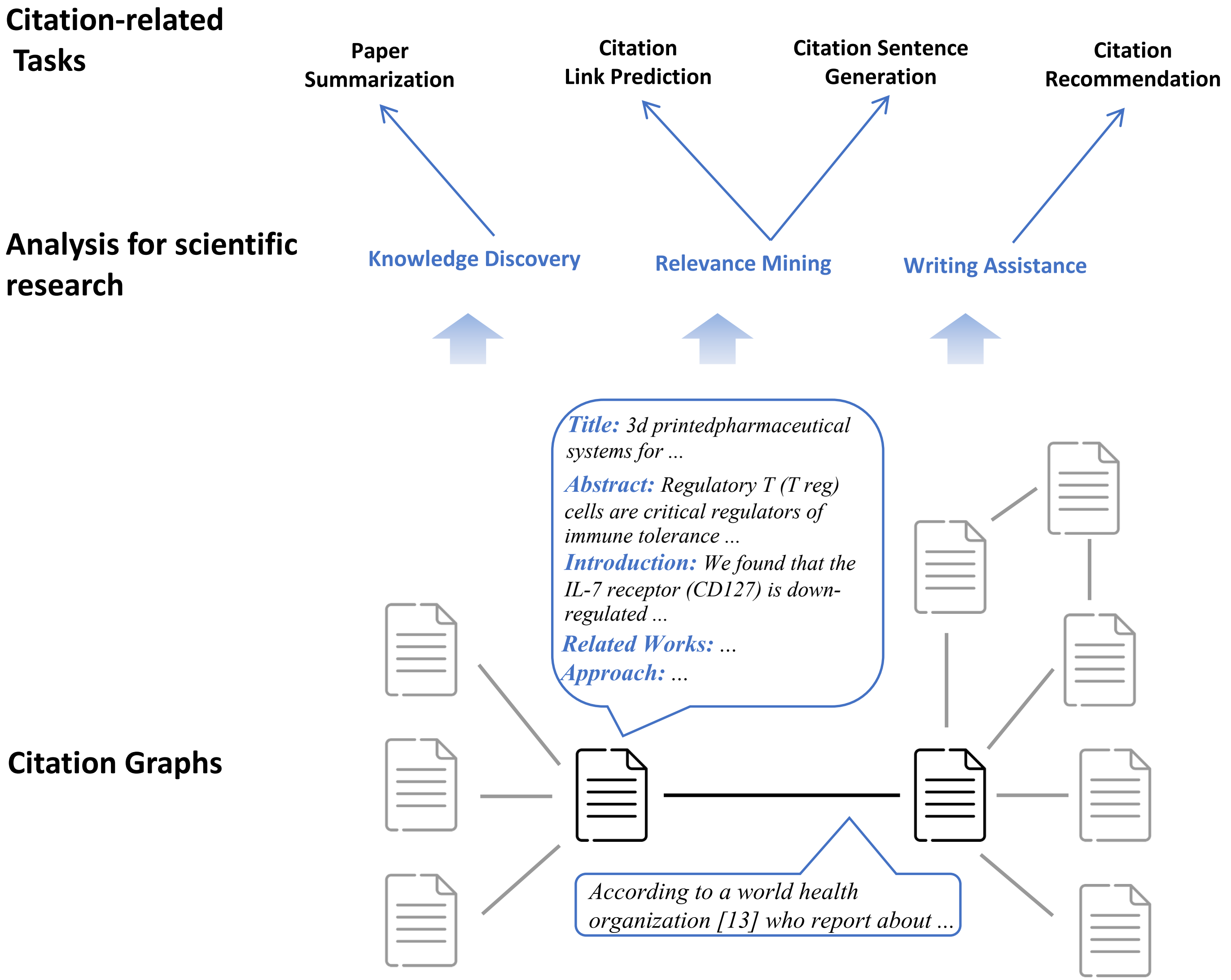
LitFM: A Retrieval Augmented Structure-aware Foundation Model For Citation Graphs

Submitted to AAAI 2025

Background



Citation-related Tasks

Paper Summarization

Citation Link Prediction

Citation Sentence Generation

Citation Recommendation

Analysis for scientific research

Knowledge Discovery

Relevance Mining

Writing Assistance

Title: *3d printedpharmaceutical systems for ...*

Abstract: *Regulatory T (T reg) cells are critical regulators of immune tolerance ...*

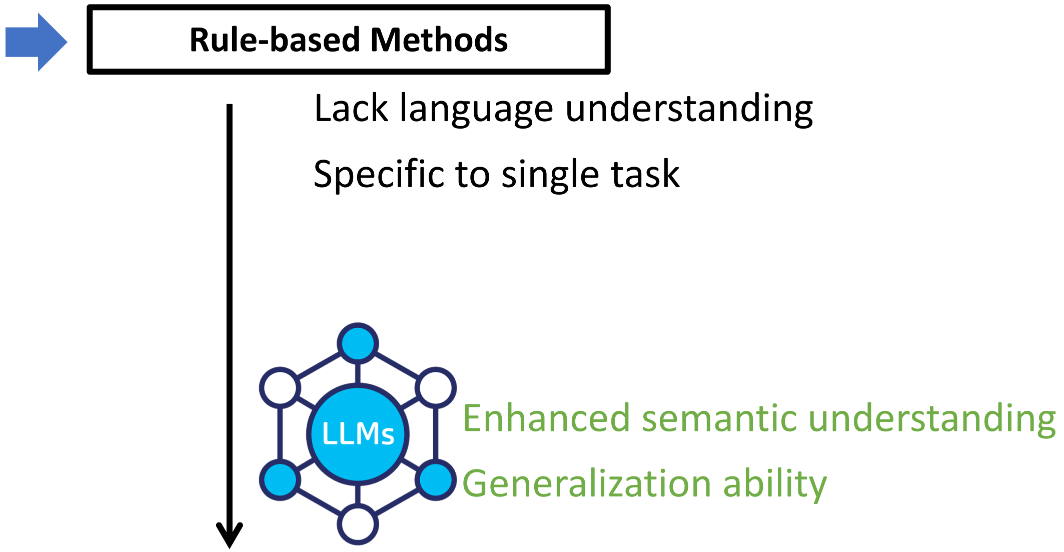
Introduction: *We found that the IL-7 receptor (CD127) is down-regulated ...*

Related Works: *...*

Approach: *...*

According to a world health organization [13] who report about ...

Citation Graphs



Rule-based Methods

Lack language understanding

Specific to single task

Enhanced semantic understanding

Generalization ability

Can we uniformly manage various citation-related tasks via LLMs?

Limitation of Existing methods

(1) Statistic-based methods

- Inverted document frequency (tf–idf), PageRank [1]
- Clustering [2]

Lack language understanding

(2) Pre-traned Language model-based methods

- SciBERT [3]
- PubMedBert [4]
- SciMult [5] ....

Unable to handle generation tasks such as paper abstract generation

---

[1] Full-text citation analysis: A new method to enhance scholarly networks

[2] Clustering Scientific Literature Using Sparse Citation Graph Analysis

[3] SciBERT: A pretrained language model for scientific text

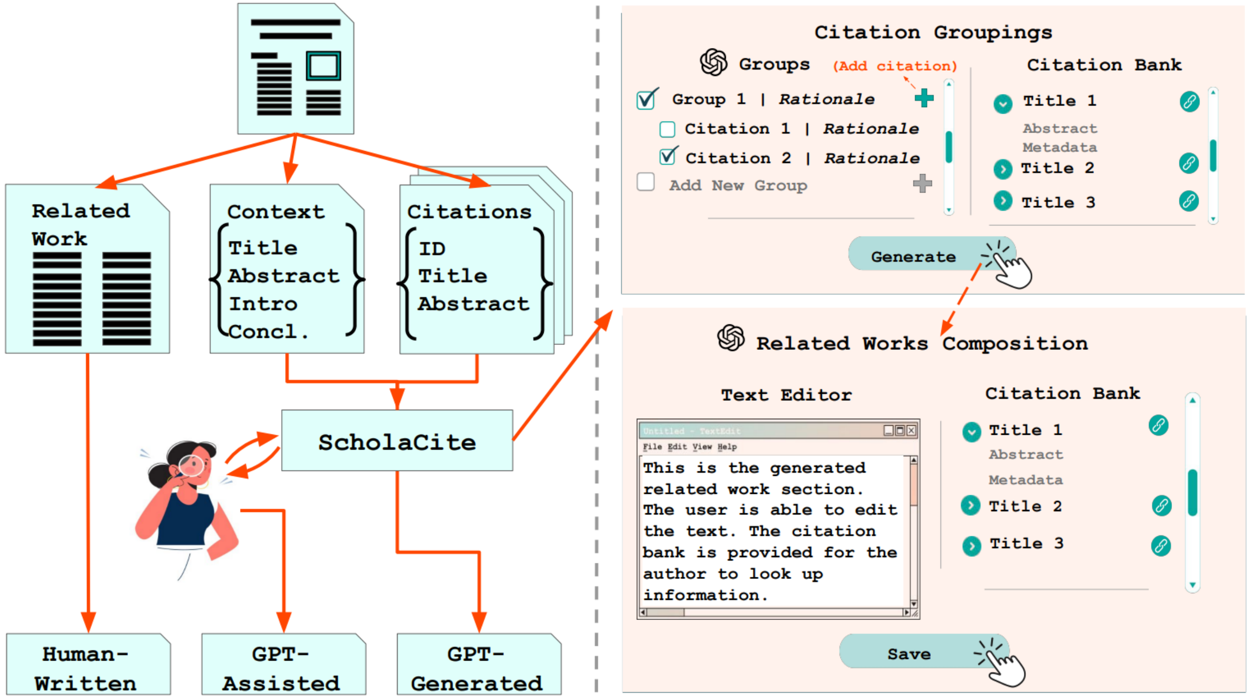
[4] Domain-specific language model pretraining for biomedical natural language processing

