Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Contribution-Statement-Supervised Author–Method Participation Attribution for Task-Driven Collaborator Recommendation

Chao Chen

cchen@seu.edu.cn

Southeast University

Chunyang Li

Southeast University

Jixiang Sun

Southeast University

Mengcheng Yao

Southeast University

Dandan Wu

Southeast University

Research Article

Keywords: collaborator recommendation, expert finding, knowledge attribution, contribution statements, weak supervision, CRediT

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11590820/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Contribution-Statement-Supervised Author–Method Participation Attribution for Task-Driven Collaborator Recommendation

Running title: Contribution-Scoped Author Attribution

Authors: Chao Chen1,*, Chunyang Li1, Jixiang Sun1, Mengcheng Yao1, Dandan Wu1

Affiliations:

1 Southeast University, 2 Southeast University Road, Jiangning District, Nanjing 211189, China

* Correspondence: cchen@seu.edu.cn

Acknowledgments: The authors would like to thank all individuals who contributed to this study. In particular, we express our sincere gratitude to the Network and Information Center of Southeast University for providing the platform and resources that supported this research.

ORCID: Chao Chen: 0009-0000-3661-546X

Abstract

Task-driven collaborator recommendation requires identifying *which author has carried out which specific methodological work*, yet prevailing author–paper–method aggregation propagates a paper's knowledge indiscriminately to all its authors, causing knowledge conflation. This paper proposes a contribution-evidence-constrained attribution layer: coarse contribution statements and fine human-adjudicated text anchors provide hierarchical weak supervision (automatic anchor extraction is specified for scale-up, not used in the reported runs) which, with author-history priors, learns many-to-many latent author–method–role participation associations — role-resolved internally, verified at the role-aggregated participation level; the corpus-aggregated attribution profile feeds an analytic factorized scoring function for task-driven ranking, trained with supervised attribution losses alone. On a public computational-biology corpus with 3,306 doubly annotated and adjudicated ground-truth cells, the method outperforms the allocation baselines compared on the attribution task (B3 and B5) in role-aggregated attribution accuracy (precision–recall area 1.000 vs 0.980 for the strongest competitor; hard-evidence 1.000 vs 0.890) and improves recommendation over the strongest prior-only baseline under identical inputs (NDCG@10 0.040 vs 0.037, Holm-corrected $p < 0.01$). Two boundary findings are reported with equal prominence: on the adversarially constructed main set all history-based rankers score below random ordering, and an author-level role-

histogram ablation matches the full attribution layer on ranking — the layer's evidenced value is auditable, calibrated, fine-grained participation attribution rather than ranking gain. A per-role analysis bounds the attribution claim to the participation level, and ablations identify text anchors as the lifeline of attribution accuracy.

Keywords: collaborator recommendation; expert finding; knowledge attribution; contribution statements; weak supervision; CRediT

---

1. Introduction

1.1 Background and motivation

The efficiency with which research collaborations are formed directly affects the quality and pace of innovation. As research becomes increasingly interdisciplinary, demand is shifting from "finding someone who works in a field" to "finding someone who can carry out a specific piece of work": a principal investigator needs not "a scholar studying single-cell sequencing" but "a collaborator who can develop graph neural network models for single-cell data". We refer to this demand as *task-driven collaborator recommendation*: queries centre on a concrete method and a required role (e.g., method design, software implementation, data analysis) rather than on paper topics.

At the same time, the CRediT contribution statements now mandated by publishers such as PLOS and the Nature Portfolio record, in machine-readable form, the role categories undertaken by each author — a supervision signal that remains largely unexploited. Whether such statements can move recommender systems from "topical relevance" to "competence matching" motivates this paper.

1.2 Problem statement

Existing collaborator-recommendation and expert-finding methods generally aggregate information along the "author–paper–method" path: once a paper involves a method, knowledge of it is propagated to all of its authors under some weighting, and even scalar schemes such as author contribution indices attribute all methods of a paper in the same proportion. Consider a paper involving single-cell experiments, graph neural network modelling, and clinical sample collection, with three authors each responsible for one of these tasks: path aggregation propagates all three competences to every author, and scalar contribution weighting cannot distinguish *the different roles that the same author plays for different methods within the same paper*.

This paper therefore asks: **how can full paper texts, contribution statements, and author histories be exploited to learn fine-grained author–method–role knowledge attribution, and can this attribution directly improve task-driven collaborator recommendation?** Formally, we infer the latent association Z_{apkr} — author a participated in method k in paper p with role r — and use it to gate knowledge propagation and build

role-resolved author representations (Section 3). The role dimension is an internal modelling structure: the verified primary output is the role-aggregated participation relation (Sections 5.5 and 6.1), with per-role resolution reported separately as a finer-grained check (Section 6.1). Two modelling facts — many-to-many participation (probabilities must not be normalized) and missing declarations (which must not be treated as negative examples) — jointly dictate the asymmetric design of every loss function in this paper (Section 3.4).

1.3 Research questions

- **RQ1 (attribution quality):** Does combining coarse-grained contribution statements with fine-grained text anchors yield more accurate author–method participation attribution (evaluated at the role-aggregated level, Section 5.5) than unsupervised allocation rules and generic weakly supervised models?
- **RQ2 (downstream value):** Under identical input information, temporal splits, and candidate sets, can fine-grained attribution yield better task-driven recommendation than paper-level contribution weighting and role-feature GNNs?

The fairness constraint of RQ2 (Section 5.4) is central to our experimental design: all compared methods receive identical textual and contribution information, so information access is controlled across models; the B7 ablation and the same-scorer control (Table 4) then test the independent contribution of the entity-level attribution layer — and the current results show no additional ranking gain from it (Section 6.2).

1.4 Contributions

1. **Method:** A contribution-evidence-constrained latent-association learning layer unifies coarse CRediT statements and fine text anchors (author–action–object triples) as hierarchical weak supervision over author–method–role associations $\hat{Z}_{apkr}$, handling many-to-many participation and missing declarations with noisy-or aggregation, prior regularization, and PU-style asymmetric losses (Section 4.3); in the reported configuration the layer is trained with supervised attribution losses only (statement consistency plus human-adjudicated anchors), and its corpus-aggregated profile P_6 feeds an analytic factorized scoring function (Section 4.5), while the knowledge-propagation pathway into role-resolved representations (Section 4.4) and the joint-optimization pathway (Section 4.5.4) are specified but not enabled.
2. **Resource:** We release a reproducible author–method–role evaluation set built on public contribution-statement corpora (PLOS / Nature Portfolio, following the Behind the Byline pipeline), with fine-grained ground truth from dual independent annotation (Cohen's $\kappa \geq 0.7$); the annotation guideline and data are publicly available (Section 5.2).
3. **Empirical evidence:** We provide systematic comparisons on attribution accuracy (RQ1) and task-driven recommendation (RQ2), including five same-information baselines, a random-ordering reference, an author-level role-histogram ablation

(B7), an independently verified query subset, cluster-bootstrap significance, four ablations, and case studies; the evidence shows that role-aware scoring improves recommendation over role-blind scoring, while localizing the attribution layer's own contribution to participation-attribution fidelity, calibration, and auditability rather than ranking gain (Section 6).

The remainder of this paper is structured as follows. Section 2 reviews related work; Section 3 formalizes the problem and its constraints; Section 4 presents the attribution layer and the scoring function; Section 5 describes the dataset, annotation, and experimental setup; Section 6 reports results and analyses; Section 7 discusses limitations; Section 8 concludes.

---

2. Related Work

2.1 Expert finding and collaborator recommendation

Expert finding retrieves researchers with relevant expertise given a topic or task description. Early methods modelled author profiles with language and topic models via the "author–paper–word" path; recent work learns embeddings on "author–paper–topic/keyword" heterogeneous graphs for collaborator ranking. Most representatively, GraphRec [3] incorporates an author contribution index into author–abstract connections for contribution-weighted expert recommendation — the closest work to ours in application setting.

The common premise of this line is that paper-level topical relevance suffices to represent author competence; scalar contribution weighting distinguishes "who contributed more" but not "who did which task". We show (Sections 6.2–6.3) that adding role information to the shared history prior yields a moderate re-weighting benefit under task-driven queries; a role-histogram ablation and a same-scorer control localize this benefit to author-level declared-role statistics, while the entity-level attribution layer itself shows no additional ranking gain (Section 6.2).

2.2 Author topic modelling and knowledge attribution

A second related line asks to whom the knowledge in a paper should be attributed. HDPauthor [12] introduces a latent author label for each word in a topic model, jointly learning author topic distributions and word-level contribution allocation — a foundational author-level attribution work whose granularity stays at the word/topic level with no real contribution evidence. HeteAlloc [4], closest to our motivation, allocates topic knowledge to authors on an "author–paper–MeSH" network by path similarity combined with author history; its allocation rule is a fixed unsupervised heuristic that cannot adapt to tasks and evidence. AT(credit) [11] compares authorship credit allocation schemes (equal, harmonic, axiomatic) for author-topic models and explicitly names exploiting the *content* of contribution statements as an important future direction — this paper advances along that

direction, upgrading contribution statements from authorship weights to role–method association supervision supported by text anchors.

The shared limitation of all three is that "attribution" operates at the topic or word level and that the allocation mechanism is a generative prior or a fixed rule; our attribution target is instead a four-way (author × method × role) association whose allocation coefficients are learned from evidence supervision.

2.3 Exploiting contribution statements and the CRediT taxonomy

CRediT (an ANSI/NISO standard [6]) standardizes research contributions into 14 roles and is now widely mandated by publishers such as PLOS and the Nature Portfolio. Recent work has begun to mine this resource: Behind the Byline [1] studied the extraction, disambiguation, and role mapping of author contributions on a large-scale contribution-statement corpus, releasing its pipeline and data; "Task specialization across research careers" [8] used contribution data to analyse task division across research careers, validating the reliability of contribution statements for individual-level behavioural analysis. On the modelling side, LLM-based role recognition [9] showed that large models can identify leadership and direct-support role categories from contribution statements.

This line has addressed the acquisition, cleaning, and coarse role recognition of statements, but stops at the role-category level: the correspondence between roles and method entities — which author undertakes which role for which method — has not yet been modelled as a learnable inference problem. Our contribution-statement parsing (Section 4.2.2) directly reuses the public pipeline of this line and pushes the inference target to method–role granularity.

2.4 Summary of differences

Table 1 summarizes how the nearest works differ from ours. One verification caveat: our examination of GraphRec is based on its published description, and we cannot exclude implementation details beyond it; the experiments (baseline B2 in Section 5.3) therefore use its publicly reproducible "contribution-index weighting" mechanism as the comparison basis.

