A Closed-Loop Framework for Material Text Annotation Correction and Validation Based on Large Language Models

Yue Liu*
School of Computer Engineering and Science
Shanghai University
Shanghai, China
yueliu@shu.edu.cn

Zi Liu
School of Computer Engineering and Science
Shanghai University
Shanghai, China
23721510@shu.edu.cn

Wei Zuo
School of Computer Engineering and Science
Shanghai University
Shanghai, China
zuowei@shu.edu.cn

Zhengwei Yang
School of Computer Engineering and Science
Shanghai University
Shanghai, China
zhw_yann@shu.edu.cn

Jie Wu
School of Computer Engineering and Science
Shanghai University
Shanghai, China
24721499@shu.edu.cn

Tianlin Dai
School of Computer Engineering and Science
Shanghai University
Shanghai, China
daitianlin@shu.edu.cn

Siqi Shi
State Key Laboratory of Materials for Advanced Nuclear Energy & School of Materials Science and Engineering
Shanghai University
Shanghai, China
sqshi@shu.edu.cn

***Abstract*—High-quality and standardized annotations are a prerequisite for extracting reliable structure–activity relationships in materials science. Despite rapid progress in extraction models, their performance is fundamentally bottlenecked by annotation noise and inconsistency inherent in domain-specific corpora. While Large Language Models (LLMs) offer a potential solution for automated annotation, they typically fail to generate high-fidelity data in specialized vertical domains due to severe hallucinations and the lack of explicit boundary constraints. To break this quality-efficiency trade-off, we propose CoMETAM (Collaborative Model and Expert-knowledge-constrained Text Annotation for Materials), a closed-loop automated data accuracy governance framework designed to refine noisy annotations into high-quality datasets via multi-model collaboration. CoMETAM first introduces a generic insight metric, operationalized by high-precision extraction models, to precisely identify low-quality datasets that degrade downstream learning. Subsequently, it employs a knowledge-guided structured prompting strategy to inject domain knowledge as rigorous constraints into the inference process, thereby effectively suppressing LLM hallucinations. Finally, a consistency verification module ensures label alignment, closing the loop for automated data hygiene. Experiments on four diverse materials datasets show that CoMETAM consistently boosts downstream extraction performance, delivering F1 gains of up to 13.98%. These results indicate that CoMETAM provides a robust basis for constructing reliable domain knowledge bases.**

Keywords—Material Text Data Accuracy Governance, Large Language Models, Collaborative Annotation Correction.

I. INTRODUCTION

A vast amount of materials science knowledge is embedded in the literature [1–4]. Extracting this knowledge into computable and reusable forms, such as structure–property or composition–processing–property relations, is essential for materials discovery and requires information extraction and knowledge graph technologies to transform unstructured text into structured knowledge [5–8]. Although extraction models have advanced rapidly, their practical deployment is often limited not by model capability, but by the scarcity of reliable annotated data [9]. Building and maintaining high-quality datasets remain labor-intensive, requiring unified label systems, explicit boundary rules, and long-term consistency control across annotators and evolving task definitions [10–12]. As a result, data construction has become one of the most costly and fragile stages in the materials text mining pipeline [7,10].

Recently, LLMs have been increasingly used for automated annotation, data cleaning, and correction, offering substantial efficiency gains [3,7]. However, in specialized domains such as materials science, they are prone to instruction drift and hallucination [13,14], often producing outputs that appear plausible but violate annotation rules or lack textual support [13,15–17]. Existing solutions either rely on expensive manual checking or use unconstrained LLM-based correction without systematic verification [7]. This makes it difficult to align outputs with domain ontologies and prevent explicit annotation noise from being converted into subtler semantic inconsistencies. Therefore, materials annotation governance requires a framework that can both constrain LLM generation with domain knowledge and automatically verify corrected results before write-back, so as to systematically suppress drift and hallucination [14–17].

In this paper, we propose a domain-knowledge-constrained multi-model collaborative closed-loop data accuracy governance framework, named CoMETAM, for the automated correction and verification of materials text annotation data. The core philosophy of CoMETAM is to decouple "generative capability" from "discriminative constraints". By converting

*Corresponding author

domain ontologies and annotation specifications into executable constraints via knowledge-guided structured prompting, we explicitly limit the LLM's output space, reducing instruction drift at the source. Simultaneously, we introduce consistency verification and fallback re-prompting mechanisms to automatically audit entity–relation structures, type constraints, logical consistency, and formatting specifications. Upon detecting conflicts, correction and re-generation are triggered, systematically suppressing hallucinations and ensuring the reliability of the written-back results. Ultimately, the corrected high-consistency annotation data flows back into the dataset, forming an iterative closed loop for quality improvement.

