# Cross-Architecture Distillation of LLM Knowledge for Non-LLM Model Enhancement

**Jimyeung Seo (서지명)**

Supervisor: Byungkook Oh

Graph & Language Intelligence Laboratory
Department of Computer Science and Engineering
Konkuk University

2026.08.12



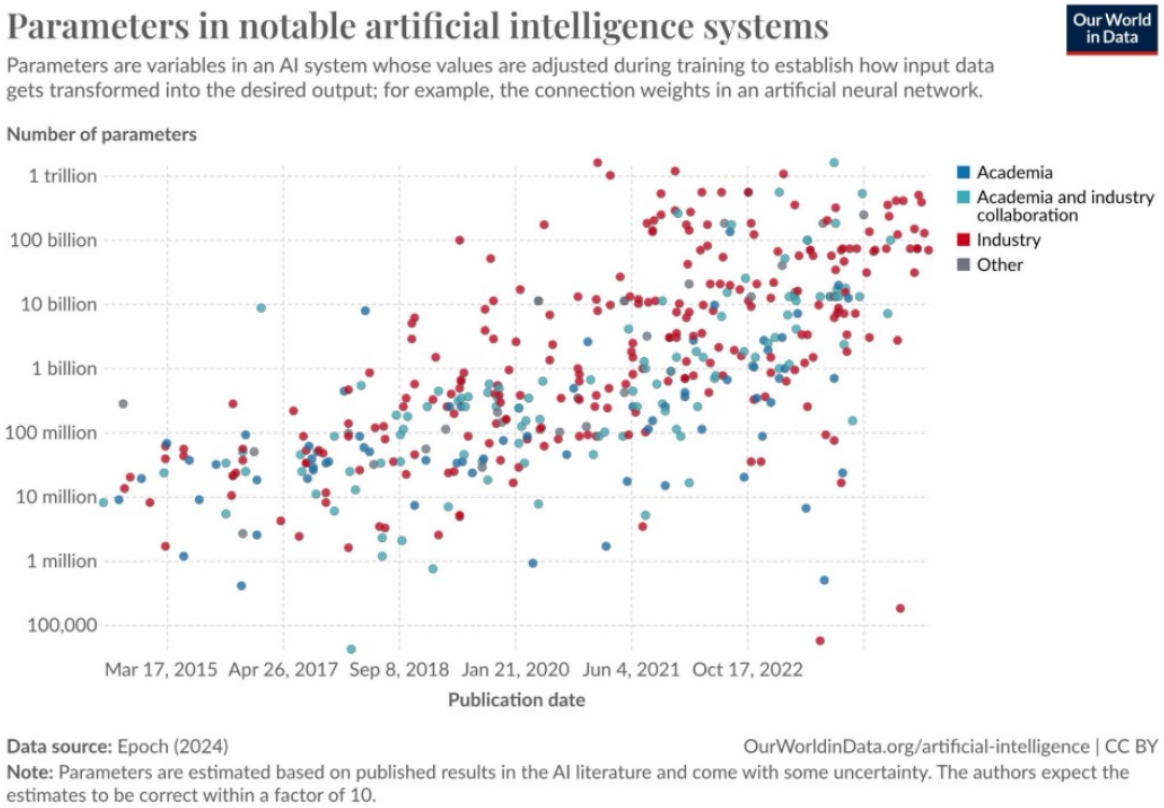
# CONTENTS



# CONTENTS

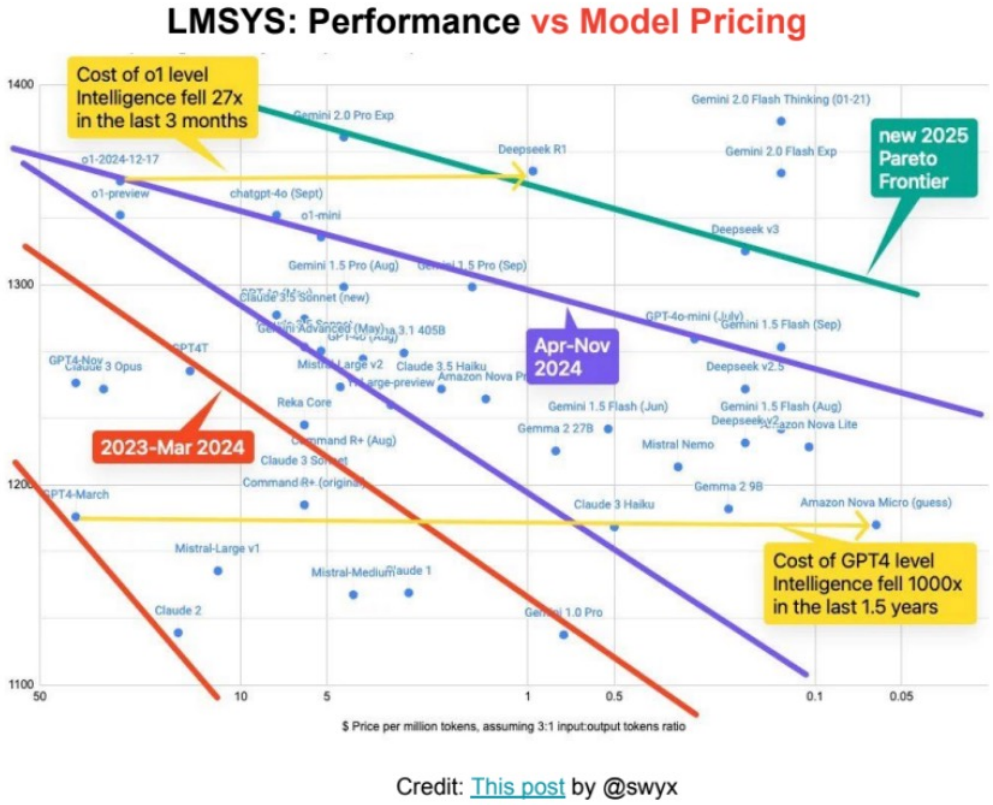
# Bigger Models Always Better?

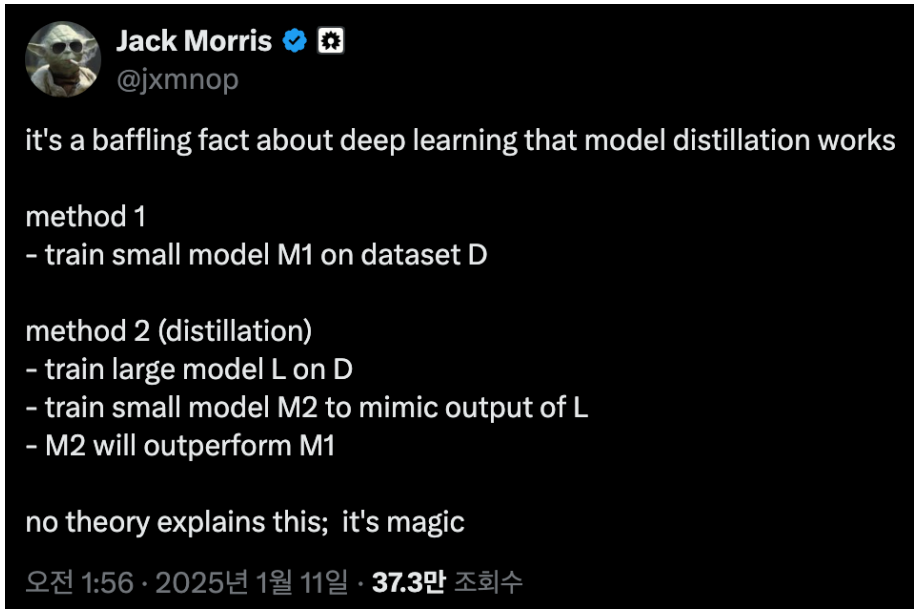
- Scaling law
  - ✓ 기존에는 Model size가 커질수록 일반적으로 높은 성능을 보임



# Bigger Models Always Better?

- Performance vs. Model Pricing
  - ✓ 그러나 최근에는 작고 싼 모델들이 크고 비싼 모델들을 따라잡고 있음
  - ✓ 강력한 큰 모델이 출시된 이후 Knowledge Distillation 기법을 이용해 작은 모델들을 학습시켰고, 이로 인해 성능 격차가 크게 줄어들었음



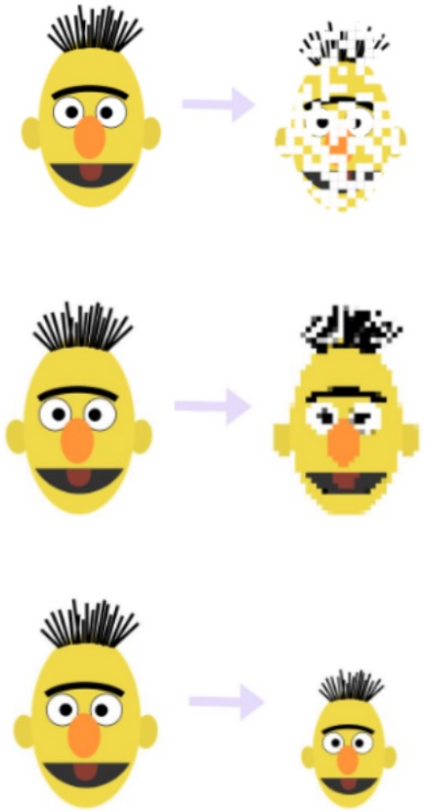
# Model Lightweight

- 높은 비용
  - ✓ 학습: 거대한 모델을 학습시키기 위해서는 많은 계산 자원이 필요함
  - ✓ 추론: 학습된 모델이 추론하는 과정에서도 많은 계산 자원이 필요하므로, 실시간으로 빠른 응답이 요구되는 서비스에 적용하기 어려움
  - ✓ 예) Gemini 1.5 Flash

- 접근성
  - ✓ 거대한 모델은 모바일이나 임베디드 기기에 적용하기 어려움
  - ✓ 예) Youtube Shorts

# Model Lightweight

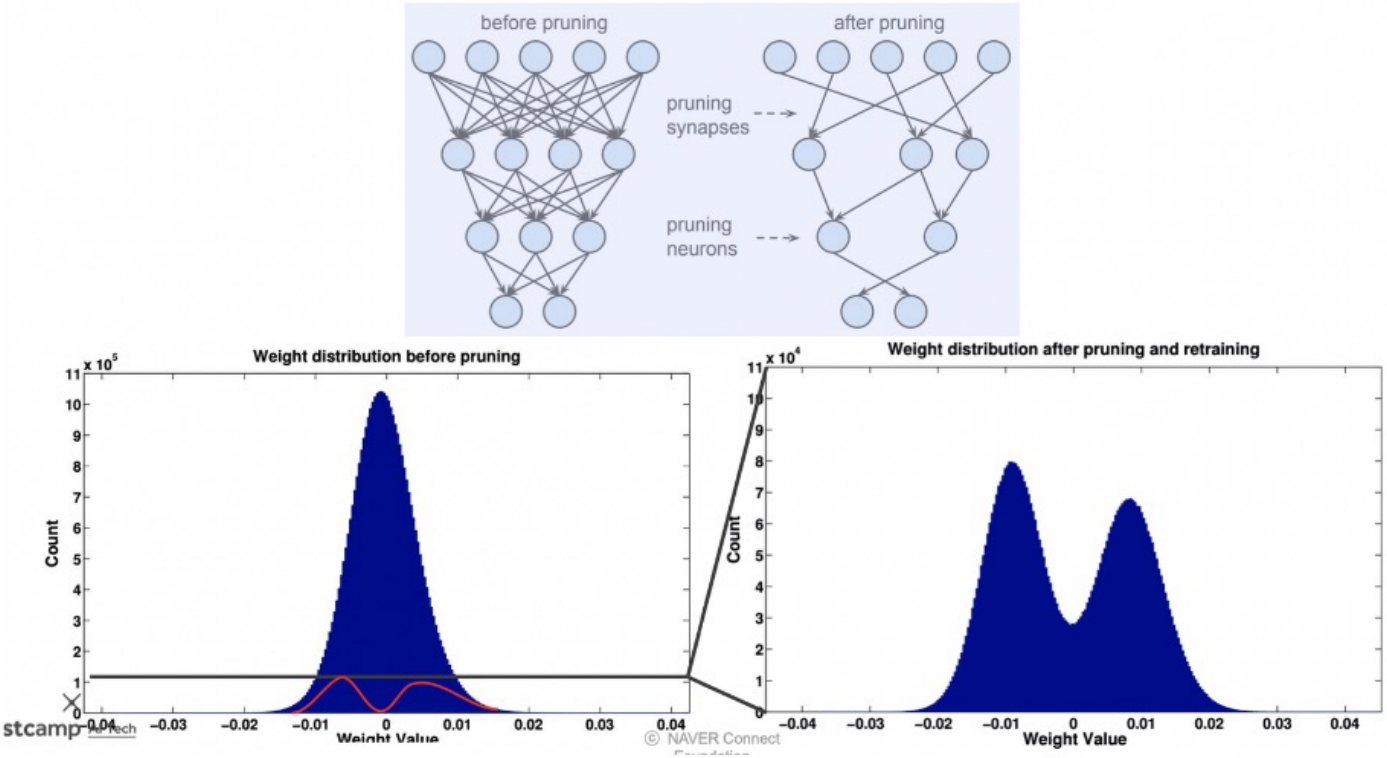
- ✓ Pruning: 딥러닝 모델의 많은 layer 중, 중요하지 않은 파라미터를 지우는 기법
- ✓ Quantization: 학습된 모델이 weight값을 저장할 때 사용하는 bit 수를 줄이는 기법
- ✓ Distillation: 거대 모델의 지식을 더 작은 모델에 전달하는 과정으로, 작은 모델이 큰 모델과 유사한 성능을 내면서도 훨씬 적은 자원으로 동작하도록 함



