# 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†

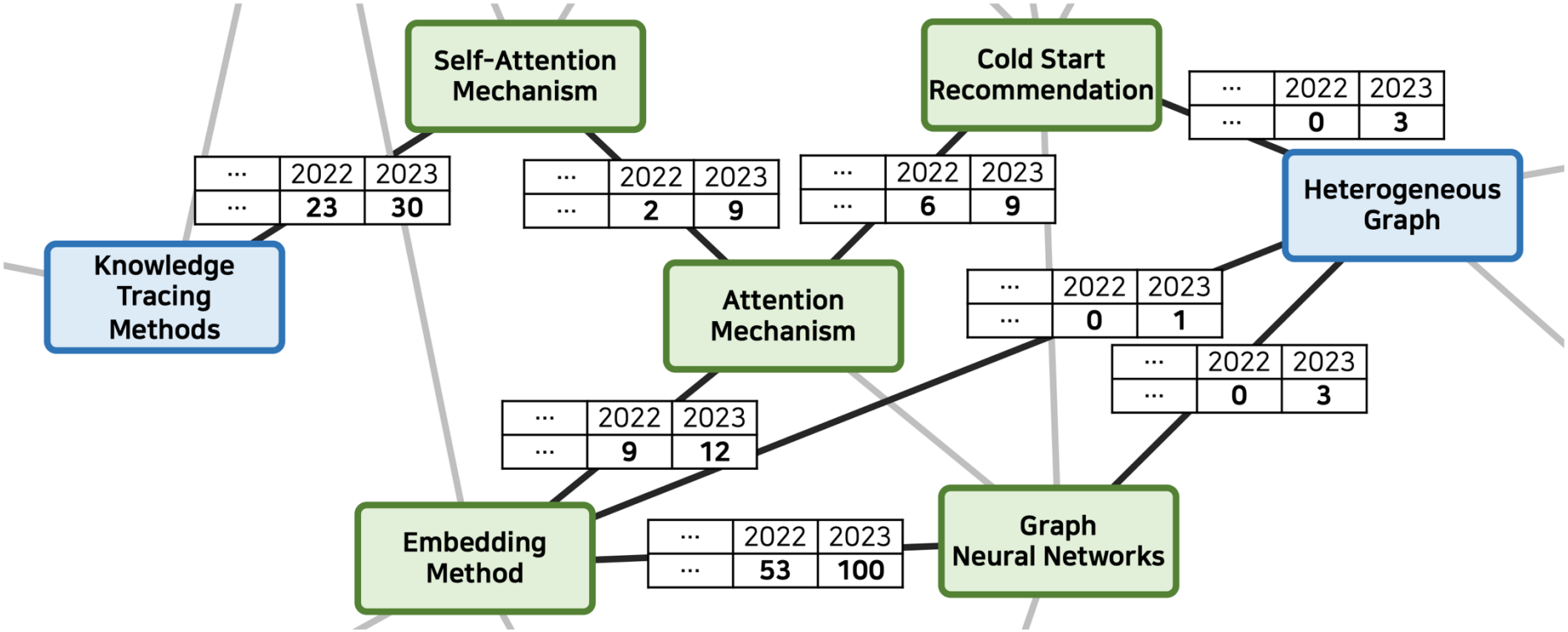
KAIST * Equal Contribution, † Corresponding Author