[5] Pre-training multi-task contrastive learning models for scientific literature understanding

Limitation of Existing methods

(3) Large Language model-based methods

- Shallow Synthesis of Knowledge in GPT-Generated Texts: A Case Study in Automatic Related Work Composition. ACL 2024



This is the generated related work section. The user is able to edit the text. The citation bank is provided for the author to look up information.

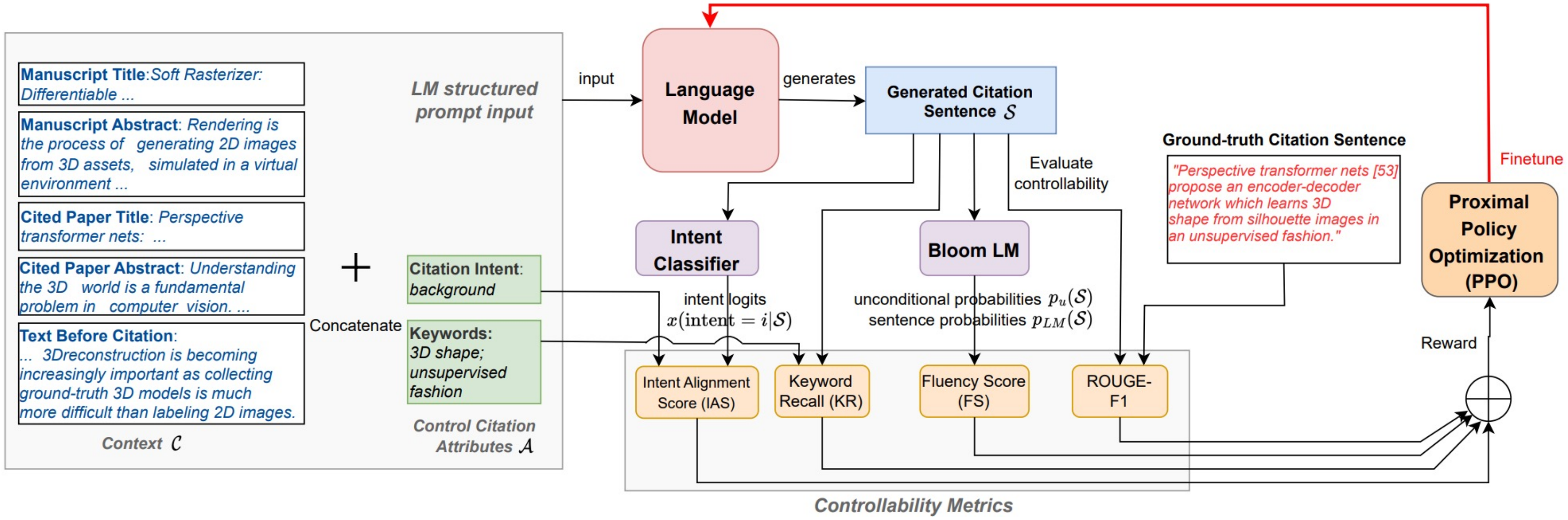
Limitations

- Without fine-tunning, **lackinng domain-specific knowledge**
- Human intervention required

Limitation of Existing methods

(3) Large Language model-based methods

- Controllable Citation Sentence Generation with Language Models. ACL 2024



Manuscript Title: *Soft Rasterizer: Differentiable ...*

Manuscript Abstract: *Rendering is the process of generating 2D images from 3D assets, simulated in a virtual environment ...*

Cited Paper Title: *Perspective transformer nets: ...*

Cited Paper Abstract: *Understanding the 3D world is a fundamental problem in computer vision. ...*

Text Before Citation: *... 3Dreconstruction is becoming increasingly important as collecting ground-truth 3D models is much more difficult than labeling 2D images.*

Citation Intent: *background*

Keywords: *3D shape; unsupervised fashion*

Ground-truth Citation Sentence

"Perspective transformer nets [53] propose an encoder-decoder network which learns 3D shape from silhouette images in an unsupervised fashion."

Context $\mathcal{C}$
Concatenate
Control Citation Attributes $\mathcal{A}$
LM structured prompt input
input
Language Model
generates
Generated Citation Sentence $\mathcal{S}$
Evaluate controllability
Intent Classifier
Bloom LM
intent logits $x(\text{intent} = i|\mathcal{S})$
unconditional probabilities $p_u(\mathcal{S})$
sentence probabilities $p_{LM}(\mathcal{S})$
Intent Alignment Score (IAS)
Keyword Recall (KR)
Fluency Score (FS)
ROUGE-F1
Controllability Metrics
Reward
Proximal Policy Optimization (PPO)
Finetune

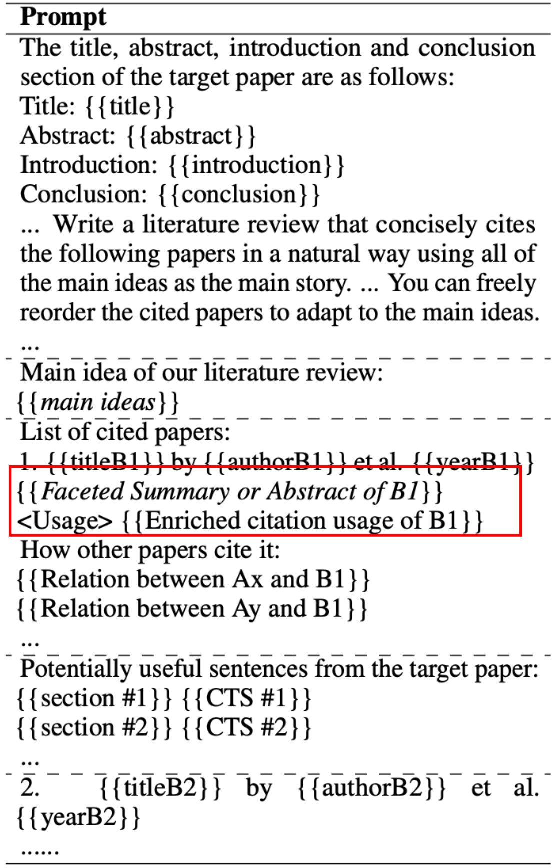
Limitations

- Specific to citation sentence generation task
- Lack citation context

Limitation of Existing methods

(3) Large Language model-based methods

- Explaining Relationships Among Research Papers. Arxiv 2024



Prompt

The title, abstract, introduction and conclusion section of the target paper are as follows:
Title: {{title}}
Abstract: {{abstract}}
Introduction: {{introduction}}
Conclusion: {{conclusion}}
... Write a literature review that concisely cites the following papers in a natural way using all of the main ideas as the main story. ... You can freely reorder the cited papers to adapt to the main ideas.
...

Main idea of our literature review:
{{*main ideas*}}

List of cited papers:
1. {{titleB1}} by {{authorB1}} et al. {{yearB1}}
{{*Faceted Summary or Abstract of B1*}}
<Usage> {{Enriched citation usage of B1}}
How other papers cite it:
{{Relation between Ax and B1}}
{{Relation between Ay and B1}}
...

Potentially useful sentences from the target paper:
{{section #1}} {{CTS #1}}
{{section #2}} {{CTS #2}}
...

2. {{titleB2}} by {{authorB2}} et al. {{yearB2}}
......