The main contributions of this paper are summarized as follows:

- In order to operationalize automated data hygiene, we propose the CoMETAM closed-loop data accuracy governance framework. This framework integrates "knowledge-guided structured prompting strategy", "consistency verification", and "flow-back iteration" into a viable pipeline for the automated correction and verification of materials annotation data.
- In order to effectively mitigate instruction drift in specialized domains, we propose a knowledge-guided structured prompting strategy. By explicitly encoding domain ontologies and annotation rules as constraints, this strategy significantly alleviates the deviation of LLMs during professional annotation correction.
- In order to systematically suppress hallucinations and implicit semantic noise, we design a consistency verification module and a fallback re-prompting mechanism. These components automatically verify and rectify outputs before write-back, ensuring data reliability at the mechanism level.
- Finally, experiments on four materials datasets demonstrate consistent improvements in downstream performance (achieving a maximum F1 score increase of 13.98%), alongside enhanced intrinsic consistency. Furthermore, the results reveal that logical rectification is predominantly concentrated on samples with medium-to-high triplet density, confirming the framework's capability for the targeted correction of hard-to-annotate cases.

The remainder of this paper is organized as follows: Section II introduces the CoMETAM framework and key module designs; Section III presents the experimental setup and results analysis; Section IV concludes the paper and outlines future work.

II. METHODOLOGY

The overall framework of CoMETAM is illustrated in Fig. 1. The framework is designed to transform noisy raw materials text datasets into high-quality knowledge bases via multi-model collaboration and closed-loop verification. The workflow comprises three primary phases. First, the insight of the dataset is quantified using a joint entity–relation extraction model to identify low-quality datasets (see Section II-A). Subsequently, knowledge-guided structured prompts are employed to steer Large Language Models (LLMs) in performing annotation correction under the constraints of domain knowledge (see Section II-B). Finally, a consistency verification module performs source tracing and logical verification on the generated results, triggering a fallback re-prompting mechanism for samples that fail validation (see Section II-C).

A. Insight-based Quality Assessment

To evaluate the potential learning value of the annotated data for downstream information extraction tasks, this study introduces an insight metric, operationalized as the F1 score achieved by a target extraction model (the evaluator) under a fixed data split. Let the j-th annotated dataset be defined as:

$$S_j = \{(x_i, \tilde{y}_i)\}_{i=1}^{N_j} \quad (1)$$

where x_i denotes an input sentence (or text span) and $\tilde{y}_i$ denotes the corresponding entity–relation annotations. Let $(X_{\text{test}}^{(j)}, \tilde{Y}_{\text{test}}^{(j)})$ be the test split of S_j, and let θ^* be the parameters learned from the training split. The insight metric is defined as:

$$I(S_j) = \text{F1}\big(M_{\theta^*}(X_{\text{test}}^{(j)}), \tilde{Y}_{\text{test}}^{(j)}\big) \quad (2)$$

We quantify the proposed insight metric $I(S_j)$ using the F1 of a joint entity–relation extractor under consistent pre-/post-governance settings. A higher $I(S_j)$ indicates that the annotations better support learning stable, generalizable

