}_{\text{25Polished}}$ as system-level comparisons; $\text{Agentic}_{\text{25Polished}}$ remains diagnostic because all ***25*** papers may inform workflow polishing.

C Additional Results for the Calibration Claim

The calibration results show that a few development papers mainly improve convention-heavy fields rather than record discovery. Accepted field accuracy is the micro average over aligned record-field comparisons: a field counts when the extracted value matches the human reference annotation exactly, after normalization, or within numeric tolerance.

The field-family breakdown makes the effect concrete. In the gold-comparable view, moving from $\text{Agentic}_{\text{Uncal}}$ to $\text{Agentic}_{\text{3Dev}}$ barely changes raw effect reporting (***0.8943*** to ***0.8868***), which was already strong. It sharply improves convention-heavy fields: model/specification context rises from ***0.3738*** to ***0.6621***, effect-parameter conventions from ***0.4587*** to ***0.7848***, means/SDs/arm-level SEs from ***0.0000*** to ***0.5618***, and derived standardized effects from ***0.2390*** to ***0.5774***. These fields depend on source-binding policy, normalization, and implicit annotator conventions. Table 8 gives the field-family breakdown.

The same pattern helps explain why field accuracy lags record F1. Many residual errors involve a correct evidence unit but incomplete binding of model labels, uncertainty conventions, standardization inputs, or optional reference-policy fields.

Field dimension	$\text{Agentic}_{\text{Uncal}}$	$\text{Agentic}_{\text{3Dev}}$	$\text{Agentic}_{\text{5Dev}}$	$\text{Agentic}_{\text{10Dev}}$	$\text{Agentic}_{\text{25Polished}}$
Core record identity	0.4395	0.5236	0.5248	0.5214	0.6280
Raw effect reporting	0.8943	0.8868	0.8776	0.8866	0.8538
Model/specification context	0.3738	0.6621	0.6750	0.6765	0.4700
Effect-parameter conventions	0.4587	0.7848	0.7833	0.7851	0.6998
Adjusted-SE context	0.0492	0.1821	0.1828	0.1809	0.1843
Sample-size binding	0.8119	0.8672	0.8578	0.8638	0.8508
Means, SDs, and arm-level SEs	0.0000	0.5618	0.5495	0.5603	0.5598
Derived standardized effects	0.2390	0.5774	0.5688	0.5710	0.5916
Other/unclassified	0.4964	0.4963	0.4963	0.4963	0.8029

Table 8: Gold-comparable field-dimension coverage. Calibration mainly improves model/specification context, effect-parameter conventions, means/SDs/arm-level SEs, and derived standardized effects.

Quantity	Count
Broad source-scope records	1,772
Reference-comparable extracted records	1,195
Reference records	1,210
Matched records	1,192
Ambiguous alignment records	5
Unmatched reference records	13
Unmatched extracted records	3
Field-level Source-A/B packets	15,267
Human-audit packets	711
Source-scope records audited	577

Table 9: Source-A/B denominator flow for the ***25-paper*** audit. Records, aligned records, packets, and audit samples are different units.

Comparison	Agreement	κ
Methods vs source binding	71.6%	0.649
Methods vs statistical extraction	98.5%	0.981
Source binding vs statistical extraction	70.3%	0.635
Human consensus vs automatic labels	43.2%	–
Reviewer-conflict packets	29.8%	–

Table 10: Human reliability for the ***711-packet*** Source-A/B audit. Automatic labels are triage, not final adjudication.

D Source-A/B Audit Details

Source-A/B review audits strict disagreements. Each packet contains the focal field, record context, field family, two blinded candidate values, and indexed source evidence such as table cells, captions, notes, pages, or text spans. The hidden key records whether candidate A or B came from the extraction or the human reference annotations; reviewers do not see that provenance. Tables 9, 10, and 11 report the packet denominator flow, human reliability, and audited field groups.

Field group	Pkts.	A/H agr.	Rev. conf.
Adjusted SE value	38	0.0%	2
Estimation parameter	24	79.2%	3
Model/specification	105	29.5%	22
P-values and CIs	43	39.5%	1
Means, SDs, SEs, Ns	247	67.6%	98
Derived Hedges' g / g_{se}	39	17.9%	20
Other fields	215	30.7%	66

Table 11: Priority field groups in the Source-A/B human audit. Agreement varies by field family, with model/specification, adjusted-SE, and derived-effect cases requiring more human interpretation. Pkts. = packets, A/H agr. = automatic-human agreement, Rev. conf. = reviewer conflicts.

E Public Benchmark Adapters

The public benchmarks are stress tests rather than the main evidence for the paper. ExtractBench tests general PDF-to-JSON extraction and does not require RCT evidence-record discovery or source-pointer binding. Yun et al. is closer to RCT numerical extraction, but its input contract conditions on intervention-comparator-outcome fields and does not require end-to-end full-PDF record construction. Full public-adapter outputs, benchmark-specific adapter policies, evaluator status notes, and aggregate TEST metrics are release-package material rather than primary appendix evidence.