Limitations

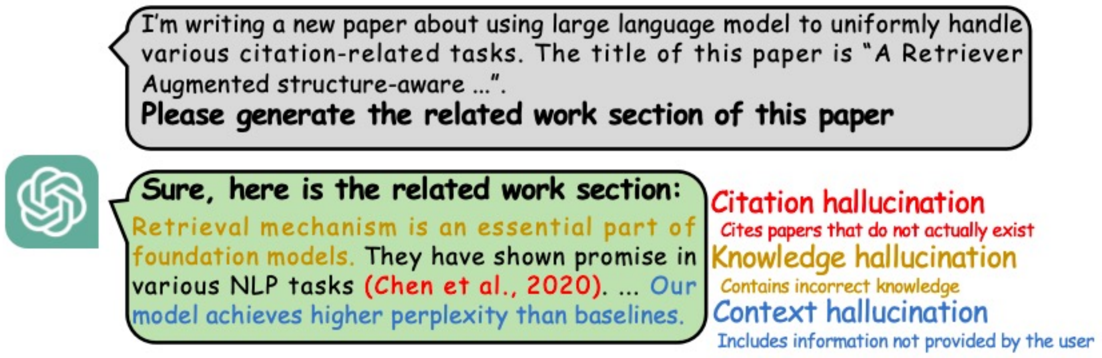
- Require the ground-truth cited papers to be provided.
- Without fine-tunning, **lackinng domain-specific knowledge**

Challenges and Motivations

Challenge 1: Unaware of Domain knowledge → **Hallucinations**

Challenge 2: Unaccess of real-world citation graph → **Low Generalization**

Challenge 3: Designed specific to single task → **Lack Unification**

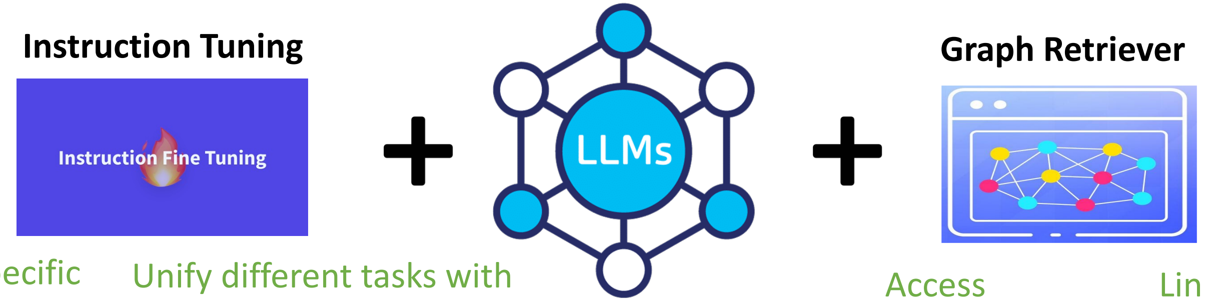


I'm writing a new paper about using large language model to uniformly handle various citation-related tasks. The title of this paper is "A Retriever Augmented structure-aware …".
Please generate the related work section of this paper

Sure, here is the related work section:
Retrieval mechanism is an essential part of foundation models. They have shown promise in various NLP tasks (Chen et al., 2020). … Our model achieves higher perplexity than baselines.

Citation hallucination
Cites papers that do not actually exist
Knowledge hallucination
Contains incorrect knowledge
Context hallucination
Includes information not provided by the user

It's urgent to develop a foundation model for citations, that can effectively handle various citation-related tasks across different domains, so that it can provide practical values in real-world usage and serve as a backbone of more complex applications.



Instruction Tuning + LLMs + **Graph Retriever**

Infuse domain-specific knowledge

Unify different tasks with prompts

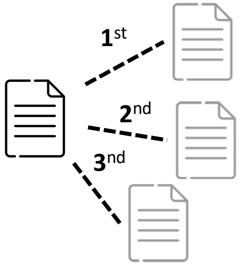
Access graph structure

Linking new papers

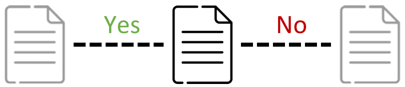
Citation Benchmark Tasks

Edge-level

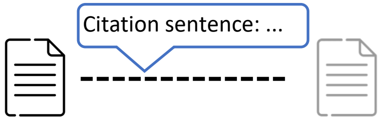
Citation Recommendation : Hits@k



Citation Link Prediction: Accuracy



Citation Sentence Generation: Bert score



Node-level

Title Generation: Bert score

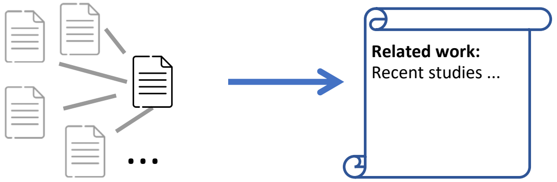
Abstract: *Regulatory T (T reg) cells are critical regulators of immune tolerance ... Finally, we show that CD127 can be used to quantitate T reg cell subsets in individuals with type 1 diabetes supporting the use of CD127 as a biomarker for human T reg cells.* → *Title:* *3d printedpharmaceutical systems for ...*

Abstract Completion: Bert score

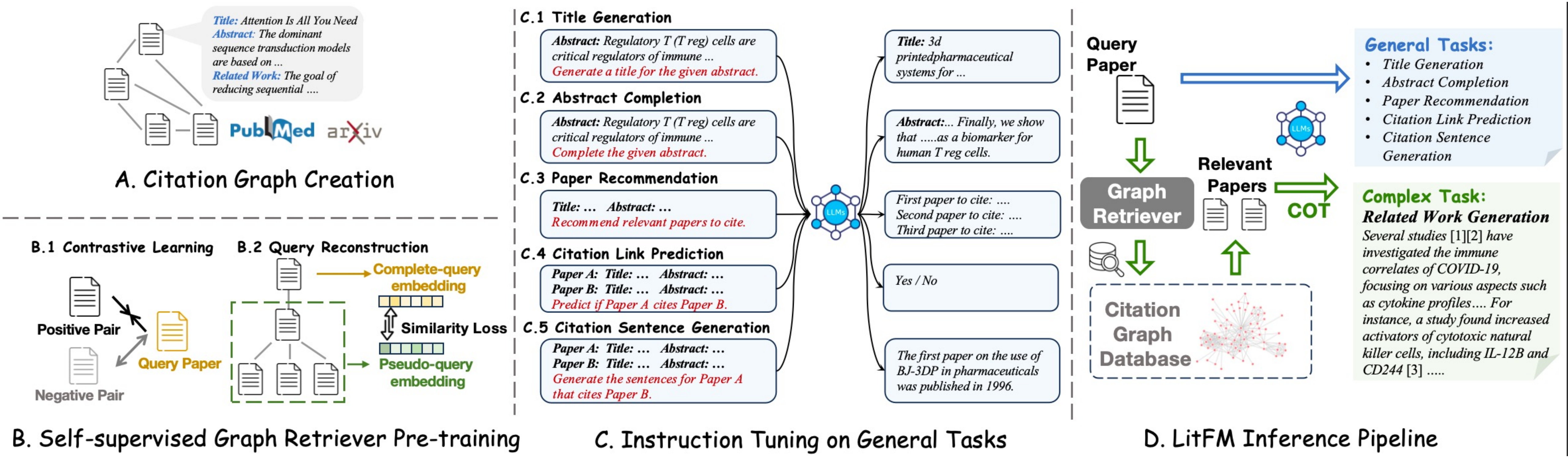
Abstract: *Regulatory T (T reg) cells are critical regulators of immune tolerance ...* → *Abstract:* *Regulatory T (T reg) cells are critical regulators of immune tolerance ... Finally, we show that CD127 can be used to quantitate T reg cell subsets in individuals with type 1 diabetes supporting the use of CD127 as a biomarker for human T reg cells.*

Subgraph-level

Related Work Generation: Bert score / ROUGE score



Overall Architecture of LitFM



Title: Attention Is All You Need
Abstract: The dominant sequence transduction models are based on ...
Related Work: The goal of reducing sequential ....

C.1 Title Generation

Abstract: *Regulatory T (T reg) cells are critical regulators of immune ...*
Generate a title for the given abstract.

Title: *3d printedpharmaceutical systems for ...*

C.2 Abstract Completion

Abstract: *Regulatory T (T reg) cells are critical regulators of immune ...*
Complete the given abstract.

Abstract:*... Finally, we show that .....as a biomarker for human T reg cells.*

C.3 Paper Recommendation

Title: ... Abstract: ...
Recommend relevant papers to cite.