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

KAIST BDILab BIG DATA INTELLIGENCE

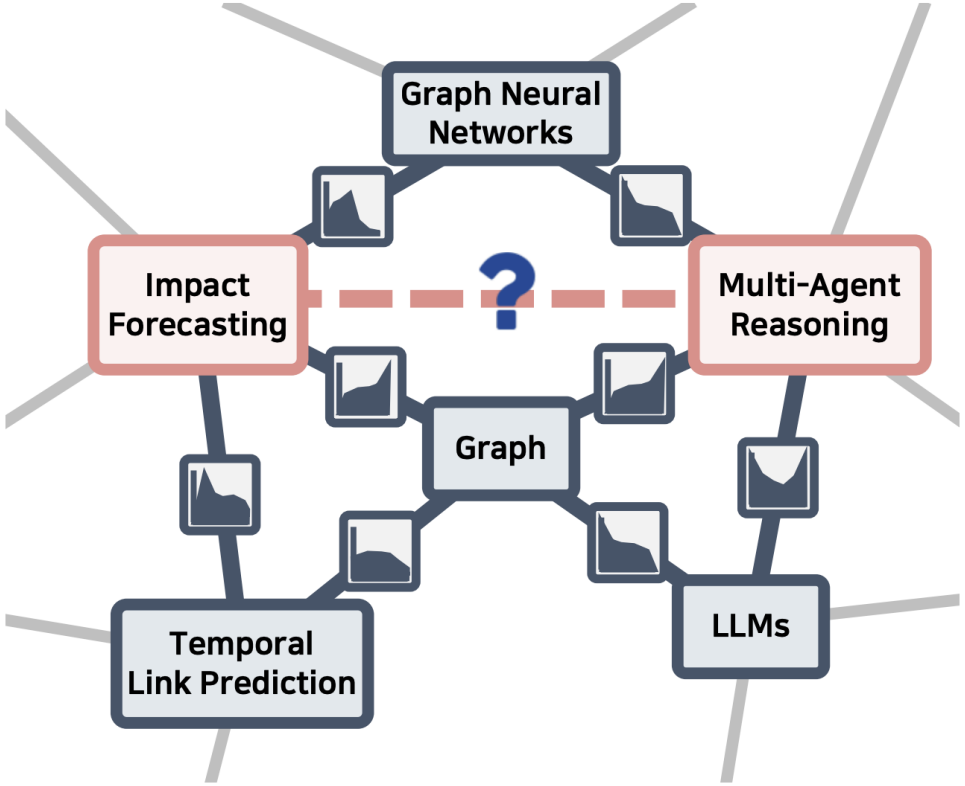
# 01 Scientific Concept Graphs (SCGs)

- Influential research often arises from **combinations of distinct research concepts**
- An SCG represents **research concepts** as nodes and their **co-occurrence in papers** as edges
  - Each edge has yearly citation counts as attributes



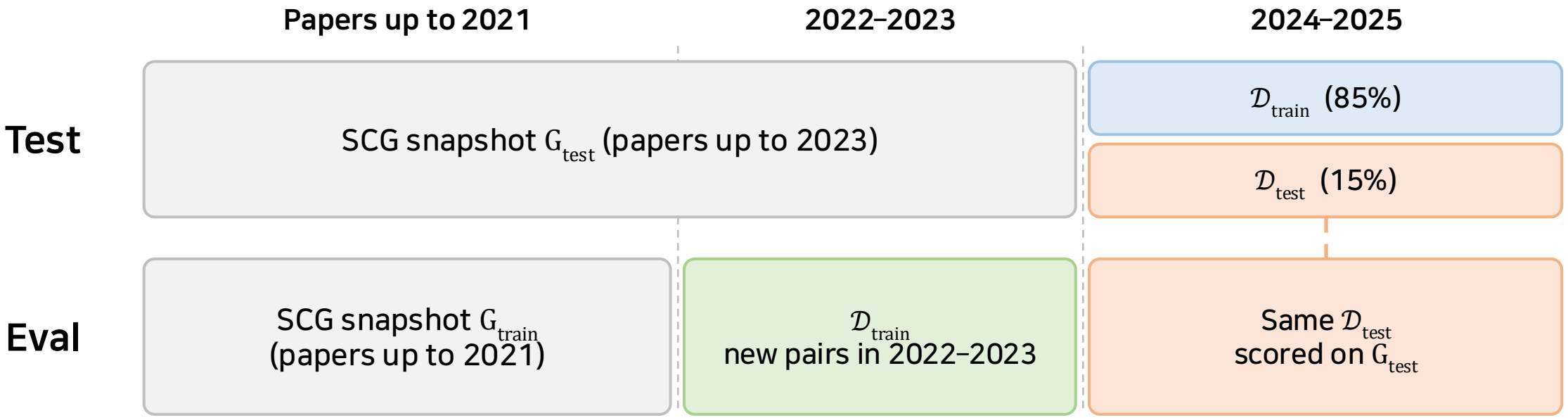
# 01 High-Impact Concept Pair Prediction

- Predict whether a **currently unconnected** concept pair will later form a **high-impact connection**
  - Positive: the pair newly appears in papers whose accumulated citations reach a threshold IR
  - IR (Impact Range): the citation threshold, set to 5 (about the top 15% of newly formed pairs)



# 01 Prediction Settings

- **Test**: snapshot up to 2023, predicting pairs that emerge in 2024–2025
- **Eval**: trained on an earlier snapshot up to 2021, evaluated on the same test set
  - Requires generalization to **concepts unseen during training**




| | SCG-CIKM | SCG-KDD | SCG-SIGIR |
|---|---|---|---|
| Unseen concepts in test set (Eval) | 51.00% | 28.00% | 41.96% |

