Diagnosing Structural Failures in LLM-Based Evidence Extraction for Meta-Analysis

Zhiyin Tan[1,*], Jennifer D'Souza[2]

[1]L3S Research Center, Leibniz University Hannover, Hannover, Germany
[2]TIB Leibniz Information Centre for Science and Technology, Hannover, Germany

Abstract

Systematic reviews and meta-analyses rely on converting narrative articles into structured, numerically grounded study records. Despite rapid advances in large language models (LLMs), it remains unclear whether they can meet the structural requirements of this process, which hinge on preserving roles, methods, and effect-size attribution across documents rather than on recognizing isolated entities. We propose a structural, diagnostic framework that evaluates LLM-based evidence extraction as a progression of schema-constrained queries with increasing relational and numerical complexity, enabling precise identification of failure points beyond atom-level extraction. Using a manually curated corpus spanning five scientific domains, together with a unified query suite and evaluation protocol, we evaluate two state-of-the-art LLMs under both per-document and long-context, multi-document input regimes. Across domains and models, performance remains moderate for single-property queries but degrades sharply once tasks require stable binding between variables, roles, statistical methods, and effect sizes. Full meta-analytic association tuples are extracted with near-zero reliability, and long-context inputs further exacerbate these failures. Downstream aggregation amplifies even minor upstream errors, rendering corpus-level statistics unreliable. Our analysis shows that these limitations stem not from entity recognition errors, but from systematic structural breakdowns, including role reversals, cross-analysis binding drift, instance compression in dense result sections, and numeric misattribution, indicating that current LLMs lack the structural fidelity, relational binding, and numerical grounding required for automated meta-analysis. The code and data are publicly available at GitHub. [1]

Keywords

Evidence extraction, Large language models, Meta-analysis, Schema-grounded evaluation

1. Introduction

Systematic reviews and meta-analyses are foundational to evidence synthesis across the sciences, providing structured methods for integrating empirical results at scale [1, 2]. Despite advances in digital libraries and scholarly infrastructures, systematic review workflows remain dominated by manual data extraction, where study-level elements, such as populations, variables, measurement units, association methods, and effect sizes, must be identified, normalized, and encoded for analysis [3]. This step is highly labor-intensive and remains a major bottleneck in creating reproducible, machine-actionable meta-analytic datasets.

Parallel work in semantic publishing and structured scholarly knowledge has demonstrated that scientific studies can be represented through explicit schemas and ontologies. Examples range from publication-centric models such as SPAR [4], to domain-specific controlled vocabularies in biomedicine (e.g., MeSH [5], GO [6], UMLS [7]), and large-scale scholarly knowledge graphs such as Semantic Scholar's Literature Graph [8] and the Computer Science KG [9]. The Open Research Knowledge Graph (ORKG) [10] further shows how paper-level claims and methodological details can be captured via human-authored semantic templates. Large language models (LLMs) appear promising for automating parts of this process. They show strong performance on isolated extraction tasks and domain-adapted

[1]https://github.com/zhiyintan/LLM-Meta-Analysis

IRCDL 2026: 22nd Conference on Information and Research Science Connecting to Digital and Library Science, February 19-20, 2026, Modena, Italy

*Corresponding author.
zhiyin.tan@l3s.de (Z. Tan); jennifer.dsouza@tib.eu (J. D'Souza)
0009-0002-4166-5810 (Z. Tan); 0000-0002-6616-9509 (J. D'Souza)

scientific text processing [11, 12, 13, 14], and recent work has explored their use in systematic review screening and semi-automated data extraction pipelines [15, 16]. However, a growing body of evidence highlights limitations that are directly relevant to the structural demands of meta-analytic workflows. LLMs struggle with compositional generalization, the systematic combination of familiar elements into novel but valid structures [17, 18]. They also exhibit failures in entity and attribute binding, where recognized elements become misaligned or incorrectly associated [19, 20]. Long-context processing introduces further fragility, with models often ignoring or misusing information located in the middle of extended inputs [21]. Finally, numerically grounded reasoning remains brittle, particularly when values must be correctly tied to specific variables, units, or statistical methods [22, 23, 24]. Meta-analysis provides a particularly stringent testbed for these weaknesses. Constructing meta-analytic inputs requires extracting and binding multi-entity, numerically annotated, and method-conditioned tuples, such as population, independent variable, dependent variable, association method, sample size, and effect size, across multiple documents. Success therefore depends not only on recognizing individual entities, but on preserving their structural relationships under aggregation.

In this work, we take a structural and diagnostic perspective on LLM-based scientific evidence extraction. Recent work has shown that LLMs can be used for semi-automated evidence extraction in systematic reviews, particularly in biomedical settings [25]. Rather than proposing a new extraction model or end-to-end pipeline, we study how current LLMs behave when evidence extraction is framed as a sequence of increasingly structured, schema-constrained tasks. This framing allows us to explicitly control the relational and binding demands imposed on the model, ranging from isolated entity extraction to higher-order association structures characteristic of meta-analytic inputs. We further consider lightweight aggregation operations to probe whether extracted evidence can be coherently integrated at the corpus level. We instantiate this perspective using a generalized semantic study schema and evaluate it across five scientific domains under controlled multi-document settings. By combining full-text inputs with human-curated gold annotations, our evaluation enables a fine-grained analysis of how model behavior evolves as structural complexity increases.

This study **makes four main contributions**:

1. a structural and diagnostic perspective on LLM-based scientific evidence extraction, which frames extraction as a progression of increasingly constrained, schema-guided tasks rather than as an end-to-end pipeline problem;
2. a tiered query framework that operationalizes this perspective through controlled variations in tuple arity and relational binding complexity, covering both direct extraction and derived statistical reasoning;
3. an manually curated, multi-domain, schema-aligned evaluation benchmark spanning five scientific domains (civil engineering, medical and health science, agricultural science, earth and environmental sciences, and social science) designed to systematically investigate LLM performance under increasing structural demands;
4. a comprehensive empirical analysis identifying where along this structured progression modern LLMs begin to fail, revealing specific weaknesses in relational binding, numerical grounding, and document-level attribution.

Together, these contributions clarify the structural requirements of automated meta-analytic evidence extraction and provide a foundation for designing future LLM-based systems that better align with the needs of evidence synthesis and digital library infrastructure.

2. Related Work

Semantic Modeling and Structured Representations of Scientific Studies. Prior work in semantic publishing treats scientific articles as structured objects rather than purely narrative text, modeling key study components such as populations, variables, interventions, outcomes, and measurements using explicit schemas and ontologies [26, 27, 28]. This perspective enables machine-interpretable

access to scientific content beyond surface-level text. This idea has been instantiated at scale through ontology suites and scholarly knowledge graphs that formalize document structure, citation relations, and scientific statements [26, 4, 28, 29, 30]. The Open Research Knowledge Graph further demonstrates paper-level contribution modeling via human-authored schemas [31], while autonomously constructed graphs such as MatKG show that complex scientific entities and relations can be populated at corpus scale [32]. Recent work also explores LLM-assisted schema discovery and refinement [33]. In this setting, a central question is how such structured study representations can be reliably instantiated from unstructured scientific articles at scale when schemas require preserving multi-entity relations and numerically grounded attributes.

Automating Information Extraction in Systematic Reviews and Meta-Analyses. Information extraction (IE), often termed data or evidence abstraction in systematic reviews, encodes study-level information into structured forms for downstream synthesis [1, 34]. Owing to its cost and error-proneness, IE has been a long-standing target for automation. Most prior work addresses localized extraction problems, targeting individual elements or simple relations such as populations, interventions, outcomes, or trial characteristics using rule-based, classical machine learning, and neural methods [35, 36, 37, 38]. These approaches reduce manual effort but operate primarily on isolated fields. More recent systems integrate IE into broader review pipelines, including screening, eligibility assessment, and risk-of-bias evaluation, with LLMs increasingly explored as assistants for such tasks [39, 40, 34, 41, 42]. Across this literature, task formulations and evaluations remain aligned with low-arity outputs, rather than with the structured, multi-field records required as direct inputs to meta-analysis [43, 34]. In particular, few studies assess whether automated IE preserves the joint binding between populations, variables, methods, and numerically grounded effect measures across documents. We target this gap by examining LLM-based IE under progressively increasing structural constraints at the pre–meta-analytic stage.