Table 1. Positioning of the nearest works and this paper

Work	Attribution granularity	Attribution mechanism	Uses contribution statements	Oriented to recommendation learning
HDPauthor [12]	word/topic–author	generative topic model	No	No
HeteAlloc [4]	topic–author	path similarity (fixed rule)	No (MeSH paths)	No
AT(credit) [11]	paper–author (scalar credit	byline credit-allocation	No	No

Work	Attribution granularity	Attribution mechanism	Uses contribution statements	Oriented to recommendation learning
	share; byline-level, not author–topic inference)	(counting) schemes		
GraphRec [3]	paper–author (scalar weight)	contribution-index weighting	Indirectly	Yes (expert recommendation)
LLM role recognition [9]	role–author	LLM classification	Yes	No
This paper	**method–author participation (verified output); method × role–author as internal structure**	**learned inference constrained by contribution evidence**	**Yes (coarse supervision + fine anchors)**	**Yes (attribution-injected scoring; joint training specified as future work)**

As Table 1 shows, neither "author contribution weighting", "author knowledge allocation", nor "LLM role recognition" alone leaves room for innovation; our contribution concentrates on the **fine-grained, contribution-evidence-constrained attribution inference layer** and its injection into task-driven recommendation scoring. Three neighbouring settings are explicitly out of scope: author name disambiguation (within-paper name matching handled by the Behind the Byline pipeline, cross-paper identity resolution by OpenAlex records), authorship credit allocation (a scalar share problem, versus our method × role structure), and expert finding (topical relevance, versus method–role competence matching).

---

3. Problem Formulation

3.1 Notation

This paper involves five entity sets:

Symbol	Meaning
A	author set, $\lvert A\rvert = N_a$
P	paper set, $\lvert P\rvert = N_p$; $A_p \subseteq A$ denotes the author set of paper p
K	method entity set (shared entity space);

Symbol	Meaning
	$K_p \subseteq K$ denotes the method entities used in paper p
R	contribution role set, following the CRediT taxonomy, $\lvert R\rvert = 14$
Q	task query set

Observed evidence (extraction is handled in Section 4.2; only abstract definitions are given here):

Symbol	Meaning	Property
$C_{apr} \in \{0,1\}$	author a is declared in the contribution statement of paper p to undertake role r	large-scale, coarse-grained, positive-only
T_{apkr}	text anchor: concrete textual evidence (with the original location and a confidence γ) that author a undertakes role r for method k	sparse, fine-grained, noisy
$\pi_{ak} \in [0,1]$	author-history prior: the participation prior probability mapped from the time-decayed historical strength w_{ak} (Eq. (1))	dense, auxiliary
$e_{pk} \in \mathbb{R}^d$	contextual evidence representation of method entity k in paper p	encoded from text

The granularity gap across the three levels is the central tension of the problem: C is observed at the author–paper–role level and π at the author–method level, whereas the knowledge attribution that is actually needed resides at the four-way author–paper–method–role level.

3.2 Latent association variable

Definition 1 (participation association). For any tuple (a, p, k, r), define the latent binary variable

$$Z_{apkr} = \begin{cases} 1, & \text{if author } a \text{ actually participated in the work on method } k \text{ in paper } p \text{ with role } r \\ 0, & \text{otherwise} \end{cases}$$

Z is not directly observable and constitutes the core inference target of this paper; its relationship to the observed evidence is as follows:

- $C_{apr} = 1 \Rightarrow \exists\, k \in K_p,\ Z_{apkr} = 1$ (declaration constraint: declaring a role entails having undertaken at least one method's work);
- the existence of T_{apkr} significantly raises the likelihood that $Z_{apkr} = 1$ (anchor hint, but not certain, because LLM extraction is noisy);
- $C_{apr} = 0$ **does not imply** anything about Z_{apkr} (see Constraint 2 in Section 3.4).

3.3 Task definitions

We study two mutually coupled tasks:

Task 1: knowledge attribution. Given full paper texts, contribution statements, and author histories, learn a mapping ϕ: $\mathcal{E}_{apkr} \mapsto [0,1]$ such that $\hat{Z}_{apkr} = P_\phi(Z_{apkr} = 1 \mid \mathcal{E}_{apkr})$ approximates the true participation association. Evaluation of this task relies on independently human-annotated tuple ground truth (Section 5.2).

Task 2: task-driven collaborator ranking. Given a query $q = (t_q, K_q, R_q)$ (defined in Section 4.5.1), output a ranking $\succ_q$ over the candidate author set such that authors with the required method competence and role experience rank at the top.

The connection is one-directional: in the reported configuration, Task 1's output enters Task 2 only through the corpus-aggregated profile P_6 in the analytic ranking score of Eq. (7); the gating of author-representation construction (Online Resource 1, Section S7) belongs to the specified-but-disabled propagation pathway (Section 4.4). The reverse direction (the ranking loss shaping attribution inference) is specified in Section 4.5.4 but disabled in the reported implementation ($\lambda_3 = 0$); we present it as future work, not as a contribution.

3.4 Modelling constraints

Constraint 1 (many-to-many; no normalization). The same method k in the same paper p may be undertaken by multiple authors, and the same author may undertake multiple methods. Therefore, for any (p, k, r) we **do not impose** the normalization $\sum_{a \in A_p} P\,(Z_{apkr}) = 1$. All participation probabilities are produced independently through a sigmoid (Eq. (3)). Forcing normalization would artificially create competition among authors, contradicting the actual form of scientific collaboration.

Constraint 2 (PU property; missing declarations are not negatives). Contribution statements are "minimal disclosure": journals only require listing the role categories undertaken by each author, and absence from the list does not equal non-participation (common reasons: authorship conventions, statement length limits, omission of minor contributions). Therefore:

- tuples with $C_{apr} = 0$ **do not enter any negative loss**;
- tuples not hit by anchors **do not enter the anchor negative loss**;
- the semantics learned by the model is "**participation associations supported by evidence**", not a complete enumeration of true participation.

This constraint determines the asymmetric design of all loss functions in Section 4.3.3 (positive-side constraints only, plus prior regularization), and its necessity is examined in the ablation of Section 6.3 (a variant that treats missing declarations as negatives).

---

4. Methodology

4.1 Overall framework

The model consists of four sequentially connected layers (Fig. 1):

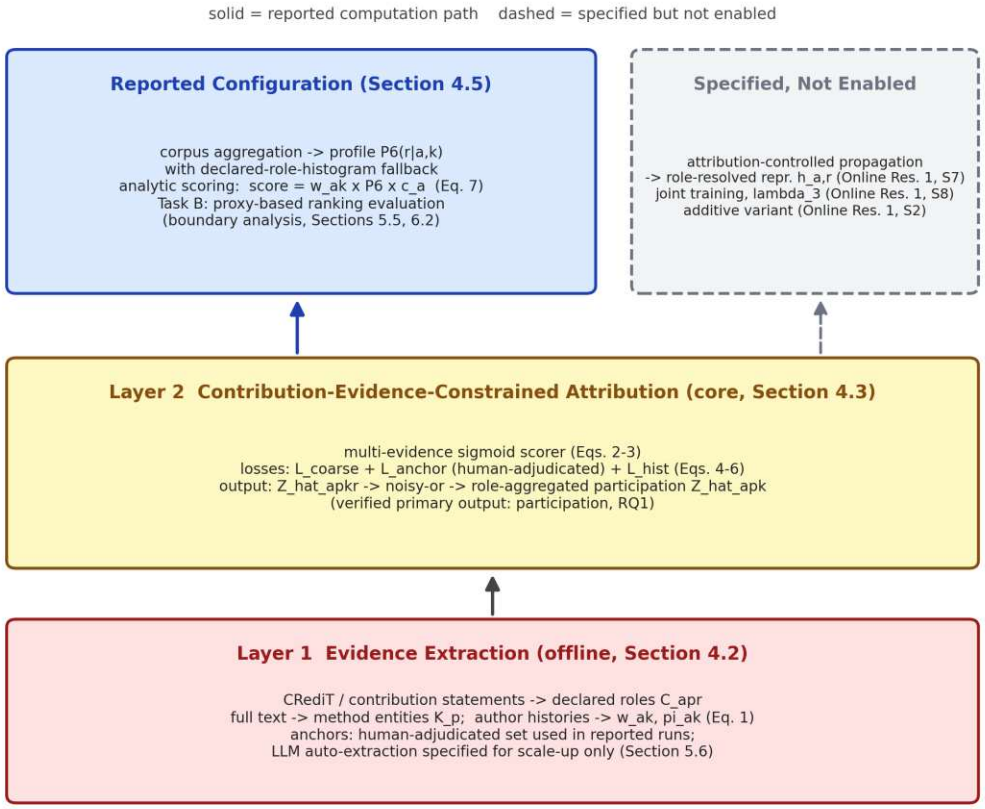


Fig. 1. *Architecture and the actual computation path. Solid arrows show the reported configuration: Layer 1 (offline evidence extraction) feeds Layer 2 (contribution-evidence-constrained attribution, trained with supervised losses over human-adjudicated anchors and CRediT statements), whose outputs — role-aggregated participation probabilities (used for the participation evaluation, RQ1) and per-role channels (used for profile construction) — are corpus-aggregated into the profile P_6 and consumed by the analytic scoring of Eq. (7) (Section 4.5). The dashed branch — attribution-controlled propagation into role-resolved representations (Section 4.4) and joint training (Section 4.5.4) — is specified but not enabled in the reported runs.*

The pipeline (Fig. 1) takes full paper texts, contribution statements, and author publication histories as input. Layer 1 (evidence extraction, Section 4.2) is offline preprocessing executed once per paper, producing C_{apr}, anchors (a, p, k, r, γ), π_{ak}, and e_{pk}; Layer 2 (contribution-evidence-constrained attribution learning, Section 4.3, the core layer) is trained with the supervised attribution losses of Section 4.3.3 and produces $\hat{Z}_{apkr}$; the ranking layer (Section 4.5) consumes these outputs only through the corpus-aggregated profile P_6 in the analytic scoring of Eq. (7). The propagation pathway into role-resolved representations (Layer 3, Section 4.4) and the joint-training variant (Section 4.5.4) are specified but disabled in the reported implementation. Although Layer 1 uses an LLM, all outputs carry original-text locations and confidence scores, and the **fine-grained ground-truth evaluation (Section 5.2) relies on independent manual annotation** — the LLM never validates its own outputs; the anchor supervision of the reported runs likewise comes from the human-adjudicated set (Section 5.2), the LLM anchor-extraction pathway being specified for scale-up only (Section 5.6).

4.2 Evidence extraction

4.2.1 Method entity recognition and linking

Recognition. For each full text (focusing on the Methods/Materials sections), an LLM extracts entity mentions of "methods, models, algorithms, experimental techniques, and dataset-construction procedures", keeping the original span and enclosing section of each mention (few-shot prompt; template and examples in the code repository).