# About KD

## Pruning

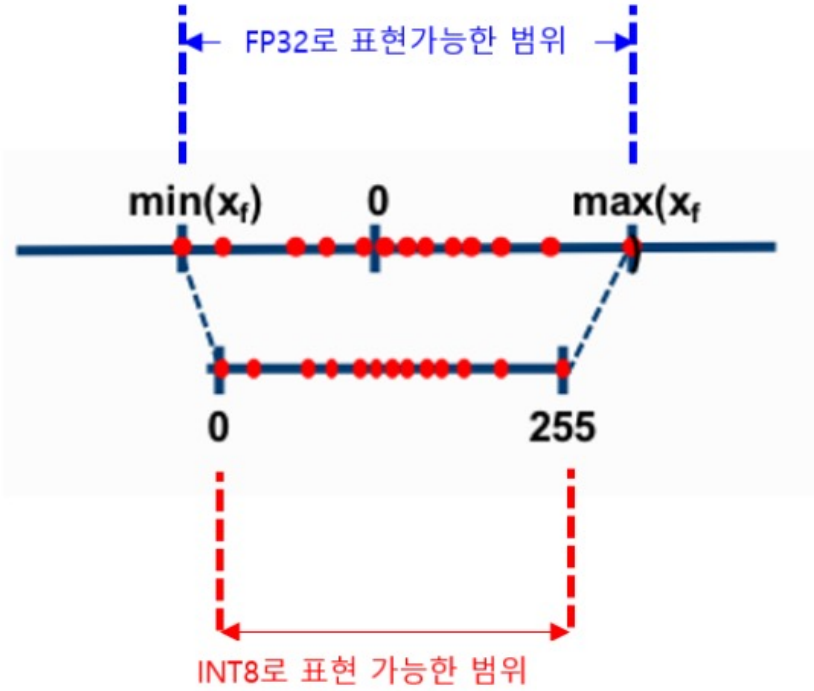
- 딥러닝 모델의 많은 layer 중, 중요하지 않은 파라미터를 지움



# About KD

## Quantization

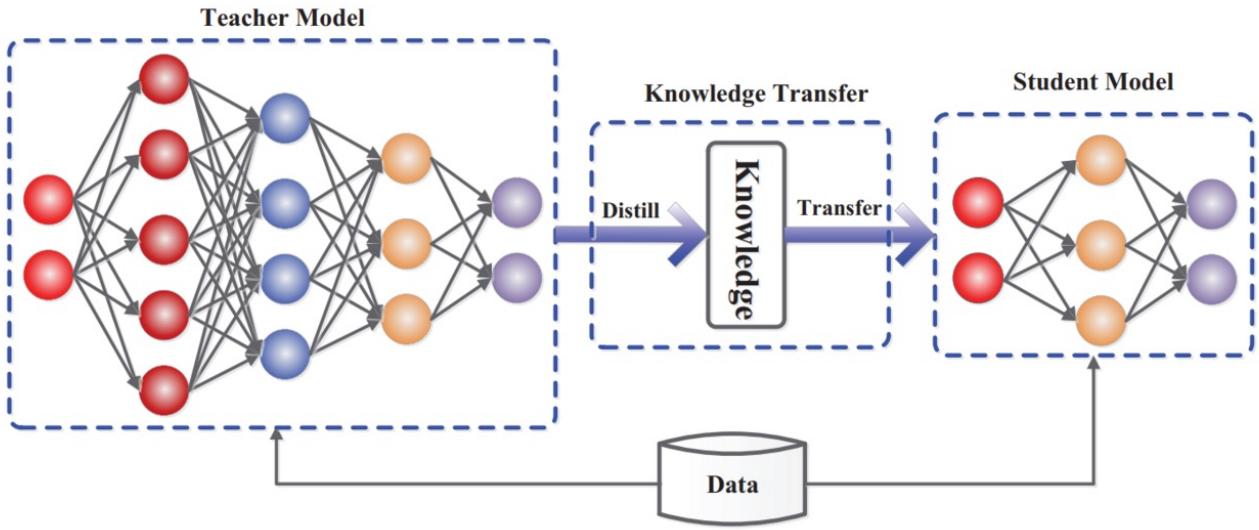
- 학습된 모델이 weight값을 저장할 때 사용하는 bit 수를 줄이는 방법



# About KD

## Distillation

- 거대 모델의 지식을 더 작은 모델에 전달하는 과정으로, 작은 모델이 큰 모델과 유사한 성능을 내면서도 훨씬 적은 자원으로 동작하도록 함



# About KD: Example

## Distillation

- Gemini 1.5 Pro -> Gemini 1.5 Flash

[1]

> *1.5 Flash excels at summarization, chat applications, image and video captioning, data extraction from long documents and tables, and more. This is because it's been trained by 1.5 Pro through a process called "distillation," where the most essential knowledge and skills from a larger model are transferred to a smaller, more efficient ···*

[1] https://blog.google/innovation-and-ai/products/google-gemini-update-flash-ai-assistant-io-2024/

# About KD: Example

## Distillation

- StyleGAN2와 이후 Imagen 같은 대형 생성모델을 teacher로 사용하고, U-Net/MobileNet 기반 student를 학습해 스마트폰에서 실시간으로 실행
  - ✓ 20개 이상의 Shorts 효과 출시
  - ✓ Pixel 8 Pro에서는 약 6ms, iPhone 13 GPU에서는 10.6ms/frame이라 보고

[1]

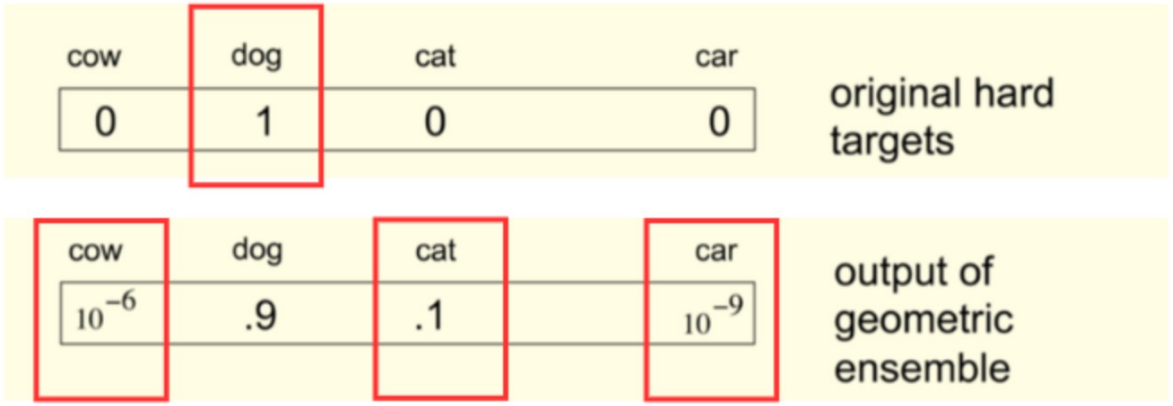
*Our approach revolves around a concept called knowledge distillation, which uses a "teacher–student" model training method. We start with a "teacher" — a large, powerful, pre-trained generative model that is an expert at creating the desired visual effect but is far too slow for real-time use. The type of teacher model varies depending on the goal. Initially, we used a custom-trained StyleGAN2 model, …*

[1] https://research.google/blog/from-massive-models-to-mobile-magic-the-tech-behind-youtube-real-time-generative-ai-effects/

# Classical KD

## Supervised KD

- Geoffrey Hinton's KD[1]
  - ✓ Image classification과 같은 task는 NN의 마지막 softmax layer를 통해 각 class의 확률값을 출력




  - ✓ 이때, target을 제외한 다른 출력값들이 모델의 지식이 될 수 있다고 생각
  - ✓ But, softmax에 의해 값이 너무 작아 모델에 반영하기 어려움

[1] Hinton, Vinyals et al., "Distilling the Knowledge in a Neural Network", NIPS'14 Deep Learning Workshop, Cited: 34296

# Classical KD

## Supervised KD

- Geoffrey Hinton's KD[1]
  - ✓ 따라서, 다음과 같이 출력값의 분포를 조금 더 soft하게 만듦
    - 기존 softmax 함수에 temperature를 넣음

| cow | dog | cat | car |
|---|---|---|---|
| $10^{-6}$ | .9 | .1 | $10^{-9}$ |

$$qi = \frac{exp(z_i)}{\sum_j exp(z_j)} \rightarrow qi = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}$$

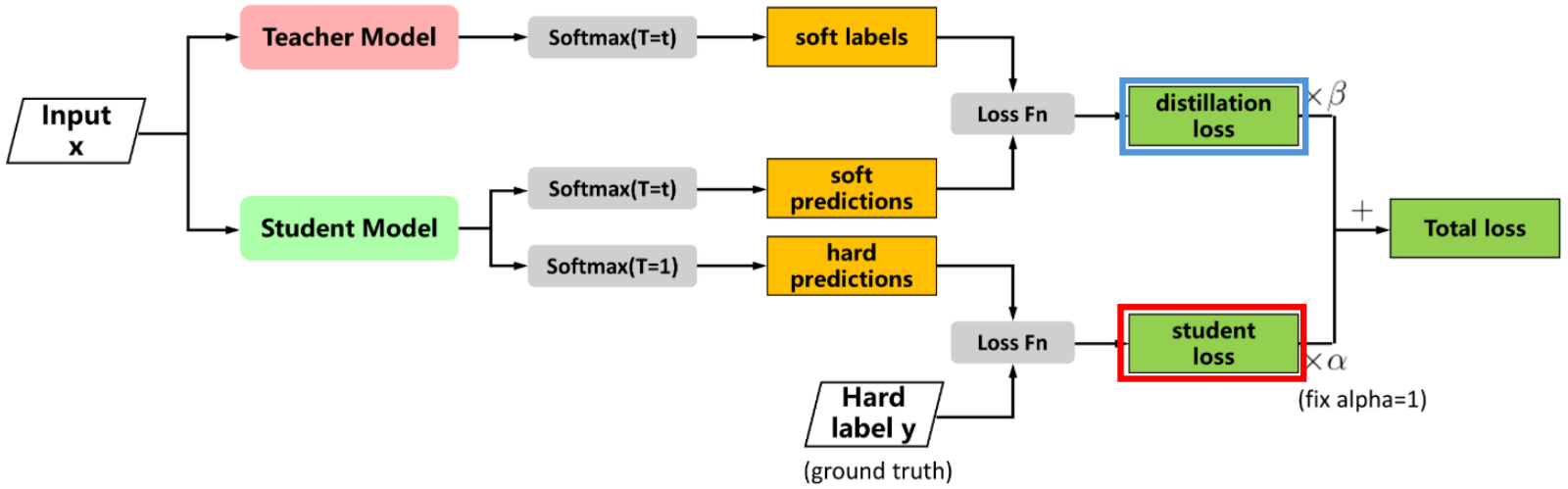
| cow | dog | cat | car |
|---|---|---|---|
| .05 | .3 | .2 | .005 |

[1] Hinton, Vinyals et al., "Distilling the Knowledge in a Neural Network", NIPS'14 Deep Learning Workshop, Cited: 34296

# Classical KD

## Supervised KD

- Geoffrey Hinton's KD[1]
  - ✓ Soft target은 결국 Teacher model의 Knowledge
  - ✓ DistillBert, Gemma 2



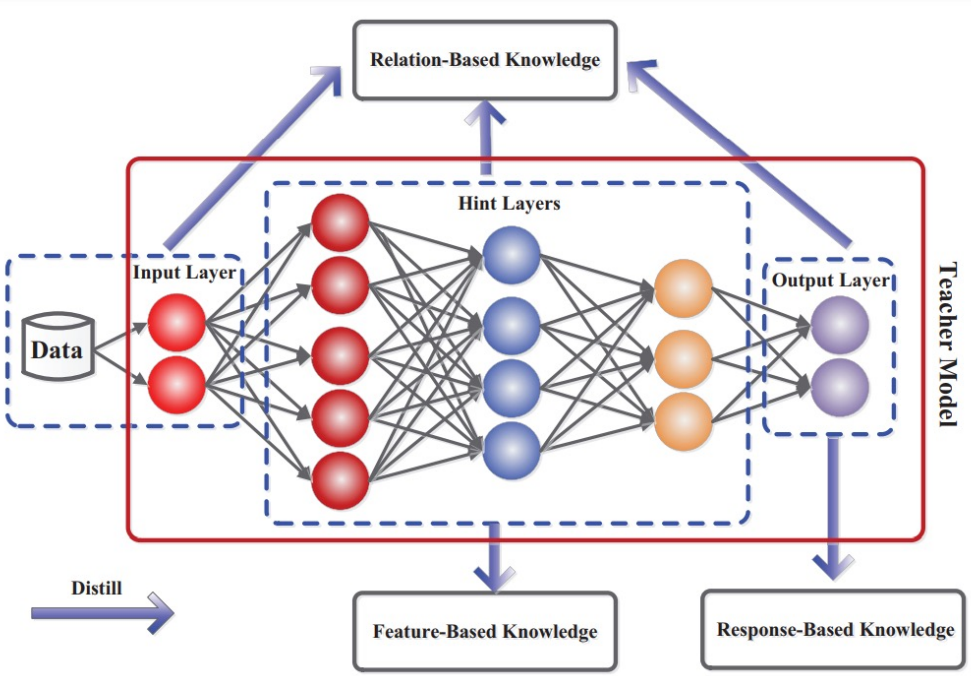

$$L = \sum_{(x,y)\in D} L_{KD}(S(x,\Theta_S,\tau), T(x,\Theta_T,\tau)) + \lambda L_{CE}(\hat{y}_S, y)$$

[1] Hinton, Vinyals et al., "Distilling the Knowledge in a Neural Network", NIPS'14 Deep Learning Workshop, Cited: 34296

# CONTENTS

# Towards LLM KD

- LLM은 Teacher로서 어떤 Knowledge를 제공할 수 있는가?
  - ✓ Black-box
    - teacher가 생성한 텍스트만 접근
  - ✓ White-box
    - Output distribution: logits, token probabilities
    - Intermediate representation: hidden states, features
    - Relation knowledge: attention, similairity, geomerty



# Towards LLM KD

LLM을 Teacher로서 어떤 Knowledge를 제공할 수 있는가?

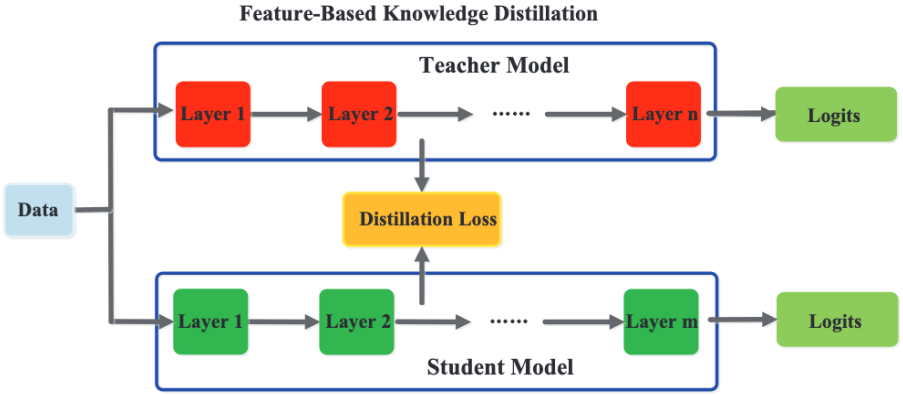



**Fig. 6** The generic feature-based knowledge distillation.

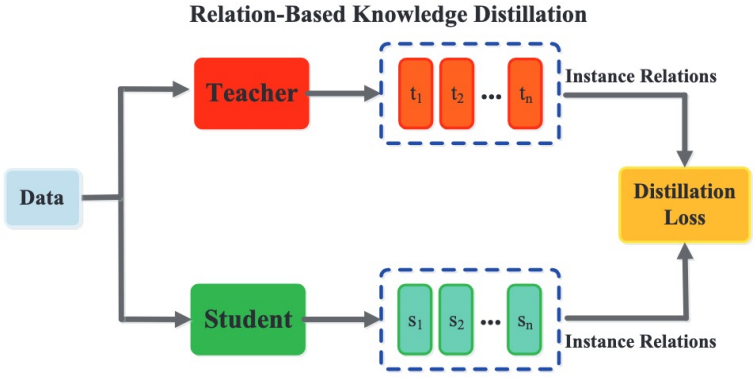
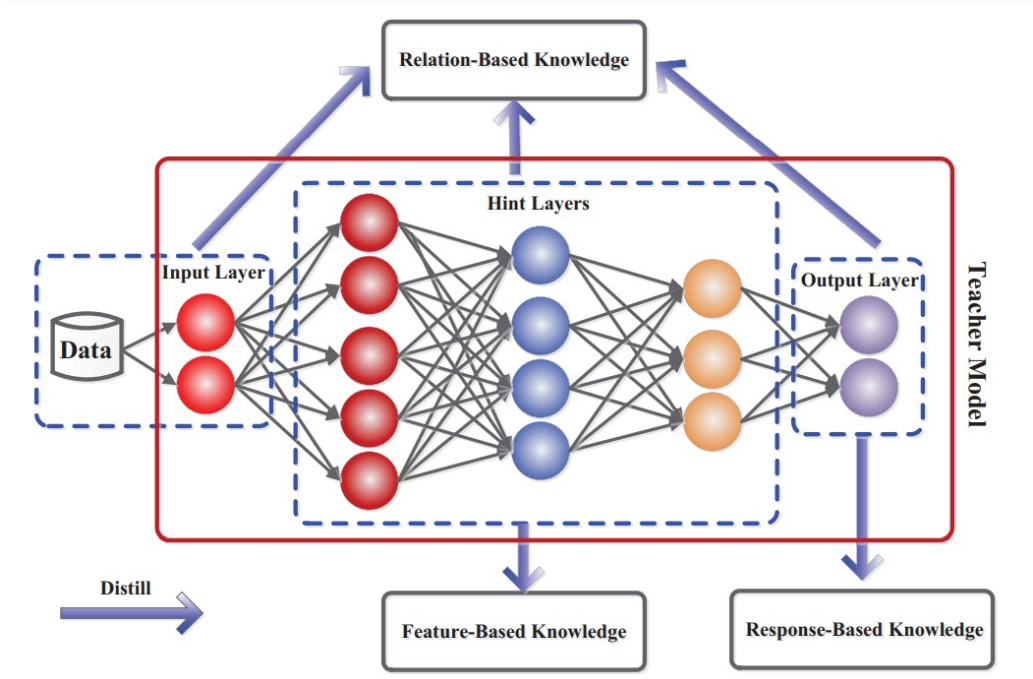