LLMs for Scientific Information Extraction and Multi-Document Reasoning. Recent work applies LLMs to scientific information extraction (SciIE), including schema-driven extraction of entities (e.g., tasks, methods, datasets, variables) and relations from full-text articles [44, 14, 45, 46, 47]. These tasks often require assembling evidence distributed across sections, rather than extracting isolated sentence-level spans. In parallel, research on multi-document and multi-hop question answering examines how models integrate evidence across passages or documents. Many approaches construct structured intermediates, such as passage-level or document-level knowledge graphs, to retrieve, align, and compose supporting evidence [48, 49]. Emerging researches spanning scientific articles, citation networks, and multimodal inputs consistently report substantial gaps between LLM and expert performance in multi-document SciIE [50, 51, 52]. Related DocVQA-style benchmarks highlight similar challenges, showing persistent failures in selecting, aligning, and integrating evidence across pages, modalities, and documents, even with retrieval augmentation [53]. Together, these results indicate that long-context ingestion alone is insufficient for reliable cross-document evidence integration. Most existing formulations nonetheless frame queries as ad-hoc questions or loosely structured prompts, with evaluation centered on answer correctness or evidence selection. Much less is known about whether LLMs can reliably construct schema-conformant, multi-field scientific records that preserve entity bindings and numerically grounded attributes across documents. We address this gap by grounding all queries in an explicit study schema and systematically varying the arity and structural constraints of the target output, reframing multi-document reasoning as structured evidence assembly rather than question answering.

Compositionality, Binding, and Numeric Reasoning in LLMs. A growing body of work documents systematic limitations of LLMs in compositional generalization, entity–attribute binding, and numerically grounded reasoning. Studies on compositionality show that models often fail to generalize to novel combinations of familiar primitives, instead relying on surface patterns rather than

systematic structure [54, 55]. Related work further shows that LLMs struggle to reliably bind attributes, intermediate conclusions, or numeric values to the correct entities, producing fluent but structurally misaligned outputs even when individual facts are known [56, 57]. Such binding errors become more pronounced in long-context or multi-document settings, where models may ignore, conflate, or misattribute evidence across sources [56]. Numeric reasoning poses additional challenges. Empirical evaluations consistently report brittle performance on arithmetic operations, comparisons, and value grounding, particularly when numeric quantities must be correctly associated with variables, units, or experimental conditions [58, 59]. These failures reflect not only difficulties with computation, but with preserving structured relationships between numbers and the entities they describe. These limitations are directly relevant to meta-analytic evidence extraction. Constructing meta-analytic inputs requires composing multiple entities, binding attributes and numeric measures to the correct study components, and maintaining these relationships across documents. As a result, compositionality, binding fidelity, and numeric grounding function not as secondary concerns, but as necessary conditions for producing valid, schema-conformant study records.

Neural–Symbolic and Schema-Aware Approaches to Evidence Synthesis. Motivated by limitations of purely neural approaches, a growing line of work explores neural–symbolic and schema-aware methods for scientific information extraction and evidence synthesis. These approaches combine LLMs with explicit symbolic structure, such as ontologies, knowledge graphs, or programmatic constraints, to guide extraction, enforce schema conformity, or support structured reasoning across documents [60, 61, 62]. Prior work investigates a range of such integrations, including ontology- or KG-constrained extraction, program-aided pipelines that produce intermediate structured representations for validation and post-processing, and retrieval-augmented or constrained-decoding methods designed to improve fidelity under complex output structures [63, 64]. Collectively, these efforts reflect a broader shift toward leveraging explicit structure to improve robustness, interpretability, and aggregation in multi-document scientific settings. Against this backdrop, an open question is how reliably current LLMs can instantiate structured study representations required for meta-analysis as structural, relational, and numerical demands increase. By holding the target schema fixed, we examine how extraction behavior degrades as evidence assembly becomes more compositional and cross-document, providing a diagnostic perspective that complements existing schema-aware and neural–symbolic approaches. This perspective motivates the evaluation framework introduced in the following section.

3. Methodology

3.1. Semantic Study Schema and Application Scope

Our methodological starting point is a unified semantic representation of empirical studies that supports controlled extraction and aggregation of association-based evidence across domains. To this end, we adopt a study-level schema that abstracts over discipline-specific reporting practices while retaining the structural elements required for quantitative evidence synthesis.

The schema used in this study builds on an earlier, domain-specific semantic modeling effort within the Open Research Knowledge Graph (ORKG). That work introduced a highly specialized template (Figure 1) tailored to a single research question, and demonstrated how rich, nested semantic structures can support structured cross-paper comparison [1]. By design, however, both the template and the resulting comparison were domain-specific and not intended to generalize.

In the present work, we substantially simplify and generalize this schema by discarding fine-grained domain semantics and retaining only study elements that recur across association-based empirical research, independent of discipline. The resulting semantic study schema captures a minimal yet expressive set of entities and relations: study populations and geographical settings, sample sizes, variables participating in statistical associations, their roles as independent or dependent variables, associated

[1] ORKG template: https://orkg.org/templates/R1471878, Comparison: https://doi.org/10.48366/R1563383

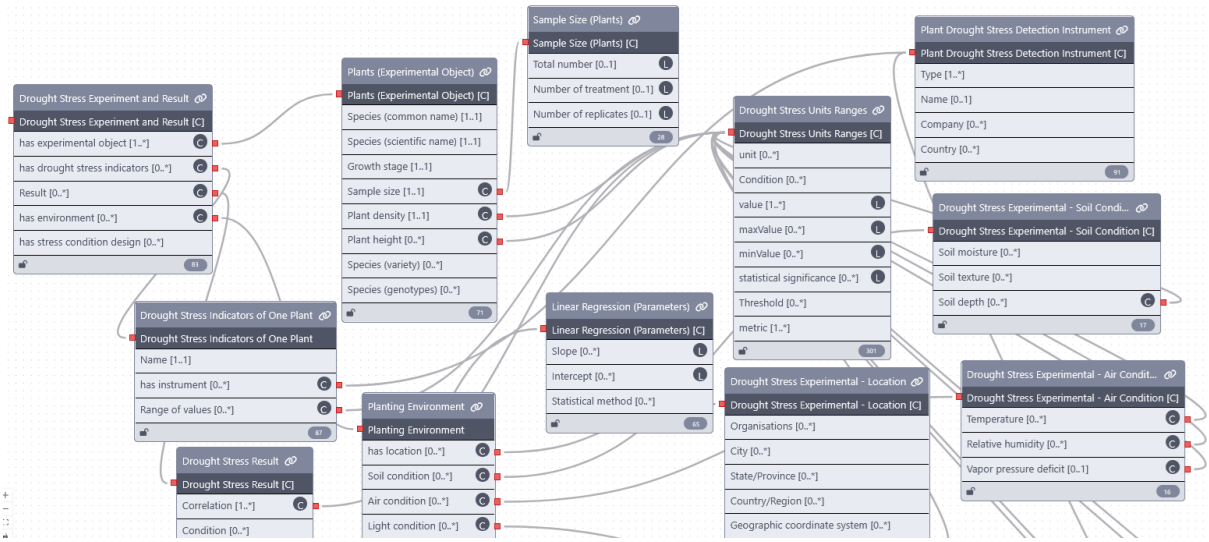

Figure 1: Partial view of the original ORKG template for *Early Drought Stress Indicators in Plants (Physiological & Multi-Sensor)* https://orkg.org/templates/R1471884, showing a richly nested, domain-specific schema (plant/specimen descriptors, multi-sensor indicators and instruments, environment/stress conditions, and structured results with value–range–unit measurements) that we generalize, in this study, into a cross-domain study schema for association-based evidence extraction.

measurement descriptors, statistical methods, and reported effect sizes. This abstraction deliberately trades domain specificity for cross-domain applicability and structural regularity. The schema serves as the unifying backbone for all extraction and derivation tasks in our framework, supporting both object-centric queries over study settings and method-centric queries over reported statistical evidence, while consistently preserving document-level attribution. Within this framework, we focus on a broad but well-structured class of empirical studies commonly used in meta-analysis, covering effect-size families such as correlation coefficients, standardized regression coefficients, coefficients of determination, and odds ratios [1, 65]. Concretely, we restrict attention to studies that report explicit statistical associations between variables (e.g., Pearson or Spearman correlations, β coefficients, R or R^2, OR), which constitute core inputs for effect-size–based synthesis across the behavioral, social, environmental, and agricultural sciences. This class of studies provides a suitable testbed for structured multi-document evidence extraction: while reporting conventions admit schema-based modeling, correct extraction still requires accurate entity recognition, role assignment, numerical grounding, and consistent document attribution, particularly under many-to-many variable relations.

Table 1 summarizes the study-level schema $\mathcal{S}$ used in our evaluation. Each element corresponds to a schema property that can be instantiated as a document-attributed atom and serve as the basic unit for extraction and derived statistical queries.

3.2. Problem Formulation

Let $\mathcal{D} = \{d_1, \ldots, d_n\}$ denote a corpus of empirical studies, where each d_i is a single source document with identifier $i \in \{1, \ldots, |\mathcal{D}|\}$. We evaluate extraction and derivation over a study schema $\mathcal{S}$ (Table 1), whose elements are *properties* such as population (P), geolocation (G), sample size (N), statistical method (A), variables (IV, DV, V, S, U), conditions (C), effect size (E), and measurement descriptors.

