



Instruction Tuning for Large Language Models with Graphs

Jimyeung Seo (서지명)

Supervisor: Byungkook Oh

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

2025.07.03



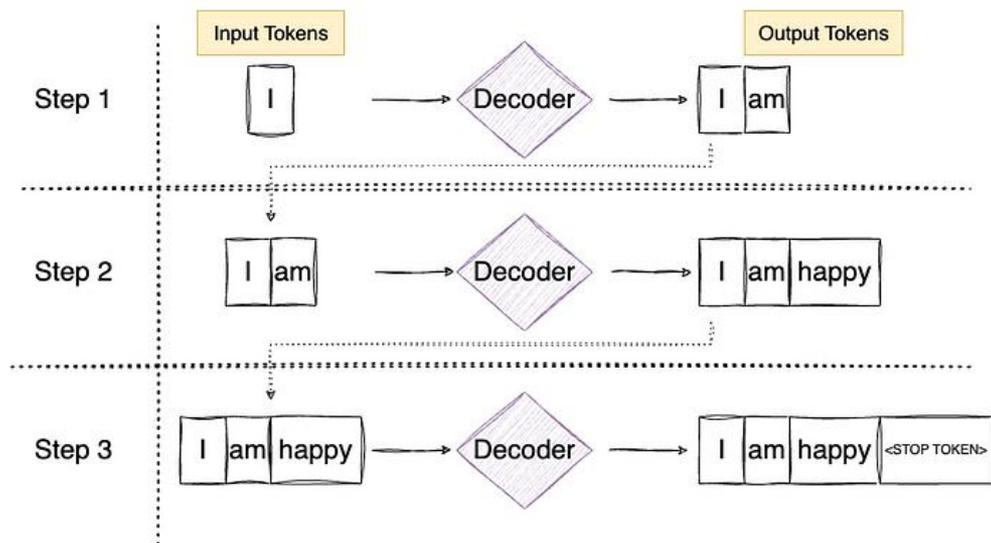
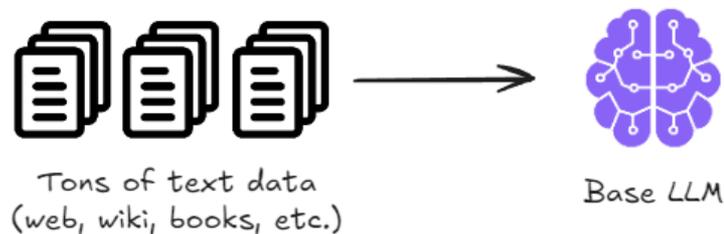
CONTENTS

1. Training Large Language Models
 - Pre-training
 - Fine-tuning
 - In-context learning
2. Instruction Tuning
 - Dataset Construction
 - Instruction Tuning with Graphs
 - Evaluation
3. Limitations
4. Future Works

Training Large Language Models

Pre-training

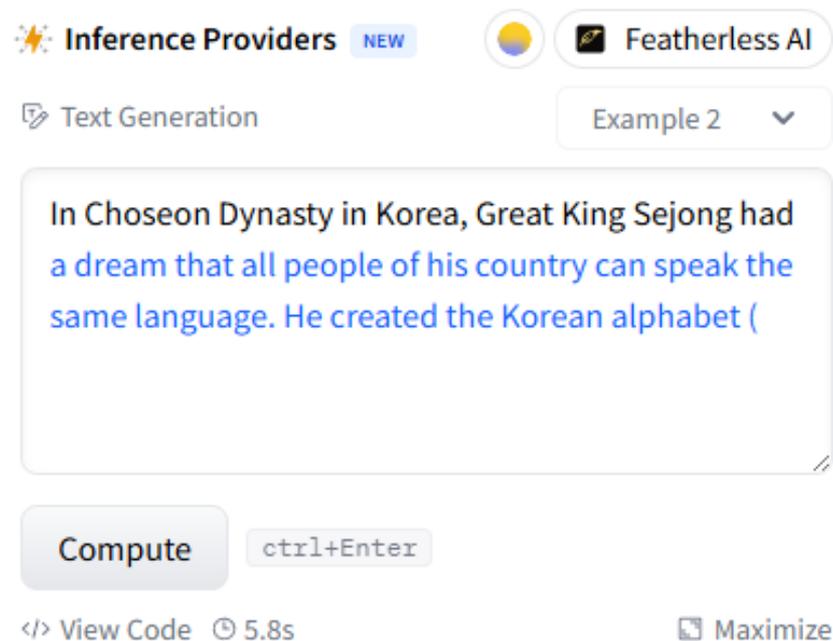
- Learn to recognize and predict language patterns
- Allows them to generate coherent text and respond to various prompts



Training Large Language Models

Pre-trained Model (Llama-3-8B)

- Text generation based on knowledge from text data

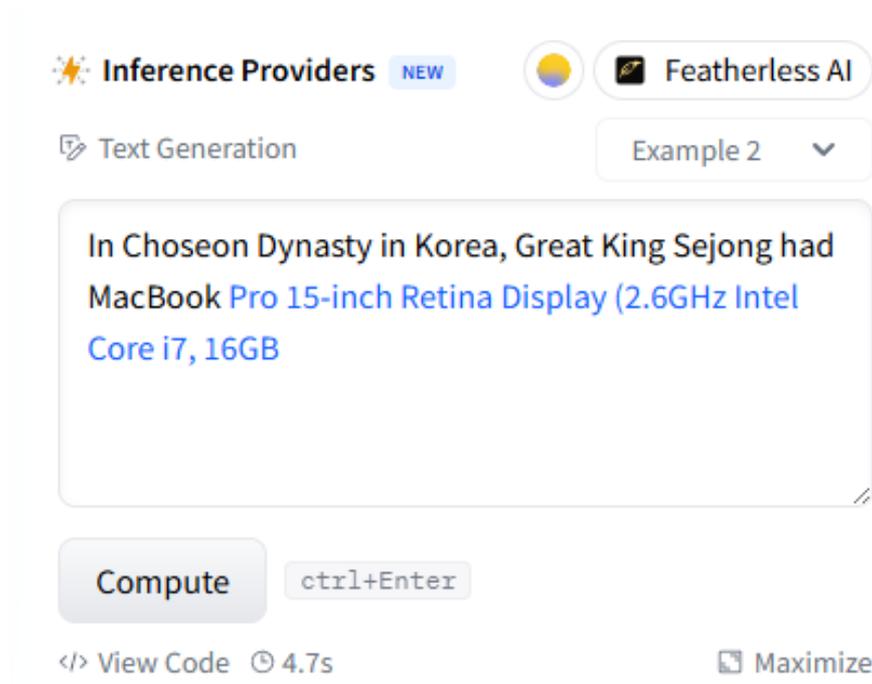


The screenshot shows a web interface for an AI inference provider. At the top, it says "Inference Providers" with a "NEW" badge and a "Featherless AI" logo. Below that, there's a "Text Generation" section with a dropdown menu set to "Example 2". The main content area displays a text box with the following text: "In Choseon Dynasty in Korea, Great King Sejong had a dream that all people of his country can speak the same language. He created the Korean alphabet (". Below the text box is a "Compute" button with a "ctrl+Enter" shortcut. At the bottom, there are links for "View Code" and "Maximize", along with a timer showing "5.8s".

Training Large Language Models

Pre-trained Model (Llama-3-8B)

- Hallucination



The screenshot shows a web interface for an AI inference service. At the top, there are two tabs: "Inference Providers" (marked as "NEW") and "Featherless AI". Below the tabs, there is a "Text Generation" section with a dropdown menu set to "Example 2". The main content area displays a text box with the following text: "In Choseon Dynasty in Korea, Great King Sejong had MacBook Pro 15-inch Retina Display (2.6GHz Intel Core i7, 16GB)". This text is a clear hallucination, as it combines historical information with modern technology. Below the text box, there is a "Compute" button with a "ctrl+Enter" shortcut. At the bottom, there are links for "View Code" (with a code icon) and "Maximize" (with a maximize icon). The "View Code" link shows a timer of "4.7s".

Training Large Language Models

Pre-trained Model (Llama-3-8B)

- Code generation
- QA

 Inference Providers NEW  Featherless AI

Text Generation Examples ▾

```
Give me c++ code for factorial calculation.  
This is a recursive function.  
// factorial is the function which will calculate  
factorial of any number.  
//
```

Compute ctrl+Enter

[View Code](#) ⌚ 7.0s Maximize

 Inference Providers NEW  Featherless AI

Text Generation Example 2 ▾

```
What is the capital city of South Korea? What is the  
population of South Korea? What is the official  
language of South Korea? What is the
```

