THE TMEX BENCHMARK AND A FIXED-TOPOLOGY MULTI-AGENT EXTRACTION FRAMEWORK FOR BUILDING AN AI-READY TOPOLOGICAL MAGNETISM DATABASE

Anonymous authors
Paper under double-blind review

ABSTRACT

Traditional scientific databases have largely relied on experimental data collected and curated by researchers. As data-driven and AI-assisted research has expanded, large language models and multi-agent systems have made it feasible to construct AI-ready databases by extracting structured information from scientific literature at scale. However, topological-magnetism papers span diverse material systems, experimental methods, and physical descriptions, so deciding which scientific dimensions to preserve and how to organize the evidence still requires domain expertise. Together with physics experts, we develop TMEX (Topological Magnetism EXtraction), an expert benchmark that integrates a domain schema, expert-annotated gold records, and a semantic judge. Guided by TMEX, we iteratively refine a fixed-topology multi-agent workflow that routes single-material and multi-material papers, indexes figure evidence to schema fields, and assigns seven scientific dimensions to parallel specialists. The judge's field-level scores localize extraction errors and guide targeted revisions to the agent topology and prompts. Instantiated on GPT-5.5, the fixed topology reaches 0.924 F1 on TMEX (0.942 precision, 0.910 recall). On Eval 53, with the backbone held fixed, DeepSeek-V4-Pro improves from 0.773 to 0.927 F1 (+15.4) at $0.19 per paper. We release TMEX, the pipeline code, and the TMEX database of 8,082 papers.

1 INTRODUCTION

Scientific databases have traditionally been built by hand. The Inorganic Crystal Structure Database took four decades of expert curation to reach its present size (Belsky et al., 2002), and MAGNDATA, the reference collection of experimentally determined magnetic structures, holds more than two thousand entries after a decade of work (Gallego et al., 2016). Large language models have lowered this cost: structured records can now be extracted from the literature at scale, and several materials databases have already been built this way (Court & Cole, 2018; Kononova et al., 2019; Dagdelen et al., 2024; Polak & Morgan, 2024). For topological magnetism, experimental observations are of particular value. Lorentz TEM and magnetic force microscopy confirm that a skyrmion or antiskyrmion lattice exists, magnetometry and transport determine the field–temperature window in which it is stable, and small-angle neutron scattering gives its period. These observations are scattered across thousands of papers with no common record format, and no database yet gathers them for computation, screening, or machine learning across materials. A domain-grade extraction framework could therefore assemble that record at a fraction of the manual cost, giving topological magnetism the structured foundation that the Materials Project and NOMAD already provide for computed materials data (Jain et al., 2013; Draxl & Scheffler, 2018).

In topological magnetism this route meets a domain bottleneck before any model is called. Papers span diverse material systems, experimental methods, and theoretical or simulation approaches. A condensed-matter researcher compares them without effort, but a database can compare them only when the same kind of fact lands in the same field. Deciding which scientific dimensions a topological-magnetism database should preserve, how the evidence should be organized across fields and materials, and where the evidence for each field comes from, still requires domain expertise

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

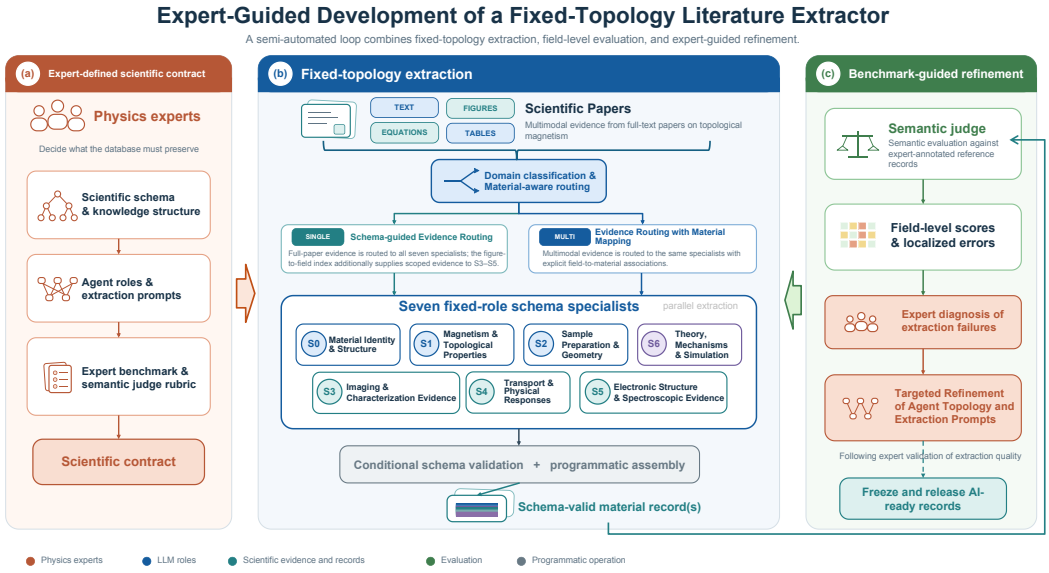


Figure 1: Expert-guided development of the fixed-topology extractor. (a) Physics experts define the schema and knowledge organization, the fixed roles and prompts, and the benchmark together with the semantic-judge rubric. These choices are the scientific contract. (b) The fixed topology classifies a paper as single-material or multi-material, indexes figures onto schema fields for Sections 3–5, assigns the seven scientific dimensions to parallel specialists, and assembles a schema-valid record in code. (c) The semantic judge scores each field against the expert reference and localizes the errors. Experts use that evidence to revise the topology and the prompts. The loop in (c) runs only during development. The topology is held fixed once extraction on the 112 gold papers is stable.

and scientific judgement. On this basis, we investigate the hypothesis that expert-defined scientific structure can serve as an effective inductive bias for LLM-based extraction.