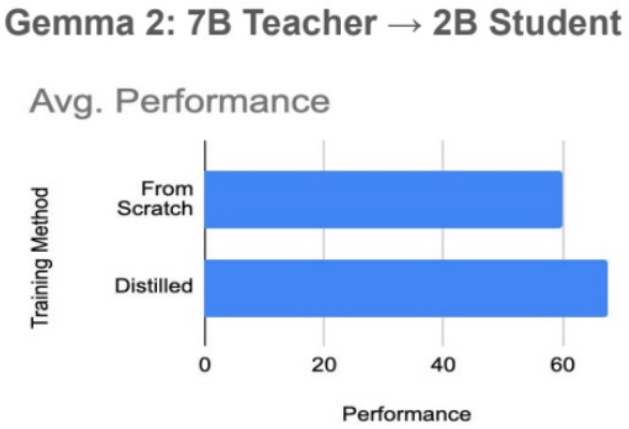

**Fig. 7** The generic instance relation-based knowledge distillation.

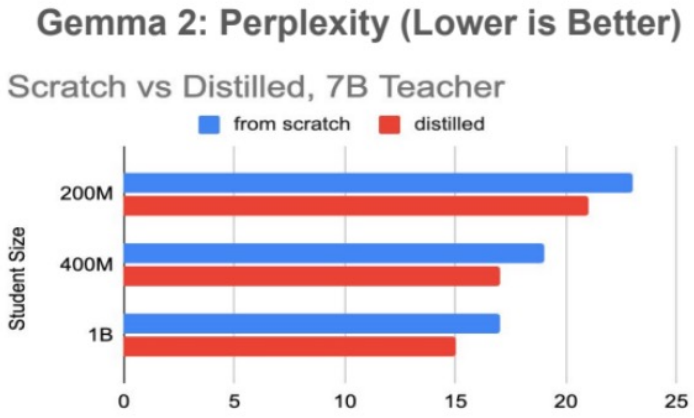


# Generative KD

## Soft Generative KD

- Teacher의 token distribution을 학습
  - ✓ DistillBert[1], Gemma 2[2]



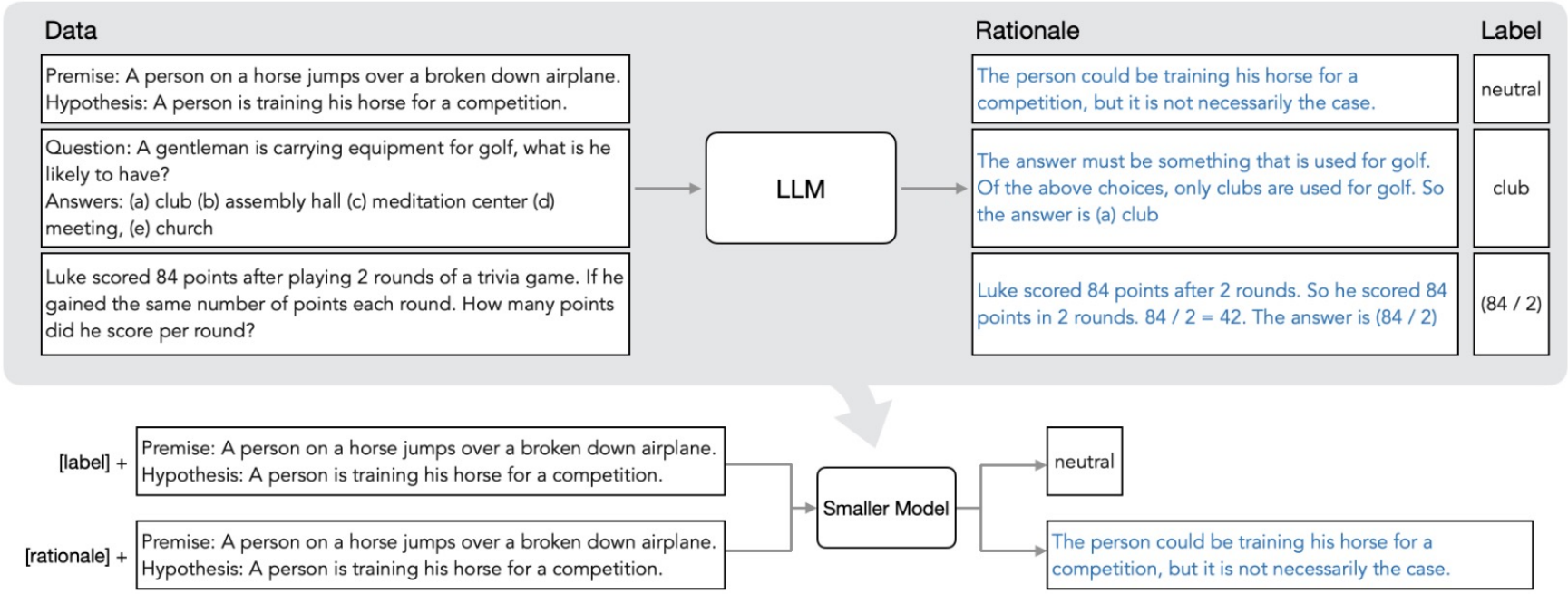




[1] https://huggingface.co/docs/transformers/en/model_doc/distilbert
[2] https://arxiv.org/abs/2408.00118

# Generative KD

## Hard / Sequence Generative KD

- Teacher가 생성한 hard sequence를 student가 학습
  - ✓ instruction-response, rationale, CoT, etc.



# Generative KD

## Reasoning KD

- Teacher의 추론 과정까지 증류
  - ✓ Counterfactual Distillation (EMNLP'24)[1]
    - 하나의 정답만 주는 대신 여러 counterfactual reasoning view를 제공
  - ✓ QCRD (EMNLP'25)[2]
    - 아무 rationale나 쓰는 게 아니라 좋은 rationale와 나쁜 rationale를 구분
  - ✓ Curriculum Distillation (EMNLP'25)[3]
    - Student 수준에 맞게 reasoning 난이도를 조절함

[1] Feng et al., "Teaching Small Language Models Reasoning through Counterfactual Distillation", EMNLP'24, Cited: 33
[2] Wang et al., "QCRD: Quality-guided Contrastive Rationale Distillation for Large Language Models", EMNLP'25, Cited: 15
[3] Jiang et al., "Teach Small Models to Reason by Curriculum Distillation", EMNLP'25, Cited: 5



# CONTENTS

# Linguistic Graph KD (AAAI'25)

**Large Language Model Meets Graph Neural Network in Knowledge Distillation**

**Shengxiang Hu[1], Guobing Zou[1*], Song Yang[1], Shiyi Lin[1], Yanglan Gan[2], Bofeng Zhang[3], Yixin Chen[4]**

[1]School of Computer Engineering and Science, Shanghai University, Shanghai, China
[2]School of Computer Science and Technology, Donghua University, Shanghai, China
[3]School of Computer and Information Engineering, Shanghai Polytechnic University, Shanghai, China
[4]Department of Computer Science and Engineering, Washington University in St. Louis, St. Louis, MO 63130 USA
{shengxianghu, gbzou, yangsong, sylin}@shu.edu.cn, ylgan@dhu.edu.cn, bfzhang@sspu.edu.cn, chen@cse.wustl.edu

Cited: 42

- Motivation: "LLM의 풍부한 언어 지식을 가벼운 GNN에 옮길 수 없을까?"
  - ✓ 기존 GNN은 graph structure 처리에는 효율적
    - 하지만 복잡한 node text의 semantics를 충분히 이해하지 못함
  - ✓ LLM은 semantics와 entity relationship을 잘 이해함
    - 하지만 수십억 개 parameter와 긴 inference latency 때문에 graph task에 직접 배포하기 어려움

# Linguistic Graph KD (AAAI'25)

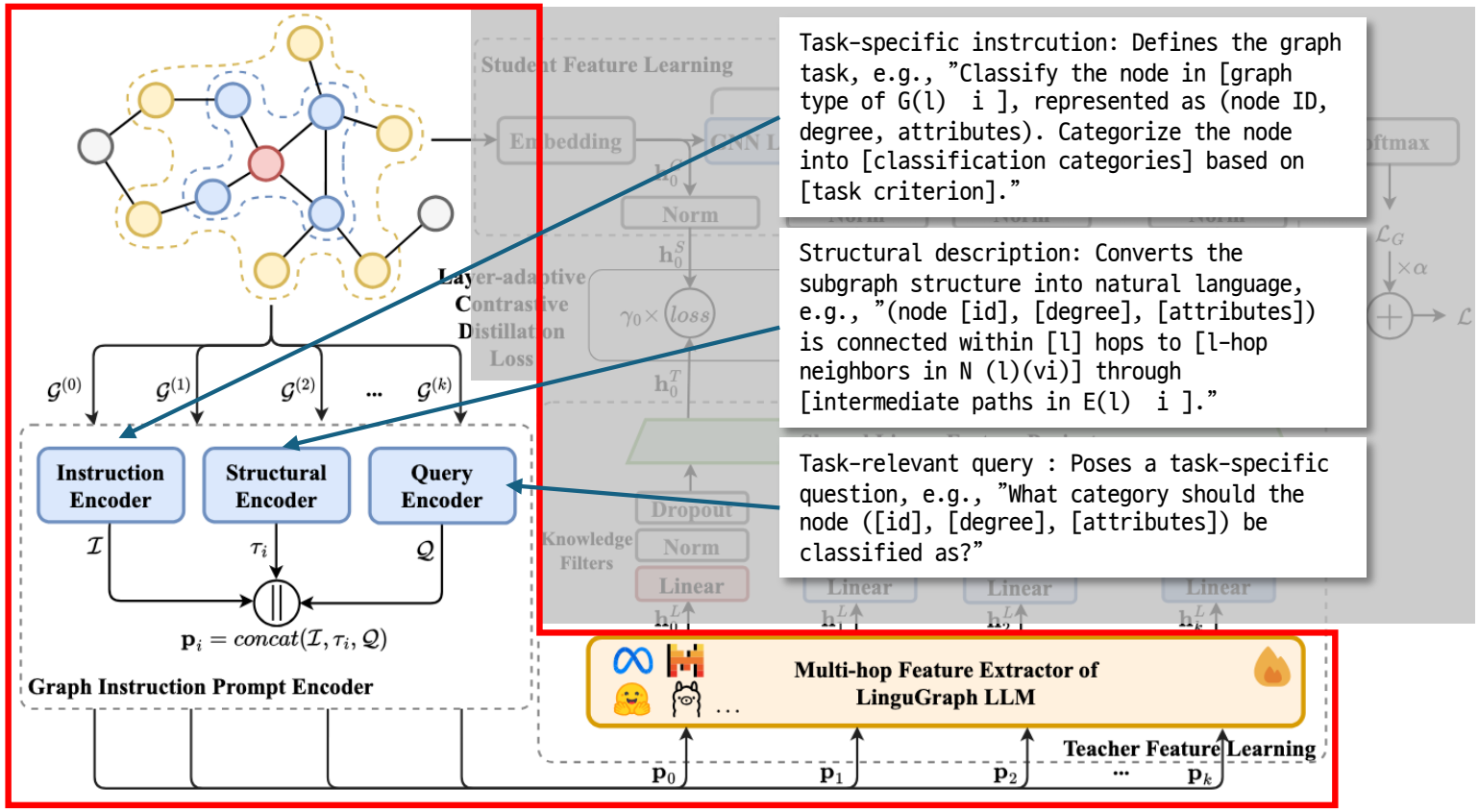
- Goal: LLM의 semantic knowledge + GNN의 structural efficiency

- Challenges
  - ✓ C1: Graph-specific teacher를 어떻게 만들 것인가?
  - ✓ C2: LLM과 GNN의 representation을 어떻게 대응시킬 것인가?
  - ✓ C3: 어떤 scale과 layer의 지식을 옮길 것인가?

