# Bridging Concepts: Multi-Agent Reasoning over Scientific Concept Graphs for High-Impact Research Prediction

**Donggyu Yoon***, **Minsung Hwang***, **Juneha Baek**, **Donghyuk Shin**, and **Joyce Jiyoung Whang**†

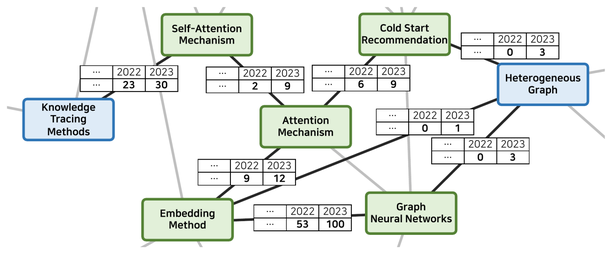
KAIST * Equal Contribution

The 3rd Workshop on Knowledge-Enhanced Information Retrieval (KEIR@CIKM 2026)

▼ BDILab ▼ GitHub

## Scientific Concept Graphs (SCGs)

- Influential research often arises when **research concepts studied apart** are first **combined in a single study**
- SCG: **research concepts** as nodes and their **co-occurrence in papers** as edges between them
  - Each edge stores the yearly citations of the papers where both concepts appear

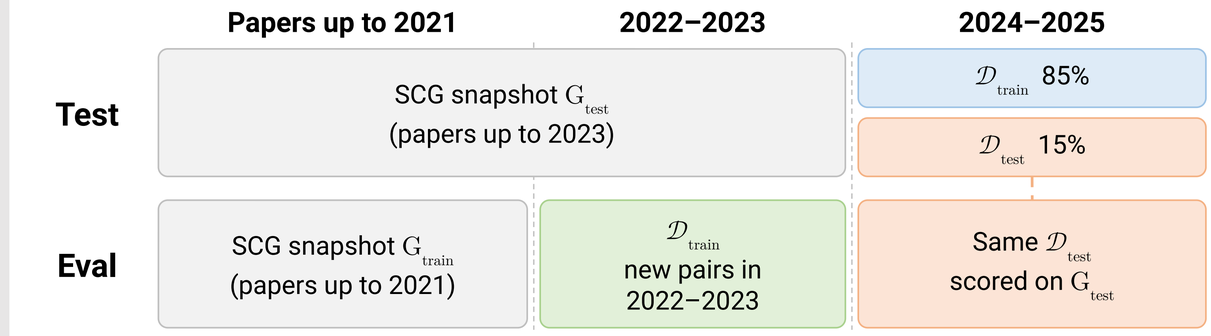



- Prior work relies on **keyword-level concepts and 141 handcrafted features**, or ignores concept semantics
- We construct new SCG benchmarks from CIKM, KDD, and SIGIR papers with **fine-grained, LLM-extracted concepts**

2.9K–6.5K concepts and 27K–75K edges per snapshot

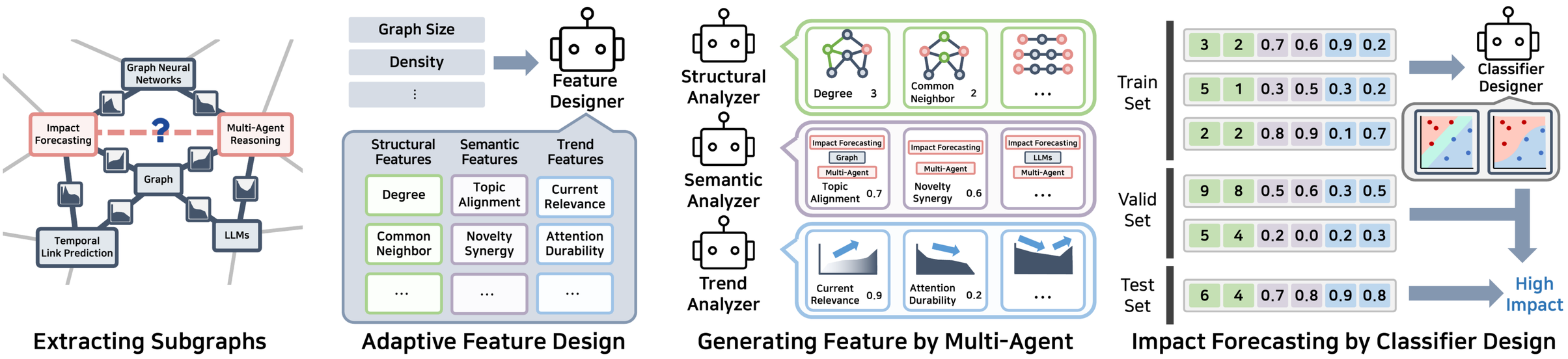
## Task and Prediction Settings

- Predict whether two **currently unconnected** research concepts will form a **high-impact connection** in the target period
  - Positive: newly appears with accumulated citations ≥ IR = 5; hard negative: newly appears but stays below IR