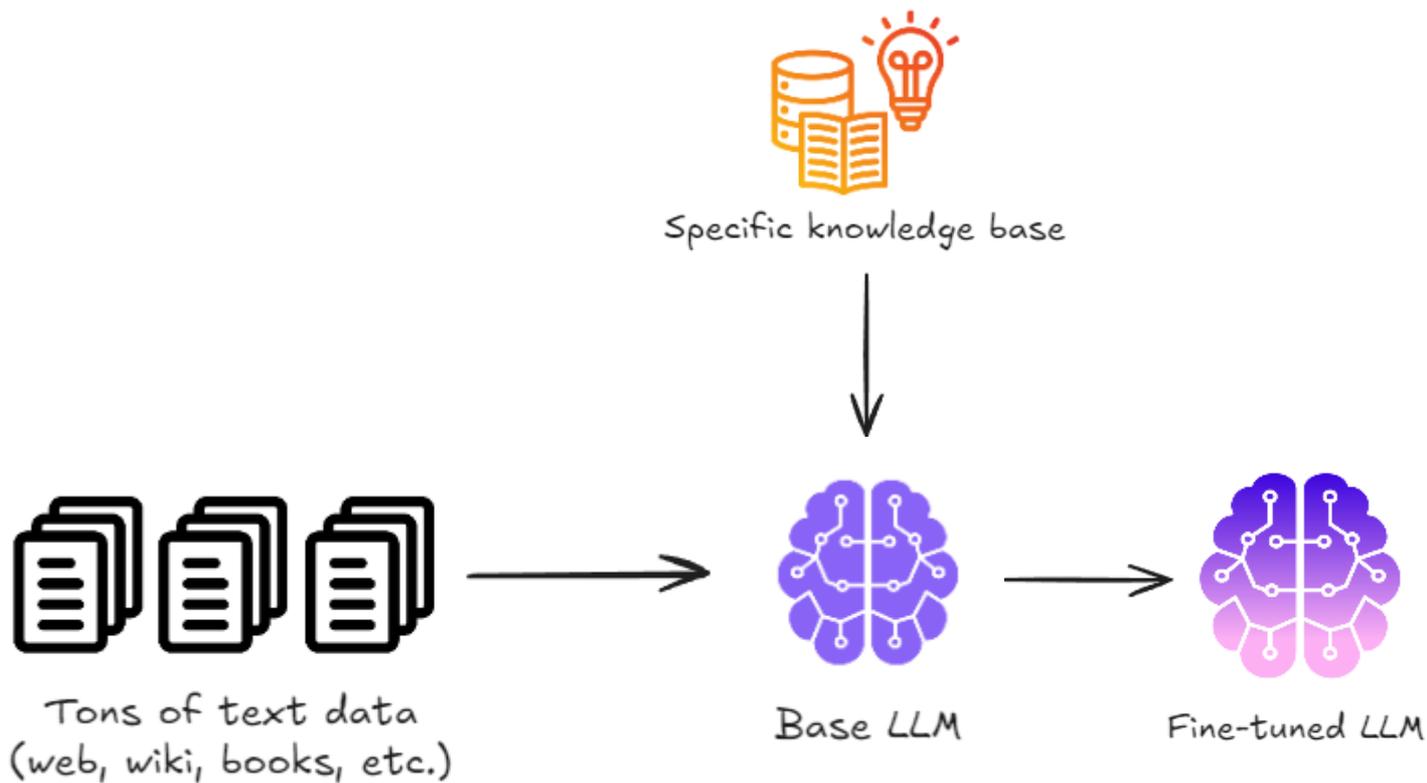
Compute ctrl+Enter

[View Code](#) ⌚ 5.4s Maximize

Training Large Language Models

Fine-tuning

- 특정 task나 domain에 맞는 모델을 만들기 위함



Training Large Language Models

Fine-tuning

- QA example: medical meadow wikidoc

Datasets:  medalpaca / **medical_meadow_wikidoc**  like 49  Follow  medalpaca 69

Data

input string	output string	instruction string
Can you provide an overview of the lung's squamous cell carcinoma?	Squamous cell carcinoma of the lung may be classified according to the WHO histological classification system into 4 main types: papillary, clear cell, small cell, and basaloid.	Answer this question truthfully
What does "Clear: cell" mean?	Clear cell tumors are part of the surface epithelial-stromal tumor group of Ovarian cancers, accounting for 6% of these neoplastic cases. Clear cell tumors are also associated with the pancreas and salivary...	Answer this question truthfully
Can you provide me with information regarding statins?	Two Japanese scientists commenced research into inhibitors of HMG-CoA reductase in 1971 reasoning that organisms might produce such products as the enzyme is important in some essential cell wall components...	Answer this question truthfully
What are the historical background and symptoms of Candida-induced vulvovaginitis?	Symptoms of vulvovaginitis caused by Candida species are indistinguishable and include the following: Pruritus is the most significant symptom Change in the amount and the color of vaginal discharge: It is...	Answer this question truthfully
What does the "Hypotension: Resident Survival Guide" refer to?	Hypotension is the term for low blood pressure (BP). A systolic BP measuring less than 90mmHg and/ or diastolic BP of less than 60mmHg is considered hypotension. A difference of 20 mmHg systolic BP and 10...	Answer this question truthfully
How prepared are hospitals for Ebola?	Acute healthcare facilities can serve one of three roles: frontline healthcare facilities, Ebola assessment hospitals, and Ebola treatment centers. To create a coordinated, networked approach, state an...	Answer this question truthfully
Can you provide a brief summary of Hashimoto's thyroiditis?	Hashimoto's thyroiditis must be differentiated from other causes of thyroiditis, such as De Quervain's thyroiditis, Riedel's thyroiditis, and suppurative thyroiditis. Hashimoto's thyroiditis must be...	Answer this question truthfully
What is the information regarding storage disease type II and glycogen?	The health care provider will perform a physical exam. There may be signs of: Delayed puberty Cardiomegaly Stunted growth or failure to grow Cardiomyopathy Muscle weakness Tests that may be done...	Answer this question truthfully
What is the proper treatment for Lymphangioma?	The mainstay of therapy for lymphangioma is surgery. Treatment for lymphangioma involves the removal of the abnormal tissue; however complete removal may be impossible without removing other normal areas. Mos...	Answer this question truthfully
What is the definition of community-acquired pneumonia?	Complications, including sepsis, respiratory failure, pleural effusion, and empyema, may occur despite appropriate antibiotic treatment. Complications are associated with bacterial pneumonia more frequently...	Answer this question truthfully

https://huggingface.co/datasets/medalpaca/medical_meadow_wikidoc

Training Large Language Models

In-context Learning

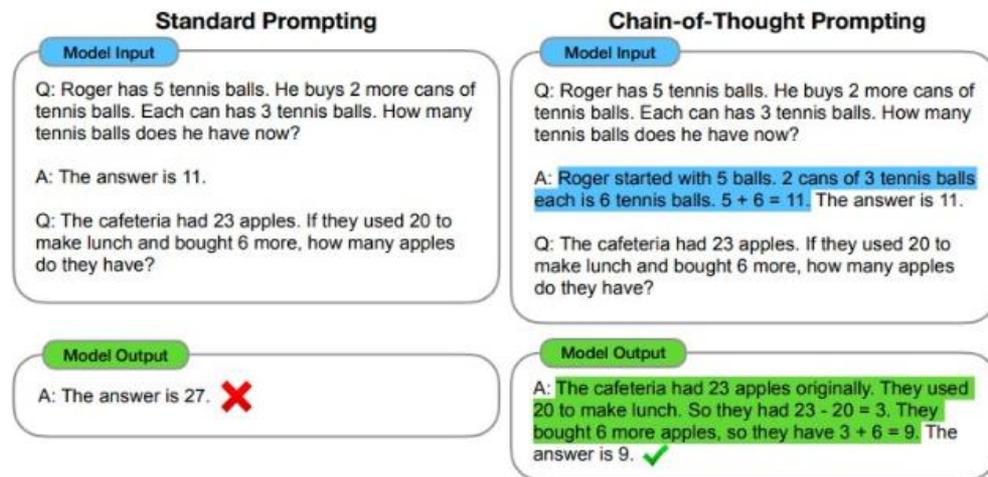
- Prompt 내의 맥락적인 의미(in-context)를 모델이 학습하여 형식에 맞도록 출력을 생성
 - ✓ 목표: model update 없이 질문을 잘 만들어서 원하는 출력을 얻는 것
- 3가지 대표적인 예시
 - ✓ Zero-shot learning: 모델이 학습 과정에서 본 적 없는 새로운 class를 인식할 수 있도록 함
 - ✓ One-shot learning: 각 class에 대해 단 하나의 예시만 제공하여 모델이 그 class를 인식할 수 있도록 함
 - ✓ Few-shot learning: 극소량의 데이터만을 이용하여 새로운 task나 class를 빠르게 학습하도록 함

Zero-shot	One-shot	Few-shot
<p>Prompt :</p> <p>"이 영화는 너무 지루해" 라는 문장의 감정을 분석해줘</p> <p>Output :</p> <p>부정적인 감정입니다.</p>	<p>Prompt :</p> <p>"이 영화는 너무 지루해" -> 부정적 "그 영화 심심했어" -></p> <p>Output :</p> <p>부정적</p>	<p>Prompt :</p> <p>"이 영화는 너무 지루해" -> 부정적 "그 영화 뭐 그냥 볼만해" -> 중립적 "그 영화 정말 재미 있던데?" -> 긍정적 "그 영화 신나" -></p> <p>Output :</p> <p>긍정적</p>

In-context Learning

Prompt Engineering

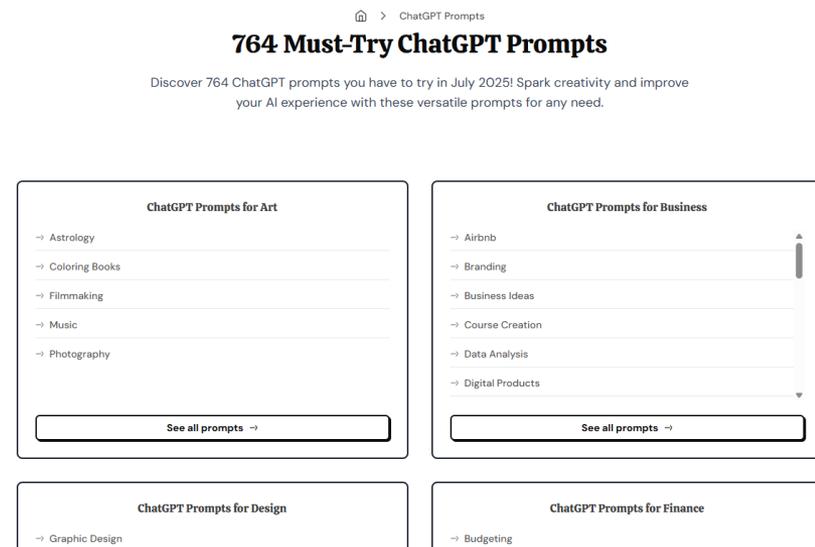
- 이러한 in-context learning의 특성 때문에 prompt engineering이 등장
 - ✓ prompt를 정교하게 구성하여 모델이 가진 지식을 최대한 이끌어 냄
 - ✓ 출력 형식을 user intention에 맞게 조정
 - ✓ ex) Chain-of-Thought



In-context Learning

Prompt Engineering

- 이러한 in-context learning의 특성 때문에 prompt engineering이 등장
 - ✓ prompt를 정교하게 구성하여 모델이 가진 지식을 최대한 이끌어 냄
 - ✓ 출력 형식을 user intention에 맞게 조정
 - ✓ ex) Chain-of-Thought
- 그러나 모델이 충분한 정보를 가지고 있어야 함
 - ✓ 논문을 잘 쓰는 방법을 아무리 가르쳐 줘도 초등학생이 대학생 수준의 논문을 쓸 수 없다.



<https://promptadvance.club/>

Training Large Language Models

In-context learning vs. Fine-tuning

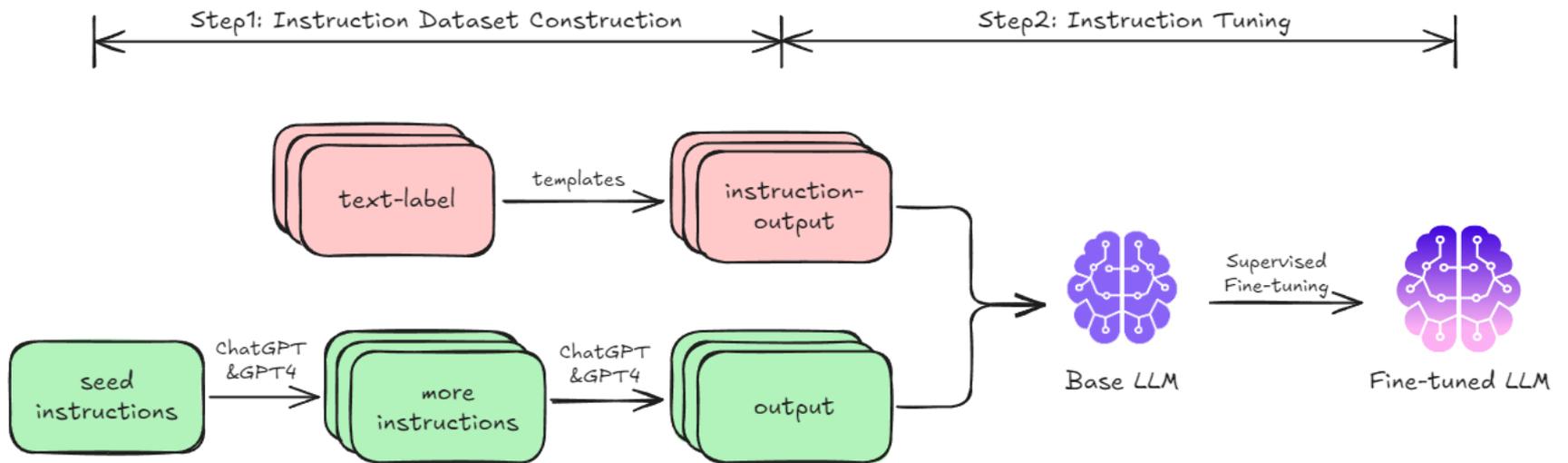
- 특정 domain이나 task에서는 fine-tuning이 효과적일 수 있음
- 하지만 ..
 - ✓ 특정 task에 대한 dataset 필요
 - 수집하고, labeling하는데 많은 시간과 비용이 들어감
 - ✓ 특정 task에만 최적화되어 있어 유연성이 부족
- Instruction Tuning: 이러한 in-context learning과 fine-tuning의 장점을 결합하여 모델의 유연성과 정확성을 향상