- LLM Teacher의 역할: Representation teacher
  - ✓ LLM의 hop별 hidden representation과 관계 구조를 GNN 표현에 정렬

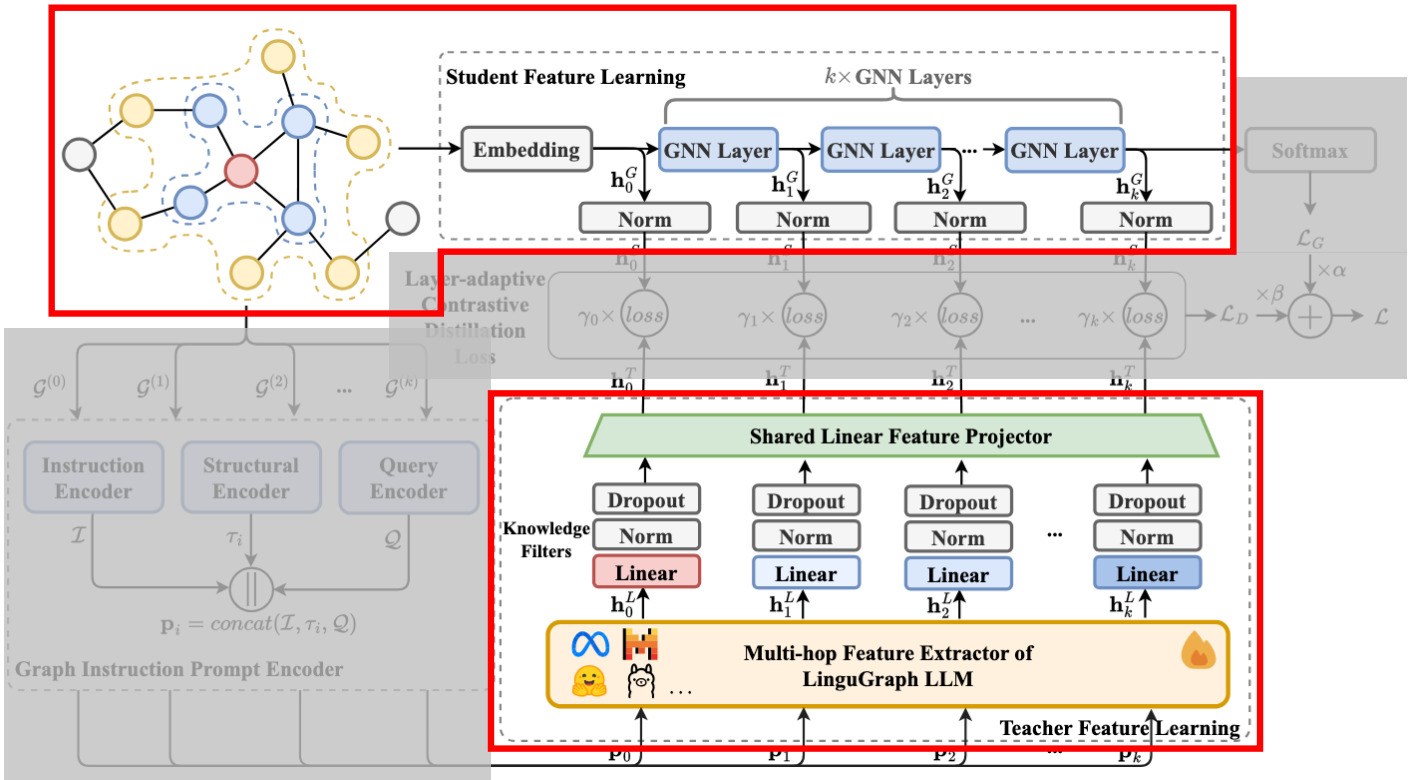
# Linguistic Graph KD (AAAI'25)

- C1: Graph-specific teacher를 어떻게 만들 것인가?
  - ✓ Graph-aware Teacher 구성
    - 중심 노드 주변을 hop별 subgraph로 구성
    - Instruction/structure/query를 결합해 graph prompt 생성
    - Graph instruction tuning으로 LinguGraph LLM 구축



# Linguistic Graph KD (AAAI'25)

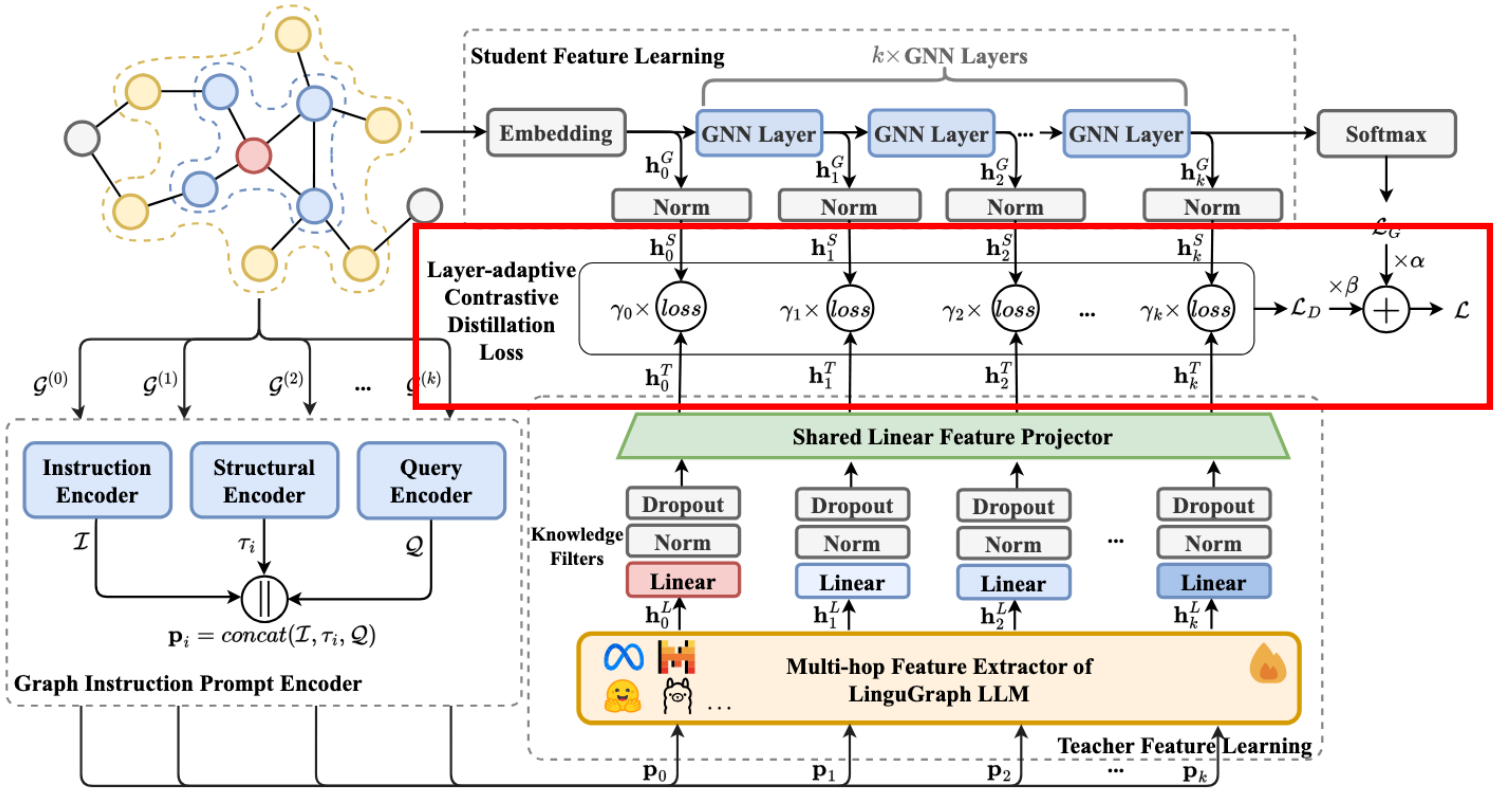
- C2: LLM과 GNN의 representation을 어떻게 대응시킬 것인가?
  - ✓ LLM과 GNN의 Hierarchical Feature 추출
    - Teacher Feature Learning: LLM은 각 hop의 semantic knowledge 추출
    - Student Feature Learning: GNN은 각 layer의 structural knowledge 추출
    - Knowledge Filter와 projector로 동일 hop끼리 연결



# Linguistic Graph KD (AAAI'25)

- C3: 어떤 scale과 layer의 지식을 옮길 것인가?
  - ✓ Multi-scale KD
    - Local alignment로 node-level semantic knowledge 전달
    - Global alignment로 전체 representation 관계 보존
    - Layer-adaptive weight로 중요한 graph depth에 집중

$$\mathcal{L}_D = \sum_{l=0}^{k} \gamma_l \left( \mathcal{L}_{D_l}^{l} + \mathcal{L}_{D_g}^{l} \right)$$



# Linguistic Graph KD (AAAI'25)

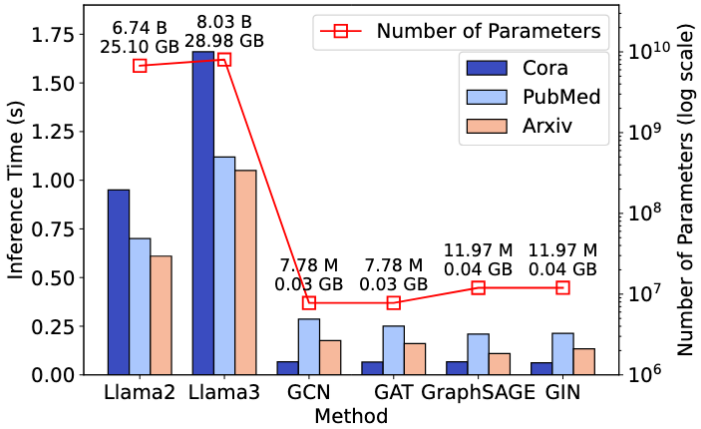
- Experimental Results
  - ✓ Dataset: Cora, PubMed, ogbn-arxiv
  - ✓ LinguGraph-Llama3 (8B)는 모든 데이터셋에서 SOTA 달성
  - ✓ 데이터셋 전체에서 student GNN 성능을 크게 향상시키며, vanilla GNN을 항상 능가

| Methods | Cora | | PubMed | | Methods | Arxiv | |
|---|---|---|---|---|---|---|---|
| | Acc.↑ | F1↑ | Acc.↑ | F1↑ | | Acc.↑ | F1↑ |
| GCN | 86.53±0.92 | 85.66±0.78 | 86.12±0.93 | 85.64±0.82 | GCN | 71.74±0.21 | 71.04±0.37 |
| GAT | 86.12±0.95 | 85.05±0.88 | 85.49±0.76 | 84.89±0.71 | GAT | 73.66±0.33 | 72.44±0.19 |
| GraphSAGE | 87.08±0.85 | 85.96±0.73 | 87.69±0.92 | 87.38±0.68 | GraphSAGE | 71.19±0.26 | 70.87±0.45 |
| GIN | 86.60±0.91 | 85.37±0.74 | 85.84±0.92 | 85.31±0.63 | GIN | 71.62±0.47 | 71.13±0.33 |
| BernNet (He et al. 2021) | 88.52±0.95 | 87.96±0.85 | 88.48±0.41 | 87.52±0.79 | GTAN (Wu and Wang 2022) | 72.97±0.17 | 71.77±0.22 |
| FAGCN (Bo et al. 2021) | 88.85±1.36 | 87.92±0.65 | 89.98±0.54 | 88.72±0.53 | UniMP (Shi et al. 2021) | 73.11±0.20 | 72.14±0.38 |
| GCNII (Chen et al. 2020) | 88.93±1.37 | 87.58±0.71 | 89.80±0.30 | 88.96±0.62 | GCNII (Chen et al. 2020) | 72.74±0.00 | 72.22±0.44 |
| RevGAT (Li et al. 2021) | 89.11±0.00 | 87.65±0.58 | 88.50±0.05 | 87.12±0.73 | RevGAT (Li et al. 2021) | 74.02±0.18 | 73.56±0.29 |
| ACM-GCN+ (Luan et al. 2022) | 89.75±1.16 | 88.94±0.54 | 90.96±0.62 | 89.77±0.51 | E2EG (Dinh et al. 2023) | 73.62±0.14 | 72.96±0.26 |
| Graphormer (Ying et al. 2021) | 80.41±0.30 | 79.98±0.56 | 88.24±1.50 | 87.52±0.71 | SGFormer (Wu et al. 2024) | 72.63±0.13 | 71.58±0.42 |
| LinguGraph-Llama2 (7B) | 88.19±0.83 | 88.12±0.73 | 94.09±0.78 | 93.55±0.61 | LinguGraph-Llama2 (7B) | 75.67±0.52 | 75.60±0.41 |
| LinguGraph-Llama3 (8B) | 91.51±0.46 | **91.53±0.18** | **95.59±0.29** | **95.55±0.10** | LinguGraph-Llama3 (8B) | **79.73±0.18** | **79.29±0.56** |
| $GCN_{(Llama2)}$ | 90.59±0.71 | 89.62±0.66 | 88.97±0.82 | 88.56±0.71 | $GCN_{(Llama2)}$ | 73.87±0.22 | 73.87±0.61 |
| $GCN_{(Llama3)}$ | 90.77±0.28 | 90.35±0.37 | 89.76±0.44 | 89.46±0.37 | $GCN_{(Llama3)}$ | 74.68±0.45 | 74.29±0.32 |
| $GAT_{(Llama2)}$ | 90.33±0.67 | 89.72±0.59 | 87.93±0.28 | 87.42±0.36 | $GAT_{(Llama2)}$ | 74.92±0.14 | 74.48±0.28 |
| $GAT_{(Llama3)}$ | 91.51±0.35 | 91.45±0.58 | 88.31±0.76 | 87.93±0.65 | $GAT_{(Llama3)}$ | 75.71±0.41 | 75.06±0.36 |
| $GraphSAGE_{(Llama2)}$ | 90.22±0.77 | 89.89±0.19 | 89.96±0.50 | 89.67±0.34 | $GraphSAGE_{(Llama2)}$ | 72.53±0.61 | 72.42±0.49 |
| $GraphSAGE_{(Llama3)}$ | **91.70±0.51** | 91.08±0.62 | 90.14±0.56 | 89.96±0.48 | $GraphSAGE_{(Llama3)}$ | 75.38±0.38 | 75.22±0.32 |
| $GIN_{(Llama2)}$ | 90.26±0.67 | 89.20±0.48 | 87.73±0.29 | 87.20±0.30 | $GIN_{(Llama2)}$ | 73.71±0.25 | 73.42±0.10 |
| $GIN_{(Llama3)}$ | 91.33±0.28 | 91.05±0.53 | 89.22±0.79 | 88.87±0.61 | $GIN_{(Llama3)}$ | 75.64±0.46 | 75.28±0.39 |
| *Avg. Dist. Gains* | 4.61% | 5.22% | 2.79% | 2.84% | *Avg. Dist. Gains* | 3.85% | 4.22% |

# Linguistic Graph KD (AAAI'25)

- Experimental Results
  - ✓ LinguGKD의 효율성을 평가하기 위해 기존 SOTA KD 프레임워크와 비교
    - 지속적으로 Cora, Arxiv에서 Accuracy 향상
  - ✓ LinguGraph LLM과 distilled GNN에 대한 모델 크기와 추론 시간 간의 trade-off
    - LLM은 뛰어난 성능을 제공하지만 모델 크기가 매우 커지고 추론 시간이 길어짐
      - 리소스가 충분하고 최대 정확도가 중요한 경우에 선호
    - Distilled GNN은 훨씬 더 작은 공간과 빠른 추론으로 비슷한 정확도를 달성함
      - 리소스가 제한된 실시간 application에 이상적

| Model | Cora | Arxiv |
|---|---|---|
| Teacher of baselines | 88.93 | 73.91 |
| Teacher of LinguGKD | **91.51** | **79.73** |
| KD (Hinton, Vinyals, and Dean 2015) | 86.38 | 71.55 |
| FitNet (Romero et al. 2015) | 85.40 | 71.38 |
| LSP (Yang et al. 2020) | 84.92 | 71.52 |
| GraphAKD (He et al. 2022) | 86.39 | - |
| G-CRD (Joshi et al. 2024) | - | 71.64 |
| LinguGKD | 90.77 | 74.68 |



# Dual KD for Robust GNN (AAAI'26)

**DuoKD: Dual Knowledge Distillation from Large Language Models for Robust Graph Neural Networks**

**Cuiying Huo[1], Xiaotong Huang[2], Dongxiao He[1], Yixuan Du[1], Wenhuan Lu[1,3,*] Di Jin[1,3,*]**

[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]School of Mathematics, Tianjin University, Tianjin, China
[3]School of Intelligence Science and Engineering, Qinghai Minzu University, Xining, China
{huocuiying, huangxt_3604, hedongxiao, wenhuan, jindi}@tju.edu.cn

Cited: 1

- 비슷한 정보만 distillation하면 충분할까?
  - ✓ 기존 LLM-GNN 방식은 주로 class와 의미적으로 유사한 positive knowledge만 전달
  - ✓ 하지만 정확한 decision boundary를 만들려면 negative knowledge도 필요
  - ✓ 노이즈가 환각이 섞인 LLM 지식을 그대로 전달하면 오류가 증폭되므로, semantic/structural rationale을 평가하고 선별해야 함