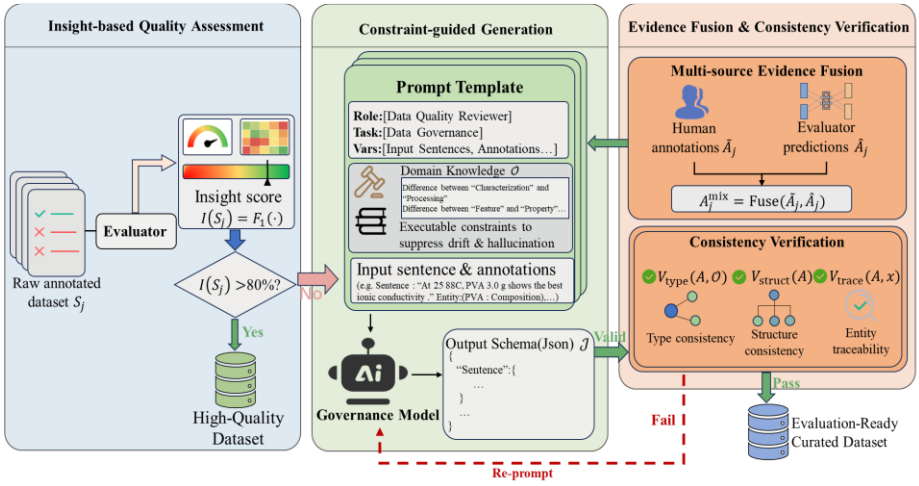
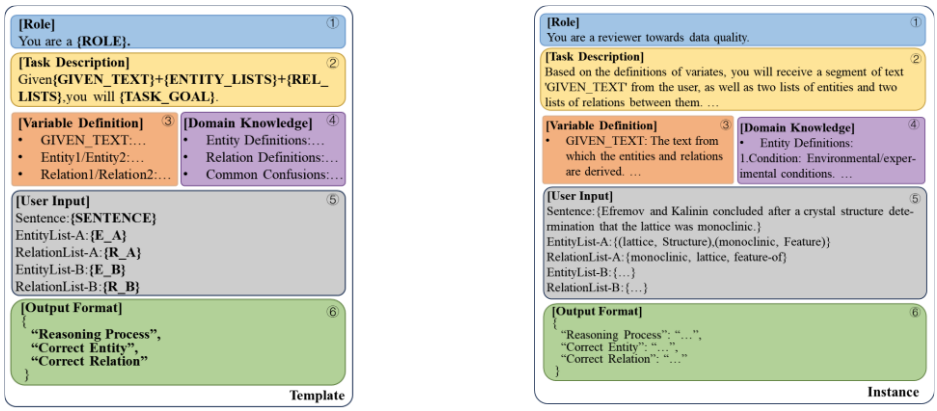


Fig. 1. An overview of CoMETAM.



(a) (b)

Fig. 2. Structured prompt design for annotation correction. (a) Prompt Template; (b) An instantiated prompt example.

patterns, while lower scores suggest noise or semantic ambiguity and thus motivate further governance. In practice, we use an NER-based heuristic: NER F1 ≥ 80% indicates sufficient quality; otherwise, additional governance is likely needed. The model predictions are used only as auxiliary evidence for the subsequent multi-source fusion stage.

B. *Constraint-guided Generation*

During annotation correction, general-purpose LLMs often lack domain expertise and explicit constraint enforcement, leading to hallucinations and guideline drift. As illustrated in Fig. 2, we therefore propose a knowledge-guided structured prompting strategy that encodes entity/relation definitions and decision rules as executable constraints to guide standardized correction, and formalize prompt construction as a template function:

$$p = \mathcal{T}(\text{Role}, \text{Task}, \text{Vars}, \mathcal{O}, \text{Input}, \text{Schema}) \quad (3)$$

where $\mathcal{T}(\cdot)$ denotes the structured prompt template, Role/Task/Vars specify the LLM's role, task description, and variable definitions, $\mathcal{O}$ represents ontology-derived definitions and boundary rules for entities and relations, Input is the context to be corrected, and Schema constrains the output format. Given the prompt p, the LLM produces a correction candidate:

$$\hat{A}_j^{\text{LLM}} = LLM(p) \quad (4)$$

To ensure the output is reliably parsable and suitable for downstream verification, we restrict it to the valid output space $\mathcal{J}$ induced by a predefined JSON schema, and define schema validity using an indicator function:

$$\text{Valid}(\hat{A}_j^{\text{LLM}}) = \mathbf{1}\{\hat{A}_j^{\text{LLM}} \in \mathcal{J}\} \quad (5)$$

When the output violates the schema constraints, fallback re-prompting is triggered to encourage the LLM to produce a schema-compliant correction candidate. By combining explicit domain-knowledge injection with strict output-structure constraints, the proposed prompt design imposes executable restrictions on instruction drift and unsupported inference, thereby improving controllability and verifiability during correction.

C. *Evidence Fusion & Consistency Verification*

To avoid relying solely on the LLM's internal parametric knowledge, which can introduce uncontrolled inferences, we perform correction under multi-source information fusion. Specifically, we organize predictions from a joint entity–relation extraction model, original human annotations, and ontology-grounded definitions into a context-rich and constraint-explicit prompt input, guiding the LLM to correct annotations under professional constraints. Let the human annotations be $\tilde{A}_j = (\tilde{\mathcal{E}}_j, \tilde{\mathcal{R}}_j)$ and the model predictions be $\hat{A}_j = (\hat{\mathcal{E}}_j, \hat{\mathcal{R}}_j)$. We combine them into a unified input structure:

$$A_j^{\text{mix}} = \text{Fuse}(\tilde{A}_j, \hat{A}_j) \quad (6)$$

where $\text{Fuse}(\cdot)$ denotes the information organization and concatenation procedure used in our implementation. This design aims to provide higher-confidence, multi-view evidence to the LLM, thereby reducing entity-type errors and inappropriate relation inferences induced by instruction drift and hallucination.