First paper to cite: ....
Second paper to cite: ....
Third paper to cite: ....

C.4 Citation Link Prediction

Paper A: Title: ... Abstract: ...
Paper B: Title: ... Abstract: ...
Predict if Paper A cites Paper B.

Yes / No

C.5 Citation Sentence Generation

Paper A: Title: ... Abstract: ...
Paper B: Title: ... Abstract: ...
Generate the sentences for Paper A that cites Paper B.

The first paper on the use of BJ-3DP in pharmaceuticals was published in 1996.

General Tasks:

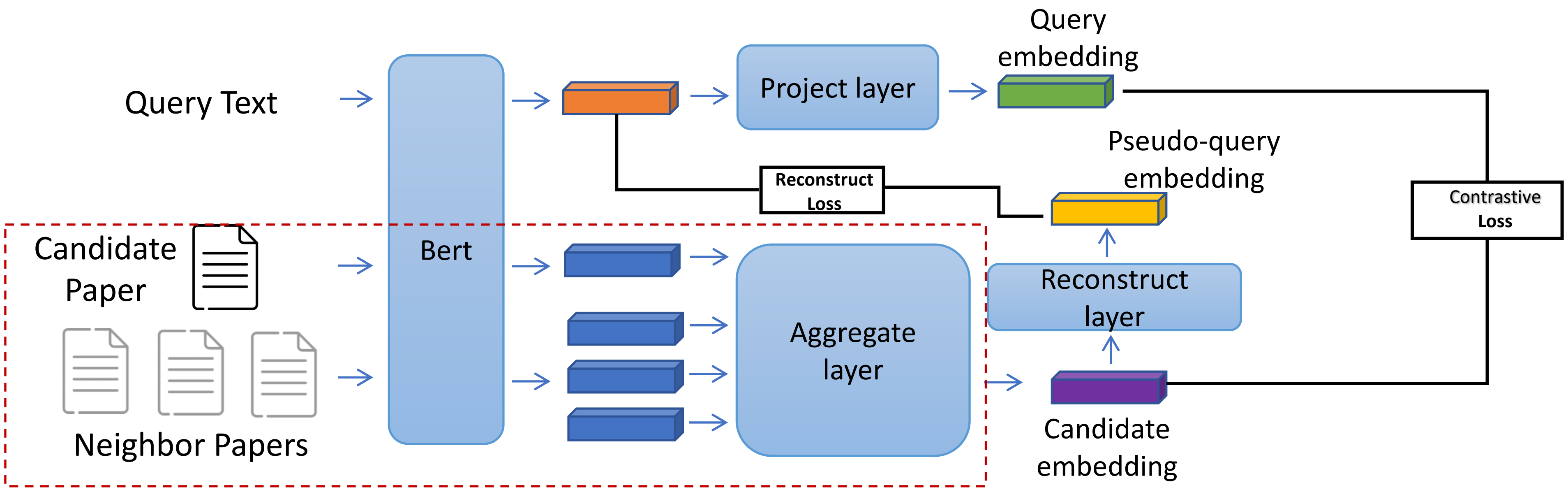
- *Title Generation*
- *Abstract Completion*
- *Paper Recommendation*
- *Citation Link Prediction*
- *Citation Sentence Generation*

Complex Task:

Related Work Generation

Several studies [1][2] have investigated the immune correlates of COVID-19, focusing on various aspects such as cytokine profiles.... For instance, a study found increased activators of cytotoxic natural killer cells, including IL-12B and CD244 [3] .....

B. Graph Retriever



Candidate embedding generation

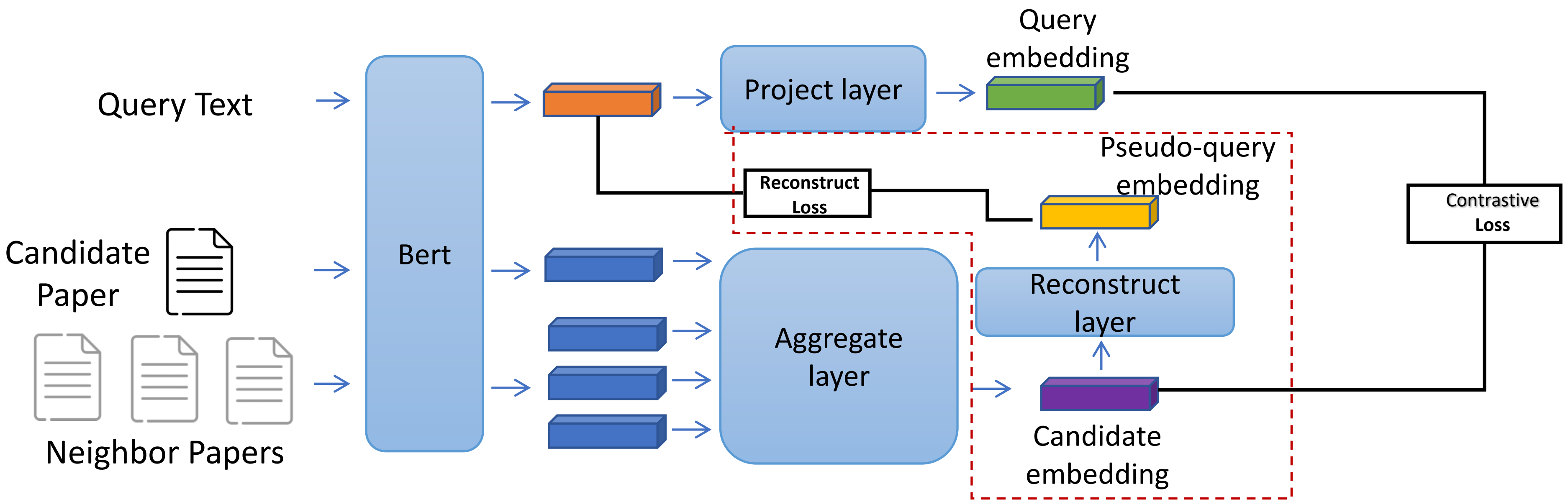
①Indexing

$$\mathbf{z}_i = \underbrace{LM}_{\text{Bert Model}}(\underbrace{T_i}_{\text{Title}}) + LM(\underbrace{A_i}_{\text{Abstract}})$$

②Neighbor Modeling

$$\underbrace{\mathbf{c}_j}_{\text{Candidate Embedding}} = \mathbf{W}^{c1}\mathbf{z}_j + \frac{1}{|\mathcal{N}(j)|}\sum_{\underbrace{k\in\mathcal{N}(j)}_{\text{Neighboring papers of paper j}}} \mathbf{W}^{c2}\mathbf{z}_k + \mathbf{b}^c$$

B. Graph Retriever



Pseudo-query Reconstruction

①Pseudo-query embedding generation

$$\underline{\mathbf{p}_j} = \mathrm{MLP}(\mathbf{c}_j)$$

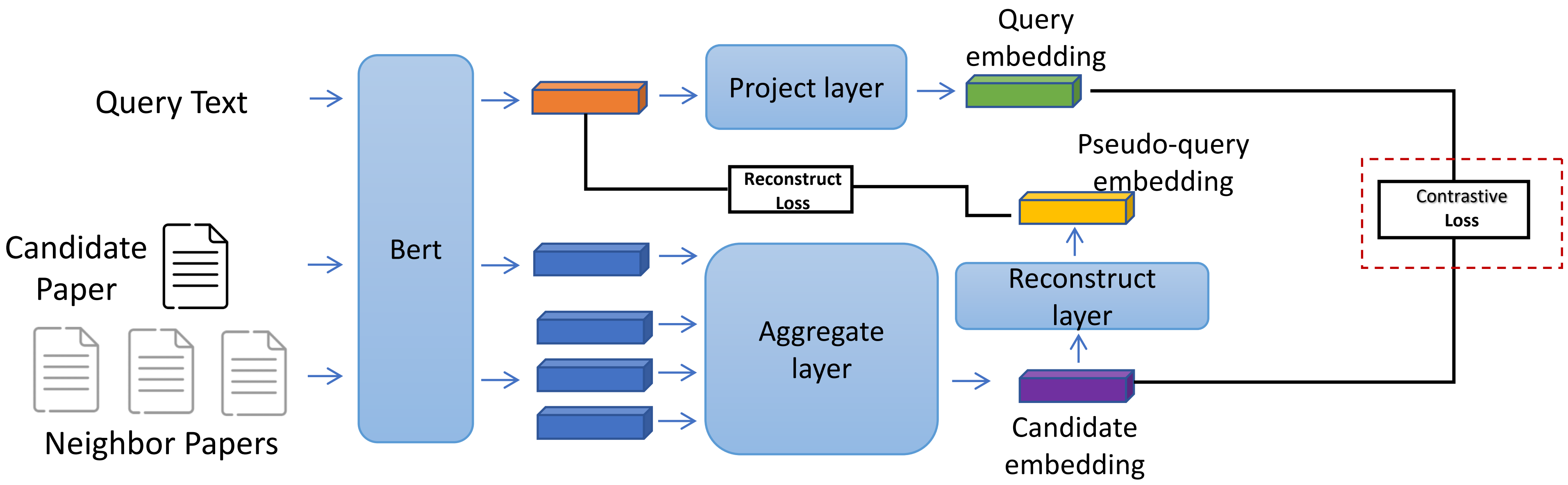
Pseudo-query Embedding of paper j

②Reconstruction loss

$$\mathcal{L}_{\mathrm{re}} = \sum_{\underline{(i,j)\in\mathcal{G}}} ||\mathbf{z}_i - \mathbf{p}_j||_1.$$