Together with experts in topological-magnetism research, we write the organization required above as TMEX (Topological Magnetism EXtraction). TMEX is a scientific contract with three parts: a seven-dimension domain schema, expert-annotated gold records, and a semantic judge. Its schema is not merely an output format. It specifies which scientific dimensions the database must preserve, how evidence is associated with each dimension, and how extracted content is to be judged, as listed in Table 1. Against this schema, the experts assembled and checked ground-truth records for 112 representative papers in the field. The semantic judge decides whether an extracted answer and its reference express the same scientific fact, and assigns a verdict to each field.

Guided by TMEX, we iteratively refine a fixed-topology multi-agent workflow (Figure 1). The workflow divides the knowledge in the literature into seven scientific dimensions and assigns them to parallel specialist agents for analysis and extraction. The extracted records then enter the TMEX judge, which analyzes them and assigns a field-level verdict. During development, experts inspect the distribution of those verdicts, decide whether a failure comes from evidence routing, role boundaries, agent topology, or prompts, and revise the corresponding part. Through this expert-guided, semi-automated refinement loop, we converge to a framework that can produce a high-quality topological-magnetism database.

Instantiated on GPT-5.5, the fixed topology reaches 0.924 F1 on TMEX (0.942 precision, 0.910 recall) over 112 gold papers. To isolate what the topology itself contributes, we use Eval 53, a comparison set of 53 papers and 1,329 gold leaves selected from the 112-paper gold so that their host-family distribution follows the gold and each paper has sufficient schema-supported fields for scoring. Every baseline and control is scored on this identical set. On Eval 53, the same DeepSeek-V4-Pro backbone improves from 0.773 F1 in a single schema-conditioned call to 0.927 in the fixed topology (+15.4), at $0.19 per paper. The model is unchanged, so the gain comes from the organiza-

tion. On this backbone the fixed topology is also stronger than frontier general-purpose single-shot models such as GPT-5.6 Sol and Claude Fable (0.869 and 0.872), and stronger than multi-agent systems that do not fix the topology.

Our contributions are fourfold. First, we introduce TMEX, an expert benchmark for scientific literature extraction in topological magnetism, comprising a seven-dimension scientific schema, expert reference records, and a field-level semantic judge; together they make correct extraction explicit, checkable, and reproducible. Second, we develop a fixed-topology extraction framework that compiles the contract into an executable system—material-aware routing, figure-to-field evidence indexing, specialized extraction roles, and programmatic assembly—refined through expert-guided error analysis and held fixed before evaluation. Third, we provide controlled evidence that explicit scientific organization is a source of extraction quality complementary to model scale: the same models perform substantially better inside an expert-defined organization, whereas unconstrained agent autonomy does not recover this performance. Fourth, we apply the framework to 8,082 papers and release the TMEX database, with an interactive browser, together with the pipeline code and evaluation commands.

2 Related Work

Multimodal databases in topological magnetism. Topological magnetism still lacks a large multimodal database that machine learning can use across materials. MAGNDATA, the reference collection of experimentally determined magnetic structures, holds more than two thousand entries after a decade of curation (Gallego et al., 2016). NEMAD tabulates 67,573 magnetic materials, with 15 features on each entry rather than the imaging, transport, and scattering evidence used to establish a topological spin texture (Itani et al., 2025). Computed catalogues organize electronic band topology, not those experiments (Xu et al., 2020). Lorentz images, field–temperature windows, and helix periods remain in individual papers, without a common record a model can train on.

Literature extraction. Early extraction from the literature took pieces out of a paper: entities and relations in abstracts (Luan et al., 2018), chemical names and properties (Swain & Cole, 2016), and synthesis steps (Kononova et al., 2019). Later methods specify the output more tightly. GoLLIE places annotation guidelines in the prompt (Sainz et al., 2024). Grammar-constrained decoding forces the generated string to be syntactically legal (Geng et al., 2023). A substantial subsequent literature applies language models to this extraction. Three problems remain. First, large collections are released without a public benchmark. Polymer Scholar reports 24 properties drawn from about 681,000 papers (Gupta et al., 2024). Alloy extraction uses a vocabulary of 354 properties (Kamatchi Sundaram et al., 2026). NEMAD lists 67,573 magnetic materials (Itani et al., 2025). The checks stay with the authors, so another group cannot rerun them. Second, where a public benchmark does exist, the schema is not yet a rich description of multi-dimensional scientific knowledge, and the score is not a judgment of scientific meaning. MatSci-NLP and POLYIE score entities, relations, and property values (Song et al., 2023; Cheung et al., 2024). Dagdelen et al. (2024) map sentences, paragraphs, and abstracts to records. MatViX uses full papers, expert JSON, and figures, and its public benchmark scores composition and property curves (Khalighinejad et al., 2024). Concrete extraction publishes both a wide attribute list and a human benchmark, and what it scores are mix proportions (Li et al., 2026). Third, where the schema is wide, the data stay small. Beaver lists more than 400 attributes, and the reported experiments use 23 biomedical papers and 20 attributes (Zhang et al., 2026). ExtractBench scores PDF-to-JSON extraction on 35 documents (Ferguson et al., 2026). No work yet publishes all three together: a record suited to complex, multi-dimensional scientific knowledge, an LLM extraction that is scientifically strict and reproducible, and coverage of a large set of literature entries. TMEX carries out these three dimensions for topological magnetic materials.

Field-level scoring. Field verdicts use the MUC-5 categories COR, PAR, INC, MIS, and SPU, with the corresponding denominators (Chinchor & Sundheim, 1993). The primary F1 counts a factually supported but underspecified PAR as correct, the same decision as SemEval type matching (Segura-Bedmar et al., 2013).