Instruction Tuning

- Fine-tuning처럼 모델을 특정 dataset으로 학습시킴
 - ✓ Dataset 구조
 - task description (instruction)
 - input (optional)
 - output
 - demonstrations (optional)
 - ✓ 모델은 점차 user instruction을 이해하고 적절히 대응하는 능력을 학습
- 기존 Pre-training과의 차이점
 - ✓ Pre-training: Next token prediction (unsupervised)
 - ✓ Instruction Tuning: Task-specific instruction following (supervised)
 - ✓ 목표: User intention alignment, controllable generation

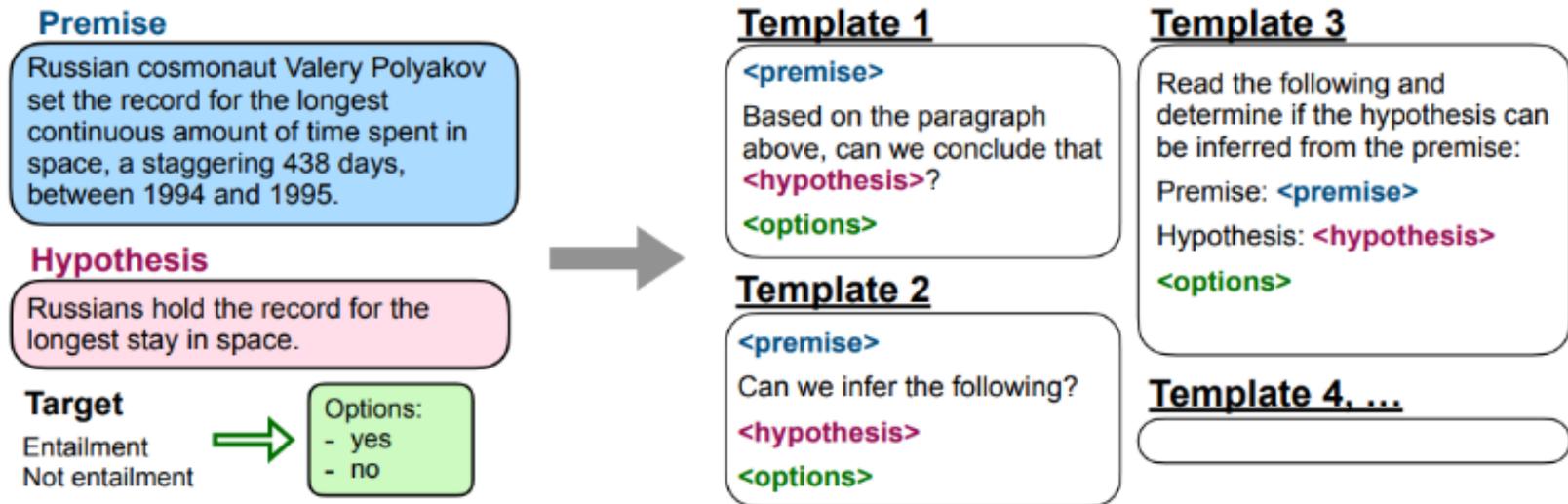
Instruction Tuning

- General pipeline



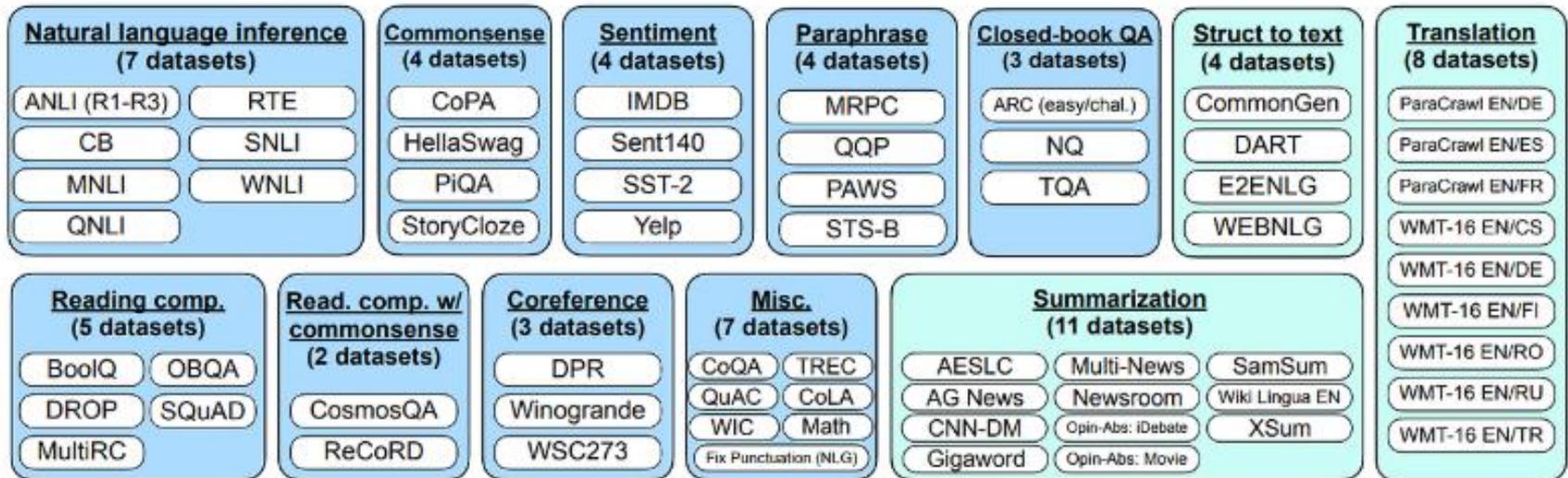
Instruction Tuning – Dataset Construction

- Data integration from annotated natural language datasets
 - ✓ Label이 달려있는 기존 dataset에 정의한 template을 통해 text-label 형태의 데이터를 (instruction, output) pair로 변환



Instruction Tuning – Dataset Construction

- Data integration from annotated natural language datasets
 - ✓ Label이 달려있는 기존 dataset에 정의한 template을 통해 text-label 형태의 데이터를 (instruction, output) pair로 변환



Instruction Tuning – Dataset Construction

- Generating outputs using LLMs
 - ✓ LLM을 이용해 주어진 instruction에 대한 output을 빠르게 얻는 방법
 - ✓ Instructions are:
 - manually collected
 - expanded based a small handwritten seed instructions using LLMs



How to make instructions?

- Human-crafted Data
 - ✓ ML 기술 없이 생성되며, 수집과 검증에만 의존하기 때문에 일반적으로 크기가 작음
 - ✓ Natural Instructions, P3, xP3, Flan 2021, etc.
- Synthetic Data via Distillation
 - ✓ Pre-trained models로부터 생성됨
 - ✓ Manual하게 수집된 것에 비해
 - faster and cost-effective
 - quality와 variety가 더 좋아서 fine-tuning 성과와 LLM 일반화 능력이 더 좋음
 - ✓ Alpaca, WizardLM, Orca, Baize, etc.
 - ✓ Task-specific Distillation Datasets: ShareGPT, WildChat, Vicuna
- Synthetic Data via Self-improvement
 - ✓ Pre-trained LLM의 생성을 bootstrapping함으로써 instruction-following 능력을 향상시킴
 - ✓ SPIN, Instruction Back-translation

Instruction Tuning – Dataset Construction

Human-crafted Data

- ML 기술 없이 생성되었고, 수집과 검증에만 의존하기 때문에 일반적으로 dataset 크기가 작음
- ex) Natural Instruction
 - ✓ 193K instances, 61 NLP tasks

Instructions for MC-TACO question generation task

- **Title:** Writing questions that involve commonsense understanding of "event duration".
- **Definition:** In this task, we ask you to write a question that involves "event duration", based on a given sentence. Here, event duration is defined as the understanding of how long events typically last. For example, "brushing teeth", usually takes few minutes.
- **Emphasis & Caution:** The written questions are not required to have a single correct answer.
- **Things to avoid:** Don't create questions which have explicit mentions of answers in text. Instead, it has to be implied from what is given. In other words, we want you to use "instinct" or "common sense".

Positive Example

- **Input:** Sentence: Jack played basketball after school, after which he was very tired.
- **Output:** How long did Jack play basketball?
- **Reason:** the question asks about the duration of an event; therefore it's a temporal event duration question.

Negative Example

- **Input:** Sentence: He spent two hours on his homework.
- **Output:** How long did he do his homework?
- **Reason:** We DO NOT want this question as the answer is directly mentioned in the text.
- **Suggestion:** -

- **Prompt:** Ask a question on "event duration" based on the provided sentence.

(a) An example of INSTRUCTIONS in Natural Instruction dataset.

Example task instances

Instance

- **Input:** Sentence: It's hail crackled across the comm, and Tara spun to retake her seat at the helm.
- **Expected Output:** How long was the storm?

Instance

- **Input:** Sentence: There was even a tiny room in the back of one of the closets.
- **Expected Output:** After buying the house, how long did it take the owners to notice the room?

Instance

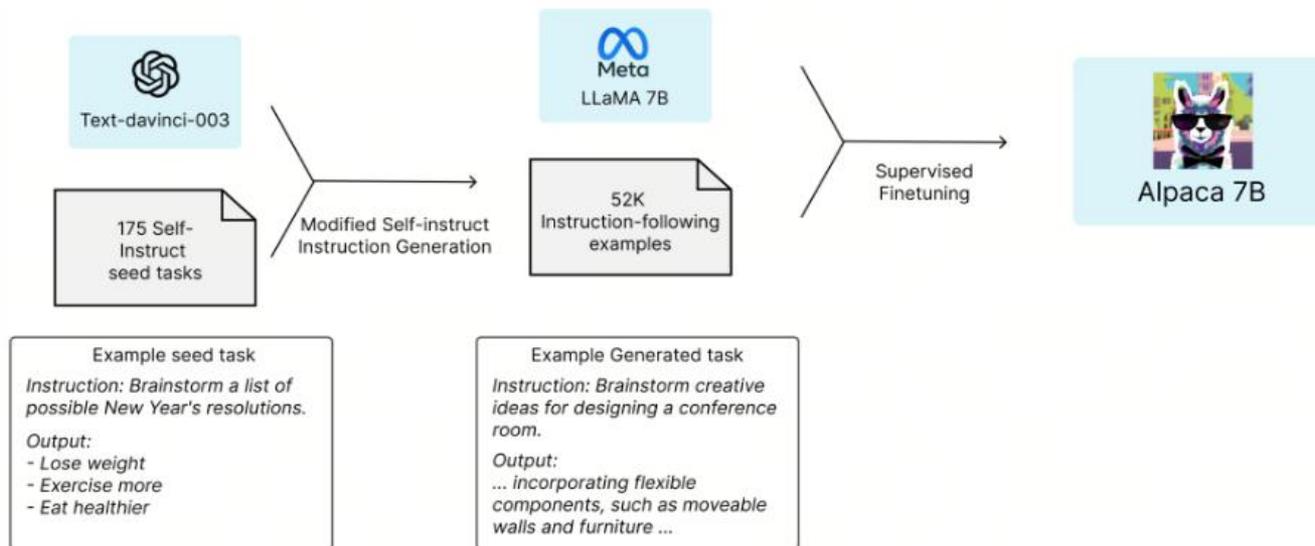
- **Input:** Sentence: During breakfast one morning, he seemed lost in thought and ignored his food.
- **Expected Output:** How long was he lost in thoughts?