# 01 Limitations of Existing Methods

- Impact4Cast
  - Depends on **keyword-level concept extraction** and manual processing
  - Applies a simple neural network to **141 handcrafted features**
- Concept co-occurrence prediction on SCGs
  - Does not use citation counts, so it **cannot predict high impact**
- Temporal link prediction
  - **Cannot consider the semantic meanings** of concepts

# 01 Goals of This Work

- Forecast high-impact concept pairs that **generalize to concepts unseen during training**
- Address each limitation of existing methods with a dedicated design choice

| Limitation | Goal | Our approach |
|---|---|---|
| Keyword-level concepts, manual processing | Scalable SCGs with fine-grained concepts | **Automated SCG construction** |
| 141 fixed handcrafted features | Features adapted to each SCG | **Feature designer agent** |
| No citations or no concept semantics | Combine structural, semantic, trend signals | **Three analyzer agents** |
| Fixed classifier architecture | Classifier adapted to each dataset | **Classifier designer agent** |

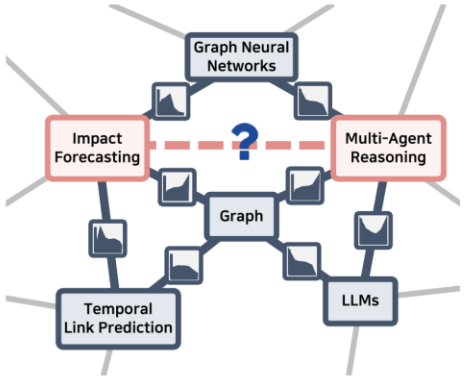
# 02 Construction of SCG Datasets

- An automated pipeline builds SCGs with **fine-grained concepts** from CIKM, KDD, and SIGIR papers
  - GPT-4o-mini extracts a **five-level concept hierarchy**, followed by LLM-based canonicalization
  - Citation counts are collected from OpenAlex

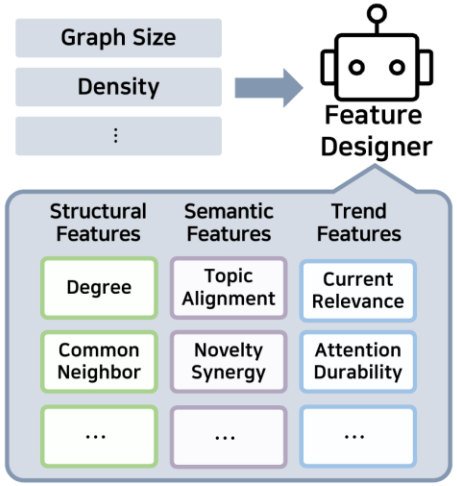
| Venue | Setting | # Concepts | # Edges | # Papers | # Train pairs | # Test pairs |
|---|---|---|---|---|---|---|
| SCG-CIKM | Test | 6,544 | 74,887 | 2,254 | 1,916 | 338 |
| | Eval | 6,059 | 61,862 | 1,948 | 1,610 | 338 |
| SCG-KDD | Test | 3,833 | 41,685 | 3,196 | 2,716 | 480 |
| | Eval | 3,529 | 34,561 | 2,490 | 2,010 | 480 |
| SCG-SIGIR | Test | 3,188 | 34,445 | 2,734 | 2,324 | 410 |
| | Eval | 2,925 | 27,465 | 2,618 | 2,208 | 410 |

Concepts and edges: nodes and edges of each snapshot SCG
Papers: papers used to build the snapshot
Train / test pairs: labeled concept pairs used for training and testing in each setting

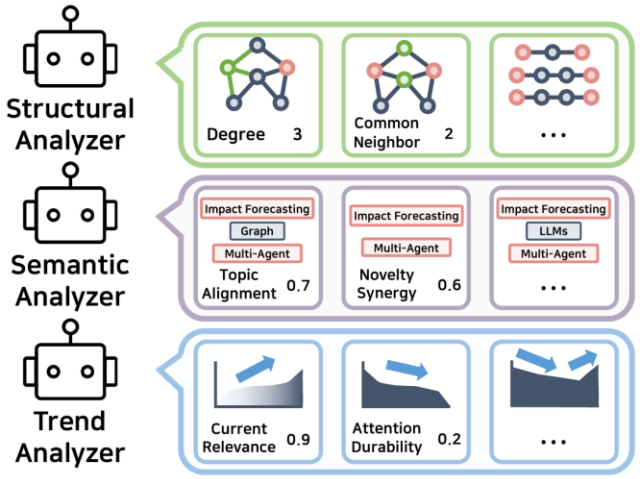
# 03 MiroGraph: Multi-Agent Reasoning with Scientific Concept Graphs