For a given document d_i, an *atom* is a document-attributed instantiation of a schema property, written as a pair $(i, s = v)$ where $s \in \mathcal{S}$ and v is the extracted value for that property in d_i (e.g., $(i, N = 5417)$ or $(i, A = \text{Pearson})$). We denote the set of all atoms extractable from d_i by

$$\mathcal{A}(d_i) = \{\, (i, s = v) \;:\; s \in \mathcal{S} \,\}.$$

Structural extraction targets not just individual atoms but *bindings* of multiple atoms that must

Table 1
Study-level schema properties used for object- and method-centric evidence extraction. Each schema atom is explicitly indexed by its source document identifier.

Property	Definition	Notation
Document ID	Index of the source document in the corpus	$i \in \{1, \ldots, \lvert\mathcal{D}\rvert\}$
Study Population	Study population or unit of analysis	$P \in \mathcal{P}$
Geolocation	Country of study conduct (normalize cities/regions to countries)	$G \in \mathcal{G}$
Sample Size	Main reported sample size for the study	$N \in \mathcal{N}$
Statistical Method	Statistical or association method (e.g., Pearson, regression, OR)	$A \in \mathcal{A}$
Independent Variable	Predictor variable (role-aware)	$IV \in \mathcal{IV}$
Dependent Variable	Outcome variable (role-aware)	$DV \in \mathcal{DV}$
Variable	Role-agnostic variable (union of IV and DV, de-duplicated)	$V \in \mathcal{V}$
Scale and Unit	Atomic pair describing measurement scale and unit	$(S, U) \in \mathcal{SU}$
Conditions	Optional contextual qualifier for an association estimate	$C \in \mathcal{C}$
Effect Size	Reported association estimate (e.g., r, β, OR, etc.)	$E \in \mathcal{E}$

co-refer to the same underlying study statement. We represent each bound output as a *tuple* $t \in \mathcal{T}$, i.e., an ordered record

$$t = (i; X_1, \ldots, X_k),$$

where each X_ℓ is a schema-specific slot-value assignment (e.g., $P = \cdot$, $IV = \cdot$, $DV = \cdot$, $A = \cdot$, $E = \cdot$) drawn from atoms in $\mathcal{A}(d_i)$. The arity k captures how many schema fields must be jointly bound for that output. The set of all target tuples for document d_i is denoted

$$\mathcal{I}(d_i) = \{\, t_{i1}, t_{i2}, \ldots \},$$

where t_{ij} indexes the j-th extracted tuple associated with document d_i.

We organize evaluation queries along two orthogonal dimensions: semantic focus and structural complexity.

Table 2
Object-centric (O) list-style extraction queries. All outputs preserve the source document identifier i.

ID	Tuple Pattern	Natural-Language Query
O_L1_Q1	(i, G)	Extract the study geolocation as a country name.
O_L1_Q2	(i, N)	Extract the reported sample size (use `null` if unavailable).
O_L1_Q3	(i, P)	Extract the study population or unit of analysis.
O_L2_Q1	(i, P, N)	Extract each study population together with its sample size.
O_L2_Q2	(i, P, G)	Extract each study population together with the study country.
O_L2_Q3	(i, P, G, N)	Extract each study population with both country and sample size.

Task families. Along the semantic dimension, we distinguish two families. *Object-centric tasks (O)* target properties of the study setting and experimental subjects (e.g., P, G, N), investigating identification, normalization, and correct binding under document-level attribution. *Method-centric tasks (M)* target methodological and statistical evidence structures (e.g., A, V, IV, DV, S, U, C, E), including role assignment and association tuple extraction. To disentangle variable identification from role assignment, this family includes a role-agnostic variable inventory task that extracts V as the de-duplicated union of IV and DV within each document.

Query types and levels. We define two types of queries based on whether aggregation is performed. *List-style extraction queries* directly enumerate document-attributed atoms or bound tuples according to predefined patterns, without aggregation. To make structural difficulty explicit, these queries are further organized into two *levels* of increasing relational complexity. L1 queries target single schema

Table 3
Method-centric (M) list-style extraction queries. All outputs preserve the source document identifier i.

ID	Tuple Pattern	Natural-Language Query
M_L1_Q1	(i, A)	Extract the statistical method used to quantify associations.
M_L1_Q2	(i, V)	Extract all variables.
M_L1_Q3	(i, IV)	Extract all independent variables.
M_L1_Q4	(i, DV)	Extract all dependent variables.
M_L2_Q1	(i, V, S, U)	Extract each variable together with its Scale and Unit.
M_L2_Q2	(i, IV, S, U)	Extract each independent variable with its Scale and Unit.
M_L2_Q3	(i, DV, S, U)	Extract each dependent variable together with its Scale and Unit.
M_L2_Q4	(i, IV, DV)	Extract independent–dependent variable pairs.
M_L2_Q5	(i, IV, DV, A)	Extract variable pairs together with the statistical method.
M_L2_Q6	(i, IV, DV, A, C, E)	Extract variable pairs, statistical method, effect size (conditions).

properties and require only isolated atom extraction, whereas L2 queries require binding multiple schema properties into structured tuples, involving higher-arity relational constraints (Tables 2 and 3).

Table 4
Derived statistical queries over document-attributed extraction results. Each query operates on specific schema atoms and applies a well-defined computation.

ID	Target Atoms	Computation
O_C_Q1	$(N \rightarrow \text{count})$	Count documents with $N > 100$.
O_C_Q2	$(N \rightarrow \text{mean})$	Compute the mean of N across all documents.
O_C_Q3	$(N \rightarrow \text{median})$	Compute the median of N across all documents.
M_C_Q1	$(i, A \rightarrow \text{count})$	For each document, count reported statistical methods.
M_C_Q2	$(i, V \rightarrow \text{count})$	For each document, count unique variables (union of IV and DV).
M_C_Q3	$(i, IV \rightarrow \text{count})$	For each document, count independent variables.
M_C_Q4	$(i, DV \rightarrow \text{count})$	For each document, count dependent variables.
M_C_Q5	(i, IV, DV, E)	List (IV, DV) pairs with $E > 0.7$.

In contrast, *derived statistical queries (C)* operate on the outputs of list-style extraction. They apply aggregation, counting, or conditional filtering to extracted atoms or tuples, yielding either corpus-level summaries or document-level derived attributes while preserving explicit document identifiers (Table 4).

Taken together, this formulation induces a controlled progression from single-atom extraction, to multi-atom binding, to structured statistical derivation. Our methodological objective is to examine how model behavior changes as relational binding complexity increases and as tasks move from direct extraction to derived statistical reasoning.

3.3. Evaluation Principles

We evaluate model outputs using tuple-level precision, recall, and F1, following established practices in structured information extraction.

Tuple-Level Metrics. Given gold tuples $\mathcal{G}$ and predicted tuples $\mathcal{P}$, we compute: Precision $= \frac{|\text{correct}|}{|\mathcal{P}|}$, Recall $= \frac{|\text{correct}|}{|\mathcal{G}|}$, and F1 $= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$. A predicted tuple is counted as correct only if (i) all required fields semantically match a gold tuple, and (ii) it is correctly attributed to the originating paper.

Matching Procedure. For each predicted tuple, we identify the best-matching gold tuple via a two-stage process: (1) *Paper constraint*: A predicted tuple may only match a gold tuple from the same source paper, enforced via citation markers (e.g., [N]). (2) *Semantic verification*: Candidate matches are first scored using character-level sequence similarity. Borderline candidates (similarity < 0.95) are then verified by an LLM judge, which determines whether the prediction and gold entry refer to the same

entity. Only LLM-confirmed matches are counted as correct. For composite tuples (e.g., `(variable, unit)`), all required fields must match independently.

Numerical Queries. For counting and aggregation queries (e.g., MC.1–MC.5), correctness requires exact numerical equality between the predicted value and the value computed from gold tuples.

Table 5
Descriptive statistics of five domain-specific corpora, reporting summary statistics over counts of document-attributed schema atoms across domains.

Papers	Source[a]	Metric	Sample Size	Stat. Methods	Variables	Effect Size
Civil Engineering: Building characteristics vs. Hotel energy use						
11	[66]	Min/Median/Max	6/28/51	1/2/5	2/10/32	2/24/46
		Total	292	23	139	258
Medical and Health Science: Anthropometric indicators vs. Ophthalmic outcomes						
9	[67]	Min/Median/Max	184/5,417/1,323,052	1/2/3	2/8/16	16/35/78
		Total	1,391,776	18	71	328
Agricultural Science: Environmental stress vs. Plant responses						
11	Journal[b]	Min/Median/Max	8/45/232	1/1/6	4/7/11	4/18/54
		Total	871	20	72	226
Earth & Environmental Sciences: Carbon storage vs. Biodiversity metrics						
10	[68]	Min/Median/Max	131/8,909/79,555	1/1/2	4/6/19	3/21/35
		Total	142,070	12	84	198
Social Science: Compassion vs. Well-being and burnout						
11	[69]	Min/Median/Max	37/303/1742	1/1/2	2/6/18	2/14/28
		Total	4,599	15	81	137

[a] All papers are from meta-analysis review articles except specifically noted.
[b] Retrieved *Agricultural Water Management* (2005–2025) with query *"(drought stress OR water stress) AND plant"*.