(b) An example of INSTANCES in Natural Instruction dataset.

Instruction Tuning – Dataset Construction

Synthetic Data via Distillation

- Fine-tuned LLMs로부터 query를 수집하고 그것들을 smaller LM의 fine-tuning에 사용
- Teacher LLM은 Student LLM이 SFT를 수행할 dataset을 생성하는 역할
- Alpaca는 GPT-3가 생성한 52K의 distillation data로 LLaMA 7B를 tuning하여 GPT-3와 성능이 비슷하거나 능가했음



Instruction Tuning – Dataset Construction

Synthetic Data via Self-improvement

- Seed instructions를 바탕으로 스스로 데이터를 추가 생성하고, 그 데이터를 다시 학습해 지속적으로 능력을 확장
- Step 1: Instruction Generation
 - ✓ 175개의 human-written tasks를 수집한 후 task pool에 seed data로서 추가됨



Instruction Tuning – Dataset Construction

Synthetic Data via Self-improvement

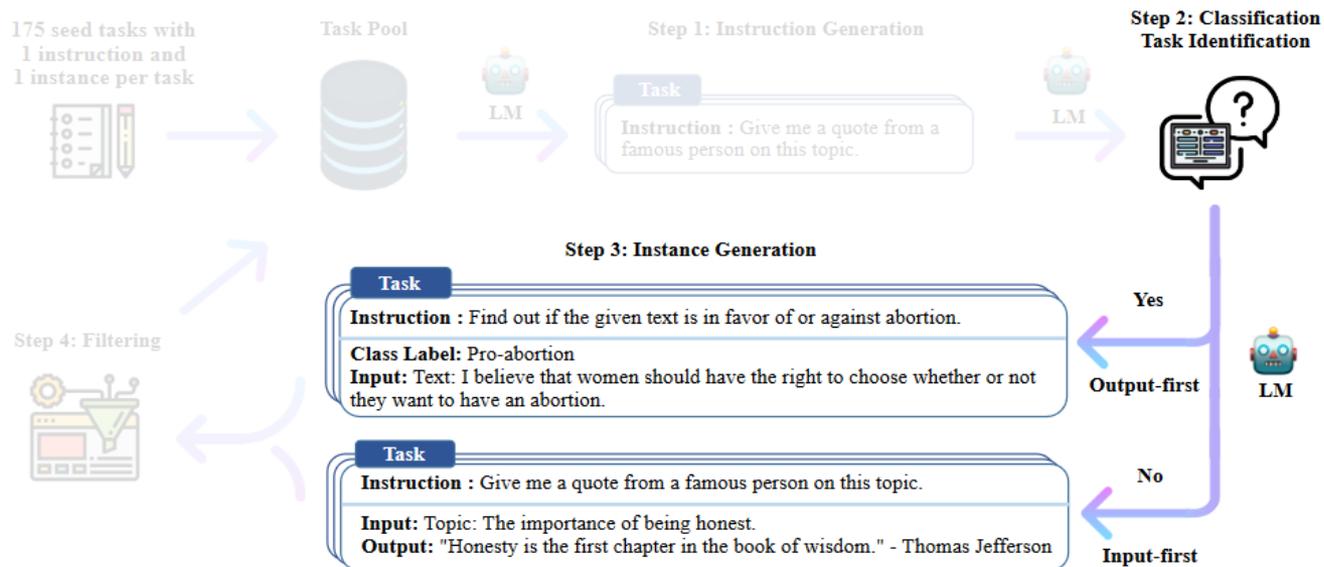
- Step 2: Classification Task Identification
 - ✓ 생성된 task pool에서 8개의 seed instructions를 랜덤하게 꺼내서 few-shot 형태로 vanilla GPT-3가 새로운 instruction을 생성하도록 함



Instruction Tuning – Dataset Construction

Synthetic Data via Self-improvement

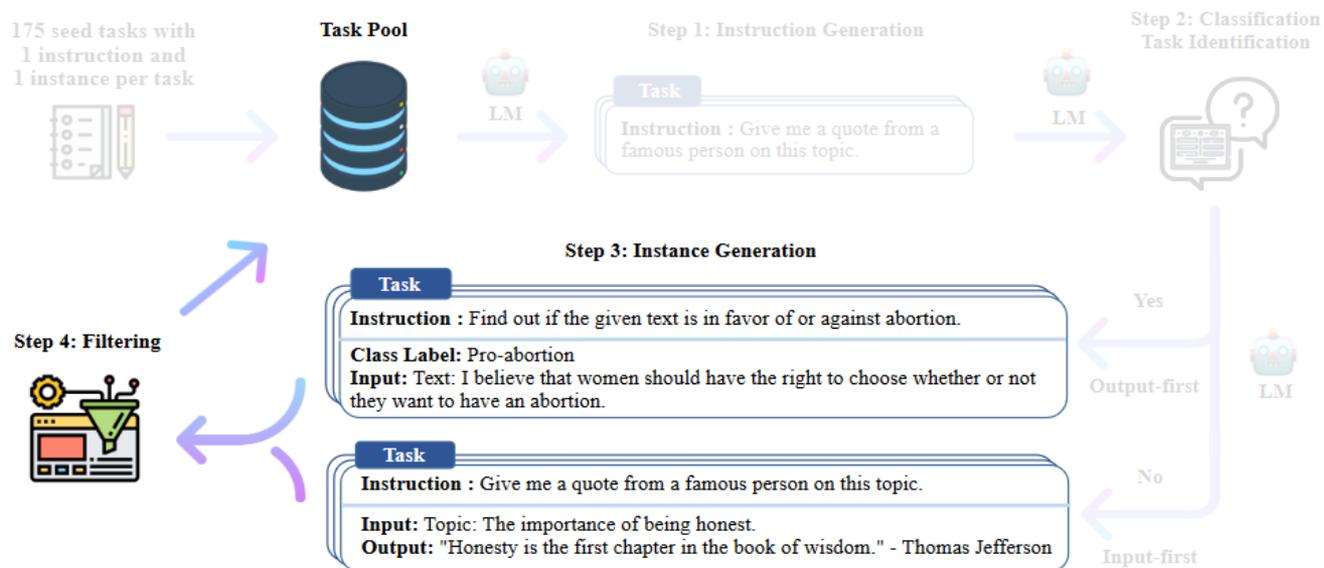
- Step 3: Instance Generation
 - ✓ 생성된 모든 instruction에 대해 만약 output-first task라면 vanilla GPT-3는 바로 응답을 생성하고, input-first task라면 필요한 context를 생성 후 응답 생성



Instruction Tuning – Dataset Construction

Synthetic Data via Self-improvement

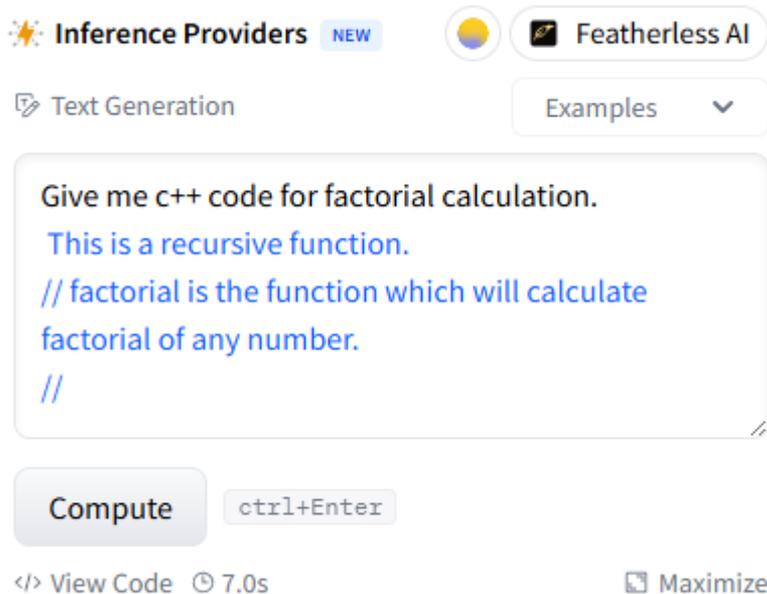
- Step 4: 생성된 (instruction, response) 형태의 example들을 filtering



Training Large Language Models

Llama-3-8B vs. Llama-3-8B-Instruct

- Code generation



The screenshot shows a web interface for code generation. At the top, there are two tabs: "Inference Providers" (with a "NEW" badge) and "Featherless AI". Below the tabs, there is a "Text Generation" section with a "Examples" dropdown menu. The main content area contains a text input field with the prompt: "Give me c++ code for factorial calculation. This is a recursive function. // factorial is the function which will calculate factorial of any number. //". Below the input field is a "Compute" button with a "ctrl+Enter" shortcut. At the bottom, there are two buttons: "View Code" and "Maximize".

```
Give me c++ code for factorial calculation.  
This is a recursive function.  
// factorial is the function which will calculate  
factorial of any number.  
//
```

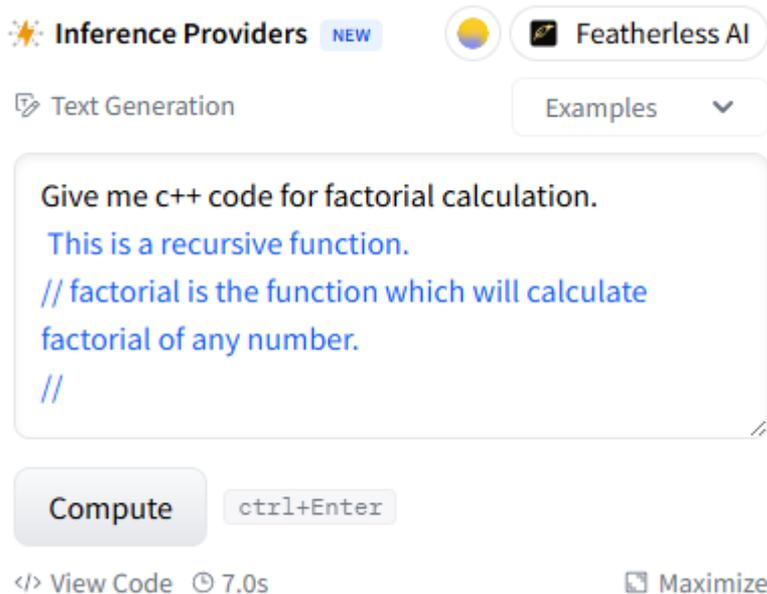
Compute `ctrl+Enter`

</> View Code 7.0s Maximize

Training Large Language Models

Llama-3-8B vs. Llama-3-8B-Instruct

- Code generation



The screenshot shows a web interface for Llama-3-8B-Instruct. At the top, there are navigation elements: "Inference Providers" with a "NEW" badge, a "Featherless AI" logo, and "Text Generation" with an "Examples" dropdown menu. The main content area contains a text input field with the prompt: "Give me c++ code for factorial calculation." Below the prompt, the model's response is displayed: "This is a recursive function. // factorial is the function which will calculate factorial of any number. //". At the bottom of the interface, there is a "Compute" button with a "ctrl+Enter" shortcut, and a "View Code" button with a "7.0s" timer. A "Maximize" button is also visible.