Linking and representation. Mentions are vectorized (same encoder as Section 4.3) and merged by nearest neighbour into a shared method-entity space K when their similarity to an existing entity exceeds τ_{link} (otherwise a new entity is created); a shared space is a prerequisite for the author-history prior (Section 4.2.3) and for query matching (Section 4.5). The evidence representation e_{pk} encodes the context windows of all mentions of k in p with attention pooling (weights learnable, part of ϕ). Recognition precision/recall and linking accuracy are manually verified on a subsample of the Section 5.2 annotation set and feed the noise-sensitivity analysis of Section 7.2.

4.2.2 Contribution-statement parsing and triple extraction

Coarse-grained role annotation C_{apr}. Journal contribution statements (in CRediT or free-text format) are parsed by a hybrid rule-and-LLM procedure: standard CRediT formats are mapped to the 14 roles by rules, and free-text formats are mapped to the nearest roles by the LLM with mapping confidences. We directly reuse the within-paper author-name matching and role-mapping components of the Behind the Byline repository; cross-paper author identity resolution uses OpenAlex records (Section 5.1).

Fine-grained anchor triples. (author, action, object) triples are extracted from contribution statements (e.g., "J.S. implemented the graph neural network" → (J.S., implement, graph neural network)) and from authorship-related main-text sentences (acknowledgements, contribution footnotes, aligned first-person statements); objects are

linked to method entities k (Section 4.2.1) and action verbs are clustered and mapped to roles r. Each triple carries a confidence $\gamma \in [0,1]$ (LLM self-assessment combined with rule checks; composition formula in the repository); only triples with $\gamma \geq \tau$ (0.8) enter the anchor loss of Section 4.3.3; the remainder serve as features only. The reported runs use only the human-adjudicated anchor set of Section 5.2 in the anchor loss; the LLM extraction pathway above is retained for scale-up (Section 5.6).

4.2.3 Author-history prior construction

For all papers $\{p' : t_{p'} < t_p\}$ of author a before the publication time t_p of paper p, we tally method-entity usage with exponential time decay into a raw strength w_{ak}, and map it to a probability π_{ak}:

$$\begin{aligned} w_{ak} &= \sum_{p' \ni a,\, t_{p'} < t_p} 1[k \in K_{p'}] \cdot exp\left(-\frac{t_p - t_{p'}}{\lambda_t}\right), \\ \pi_{ak} &= \begin{cases} clip(1 - e^{-w_{ak}},\ 10^{-4},\ 0.99), & w_{ak} > 0, \\ 0.01, & w_{ak} = 0 \text{ (no historical record).} \end{cases} \end{aligned} \tag{1}$$

where λ_t is the decay time constant (selected on the validation set; initialized to 3 years). The two quantities serve different consumers: the ranking score of Eq. (7) uses the raw strength w_{ak} directly, while the attribution layer injects π_{ak} as log-odds (Eq. (3)) and as the regularization target (Eq. (6)). Three remarks are in order: the fallback branch triggers exactly when author a has no recorded usage of entity k before t_p ($w_{ak} = 0$), and the small default probability (0.01) injects a deliberately *conservative negative bias* through its log-odds — absent history is weak counter-evidence, not evidence of participation, consistent with Constraint 2; π_{ak} is a role-blind method-usage probability shared across the role channels of (a, k), and must not be read as a role-conditioned participation probability; and w_{ak} uses only information preceding t_p, which is consistent by construction with the anti-leakage requirement of the temporal-split evaluation in Section 5.5.

4.3 Contribution-evidence-constrained latent association learning layer

4.3.1 Objective, inputs, and outputs of the layer

The attribution learning layer estimates, for every tuple (author $a \in A$, paper $p \in P$, method entity $k \in K_p$, role $r \in R$), the participation probability $P_\phi(Z_{apkr} = 1 \mid \mathcal{E}_{apkr})$ from the three evidence kinds of Section 4.2 (coarse statements C_{apr}; fine anchors T_{apkr} with confidence γ; history prior π_{ak}). Its output is a sparse author–method–role participation-probability tensor $\hat{Z}$ that gates knowledge propagation in Section 4.4.

4.3.2 Parameterization of participation probabilities

For each candidate tuple (a, p, k, r), we parameterize the participation probability with a multi-evidence scoring sigmoid:

$$s_{apkr} = \underbrace{u_a^{\top} W_r\, e_{pk}}_{\text{author–method–role compatibility}} + \underbrace{w_r^{\top}\left[\, e_{pk} \,\|\, h_{ak}^{\text{hist}} \,\right]}_{\text{role-conditioned historical evidence}} + \underbrace{b_r}_{\text{role bias}} \qquad (2)$$

$$P_{\phi}(Z_{apkr} = 1 \mid \mathcal{E}_{apkr}) = \sigma\left(s_{apkr} + log\,\frac{\pi_{ak}}{1-\pi_{ak}}\right) \qquad (3)$$

where:

- $e_{pk} \in \mathbb{R}^d$: the evidence representation of method entity k in paper p, obtained by encoding the text anchors extracted in Section 4.2 with a pre-trained language model and attention-pooling over multiple mention positions;
- $u_a \in \mathbb{R}^d$: the learnable embedding of author a;
- $W_r \in \mathbb{R}^{d\times d}$, $w_r \in \mathbb{R}^{2d}$, b_r: role-specific learnable parameters that capture "how the participation pattern of the same author–method pair differs across roles" (e.g., Methodology emphasizes design semantics, while Software emphasizes implementation semantics);
- h_{ak}^{hist}: the author-history evidence vector, aggregated from the contextual representations of method k in the author's historical papers;
- $log\,\frac{\pi_{ak}}{1-\pi_{ak}}$: the author-history prior, injected in log-odds form and equivalent to a prior factor multiplied with the evidence term: it yields to the data when evidence is sufficient and acts as a fallback when evidence is sparse.

Two design notes (the modelling constraints of Section 3.4). All probabilities pass through the sigmoid independently with no normalization across authors (Constraint 1: multi-author participation is the norm); and $C_{apr} = 0$ never forces a probability to 0 (Constraint 2: a missing declaration does not imply non-participation), enforced softly through the losses below.

4.3.3 Supervision signals and loss functions

Since Z_{apkr} is not directly observable, we design three supervision signals of differing noise levels, plus a prior regularizer.

(1) Coarse-grained statement-consistency loss $\mathcal{L}_{\text{coarse}}$. When $C_{apr} = 1$, author a participated in some work of paper p under role r, but the specific method is unknown. We treat all method entities K_p of paper p as a candidate bag and require that **at least one** method obtains a high participation probability, aggregated in noisy-or form:

$$\mathcal{L}_{\text{coarse}} = -\sum_{(a,p,r):\, C_{apr}=1} log\left(1 - \prod_{k\in K_p}(1-\hat{Z}_{apkr})\right) \qquad (4)$$

Noisy-or is preferred over max/mean pooling: it keeps non-zero gradients for all candidates and admits multi-author multi-method solutions without dilution. When $C_{apr} = 0$, **no loss is applied** (PU setting: unlabelled ≠ negative).

(2) Fine-grained anchor loss $\mathcal{L}_{\text{anchor}}$. Anchor supervision comes from two sources: a human-annotated anchor set $\mathcal{D}^{\text{hum}}$ (small, clean) with labels $y \in \{0,1\}$, and an LLM-extracted anchor set $\mathcal{D}^{\text{llm}}$ (larger, noisy) where each anchor carries a confidence γ. Standard binary cross-entropy is used for human anchors; LLM anchors are weighted by confidence, and only high-confidence anchors are treated as positives:

$$\mathcal{L}_{\text{anchor}} = \sum_{(i,y)\in\mathcal{D}^{\text{hum}}} BCE\,(\hat{Z}_i, y) \;-\; \sum_{\substack{(i,\gamma)\in\mathcal{D}^{\text{llm}} \\ \gamma\geq\tau}} \gamma\, log\hat{Z}_i \qquad (5)$$

where $BCE(\hat{z}, y) = -[ylog\hat{z} + (1-y)log(1-\hat{z})]$ is the standard non-negative cross-entropy loss, minimized directly as in the released implementation, and τ is the confidence threshold (0.8 in the experiments; see Section 5.6); in the reported runs $\mathcal{D}^{\text{llm}}$ is empty — the anchor loss consumes human-adjudicated anchors only.

(3) Author-history prior regularization $\mathcal{L}_{\text{hist}}$. The predicted distribution is required not to deviate substantially from the author-history prior where evidence is absent, preventing the attribution layer from assigning knowledge to authors who have never worked in the method area:

$$\mathcal{L}_{\text{hist}} = \sum_{(a,p,k,r)} (\,1-\pi_{ak}) \cdot KL\big(\,Bern(\pi_{ak}) \;\|\; Bern(\hat{Z}_{apkr})\,\big) \qquad (6)$$

The weight $(1-\pi_{ak})$ concentrates the regularizer on pairs with little or no history: under Eq. (1), the exact relation for documented pairs is $1-\pi_{ak} = clip(e^{-w_{ak}},\ 0.01,\ 0.9999)$ (and 0.99 for pairs without record) — the simpler equality $1-\pi_{ak} = e^{-w_{ak}}$ holds only away from the clipping bounds — so, up to those bounds, the regularization strength decays exponentially with the accumulated historical mass w_{ak} itself rather than with a relative frequency, and well-documented pairs are dominated by the data rather than by the prior.

(4) Total loss. The attribution layer is optimized with the three supervised terms above (their combination with the ranking loss is specified in Online Resource 1, Section S8; the reported implementation trains with attribution losses only, $\lambda_3 = 0$). λ_1, λ_2 are selected on the validation set; Section 6.3 removes each term in turn.

4.3.4 Training procedure

Training alternates between the attribution forward pass (Eqs. (2)(3)) and the three supervised losses (Eqs. (4)–(6)), optimized with Adam; full pseudocode is given in Section S5 of Online Resource 1.

Complexity note. Per-paper candidate tuples number $|A_p| \cdot |K_p| \cdot |R|$ (typically $\lesssim 10 \times 15 \times 14$), so training cost is comparable to an ordinary graph recommender.

4.4 Attribution-controlled knowledge propagation

Status note. This pathway belongs to the full design and is *not enabled* in the reported configuration: ranking consumes the attribution layer only through the corpus-aggregated profile P_6 (Eq. (7), Section 4.5.2). In brief, the specification replaces the 0/1 hard connections of the author–paper–method graph with soft connections gated by $\hat{Z}_{apkr}$, producing role-resolved author representations $h_{a,r}$ through normalized gated aggregation with a residual identity term, plus an optional two-hop co-author smoothing; the full specification is given in Online Resource 1, Section S7.