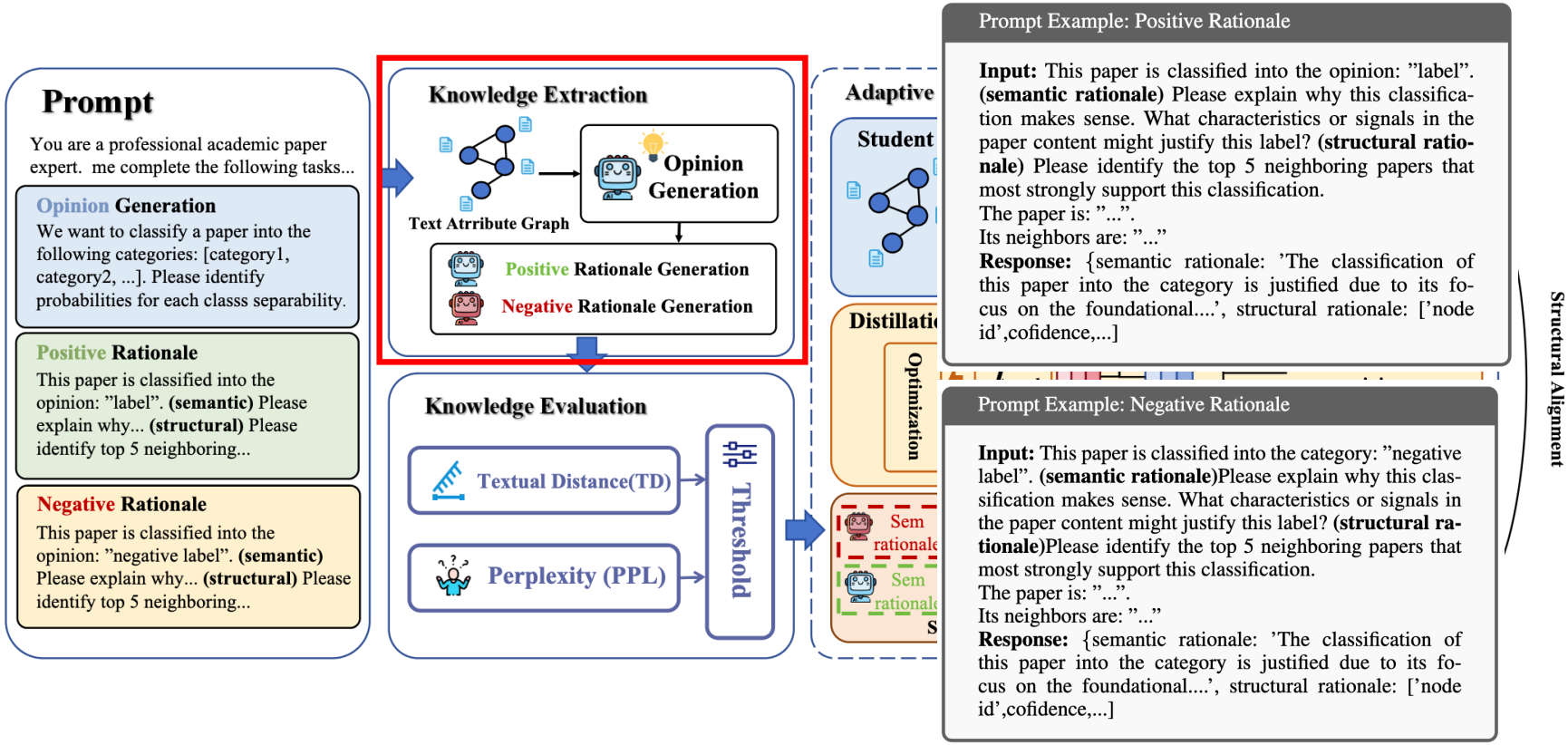
# Dual KD for Robust GNN (AAAI'26)

- Challenges
  - ✓ C1: Intra-class similarity뿐 아니라 inter-class boundary도 학습해야 함
  - ✓ C2: LLM이 생성한 supervision이 신뢰할 만한지 평가해야 함
  - ✓ C3: 단순 label 외에 semantic/structural reasoning 지식도 옮겨야 함

- LLM Teacher의 역할: Multi-signal rationale teacher
  - ✓ verbalized soft opinion, positive/negative semantic rationale, structural rationale

# Dual KD for Robust GNN (AAAI'26)

- C1: Intra-class similarity뿐 아니라 inter-class boundary도 학습해야 함
  - ✓ Positive/Negative Knowledge 생성
    - 가장 가능성 높은 class에서 positive opinion 생성
    - 가장 가능성 낮은 class에서 negative opinion 생성
    - 각각 semantic/structural rationale 생성

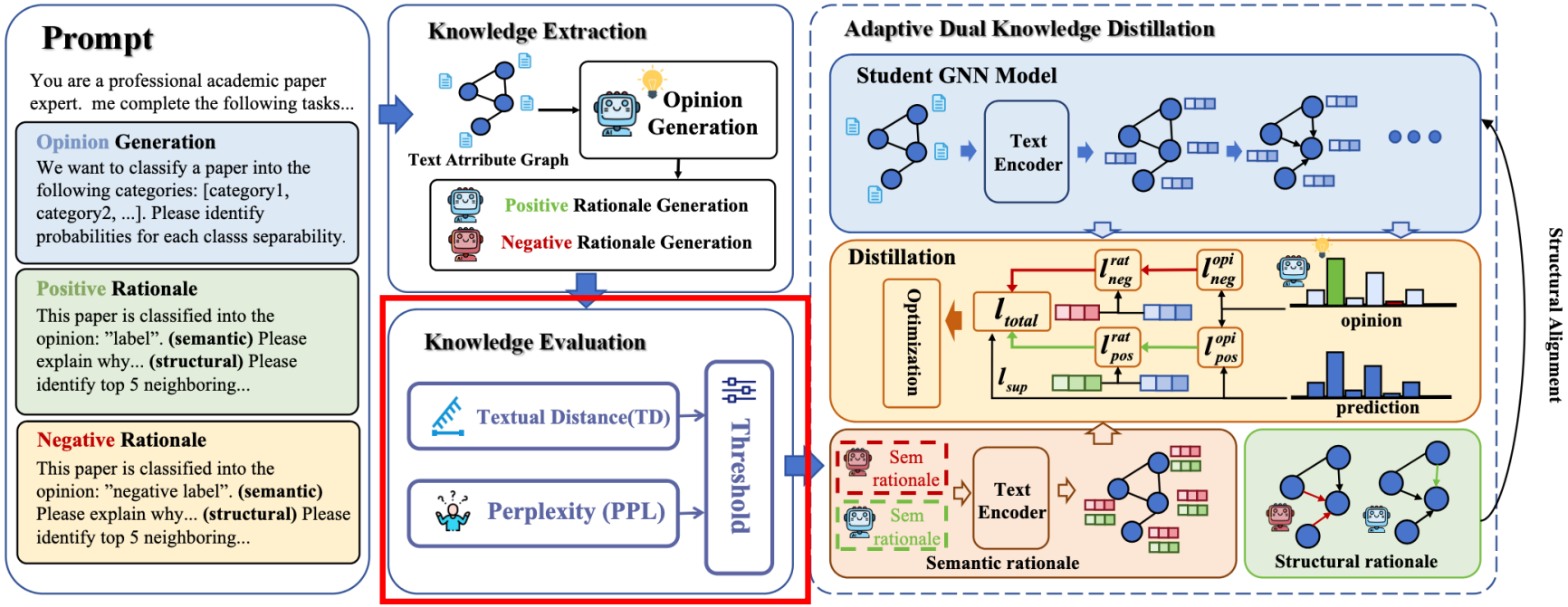


# Dual KD for Robust GNN (AAAI'26)

- C2: LLM이 생성한 supervision이 신뢰할 만한지 평가해야 함
  - ✓ 생성된 knowledge 검증
    - Textual Distance로 rationale 간 deviation 평가
    - Perplexity로 생성 문장의 품질 평가
    - 신뢰할 수 있는 semantic rationale만 유지

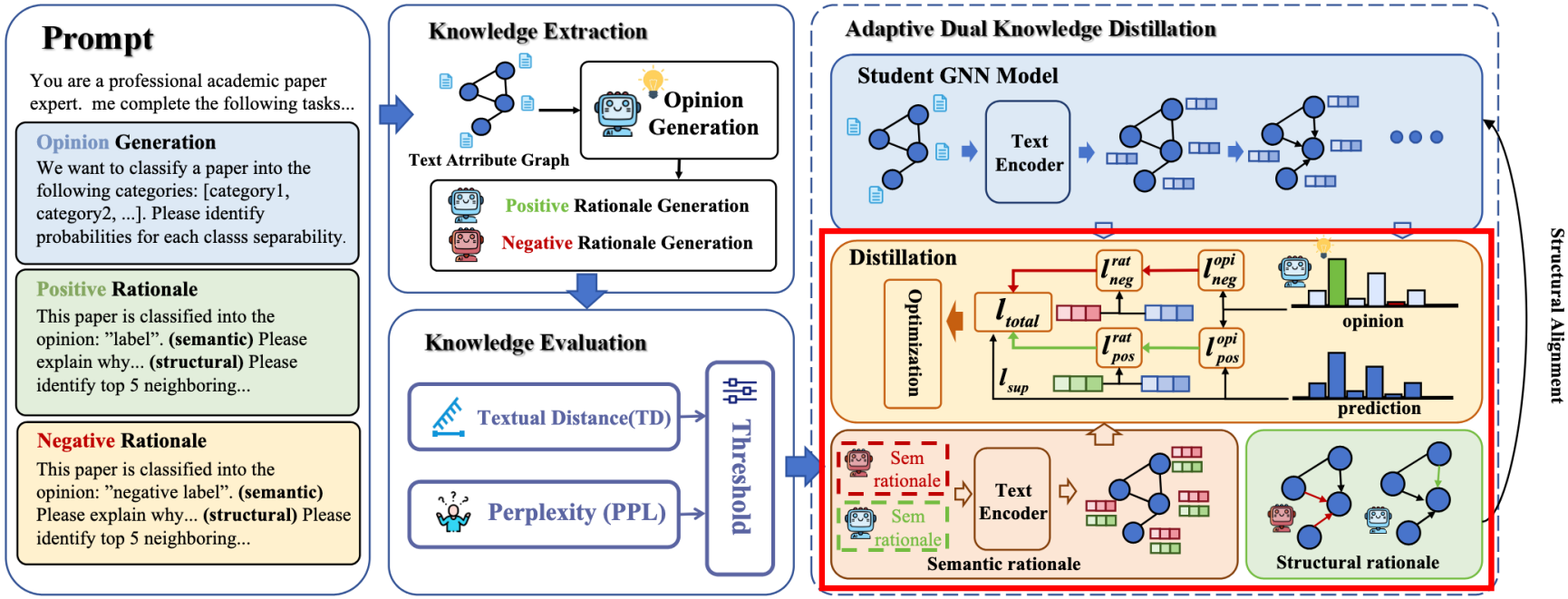
$$\mathrm{TD}(x, x_{cf}) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{Levenshtein}(x^{(i)}, x_{cf}^{(i)})$$

$$\mathrm{PPL}(x_{cf}) = \exp\left(-\frac{1}{T}\sum_{t=1}^{T} \log P_{\text{GPT-2}}(w_t \mid w_{<t})\right)$$



# Dual KD for Robust GNN (AAAI'26)

- C3: 단순 label 외에 semantic/structural reasoning 지식도 옮겨야 함
  - ✓ Positive는 Pull, Negative는 Push
    - Positive opinion/rationale와 GNN representation을 정렬
    - Negative opinion/rationale에서는 멀어지도록 학습
    - Structural rationale는 중요한 neighbor의 message passing에 반영




$$\mathcal{L}_{pos}^{opi} = \mathrm{KL}\big(\sigma(\mathbf{z}^{(LLM)}/\tau) \,\|\, \sigma(\mathbf{z}^{(GNN)}/\tau)\big)$$

$$\mathcal{L}_{neg}^{opi} = \sum_{i \in \mathcal{I}_{neg}} \max\left(0, \cos(\mathbf{z}_i^{(GNN)}, \mathbf{z}_i^{(LLM)}) - m\right)$$

$$\mathcal{L}_{pos}^{rat} = \mathbb{1}\left[\mathrm{Score}_{\mathrm{sem}}(r_i^+) < \tau\right] \cdot \mathrm{MSE}\big(\tilde{\phi}^{GNN}(v_i), \phi^{LLM}(r_i^+)\big) \quad (11)$$

$$\mathcal{L}_{neg}^{rat} = \mathbb{1}\left[\mathrm{Score}_{\mathrm{sem}}(r_i^-) < \tau\right] \cdot \max\left(0, \cos(\tilde{\phi}^{GNN}(v_i), \phi^{LLM}(r_i^-)) - m\right)$$

# Dual KD for Robust GNN (AAAI'26)

- Experimental Results
  - ✓ Dataset: Cora, PubMed, ogbn-arxiv
  - ✓ 기존 GNN, LLM 모델 및 KD에서 일관된 개선을 보임

| Method | Cora | | PubMed | | ogbn-arxiv | |
|---|---|---|---|---|---|---|
| | Acc | F1 | Acc | F1 | Acc | F1 |
| GCN | 86.53 ± 0.92 | 85.36 ± 1.01 | 86.15 ± 0.87 | 85.64 ± 0.82 | 71.74 ± 0.21 | 70.44 ± 0.33 |
| GAT | 86.10 ± 1.00 | 85.12 ± 1.03 | 86.20 ± 0.93 | 85.31 ± 0.88 | 73.66 ± 0.35 | 71.04 ± 0.47 |
| GraphSAGE | 86.92 ± 0.85 | 85.85 ± 0.79 | 87.12 ± 0.76 | 86.90 ± 0.70 | 73.85 ± 0.43 | 72.31 ± 0.50 |
| GIN | 86.60 ± 0.91 | 85.41 ± 0.88 | 86.40 ± 0.95 | 85.78 ± 0.84 | 71.62 ± 0.42 | 70.33 ± 0.47 |
| FAGCN | 88.32 ± 0.55 | 86.93 ± 0.61 | 88.64 ± 0.63 | 87.96 ± 0.58 | 74.21 ± 0.58 | 72.97 ± 0.62 |
| GIANT (GCN) | 87.74 ± 0.85 | 86.30 ± 0.78 | 88.32 ± 0.70 | 87.05 ± 0.69 | 75.61 ± 0.67 | 74.25 ± 0.65 |
| GLEM (GCN) | 88.71 ± 0.61 | 87.51 ± 0.65 | 88.74 ± 0.62 | 88.04 ± 0.57 | 76.20 ± 0.44 | 75.20 ± 0.47 |
| GLEM (GAT) | 89.12 ± 0.54 | 88.04 ± 0.60 | 89.01 ± 0.55 | 88.36 ± 0.48 | 76.42 ± 0.52 | 75.43 ± 0.50 |
| GLEM (SAGE) | 88.90 ± 0.47 | 87.92 ± 0.51 | 89.30 ± 0.58 | 88.62 ± 0.53 | 76.80 ± 0.41 | 75.91 ± 0.43 |
| LinguGKD (GCN) | 90.50 ± 0.42 | 88.90 ± 0.43 | 88.78 ± 0.54 | 88.66 ± 0.57 | 76.15 ± 0.51 | 76.92 ± 0.50 |
| LinguGKD (GAT) | 90.83 ± 0.46 | 89.21 ± 0.50 | 88.93 ± 0.47 | 87.90 ± 0.45 | 77.62 ± 0.40 | 77.21 ± 0.39 |
| LinguGKD (SAGE) | 90.91 ± 0.43 | 89.17 ± 0.44 | 88.10 ± 0.42 | 88.14 ± 0.38 | 77.05 ± 0.39 | 77.89 ± 0.41 |
| **Ours (GCN)** | **91.91 ± 0.57** | **89.10 ± 0.63** | **90.93 ± 0.49** | **89.89 ± 0.55** | **78.84 ± 0.66** | **78.52 ± 0.61** |
| **Ours (GAT)** | **92.30 ± 0.51** | **89.60 ± 0.58** | **89.21 ± 0.42** | **89.04 ± 0.48** | **80.13 ± 0.59** | **78.81 ± 0.50** |
| **Ours (SAGE)** | **91.85 ± 0.63** | **88.92 ± 0.71** | **88.05 ± 0.44** | **88.78 ± 0.52** | **80.01 ± 0.47** | **78.66 ± 0.46** |