4. Experiments

This section describes the experimental setup, including the corpora, annotation protocol, preprocessing pipeline, model configurations, and input regimes. Evaluation is performed according to the tuple-level metrics and procedures defined in Section 3.

4.1. Corpora and Annotation

We construct five domain-specific corpora of primary empirical association studies spanning engineering, health, agricultural, environmental, and social sciences. The corpora are intentionally diverse in study scale, methodological complexity, and reporting practices, providing a challenging testbed for structured evidence extraction and derived statistical analysis. All included studies are primary empirical articles reporting quantitative associations with explicit statistical methods and effect-size estimates, reviews and secondary analyses are excluded.

Table 5 summarizes the key descriptive statistics of the five corpora. For each domain, the table reports corpus size as well as summary statistics over counts of document-attributed schema atoms, including sample sizes, statistical methods, variables, and reported effect-size instances. These statistics highlight substantial cross-domain variation in both study scale and structural complexity of reported evidence. Gold-standard annotations are created manually by a single expert annotator. All schema properties and association structures are extracted directly from the text, preserving explicit document identifiers for every extracted instance. Annotations include all reported variables, methods, and

association estimates, regardless of statistical significance. As the task focuses on factual extraction rather than interpretive labeling, no inter-annotator agreement procedure is required.

We release all gold-standard JSON annotations, the full query suite, experiment scripts, and the DOIs of all included articles. Due to copyright restrictions, article PDFs and their derived text representations are not redistributed, ensuring full methodological reproducibility while respecting publisher's license.

4.2. Experimental Setup

PDF-to-text preprocessing. All source PDFs are converted to structured markdown using MinerU 2.5 [2] [70] with default settings. This choice is motivated by three experimental requirements: (i) faithful preservation of document structure, including multi-column layouts, tables, and mathematical expressions; (ii) retention of section hierarchy and reading order to provide models with inputs that reflect the original article organization; and (iii) deterministic, open-source preprocessing to ensure full reproducibility. MinerU employs vision–language models for layout detection and produces markdown outputs in which tables are rendered as HTML and formulas as LaTeX, avoiding the loss of semantically relevant content. No manual editing, segmentation, or truncation is applied after conversion. For each domain, all converted articles are concatenated into a single markdown input, resulting in inputs of approximately 20k tokens per domain.

Models. We evaluate two state-of-the-art large language models representing complementary system classes: a closed-source model, GPT-5.2 (version 2025-12-11) [3], and a large open-source vision–language model, Qwen3-VL (Qwen3-VL-235B-A22B-Instruct-FP8) [4] [71]. Both models are evaluated under identical decoding settings (temperature $= 0.1$, default parameters). No external tools, retrieval components, or code execution are allowed, ensuring that all results reflect the models' intrinsic extraction and reasoning capabilities. All inputs fall within the supported context limits of both models (Qwen3-VL: 262k tokens; GPT-5.2: 400k tokens). Qwen3-VL is served via vLLM [72] on $4\times$H100 GPUs.

Queries and input regimes. The same tiered query suite is applied uniformly across all corpora. Each query is written once in natural language and reused unchanged in all experimental conditions. For per-document evaluation, corpus-level queries are programmatically rewritten into single-document variants while preserving their semantics. All models receive identical instructions and query formulations, ensuring strict control over prompt-level variables. We consider two document presentation regimes. **(1) Global input.** All article markdown files within a domain are concatenated into a single input, preceded by document identifiers and followed by the query. **(2) Per-paper input.** Following prior work on decomposed and least-to-most prompting [73, 74, 75], each query is applied independently to individual documents. After per-paper extraction, a separate aggregation step uses the model to combine document-level outputs into a final answer. No intermediate outputs are provided as context when processing subsequent documents. Across all settings, a total of 2,976 LLM calls are executed.

5. Results

5.1. Structural Trends and Domain-Level Effects

Table 6 shows systematic differences across input regimes and task groups, with additional variation across domains and models. On macro-averaged performance over all tasks, the *Per-Paper* regime yields higher F1 scores than the *Global* regime for both models: Qwen3-VL decreases from 0.35 to 0.18, and GPT-5.2 decreases from 0.28 to 0.24. This pattern is observed consistently across all domains.

Across task groups, F1 scores are generally higher for single-atom extraction (L1) than for multi-atom binding (L2) under the Per-Paper regime, and this relationship holds for both object-centric and method-centric tasks. For object-centric extraction, Qwen3-VL decreases from O1 = 0.62 to O2 = 0.30,

[2] https://github.com/opendatalab/MinerU
[3] https://platform.openai.com/docs/models/gpt-5.2
[4] https://huggingface.co/Qwen/Qwen3-VL-235B-A22B-Instruct-FP8

Table 6
Comparison of F1 scores by extraction group across domains, models, and input regimes (Per-Paper vs. Global). Column labels use abbreviated task group names (e.g., O1 = O_L1, O2 = O_L2, OC = O_C, M1 = M_L1, M2 = M_L2, MC = M_C).

		Per-Paper							**Global**						
		Object			Method				Object			Method			
Domain	**Model**	O1	O2	OC	M1	M2	MC	All	O1	O2	OC	M1	M2	MC	All
Agricultural Sci.	Qwen3-VL	.56	.30	.00	.39	.20	.23	.27	.48	.36	.33	.22	.12	.00	.22
	GPT-5.2	.49	.09	.00	.38	.16	.27	.23	.66	.06	.33	.40	.23	.09	.27
Social Sci.	Qwen3-VL	.47	.26	.00	.28	.22	.27	.25	.46	.09	.00	.00	.05	.00	.08
	GPT-5.2	.46	.23	.00	.26	.16	.10	.19	.48	.26	.00	.26	.18	.00	.18
Medical & Health Sci.	Qwen3-VL	.80	.37	.67	.55	.17	.28	.42	.67	.37	.33	.27	.07	.00	.23
	GPT-5.2	.56	.19	.67	.52	.23	.19	.36	.67	.11	.00	.42	.20	.00	.22
Earth & Environ. Sci.	Qwen3-VL	.62	.03	.00	.56	.23	.37	.31	.53	.00	.00	.35	.00	.00	.13
	GPT-5.2	.42	.00	.00	.52	.23	.19	.24	.56	.00	.00	.53	.21	.11	.24
Civil Engineering	Qwen3-VL	.66	.52	1.0	.48	.27	.26	.47	.76	.55	.33	.21	.01	.01	.24
	GPT-5.2	.57	.48	.67	.56	.33	.21	.43	.91	.45	.00	.36	.18	.13	.30
Average	*Qwen3-VL*	*.62*	*.30*	*.33*	*.45*	*.22*	*.28*	*.35*	*.57*	*.27*	*.20*	*.21*	*.05*	*.00*	*.18*
	GPT-5.2	*.50*	*.20*	*.27*	*.45*	*.22*	*.19*	*.28*	*.66*	*.18*	*.07*	*.39*	*.20*	*.07*	*.24*

and GPT-5.2 decreases from O1 = 0.50 to O2 = 0.20. For method-centric extraction, both models achieve M1 = 0.45 and decrease to M2 = 0.22.

Comparing input regimes within task families, the Global regime affects L2 tasks more strongly than L1 tasks, with the largest drops observed in method-centric extraction. For Qwen3-VL, M2 decreases from 0.22 (Per-Paper) to 0.05 (Global), and MC decreases from 0.28 to 0.00, whereas O1 decreases from 0.62 to 0.57 and O2 from 0.30 to 0.27. For GPT-5.2, Global input yields M1 = 0.39, M2 = 0.20, and MC = 0.07, alongside O1 = 0.66 and O2 = 0.18. In contrast, differences between Per-Paper and Global input for L1 tasks are comparatively smaller across domains.

Precision and recall both decrease as structural complexity increases, but recall more frequently reaches near-zero values first, particularly for method-centric L2 and derived tasks. This recall-dominant degradation pattern appears consistently across models and domains.

When comparing object-centric and method-centric tasks at matched tuple arity, object-centric tasks are higher on average. At the single-atom level (L1), macro-averaged F1 is 0.59 for O1 versus 0.38 for M1. Within object-centric tasks, performance varies systematically by entity type: geolocation and sample size yield higher F1 scores (0.72 and 0.70, respectively), while population or unit-of-analysis extraction is substantially lower (0.35). At the multi-atom level (L2), macro-averaged F1 remains higher for object-centric tasks (O2 = 0.24) than for method-centric tasks (M2 = 0.17).