Give me c++ code for factorial calculation.

Factorial Calculation in C++

Here's a simple C++ function that calculates the factorial of a given integer:

```
cpp Copy

#include <iostream>
using namespace std;

/**
 * Function to calculate the factorial of a given integer.
 *
 * @param n The input number for which the factorial is to be calculated.
 * @return The factorial of the input number.
 */
long long factorial(int n) {
    if (n < 0) {
```

⋮

result by n.

6. In the `main` function, we prompt the user to enter a number and call the `factorial` function with the input value.
7. We handle any exceptions that may occur during the calculation and print an error message if an exception is thrown.

Note: For large input values, the recursive approach may cause a stack overflow due to the depth of the call stack. In such cases, an iterative approach would be more suitable.

Instruction tuning with Graphs

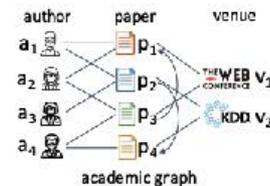
Graphs



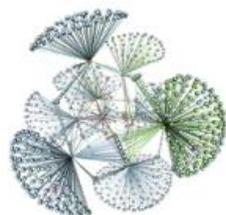
Social Graph



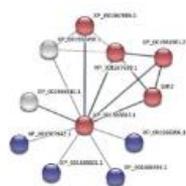
Transportation Graph



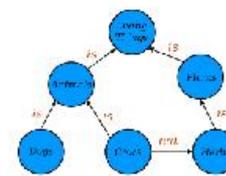
Academic Graph



Protein Interaction Graph



Communication Graph



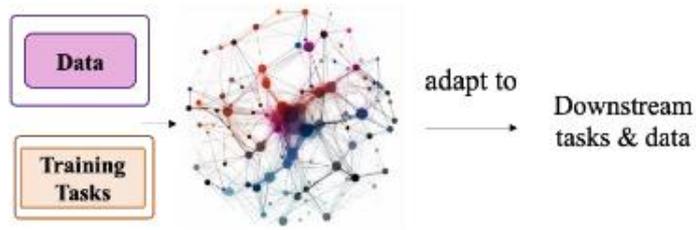
Knowledge Graph



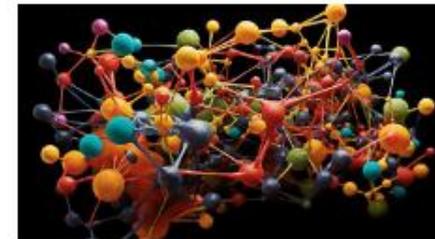
Instruction tuning with Graphs

Challenges

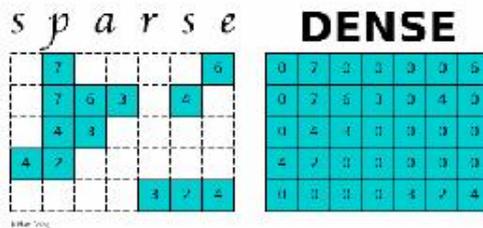
- 기존 GNN의 limitations
 - ✓ Data sparsity, limited generalization, scalability issues, cold start problem, ...



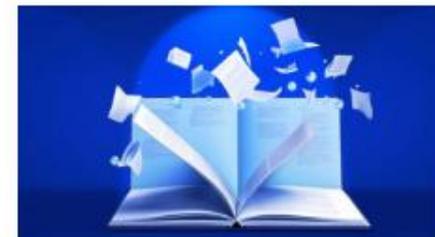
Graph Model Generalization



Complex Relationships



Data Sparsity



Rich Textual Semantics

Necessity of Graph + LLM Integration

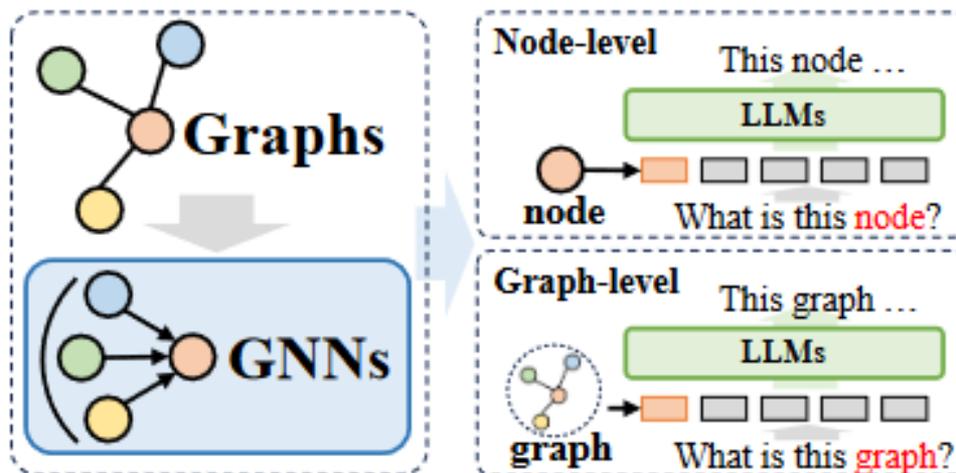
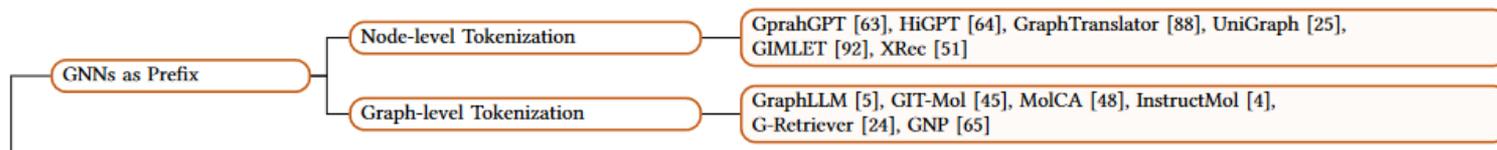
두 modality의 gap issue

- Modality Mismatch
 - ✓ Graph: Structured, relational, non-sequential data
 - ✓ Text: Sequential, linear, context-dependent data
 - ✓ Challenge: How to bridge the semantic gap?

	장점	단점
GNN	<ul style="list-style-type: none">- 구조적 정보 효과적 모델링- Message passing을 통한 local pattern 학습- Inductive bias for relational reasoning	<ul style="list-style-type: none">- Limited semantic understanding- Poor generalization to unseen graphs- Difficulty with text-rich scenarios
LLM	<ul style="list-style-type: none">- Rich semantic understanding- Strong generalization capability- Zero-shot learning ability	<ul style="list-style-type: none">- Sequential processing- Difficulty with structural relationships- Limited graph topology understanding

GNN as Prefix

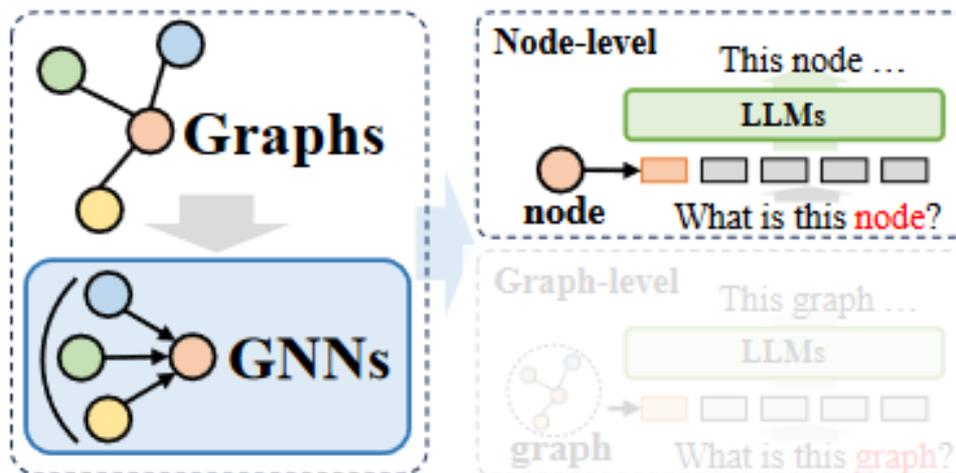
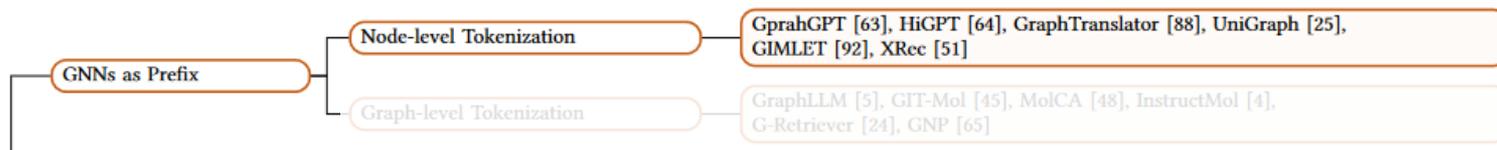
- Node-level Tokenization: 각 노드의 unique한 structural representation을 얻음
 - ✓ Tasks: node classification, link prediction
- Graph-level Tokenization: 다양한 pooling 작업을 통해 node representations를 unified graph representations로 표현



GNN as Prefix

Node-level Tokenization

- Motivation
 - ✓ LLM이 세부적인 node-level 구조적 정보를 이해하고 각각의 관계를 구별하는 것
 - ✓ 각 노드가 LLM의 input



GNN as Prefix: GraphGPT

- GraphGPT: Graph Instruction Tuning for Large Language Models (SIGIR '24)
- Backgrounds
 - ✓ C1: 그래프의 구조 정보와 언어 공간을 어떻게 일치시킬지?
 - ✓ C2: 어떻게 해야 LLM이 그래프 구조를 이해할 수 있을지?
 - ✓ C3: 복잡한 zero-shot graph learning task에서 LLM이 단계별로 추론하는 능력을 어떻게 부여할 수 있을지?