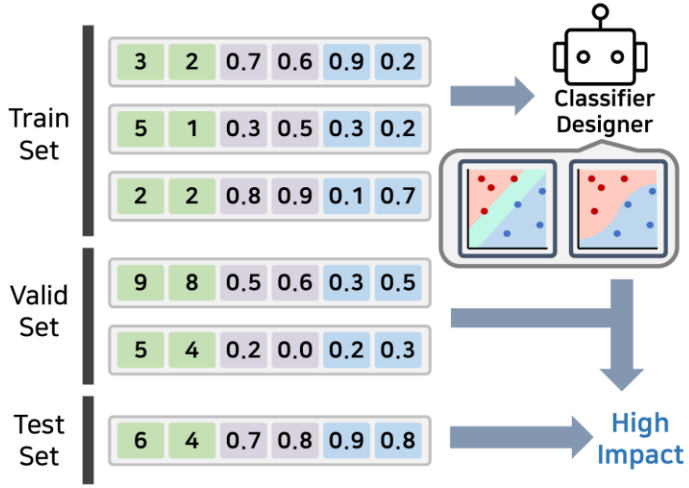

Extracting Subgraphs




Adaptive Feature Design



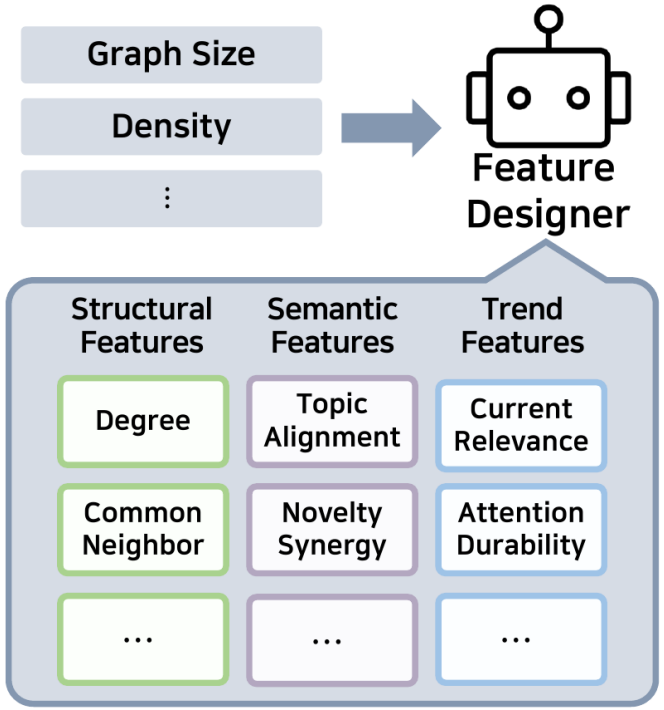

Generating Feature by Multi-Agent




Impact Forecasting by Classifier Design

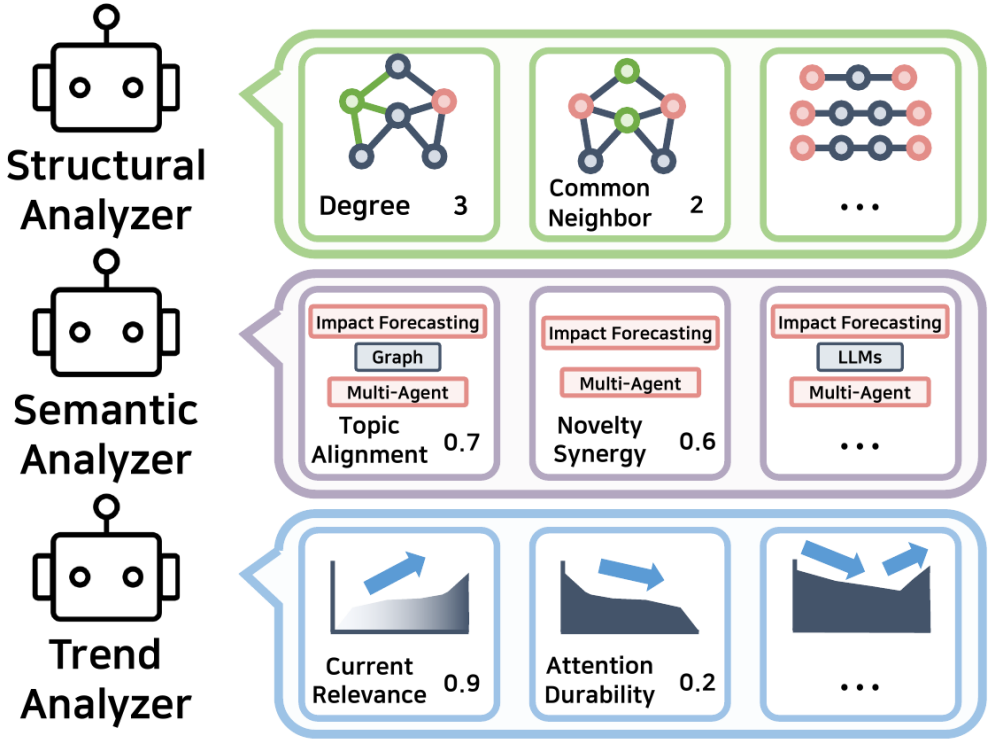
# 03 Adaptive Feature Design

- A **feature designer agent** (GPT-5.4) designs a compact feature schema for each dataset
  - Input: **SCG summary statistics** such as graph size, density, degree distribution, and connected components
  - Output: structural, semantic, and trend features, each with a description, direction, and value range
  - The three perspectives are kept **mutually disentangled**, and labels or future information are not allowed