# Dual KD for Robust GNN (AAAI'26)

- Robustness analysis
  - ✓ non-targeted DICE attack, random text replacement를 활용
  - ✓ positive뿐 아니라 negative knowledge도 통합함으로써 더 robust한 결과를 얻음
  - ✓ <u>negative knowledge는 모델이 오해 소지가 있는 신호를 인식하고 거부하도록 명시적으로 가르침으로써 decision boundary를 강화함</u>

| | | Structural Perturbation | | | | Attribute Perturbation | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Dataset** | **Ptb Rate** | **GCN** | **GIANT** | **GLEM** | **Ours** | **GCN** | **GIANT** | **GLEM** | **Ours** |
| Cora | 0% | 86.53 ± 0.92 | 87.74 ± 0.85 | 88.78 ± 0.46 | 91.91 ± 0.57 | 86.53 ± 0.92 | 87.74 ± 0.85 | 88.78 ± 0.46 | 91.91 ± 0.57 |
| | 10% | 81.38 ± 0.65 | 84.21 ± 0.53 | 85.92 ± 0.42 | 86.35 ± 0.34 | 82.80 ± 0.58 | 85.45 ± 0.48 | 87.25 ± 0.39 | 89.42 ± 0.33 |
| | 20% | 78.38 ± 0.70 | 82.42 ± 0.58 | 84.63 ± 0.48 | 85.90 ± 0.36 | 78.60 ± 0.67 | 82.10 ± 0.53 | 85.15 ± 0.45 | 87.30 ± 0.37 |
| | 40% | 71.62 ± 0.76 | 77.93 ± 0.64 | 80.78 ± 0.52 | 84.21 ± 0.38 | 72.30 ± 0.74 | 76.45 ± 0.60 | 80.50 ± 0.48 | 85.00 ± 0.35 |
| Pubmed | 0% | 86.15 ± 0.87 | 88.32 ± 0.70 | 88.68 ± 0.45 | 90.93 ± 0.49 | 86.15 ± 0.87 | 88.32 ± 0.70 | 88.68 ± 0.45 | 90.93 ± 0.49 |
| | 10% | 82.86 ± 0.63 | 85.60 ± 0.50 | 87.02 ± 0.43 | 88.12 ± 0.33 | 83.25 ± 0.56 | 86.15 ± 0.44 | 87.90 ± 0.38 | 88.65 ± 0.32 |
| | 20% | 80.11 ± 0.68 | 83.53 ± 0.55 | 85.84 ± 0.47 | 87.75 ± 0.35 | 78.95 ± 0.64 | 83.20 ± 0.51 | 86.10 ± 0.43 | 86.95 ± 0.36 |
| | 40% | 76.15 ± 0.73 | 80.92 ± 0.61 | 83.67 ± 0.50 | 86.24 ± 0.37 | 73.15 ± 0.72 | 78.85 ± 0.57 | 83.30 ± 0.46 | 84.90 ± 0.34 |
| ogbn-arxiv | 0% | 71.74 ± 0.21 | 75.61 ± 0.67 | 76.12 ± 0.42 | 78.84 ± 0.66 | 71.74 ± 0.21 | 75.61 ± 0.67 | 76.12 ± 0.42 | 78.84 ± 0.66 |
| | 10% | 64.23 ± 0.69 | 67.54 ± 0.53 | 69.90 ± 0.45 | 71.12 ± 0.32 | 67.50 ± 0.62 | 71.55 ± 0.50 | 74.00 ± 0.41 | 76.85 ± 0.30 |
| | 20% | 62.35 ± 0.72 | 65.83 ± 0.56 | 67.74 ± 0.47 | 69.80 ± 0.34 | 63.20 ± 0.66 | 68.20 ± 0.52 | 71.60 ± 0.44 | 74.35 ± 0.30 |
| | 40% | 56.44 ± 0.76 | 61.29 ± 0.61 | 64.88 ± 0.50 | 67.65 ± 0.36 | 57.90 ± 0.71 | 63.40 ± 0.59 | 67.80 ± 0.47 | 72.00 ± 0.33 |

# LLM-Augemented Graph Learning (WWW'26)

## Detecting Miscitation on the Scholarly Web through LLM-Augmented Text-Rich Graph Learning


Huidong Wu
Chinese Academy of Sciences & City University of Hong Kong
Beijing, China
wuhuidong21@mails.ucas.ac.cn

Haojia Xiang
City University of Hong Kong
Hong Kong, China
haojxiang2-c@my.cityu.edu.hk

Jingtong Gao
City University of Hong Kong
Hong Kong, China
jt.g@my.cityu.edu.hk

Xiangyu Zhao*
City University of Hong Kong
Hong Kong, China
xianzhao@cityu.edu.hk

Dengsheng Wu*
Shenzhen University & Beijing Dongbi Data Technology
Shenzhen, China
wds@szu.edu.cn

Jianping Li
University of Chinese Academy of Sciences
Beijing, China
ljp@ucas.ac.cn


Cited: 3

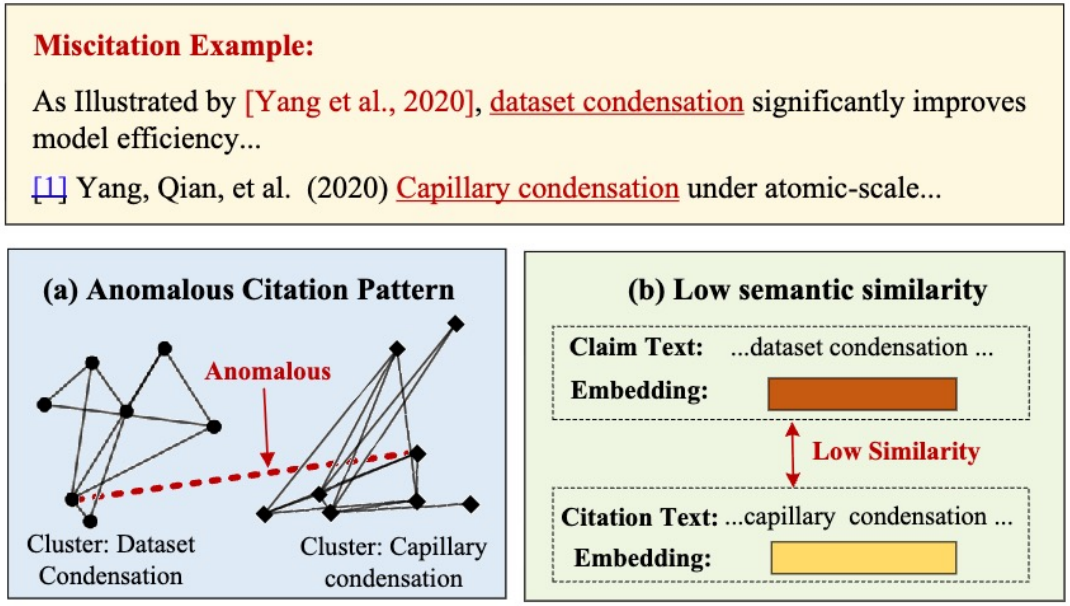
- ✓ **Task-specific Reasoning-state KD**: multi-hop reasoning state를 GNN layer에 정렬

- Citation 하나의 진위를 판단하려면 그 citation만 보면 될까?
  - ✓ 기존 structure 기반 방식은 이상 인용 패턴만 보고, text 기반 방식은 표면적 의미 유사도에 의존
  - ✓ 실제 인용 타당성은 여러 문헌을 거치는 multi-hop evidence chain을 통해 검증해야 함
  - ✓ 모든 citation에 LLM 추론을 수행하기에는 비용이 크므로, task-specific reasoning state를 GNN에 distillation함

# LLM-Augemented Graph Learning (WWW'26)

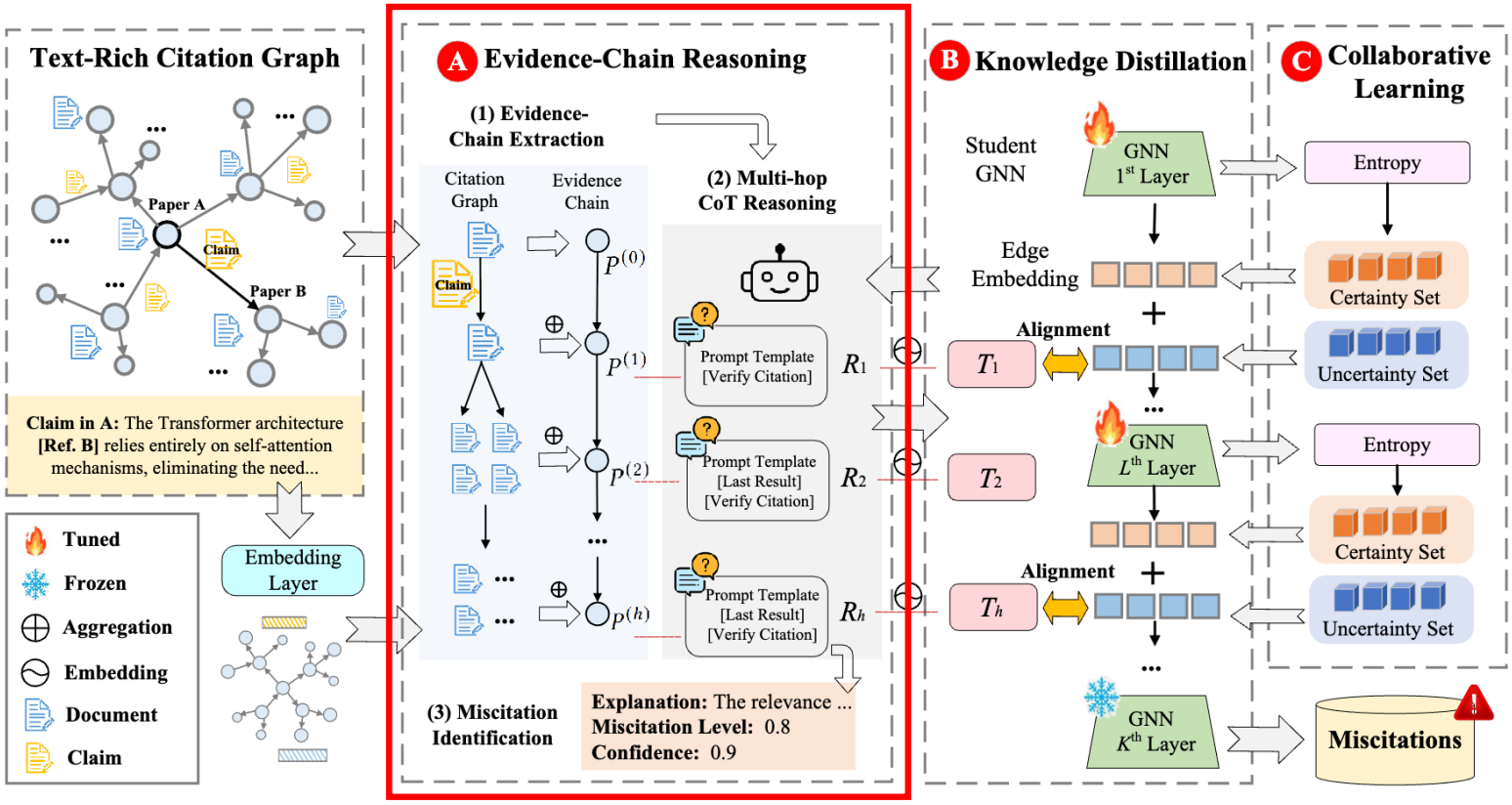
- Challenges
  - ✓ C1: Local citation pair만으로 안정적인 판단이 가능한가?
  - ✓ C2: Multi-hop LLM reasoning을 web scale에서 실행할 수 있는가?
  - ✓ C3: 모든 training example에 동일하게 KD해야 하는가?

- LLM Teacher의 역할: Task-specific reasoning teacher
  - ✓ evidence-chain의 hop별 reasoning state를 GNN layer별 edge representation에 정렬



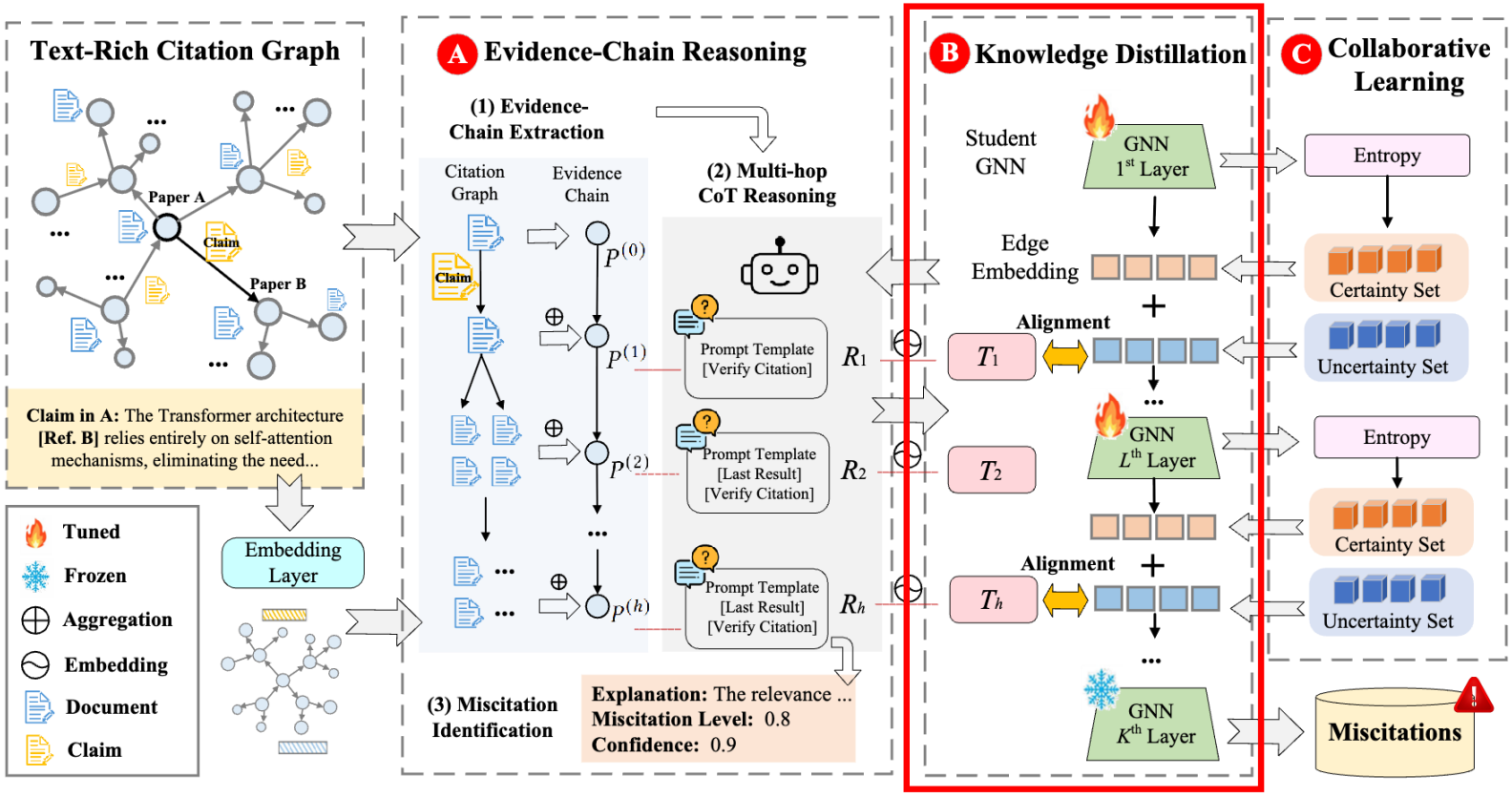
# LLM-Augemented Graph Learning (WWW'26)