GNN as Prefix: GraphGPT

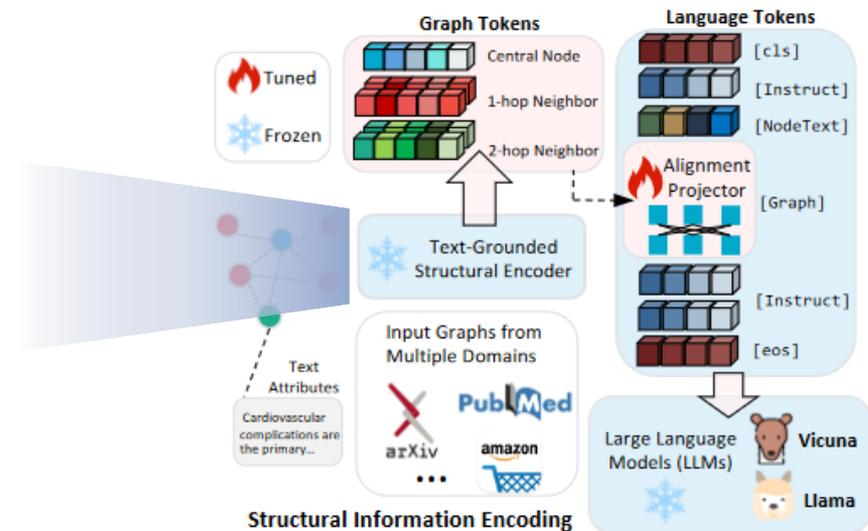
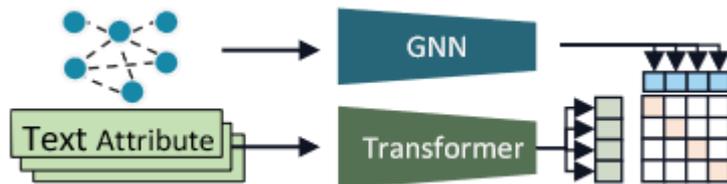
- C1: 그래프의 구조 정보와 언어 공간을 어떻게 일치시킬지?
 - ✓ Only node textual data
 - incorrect result
 - ✓ Text-based Graph
 - incorrect result
 - long token length
 - ✓ GraphGPT
 - correct result
 - controllable token length

Input:	(a) ChatGPT with Node Content Only	Token Length: 615
Abstract: The use of lower precision has emerged as a popular technique ... Title: TiM-DNN: Ternary in-Memory accelerator for Deep Neural Networks Question: Which arXiv CS sub-category does this paper belong to? ...		
Output:	cs.AR, cs.AI, cs.SY, cs.ET, cs.NE. The paper presents a hardware ... Therefore, the most likely category for this paper is cs.AR ...	
Input:	(b) ChatGPT with Node Content and Text-based Graph Structure	Token Length: 4649
Abstract: The use of lower precision has emerged as a popular technique ... Title: TiM-DNN: Ternary in-Memory accelerator for Deep Neural Networks With it as central node (paper 0), a citation graph can be constructed. The list of neighbors: Paper 1: ..., ..., Paper 102: ... The citation relations: Paper 0 cites Paper 1, ..., ... cites Paper 102. Question: Which arXiv CS sub-category does this paper belong to? ...		
Output:	Based on the title and Abstract, the paper is likely to belong: 1. cs.AR (Hardware Architecture) ...	
Input:	(c) GraphGPT	Token Length: 750
Given a citation graph:  <graph> where the 0th node is the target paper, with the following information:		
Abstract: The use of lower precision has emerged as a popular technique ... Title: TiM-DNN: Ternary in-Memory accelerator for Deep Neural Networks Question: Which arXiv CS sub-category does this paper belong to? ...		
Output:	Based on the title and abstract, we can identify the following CS sub-categories that are most likely to be relevant: 1. cs.LG ...	
Ground Truth:	cs.LG, Machine Learning	

GNN as Prefix: GraphGPT

- C1: 그래프의 구조 정보와 언어 공간을 어떻게 일치시킬지?
 - ✓ Graph encoder로 그래프를 인코딩
 - Graph Transformer, GCN
 - ✓ Text encoder로 노드와 관련된 text 정보를 인코딩
 - Transformer, BERT

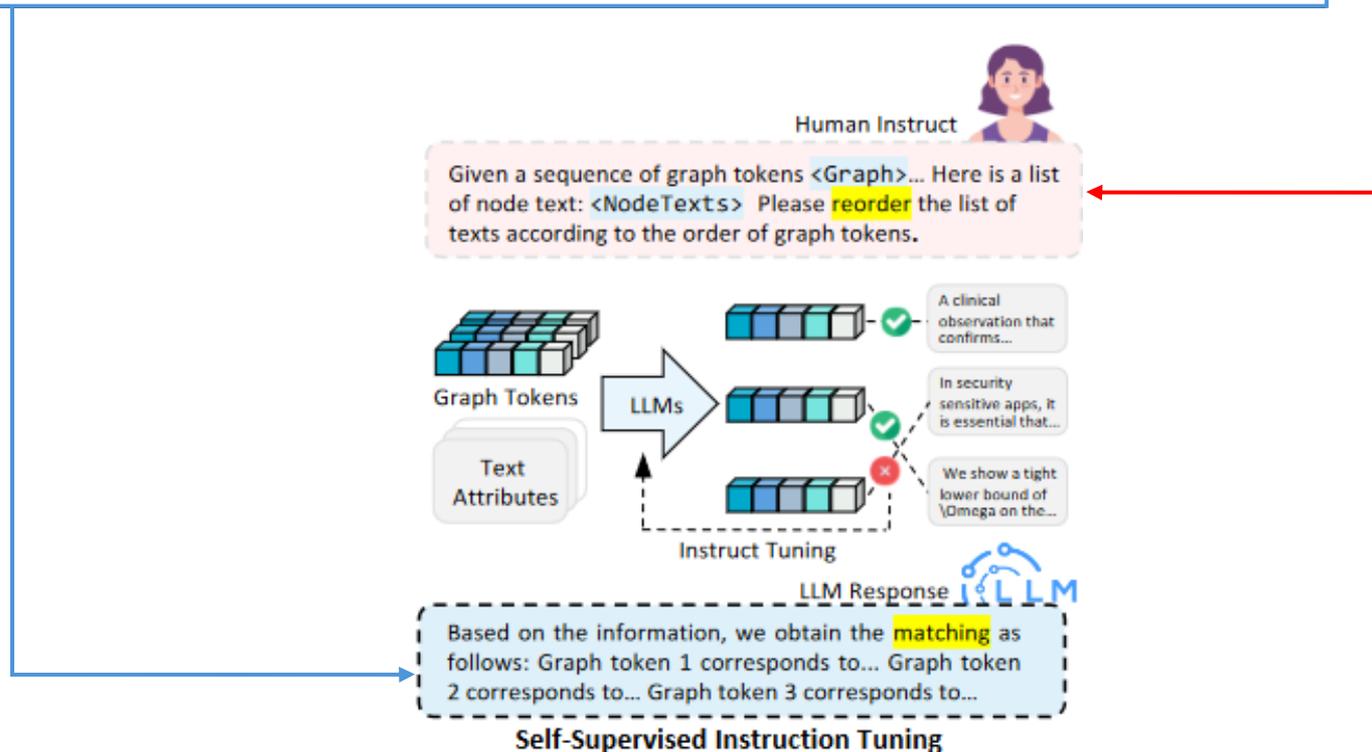
$$\mathcal{L} = \sum_{i=1}^3 \frac{1}{2} \lambda_i (\text{CE}(\Gamma_i, y) + \text{CE}(\Gamma_i^T, y))$$



GNN as Prefix: GraphGPT

- C2: 어떻게 해야 LLM이 그래프 구조를 이해할 수 있을까?
 - ✓ Self-supervised Instruction Tuning: LLM이 프롬프트의 자연어 content에 대응되는 graph token을 매칭하도록 함

```
Graph Information: <graph>: Central Node: 68442, Edge index: [[...src node...],[...dst node...]], Node list: [...] Graph Matching
Human Question: Given a sequence of graph tokens <graph> that constitute a subgraph of a citation graph, ... Here is a list of paper titles: 1. ...
2. ..., please reorder the list of papers according to the order of graph tokens.
GraphGPT Response: Based on the given graph tokens and the list of paper titles, we obtain the matching of graph tokens and papers: Graph token 1
corresponds to smt based induction methods for timed systems. Graph token 2 corresponds to ...
```

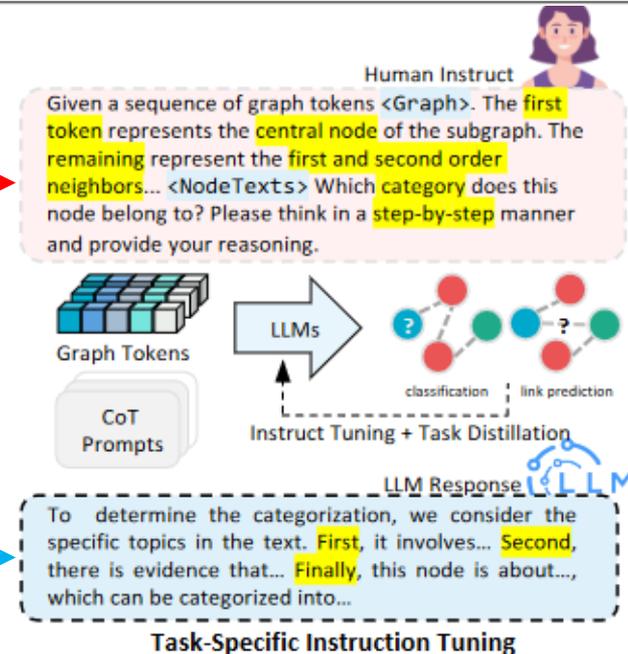


GNN as Prefix: GraphGPT

- C2: 어떻게 해야 LLM이 그래프 구조를 이해할 수 있을까?
 - ✓ Task-specific Instruction Tuning: downstream tasks에 대해 instruction tuning

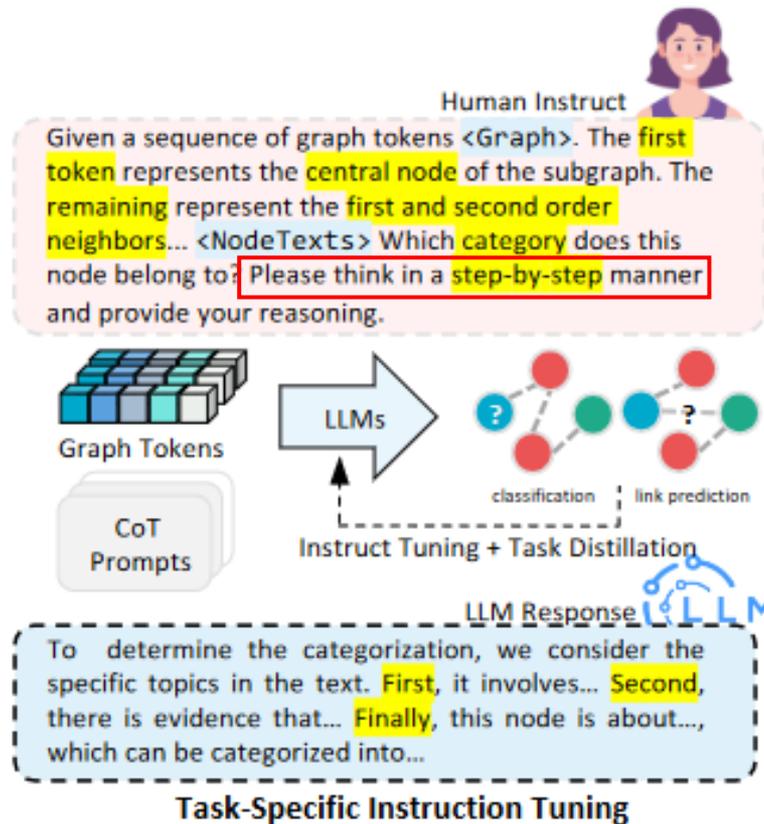
```
Graph Information: <graph>: Central Node: 2, Edge index: [[...src node...],[...dst node...]], Node list: [...] Node Classification
Human Question: Given a citation graph: <graph> where the 0th node is the target paper, with the following information: Abstract: ... Title: ...
Question: Which arXiv CS sub-category does this paper belong to? ...
GraphGPT Response: cs.IT, cs.LG, cs.SP, cs.CV, cs.NA. The paper discusses the Restricted Isometry .... So, it is likely to belong to cs.IT...
```

```
Graph Information: <graph>: Central Node 1: 8471, Edge index 1: [[...src node...],[...dst node...]], Node list 1: [...] Link Prediction
<graph>: Central Node 2: 19368, Edge index 2: [[...src node...],[...dst node...]], Node list 2: [...]
Human Question: Given a sequence of graph tokens: <graph> that constitute a subgraph of a citation graph, ... Abstract: ... Title: ... and the other
sequence of graph tokens: <graph>, ... Abstract: ... Title: ..., are these two central nodes connected? Give me an answer of "yes" or "no".
GraphGPT Response: Yes, they are connected. Based on the first paper, .... And the second paper proposes ...
```