After the LLM generates a correction candidate, we apply a consistency verification module to ensure that only reliable outputs are written back and that implicit semantic noise does not propagate. We formalize the verification as a conjunctive predicate:

$$V(A, x, \mathcal{O}) = V_{\text{type}}(A, \mathcal{O}) \wedge V_{\text{struct}}(A) \wedge V_{\text{trace}}(A, x) \quad (7)$$

where $\mathcal{O}$ is the ontology constraint set and x is the original input sentence (or text span). The type-compliance term V_{type} checks whether entity and relation types satisfy ontology-grounded definitions and constraints. The structural-consistency term V_{struct} checks whether the head and tail entities in each relation triple are contained in the entity list. Let $A = (\mathcal{E}, \mathcal{R})$, then

$$V_{\text{struct}}(A) = \bigwedge_{(e_h, t_r, e_t) \in \mathcal{R}} (e_h \in \mathcal{E} \wedge e_t \in \mathcal{E}) \quad (8)$$

The source-tracing term V_{trace} verifies that each entity in the entity list has explicit evidence in the original sentence:

$$V_{\text{trace}}(A, x) = \bigwedge_{e \in \mathcal{E}} \text{Support}(e, x) \quad (9)$$

III. EXPERIMENTS

A. *Datasets*

To evaluate generalizability, we conduct experiments on four materials text datasets that differ in materials categories, annotation standards, and knowledge density (Table I). NASICON, AZIB, and CSEM share a unified entity–relation

schema (8 entity types and 9 relation types), while NBSCS follows an alternative annotation standard with a larger label space (11 entity types and 10 relation types), reflecting its distinct domain of nickel-based single-crystal superalloys. In terms of domain coverage, the four datasets span NASICON solid electrolytes, nickel-based single-crystal superalloys, aqueous zinc-ion batteries, and composite solid-state electrolyte materials, thus covering diverse terminology distributions and relation patterns. Moreover, Table I indicates substantial variation in knowledge density, as reflected by the average number of relation triplets per sentence, which further challenges governance robustness across corpora with different information richness. These four datasets are initially annotated corpora and have not undergone any further refinement.

TABLE I. DATASETS STATISTICS.

Dataset	Entity Types	Relation Types	Sentences	Relation Triplets
Nasicon	8	9	2890	10560
NBSCS	11	10	3994	16389
AZIB	8	9	2843	9817
CSEM	8	9	2698	13838

B. Experimental Setup

Models Selection. Our model selection consists of two components: a downstream extraction evaluator used to assess the potential learning value of the dataset for downstream tasks, and governance models used as fully automated correction engines.

- **Evaluator model.** To indirectly assess improvements in data quality, we adopt PURE [18] and SpERT [19] as downstream evaluators. PURE is a strong pipeline baseline that performs span-based entity recognition and relation classification in two decoupled stages, while SpERT is a span-based joint extraction model built on a pre-trained Transformer, which jointly classifies entity spans and predicts relations with localized context. Evaluating governed data on both models allows us to examine the effectiveness of our framework across two complementary extraction paradigms. Importantly, the proposed governance framework is model-agnostic and operates purely at the data level by enforcing ontology-grounded constraints to correct annotation inconsistencies; therefore, the observed gains mainly reflect improved data quality rather than evaluator-specific adaptation.
- **Governance Models.** We employ GPT-4o-mini (GPT) and DeepSeek-R1-Distill-Qwen-14B (DS) as fully automated correction engines. This selection is intended to validate the effectiveness and robustness of the proposed framework across LLM backbones with different parameter scales and capability profiles, while keeping the overall governance procedure unchanged, such that observed differences primarily reflect backbone variance rather than procedural changes.

Parameter settings. For governance models, all generation/inference hyperparameters are kept at their default settings. For the downstream PURE evaluator, we follow a standard training configuration with SciBERT [6] (scibert_scivocab_uncased) as the backbone, using a learning rate of 1e−5 for the encoder and a higher task learning rate of 5e−4 for the task-specific layers, a train batch size of 16, and context_window = 0. The entity recognition module is trained for 50 epochs, while the relation extraction module is trained for 10 epochs. For SpERT, we use MatSciBERT [20] as the backbone, with a learning rate of 5e−5, batch size of 1, 20 training epochs, and 100 negative samples for both entities and relations. We adopt an 8:1:1 train/validation/test split and use Adam as the optimizer. We implement CoMETAM and train all models on an NVIDIA RTX A6000 GPU with 48 GB memory.