3 METHOD

3.1 TASK AND CONTRACT

The input is one complete scientific paper, including the body text, figures, tables, and equations. The output is a set of schema-valid material records, with one record per material.

Each paper is classified as single-material or multi-material. Although a manuscript may mention many chemical formulae, a record is created only for a material on which the paper reports experiments or observations, and a formula that appears only in passing receives no record, so that unsupported entries do not enter the database. In the single-material case, all extracted fields belong to one record. In the multi-material case, each experiment, figure, and measurement must be attributed to the material it describes and stored as a field-to-material mapping, which treats collapsing the evidence for two compounds into one record as an error.

Together with physics experts, we define a seven-dimension schema that specifies which scientific information a topological-magnetism database should preserve. The dimensions follow scientific roles—material identity and magnetism, sample preparation, multimodal experimental evidence, and theory—rather than the table of contents of a paper. Sections 0 and 1 constitute the material card; Section 2 records how the sample was prepared; Sections 3–5 record experimental evidence as images, curves, and spectra; and Section 6 records theory and mechanism. The schema is given in Table 1: each row lists the scientific dimension, its scope, representative field paths, and the form of evidence. In the fixed-topology framework, each section is handled by a dedicated agent. The complete machine-readable tree is part of the release.

Table 1: Expert-locked seven-section schema used by the extractor and the TMEX gold. The table gives the scope and representative field paths of each section; the complete field tree is released with the benchmark.

Section	Scientific dimension	Scope	Representative field paths	Evidence
S0	Material Identity & Structure	Host system, crystal symmetry, inversion, and magnetic sites.	`material_system`; `space_group`; `magnetic_atoms`	Text
S1	Magnetism & Topological Properties	Magnetic order, transitions, anisotropy, and topological spin texture.	`magnetic_order`; `ground_state_spin_texture`; `topological_spin_texture`	Text
S2	Sample Preparation & Geometry	As-grown morphology, synthesis, and device patterning.	`sample_morphology`; `growth_method`; `patterning`	Text
S3	Imaging & Characterization Evidence	Diffraction, SANS, and real-space magnetic imaging.	`magnetic_imaging.Lorentz_TEM`; `SANS`; `XRD`	Figures; text
S4	Transport & Physical Responses	Magnetization, susceptibility, resistivity, and Hall responses.	`M_H`; `M_T`; `Hall_effect`	Figures; text
S5	Electronic Structure & Spectroscopic Evidence	Electronic, STM/XMCD, and magnon spectroscopic measurements.	`band_structure`; `STM_STS`; `magnon_spectrum`	Figures; text
S6	Theory, Mechanisms & Simulation	Hamiltonian, competing phases, and numerical modelling settings.	`Hamiltonian`; `competing_phases`; `micromagnetic_simulation`	Text; equations

TMEX supplements this schema with expert-annotated ground truth and a field-level scoring rubric. Physics experts compiled and verified ground-truth records for 112 representative papers, including both single-material and multi-material cases. The semantic judge compares each extracted field with the corresponding ground-truth entry at the level of scientific meaning and assigns a field-level verdict. Following the MUC-5 categories and denominators (Chinchor & Sundheim, 1993), each predicted leaf is labelled COR, PAR, INC, MIS, or SPU. COR denotes a match in scientific fact at comparable specificity; PAR denotes a factually supported but underspecified statement; INC denotes an incorrect value; MIS denotes a supported field that is absent; and SPU denotes a prediction that the paper does not support. We compute precision, recall, and F1 from these labels as

$$P = \frac{\mathrm{COR} + \mathrm{PAR}}{\mathrm{COR} + \mathrm{INC} + \mathrm{PAR} + \mathrm{SPU}}, \quad R = \frac{\mathrm{COR} + \mathrm{PAR}}{\mathrm{COR} + \mathrm{INC} + \mathrm{PAR} + \mathrm{MIS}}, \quad F1 = \frac{2PR}{P + R}. \quad (1)$$

The numerator counts a factually supported but underspecified PAR as correct, the same decision as SemEval type matching (Segura-Bedmar et al., 2013). The semantic judge is an agent system whose decisions the same experts iteratively revised. On a spot check of 466 fields from 10 papers, its agreement with expert judgement reaches 0.991. Experts inspect the judge's verdicts, together

with the distribution of errors over the schema, in order to identify failures, localize their source, and revise the corresponding part of the framework. This revision takes place while the topology is still under development.

3.2 FIXED-TOPOLOGY PIPELINE

A topological-magnetism paper is converted into schema-valid records through our fixed topology in the following stages: domain classification, material-aware routing, figure-to-field indexing, seven parallel dedicated agents, and programmatic assembly (Figure 1). Each step takes a structured intermediate from the previous step and passes one to the next, and dedicated agents do not read one another's outputs, so the context of each role is only the intermediate assembled by the topology.

The paper first goes through the *Category* and *Router* agent, which assigns a topic label, classifies the paper as single-material or multi-material, writes a paper-level header, and lists the materials on which the paper reports experiments or observations. In the multi-material case, later steps must map each experimental measurement and each theoretical property to the material it describes.

The *FigIndex* agent then maps materials and figures to a section and a legal field path and attaches the caption evidence in scope, so that the imaging, transport, and spectroscopy agents (S3–S5) receive only the figures assigned to them.

The seven agents then run in parallel, and each is dedicated to gathering and processing the information of one scientific dimension. Each receives the paper text, the paper-level header, the slice of the schema it owns, and a role-specific prompt written with the physics team, and S3–S5 also receive their scoped figures. They do not communicate. The split keeps any one agent from holding the entire leaf tree and narrows each role to field extraction in one scientific dimension.

For a single-material paper the seven slices are concatenated into one record, whereas for a multi-material paper the agents emit field–material–value mappings and a deterministic procedure builds one record per material. *FieldChecker* then checks the assembled record against the allow-list of legal field names and drops or relocates keys that the schema does not permit. After assembly and checking, the paper is represented by schema-valid material records.