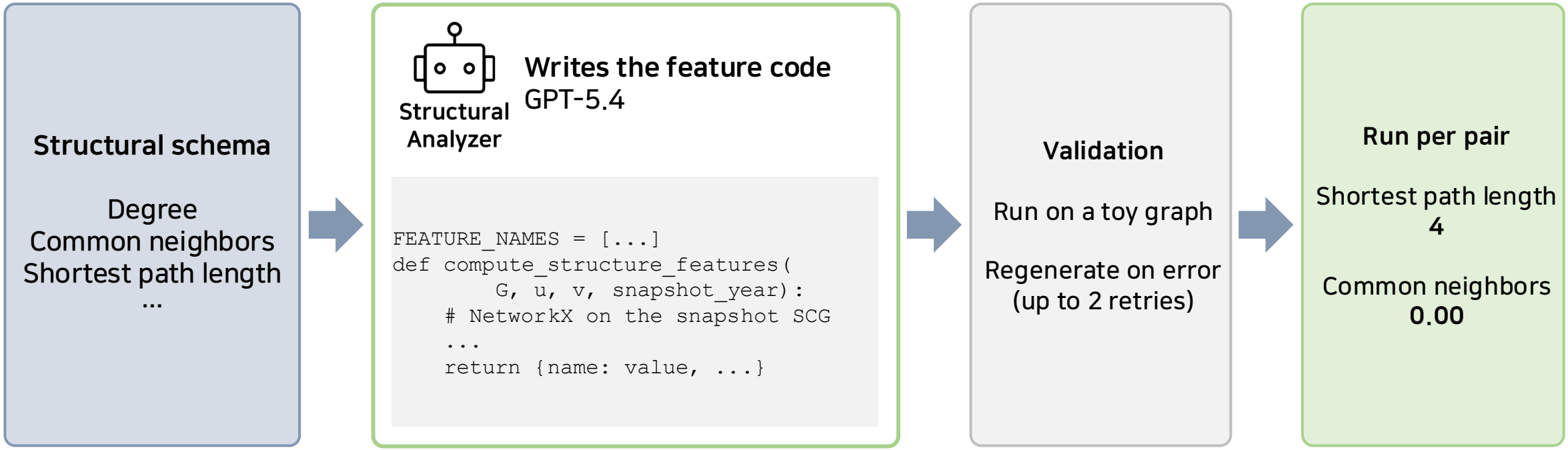
# 03 Multi-Agent Feature Computing

- For each **concept pair** (u, v), extract an **enclosing subgraph** from induced paths of length ≤ 4
  - Serialized as a JSON edge list with yearly citation counts on each edge
- Three analyzers compute the features in the **feature schema** and concatenate them



# 03 Structural Analyzer

- Computes **topological features** of the query pair on the SCG snapshot
  - e.g., degree, common neighbors, shortest path length, connectivity of intermediate concepts
- GPT-5.4 **writes executable NetworkX code once per dataset**, which is then run for every pair