Evaluation Metrics. Our evaluation metrics consist of two categories: Extrinsic Metrics are used to quantify the dataset's insight by measuring changes in model performance, while Intrinsic Metrics quantify governance-induced shifts in dataset quality and consistency.

- **Extrinsic Metrics.** We evaluate extrinsic performance using downstream extraction results measured by Precision (P), Recall (R), and F1-score (F1). Specifically, P reflects prediction accuracy, R indicates coverage, and F1, as the harmonic mean of P and R, serves as the primary indicator of overall extraction performance.
- **Intrinsic Metrics.** Beyond extrinsic evaluation, we quantify intrinsic quality shifts induced by governance using dataset-level statistics to assess data quality and consistency. Specifically, Sample Change Rate (SCR) measures revision intensity as the proportion of samples whose annotations are modified, i.e., $SCR = N_{changed}/N_{total}$. Entity/relation volume shifts are computed as net deltas $\Delta Ent = |Ent_{after}| - |Ent_{before}|$ and $\Delta Rel = |Rel_{after}| - |Rel_{before}|$. Entity Type Conflict Rate (ETCR) assesses typing consistency by measuring the share of unique entities that are assigned three or more distinct entity types when aggregating all their occurrences across the dataset, with a higher ETCR indicating more severe type drift and ambiguity. Accordingly, the change in Entity Type Conflict Rate ($\Delta ETCR$) is computed as $\Delta ETCR = ETCR_{after} - ETCR_{before}$. Finally, the Disconnected-pair Deletion Ratio (DPR) characterizes the nature of relation deletions: letting rel_del denote deleted relations and $pair\_removed$ the subset whose corresponding (undirected) entity pairs have zero remaining relations after governance, we compute $DPR = |pair\_removed|/|rel\_del|$. In addition, Entity Type Correction (ETC) counts entity-type corrections, Relation Type Correction (RTC) counts relation-type corrections, and Direction Correction (DirCorr) counts the number of relations whose head–tail direction is reversed from before to after governance.

C. Extrinsic Evaluation

Table II shows the extrinsic effectiveness of CoMETAM across both PURE and SpERT. On all four datasets, governed data consistently improves PURE over the Original annotations for both entity and relation extraction, indicating that the gains mainly arise from enhanced annotation quality rather than evaluator-specific bias. Ours (GPT) yields clear improvements across all corpora, while Ours (DS) also consistently outperforms the Original data and shows complementary strengths, achieving the largest PURE entity-level gain on NBSCS, where F1 increases from 78.72 to 89.18 (+10.46). A similar trend is observed with SpERT, with particularly notable gains on Nasicon and NBSCS. Notably, the largest overall F1 gain in Table II appears on Nasicon relation extraction under

TABLE II.

COMPARISONS OF DOWNSTREAM EXTRACTION PERFORMANCE ACROSS GOVERNANCE STRATEGIES AND EVALUATOR.