3.3 EXPERT-GUIDED REFINEMENT

The schema-valid records produced by this topology are assessed by the TMEX judge, which compares each filled leaf with the corresponding ground-truth entry and assigns a field-level verdict and score. From the distribution of these verdicts over the schema, and before the framework has matured, the experts attribute each failure to evidence routing, role boundaries, the topology, or the prompts. Revision of the corresponding structure or content improves subsequent extraction.

Early diagnostics showed that incorrect and missing leaves were concentrated in the imaging section, and that a single role was required to locate figures, assign them to fields, and extract measurement conditions in one pass. For example, the experts had previously observed that the same figure was repeatedly assigned to Sections 3, 4, and 5. A *FigIndex* agent was therefore introduced to map each figure onto a section and a legal field path before the imaging, transport, and spectroscopy agents (S3–S5) receive their evidence, which substantially reduced this class of error.

The refinement cycle is semi-automated: extraction and field-level scoring are performed automatically, whereas diagnosis and revision remain the responsibility of the experts. Iteration of this cycle improves extraction until the framework produces stable, high-quality records. The topology, the role assignments, and the prompts are then held fixed, so that the agent responsible for the next step is prescribed by the topology rather than selected by the model. The comparisons in Section 4.2 are restricted to this fixed-topology pipeline.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

On TMEX the fixed topology instantiated on GPT-5.5 reaches 0.924 F1 (0.942 precision, 0.910 recall) over 112 gold papers: 0.926 on 96 single-material papers and 0.911 on 16 multi-material papers.

With that topology fixed, all subsequent experiments evaluate systems on Eval 53, a comparison set of 53 papers and 1,329 gold leaves. Figure 2 shows the host-family distribution. All systems are evaluated on this identical set.

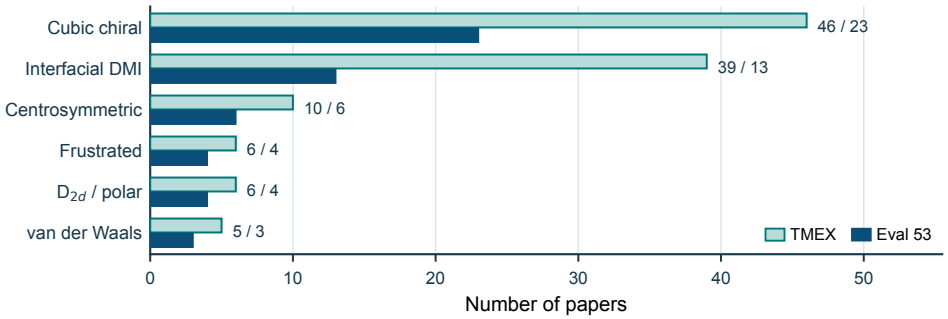


Figure 2: Host-family coverage of the 112 TMEX gold papers. Light bars are the gold set; dark bars are Eval 53.

Metrics. Each predicted leaf is labelled COR, PAR, INC, MIS, or SPU as in Section 3.1, and precision, recall, and F1 follow the formulas given there. The primary F1 treats a factually supported but underspecified PAR leaf as correct. Appendix C reports the same systems under two alternative scoring schemes: Partial, the MUC-5 half-credit numerator (Chinchor & Sundheim, 1993), and CoNLL Strict (Tjong Kim Sang & De Meulder, 2003), which counts COR only. Scores are macro-averaged: one $(P, R, F1)$ per paper, then the mean over the 53 papers. An empty extraction receives $P = R = F1 = 0$ and remains in the average. An empty record is treated as an extraction failure: the schema is heavy, and models can be unstable on this task.

Cost. Cost is the vendor list price of extracting all 53 papers. We report the total list-price cost, in USD, on Eval 53.

Three evaluations. The fixed topology is not bound to one model. We evaluate along three axes. The first comparison places the fixed topology on DeepSeek-V4-Pro against single-shot extractors, including GPT-5.6 Sol, Claude Fable, GLM-5.3, Gemini 2.5 Pro, Qwen 3.8, Kimi K3, and Doubao Seed 2.0 Pro, and against multi-agent systems that do not adopt our topology. The latter comprise an open supervisor on DeepSeek-V4-Pro, free to instantiate auxiliary agents without our role constraints, and a coding-agent harness (Claude Code) instantiated with DeepSeek-V4-Pro and with Claude Fable. We label the DeepSeek-V4-Pro instantiation of the fixed topology Ours: it is the star in Figure 3a. The second isolates the contribution of the topology by holding the model fixed. For the frontier models of three providers, DeepSeek-V4-Pro, Claude Fable, and GPT-5.6 Sol, each model is evaluated both as an independent extractor and as the backbone of the fixed topology, and we report the gain from the framework. The third assesses the dedicated roles by ablating FigIndex, FieldChecker, or the seven-section split, with the remaining topology held on DeepSeek-V4-Pro so that the model does not change. All systems are scored by the same judge against the same gold records. Table 2 lists every arm with precision, recall, empty count, and cost.

4.2 MAIN RESULTS

Figure 3 reports the first two evaluations of Section 4.1 on Eval 53. Panel (a) places the fixed topology against single-shot extractors and multi-agent systems that do not adopt our topology.

Panel (b) shows the per-paper distribution of the stronger systems in panel (a). Panel (c) holds the model fixed.