4.5 Task-driven collaborator ranking

4.5.1 Definition and representation of task queries

A task-driven query is defined as

$$q = (t_q,\ K_q,\ R_q)$$

- $t_q \in \mathbb{R}^d$: the task-description text (e.g., "develop graph neural network models for single-cell data") encoded as in Section 4.2;
- K_q: the required method entities, identified explicitly or automatically from the query (same entity space as K and entity-linking of Section 4.2);
- R_q: the role requirements implied by the query (e.g., Methodology, Software); when not stated explicitly, a lightweight classifier predicts a role distribution $\hat{p}_q(r)$ from t_q.

This covers both explicit queries ("find someone who can do method X"; K_q explicit, R_q inferred) and free-text task descriptions (both inferred).

4.5.2 Matching score function

The score of query q for candidate author a uses a product factorization of "historical strength × role matching × activity":

$$score(q,a) = \frac{1}{|K_q|} \sum_{k \in K_q} \underbrace{w_{ak}}_{\text{historical strength}} \times \underbrace{\sum_{r \in R} \omega_q(r) \cdot P_6(r \mid a,k)}_{\text{role matching}} \times \underbrace{c_a}_{\text{activity}} \qquad (9)$$

where:

- **Historical strength** w_{ak}: the author–method prior count weight from Eq. (1) (public records before t^*), role-blind — the only signal shared by all models, which guarantees the same-information fairness of Section 5.4: any gain beyond the baselines can only come from the latter two factors;
- **Role matching** $P_6(r \mid a,k)$: the role-normalized distribution of the attribution layer's per-role probabilities $\hat{Z}_{apkr}$, noisy-or aggregated over the corpus; $\omega_q(r)$ is the query's role weight (an indicator when roles are explicit). Uncovered (a,k) pairs

fall back to the author's declared-role histogram, then to a uniform distribution. This term is the only injection point of the attribution layer into recommendation — "using k for software development" is distinguished from "using k for data analysis";

- **Activity** c_a: the average proportion of work-type roles the author undertakes per paper (defaulting to the corpus-wide mean); P_6 answers "is this the right role?", while c_a answers "how specialized is the author in method work?";
- The three factors are isolated by an ablation ladder (Section 6.3): w_{ak} alone (B6-rb), $w_{ak} \times P_6$ (B6-role), and the full product (B6).

Design note. An earlier additive variant (cosine similarity of role-resolved representations + noisy-or coverage + topic) was abandoned: with within-paper query evidence absent, the neural terms saturate near zero and drown the only discriminative historical signal (Section S2 of Online Resource 1). The product factorization anchors comparability on the shared w_{ak} and confines the attribution layer's contribution to the single injection point P_6, making ablation attribution clean. To make the actual computation path explicit: in the reported system the ranking score uses exactly the three factors of Eq. (7); the role-resolved representations $\mathcal{H}$ of Section 4.4 do not enter the ranking score — they are retained for the abandoned additive variant (Section S2) and the joint-optimization pathway (Section 4.5.4), both outside the reported configuration.

4.5.3 Ranking loss (specified, not enabled)

The full design defines a BPR ranking loss over temporal-split positives (Section 5.5) with topic-related hard negatives and no temporal leakage; it is disabled in the reported implementation. Full specification: Online Resource 1, Section S8.

4.5.4 Joint-optimization pathway (specified, not enabled)

The full objective combines the attribution losses with the ranking loss under a coefficient λ_3 controlling gradient flow into ϕ. **In the reported implementation** $\lambda_3 = 0$: the attribution layer is trained with the supervised attribution losses alone, and ranking consumes its outputs through the analytic scoring of Eq. (7) with no gradient pathway — with within-paper query evidence absent, neural matching terms saturate near zero and ranking gradients add variance without measurable benefit (Section 4.5.2); stability measures for the future joint variant are given in Online Resource 1, Section S8.

4.5.5 Inference complexity

At inference time $\hat{Z}$ and $\mathcal{H}$ are precomputed offline; an online query needs one text encoding plus a constant number of dot products per candidate, i.e., cost linear in the candidate pool and independent of corpus size. With approximate nearest-neighbour indexing (HNSW [5]) we expect this to remain practical at 10^5-author scale, although we have not measured latency at that scale and report this as a complexity expectation, not a measurement.

5. Dataset and Experimental Setup

5.1 Data sources and construction

Corpus sources. We validate on a single discipline, computational biology (PLOS Computational Biology, which mandates CRediT statements and is dense in method entities). Full texts and contribution statements come from PLOS; cross-paper author identity resolution reuses OpenAlex records (evaluation-period matching rate 65.7%; history-library coverage of evaluation-period authors 91.5%); and author histories are built from the full OpenAlex publication records (3,241 authors, 285,663 papers) [7].

Scale.

Item	Count
Papers	750 (training period 500: 2015–2023; evaluation period 250: 2024–2025)
Author–paper records	4,820 (2,958 training + 1,862 evaluation)
Authors (OpenAlex-disambiguated)	3,337
Method entities (after linking)	9,436 (existing K=6,549 + 2,887 new entities in the evaluation period)
Method mentions (LLM-extracted)	14,243, of which 14,159 (99.4%) passed span-localizability verification
Author-history priors ($w_{ak} > 0$)	171,317: 137,049 for training-period entities + 34,268 for evaluation-period new entities (cross-journal history matching, covering 1,153/2,887 new entities and 3,090 authors)
Contribution-statement role records	19,963 (structured CRediT coverage 95.6%)
Human anchors (gold-standard training injection)	2,658 cells (the training portion of an 8:2 split of the human gold standard; automatic LLM extraction of author–action–object triples is planned for scale-up)

Preprocessing and anti-leakage. The temporal cut point is t^*=2024-01-01: the 500 papers before t^* are used to build attribution evidence and for training, and the 250 papers after it are reserved for evaluation (Section 5.5). The construction of π_{ak}, e_{pk}, and $\hat{Z}$ is strictly limited to information before the corresponding time of each sample; the evaluation-period corpus and the training base are physically isolated, and the author-history library uses only records published before t^*.

5.2 Fine-grained evaluation-set construction (manual annotation protocol)

Sampling. 94 papers were stratified from the training-period corpus (single/multi-method papers, single/multi-person undertakings of the same role, all high-frequency CRediT roles)

and their (author × method entity) candidate cells were enumerated, in a 50-paper initial batch and a 44-paper formal batch; **49 of these papers were doubly annotated and adjudicated into the final gold standard** — a 5-paper pilot batch (used for guideline calibration, v1 → v2) plus the 44-paper formal batch — while the remaining 45 enumerated papers of the initial batch were superseded by the formal batch after the guideline revision and are unused. For each annotated paper, the role dimension was supplied by the annotator together with the participation judgement: for every cell judged positive, annotators were required to enumerate **all** roles under which the author participated in that method (within the roles the author declared in that paper), so a positive cell whose role list excludes a role constitutes an explicit adjudicated negative on that role.

Annotation. Two annotators — both associate professors, one with a biomedical-engineering background and one with a computer-science-and-engineering background — independently judged $Z \in \{0,1,\text{uncertain}\}$ for every candidate cell, based on the full paper text (focusing on division-of-labour statements in Methods, acknowledgements, and supplementary materials). Both annotators were trained on the written guideline and calibrated on a pilot batch (5 papers, 312 cells) before the formal batch, and both followed the documented annotator requirements (released with the guideline). Annotation proceeded under blind conditions: the two annotators judged the same batch of papers independently, neither could view the other's results, and no specific judgement was discussed before reconciliation; **model outputs and LLM anchors were not shown during annotation**, to avoid anchoring bias. Disagreeing samples were resolved in a centralized reconciliation session (adjudication records released with the data), and the authors then re-reviewed and confirmed 100% of the final labels. Agreement: raw agreement 0.964 on 3,306 doubly annotated cells, Cohen's κ = **0.938** (κ = 0.914 on the 1,902 cells where both annotators gave a valid judgement), far above the preset threshold of ≥ 0.7. Disagreement composition: 48 opposing disagreements (1 vs 0), adjudicated strictly; the remaining 72 were strictness differences of "uncertain vs valid judgement". A further 424 editorial cells were reviewed with in-depth adjudication.

Products and usage. The annotated set contains 3,306 candidate cells: 2,326 valid judgements (1,541 positive / 785 negative; positives account for 66.3% of valid judgements) and 980 uncertain cells (used only for the behavioural analysis in Section 6.5, excluded from metrics). The set is split by paper 8:2 into a training-injection part (2,658 cells) and a held-out test part (648 cells, including 45 uncertain cells); the split is frozen with the data, and the per-paper split list is released alongside it. Effective supervision: 935 of the 2,658 training cells are uncertain (35.2%), leaving **1,723 cells that actually enter the anchor loss** (1,139 positive / 584 negative), while the test side contains only 6.9% uncertain. We disclose this asymmetry explicitly: uncertain judgements concentrate in a subset of formal-batch papers with genuinely ambiguous evidence, and the frozen by-paper split happens to place less ambiguous papers in the test set — the held-out evaluation is therefore conducted on *cleaner* cells than the training portion, which the hard-evidence subset (Section 6.1) and the uncertain-cell behavioural analysis (Section 6.5) partially

counterbalance. The set also serves as the manual-verification subsample for method-entity recognition quality in Section 4.2.1.

Reproducibility. The annotation guideline, the annotator requirements and work instructions, annotation-interface screenshots, disagreement-handling rules and adjudication records, and the final annotation data (de-identified) are released alongside the paper.

5.3 Baselines

Seven comparison references in total — five same-information baselines (B1–B5), one author-level role-histogram ablation (B7), and a random-ordering reference (numbering consistent with Section 6):

#	Method	Description
B1	Co-authorship sharing	paper knowledge is propagated equally to all authors
B2	Contribution scalar weighting	author–paper-level contribution weights (GraphRec-style author contribution index)
B3	HeteAlloc+LLM	HeteAlloc path-similarity allocation, with method entities replaced by the LLM entities of Section 4.2.1
B4	Role-feature GNN	CRediT roles as author node features in a homogeneous GNN, no explicit attribution layer
B5	Generic MIL attribution	a multiple-instance weakly supervised attribution model with the same inputs as ours, without contribution-constraint designs
B6	This paper	the full contribution-constrained attribution layer
B7	Author-level role histogram	the role-matching factor $P_6(r \mid a, k)$ replaced by the author's corpus-level declared-role histogram (the first-level fallback of P_6),

#	Method	Description
		keeping the same w_{ak} and c_a; isolates the entity-level attribution layer's contribution to ranking
Random	Random ordering	uniform random ranking of the candidate set (20 repetitions); a sanity reference for the absolute level of the metrics

Fairness-adaptation statement (key): B1–B5 all use the same method entities, text-encoding features, and contribution-statement information as B6 (to the extent allowed by their model structures). B3 tests "unsupervised rule-based allocation vs supervised learned allocation"; B5 tests "generic weak supervision vs contribution-evidence constraints". B7 and Random are diagnostic references rather than competitors: B7 tests whether entity-level attribution adds anything beyond author-level declared-role statistics (Section 6.2), and Random marks the score level obtainable without any signal under our candidate construction.