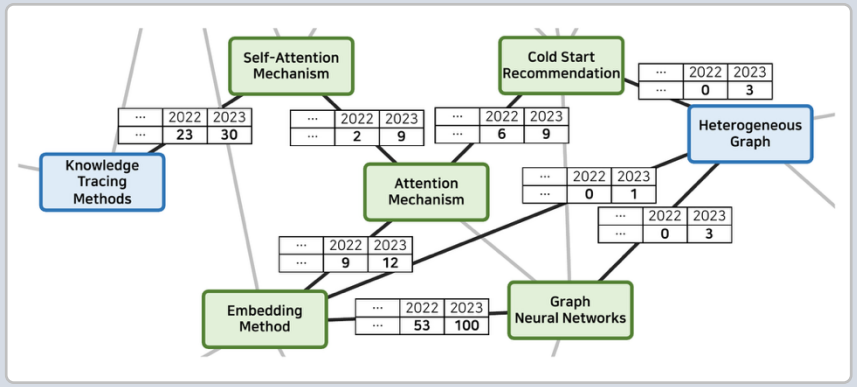



Output values: Knowledge Tracing Methods × Heterogeneous Graph on SCG-CIKM (Test)

# 03 Semantic Analyzer

- Scores **semantic features** in [0, 1]: how the research themes of the concepts relate
  - e.g., topic alignment, novelty synergy, cross-domain fit
- GPT-4o reads the **concept names and their local context** in the enclosing subgraph

**Input**
Names of concept X and concept Y
Enclosing subgraph as a JSON edge list
Semantic feature definitions from the schema




**Semantic Analyzer**

**GPT-4o**
temperature 0
JSON output in [0, 1]

**Output (example)**

Novelty synergy
**0.75**

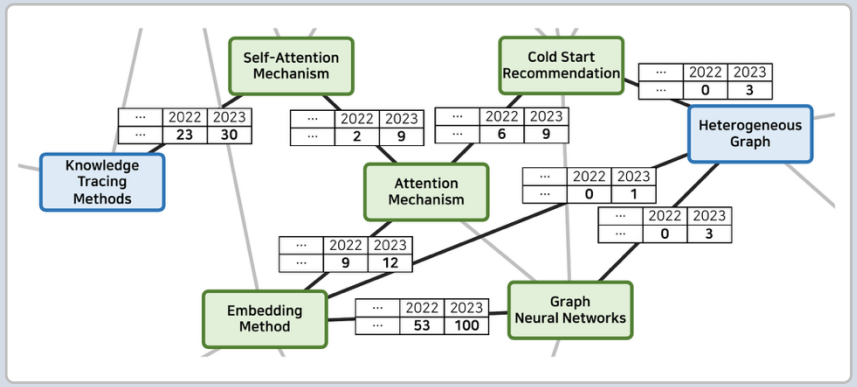
Cross-domain fit
**0.70**

Example: Knowledge Tracing Methods × Heterogeneous Graph on SCG-CIKM (Test)

# 03 Trend Analyzer

- Scores **trend features** in [0, 1]: citation-based temporal signals
  - e.g., current relevance, attention durability
- GPT-4o reads **year-by-year citations and recency patterns** on the subgraph edges

**Input**
Concept pair and enclosing subgraph
Yearly citations on each edge (small tables)
Trend feature definitions from the schema




**Trend Analyzer**

**GPT-4o**
temperature 0
JSON output in [0, 1]

**Output (example)**

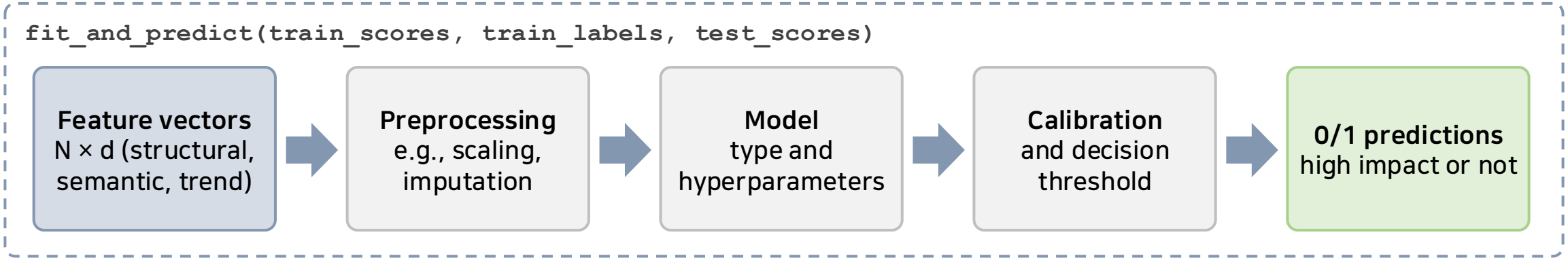
Current relevance
**0.75**

Attention durability
**0.73**

Example: Knowledge Tracing Methods × Heterogeneous Graph on SCG-CIKM (Test)