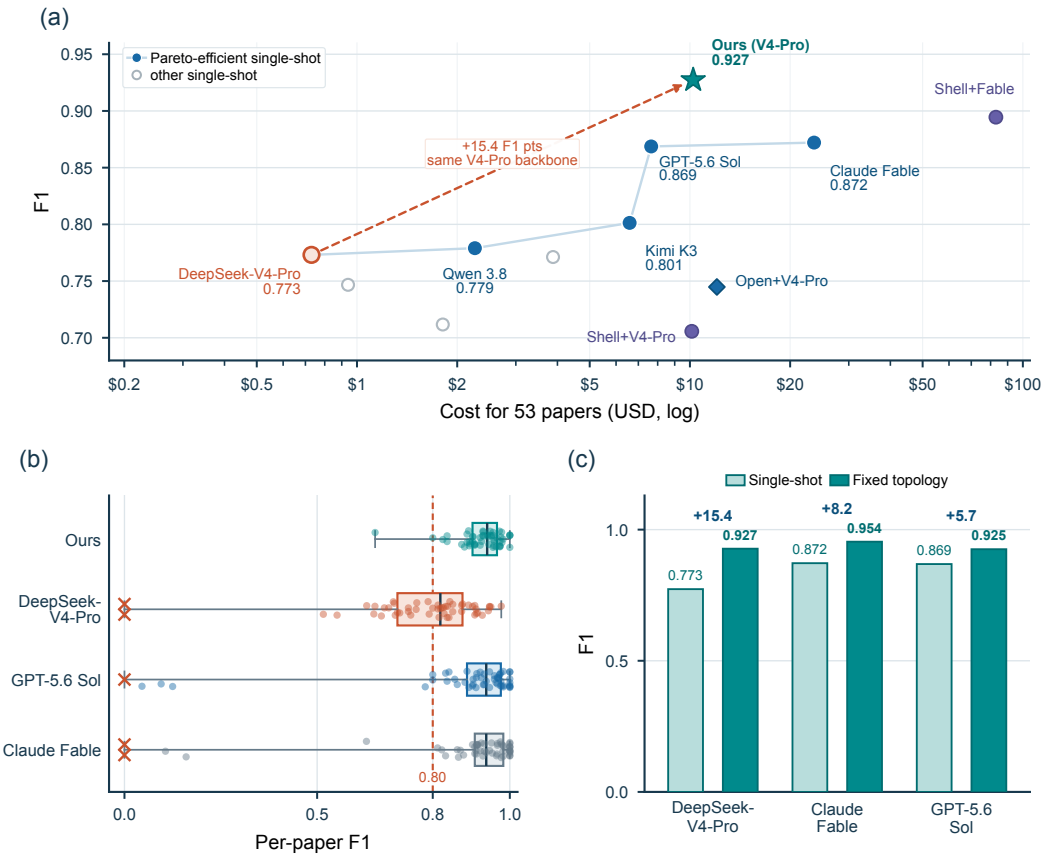


Figure 3: Results on Eval 53. F1 is that of Section 3.1: a factually supported but underspecified PAR leaf counts as correct. (a) F1 versus list-price extraction cost for all 53 papers. Cost is on a log scale, and the F1 axis is cropped to the observed range. The blue curve connects the Pareto-efficient single-shot extractors. Gray open markers are the remaining independent extractors, and the star is our fixed topology on DeepSeek-V4-Pro. Open denotes the open supervisor and Shell the coding-agent harness. The orange dashed arrow marks the F1 gain of the fixed topology on DeepSeek-V4-Pro (+15.4). (b) Per-paper F1 for the V4-Pro fixed topology, DeepSeek-V4-Pro, GPT-5.6 Sol, and Claude Fable. Boxes and whiskers span the full range. Empty extractions score zero and are marked by crosses. The dashed line marks 0.80. (c) Comparison, for each backbone, of an independent extractor against the same model as the backbone of the fixed topology.

Single-shot extractors and unconstrained multi-agent systems. The fixed topology instantiated on DeepSeek-V4-Pro, reported as Ours, reaches 0.927 at $10.23 for the 53 papers ($0.19 per paper). This score is higher than that of every single-shot extractor and every unconstrained multi-agent system in the comparison.

Among single-shot extractors, GPT-5.6 Sol reaches 0.869 at $7.65 and Claude Fable 0.872 at $23.60 (Figure 3a). DeepSeek-V4-Pro used independently reaches 0.773, and the fixed topology raises the same backbone by 15.4 points. In Figure 3a, Kimi K3 reaches 0.801 at $6.59 and Qwen 3.8 0.779. Table 2 also lists GLM-5.3 at 0.771, Gemini 2.5 Pro at 0.712, and Doubao Seed 2.0 Pro at 0.747.

The open supervisor on DeepSeek-V4-Pro scores 0.745 with six empty extracts. The coding-agent harness (Claude Code) scores 0.706 with seven empty extracts when instantiated on DeepSeek-V4-Pro, and 0.895 at $82.99 when instantiated on Claude Fable, with one empty extract remaining.

On the per-paper view in Figure 3b, Ours returns a record for every paper. Its standard deviation is 0.060, and 52 of the 53 papers lie at or above 0.80. DeepSeek-V4-Pro used independently has 22 papers below 0.80. Claude Fable and GPT-5.6 Sol perform similarly to each other, but both leave empty extracts and both have large standard deviations of 0.242 and 0.239.

Contribution of the topology. Holding the model fixed isolates the topology (Figure 3c). DeepSeek-V4-Pro rises from 0.773 as an independent extractor to 0.927 as the backbone of the fixed topology (+15.4), and both empty extracts are removed. Claude Fable rises from 0.872 to 0.954 (+8.2). GPT-5.6 Sol rises from 0.869 to 0.925 (+5.7). The Claude Fable instantiation costs $246.51 for the 53 papers, and the GPT-5.6 Sol instantiation costs $79.68.

4.3 Ablation

The third evaluation ablates the dedicated roles, with the remaining topology instantiated on DeepSeek-V4-Pro (Figure 4(a) and Table 3). Removing the seven-section split reduces F1 from 0.927 to 0.571 (−35.7) and produces 17 empty extracts. Removing FigIndex costs 3.9 points, and removing FieldChecker costs 2.1 points.