5.4 Fairness and leakage-control checklist

1. All methods use the same temporal split, the same candidate sets, and the same evaluation scripts;
2. All methods have access to the same textual and contribution information (the difference lies only in how it is used);
3. Attribution-type methods (B3, B5, B6) have the same parameter-search budget (the same number of validation-set evaluations);
4. No feature in training or evaluation may use information after t^* (the entity vocabulary of the evaluation period provides only alias sets of "what it matches in historical text"; history matching is strictly limited to records published before t^*; gold-standard labels do not participate in prior construction);
5. B6 results are means ± standard deviations over 6 runs with different random seeds (one primary seed plus five replications); B4/B5 are trained once (seed 42); B1/B2/B3/B7 are deterministic rules, run once; Random is averaged over 20 repetitions.

5.5 Evaluation tasks and metrics

Task A: knowledge-attribution quality (answering RQ1). Evaluated on the held-out test set of Section 5.2: per-tuple Precision / Recall / F1 (reported at threshold 0.5 and at the optimal threshold), AUPRC (the primary metric under class imbalance), and calibration ECE (reliability diagrams released in the code repository). **Evaluation granularity (important):** Task A scores one participation probability per (author, method) cell — the attribution layer's per-role probabilities are noisy-or aggregated (Section 4.3) before being

compared with the adjudicated participation label z, and the role lists annotated for positive cells do not enter this metric. Task A therefore measures **role-aggregated participation attribution**, not joint method–role assignment; a dedicated per-role evaluation is reported in Section 6.1.

Task B: task-driven role-matching recommendation (answering RQ2). We use a multi-positive temporal recommendation construction: from the 250 papers published after t^*, a query is constructed for each verified method mention (method entity k; the contextual role is the majority ctx of that entity's mentions in the text) and each work-type role r that has at least one declared holder; the gold standard gold(k, r) comprises **all** authors who, within the evaluation period, declared role r in papers mentioning k (aggregated across papers; multiple positives). **Proxy-label caveat.** These gold labels are paper-level declared-role proxies: an author's declaration of role r in a paper that mentions k does not by itself verify that the author performed that role *for that specific method*. Task B is therefore a proxy-based retrieval evaluation and a boundary analysis of the attribution layer, not a validation of actual method-level competence; the independently human-verified query subset (Section 6.2.1) provides the more trustworthy, size-limited complement. The main set contains 7,099 queries with $k \in K$ (the existing entity space); the 13,753 queries on evaluation-period new entities form the **method-cold-start subset** — after the cross-journal history-prior extension of Section 5.1, new entities also have history-only candidates available, and candidate construction is identical to the main set (mean 4.9 hard negatives per query; for 59.1% of cold-start queries the method has no author prior before t^*, so candidates contain only easy negatives); gold-set sizes: median 2, mean 3.7, max 44. Candidate set = all gold authors + 100 negatives: hard negatives are authors with a history prior on k before t^* ($w > 0$) but not in gold(k, r) ("history-only candidates": authors with a historical record on k who did not become observed proxy positives within the evaluation window, mean 29.1 per query), and easy negatives are sampled from the remaining authors in the prior library. Metrics: NDCG@{5,10,20}, Recall@{10,20}, MAP, MRR. Ties — including all-zero-score candidate sets — are broken by a fixed-seed pseudo-random permutation (seed 42) and are thus arbitrary but reproducible. Because the hard negatives are deliberately chosen as authors with strong history priors on k (history-only candidates, mean 29.1 per query), history-based rankers are pushed below the random-ordering level on the main set (Section 6.2); we retain this adversarial construction because it operationalizes the mismatch between historical records and observed proxy labels that task-driven recommendation must handle, and we report the random reference alongside all models.

Construction lesson (full account in Section S4 of Online Resource 1). The initial single-positive construction made 94.8% of positives prior-free while 91.6% of queries had historical experts as hard negatives — no history-based model could possibly win. The multi-positive temporal construction fixes this: an author who declares the relevant role in an evaluation-period paper mentioning the method becomes gold, and only history-only candidates form hard negatives.

Significance testing. For the primary Task B metrics, B6 is compared with each baseline by a paired permutation test (10,000 sign-flip resamples), and we report p-values and bootstrap 95% confidence intervals with Holm correction for multiple comparisons. Because multiple queries are derived from the same evaluation paper or the same entity, per-query tests overstate the effective sample size; all key comparisons are therefore re-run under a cluster bootstrap (10,000 resamples over evaluation-paper clusters and, as a stricter variant, over entity clusters), reported in Section 6.2. Robustness: the B6 attribution layer is retrained with 6 seeds and all significance tests are repeated per seed.

5.6 Implementation details

- Text encoder: TF-IDF (512 dimensions, 1–2 grams, sublinear-tf, fitted on the entity contexts and abstracts of all 49 annotated papers, including the 10 held-out test papers; the fit is unsupervised and uses no labels, so no gold information enters through the encoder, but we disclose the transductive text statistics explicitly). **Pre-trained semantic encoder comparison:** wholesale replacement with SPECTER2 base (Singh et al. [10]; frozen backbone, [CLS] 768 dimensions), followed by retraining and re-evaluation, leaves Task B on the main set on par (B6 NDCG@10 0.0398 vs 0.0400 for TF-IDF) but degrades Task A (full-layer AUPRC 0.995 vs 1.000; hard-evidence subset 0.883 vs 1.000): representations trained for paper-level semantic similarity smooth the lexical distinctions at the method-name level on which attribution evidence relies, so TF-IDF is retained (scripts and artefacts released with the code); as a control for the transductive fit, we additionally re-fitted TF-IDF on the 39 training-injection papers only and retrained both tasks (3 seeds each): Task A full-layer and hard-evidence AUPRC remain 1.000 (ECE 0.020) and Task B main-set NDCG@10 is unchanged at 0.040 (0.0399 ± 0.000 vs 0.0399 ± 0.001 under the transductive fit), so excluding test-paper text statistics leaves all conclusions intact;
- Dimension d=64 (internal projection dimension); optimizer Adam, learning rate 1e-3;
- Loss weights: λ_1=1.0 (anchor), λ_2=0.1 (prior regularization); the joint ranking loss is disabled (λ_3=0; see Section 4.5.4);
- Anchor confidence threshold τ=0.8; time decay λ_t=3 years;
- Training: 300 epochs, early stopping by gold-standard BCE on an internal validation split (8 of the 39 training-injection papers, held out with a fixed RNG seed; the 10 held-out test papers are never used for training or early stopping); hardware CPU (2–3 minutes per training run or full-corpus forward pass);
- **Reported configuration (code-aligned).** Training supervision: $\mathcal{L}_{\text{coarse}}$ over declared work-type roles + $\mathcal{L}_{\text{anchor}}$ (BCE on the 1,723 valid human training cells) + $\mathcal{L}_{\text{hist}}$ (Eq. (6)); the LLM auto-extraction of anchor triples is **not** used in the reported runs (specified for scale-up only), and the ranking loss is disabled ($\lambda_3 = 0$): Eq. (7) is an analytic scoring form without a gradient pathway, so no ranking gradient reaches ϕ. Inference inputs: author declared-role features, entity TF-IDF context features, and history priors. Enabled modules: the single-layer per-role attribution scorer, noisy-or aggregation over roles, the P_6 profile with declared-role-histogram fallback, and

the analytic scoring of Eq. (7). Specified but disabled: two-hop propagation (the α row of Table 6; Section S1), the joint-optimization pathway (Section S3), and the additive scoring variant (Section S2).

- Code and data: publicly available at https://github.com/seucchen-eng/contribution-scoped-attribution

---

6. Results and Analysis

6.1 Knowledge-attribution quality (answering RQ1)

Table 2 reports attribution-task performance on the held-out test set of Section 5.2, comparing our method with the allocation baselines evaluated on this task (B3 and B5).

Table 2. Attribution task results (held-out test set of Section 5.2; n=603 cells: 402 positive / 201 negative; role-aggregated participation evaluation, per-role probabilities noisy-or aggregated before scoring; B6 unrounded full-layer AUPRC 0.9996, paper-cluster bootstrap 95% CI [0.998, 1.000] over the 10 test papers)

Method	AUPRC	P@0.5	R@0.5	F1@0.5	ECE ↓
B3 HeteAlloc+LLM	0.709	0.900	0.022	0.044	0.543
B5 Generic MIL	0.980	0.982	0.973	0.977	0.038
B6 This paper	**1.000**	**0.985**	**0.995**	**0.990**	**0.015**
(reference) B2 contribution scalar	0.930	0.829	0.711	0.766	0.127

(B1/B2/B4 have no explicit attribution output; B2's index is an auxiliary reference row. Stratification: hard-paper subset n=299: B6/B5/B3 = 0.999/0.980/0.728; hard-evidence subset n=218 (only cells with direct textual evidence or in-depth adjudication, the most discriminative layer): **1.000**/0.890/**0.184**.)

Answering RQ1 at the role-aggregated participation level: our method improves AUPRC by +0.020 over the strongest baseline B5 on the full test set (1.000 vs 0.980), and by 11.0 absolute points on the most discriminative hard-evidence subset (1.000 vs 0.890). The gain has two sources: B3's fixed-rule allocation fails where declarations are missing or conflicting (hard-evidence AUPRC 0.184), demonstrating the value of learnable allocation, while B5's unconstrained weak supervision lags on the same subset, demonstrating the value of the contribution-constraint design. B6's ECE of 0.015 (vs 0.038 for B5) indicates

that $\hat{Z}$ is also well calibrated in probability scale, justifying the probability-weighted propagation of Section 4.4. The hard-evidence subset — where B3 and B5 degrade while B6 remains at 1.000 — concentrates cells in which declarations and full-text evidence disagree, and is the focus of the error analysis in Section 6.5.

Per-role resolution (a finer-grained check). The headline figures above are role-aggregated; we additionally score the 3,618 (author × method × role) tuples implied by the test cells (positives: the annotated roles of positive cells, 404 tuples; negatives: all remaining role channels, with z=2 cells excluded as elsewhere). Micro-AUPRC drops to 0.234 (per role: methodology 0.874 over 281 positives; formal-analysis 0.230 over 80; software 0.183 over 43; the other three work-type roles have no positive test cell), and on positive cells the argmax role matches the annotation in only 5.0% of cases (top-2: 63.9%). The diagnosis is structural: no cell-level role supervision exists — anchor supervision applies to the noisy-or-aggregated participation probability — so a positive cell raises all role channels together (e.g., investigation 0.842 on positive cells versus ≤ 0.02 on negative cells). The layer is thus an accurate **participation** detector whose per-role outputs remain prior- and bias-dominated; downstream ranking consumes them only through the smoothed, corpus-aggregated P_6 profile with a declared-role-histogram fallback (Section 4.5.2) — precisely the component that the B7 comparison (Section 6.2) identifies as the operative signal. Adding cell-level role supervision is the clearest direction for strengthening the layer (Section 7.4).