# 03 What the Classifier Designer Builds

- A classifier is a complete **fit_and_predict** program: feature vectors in, 0/1 labels out
- GPT-5.4 writes **10 diverse candidates** in one call (preprocessing, model, regularization, threshold)
- Objective: **maximize macro-F1 and accuracy** (not AUC), tuning only on training data




| Designed by the agent | Fixed by MiroGraph |
|---|---|
| Preprocessing (e.g., scaling, imputation) | Input: structural, semantic, and trend features |
| Model type and hyperparameters | 70:15 train and validation split of the training set |
| Calibration and decision threshold | Selection by validation macro-F1, then retraining |

# 04 Generalization and LLM Baselines

- Strong baselines in Test can **drop sharply in Eval**, e.g., TGN on SCG-CIKM
- Feeding the subgraph directly to LLMs **lowers performance**, so gains come from the framework

| Method | SCG-CIKM | | SCG-KDD | | SCG-SIGIR | |
|---|---|---|---|---|---|---|
| | Test | Eval | Test | Eval | Test | Eval |
| Impact4Cast | 0.6633 | 0.5865 | 0.6030 | 0.5811 | 0.6143 | 0.6106 |
| TGAT | 0.6439 | 0.5465 | 0.5645 | 0.5662 | 0.5627 | 0.5527 |
| DyGFormer | 0.6785 | 0.6264 | 0.5764 | 0.5736 | 0.6065 | 0.5951 |
| CAWN | 0.6311 | 0.5699 | 0.5642 | 0.5727 | 0.5747 | 0.6001 |
| GraphMixer | 0.5844 | 0.5162 | 0.5724 | 0.5194 | 0.5566 | 0.6118 |
| TGN | 0.7292 | 0.5312 | 0.5319 | 0.5052 | 0.5344 | 0.5649 |
| JODIE | 0.5966 | 0.5272 | 0.5256 | 0.4888 | 0.5423 | 0.5178 |
| DyRep | 0.5735 | 0.5285 | 0.5236 | 0.4877 | 0.5545 | 0.4983 |
| GPT-4o-mini | 0.5947 | 0.5976 | 0.5761 | 0.5780 | 0.5930 | 0.5995 |
| GPT-4o | 0.5829 | 0.5771 | 0.5997 | 0.6009 | 0.5751 | 0.5758 |
| GPT-4o-mini w/ subgraph | 0.3693 | 0.3718 | 0.3982 | 0.4063 | 0.4479 | 0.4333 |
| 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** |

Macro F1, mean of three runs. Bold: best, underline: second best, *: significant over the best baseline ($p < 0.05$).