- C1: Local citation pair만으로 안정적인 판단이 가능한가?
  - ✓ Evidence Chain을 따라 Citation 검증
    - Citation source를 multi-hop으로 추적
    - 각 hop에서 citation 내용의 semantic fidelity 검증
    - 누적된 reasoning으로 miscitation level과 confidence 출력



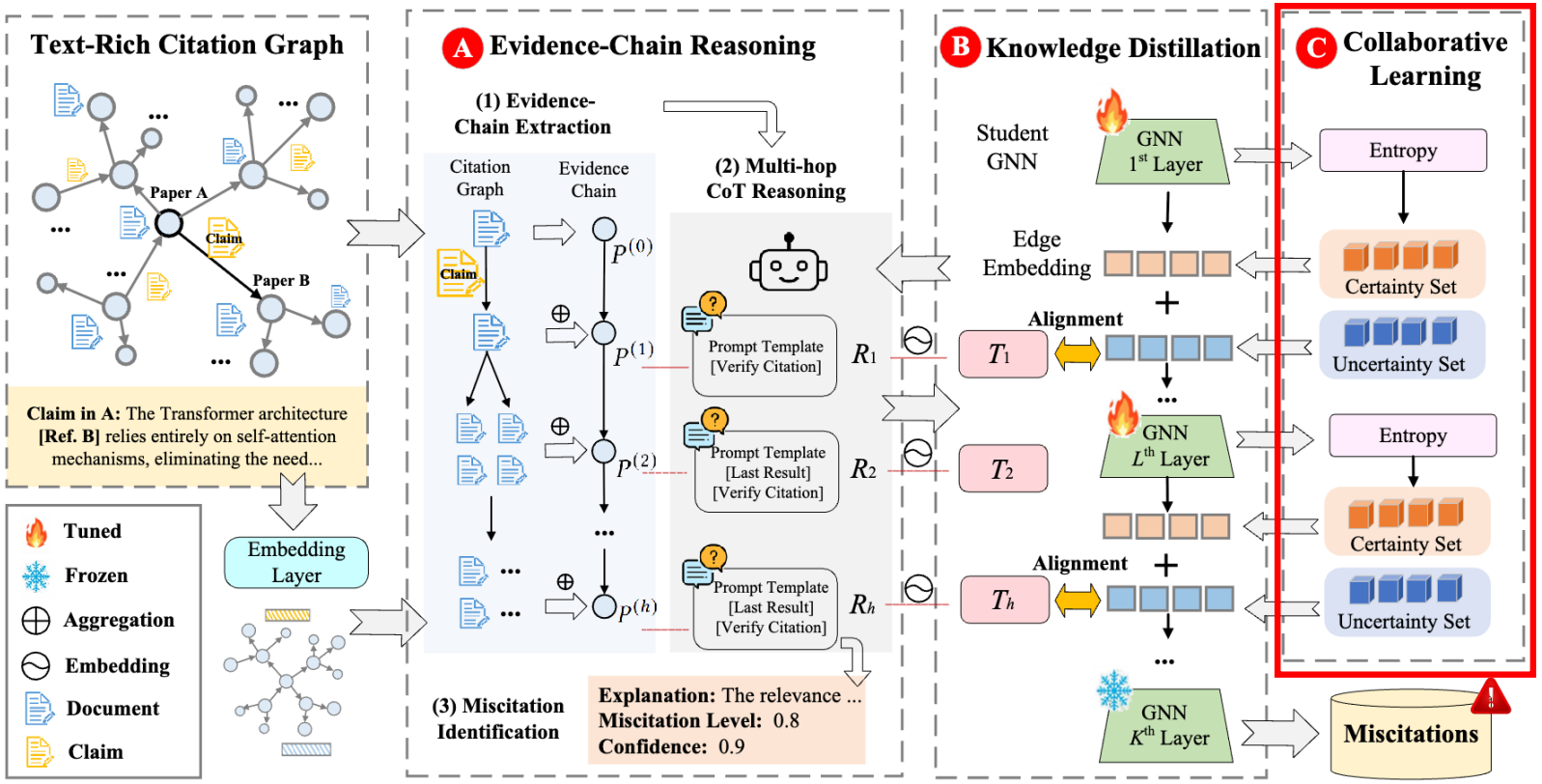
# LLM-Augemented Graph Learning (WWW'26)

- C2: Multi-hop LLM reasoning을 web scale에서 실행할 수 있는가?
  - ✓ LLM Reasoning을 GNN Layer에 전달
    - LLM의 hop별 reasoning state 추출
    - GNN의 대응 layer에서 edge representation 추출
    - 같은 hop과 layer를 contrastive learning으로 정렬



# LLM-Augemented Graph Learning (WWW'26)

- C3: 모든 training example에 동일하게 KD해야 하는가?
  - ✓ 어려운 Training Case에만 집중
    - GNN entropy로 불확실한 training edge 탐색
    - Teacher confidence로 noisy reasoning 제거
    - 선택된 edge에만 targeted distillation 적용



# LLM-Augemented Graph Learning (WWW'26)

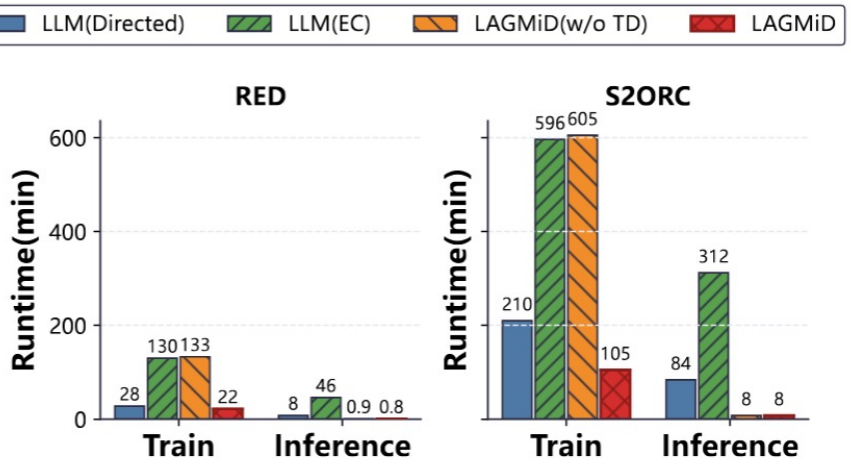
- Experimental Results
  - ✓ Dataset: RED, SciFact, S2ORC
  - ✓ 구조적 특징이나 의미적 유사도에 의존하는 전통적 GNN 및 PLM 기반 방법은 가장 약함
    - 정확한 탐지에 필수적인 구조적 인식과 의미 이해를 효과적으로 통합하는 데 있어 한계를 보임
  - ✓ LLM 기반 방법은 향상된 결과를 보이지만, 텍스트 중심 graph learning에 비해 미흡함

| Dataset | Metric | GCN | GLAD | RoBERTa | SciBERT | GLM | Qwen | AnomalyLLM | GuARD | LAGMiD (Ours) |
|---|---|---|---|---|---|---|---|---|---|---|
| RED | AUC | 0.7559 | 0.7774 | 0.7068 | 0.7651 | 0.8220 | 0.8362 | 0.8982 | 0.9100 | **0.9615*** |
| | F1 | 0.6857 | 0.6936 | 0.6082 | 0.7471 | 0.7202 | 0.8056 | 0.8354 | 0.8571 | **0.9167*** |
| | Precision | 0.7557 | 0.7774 | 0.5982 | 0.7182 | 0.8051 | 0.7682 | 0.8327 | 0.8753 | **0.9167*** |
| SciFact | AUC | 0.6045 | 0.6244 | 0.5677 | 0.6211 | 0.7909 | 0.7832 | 0.8024 | 0.7906 | **0.8208*** |
| | F1 | 0.5138 | 0.6006 | 0.5252 | 0.6622 | 0.7242 | 0.7213 | 0.7958 | 0.7714 | **0.8205*** |
| | Precision | 0.6673 | 0.6677 | 0.6360 | 0.7102 | 0.7250 | 0.7536 | 0.7485 | 0.7453 | **0.7724*** |
| S2ORC | AUC | 0.6971 | 0.7883 | 0.7459 | 0.7024 | 0.7129 | 0.7506 | 0.7547 | 0.7528 | **0.8100*** |
| | F1 | 0.6871 | 0.7932 | 0.6330 | 0.7094 | 0.7621 | 0.7678 | 0.7381 | 0.7687 | **0.8256*** |
| | Precision | 0.6606 | 0.7733 | 0.7081 | 0.7653 | 0.7681 | 0.7324 | 0.7420 | 0.7740 | **0.8419*** |

# LLM-Augemented Graph Learning (WWW'26)

- Ablation & Efficiency
  - ✓ Evidence-chain reasoning을 제거했을 때 성능이 가장 크게 감소
  - ✓ Distilled GNN으로 LLM reasoning보다 빠른 추론 가능

| Model | RED | | | S2ORC | | |
|---|---|---|---|---|---|---|
| | AUC | F1 | Precision | AUC | F1 | Precision |
| w/o EC | 0.9295 | 0.8412 | 0.8571 | 0.7480 | 0.7855 | 0.7728 |
| w/o KD | 0.9487 | 0.8511 | 0.8696 | 0.7586 | 0.7935 | 0.7885 |
| w/o LD | 0.9503 | 0.8696 | 0.8571 | 0.7700 | 0.8207 | 0.8035 |
| w/o TD | 0.9487 | 0.8511 | 0.8696 | 0.7625 | 0.7968 | 0.8277 |
| LAGMiD | **0.9615** | **0.9167** | **0.9167** | **0.8100** | **0.8256** | **0.8419** |





# CONTENTS

# Recent Trends

- LLM Teacher의 역할은 지식을 더 정교하게 다루는 것까지 확장됨
  - ✓ judge, generator ..

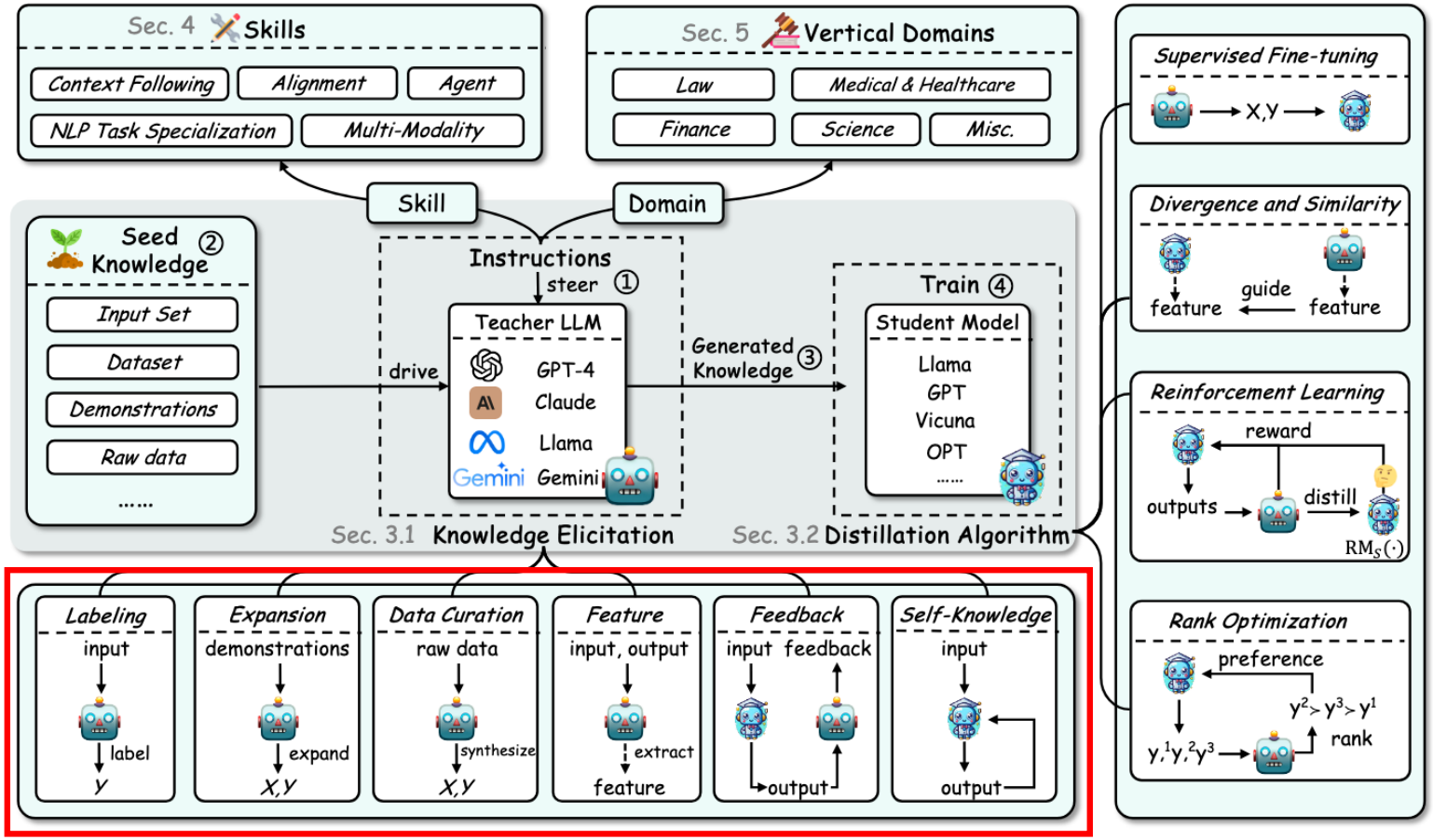
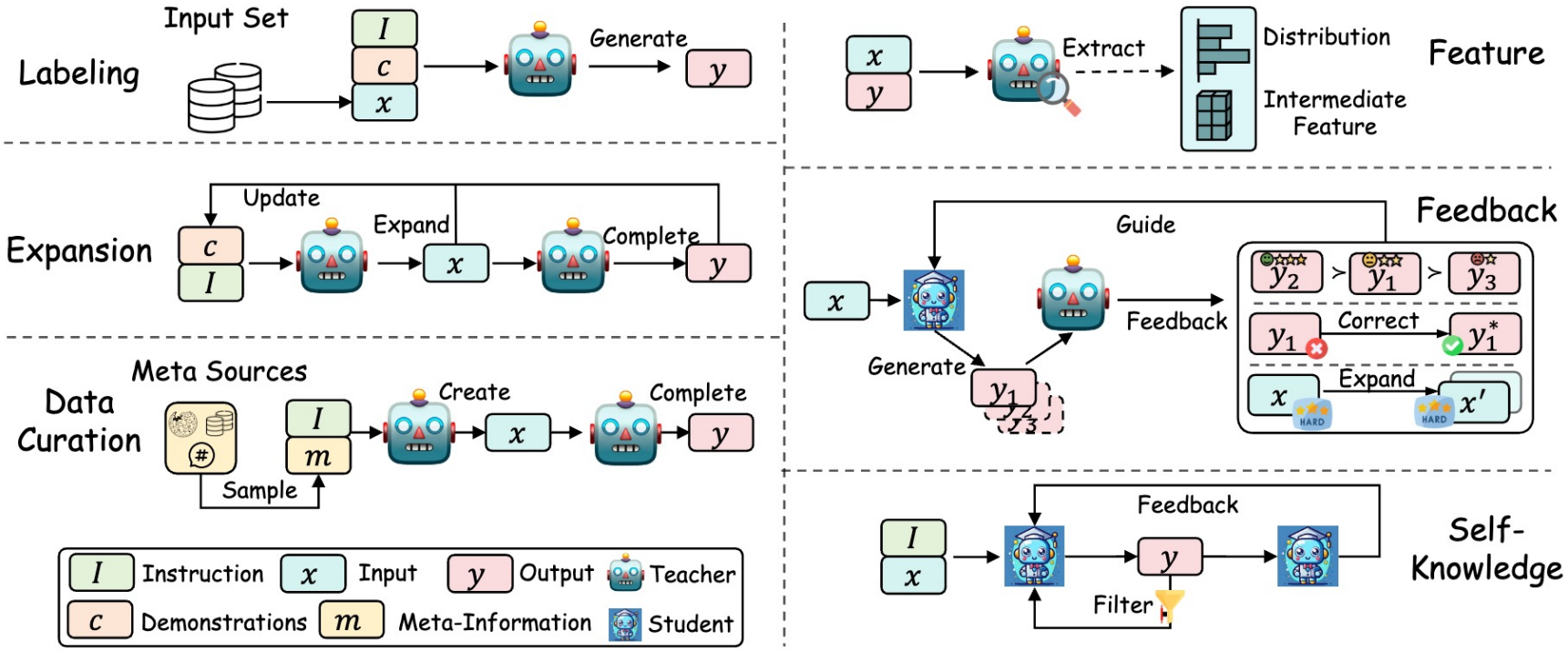