- **Eval** trains earlier, so **51.00%, 28.00%, and 41.96%** of test concepts are unseen in training (CIKM, KDD, SIGIR)

## MiroGraph: Multi-Agent Reasoning over Scientific Concept Graphs




## 1. Adaptive Feature Design

- **Why**: useful features **differ across venues and snapshots**, so a fixed set of handcrafted features cannot fit every SCG
- **Feature designer** (GPT-5.4, once per dataset) sees only **SCG statistics**
  - e.g., size, density, degree distribution, …
- Outputs a **feature schema**: structural, semantic, and trend features, each with a description, direction, and value range
  - Perspectives are disentangled, with no labels or future information

## 2. Multi-Agent Feature Computing

- Input: **enclosing subgraph** of (u, v) from induced paths of length ≤ 4, as a JSON edge list with yearly citations
- **Structural analyzer** (GPT-5.4) **writes NetworkX code once**, validated on a toy graph and run for every pair
  - e.g., degree, common neighbors, shortest path length
- **Semantic analyzer** (GPT-4o) scores how research themes relate, in [0, 1]
  - e.g., topic alignment, novelty synergy
- **Trend analyzer** (GPT-4o) scores **yearly citation dynamics**, in [0, 1]
  - e.g., current relevance, attention durability

## 3. Impact Forecasting by Classifier Design

- Classifier = a complete **fit_and_predict** program: features in, 0/1 labels out
- **Classifier designer** (GPT-5.4) writes **10 candidates** varying preprocessing, model, hyperparameters, and threshold, to maximize **macro-F1**
- Each candidate is fit on a 70 : 15 train/validation split; the **best on validation** is retrained on all training pairs
  - Test labels are never used to generate or select classifiers

## Preventing Knowledge Leakage

- Agents that **never see concept names** (feature designer, structural analyzer, classifier designer) use GPT-5.4
- Agents that **see concept names** use **GPT-4o** (semantic and trend analyzers), whose cutoff precedes the target period




- Snapshots keep only citations received up to the snapshot year

## Experimental Results (Macro F1)

- MiroGraph achieves **the best Macro F1 in all six settings**
- Strong Test baselines **degrade in Eval**, and feeding subgraphs directly to LLMs **lowers performance**

| Method | SCG-CIKM | | SCG-KDD | | SCG-SIGIR | |
|---|---|---|---|---|---|---|
| | Test | Eval | Test | Eval | Test | Eval |
| Best non-LLM baseline | 0.7292 | 0.6264 | 0.6030 | 0.5811 | 0.6143 | 0.6118 |
| GPT-4o | 0.5829 | 0.5771 | 0.5997 | 0.6009 | 0.5751 | 0.5758 |
| GPT-4o w/ subgraph | 0.4408 | 0.4198 | 0.4528 | 0.4455 | 0.4901 | 0.4745 |
| **MiroGraph** | **0.7501*** | **0.6435*** | **0.6878*** | **0.6334*** | **0.6818*** | **0.6157** |

*: significant over the best baseline ($p < 0.05$). Best non-LLM baseline: TGN/DyGFormer (CIKM), Impact4Cast (KDD), Impact4Cast/GraphMixer (SIGIR)

- **Ablation** (SCG-CIKM, Eval): 0.6435 vs. 0.5355 without the feature designer and 0.5731 with an MLP classifier

## Conclusion

- **SCG benchmarks**, including a setting with **unseen concepts**
- **MiroGraph**: agents **design and compute features and build classifiers**
- **Best performance in all evaluation settings**

† Corresponding Author (jjwhang@kaist.ac.kr) | GitHub: https://github.com/bdi-lab/MiroGraph | Lab Homepage: https://bdi-lab.kaist.ac.kr