Dataset	Version	PURE (Entity)			PURE (Relation)			SpERT (Entity)			SpERT (Relation)		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Nasicon	Original	74.21 ± 1.38	71.70 ± 0.59	72.93 ± 0.64	47.85 ± 2.20	43.15 ± 2.07	45.32 ± 0.68	72.33 ± 1.85	71.05 ± 1.49	71.68 ± 1.62	42.43 ± 3.18	38.89 ± 0.24	40.55 ± 1.58
	Ours (GPT)	**82.69 ± 0.44**	**82.11 ± 0.47**	**82.40 ± 0.07**	**56.52 ± 1.09**	**54.33 ± 0.36**	**55.40 ± 0.57**	**80.65 ± 0.98**	**78.16 ± 0.67**	**79.39 ± 0.82**	**57.80 ± 1.05**	**51.62 ± 0.71**	**54.53 ± 0.13**
	Ours (DS)	81.05 ± 0.90	80.23 ± 0.61	80.64 ± 0.18	54.87 ± 1.13	51.87 ± 0.39	53.33 ± 0.46	76.27 ± 1.47	75.22 ± 0.43	75.74 ± 0.90	51.39 ± 1.61	46.29 ± 2.82	48.67 ± 1.65
NBSCS	Original	77.72 ± 0.36	79.75 ± 0.09	78.72 ± 0.14	67.48 ± 1.77	60.07 ± 0.85	63.54 ± 0.36	76.45 ± 0.40	78.30 ± 0.22	77.37 ± 0.31	61.87 ± 0.67	59.70 ± 1.11	60.76 ± 0.84
	Ours (GPT)	85.97 ± 0.86	86.14 ± 0.45	86.05 ± 0.31	**71.96 ± 0.55**	**68.54 ± 1.56**	**70.20 ± 0.88**	85.45 ± 0.44	84.96 ± 0.44	85.20 ± 0.18	**68.17 ± 1.38**	**66.37 ± 0.50**	**67.25 ± 0.60**
	Ours (DS)	**90.02 ± 0.79**	**88.35 ± 0.19**	**89.18 ± 0.32**	70.99 ± 1.01	66.83 ± 1.05	68.84 ± 0.48	**85.82 ± 0.52**	**86.67 ± 0.32**	**86.24 ± 0.34**	64.89 ± 1.42	65.78 ± 0.70	65.32 ± 0.54
AZIB	Original	78.33 ± 0.69	79.37 ± 0.57	78.84 ± 0.22	57.39 ± 1.66	53.78 ± 0.98	55.51 ± 0.29	74.17 ± 0.63	72.97 ± 0.30	73.56 ± 0.45	52.14 ± 0.37	46.34 ± 0.24	49.07 ± 0.27
	Ours (GPT)	80.85 ± 0.27	81.40 ± 0.18	81.12 ± 0.15	**63.94 ± 1.35**	58.55 ± 0.42	**61.13 ± 0.77**	75.26 ± 0.40	73.64 ± 0.84	74.44 ± 0.50	57.50 ± 1.72	**54.83 ± 1.20**	56.11 ± 0.39
	Ours (DS)	**81.97 ± 0.60**	**82.65 ± 0.52**	**82.31 ± 0.19**	60.31 ± 1.62	**61.38 ± 1.76**	60.81 ± 0.55	**75.70 ± 0.66**	**74.75 ± 0.59**	**75.22 ± 0.61**	**58.43 ± 1.57**	55.99 ± 1.13	**57.16 ± 0.29**
CSEM	Original	68.95 ± 1.44	73.85 ± 1.62	71.30 ± 0.18	50.38 ± 2.70	50.59 ± 1.20	50.43 ± 0.76	73.07 ± 0.91	72.83 ± 0.47	72.94 ± 0.62	49.73 ± 1.18	51.49 ± 1.30	50.58 ± 1.00
	Ours (GPT)	75.55 ± 0.74	76.82 ± 0.37	76.18 ± 0.23	55.97 ± 1.67	50.43 ± 1.39	53.03 ± 0.06	76.61 ± 0.78	76.73 ± 0.99	76.66 ± 0.14	48.50 ± 2.04	50.02 ± 1.21	49.22 ± 0.82
	Ours (DS)	**79.62 ± 0.61**	**79.99 ± 0.19**	**79.80 ± 0.24**	**57.87 ± 2.61**	**52.61 ± 1.85**	**55.05 ± 0.19**	**80.84 ± 0.89**	**80.14 ± 0.35**	**80.49 ± 0.51**	**52.73 ± 0.45**	**54.37 ± 1.14**	**53.53 ± 0.61**

Note: We denote GPT-4o-mini as GPT, and abbreviate DeepSeek-R1-Distill-Qwen-14B as DS.

SpERT, where Ours (GPT) improves F1 from 40.55 to 54.53 (+13.98). Overall, the improvements across PURE and SpERT suggest that CoMETAM improves data quality in a largely model-robust manner and generalizes across both pipeline and joint extraction paradigms.

D. Intrinsic Evaluation

Table III shows that CoMETAM induces large-scale and systematic annotation restructuring rather than random perturbations. The following content is only compared with the dataset version after GPT governance in this article. SCR exceeds 60% across all corpora (up to 82.7% on CSEM), while ETCR@3 consistently decreases, indicating pervasive type drift and semantic ambiguity in the raw data that are rectified by constraint-driven governance. The revision pattern is dataset-adaptive: CSEM is dominated by aggressive denoising with substantial net removals (−1,109 entities; −3,493 relations), whereas Nasicon is more completion-oriented with a net entity increase (+570). Moreover, the consistently high DPR (e.g., 93.2% on Nasicon) suggests that most deleted relations are structurally weak links whose removal fully disconnects the corresponding entity pairs, highlighting a preference for pruning unstable connections to simplify topology and improve graph usability.

TABLE III.