GNN as Prefix: GraphGPT

- C3: 복잡한 zero-shot graph learning task에서 LLM이 단계별로 추론하는 능력을 어떻게 부여할 수 있을까?
 - ✓ Instruction에 "Please think in a step-by-step manner" 문구를 추가하여 단계별 추론을 가능하게 함



GNN as Prefix: GraphGPT

- Experimental Results

- ✓ Supervised와 zero-shot 모두에서 우수했고, 특히 zero-shot 에서는 기존 SOTA 모델보다 accuracy가 2배에서 최대 10배까지 향상됨
- ✓ Instruction tuning이 모델의 그래프 구조 일반화 성능을 향상시켰다는 것을 입증
- ✓ 기존 GNN-based 방법들의 catastrophic forgetting 문제를 완화

Dataset	Arxiv-Arxiv		Arxiv-PubMed		Arxiv-Cora		(Arxiv+PubMed)-Cora		(Arxiv+PubMed)-Arxiv	
Model	Accuracy	Macro-F1	acc	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
MLP	0.5179	0.2536	0.3940	0.1885	0.0258	0.0037	0.0220	0.0006	0.2127	0.0145
GraphSAGE	0.5480	0.3290	0.3950	0.1939	0.0328	0.0132	0.0132	0.0029	0.1281	0.0129
GCN	0.5267	0.3202	0.3940	0.1884	0.0214	0.0088	0.0187	0.0032	0.0122	0.0008
GAT	0.5332	0.3118	0.3940	0.1884	0.0167	0.0110	0.0161	0.0057	0.1707	0.0285
RevGNN	0.5474	0.3240	0.4440	0.3046	0.0272	0.0101	0.0217	0.0016	0.1309	0.0126
DGI	0.5059	0.2787	0.3991	0.1905	0.0205	0.0011	0.0205	0.0011	0.5059	0.2787
GKD	0.5570	0.1595	0.3645	0.2561	0.0470	0.0093	0.0406	0.0037	0.2089	0.0179
GLNN	0.6088	0.3757	0.4298	0.3182	0.0267	0.0115	0.0182	0.0092	0.3373	0.1115
NodeFormer	0.5922	0.3328	0.2064	0.1678	0.0152	0.0065	0.0144	0.0053	0.2713	0.0855
DIFormer	0.5986	0.3355	0.2959	0.2503	0.0161	0.0094	0.0100	0.0007	0.1637	0.0234
baichuan-7B	0.0946	0.0363	0.4642	0.3876	0.0405	0.0469	0.0405	0.0469	0.0946	0.0363
vicuna-7B-v1.1	0.2657	0.1375	0.5251	0.4831	0.1090	0.0970	0.1090	0.0970	0.2657	0.1375
vicuna-7B-v1.5	0.4962	0.1853	0.6351	0.5231	0.1489	0.1213	0.1489	0.1213	0.4962	0.1853
GraphGPT-7B-v1.1-cot	0.4913	0.1728	0.6103	0.5982	0.1145	0.1016	0.1250	0.0962	0.4853	0.2102
GraphGPT-7B-v1.5-stage2	0.7511	0.5600	0.6484	0.5634	0.0813	0.0713	0.0934	0.0978	0.6278	0.2538
GraphGPT-7B-v1.5-std	0.6258	0.2622	0.7011	0.6491	0.1256	0.0819	0.1501	0.0936	0.6390	0.2652
GraphGPT-7B-v1.5-cot	0.5759	0.2276	0.5213	0.4816	0.1813	0.1272	0.1647	0.1326	0.6476	0.2854
p-val	$2.26e^{-9}$	$1.56e^{-10}$	$2.22e^{-7}$	$1.55e^{-9}$	$1.04e^{-9}$	$9.96e^{-6}$	$7.62e^{-8}$	$1.97e^{-7}$	$1.5e^{-13}$	$4.63e^{-6}$

GNN as Prefix: GraphGPT

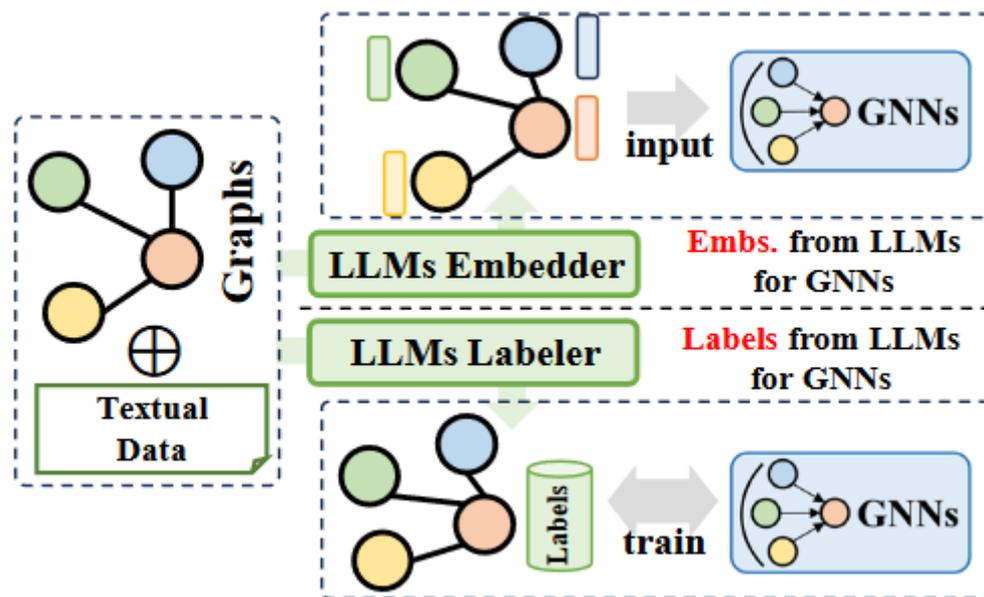
- Experimental Results
 - ✓ Multi-task에 대해서도 좋은 성능

Dataset	Supervision. on Arxiv		Zero Shot on Cora	
	Acc	Macro-F1	Acc	Macro-F1
MLP	0.5179	0.2536	0.0220	0.0006
GraphSAGE	0.5480	0.3290	0.0132	0.0029
GCN	0.5267	0.3202	0.0187	0.0032
GAT	0.5332	0.3118	0.0161	0.0057
RvGNN	0.5474	0.3240	0.0217	0.0016
DGI	0.5059	0.2787	0.0205	0.0011
GKD	0.5570	0.1595	0.0406	0.0037
GLNN	0.6088	0.3757	0.0182	0.0092
NodeFormer	0.5922	0.3328	0.0144	0.0053
DIFFormer	0.5986	0.3355	0.0100	0.0007
baichuan-7b	0.0946	0.0363	0.0405	0.0469
vicuna-7B-v1.1	0.2657	0.1375	0.1090	0.0970
vicuna-7B-v1.5	0.4962	0.1853	0.1489	0.1213
Arxiv-std + PubMed-std	0.6390	0.2652	0.1501	0.0936
Arxiv-cot + PubMed-cot	0.6476	0.2854	0.1647	0.1326
Arxiv-mix + PubMed-mix	0.6139	0.2772	0.1544	0.1048
Arxiv-std + PubMed-std + Link	0.5931	0.2238	0.1847	0.1579
Arxiv-mix + Pubmed-mix + Link	0.6874	0.3761	0.1836	0.1494

Dataset	PubMed	
Model	AUC	AP
MLP	0.5583	0.5833
GAT	0.5606	0.6373
GraphSAGE	0.5041	0.5813
RevGNN	0.4538	0.5083
Node2Vec	0.6535	0.6885
w/o Link	0.5010	0.5005
only Link	0.6704	0.6087
Arxiv-std + PubMed-std + Link	0.8246	0.8026
Arxiv-mix + PubMed-mix + Link	0.6451	0.5886

LLMs as Prefix

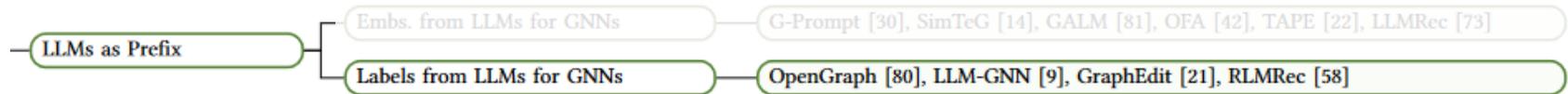
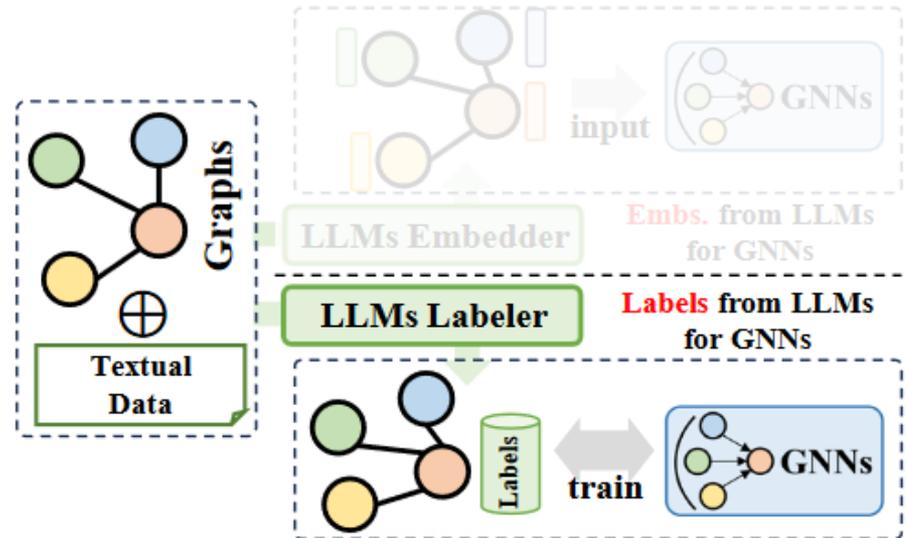
- LLM이 GNN을 위한 graph embedding을 제공
- LLM이 GNN을 위한 graph label을 제공