Citing paper of j as ground-truth query

B. Graph Retriever



Self-supervised Training

①similarity caculation

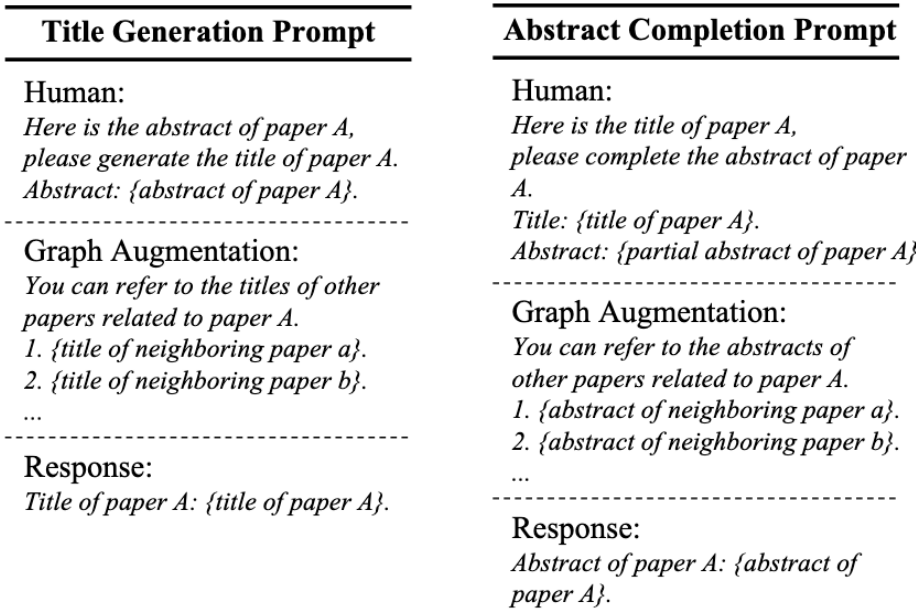
$$\mathrm{sim}(i,j)=\cos(\underbrace{\mathbf{q}_i}_{\text{Embedding of query}},\underbrace{\mathbf{c}_j}_{\text{Embedding of candidate paper j}}+\underbrace{\mathbf{p}_j}_{\text{Pseudo-query Embedding of j}})$$

②Contrastive loss

$$\mathcal{L}_{\mathrm{nce}}=-\frac{1}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)}\log\frac{\exp(\mathrm{sim}(i,j))}{\underbrace{\sum_{j'\notin\mathcal{N}(i)}}_{\text{Negative samples}}\exp(\mathrm{sim}(i,j'))}$$

C. Instruction-tuning Paradigm

Paper Content Understanding



Title Generation Prompt

Human:
Here is the abstract of paper A,
please generate the title of paper A.
Abstract: {abstract of paper A}.

Graph Augmentation:
You can refer to the titles of other papers related to paper A.
1. {title of neighboring paper a}.
2. {title of neighboring paper b}.
…

Response:
Title of paper A: {title of paper A}.

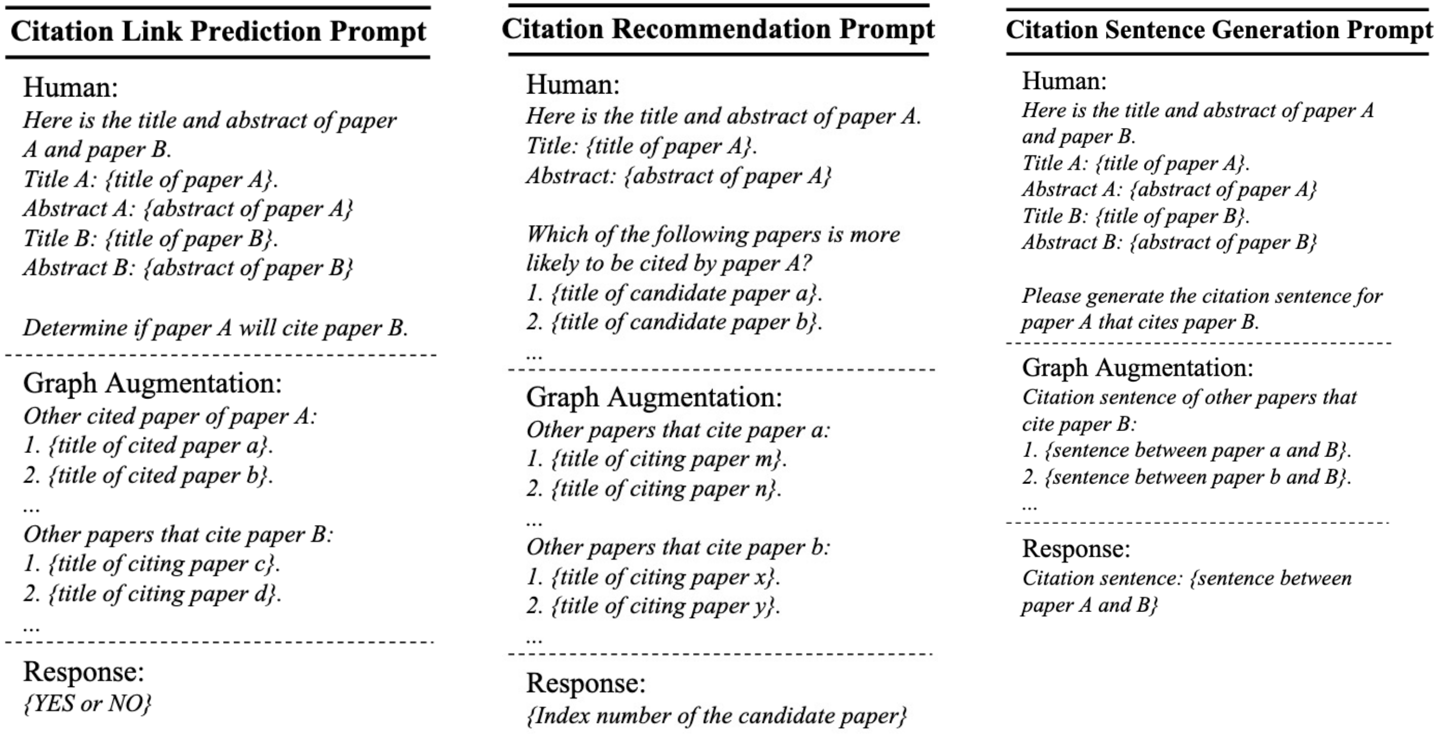
Abstract Completion Prompt

Human:
Here is the title of paper A,
please complete the abstract of paper A.
Title: {title of paper A}.
Abstract: {partial abstract of paper A}

Graph Augmentation:
You can refer to the abstracts of other papers related to paper A.
1. {abstract of neighboring paper a}.
2. {abstract of neighboring paper b}.
…

Response:
Abstract of paper A: {abstract of paper A}.

Paper Relevance Understanding



Citation Link Prediction Prompt

Human:
Here is the title and abstract of paper A and paper B.
Title A: {title of paper A}.
Abstract A: {abstract of paper A}
Title B: {title of paper B}.
Abstract B: {abstract of paper B}

Determine if paper A will cite paper B.