Within method-centric extraction, task difficulty varies depending on the presence of role and relational constraints, and similar patterns recur across arity levels. At L1, role-agnostic variable extraction (i, V) achieves an F1 of 0.44, compared to 0.37 for (i, IV) and 0.42 for (i, DV). At L2, the role-agnostic task (i, V, S, U) reaches 0.25, whereas (i, IV, S, U) and (i, DV, S, U) reach 0.15 and 0.16, respectively. For variable-pair extraction, (i, IV, DV) yields an F1 of 0.26, and adding statistical method binding further reduces performance to 0.21 for (i, IV, DV, A). The same ordering is observed across both L1 and L2 tasks, with role-agnostic variants consistently achieving higher F1. The highest-arity method-centric task (i, IV, DV, A, C, E) attains an F1 of 0.00 under both input regimes and both models.

Domain-level differences are also visible in overall F1 scores and vary with the input regime. Under the Per-Paper regime, overall performance across domains is relatively concentrated, whereas under the Global regime domain-level variance increases substantially. This contrast is further illustrated by the Per-Paper $\rightarrow$ Global F1 drop across domains (Figure 2), where all domains exhibit performance

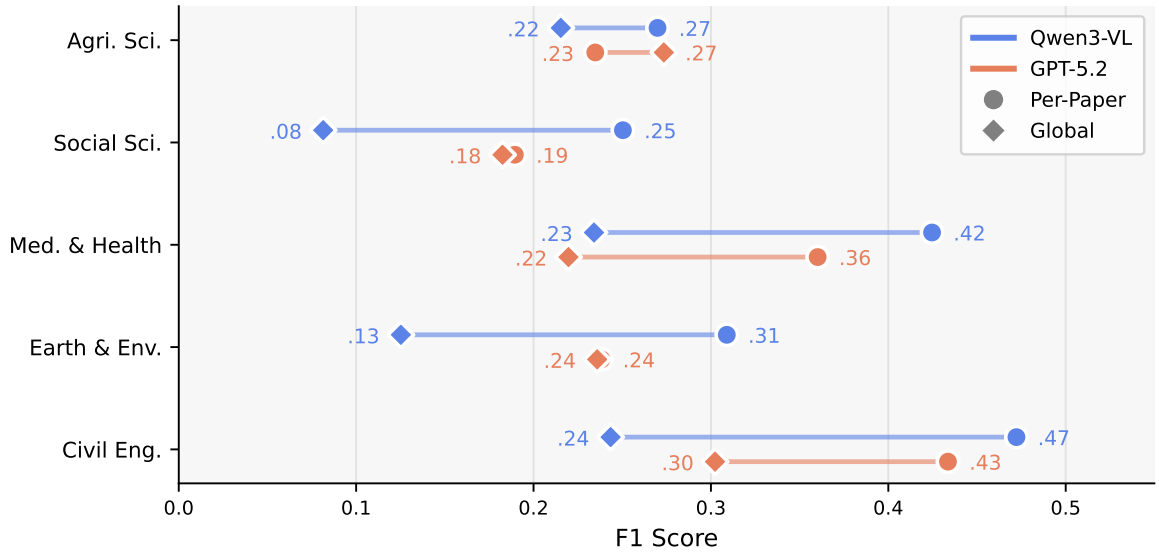

Figure 2: Domain-level comparison of macro-averaged F1 scores under the Per-Paper and Global input regimes. Each line connects Per-Paper and Global performance for the same domain and model, illustrating the magnitude of performance change induced by input regime variation.

degradation but with markedly different magnitudes. Civil Engineering achieves the highest overall performance under both regimes (Per-Paper: 0.47/0.43; Global: 0.24/0.30 for Qwen3-VL/GPT-5.2). Medical & Health Science ranks second under Per-Paper input (0.42/0.36), while under Global input it is close to Agricultural Science (Qwen3-VL: 0.23 vs. 0.22; GPT-5.2: 0.22 vs. 0.27). Social Science yields the lowest overall F1, particularly under Global input (0.08/0.18). Earth & Environmental Science shows a large model gap under Global input, with Qwen3-VL at 0.13 and GPT-5.2 at 0.24. These regime-dependent domain shifts motivate a closer inspection of domain-specific failure cases in the subsequent analysis.

5.2. Error Analysis

The quantitative patterns observed above indicate that performance degradation is not primarily attributable to failures in identifying individual entities, but instead to systematic breakdowns in preserving relational structure as task demands increase. To better understand these trends, we analyze model outputs at the error level and identify four recurring error patterns that account for the sharp declines observed in method-centric and derived tasks.

A first prominent failure mode is *role confusion in variable directionality*. In variable-pair extraction tasks (M2.4–M2.6), 15.5% of spurious predictions (688 out of 4,430) correspond to exact swaps of independent and dependent variables relative to the gold standard. These errors occur despite explicit role cues in the source documents, including table headers (e.g., *predictor* vs. *outcome*), regression formulations, and methodological descriptions. While models reliably detect the co-occurrence of relevant variables, they frequently fail to consistently bind role-specific information to those variables. Instead, roles are often assigned based on superficial cues such as surface order of mention, indicating that role assignment remains brittle even when variable identification is correct.

Beyond local role assignment, models also exhibit systematic *binding drift* across distinct analytical contexts within the same document. Among spurious predictions in M2.5 and M2.6 under the per-paper input setting, 21.6% (518 out of 2,403) involve variable pairs that are present in the gold annotations but are incorrectly associated with an unrelated statistical method, effect size, or experimental condition. These errors arise most frequently in papers that reuse the same variables across multiple analyses, tables, or models. In such cases, models successfully extract the relevant components in isolation, but fail to preserve the boundaries between separate analytical units, resulting in cross-analysis conflation.

Binding errors are further amplified in documents with high densities of reported results, leading to a *multi-instance compression* effect. For the highest-arity task (M2.6), per-paper recall decreases monotonically as the number of gold-standard tuples increases. Papers containing 1–5 annotated instances achieve an average recall of 0.45, whereas papers with 31 or more instances drop to 0.32, with several dense papers yielding no correct extractions. This degradation is specific to method-centric tasks, where a single document may report dozens of associations. Object-centric queries, which typically admit only one or a small number of instances per paper, do not exhibit a comparable decline.

Finally, these upstream extraction failures propagate into pronounced *error amplification* in derived statistical queries. Aggregation tasks such as counts, means, and medians require complete and accurate upstream extractions; as a result, even limited omission or misbinding renders the final scalar output incorrect. For example, OC.1 succeeds in only 20% of cases despite its underlying extraction achieving an F1 of 0.66, and OC.2 fails in over 60% of cases because excluding a single document alters the computed mean. These results illustrate that extraction quality sufficient for qualitative inspection is inadequate for quantitative synthesis, where exhaustive recall is a strict requirement.

Taken together, these error patterns show that current models do not primarily fail at recognizing scientific entities in isolation. Rather, failures arise from maintaining role consistency, analytical scope, and instance completeness as relational complexity increases. These limitations directly align with the performance ceilings observed in high-arity method-centric and derived tasks, and delineate the core challenges that must be addressed for reliable meta-analytic evidence extraction.

6. Conclusion

We presented a structural, diagnostic study of LLM-based evidence extraction for meta-analysis, framing the problem not as an end-to-end pipeline but as a controlled progression from single-property extraction, to multi-field binding, to derived statistical computation under a fixed semantic study schema. Across five domains and two state-of-the-art models, the results draw a clear boundary between what current LLMs can do reliably and what meta-analytic workflows actually demand. Single-atom queries are often handled competently, but performance deteriorates sharply once outputs must preserve higher-arity relational structure, especially when variable roles, statistical methods, conditions, and effect sizes must be jointly bound and attributed to the correct source paper. This gap widens further in long-context, multi-document settings, where global ingestion consistently underperforms compared to per-paper decomposition, and where downstream aggregations exhibit severe error amplification, even when upstream extraction appears sufficient by conventional IE metrics. Our error analysis shows that the dominant failures are not failures of recognition, but failures of structure: role directionality swaps, cross-analysis binding drift, instance compression in dense result sections, and cascading brittleness in numeric summaries. These patterns explain why the highest-arity association tuples collapse to near-zero performance and why seemingly simple corpus-level statistics become unreliable. The implication is that meta-analytic evidence extraction is constrained less by surface-level scientific language understanding than by the ability to maintain stable bindings over many-to-many relations and to preserve analytical scope boundaries under accumulation. The query suite and evaluation protocol introduced here provide a principled way to measure these structural capabilities and to localize where failures begin. More importantly, they suggest a concrete agenda for the next generation of systems: schema-aware extraction that treats binding as a first-class objective, explicit representations of analytical units to prevent cross-model conflation, and verification layers that can enforce consistency between variables, roles, methods, and numeric values before aggregation. By making the structural requirements explicit and empirically mapping current failure regimes, this work offers a foundation for neural–symbolic and constraint-guided approaches that are better aligned with the precision, attribution, and completeness needs of evidence synthesis at scale.