5 Discussion

When used as an independent extractor, DeepSeek-V4-Pro attains an F1 of 0.773 and already fails on two of the 53 papers. In the fixed topology on DeepSeek-V4-Pro, it reaches 0.927 at $10.23 and returns a record for every paper, exceeding independent GPT-5.6 Sol (0.869, one empty extract) and Claude Fable (0.872, two empty extracts). The improvement is not an increase in model capacity: without the topology, V4-Pro remains well below Sol. It follows from assigning a seven-dimension record, whose supporting evidence is distributed across text, figures, and tables, to dedicated roles, so that this backbone is no longer required to reconstruct that organization in a single pass.

On Eval 53, the stability of the fixed topology on DeepSeek-V4-Pro shows up paper by paper. It returns a record for every paper, and 52 of the 53 papers lie at or above 0.80. DeepSeek-V4-Pro used independently has 22 papers below 0.80. Claude Fable and GPT-5.6 Sol reach an acceptable F1, but their lower tails are wider and include empty extracts.

Unconstrained multi-agent search does not improve on the fixed topology. On DeepSeek-V4-Pro, the open supervisor scores 0.745 with six empty extracts at $12.05, and the coding-agent harness scores 0.706 with seven empty extracts at $10.13, against 0.927 and no empty extract for the fixed topology on DeepSeek-V4-Pro at $10.23. Moving the coding-agent harness to Claude Fable raises the score only to 0.895, at $82.99, and one paper still returns no record.

Holding the backbone fixed shows that the same organization remains beneficial when the model is already strong. The topology supplies a domain-specific organization of this literature: seven scientific dimensions, figure-to-field indexing, and programmatic assembly. A single general-purpose call does not itself produce that organization, so Claude Fable still gains 8.2 points and GPT-5.6 Sol 5.7 (Figure 3c).

Ablating dedicated roles, with the remainder of the topology held on DeepSeek-V4-Pro, isolates which part of this organization carries the load. Removing the seven-section split reduces F1 from 0.927 to 0.571 (−35.7) and yields 17 empty extracts: a single call is again required to populate the entire heavy schema, which is the failure the parallel assignment was introduced to prevent. FigIndex and FieldChecker are narrower. Their removal costs 3.9 and 2.1 points, concentrated in recall. They do not substitute for the split; they reduce the omission of observational and experimental leaves, for example Lorentz images and transport measurements.

The Eval 53 comparison therefore leads us to the DeepSeek family. As reported in Table 2, the fixed topology on DeepSeek-V3.2-Exp reaches 0.878 F1, above the GPT-5.6 Sol single-shot score of 0.869 and below the fixed topology on DeepSeek-V4-Pro at 0.927. However, the fixed topology on DeepSeek-V3.2-Exp costs $3.69 for the 53 papers, compared with $10.23 on DeepSeek-V4-Pro.

We select this lower-cost backbone for large-scale production because its F1 remains satisfactory. We apply it to 8,082 papers and release the TMEX database. Each record includes a chemical formula, magnetic order, basic sample information, multimodal experimental evidence from the source paper—for example field-indexed imaging figures—and a DOI linking to the original article. An anonymized interactive browser is provided in the supplementary material. The records are LLM extractions and should be checked against the original articles.

6 LIMITATIONS

The size of Eval 53 was set by experimental time and cost; a larger comparison might yield more persuasive results.

The topology is weaker on review articles. A review covers many materials, and figures and experimental observations are interleaved across compounds, so routing and figure-to-field assignment—designed for a paper that reports its own measurements—become less reliable. The TMEX gold is assembled mainly from experimental papers.

The released TMEX database contains 8,082 papers extracted with DeepSeek-V3.2-Exp. This public set is not a second gold set. Neither this extractor, nor a human expert, nor a current model can ensure that every field is correct, and the extracted record inherits deficiencies of the source article. The corpus is therefore drawn from well-recognized journals in the field. The records should still be checked against the original articles.

7 CONCLUSION

TMEX makes correct extraction explicit for topological magnetism: a seven-dimension schema, expert reference records, and a field-level semantic judge. A fixed topology carries that contract out and compiles the expert organization into an inductive bias for a general-purpose model. On Eval 53, with the backbone held fixed, DeepSeek-V4-Pro gains 15.4 F1 points and returns a record for every paper. The model is unchanged, so the gain comes from this scientific organization and is complementary to model scale, while systems that leave the organization unfixed do not recover it. We release the TMEX benchmark, the pipeline code, and the TMEX database of 8,082 papers. Each entry includes material identity, magnetic order, multimodal experimental evidence from the source paper, and a DOI, together with an interactive browser.

DATA AVAILABILITY

An anonymized snapshot of the public papers and their browser is included in the supplementary material. The website address will be included in the camera-ready version.

AI USE STATEMENT

In this work we used generative AI tools to draft and edit the manuscript. We have not used generative AI tools to generate synthetic evaluation data, to develop theoretical models or proofs, to propose hypotheses, or to design the Eval 53 experiments. The remaining required-disclosure tasks are not applicable. We have reviewed the AI-assisted text and take responsibility for the final content of this work.

ETHICS STATEMENT

The TMEX schema and gold were written by domain experts. We release the TMEX benchmark and the TMEX database of extracted records; we do not redistribute publisher PDFs. The bulk corpus is drawn from journals that the authors' institutions are licensed to read. The task is scientific information extraction. Released records are model outputs and should be checked against the original articles.