Graph Augmentation:
Other cited paper of paper A:
1. {title of cited paper a}.
2. {title of cited paper b}.
…
Other papers that cite paper B:
1. {title of citing paper c}.
2. {title of citing paper d}.
…

Response:
{YES or NO}

Citation Recommendation Prompt

Human:
Here is the title and abstract of paper A.
Title: {title of paper A}.
Abstract: {abstract of paper A}

Which of the following papers is more likely to be cited by paper A?
1. {title of candidate paper a}.
2. {title of candidate paper b}.
…

Graph Augmentation:
Other papers that cite paper a:
1. {title of citing paper m}.
2. {title of citing paper n}.
…
Other papers that cite paper b:
1. {title of citing paper x}.
2. {title of citing paper y}.
…

Response:
{Index number of the candidate paper}

Citation Sentence Generation Prompt

Human:
Here is the title and abstract of paper A and paper B.
Title A: {title of paper A}.
Abstract A: {abstract of paper A}
Title B: {title of paper B}.
Abstract B: {abstract of paper B}

Please generate the citation sentence for paper A that cites paper B.

Graph Augmentation:
Citation sentence of other papers that cite paper B:
1. {sentence between paper a and B}.
2. {sentence between paper b and B}.
…

Response:
Citation sentence: {sentence between paper A and B}

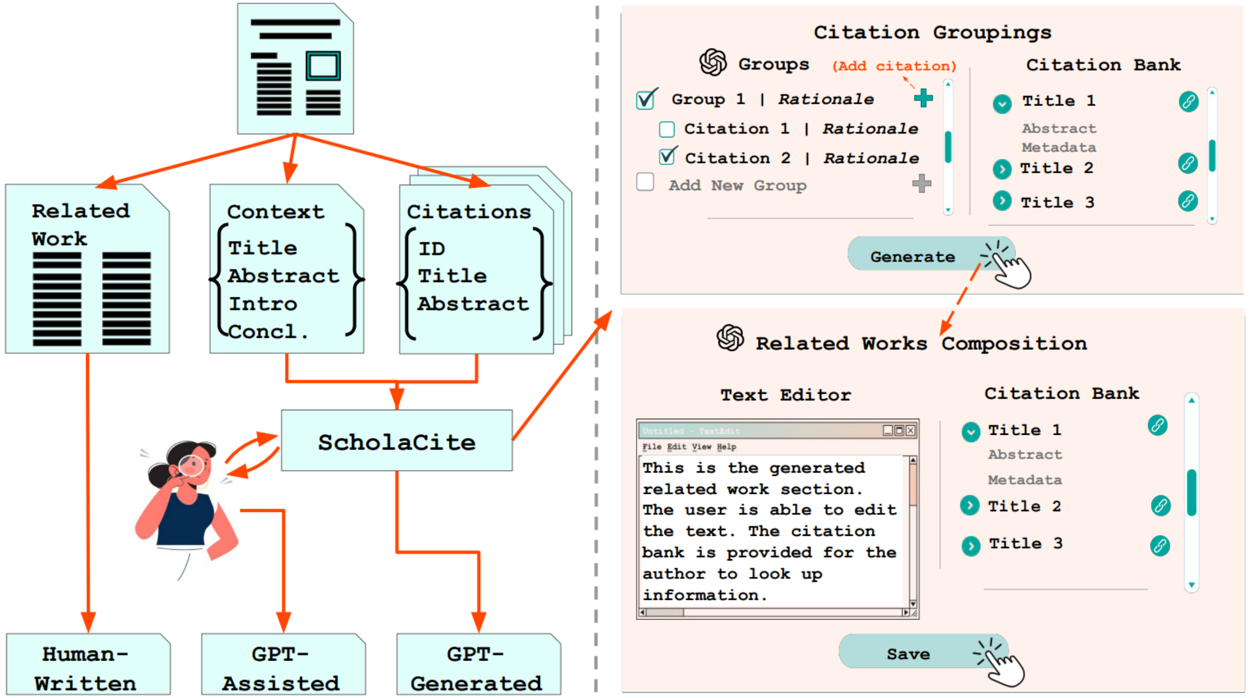
D. Chain-of-Thought Strategy

Challenges of existing methods in generating related work sections

1. Generate fake citations.
2. Cannot handle new papers that not exist in citation graphs.
3. Human intervention required.

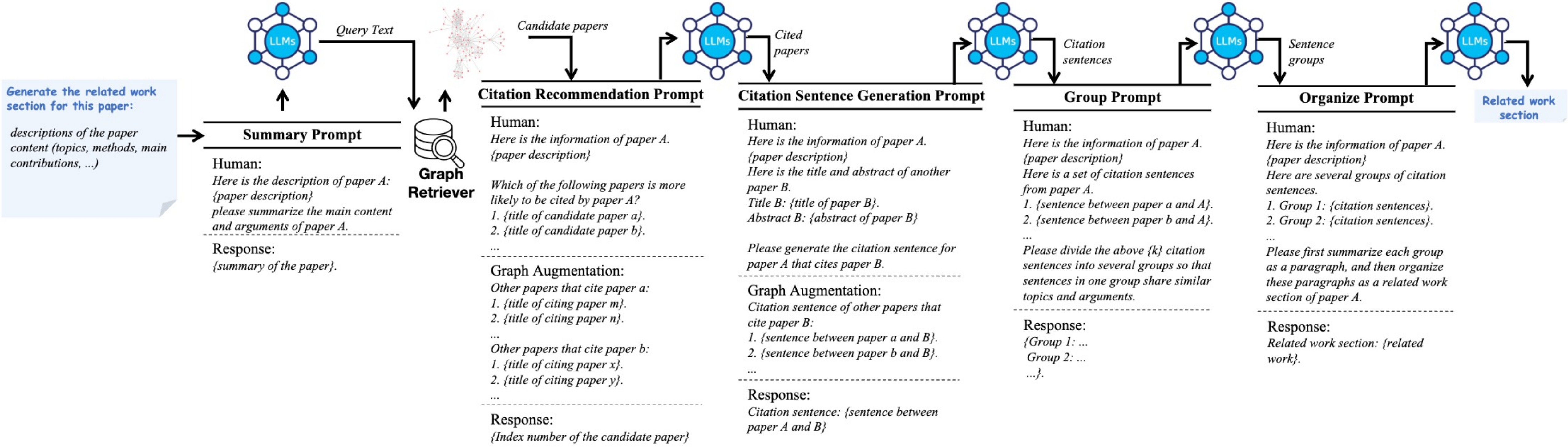
} **Graph Retriever** (1, 2)

→ **Chain-of-Thought Strategy** (3)



This is the generated related work section. The user is able to edit the text. The citation bank is provided for the author to look up information.

D. Chain-of-Thought Strategy



Generate the related work section for this paper:

descriptions of the paper content (topics, methods, main contributions, ...)

Summary Prompt

Human:
Here is the description of paper A:
{paper description}
please summarize the main content and arguments of paper A.

Response:
{summary of the paper}.

Citation Recommendation Prompt

Human:
Here is the information of paper A.
{paper description}

Which of the following papers is more likely to be cited by paper A?
1. {title of candidate paper a}.
2. {title of candidate paper b}.
...

Graph Augmentation:
Other papers that cite paper a:
1. {title of citing paper m}.
2. {title of citing paper n}.
...
Other papers that cite paper b:
1. {title of citing paper x}.
2. {title of citing paper y}.
...

Response:
{Index number of the candidate paper}

Citation Sentence Generation Prompt

Human:
Here is the information of paper A.
{paper description}
Here is the title and abstract of another paper B.
Title B: {title of paper B}.
Abstract B: {abstract of paper B}

Please generate the citation sentence for paper A that cites paper B.

Graph Augmentation:
Citation sentence of other papers that cite paper B:
1. {sentence between paper a and B}.
2. {sentence between paper b and B}.
...

Response:
Citation sentence: {sentence between paper A and B}

Group Prompt

Human:
Here is the information of paper A.
{paper description}
Here is a set of citation sentences from paper A.
1. {sentence between paper a and A}.
2. {sentence between paper b and A}.
...
Please divide the above {k} citation sentences into several groups so that sentences in one group share similar topics and arguments.

Response:
{Group 1: ...
Group 2: ...
...}.