Acknowledgments

This work was supported by the project "HybrInt – Hybrid Intelligence through Interpretable AI in Machine Perception and Interaction" (Zukunft.Niedersachsen; Lower Saxony Ministry for Science and Culture; Grant ID: ZN4219). Computational resources were provided by the TIB – Leibniz Information Centre for Science and Technology. The lead author thanks Kheir Eddine Farfar (ORKG Development Team) for support and guidance on the semantic modeling associated with the ORKG template https://orkg.org/templates/R1471884 and the corresponding comparison https://doi.org/10.48366/R1563383.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Grammar and spelling check. After using the tool, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

References

[1] J. P. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, V. A. Welch (Eds.), Cochrane Handbook for Systematic Reviews of Interventions, The Cochrane Collaboration, 2019. URL: https://onlinelibrary.wiley.com/doi/book/10.1002/9781119536604. doi:10.1002/9781119536604.

[2] H. Cooper, Research synthesis and meta-analysis: A step-by-step approach, volume 2, Sage publications, 2015.

[3] G. Tsafnat, P. Glasziou, M. K. Choong, A. Dunn, F. Galgani, E. Coiera, Systematic review automation technologies, Systematic reviews 3 (2014) 74.

[4] S. Peroni, D. Shotton, The spar ontologies, in: D. Vrandečić, K. Bontcheva, M. C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, L.-A. Kaffee, E. Simperl (Eds.), The Semantic Web – ISWC 2018, Springer International Publishing, Cham, 2018, pp. 119–136.

[5] C. E. Lipscomb, Medical subject headings (mesh), Bulletin of the Medical Library Association 88 (2000) 265.

[6] M. Ashburner, C. A. Ball, J. A. Blake, et al., Gene ontology: tool for the unification of biology, Nature Genetics 25 (2000) 25–29. URL: https://doi.org/10.1038/75556. doi:10.1038/75556.

[7] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucleic Acids Research 32 (2004) D267–D270. URL: https://doi.org/10.1093/nar/gkh061. doi:10.1093/nar/gkh061.

[8] W. Ammar, D. Groeneveld, C. Bhagavatula, I. Beltagy, M. Crawford, D. Downey, J. Dunkelberger, A. Elgohary, S. Feldman, V. Ha, R. Kinney, S. Kohlmeier, K. Lo, T. Murray, H.-H. Ooi, M. Peters, J. Power, S. Skjonsberg, L. L. Wang, C. Wilhelm, Z. Yuan, M. van Zuylen, O. Etzioni, Construction of the literature graph in semantic scholar, in: S. Bangalore, J. Chu-Carroll, Y. Li (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers), Association for Computational Linguistics, New Orleans - Louisiana, 2018, pp. 84–91. URL: https://aclanthology.org/N18-3011/. doi:10.18653/v1/N18-3011.

[9] D. Dessí, F. Osborne, D. Buscaldi, et al., Cs-kg 2.0: A large-scale knowledge graph of computer science, Scientific Data 12 (2025) 964. URL: https://doi.org/10.1038/s41597-025-05200-8. doi:10.1038/s41597-025-05200-8.

[10] S. Auer, J. D'Souza, K. E. Farfar, M. Y. Jaradeh, A. Jiomekong, O. Karras, A. Oelen, L. Snyder, M. Stocker, L. Vogt, Open research knowledge graph: a large-scale neuro-symbolic knowledge organization system, in: Handbook on Neurosymbolic AI and Knowledge Graphs, IOS Press, 2025, pp. 385–420.

[11] M. Shamsabadi, J. D'Souza, S. Auer, Large language models for scientific information extraction:

An empirical study for virology, in: Findings of the Association for Computational Linguistics: EACL 2024, 2024, pp. 374–392.

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

[13] W. Lu, R. K. Luu, M. J. Buehler, Fine-tuning large language models for domain adaptation: exploration of training strategies, scaling, model merging and synergistic capabilities, npj Computational Materials 11 (2025) 84. URL: https://doi.org/10.1038/s41524-025-01564-y. doi:10.1038/s41524-025-01564-y.

[14] J. D'Souza, Z. Laubach, T. Al Mustafa, S. Zarrieß, R. Frühstückl, P. Illari, Mining for species, locations, habitats, and ecosystems from scientific papers in invasion biology: A large-scale exploratory study with large language models, in: Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025), 2025, pp. 16–23.

[15] Y. Wang, Q. Guo, W. Yao, H. Zhang, X. Zhang, Z. Wu, M. Zhang, X. Dai, M. Zhang, Q. Wen, et al., Autosurvey: large language models can automatically write surveys, in: Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS 2024), Curran Associates, Inc., New Orleans, LA, USA, 2024, pp. 115119–115145. Published: 10 December 2024.

[16] B. Guo, Z. Wen, Y. Yang, P. Gao, R. Yang, J. Shen, Sgsimeval: A comprehensive multifaceted and similarity-enhanced benchmark for automatic survey generation systems, in: Advanced Data Mining and Applications: 21st International Conference, ADMA 2025, Kyoto, Japan, October 22–24, 2025, Proceedings, Part II, Springer, Cham, 2025, pp. 393–407. URL: https://doi.org/10.1007/978-981-95-3456-2_27. doi:10.1007/978-981-95-3456-2_27, published: 22 October 2025.

[17] N. Dziri, X. Lu, M. Sclar, X. L. Li, L. Jiang, B. Y. Lin, S. Welleck, P. West, C. Bhagavatula, R. Le Bras, J. Hwang, S. Sanyal, X. Ren, A. Ettinger, Z. Harchaoui, Y. Choi, Faith and fate: Limits of transformers on compositionality, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 70293–70332. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/deb3c28192f979302c157cb653c15e90-Paper-Conference.pdf.

[18] M. Ismayilzada, D. Circi, J. Sälevä, H. Sirin, A. Köksal, B. Dhingra, A. Bosselut, D. Ataman, L. V. D. Plas, Evaluating morphological compositional generalization in large language models, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), 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), Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 1270–1305. URL: https://aclanthology.org/2025.naacl-long.59/. doi:10.18653/v1/2025.naacl-long.59.

[19] J. Feng, J. Steinhardt, How do language models bind entities in context?, in: The Twelfth International Conference on Learning Representations, 2024. URL: https://openreview.net/forum?id=zb3b6oKO77.

[20] Y. Gur-Arieh, M. Geva, A. Geiger, Mixing mechanisms: How language models retrieve bound entities in-context, 2025. URL: https://arxiv.org/abs/2510.06182. arXiv:2510.06182.

[21] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, P. Liang, Lost in the middle: How language models use long contexts, Transactions of the Association for Computational Linguistics 12 (2024) 157–173. URL: https://aclanthology.org/2024.tacl-1.9/. doi:10.1162/tacl_a_00638.

[22] M. Akhtar, A. Shankarampeta, V. Gupta, A. Patil, O. Cocarascu, E. Simperl, Exploring the numerical reasoning capabilities of language models: A comprehensive analysis on tabular data, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 15391–15405. URL: https://aclanthology.org/2023.findings-emnlp.1028/. doi:10.18653/v1/2023.findings-emnlp.1028.

[23] K. Ahmed, M. Osama, O. Sharif, E. Hossain, M. M. Hoque, BenNumEval: A benchmark to assess LLMs' numerical reasoning capabilities in Bengali, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Findings of the Association for Computational Linguistics: ACL 2025,

Association for Computational Linguistics, Vienna, Austria, 2025, pp. 17782–17799. URL: https://aclanthology.org/2025.findings-acl.915/. doi:10.18653/v1/2025.findings-acl.915.

[24] F. Oyesanmi, P. O. Olukanmi, How good are large language models at arithmetic reasoning in low-resource language settings?—a study on yorùbá numerical probes with minimal contamination, Applied Sciences 15 (2025) 4459.

[25] L. Li, A. Mathrani, T. Susnjak, Automated risk-of-bias assessment of randomized controlled trials: A first look at a gepa-trained programmatic prompting framework, arXiv preprint arXiv:2512.01452 (2025).

[26] S. Peroni, D. Shotton, Fabio and cito: Ontologies for describing bibliographic resources and citations, Journal of Web Semantics 17 (2012) 33–43. URL: https://www.sciencedirect.com/science/article/pii/S1570826812000790. doi:https://doi.org/10.1016/j.websem.2012.08.001.