Fig. 2: An overview of this survey on knowledge distillation of large language models. Note that 'Section' is abbreviated as 'Sec.' in this figure. $\mathrm{RM}_S(\cdot)$ denotes the student reward model. ①②③④ denote the steps in KD of LLMs.

# Recent Trends

- Knowledge carrier가 task-specific해지고 있음
  - ✓ 초기에는 LLM의 answer·label을 모방하는 방식이 중심
  - ✓ 최근에는 feature, preference, ranking, rationale 등 student task에 맞는 지식을 전달
  - ✓ 즉, 무엇을 전달할지는 teacher가 아니라 student의 inductive bias와 downstream task가 결정



# Limitations

- 서로 다른 architecture의 feature는 자연스럽게 대응되지 않음
  - ✓ 같은 architecture에서는 layer 간 대각선 대응이 비교적 명확
  - ✓ CNN–ViT–MLP처럼 architecture가 달라지면 representation correspondence가 크게 약화
  - ✓ 따라서 LLM layer와 GNN layer를 단순 MSE로 맞추는 것이 의미 있는 knowledge transfer를 보장하지 않음

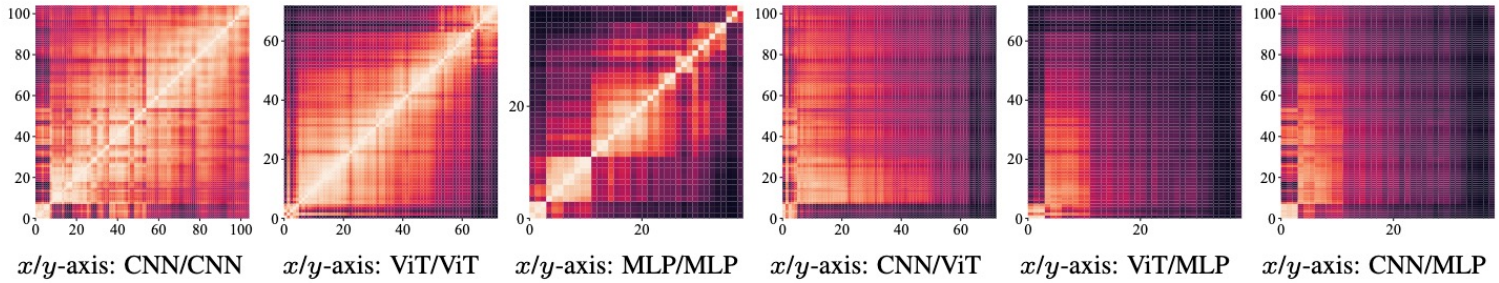



Figure 1: Similarity heatmap of intermediate features measured by CKA. We compare features from **MobileNetV2** (CNN), **ViT-Small** (Transformer) and **Mixer-B/16** (MLP model). The first three figures illustrate similarity between homogeneous models, and the last three illustrate feature similarity between heterogeneous models. Coordinate axes indicate the corresponding layer index.

[1] Hao et al., "One-for-All: Bridge the Gap Between Heterogeneous Architectures in Knowledge Distillation", NIPS'23, Cited: 254

# Limitations

- 좋은 predictor가 항상 좋은 teacher는 아님
  - ✓ teacher의 높은 task accuracy가 student improvement를 보장하지 않음
  - ✓ teacher의 hallucination/tail noise/과도한 믿음이 그대로 distillation signal이 될 수 있음
  - ✓ KD 성공 여부는 teacher–student–data의 조합과 teacher calibration에 의존

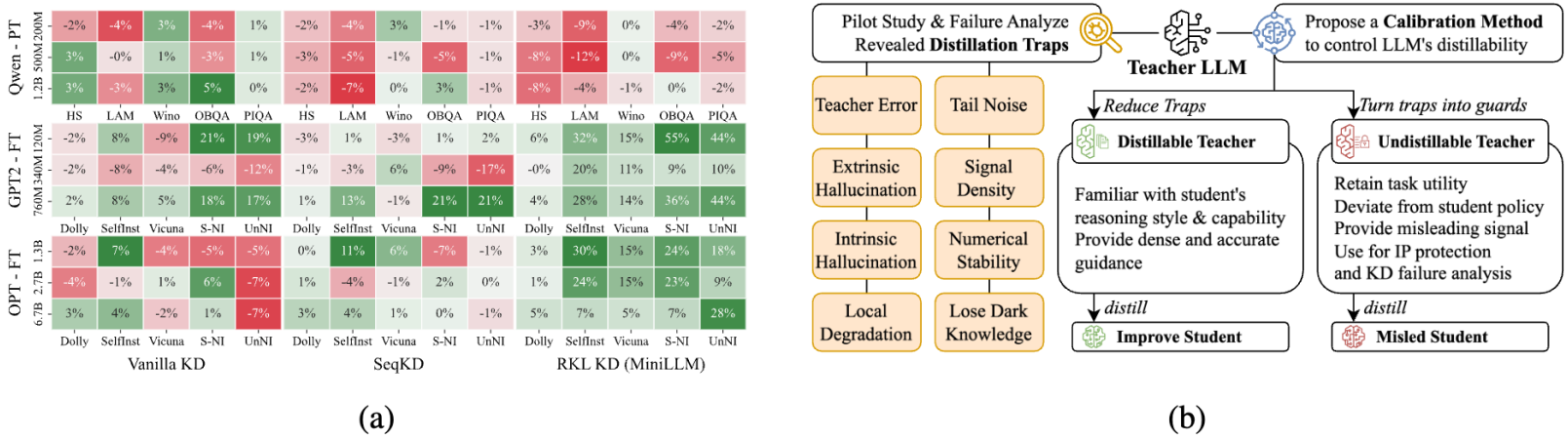
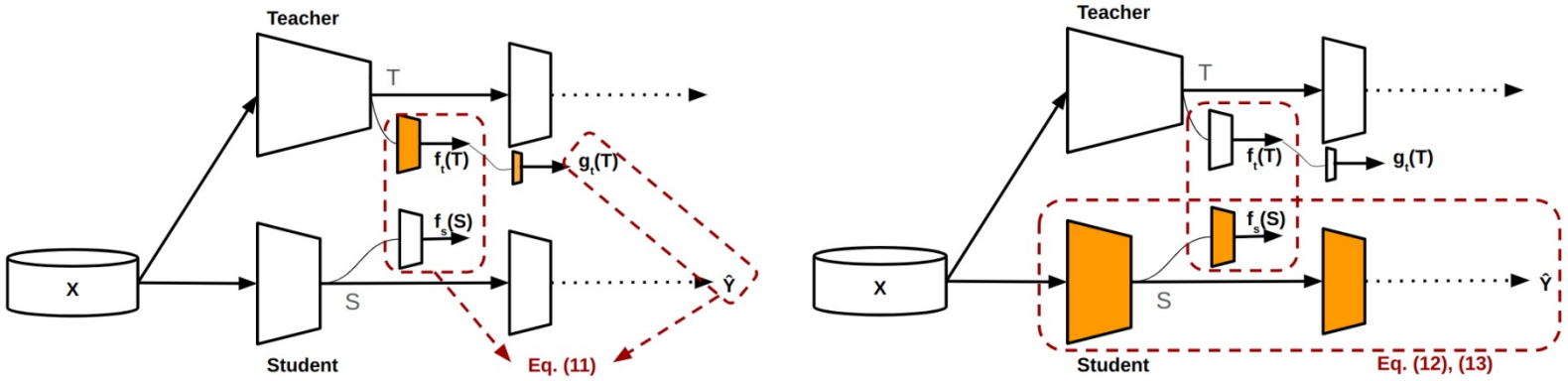



Figure 1: (a) The efficacy of KD is not always guaranteed, with certain combinations of models and datasets leading to unexpected failure. This heatmap shows relative performance gain and loss from employing various KD methods in pre-training (PT) and fine-tuning (FT) compared to training without KD loss. (b) We identify several "Distillation Traps" and propose a calibration method to control models' distillability. Reducing the traps yields distillable teachers that can give dense and accurate guidance, whereas amplifying them can turn traps into guards, yielding "undistillable teachers" preventing unauthorized distillation.

[1] Zhan et al., "Distillation Traps and Guards: A Calibration Knob for LLM Distillability", ACL'26, Cited: 0

# Future Works

- 모든 representation이 아니라 공통되고 task-relevant한 지식만 전달
  - ✓ teacher feature 전체를 복제하기보다 architecture-independent component를 식별
  - ✓ student의 고유한 inductive bias는 제거하지 않고 보존




(a) Phase 1: Train the teacher's filters while the student is kept frozen

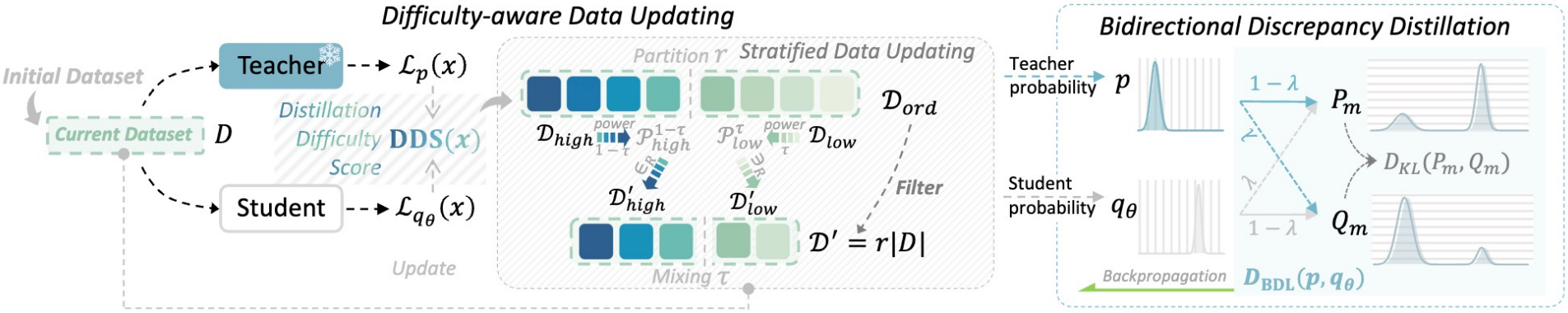
(b) Phase 2: Train the student along with student's filters while the teacher is kept frozen

Figure 3: Redundant Information Distillation (RID) framework: $f_t(\cdot)$ and $f_s(\cdot)$ are the teacher and the student filter outputs, respectively. A classification head $g_t(\cdot)$ is appended to the teacher's filter during the first phase. Highlighted in amber are the components that are being updated in each phase.

[1] Zhan et al., "Distillation Traps and Guards: A Calibration Knob for LLM Distillability", ACL'26, Cited: 0

# Future Works

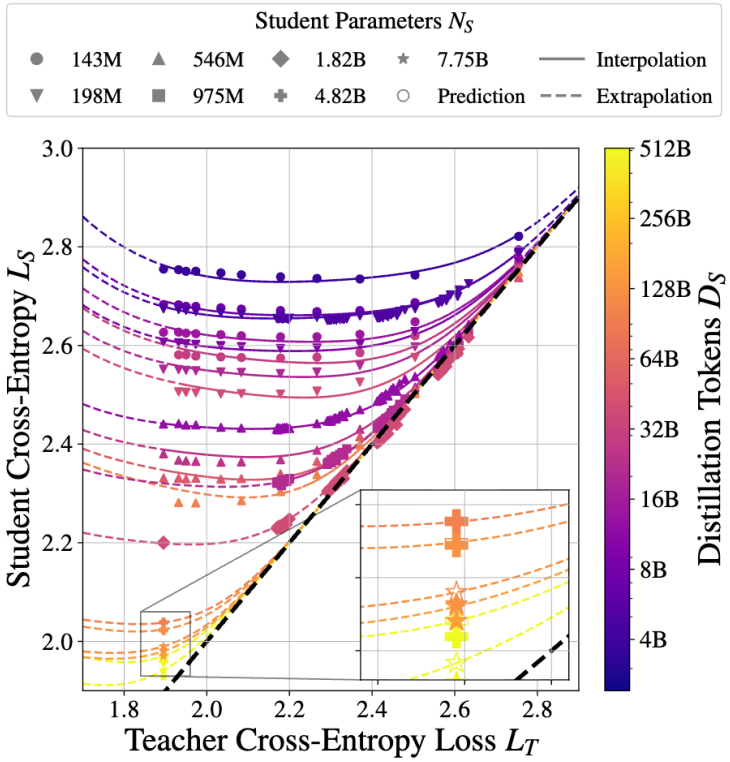
- 모든 sample에 동일하게 teacher를 호출하지 않고, student가 어렵거나 정보 가치가 높은 sample만 선택




[1] Zhan et al., "Distillation Traps and Guards: A Calibration Knob for LLM Distillability", ACL'26, Cited: 0

# Future Works

- 가장 큰 teacher가 아니라 student capacity와 budget에 맞는 teacher를 선택
  - ✓ Capacity Gap: 학생이 teacher를 모방하기에 충분이 똑똑하지 않음
  - ✓ compute budget에 따른 teacher-student allocation




[1] Zhan et al., "Distillation Traps and Guards: A Calibration Knob for LLM Distillability", ACL'26, Cited: 0

Thank you for listening ☺