INTRINSIC QUALITY SHIFT STATISTICS BEFORE AND AFTER GOVERNANCE.

Dataset	SCR	ΔEnt	ΔRel	ΔETCR@3	DPR
Nasicon	80.6%	570	-492	-0.17%	93.2%
NBSCS	76.4%	280	-172	-2.9%	60.3%
AZIB	63.8%	59	-388	-0.06%	82.5%
CSEM	82.7%	-1109	-3493	-0.03%	88.0%

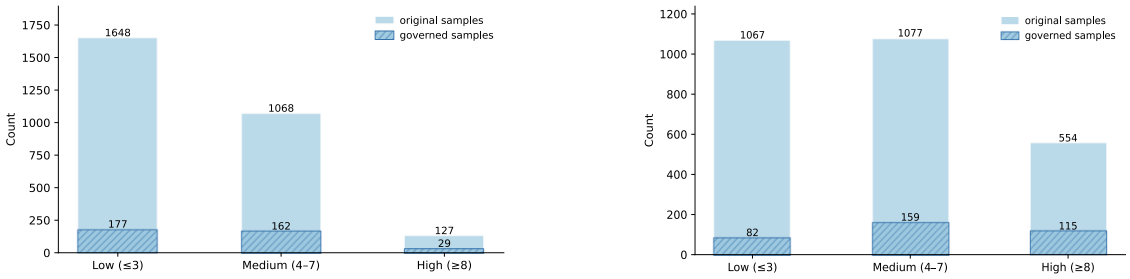
Combined with the fine-grained action breakdown in Table IV, the decrease in ETCR is not primarily driven by simple add/delete operations, but by the joint effect of large-scale type corrections (ETC) and relation-logic–level repairs (RTC and DirCorr). For example, NBSCS exhibits the largest reduction in ETCR@3 (−2.9%), accompanied by substantial cross-type boundary corrections, indicating that knowledge-constrained prompting can systematically rectify entity type drift. Meanwhile, the method also performs high-precision relation-level repairs; for instance, on Nasicon it corrects 310 relation types and fixes 140 direction errors. These repairs eliminate semantic and structural inconsistencies and are therefore critical for improving the usability of the resulting knowledge graph.

TABLE IV.

FINE-GRAINED CORRECTION BREAKDOWN OF ENTITY AND RELATION UPDATES.

Dataset	ETC	RTC	DirCorr	Total
Nasicon	1183	310	140	1633
NBSCS	1502	562	41	2105
AZIB	291	102	62	455
CSEM	274	163	58	495

Taking AZIB and CSEM as the cases, Fig. 3 shows that governance repairs concentrate on medium-to-high triplet-density samples, consistent with the intuition that manual annotation becomes more error-prone as information density and semantic complexity increase. Triplet density is defined by the number of relation triplets in the Original annotation (Low ≤3, Medium 4–7, High ≥8). We focus on fine-grained correction actions and define governance contribution as ETC + RTC + DirCorr; a sample is counted as governed if it triggers at least one such repair. Under this definition, the governance trigger rate rises with density (AZIB: 10.7%/15.2%/22.8%; CSEM: 7.7%/14.8%/20.8%). Moreover, repair operations are skewed toward higher-density groups rather than scaling with sample volume: in AZIB, Medium+High contribute 55.6% of repairs (253/455) while comprising 42.1% of samples; in CSEM, High



(a) (b)

Fig. 3. Original vs. Governed Sample Counts Across Triplet-Density Groups. (a) AZIB dataset. (b) CSEM dataset Triplet density is measured by the number of relation triplets in the original annotation of each sample and used to form three density groups: Low (≤3 triplets), Medium (4–7 triplets), and High (≥8 triplets). "Original samples" denotes the total number of samples falling into each density group. "Governed samples" denotes the number of samples within the same group that were actually modified by governance, i.e., at least one logic-repair operation was triggered when comparing the governed version against the original annotation. Specifically, a sample is counted as governed if it contains any non-zero fine-grained repairs, aggregated as ETC + RTC + DirCorr.

contributes 32.1% of repairs (159/495) with 20.5% of samples. Together, these results indicate that logic-level errors preferentially accumulate in more complex samples, and CoMETAM targets these hard cases with more concentrated corrections.