[27] A. Constantin, S. Peroni, S. Pettifer, D. Shotton, F. Vitali, The document components ontology (doco), Semantic Web 7 (2016) 167–181. URL: https://journals.sagepub.com/doi/abs/10.3233/SW-150177. doi:10.3233/SW-150177.

[28] P. Ciccarese, E. Wu, G. Wong, M. Ocana, J. Kinoshita, A. Ruttenberg, T. Clark, The swan biomedical discourse ontology, Journal of Biomedical Informatics 41 (2008) 739–751. URL: https://www.sciencedirect.com/science/article/pii/S1532046408000580. doi:https://doi.org/10.1016/j.jbi.2008.04.010, semantic Mashup of Biomedical Data.

[29] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, T. C. Rindflesch, Semmeddb: a pubmed-scale repository of biomedical semantic predications, Bioinformatics 28 (2012) 3158–3160. URL: https://doi.org/10.1093/bioinformatics/bts591. doi:10.1093/bioinformatics/bts591.

[30] D. S. Himmelstein, A. Lizee, C. Hessler, L. Brueggeman, S. L. Chen, D. Hadley, A. Green, P. Khankhanian, S. E. Baranzini, Systematic integration of biomedical knowledge prioritizes drugs for repurposing, eLife 6 (2017) e26726. URL: https://doi.org/10.7554/eLife.26726. doi:10.7554/eLife.26726.

[31] A. Oelen, M. Y. Jaradeh, K. E. Farfar, M. Stocker, S. Auer, Comparing research contributions in a scholarly knowledge graph, in: SciKnow 2019, Third International Workshop on Capturing Scientific Knowledge: proceedings of the Third International Workshop on Capturing Scientific Knowledge, co-located with the 10th International Conference on Knowledge Capture (K-CAP 2019), volume 2526, Aachen: RWTH Aachen, 2019, pp. 21–26.

[32] V. Venugopal, E. Olivetti, Matkg: An autonomously generated knowledge graph in material science, Scientific Data 11 (2024) 217.

[33] S. Sadruddin, J. D'Souza, E. Poupaki, A. Watkins, H. Babaei Giglou, A. Rula, B. Karasulu, S. Auer, A. Mackus, E. Kessels, Llms4schemadiscovery: A human-in-the-loop workflow for scientific schema mining with large language models, in: European Semantic Web Conference, Springer, 2025, pp. 244–261.

[34] L. Schmidt, A. N. Finnerty Mutlu, R. Elmore, et al., Data extraction methods for systematic review (semi)automation: Update of a living systematic review, F1000Research 10 (2025) 401. URL: https://doi.org/10.12688/f1000research.51117.3. doi:10.12688/f1000research.51117.3, version 3; peer review: 3 approved.

[35] D. Demner-Fushman, J. Lin, Answering clinical questions with knowledge-based and statistical techniques, Computational Linguistics 33 (2007) 63–103. URL: https://doi.org/10.1162/coli.2007.33.1.63. doi:10.1162/coli.2007.33.1.63.

[36] B. C. Wallace, J. Kuiper, A. Sharma, M. B. Zhu, I. J. Marshall, Extracting pico sentences from clinical trial reports using supervised distant supervision, Journal of Machine Learning Research 17 (2016) 1–25. URL: http://jmlr.org/papers/v17/15-404.html.

[37] S. Kiritchenko, B. de Bruijn, S. Carini, et al., Exact: automatic extraction of clinical trial characteristics from journal publications, BMC Medical Informatics and Decision Making 10 (2010) 56. URL: https://doi.org/10.1186/1472-6947-10-56. doi:10.1186/1472-6947-10-56.

[38] Y. Hu, V. K. Keloth, K. Raja, Y. Chen, H. Xu, Towards precise pico extraction from abstracts of randomized controlled trials using a section-specific learning approach, Bioinformatics 39 (2023) btad542. URL: https://doi.org/10.1093/bioinformatics/btad542. doi:10.1093/bioinformatics/btad542.

[39] J. Brassey, C. Price, J. Edwards, M. Zlabinger, A. Bampoulidis, A. Hanbury, Developing a fully automated evidence synthesis tool for identifying, assessing and collating the evidence, BMJ Evidence-Based Medicine 26 (2021) 24–27. URL: https://ebm.bmj.com/content/26/1/24. doi:10.1136/bmjebm-2018-111126.

[40] P. S. J. Jardim, C. J. Rose, H. M. Ames, et al., Automating risk of bias assessment in systematic reviews: a real-time mixed methods comparison of human researchers to a machine learning system, BMC Medical Research Methodology 22 (2022) 167. URL: https://doi.org/10.1186/s12874-022-01649-y. doi:10.1186/s12874-022-01649-y, version of record: 08 June 2022.

[41] J.-L. Lieberum, M. Toews, M.-I. Metzendorf, F. Heilmeyer, W. Siemens, C. Haverkamp, D. Böhringer, J. J. Meerpohl, A. Eisele-Metzger, Large language models for conducting systematic reviews: on the rise, but not yet ready for use—a scoping review, Journal of Clinical Epidemiology 181 (2025) 111746. URL: https://www.sciencedirect.com/science/article/pii/S0895435625000794. doi:https://doi.org/10.1016/j.jclinepi.2025.111746.

[42] M. A. Khan, U. Ayub, S. A. A. Naqvi, K. Z. R. Khakwani, Z. b. R. Sipra, A. Raina, S. Zhou, H. He, A. Saeidi, B. Hasan, et al., Collaborative large language models for automated data extraction in living systematic reviews, Journal of the American Medical Informatics Association 32 (2025) 638–647. URL: https://doi.org/10.1093/jamia/ocae325. doi:10.1093/jamia/ocae325.

[43] N. Buscemi, L. Hartling, B. Vandermeer, L. Tjosvold, T. P. Klassen, Single data extraction generated more errors than double data extraction in systematic reviews, Journal of Clinical Epidemiology 59 (2006) 697–703. URL: https://www.sciencedirect.com/science/article/pii/S0895435605004117. doi:https://doi.org/10.1016/j.jclinepi.2005.11.010.

[44] S. Jain, M. van Zuylen, H. Hajishirzi, I. Beltagy, SciREX: A challenge dataset for document-level information extraction, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7506–7516. URL: https://aclanthology.org/2020.acl-main.670/. doi:10.18653/v1/2020.acl-main.670.

[45] Q. Zhang, Z. Chen, H. Pan, C. Caragea, L. J. Latecki, E. Dragut, SciER: An entity and relation extraction dataset for datasets, methods, and tasks in scientific documents, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 13083–13100. URL: https://aclanthology.org/2024.emnlp-main.726/. doi:10.18653/v1/2024.emnlp-main.726.

[46] D. Pham, X. Ho, Q. T. Ha, A. Aizawa, Solving label variation in scientific information extraction via multi-task learning, in: C.-R. Huang, Y. Harada, J.-B. Kim, S. Chen, Y.-Y. Hsu, E. Chersoni, P. A, W. H. Zeng, B. Peng, Y. Li, J. Li (Eds.), Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation, Association for Computational Linguistics, Hong Kong, China, 2023, pp. 243–256. URL: https://aclanthology.org/2023.paclic-1.25/.

[47] S. Kabongo, J. D'Souza, S. Auer, Effective context selection in llm-based leaderboard generation: an empirical study, in: International Conference on Applications of Natural Language to Information Systems, Springer, 2024, pp. 150–160.

[48] Y. Wang, N. Lipka, R. A. Rossi, A. Siu, R. Zhang, T. Derr, Knowledge graph prompting for multi-document question answering, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 19206–19214. URL: https://doi.org/10.1609/aaai.v38i17.29889. doi:10.1609/aaai.v38i17.29889.

[49] Z. Yang, Z. Zhu, J. Zhu, CuriousLLM: Elevating multi-document question answering with LLM-enhanced knowledge graph reasoning, in: W. Chen, Y. Yang, M. Kachuee, X.-Y. Fu (Eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track), Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 274–286. URL: https://aclanthology.org/2025.naacl-industry.23/. doi:10.18653/v1/2025.naacl-industry.23.

[50] C. Li, Z. Shangguan, Y. Zhao, D. Li, Y. Liu, A. Cohan, M3SciQA: A multi-modal multi-document scientific QA benchmark for evaluating foundation models, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen

(Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 15419–15446. URL: https://aclanthology.org/2024.findings-emnlp.904/. doi:10.18653/v1/2024.findings-emnlp.904.

[51] Z. Tan, C. Duan, Multi-Disciplinary Dataset Discovery from Citation-Verified Literature Contexts, in: Proceedings of the 2025 ACM/IEEE Joint Conference on Digital Libraries (JCDL), IEEE Computer Society, Los Alamitos, CA, USA, 2025, pp. 109–118. doi:10.1109/JCDL67857.2025.00022. arXiv:https://arxiv.org/abs/2601.05099.