6.2 Task-driven recommendation performance (answering RQ2, main results)

Table 3. Task B role-matching recommendation (within-K main set, n=7,099 queries; B6 is mean ± std over 6 seeds; B1–B3/B7 are deterministic rules run once; B4/B5 are trained once (seed 42; a 6-seed retraining evaluated under B6's scoring function is reported in Table 4 and Section 6.3); Random is mean over 20 repetitions, std ≤ 0.002 on all metrics)

Method	NDCG@10	NDCG@20	Recall@10	Recall@20	MRR
Random ordering (reference)	0.079	0.113	0.094	0.189	0.165
B1 Co-authorship sharing	0.028	0.045	0.037	0.086	0.073
B2 Contribution scalar weighting	0.037	0.058	0.046	0.105	0.088
B3 HeteAlloc+LLM	0.032	0.050	0.040	0.091	0.081

Method	NDCG@10	NDCG@20	Recall@10	Recall@20	MRR
B4 Role-feature GNN	0.001	0.004	0.002	0.006	0.019
B5 Generic MIL attribution	0.003	0.005	0.003	0.007	0.023
B7 Author-level role histogram	0.041	0.068	0.054	0.129	0.095
B6 This paper	**0.040 ± 0.000**	**0.067 ± 0.000**	**0.052 ± 0.000**	**0.126 ± 0.000**	**0.094 ± 0.000**

Addendum — significance, clustering, and the role of the attribution layer: (i) B6 vs each of B1–B5: paired permutation tests (10,000 sign flips, Holm-corrected) give corrected $p \leq 0.003$ on the four primary metrics. (ii) Cluster bootstrap (10,000 resamples over 206 evaluation-paper clusters): B6 vs B2 NDCG@10 $+0.0031$, 95% CI $[+0.0021, +0.0041]$, Holm $p < 0.001$; under stricter entity clustering (792 clusters) NDCG@10 remains significant (Holm $p = 0.003$) while MRR becomes marginal ($p = 0.06$). (iii) B6 vs B7: the entity-level attribution layer adds **no** ranking gain over the author-level declared-role histogram — B6 is marginally below B7 on the main set (0.040 vs 0.041; Holm $p = 0.003$) and identical on the cold-start subset (by fallback construction). We therefore claim no ranking-accuracy contribution for it; its value lies in attribution fidelity and auditability (RQ1) and the per-recommendation evidence chain (Section 6.4). (iv) The ablation ladder increases level by level (B6-rb 0.032 → B6-role 0.034 → B6 0.040), significant per seed. (v) All history-based rankers score far below the Random row — a direct consequence of the adversarial hard-negative construction (Section 5.5). Random is an absolute-level reference, not a competitor; the informative reading is the relative ordering under identical inputs and the cold-start subset, where B6/B7 exceed Random (Section 6.6).

Table 4. Same-scorer control: B4/B5 retrained under the same 6-seed budget as B6 and evaluated through B6's analytic scoring function (Eq. (7)) with the identical profile-injection and fallback mechanism (mean over 6 seeds; per-seed std $< 10^{-4}$ throughout; cold start: NDCG@10 on the 13,753-query subset, identical across the three variants by fallback construction; B7 row repeated from Table 3 for comparison)

Model (identical scoring function)	NDCG@10	Recall@10	MAP	MRR	Cold-start NDCG@10
B4-same (GNN role-blind profile injected)	0.0405	0.0531	0.0766	0.0943	0.065

Model (identical scoring function)	NDCG@10	Recall@10	MAP	MRR	Cold-start NDCG@10
B5-same (MIL role-blind profile injected)	0.0406	0.0534	0.0768	0.0944	0.065
B6 (full attribution profile P_6)	0.0399	0.0518	0.0765	0.0938	0.065
B7 Author-level role histogram (reference)	0.0407	0.0535	0.0771	0.0947	0.065

Paired per-query permutation tests against B6 give $p = 0.0007$ (B4-same) and $p = 0.0001$ (B5-same) — statistically significant but negligible in magnitude (Section 6.3).

Answering RQ2: with exactly the same input information, B6 improves over the strongest prior-only baseline B2 by 8% on NDCG@10 (0.040 vs 0.037, Holm-corrected p=0.002; robust under cluster bootstrap), and by +12% / +7% / +6% on Recall@10 / MAP / MRR. Read by groups, the B1–B2 gap (0.028 vs 0.037) shows that the activity scalar already captures most of what binary co-authorship offers, while the B2–B6 gap is the net benefit of adding role information to the shared prior — a benefit that the B7 comparison (Table 3) localizes to author-level declared-role statistics rather than entity-level attribution. B4/B5 score near zero (NDCG@10 ≤ 0.003) because attribution models draw discriminative power from within-paper evidence, which is absent at query time: their forward outputs saturate and drown the prior signal — under the identical analytic scoring function their learned profiles essentially reproduce the author-level histogram (Table 4), so the moderate role benefit enters through profile-level re-weighting (B6-role 0.034 vs B6-rb 0.032, p=0.002), not through entity-level attribution; B3's candidate-normalized rule allocation (0.032) is likewise inferior to product-factorized role matching (0.040). By role, B6 leads on 5 of 6 work-type roles, lagging only on methodology (0.045 vs 0.052), the role with the most generic declarations (Section 6.5). Absolute main-set metrics are limited by historical coverage (Section 7.4); on the full set of 20,852 queries B6 ranks first among the history-based models (0.057 vs B2 0.054) but remains below the random-ordering reference (0.063; Section 6.6).

6.2.1 Independently verified fine-grained queries

To test whether the ranking conclusions survive independently verified ground truth, we built a verified query subset from the 10 held-out test papers of Section 5.2. For each (method entity, role) pair covered by human-annotated cells, the query's gold authors are

those adjudicated as participating in that role for that entity ($z = 1$), and the negatives are authors verified as non-participants ($z = 0$) together with authors verified as participating in a *different* role for the same entity — every candidate thus carries an independent human label, and uncertain cells are excluded. In particular, because the annotation protocol requires annotators to enumerate every role under which an author participated in a method (Section 5.2), a verified positive on a *different* role is an explicit adjudicated negative on the query role; role enumeration is bounded by the roles declared in the paper, which we note as a limitation. This yields 115 verified queries (mean 2.9 gold and 3.1 verified negatives per query), plus a harder 40-query subset requiring at least 5 verified negatives. Candidate pools are small, so Recall@10 is 1.000 for all models and is omitted; NDCG@10, MAP and MRR remain discriminative.

Table 5. Verified-query results (NDCG@10 / MAP / MRR; Random: mean over 200 repetitions, NDCG@10 std 0.016 on the full subset and 0.027 on the harder subset; paired bootstrap confidence intervals are reported in the text)

Model	All verified queries (n=115)	Harder subset (n=40)
Random ordering	0.737 / 0.621 / 0.659	0.578 / 0.414 / 0.452
B1 Co-authorship sharing	0.746 / 0.634 / 0.672	0.579 / 0.417 / 0.461
B2 Contribution scalar weighting	0.750 / 0.641 / 0.679	0.568 / 0.405 / 0.442
B6-rb (role-blind ablation)	0.744 / 0.631 / 0.669	0.578 / 0.416 / 0.462
B6-role	0.751 / 0.640 / 0.684	0.582 / 0.420 / 0.469
B6 This paper	0.751 / 0.642 / 0.681	0.572 / 0.410 / 0.448
B7 Author-level role histogram	**0.762 / 0.658 / 0.693**	**0.600 / 0.447 / 0.485**

Paired analyses sharpen the picture (per-query NDCG@10 differences with 95% bootstrap CIs, computed over queries and, to account for within-paper correlation, over the 10 source papers — every query derives from a single paper): (i) role-aware scoring beats role-blind scoring on verified labels — B7 vs B1: +0.016 on the full subset (CI [+0.005, +0.029] over queries; [+0.001, +0.038] over papers) and +0.021 on the harder subset (both CIs marginal, touching zero) — consistent with a real but modest role signal, which the B7 construction localizes to author-level declared-role statistics; (ii) the entity-level attribution layer trails the author-level histogram, significantly on the harder subset (B6 vs B7: −0.028, CI [−0.059, −0.003] over queries and [−0.069, −0.005] over papers) and by −0.011 on the full subset — corroborating Table 3 on fully independent labels; (iii) no model significantly exceeds the random reference (B6 vs Random: +0.015 on the full subset and −0.008 on the harder subset, all CIs spanning zero), so we report this subset as directional evidence only; expanding the verified pool is future work.

6.3 Ablation study

Table 6. Ablation (Task A: full-layer AUPRC on the held-out test set; Task B: NDCG@10 on the within-K main set)

Variant	AUPRC	NDCG@10	What it tests
Full model	1.000	0.040	—
− Statement supervision (remove $\mathcal{L}_{\text{coarse}}$)	1.000	0.042	coarse supervision
− Text anchors (remove $\mathcal{L}_{\text{anchor}}$)	**0.827** (hard-evidence layer 0.172)	0.040	fine anchors
− Role dimension (merge r, single author vector)	1.000	**0.037** (identical to B2 by construction)	role-resolved representation
− History prior (remove π and $\mathcal{L}_{\text{hist}}$)	0.997	0.040	the prior; on the Task B side only the prior usage inside the attribution layer is ablated — the base signal w_{ak} shared by the recommendation protocol is unchanged
Z-normalized variant (violating Constraint 1)	**0.986** (hard-evidence layer 0.857, R@0.5 collapses to 0.44)	0.041	the many-to-many design: credit conservation forces single-person exclusivity, and joint participation is systematically missed
Missing declarations as negatives (violating Constraint 2)	1.000	0.040	the PU design: with 95.6% CRediT coverage in this corpus, missing ≈ true negatives, so no drop; calibration slightly degrades (ECE 0.015→0.026)

Variant	AUPRC	NDCG@10	What it tests
$\alpha = 0$ (pure one-hop) / $\alpha > 0$	part of the full design; two-hop propagation not enabled in the current experiments	same	two-hop smoothing

Reading the ablation: **text anchors are the lifeline of attribution accuracy** — without them, hard-evidence AUPRC collapses from 1.000 to 0.172 and the model degenerates into "saying yes to everything" (R@0.5=1.000, P@0.5=0.078); removing coarse statement supervision does not drop accuracy on adjudicated cells (Task B even gains +0.002, within single-seed noise) — its value lies in coverage supervision for the 95%+ of papers without anchors — whether this yields measurable gains at larger scale is a hypothesis to be tested, not an established result. Merging the role dimension makes Task B by construction identical to B2 (0.037), i.e., role resolution contributes the entire excess gain of B6 over B2 — which the B7 comparison (Table 3) localizes to author-level declared-role statistics rather than entity-level attribution. The history prior plays a limited role inside the attribution layer (0.997 vs 1.000); its main role is the shared base signal w_{ak}. The last two variants delimit where the modelling constraints matter: the normalized variant (violating Constraint 1) degrades attribution sharply (hard-evidence AUPRC 1.000→0.857, R@0.5 collapses to 0.44, ECE 0.328), showing that single-person exclusivity systematically misses joint participation — Constraint 1 is necessary here; treating missing declarations as negatives (violating Constraint 2) does **not** drop accuracy on this corpus — with 95.6% structured CRediT coverage, missing declarations approximate true negatives, so Constraint 2 is not empirically necessary here (its calibration cost is nonetheless already visible: ECE 0.015→0.026) and is retained for disciplines with high declaration-missing rates (Section 7.4).