REPRODUCIBILITY STATEMENT

Section 3.1 defines the field-level labels and the F1. Section 4.1 defines the two evaluation scopes, the empty-extract policy, list-price cost, and the compared families. Section 3 specifies the fixed topology and the single-material and multi-material routes. The schema is given in Table 1. Appendices A–D provide the complete Eval 53 results, component and error diagnostics, stricter credits, and cost details. Anonymous supplementary material includes TMEX in full, the Eval 53 identifiers, pipeline code and prompts, commands to reproduce the Eval 53 table, and an anonymized browser for the TMEX database. The public URL will be included in the camera-ready version.

REFERENCES

Alec Belsky, Mariette Hellenbrandt, Vicky Lynn Karen, and Peter Luksch. New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design. *Acta Crystallographica Section B*, 58:364–369, 2002. doi: 10.1107/S0108768102006948.

Jerry Cheung, Yuchen Zhuang, Yinghao Li, Pranav Shetty, Wantian Zhao, Sanjeev Grampurohit, Rampi Ramprasad, and Chao Zhang. POLYIE: A dataset of information extraction from polymer material scientific literature. In *Proceedings of NAACL*, pp. 2370–2385, 2024.

Nancy Chinchor and Beth Sundheim. MUC-5 evaluation metrics. In *Fifth Message Understanding Conference (MUC-5)*, pp. 69–78, Baltimore, Maryland, 1993.

Callum J. Court and Jacqueline M. Cole. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. *Scientific Data*, 5:180111, 2018. doi: 10.1038/sdata.2018.111.

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

Claudia Draxl and Matthias Scheffler. NOMAD: The FAIR concept for big data-driven materials science. *MRS Bulletin*, 43:676–682, 2018.

Nelson Ferguson, Jeffrey Pennington, Narek Beghian, Ameet Mohan, Douwe Kiela, Shubham Agrawal, and Thang H. Nguyen. ExtractBench: A benchmark and evaluation methodology for complex structured extraction, 2026.

Samuel V. Gallego, J. Manuel Perez-Mato, Luis Elcoro, Emre S. Tasci, Robert M. Hanson, Koichi Momma, Mois I. Aroyo, and Gotzon Madariaga. MAGNDATA: towards a database of magnetic structures. I. the commensurate case. *Journal of Applied Crystallography*, 49:1750–1776, 2016. doi: 10.1107/S1600576716012863.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of EMNLP*, 2023.

Sonakshi Gupta, Akhlak Mahmood, Pranav Shetty, Aishat Adeboye, and Rampi Ramprasad. Data extraction from polymer literature using large language models. *Communications Materials*, 5: 269, 2024. doi: 10.1038/s43246-024-00708-9.

Suman Itani, Yibo Zhang, and Jiadong Zang. The northeast materials database for magnetic materials. *Nature Communications*, 16:9415, 2025. doi: 10.1038/s41467-025-64458-z.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013.

Aakaash Kamatchi Sundaram, Mouyang Chakraborty, Sushant M. K. Devathi, Bishnu P. Prusty, and Rohit Batra. Automated extraction of multicomponent alloy data using large language models for sustainable design. *Advanced Science*, pp. e75916, 2026. doi: 10.1002/advs.75916.

Ghazal Khalighinejad, Sharon Scott, Ollie Liu, Kelly L. Anderson, Rickard Stureborg, Aman Tyagi, and Bhuwan Dhingra. MatViX: Multimodal information extraction from visually rich articles, 2024.

Olga Kononova, Haoyan Huo, Tanjin He, Ziqin Rong, Tiago Botari, Wenhao Sun, Vahe Tshitoyan, and Gerbrand Ceder. Text-mined dataset of inorganic materials synthesis recipes. *Scientific Data*, 6:203, 2019.

Zhifei Li, Kai Yang, Qiang He, and Kai Gong. Large language model-enabled automated data extraction for concrete materials informatics. *npj Computational Materials*, 2026. doi: 10.1038/s41524-026-02304-6.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-Task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of EMNLP*, 2018.

Maciej P. Polak and Dane Morgan. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, 15:1569, 2024. doi: 10.1038/s41467-024-45914-8.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. GoLLIE: Annotation guidelines improve zero-shot information-extraction. In *Proceedings of ICLR*, 2024.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. SemEval-2013 task 9: Extraction of drug-drug interactions from biomedical texts. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval)*, pp. 341–350, Atlanta, Georgia, USA, 2013. Association for Computational Linguistics.

Yu Song, Santiago Miret, and Bang Liu. MatSci-NLP: Evaluating scientific language models on materials science language tasks using text-to-schema modeling. In *Proceedings of ACL*, pp. 3621–3639, 2023.

Matthew C. Swain and Jacqueline M. Cole. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. *Journal of Chemical Information and Modeling*, 56(10):1894–1904, 2016.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL*, pp. 142–147, 2003.

Yuanfeng Xu, Luis Elcoro, Zhi-Da Song, Benjamin J. Wieder, Maia G. Vergniory, Nicolas Regnault, Yulin Chen, Claudia Felser, and B. Andrei Bernevig. High-throughput calculations of magnetic topological materials. *Nature*, 586:702–707, 2020. doi: 10.1038/s41586-020-2837-0.

Sheng Zhang, Qian Liu, Ruixiang Luo, et al. Building agent harnesses for scientific curation from multimodal sources, 2026.

A COMPLETE EVAL 53 RESULTS

Table 2 lists precision, recall, F1, empty extractions, and list-price cost for every arm on Eval 53. The 112-paper TMEX score is reported in Section 4.1 and is not a row in this table.

Table 2: Comparison on Eval 53 among the fixed topology, unconstrained multi-agent systems, and independent extractors. Precision, recall, and F1 follow Section 3.1, empty extractions score zero and remain in the average, and cost is the list price in USD for extracting all 53 papers.