[52] J. Park, S. Pyeon, J. Kim, R. C. Cabal, Y. Ding, S. C. Han, Dochop-qa: Towards multi-hop reasoning over multimodal document collections, 2025. URL: https://arxiv.org/abs/2508.15851. arXiv:2508.15851.

[53] K. Dong, Y. Chang, S. Huang, Y. Wang, R. Tang, Y. Liu, Benchmarking retrieval-augmented multimomal generation for document question answering, in: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2025. URL: https://openreview.net/forum?id=W4b3v9jx1p.

[54] J. Zhao, J. Tong, Y. Mou, M. Zhang, Q. Zhang, X. Huang, Exploring the compositional deficiency of large language models in mathematical reasoning through trap problems, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 16361–16376. URL: https://aclanthology.org/2024.emnlp-main.915/. doi:10.18653/v1/2024.emnlp-main.915.

[55] P. Shojaee, I. Mirzadeh, K. Alizadeh, M. Horton, S. Bengio, M. Farajtabar, The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity, in: The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025. URL: https://openreview.net/forum?id=YghiOusmvw. doi:https://doi.org/10.48550/arXiv.2506.06941, neurIPS 2025.

[56] P. Song, P. Han, N. Goodman, A survey on large language model reasoning failures, in: 2nd AI for Math Workshop @ ICML 2025, 2025. URL: https://openreview.net/forum?id=hsgMn4KBFG.

[57] K. Mahowald, A. A. Ivanova, I. A. Blank, N. Kanwisher, J. B. Tenenbaum, E. Fedorenko, Dissociating language and thought in large language models, Trends in Cognitive Sciences 28 (2024) 517–540. URL: https://www.sciencedirect.com/science/article/pii/S1364661324000275. doi:https://doi.org/10.1016/j.tics.2024.01.011.

[58] J. Boye, B. Moell, Large language models and mathematical reasoning failures, 2025. URL: https://arxiv.org/abs/2502.11574. arXiv:2502.11574.

[59] Y. Yan, Y. Lu, R. Xu, Z. Lan, Do large language models truly grasp addition? a rule-focused diagnostic using two-integer arithmetic, 2025. URL: https://arxiv.org/abs/2504.05262. arXiv:2504.05262.

[60] L. Liu, Z. Wang, H. Tong, Neural-symbolic reasoning over knowledge graphs: A survey from a query perspective, SIGKDD Explor. Newsl. 27 (2025) 124–136. URL: https://doi.org/10.1145/3748239.3748249. doi:10.1145/3748239.3748249.

[61] B. Škrlj, B. Koloski, S. Pollak, N. Lavrač, From symbolic to neural and back: Exploring knowledge graph–large language model synergies, in: M. van Leeuwen, J. Vreeken (Eds.), Challenges and Algorithms for Knowledge Discovery from Data, volume 16067 of *Lecture Notes in Computer Science*, Springer, Cham, 2026. URL: https://doi.org/10.1007/978-3-032-03028-3_11. doi:10.1007/978-3-032-03028-3_11, published online: 23 September 2025.

[62] J. Z. Pan, S. Razniewski, J.-C. Kalo, S. Singhania, J. Chen, S. Dietze, H. Jabeen, J. Omeliyanenko, W. Zhang, M. Lissandrini, R. Biswas, G. de Melo, A. Bonifati, E. Vakaj, M. Dragoni, D. Graux, Large language models and knowledge graphs: Opportunities and challenges, Transactions on Graph Data and Knowledge 1 (2023) 2:1–2:38. URL: https://doi.org/10.4230/TGDK.1.1.2. doi:10.4230/TGDK.1.1.2, special Issue on Trends in Graph Data and Knowledge.

[63] B. Zhang, H. Soh, Extract, define, canonicalize: An LLM-based framework for knowledge graph construction, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 9820–9836. URL: https://aclanthology.org/2024.emnlp-main.548/.

doi:10.18653/v1/2024.emnlp-main.548.

[64] K. Li, T. Zhang, X. Wu, H. Luo, J. R. Glass, H. M. Meng, Decoding on graphs: Faithful and sound reasoning on knowledge graphs through generation of well-formed chains, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 24349–24364. URL: https://aclanthology.org/2025.acl-long.1186/. doi:10.18653/v1/2025.acl-long.1186.

[65] M. Borenstein, L. V. Hedges, J. P. Higgins, H. R. Rothstein, Introduction to meta-analysis, John wiley & sons, 2021.

[66] R. S. Arenhart, T. Martins, R. M. Ueda, A. M. Souza, R. R. Zanini, Energy use and its contributors in hotel buildings: A systematic review and meta-analysis, PloS one 19 (2024) e0309745.

[67] Z. Guan, Y. Luo, H. Liu, et al., Association of anthropometric parameter with myopia in children and adolescents: A systematic review and meta-analysis, International Journal of Obesity 49 (2025) 2395–2405. URL: https://doi.org/10.1038/s41366-025-01900-8. doi:10.1038/s41366-025-01900-8.

[68] B. Liu, L. Scherer, P. M. van Bodegom, Z. Sun, P. Behrens, Exploring the spatial relationship between carbon storage and biodiversity: A systematic review and meta-analysis, Land Degradation & Development 36 (2025) 2786–2797.

[69] M. Zhuniq, F. Winter, C. Aguilar-Raab, Compassion for others and well-being: a meta-analysis, Scientific Reports 15 (2025) 36478. URL: https://doi.org/10.1038/s41598-025-23460-7. doi:10.1038/s41598-025-23460-7.

[70] J. Niu, Z. Liu, Z. Gu, B. Wang, L. Ouyang, Z. Zhao, T. Chu, T. He, F. Wu, Q. Zhang, Z. Jin, G. Liang, R. Zhang, W. Zhang, Y. Qu, Z. Ren, Y. Sun, Y. Zheng, D. Ma, Z. Tang, B. Niu, Z. Miao, H. Dong, S. Qian, J. Zhang, J. Chen, F. Wang, X. Zhao, L. Wei, W. Li, S. Wang, R. Xu, Y. Cao, L. Chen, Q. Wu, H. Gu, L. Lu, K. Wang, D. Lin, G. Shen, X. Zhou, L. Zhang, Y. Zang, X. Dong, J. Wang, B. Zhang, L. Bai, P. Chu, W. Li, J. Wu, L. Wu, Z. Li, G. Wang, Z. Tu, C. Xu, K. Chen, Y. Qiao, B. Zhou, D. Lin, W. Zhang, C. He, Mineru2.5: A decoupled vision-language model for efficient high-resolution document parsing, 2025. URL: https://arxiv.org/abs/2509.22186. arXiv:2509.22186.

[71] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, W. Ge, Z. Guo, Q. Huang, J. Huang, F. Huang, B. Hui, S. Jiang, Z. Li, M. Li, M. Li, K. Li, Z. Lin, J. Lin, X. Liu, J. Liu, C. Liu, Y. Liu, D. Liu, S. Liu, D. Lu, R. Luo, C. Lv, R. Men, L. Meng, X. Ren, X. Ren, S. Song, Y. Sun, J. Tang, J. Tu, J. Wan, P. Wang, P. Wang, Q. Wang, Y. Wang, T. Xie, Y. Xu, H. Xu, J. Xu, Z. Yang, M. Yang, J. Yang, A. Yang, B. Yu, F. Zhang, H. Zhang, X. Zhang, B. Zheng, H. Zhong, J. Zhou, F. Zhou, J. Zhou, Y. Zhu, K. Zhu, Qwen3-vl technical report, 2025. URL: https://arxiv.org/abs/2511.21631. arXiv:2511.21631.

[72] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang, I. Stoica, Efficient memory management for large language model serving with pagedattention, in: Proceedings of the 29th Symposium on Operating Systems Principles, SOSP '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 611–626. URL: https://doi.org/10.1145/3600006.3613165. doi:10.1145/3600006.3613165.

[73] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. V. Le, E. H. Chi, Least-to-most prompting enables complex reasoning in large language models, in: The Eleventh International Conference on Learning Representations, 2023. URL: https://openreview.net/forum?id=WZH7099tgfM.

[74] O. Press, M. Zhang, S. Min, L. Schmidt, N. Smith, M. Lewis, Measuring and narrowing the compositionality gap in language models, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 5687–5711. URL: https://aclanthology.org/2023.findings-emnlp.378/. doi:10.18653/v1/2023.findings-emnlp.378.

[75] T. Khot, H. Trivedi, M. Finlayson, Y. Fu, K. Richardson, P. Clark, A. Sabharwal, Decomposed prompting: A modular approach for solving complex tasks, in: The Eleventh International Conference on Learning Representations, 2023. URL: https://openreview.net/forum?id=_nGgzQjzaRy.