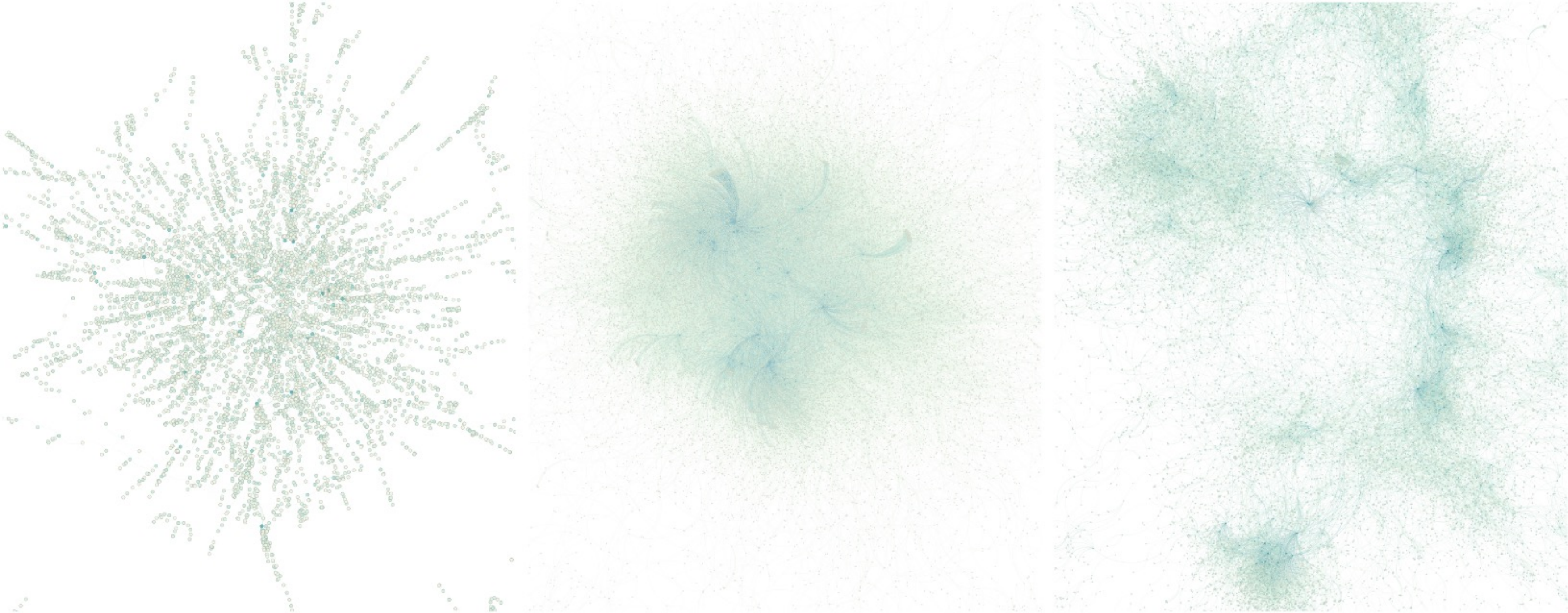
Organize Prompt

Human:
Here is the information of paper A.
{paper description}
Here are several groups of citation sentences.
1. Group 1: {citation sentences}.
2. Group 2: {citation sentences}.
...
Please first summarize each group as a paragraph, and then organize these paragraphs as a related work section of paper A.

Response:
Related work section: {related work}.

Related work section

Citation Graph Benchmark Datasets



(a) Medicine graph.

Nodes: 2.1 M
Edges: 7.4 M
Related Works: 1.5 M
Raw Data: Pubmed

(b) Computer science graph.

Nodes: 349 K
Edges: 3.2 M
Related Works: 188 K
Raw Data: Arxiv

(c) Physics graph.

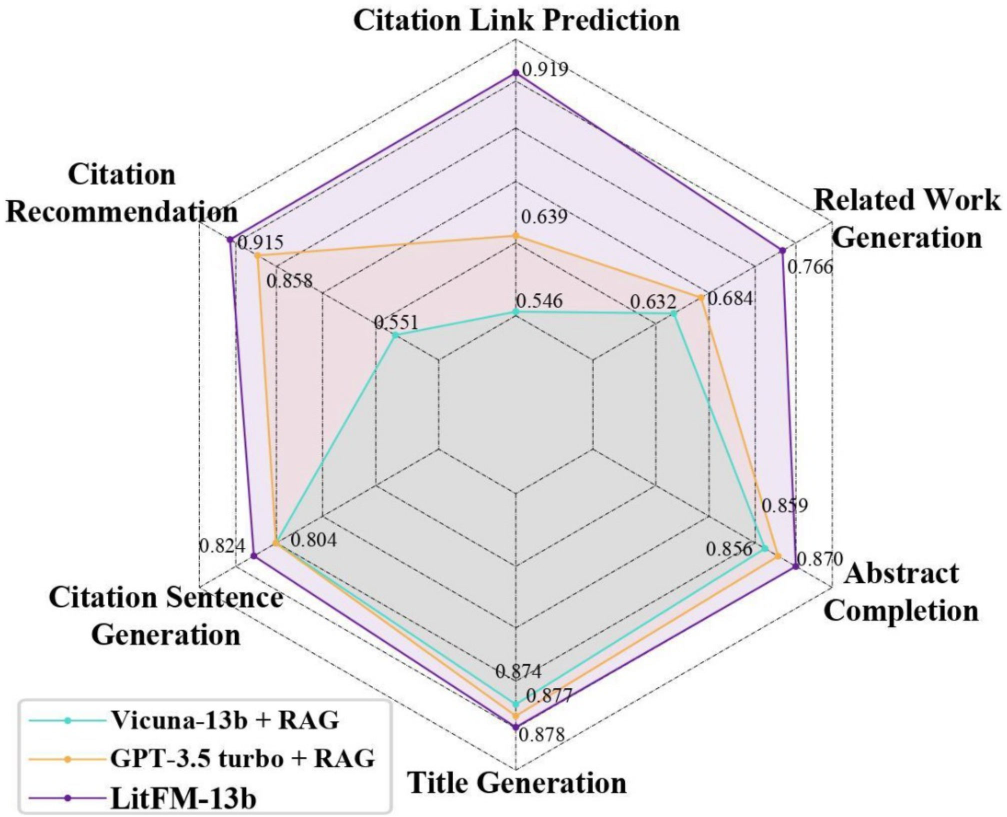
Nodes: 59 K
Edges: 120 K
Related Works: 19 K
Raw Data: Arxiv

Compared with exising datasets:

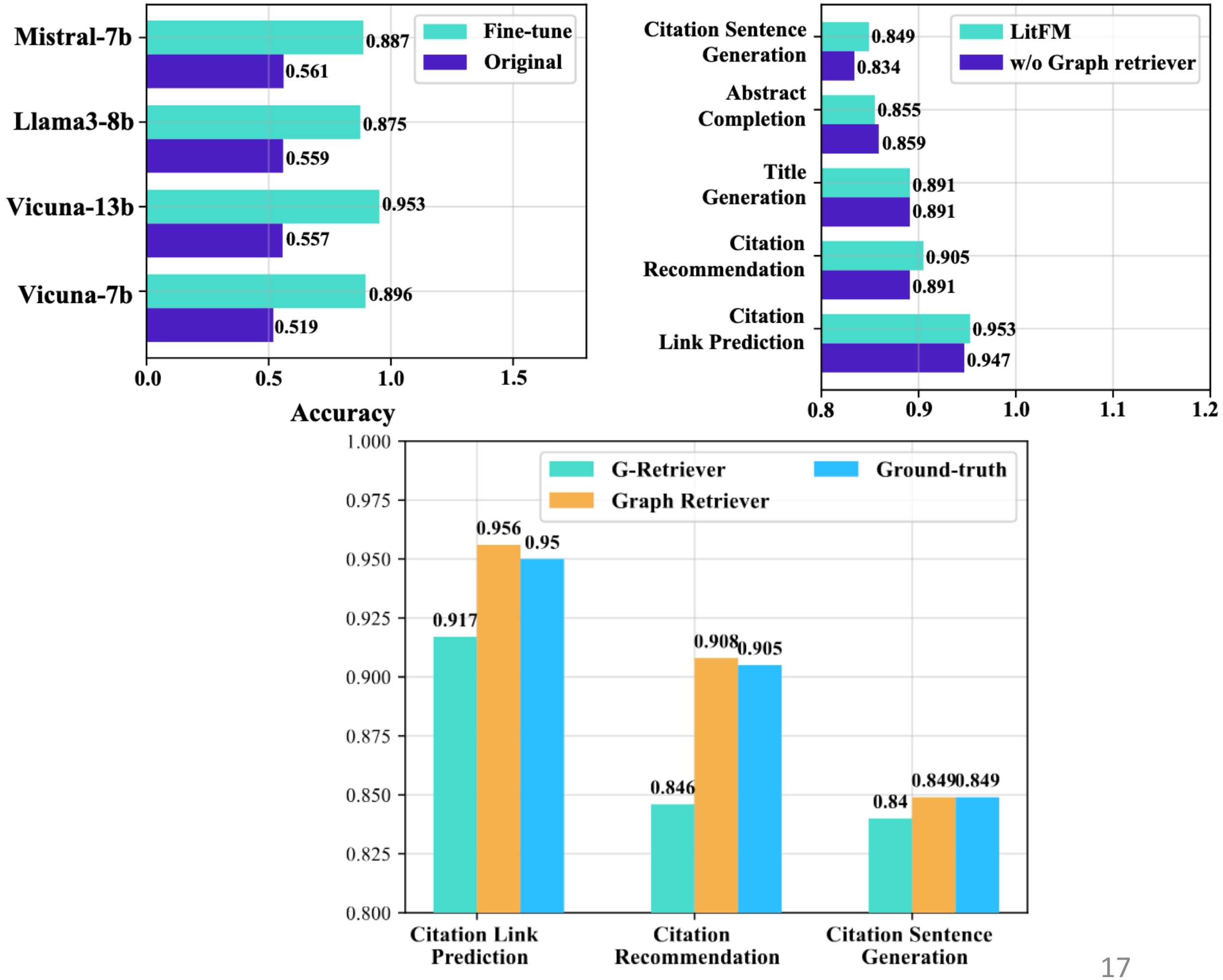
1. Node attributes include the paper’s related work section (when available)
2. Each edge is annotated with the citing sentence, with its local context.