Same-scorer control (Table 4). Retraining B4/B5 under the same 6-seed budget as B6 and re-evaluating them through B6's analytic scoring of Eq. (7) with the identical injection and fallback yields learned role-blind profiles that essentially reproduce the B7 author-level histogram (paired differences vs B6 statistically significant but negligible, $p \leq 0.0007$): under an identical scoring function the entity-level attribution layer still adds no ranking gain — corroborating that its evidenced contribution is attribution fidelity and auditability (RQ1), not ranking accuracy.

6.4 Case studies

Case 1: attribution inference. Fig. 2 compares the hard-case paper *Dynamical modelling of viral infection and cooperative immune protection in COVID-19 patients* (10.1371/journal.pcbi.1011383, held-out test set; 13 authors × 8 method entities, 4 authors simultaneously declaring Methodology, 83 gold cells): (a) the human ground truth shows a clear block structure — five authors undertake all 8 methods; (b) B3's rule allocation gives nearly uniform low scores, losing the block structure, with a false positive of 0.987 on the Mei × ELISpot cell (true z=0; B6: 0.750); (c) B6's aggregated $\hat{Z}$ fully recovers

the block structure (per-paper AUPRC 0.993 vs 0.592) and corrects B3's false negatives (e.g., Li(3) × ODE: 0.077 → 1.000); (d)(e) the role channels show B6's role resolution (Zhao × latin hypercube sampling: 0.851 vs 0.045 across the formal-analysis and validation channels).

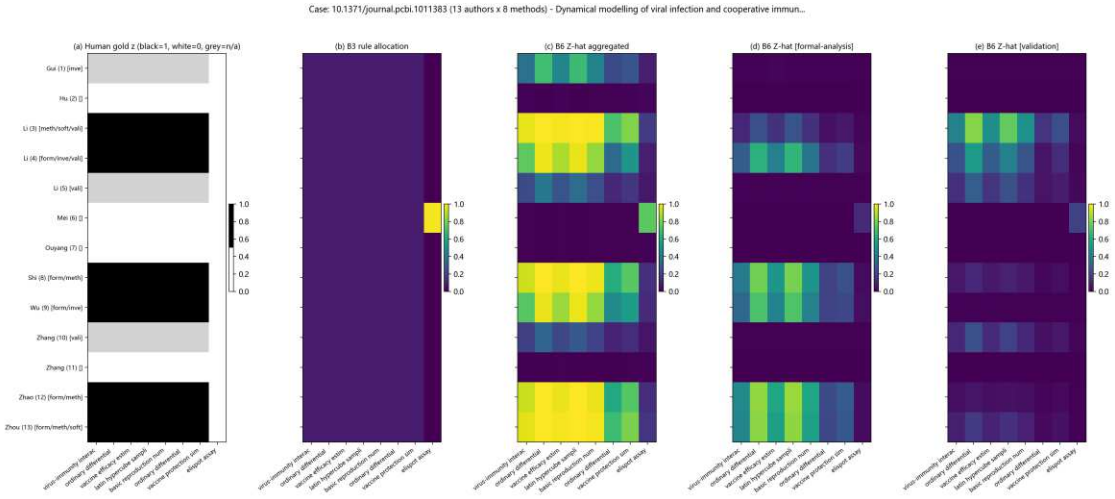


Fig. 2. *Case study on the attribution of paper 10.1371/journal.pcbi.1011383: (a) human ground truth z (black = positive, white = negative, grey = unannotated); (b) B3 rule allocation; (c) B6 $\hat{Z}$ aggregation; (d)(e) B6's formal-analysis and validation role channels.*

Case 2: task-driven query (full Top-6 comparison in Section S6 of Online Resource 1). For query Q07185 (the **data-curation** role for **deep neural network**, 12 gold authors), B2 ranks Alinejad-Rokny, the most productive historical author on this method (w=4.61), first — but his role profile shows zero inclination toward data-curation; B6 down-weights him via P_6 and lifts the gold author Yuan from rank 18 to rank 3. An honest boundary: here P_6 comes mainly from the author-level role-histogram fallback, so what is shown is the fallback-level benefit of role matching where entity-level profiles are sparse.

From prediction to evidence (what "auditable" concretely means here). Auditability rests on the additive decomposition of every cell score: each $\hat{Z}_{apkr}$ is a sigmoid of a compatibility term, an entity-side term, a role bias and a prior log-odds (Section 4.3.2), so every prediction reads as a sum of named, individually inspectable contributions, with the prior term traceable to concrete historical papers. Checking this chain on the 406 test cells scored ≥ 0.5: **every positive prediction carries at least one human-inspectable evidence item** — 98.5% a work-type role declared by the author in the paper itself, 14.5% a non-zero history prior — and Section 6.5 traces all 8 errors individually through these terms. The boundary of this guarantee: the chain explains *which evidence drove the score*, not whether that evidence is sufficient — sufficiency is what the human adjudication of Section 5.2 provides.

6.5 Error analysis

Protocol adaptation. B6 produced only **6 false positives and 2 false negatives** on the 603-cell held-out test set (threshold 0.5), fewer than the planned sampling scale; we therefore examine all 8 errors individually, together with a behavioural analysis of the 45 uncertain cells (z=2, excluded from the metrics).

False positives (6 cases, predictions 0.586–0.855): all fall on low-frequency entities — 5/6 involve entities appearing at most twice in the corpus (zigzag-hmc, phylogenetic multivariate probit model ×2, ELISpot assay). Causes: feature sparsity of long-tail entities and entity-linking errors that cross-paper evidence cannot correct. The typical case is Mei × ELISpot (Section 6.4): B3 scores 0.987, B6 converges to 0.750 but still crosses the threshold — without anchor evidence, frequency-1 entities lack counter-evidence.

False negatives (2 cases, predictions 0.350 / 0.476): both borderline cells on ordinary differential equation integration in the case paper, both with weak declaration–text correspondence; the predictions sit just below the threshold — the model did not "miss" these cells but "dared not confirm" them. This is the inherent boundary at which overly brief statements leave attribution evidence insufficient; the conservative design pushes such cells towards uncertainty rather than false positives — an expected cost of Constraint 2 rather than a defect.

Behaviour on uncertain cells (45 cases): mean score 0.602, an unexpected direction. The conservative design should press ambiguous evidence toward low scores, but the model **leans toward judging participation** — prior plus statement supervision has a positive bias in the ambiguous zone, and the calibration guarantee (ECE 0.015) holds only on adjudicated cells; uncertainty-aware calibration is needed (Section 7.2).

6.6 Method-cold-start subset analysis (the value boundary of cross-journal history priors)

Construction. 2,887 evaluation-period method entities (78% of the evaluation-period entity space) never appeared in the training corpus, corresponding to 13,753 queries whose candidate priors were initially all zero. Extending alias matching to authors' cross-journal histories (285 thousand papers, same Eq. (1) specification and t^* cut) gave 1,153 new entities (39.9%) historical coverage via 34,268 priors; candidate construction is identical to the main set (Section 5.5). The low coverage is not a matcher defect: relaxed matching (singularization, order-free) recovers only ~5% of sampled missing gold pairs — gold authors genuinely never mentioned the query method in their history.

Table 7. Results on the method-cold-start subset and the full set (B6: mean over 6 runs, std<0.0001; Random: mean over 20 repetitions; significance: Holm-corrected paired permutation tests and cluster bootstrap)

Set	Method	NDCG@10	Recall@10	MAP	MRR
Cold-start	Random	0.055	0.098	0.062	0.086

Set	Method	NDCG@10	Recall@10	MAP	MRR
subset n=13,753	ordering (reference)				
	B1 Co-authorship sharing	0.059	0.097	0.067	0.091
	B2 Contribution scalar weighting	0.062	0.103	0.069	0.093
	B3 HeteAlloc+LLM	0.061	0.101	0.068	0.092
	B7 Author-level role histogram	0.065	0.107	0.071	0.097
	B6 This paper	**0.065**	**0.107**	**0.071**	**0.097**
Full set n=20,852	Random ordering (reference)	0.063	0.096	0.074	0.113
	B1 Co-authorship sharing	0.048	0.077	0.067	0.085
	B2 Contribution scalar weighting	0.054	0.083	0.070	0.092
	B3 HeteAlloc+LLM	0.051	0.080	0.068	0.088
	B7 Author-level role histogram	0.057	0.088	0.074	0.096
	B6 This paper	**0.057**	**0.088**	**0.074**	**0.096**

On both sets, B6 is significant against B1/B2/B3 and the ablation variants on all four primary metrics (Holm-corrected $p \leq 0.0032$), surviving cluster bootstrap over papers and entities. Two boundary disclosures: B6 and B7 are identical on the cold-start subset (P_6 falls back to the histogram for every candidate by construction) and differ only in the third

decimal on the full set; and history-based ranking exceeds chance exactly where the construction allows it — cold-start (0.065 vs Random 0.055) — while the full-set mean remains below random (0.057 vs 0.063) because the adversarial main-set construction dominates the average. The cold-start subset's higher absolute metrics reflect its smaller gold sets (mean 2.3 vs 6.4) and 59.1% of queries without hard negatives; only the relative pattern within each subset is a valid reading.

Negative result: history-profile fallback is a counter-signal on cold-start queries. We tested the intuitive remedy — on all-zero-prior queries, break ties with $\varepsilon \cdot P_6 \cdot c_a$ instead of zero scores. On the 547 such queries of the main set this achieves NDCG@10 of only 0.002, **significantly below the 0.074 of a pure random ordering** (profile-only and scalar-only variants: 0.006 / 0.002). Diagnosis: gold authors of such queries are typical "introducers of new methods" whose historical volume is far smaller than the negatives' (median 13 vs 43 papers), so any history-built signal is biased toward the wrong people. The final model keeps the conservative definition $w = 0 \Rightarrow s = 0$ (Section 4.5.2): **history-evidence-based ranking reaches its boundary in the pure cold-start region; cold-start recommendation needs content/semantic signals** — but naive wholesale encoder replacement with a domain pre-trained representation hurts attribution accuracy (Section 5.6), so semantic augmentation needs a design that explicitly mixes lexical evidence (future work).