Family	System	P	R	F1	Empty	USD/53
Fixed topology	DeepSeek-V4-Pro	0.950	0.912	0.927	0	10.23
	DeepSeek-V3.2-Exp	0.894	0.865	0.878	0	3.69
	Claude Fable	0.959	0.951	0.954	0	246.51
	GPT-5.6 Sol	0.937	0.915	0.925	0	79.68
Multi-agent	Open supervisor + DeepSeek-V4-Pro	0.764	0.730	0.745	6	12.05
	Claude Code + DeepSeek-V4-Pro	0.758	0.672	0.706	7	10.13
	Claude Code + Claude Fable	0.922	0.890	0.895	1	82.99
Independent	GPT-5.6 Sol	0.903	0.862	0.869	1	7.65
	Claude Fable	0.909	0.860	0.872	2	23.60
	Kimi K3	0.837	0.790	0.801	4	6.59
	GLM-5.3	0.774	0.770	0.771	10	3.88
	Qwen 3.8	0.821	0.752	0.779	1	2.26
	DeepSeek-V4-Pro	0.830	0.735	0.773	2	0.73
	Doubao Seed 2.0 Pro	0.849	0.703	0.747	3	0.94
	Gemini 2.5 Pro	0.823	0.654	0.712	1	1.81

B COMPONENT AND ERROR DIAGNOSTICS

Figure 4 shows the P/R/F1 profiles and the INC–MIS error plane. Table 3 lists the complete component scores and costs.

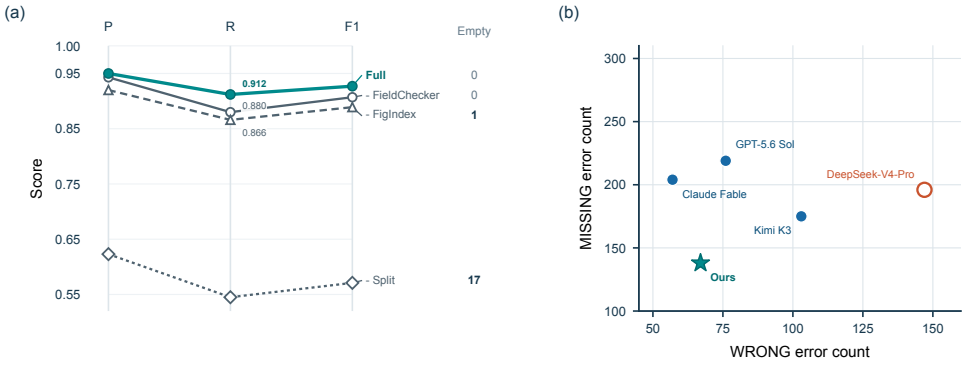


Figure 4: Component and error diagnostics on Eval 53. (a) Precision, recall, and F1 of the fixed topology on DeepSeek-V4-Pro, with one component removed at a time. The right-hand column counts empty extractions. (b) Micro INC and MIS counts (WRONG and MISSING on the axes) over the 1,329 gold leaves.

Table 3: Precision, recall, and F1 of the fixed topology on DeepSeek-V4-Pro on Eval 53, with one component removed at a time. Cost is the list price in USD for the 53 papers.

Configuration	P	R	F1	ΔF1	Empty	USD
Full	0.950	0.912	0.927	—	0	10.23
−FigIndex	0.920	0.866	0.889	−3.9	1	9.10
−FieldChecker	0.943	0.880	0.907	−2.1	0	9.72
−Split	0.623	0.545	0.571	−35.7	17	7.40

C Partial and Strict credits

Table 4 reports Partial (PAR $\times 0.5$), the MUC-5 half-credit numerator (Chinchor & Sundheim, 1993), and CoNLL Strict (COR only) on the same 53 papers, with the same macro-averaging as the Eval 53 table (empty extracts stay at 0). The main-text comparison remains the F1 of Section 3.1. The same-backbone gap of +15.4 F1 becomes +17.5 Partial and +21.3 Strict. Against the GPT-5.6 Sol single-shot, the fixed topology on DeepSeek-V4-Pro remains higher under both stricter credits (Partial +4.2, Strict +3.1).

Table 4: Macro F1 under three credits on Eval 53. F1 is the metric of Section 3.1.

System	F1	Partial	Strict
Fixed topology + Claude Fable	0.954	0.788	0.722
Fixed topology + DeepSeek-V4-Pro	0.927	0.729	0.669
Fixed topology + GPT-5.6 Sol	0.925	0.727	0.666
Claude Fable single-shot	0.872	0.710	0.662
DeepSeek-V4-Pro single-shot	0.773	0.554	0.456
GPT-5.6 Sol single-shot	0.869	0.687	0.638

D COST DETAILS

Table 5 lists the API identifiers and list prices of the backbones in Table 2. Prices are vendor list prices per million input and output tokens. Doubao is priced in CNY (3.2 / 16 CNY per million tokens) and converted at 1 USD = 6.7852 CNY (2026-08-25). The USD totals in Table 2 are computed from these list prices. Judge tokens are never added to extraction cost.

Table 5: API identifiers and list prices, in USD per million input and output tokens, of the backbones in the Eval 53 comparison.

Name in text	API identifier	List price in / out
DeepSeek-V4-Pro	`deepseek/deepseek-v4-pro`	0.66 / 1.98
DeepSeek-V3.2-Exp	`deepseek/deepseek-v3.2-exp`	0.28 / 0.42
Qwen 3.8	`qwen/qwen3.8-max`	2.00 / 6.00
GLM-5.3	`z-ai/glm-5.3`	1.40 / 4.40
Kimi K3	`moonshotai/kimi-k3`	3.00 / 15.00
Doubao	`doubao-seed-2.0-pro`	0.47 / 2.36 (from CNY)
GPT-5.6 Sol	`openai/gpt-5.6-sol`	5.00 / 30.00
Claude Fable	`anthropic/claude-fable-5`	10.00 / 50.00
Gemini 2.5 Pro	`gemini-2.5-pro`	1.25 / 10.00