IV. Conclusion

We propose CoMETAM, a closed-loop framework that refines noisy materials text annotations into high-consistency, reusable corpora. It combines knowledge-guided structured prompting, multi-source evidence fusion, and consistency verification with fallback re-prompting to reduce drift and hallucinations while enforcing ontology-grounded correctness. Across four heterogeneous datasets, CoMETAM consistently improves PURE (up to +13.98% F1) and reduces type conflicts and structural noise via relation-logic repairs, which mainly target medium-to-high triplet-density, hard cases, enabling more reliable downstream information extraction. In the future, CoMETAM can be used to automatically curate high-consistency, large-scale materials knowledge corpora for downstream applications such as robust information extraction, knowledge graph construction, and data-driven materials discovery.

Acknowledgment

This research was supported by the National Natural Science Foundation of China (NSFC) under Grant No.92270124.

References

[1] X. Jiang, W. Wang, S. Tian, H. Wang, T. Lookman, and Y. Su, "Applications of natural language processing and large language models in materials discovery," npj Comput. Mater., vol. 11, Art. no. 79, 2025.

[2] J. Choi and B. Lee, "Accelerating materials language processing with large language models," Commun. Mater., vol. 5, Art. no. 13, 2024.

[3] G. Lei, R. Docherty, and S. J. Cooper, "Materials science in the era of large language models: a perspective," Digit. Discov., vol. 3, pp. 1257–1272, 2024.

[4] J. Dagdelen *et al*., "Structured information extraction from scientific text with large language models," Nat. Commun., vol. 15, Art. no. 1418, 2024.

[5] M. C. Swain and J. M. Cole, "ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature," J. Chem. Inf. Model., vol. 56, no. 10, pp. 1894–1904, 2016.

[6] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A Pretrained Language Model for Scientific Text," in Proc. 2019 Conf. Empirical Methods in Natural Language Processing and 9th Int. Joint Conf. Natural Language Processing (EMNLP-IJCNLP), pp. 3615–3620, 2019.

[7] V. Venugopal and E. Olivetti, "MatKG: An autonomously generated knowledge graph in Material Science," Sci. Data., vol. 11, Art. no. 217, 2024.

[8] Y. Qi, H. Peng, X. Wang, B. Xu, L. Hou, and J. Li, "ADELIE: Aligning Large Language Models on Information Extraction," in *Proc.* 2024 Conf. Empirical Methods in Natural Language Processing (EMNLP), pp. 7371–7387, 2024.

[9] F. Bai *et al*., "Schema-Driven Information Extraction from Heterogeneous Tables," in Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 10252–10273, 2024.

[10] A. N. Rubungo, K. Li, J. Hattrick-Simpers, and A. B. Dieng, "LLM4Mat-bench: benchmarking large language models for materials property prediction," Mach. Learn.: Sci. Technol., vol. 6, no. 2, Art. no. 020501, May 2025.

[11] M. Zaki, Jayadeva, Mausam, N. M. A. Krishnan, "MaScQA: investigating materials science knowledge of language models," *Digit. Discov*., vol. 3, no. 2, pp. 313–327, 2024.

[12] J. Zhang, X. Chen, X. Ye, Y. Yang, and B. Ai, "Large Language Model in Materials Science: Roles, Challenges, and Strategic Outlook," Adv. Intell. Discov., early view, Art. no. e2500085, Jul. 2025.

[13] L. Huang *et al*., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," arXiv:2311.05232, 2024.

[14] S. Geng *et al*., "Generating Structured Outputs from Language Models: Benchmark and Studies," arXiv:2501.10868, 2025.

[15] F. Raspanti, T. Ozcelebi, and M. Holenderski, "Grammar-Constrained Decoding Makes Large Language Models Better Logical Parsers," in Proc. 63rd Annu. Meeting Assoc. Comput. Linguistics (ACL 2025), vol. 6, Industry Track, pp. 485–499, 2025.

[16] K. Park, T. Zhou, and L. D'Antoni, "Flexible and efficient grammar-constrained decoding," arXiv:2502.05111, 2025.

[17] S. S. Rajan, E. Soremekun, and S. Chattopadhyay, "Knowledge-based consistency testing of large language models," in Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 10185–10196, 2024.

[18] Z. Zhong and D. Chen, "A Frustratingly Easy Approach for Entity and Relation Extraction," in Proc. 2021 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 50–61, 2021.

[19] M. Eberts and A. Ulges, "Span-Based Joint Entity and Relation Extraction with Transformer Pre-Training," in Proc. ECAI 2020, Front. Artif. Intell. Appl., vol. 325, pp. 2006–2013, 2020.

[20] T. Gupta, M. Zaki, N. M. A. Krishnan, and Mausam, "MatSciBERT: A materials domain language model for text mining and information extraction," npj Comput. Mater., vol. 8, no. 1, Art. no. 102, 2022.