F Artifact Release and Responsible Research

Artifacts, intended use, and release terms. This work uses and creates scientific artifacts: private RCT evidence-record annotations, PDF-derived extraction outputs, public benchmark adapters, redacted prompt templates, schemas, aggregate result files, evaluation pseudocode, source-fidelity rubrics, synthetic fixtures, and table-reproduction scripts. The artifact package records split mani-

fests, release-safe run-configuration notes where available, record-alignment pseudocode, source-fidelity review guidelines, and diagnostic historical records. Public external artifacts are cited in the main text, including ExtractBench and Yun et al.; the IDEAL public coding protocol is cited as the process-level annotation guidance behind the RCT evidence-record task. The anonymous release package is intended to support research review, aggregate table reproduction, and inspection of the evaluation protocol. It is not intended for deployment as an automated replacement for expert systematic-review extraction or policy evidence synthesis. The paper and non-code supplementary materials are released under CC BY 4.0; script licensing is specified in the artifact package.

Private data and anonymization boundary. The private study corpus contains published scientific PDFs and human reference annotations, but the anonymous release does not redistribute raw PDFs, full PDF-derived text, page images, source packets, real reference workbooks, validation spreadsheets, hidden Source-A/B keys, reviewer packet payloads, credentials, provider account metadata, private paths, or internal dashboards. Private document identifiers are replaced with opaque IDs in split manifests. Aggregate metric files contain system IDs, split labels, counts, scores, and hashes, but no real row text or source excerpts. Synthetic fixtures are invented examples designed to exercise row alignment, numeric tolerance, field-family scoring, and source-fidelity labels without copying real source documents or annotations.

Human review protocol. The Source-A/B audit is a validation-stage diagnostic, not an extraction input. Frozen system outputs are first compared to human reference annotations; sampled strict disagreements are then converted into blinded packets containing the focal field, row context, indexed source evidence, and two candidate values. Candidate provenance is hidden from reviewers where feasible. Reviewers assign source-fidelity labels before any unblinding, and the hidden provenance key is used only afterward to aggregate labels as source-favoring extraction, source-favoring reference, acceptable variant, source-silent, or ambiguous. The release-safe reviewer guidelines specify row-level labels, field-level labels, confidence labels, source-evidence requirements, and the separation between source-fidelity review and gold-comparable review. The completed diagnostic audit contains ***711*** field-level packets and a ***577***-row source-scope audit, with pairwise agreement and conflict counts reported in Appendix D.

Reviewer recruitment and compensation. The human reviewers were project-internal lab members with relevant methodological, source-binding, or statistical-extraction expertise. They were not recruited through a crowdsourcing platform and were not paid separately per packet; the audit was performed as part of their normal lab work. Reviewer identities, individual packet payloads, hidden answer keys, and raw rationales are not released. We report only aggregate agreement, conflict, and label-count statistics.

Human-subjects and consent boundary. The paper does not report an intervention with study participants. Human review was used to audit scientific extraction artifacts and source support, not to collect personal, demographic, behavioral, or health data from reviewers or from people described in the RCT papers. The source documents are published scientific articles, and released materials exclude real packet payloads, private workbook cells, and reviewer identities. Because the review involved internal artifact validation, no personal data collection, and aggregate-only reporting, no separate IRB protocol was sought.

Computational setup and budget. No model weights are trained or fine-tuned. The experiments use frozen prompts, schemas, source-access policies, normalizers, row-alignment procedures, and benchmark evaluators. Release manifests record provider/model identifiers, temperature, reasoning settings, prompt hashes, source-only boundaries, and frozen-output hashes where available. For the Same15 agentic variants, inspected wrapper manifests record OpenAI GPT 5.5, temperature ***0.0***, reasoning disabled, and no gold access before freeze; closed-model parameter counts are not available. The runtime/cost release file reports token, cost, completion, and wall-time metadata where telemetry was preserved. Known tracked model usage is ***39.1M*** tokens and USD ***17.89***, excluding historical full-agentic telemetry that was not preserved. Cost comparisons are therefore operational metadata rather than a primary scientific result.

Evaluation packages and parameters. Private RCT scoring uses deterministic row alignment, field-specific normalization, numeric tolerance, and source-scope filtering. Public benchmark checks

use benchmark-specific adapters and official evaluator contracts where available. The anonymous release includes row-alignment pseudocode, normalizer policy, source-fidelity rubric, aggregate artifact hashes, and scripts that reproduce the paper tables from aggregate files only.

AI assistants. This paper studies LLM-based extraction systems as the object of research. AI assistants were also used by the authors for drafting code, prompts, prose, and submission metadata. All experiments, evidence links, tables, claims, and submitted text were reviewed by the authors. AI-generated text was not treated as source evidence, and reference annotations were not used to repair frozen test outputs.