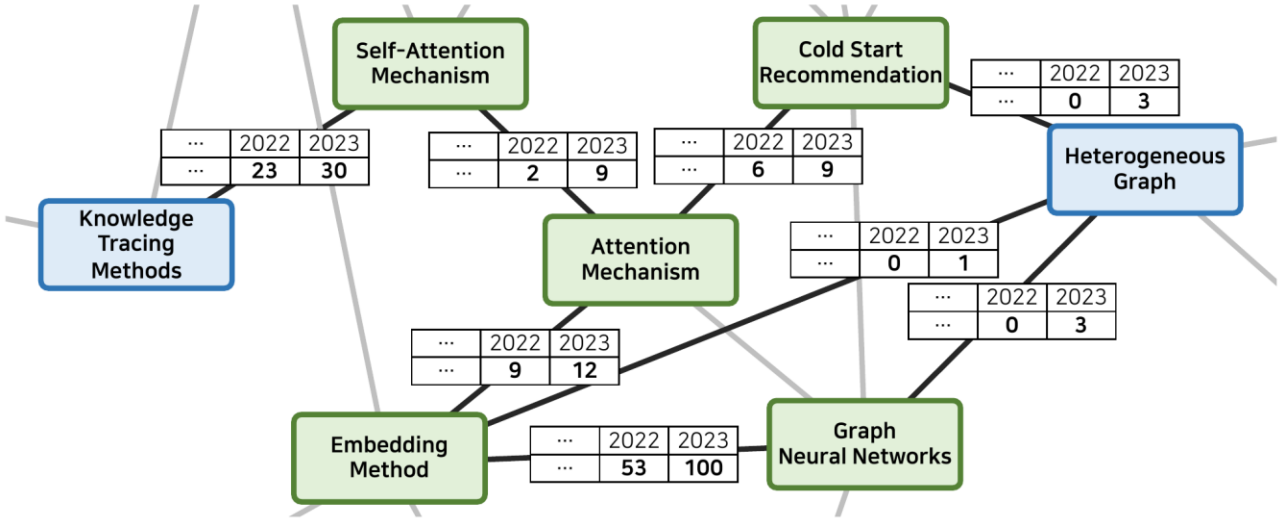
# 04 Ablation Study

- All variants underperform MiroGraph
  - **single score**: no feature designer, **MLP**: an MLP replaces the classifier designer
- Drops of single score and MLP are much larger in Eval than in Test
  - Automated feature and classifier design matter for generalization

| Method | SCG-CIKM (Test) | SCG-CIKM (Eval) |
|---|---|---|
| MiroGraph | **0.7501** | **0.6435** |
| w/o structure | 0.5844 | 0.5805 |
| w/o semantic | 0.7297 | 0.6192 |
| w/o trend | 0.7373 | 0.6065 |
| single score | 0.6939 | 0.5355 |
| MLP | 0.6928 | 0.5731 |

# 04 Case Study

- Enclosing subgraph for **Knowledge Tracing Methods** and **Heterogeneous Graph**
  - Linked only through distant bridges, but a **novel cross-area combination**




| Perspective | Feature | Value |
|---|---|---|
| Structural | Shortest path length | 4 |
| | Common neighbors | 0.00 |
| Semantic | Novelty synergy | 0.75 |
| | Cross-domain fit | 0.70 |
| Trend | Current relevance | 0.75 |
| | Attention durability | 0.73 |

# 04 Conclusion

- Construct **SCG benchmark datasets** from CIKM, KDD, and SIGIR with fine-grained concepts
  - Including a setting with concepts unseen during training
- Propose **MiroGraph**, where LLM agents **design features, compute them, and construct classifiers**
- MiroGraph achieves **the best performance in all evaluation settings**
- Future work: extend SCGs into **richer knowledge graphs** to better generalize to unseen concepts

# Thank You!



▲ GitHub



▲ BDILab

**Our datasets and codes are available at:**
https://github.com/bdi-lab/MiroGraph

**You can find us at:**
{justinyoonwk, hminsung, jjwhang}@kaist.ac.kr
https://bdi-lab.kaist.ac.kr

KAIST BDILab
BIG DATA INTELLIGENCE

# 05 Why Let LLM Agents Design the Pipeline

- Signals for a concept pair are **dispersed across its relational and temporal context**
- Handcrafted features fix **what to measure** in advance (141 predefined features in Impact4Cast)
- SCGs **differ across venues and snapshots**, so useful features and classifiers can differ
  - MiroGraph lets agents **design the feature schema and the classifier** for each dataset

| | Fixed pipeline | MiroGraph |
|---|---|---|
| Features | Predefined handcrafted metrics | Schema designed from SCG statistics |
| Feature values | Fixed formulas | Generated code and LLM analysis of the subgraph |
| Classifier | Fixed architecture | Candidates designed and selected on validation |

- Replacing either design step hurts most in **Eval** (SCG-CIKM)
  - Single score 0.5355 and MLP 0.5731, compared with 0.6435 for MiroGraph

# 05 Preventing Knowledge Leakage

- An LLM may already know the outcome if the target period is in its training data
- We **match each agent's information access to its knowledge cutoff**

| Agent | Input | Concept names | Model |
|---|---|---|---|
| Feature designer | SCG summary statistics | No | GPT-5.4 |
| Structural analyzer | Schema and graph interface for code | No | GPT-5.4 |
| Semantic analyzer | Concept pair and enclosing subgraph | Yes | GPT-4o |
| Trend analyzer | Concept pair and yearly citations | Yes | GPT-4o |
| Classifier designer | Numeric feature columns only | No | GPT-5.4 |




- Each snapshot keeps only citations received up to its snapshot year
- GPT-4o alone reaches only **0.58–0.60 Macro F1**, far below MiroGraph