Performance

Overall Comparsion



Ablation Study



Case Study (Citation Sentence Generation)

Title A	Abstract A	Title B	Abstract B	Ground-Truth	GPT-3.5 Turbo	LitFM
PtyLab: a cross-platform, open-source inverse modeling toolbox for conventional and Fourier ptychography	Conventional (CP) and Fourier (FP) ptychography have emerged as versatile quantitative phase imaging techniques. While the main application cases for each technique are different, namely lens-less short wavelength imaging for CP and lens-based visible light imaging for FP, both methods share a common algorithmic ground. CP and FP have in part independently evolved to include experimentally robust forward models and inversion techniques. This separation has resulted in a plethora of algorithmic extensions, some of which have not crossed the boundary from one modality to the other. Here, we present an open source, cross-platform software, called PtyLab, enabling both CP and FP data analysis in a unified framework. With this framework, we aim to facilitate and accelerate cross-pollination between the two techniques. Moreover, the availability in Matlab, Python, and Julia will set a low barrier to enter each field.	Using Automatic Differentiation as a General Framework for Ptychographic Reconstruction	Coherent diffraction imaging methods enable imaging beyond lens-imposed resolution limits. In these methods, the object can be recovered by minimizing an error metric that quantifies the difference between diffraction patterns as observed, and those calculated from a present guess of the object. Efficient minimization methods require analytical calculation of the derivatives of the error metric, which is not always straightforward. This limits our ability to explore variations of basic imaging approaches. In this paper, we propose to substitute analytical derivative expressions with the automatic differentiation method, whereby we can achieve object reconstruction by specifying only the physics-based experimental forward model. We demonstrate the generality of the proposed method through straightforward object reconstruction for a variety of complex ptychographic experimental models.	Another line of research has investigated the capabilities opened up by modern automatic differentiation (AD) and machine learning (ML) toolboxes \cite{Kandel2019}.	PtyLab builds on the automatic differentiation methods for ptychographic reconstruction introduced by [1].	Another example of a more recent development is the use of automatic differentiation and machine learning as a general framework for ptychographic reconstruction~\cite{Kandel2019} Correct citation Detailed information

Case Study (Related Work Generation)

Title	Cited papers	LitFM More knowledge More citations No fake papers	GPT3.5-turbo Sparse citations Hallucinations
Massive Access in Secure NOMA under Imperfect CSI: Security Guaranteed Sum-Rate Maximization with First-Order Algorithm	1. Securing Downlink Massive MIMO-NOMA Networks with Artificial Noise 2. Edge-Assisted V2X Motion Planning and Power Control Under Channel Uncertainty 3. Beamforming Design and Power Allocation for Secure Transmission with NOMA 4. Massive Connectivity with Massive MIMO-Part I: Device Activity Detection and Channel Estimation 5. On the Performance Gain of NOMA over OMA in Uplink Communication Systems 6. Secure Communications in a Unified Non-Orthogonal Multiple Access Framework 7. Practical Non-linear Energy Harvesting Model and Resource Allocation for SWIPT Systems 8. Robust and Secure Resource Allocation for Full-Duplex MISO Multicarrier NOMA Systems 9. Prospective Multiple Antenna Technologies for Beyond 5G 10. Downlink SDMA with Limited Feedback in Interference-Limited Wireless Networks 11. Optimal Resource Allocation for Power-Efficient MC-NOMA with Imperfect Channel State Information 12. A Survey of Physical Layer Security Techniques for 5G Wireless Networks and Challenges Ahead 13. Exploiting Inter-User Interference for Secure Massive Non-Orthogonal Multiple Access 14. Secure Users Oriented Downlink MISO NOMA	Non-orthogonal multiple access (NOMA) has been recognized as a promising solution for secure transmission under massive access [1][2]. However, the use of NOMA in practice faces challenges such as the uncertainty of channel state information (CSI), which affects the performance of the system [3][4]. In particular, the CSI of the legitimate users may be imperfect due to the limited feedback at the base station [5], while the CSI of the eavesdroppers remains unknown due to their passive nature [6]. As a result, the optimal power allocation and transmission rate design for a secure NOMA scheme remain unclear due to the difficulty of handling the probabilistic constraints [7]. To address this challenge, several research efforts have focused on developing algorithms to solve the security-guaranteed sum-rate maximization problem in NOMA systems with imperfect CSI [8][9][10]. One approach is to employ zero-forcing beamforming to suppress interference among clusters [11]. Another strategy is to apply successive interference cancellation (SIC) within each cluster to improve the system performance [12]. These techniques aim to enhance the security and efficiency of NOMA-based communication systems. In this context, our paper contributes to the literature by proposing a novel transformation of the probabilistic constraints and variable decoupling to solve the security-guaranteed sum-rate maximization problem in NOMA systems with imperfect CSI using branch-and-bound method and difference of convex programming [13]. Our proposed algorithm has a low computational complexity, making it suitable for large-scale deployment. The simulation results demonstrate that our proposed scheme significantly improves the security-guaranteed sum-rate compared to the orthogonal multiple access transmission and NOMA without considering the CSI uncertainty [14].	Recent works have addressed various aspects of NOMA systems with a focus on different performance metrics. For instance, [1] explored the sum-rate maximization in NOMA systems, emphasizing power allocation strategies to enhance spectral efficiency. Meanwhile, [2] investigated user fairness and proposed algorithms to balance the trade-off between system throughput and user equity. In the context of security, [3] discussed physical layer security in NOMA, proposing methods to safeguard communication against eavesdropping. The challenge of imperfect CSI has been a critical consideration in wireless communication research. Works like [4] have examined robust transmission strategies that account for uncertainties in CSI, proposing methods to mitigate their effects on system performance. Additionally, [5] explored the impact of imperfect CSI on the secrecy capacity of communication systems, highlighting the importance of robust design in maintaining secure communication. Despite these contributions, the problem of secure communication in NOMA systems under imperfect CSI, particularly with a focus on sum-rate maximization while ensuring security guarantees, has not been thoroughly investigated. Our work extends the existing literature by addressing this gap and proposing a first-order algorithm for optimizing the secure sum-rate in NOMA systems under imperfect CSI conditions. **[1]** Ding, Z., Adachi, F., & Poor, H. V. (2017). The application of non-orthogonal multiple access in LTE and 5G. IEEE Communications Magazine, 55(2), 185-191. **[2]** Wang, P., Zhu, Y., & Liang, Y. (2018). Secrecy capacity optimization for NOMA with perfect CSI.IEEE Transactions on Communications, 66(9), 3983-3996. **[3]** Zhang, R., Chen, W., & Huang, Y. (2019). Secure transmission in NOMA systems with imperfect CSI. IEEE Transactions on Vehicular Technology, 68(10), 9869-9873. **[4]** Liu, L., Zhang, J., & Ma, X. (2020). Robust beamforming for secure NOMA with imperfect CSI.IEEE Transactions on Wireless Communications, 19(4), 2698-2710. **[5]** Huang, Q., Wang, J., & Yang, J. (2021). First-order methods for energy-efficient NOMA systems.IEEE Transactions on Green Communications and Networking, 5(2), 655-667

Limitations and Future Directions

1. The graph retriever can only perform neighboring retrieval

Adaptive graph retriever that can perform different kinds of retrieval based on tasks.

2. The graph retriever is isolated from the training of LLM

A strategy to combine the optimization of graph retriever and LLM.

3. The time information of citations is not considered

Exploring the ability of LLM in doing time-aware question answering.