---

7. Discussion

7.1 The relationship between attribution improvement and recommendation improvement

We framed RQ2 as a testable question — *does entity-level attribution improve task-driven recommendation under identical inputs?* — and the honest answer returned by our experiments is **largely negative for ranking accuracy**. Although the ablation ladder first suggested a role-resolution gain (B6-role vs B6-rb, 0.034 vs 0.032, p=0.002), the B7 comparison (Table 3) localizes the excess: it is fully reproduced by the author-level declared-role histogram, and on independently verified queries the entity-level layer trails the histogram, significantly on the harder subset (Section 6.2.1). What the entity-level attribution layer demonstrably adds is attribution fidelity, calibration, and per-recommendation evidence chains (Sections 6.1, 6.4), not ranking accuracy. Per-query correlation adds a counter-detail: role-resolution sharpness is **weakly negatively correlated** with per-query NDCG@10 gain (n=6,120, Pearson r=−0.079) — P_6 falls back to the histogram for most candidates and gold authors are mostly newcomers with blunt role profiles, so aggressive role filtering hurts them. Where role information does help, the channel is therefore **"gentle role re-weighting", not "sharp role filtering"**. Observational evidence cannot fully rule out confounders; the fairness setting of Section 5.4 controls input-information confounding, and equal-budget tuning mitigates model-capacity confounding. We accordingly position the recommendation task as a downstream

application and boundary analysis of the attribution layer rather than as the paper's primary claim, and B7 as the key control that any future entity-level variant must beat.

7.2 Sensitivity to LLM extraction noise

Current extraction quality. The span-localizability rate of LLM method-mention extraction is above 99.4%; the combined JSON-parsing and LLM-call failure rate is below 1%.

Anchor-noise sensitivity experiment. On the training-injection anchors (the gold-standard training portion), labels are symmetrically flipped at $\rho \in \{10\%, 20\%\}$ and B6 is retrained (all other settings identical to the main model):

Anchor noise	AUPRC (full layer)	AUPRC (hard-evidence layer)	ECE (full layer)
0% (main model)	1.000	1.000	0.015
10%	0.982	0.729	0.082
20%	0.800	0.188	0.065

The model is robust to 10% noise on the full layer (−1.8 points), but **hard-evidence discrimination is highly sensitive to anchor noise** (−27 points; near-collapse at 20%): **the quality of human anchors is the single lifeline of attribution accuracy** — positive-only supervision gives immunity to missing labels, not to wrong labels, so the anchor quality control of Section 4.2 (τ=0.8 + manual verification) is a prerequisite for correctness.

Uncertainty-aware calibration (Section 6.5). The uncertain-cell positive bias (mean 0.602) shows that calibration holds only on adjudicated cells; the improvement direction is explicit fallback on low-evidence-density cells at inference, returning "uncertain" to human adjudication.

7.3 Ethics and privacy

Only publicly available papers and contribution-statement data are used; author-disambiguation results come from public pipelines. Attribution inference is used only for "evidence-supported participation" and never generates any negative inference of the form "an author lacks a competence" (the social meaning of Constraint 2); the recommender system is positioned as an assistive tool, with final collaboration decisions made by humans.

7.4 Limitations

(1) **Corpus coverage:** a single discipline and journal (PLOS Computational Biology, 750 papers); generalization remains to be verified. (2) **Annotation scale:** 3,306 tuples (2,326 valid judgements); reliability for rare roles is limited. (3) **Intentional cost of the conservative design:** missing declarations are not treated as negatives (Constraint 2), which may underestimate senior authors' minor participation. (4) **Historical coverage and the cold-start boundary:** 8.6% of gold author–method

pairs have a non-zero prior (25.3% of queries) — an inherent property of the setting; in the pure cold-start region history-based methods reach their boundary and the profile fallback is a counter-signal (Table 7, Section 6.6). (5) **Positive bias on uncertain cells:** under ambiguous evidence the model leans toward participation (mean score 0.602); calibration covers adjudicated cells only (Sections 6.5, 7.2). (6) **Conditionality of the PU design's value:** with 95.6% structured CRediT coverage, missing declarations approximate true negatives here; the benefit of Constraint 2 in high-declaration-missing disciplines is a testable hypothesis, not an established result (Table 6 shows no drop on this corpus). (7) **Implementation boundary:** the encoder is TF-IDF (Section 5.6); two-hop propagation and joint training are specified but not enabled (Table 6) — the reported system is a supervised attribution layer with an analytic scoring function, not an end-to-end jointly optimized recommender. (8) **Ranking contribution of the attribution layer:** an author-level declared-role histogram (B7) matches the full layer on all ranking protocols (Tables 3, 5 and 7); the layer's evidenced contribution is attribution quality (RQ1) and auditability, not ranking gain. (9) **Adversarial evaluation construction:** main-set candidates penalize history-based signals by design, so absolute scores fall below random ordering; Task B conclusions are relative comparisons under identical inputs, not absolute competence claims. (10) **Per-role resolution:** Task A metrics are role-aggregated; tuple-level discrimination is weak (micro-AUPRC 0.234; Section 6.1) because no cell-level role supervision exists — the AUPRC 1.000 headline must be read as participation attribution only.

---

8. Conclusion and Future Work

To address the knowledge-conflation problem in task-driven collaborator recommendation, this paper proposes a contribution-evidence-constrained author–method–role attribution layer: coarse CRediT statements and fine text anchors provide hierarchical weak supervision for many-to-many latent participation associations, and the per-role outputs feed an analytic factorized scoring function, the attribution layer being trained with supervised losses alone. Under same-information comparison, the method outperforms the strongest baselines in role-aggregated attribution accuracy (full-test AUPRC 1.000 vs 0.980; hard-evidence 1.000 vs 0.890) and improves recommendation over the strongest prior-only baseline (main-set NDCG@10 +8%, Holm $p<0.01$, robust over 6 seeds and cluster bootstrap; on the full query set it remains below the random-ordering reference, exceeding it only on cold-start queries). Two boundary findings are reported with equal prominence: on the adversarially constructed main set all history-based rankers score below random ordering, and an author-level role-histogram ablation (B7) matches the full attribution layer on ranking — the attribution layer's evidenced contribution is accurate, calibrated and auditable participation attribution, not ranking gain. A per-role analysis bounds the attribution claim to the participation level (tuple-level micro-AUPRC 0.234; Section 6.1). Ablations identify text anchors as the single strongly necessary supervision component; the coarse-supervision and PU-constraint terms are not

empirically necessary on this corpus and are retained as design provisions for higher-missing-rate settings, where their value remains to be tested (Section 6.3).

Future work: (1) enabling the joint-optimization pathway ($\lambda_3 > 0$) with the warm-start stability measures of Online Resource 1, Section S8; (2) extending from single-person retrieval to role-complementary team-composition optimization; (3) transferring to expert matching for grant review; (4) temporally dynamic modelling of author-competence evolution to mitigate the underestimation of senior authors' minor participation (Section 7.4); (5) human evaluation of the relevance and sufficiency of the evidence chains behind positive predictions, to validate the auditability claim beyond score decomposition (Sections 6.4–6.5); (6) cell-level role supervision to strengthen per-role resolution (Section 6.1).

References

1. Assraf, I., & Fire, M. (2025). Behind the byline: A large-scale study of scientific author contributions. *arXiv*. https://arxiv.org/abs/2505.06721

2. Chen, C., Li, C., Sun, J., Yao, M., & Wu, D. (2026). *Contribution-scoped attribution: Code, data and annotation artefacts* [Data set]. GitHub. https://github.com/seucchen-eng/contribution-scoped-attribution

3. Lee, J.-W., & Kim, J.-H. (2024). GraphRec-based Korean expert recommendation using author contribution index and the paper abstracts in marine. *Engineering Applications of Artificial Intelligence*, *133*, 108219. https://doi.org/10.1016/j.engappai.2024.108219

4. Li, X., Verginer, L., Riccaboni, M., & Panzarasa, P. (2020). A network approach to expertise retrieval based on path similarity and credit allocation. *arXiv*. https://arxiv.org/abs/2009.13958

5. Malkov, Y. A., & Yashunin, D. A. (2020). Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(4), 824–836. https://doi.org/10.1109/TPAMI.2018.2889473

6. National Information Standards Organization. (n.d.). *CRediT: Contributor Roles Taxonomy*. https://credit.niso.org/

7. Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv*. https://doi.org/10.48550/arXiv.2205.01833

8. Robinson-Garcia, N., Costas, R., Sugimoto, C. R., Larivière, V., & Nane, G. F. (2020). Task specialization across research careers. *eLife*, *9*, e60586. https://doi.org/10.7554/eLife.60586

9. Seo, W., & Bu, Y. (2025). Transforming role classification in scientific teams using LLMs and advanced predictive analytics. *arXiv*. https://arxiv.org/abs/2501.07267

10. Singh, A., D'Arcy, M., Cohan, A., Downey, D., & Feldman, S. (2023). SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. https://arxiv.org/abs/2212.09541

11. Xu, S., Li, L., Wang, C., An, X., & Yang, G. (2022). An improved author-topic (AT) model with authorship credit allocation schemes. *Journal of Information Science*. https://doi.org/10.1177/01655515221133530

12. Yang, M., & Hsu, W. H. (2016). HDPauthor: A new hybrid author-topic model using latent Dirichlet allocation and hierarchical Dirichlet processes. In *Proceedings of the 25th International Conference Companion on World Wide Web* (pp. 619–624). ACM. https://doi.org/10.1145/2872518.2890561

Statements and Declarations

Funding. This work was supported by the "AI 'Three-Synergy' Innovation Platform Service Capacity Building" project of the Department of Education of Jiangsu Province (Grant No. 7660860001).

Competing interests. The authors have no relevant financial or non-financial interests to disclose.

Author contributions. Chao Chen: conceptualization, methodology, writing—original draft, supervision, project administration. Chunyang Li: data curation, software, investigation. Jixiang Sun: software, validation, formal analysis. Mengcheng Yao: data curation, investigation. Dandan Wu: validation, writing—review and editing.

Data availability. The data and code that support the findings of this study, including the annotation guideline, the adjudication rules, and the de-identified annotation data, are openly available on GitHub [2]: https://github.com/seucchen-eng/contribution-scoped-attribution

AI-use disclosure. Kimi K3 assisted with English translation, analysis code, data preparation, and preliminary source checks. The authors reviewed the manuscript and checked the applicable AI-tool terms and publication permissions. They take responsibility for the integrity of the entire work.

Ethics approval. Not applicable.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.docx