LLMs as Prefix

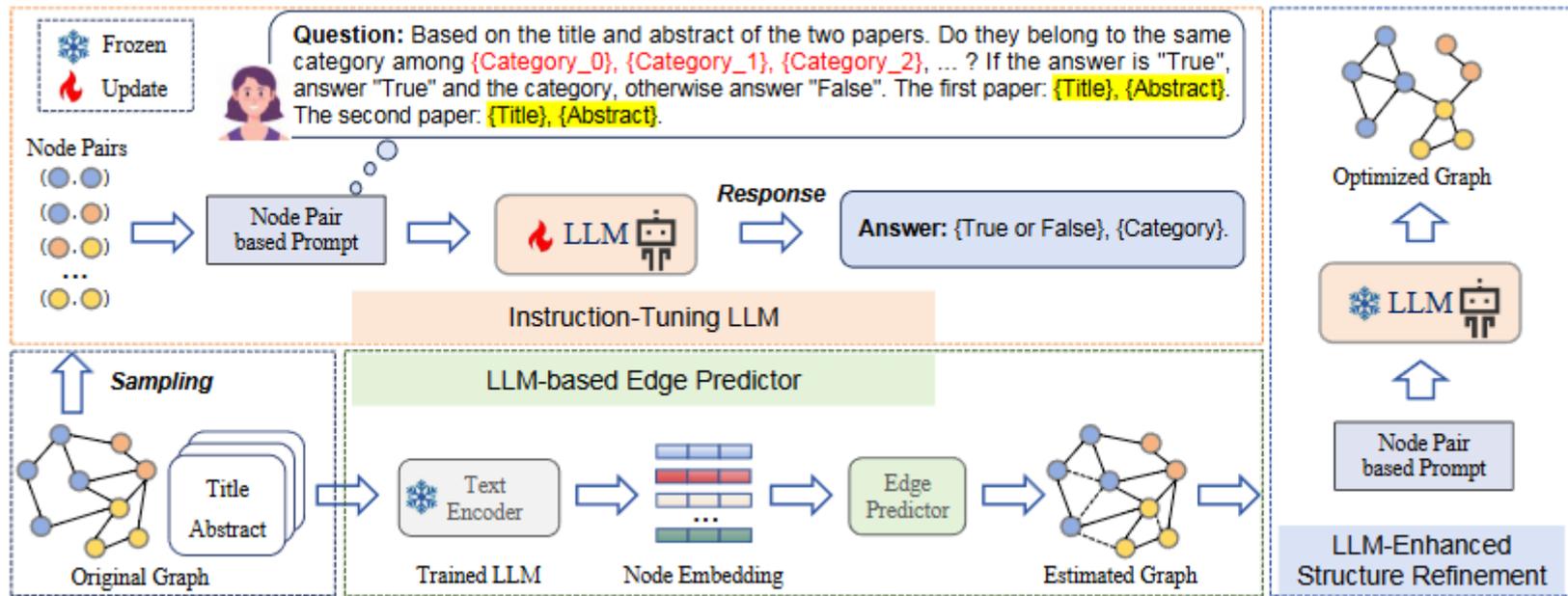
LLM as Labeler

- Motivation
 - ✓ Graph labels are insufficient
 - node class labels, link labels
 - ✓ Label generation based on
 - 언어 이해와 추론
 - 의미적 표현 학습
 - Real-world knowledge



LLMs as Labeler: GraphEdit

- GraphEdit: Large Language Models for Graph Structure Learning
- 논문의 title과 abstract를 기반으로 해당 노드들이 같은 카테고리인지(True/False)와 카테고리 이름을 판별하도록 훈련
- 노드 임베딩 pair 간 edge 존재 여부 predictor를 학습하고, top-k 후보 생성
- 후보에 대해 다시 LLM이 연결할 것인지 prompting해서 판단



LLMs as Labeler: GraphEdit

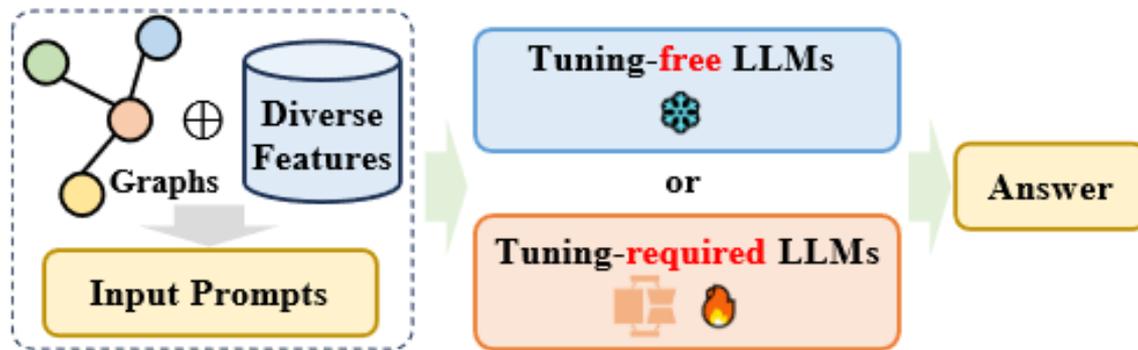
- Experimental Result
 - ✓ PubMed dataset에서는 noise를 주입했을 때가 더 성능이 좋음
 - noise를 “새로운 정보”로 포착
 - LLM의 robust한 reasoning

Model	Cora	Citeseer	PubMed
GCN	87.36 ± 1.60	78.87 ± 2.18	87.37 ± 0.77
GRCN	84.13 ± 0.37	74.23 ± 1.18	85.20 ± 0.10
IDGL	88.63 ± 0.44	80.85 ± 0.07	88.30 ± 0.12
GAug	86.72 ± 0.63	77.61 ± 1.02	84.48 ± 0.37
GEN	86.53 ± 0.63	80.38 ± 0.72	87.04 ± 0.11
SLAPS	81.99 ± 1.57	73.17 ± 0.87	85.21 ± 0.18
GT	88.34 ± 0.35	78.46 ± 0.48	86.69 ± 0.19
CoGSL	82.07 ± 0.51	78.84 ± 0.11	OOM
WSGNN	89.59 ± 0.17	80.88 ± 0.48	87.17 ± 0.19
SUBLIME	85.04 ± 0.37	43.73 ± 7.08	86.03 ± 0.33
STABLE	88.75 ± 0.35	75.67 ± 0.98	86.30 ± 0.15
Nodeformer	88.56 ± 1.01	80.28 ± 0.57	87.93 ± 0.26
GSR	87.56 ± 1.19	78.77 ± 1.56	85.61 ± 0.55
SEGS	87.49 ± 0.66	78.91 ± 0.52	87.57 ± 0.37
GraphEdit	90.90 ± 1.16	81.85 ± 1.42	94.09 ± 0.28

Method	Attack Rate				
	0.05	0.1	0.15	0.2	0.25
GCN	86.06	85.13	84.28	83.61	83.26
IDGL	86.20	85.29	83.54	84.18	82.39
WSGNN	85.94	85.24	84.59	83.64	84.21
GraphEdit	94.07	94.14	94.16	94.20	94.27

LLMs-Only

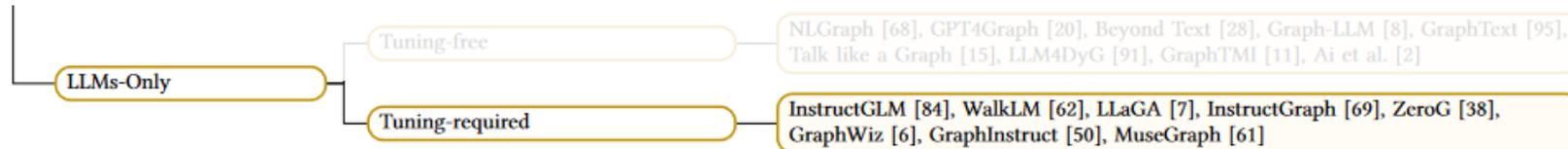
- Graph 관련 질문에 응답하기 위해 적절한 prompt를 구축
- Tuning-free: LLM이 바로 이해할 수 있는 prompt를 설계
- Tuning-required: Instruction tuning을 통해 그래프의 지식과 정렬



Tuning-required LLMs

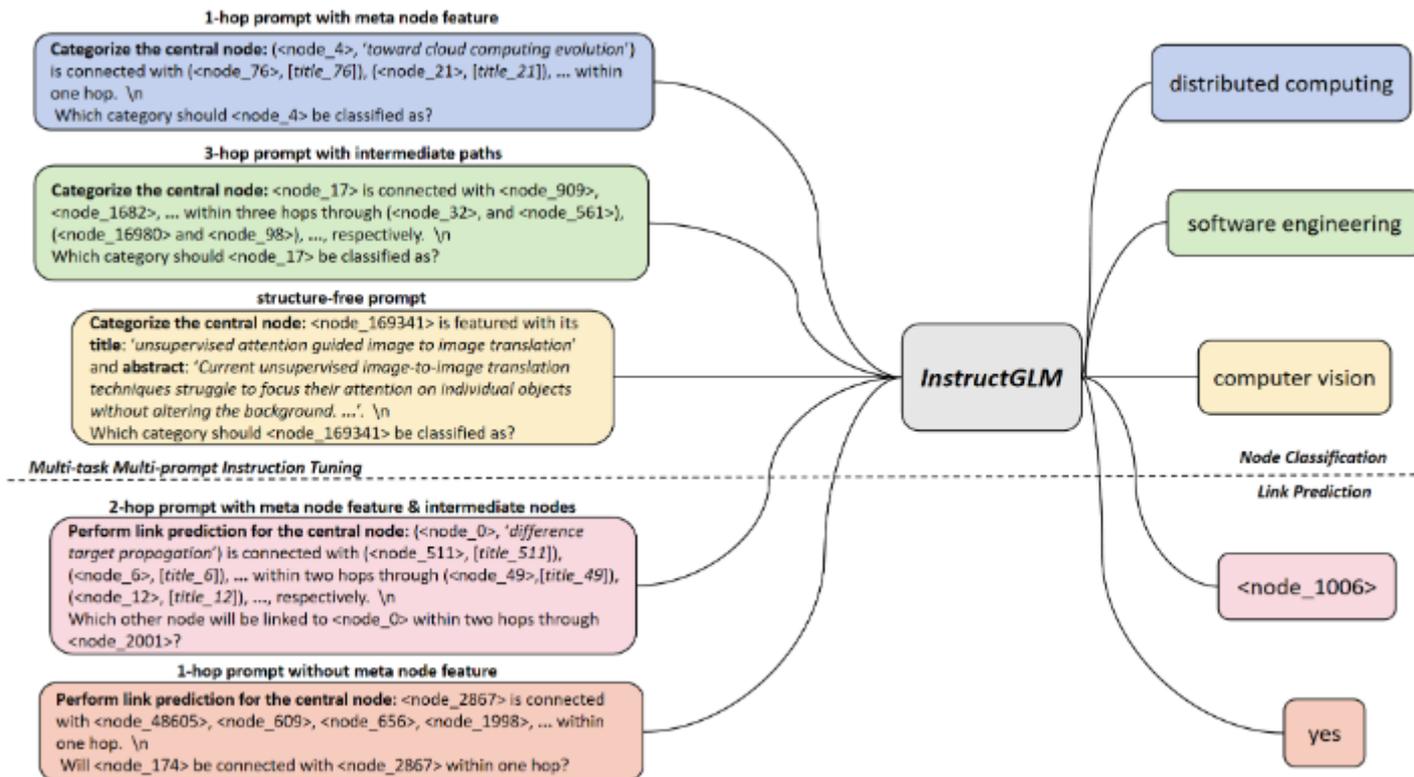
- Motivation

- ✓ 그래프를 sequence 형태로 변환한 후 fine-tuning 기법으로 graph token sequences와 natural language token sequences를 정렬



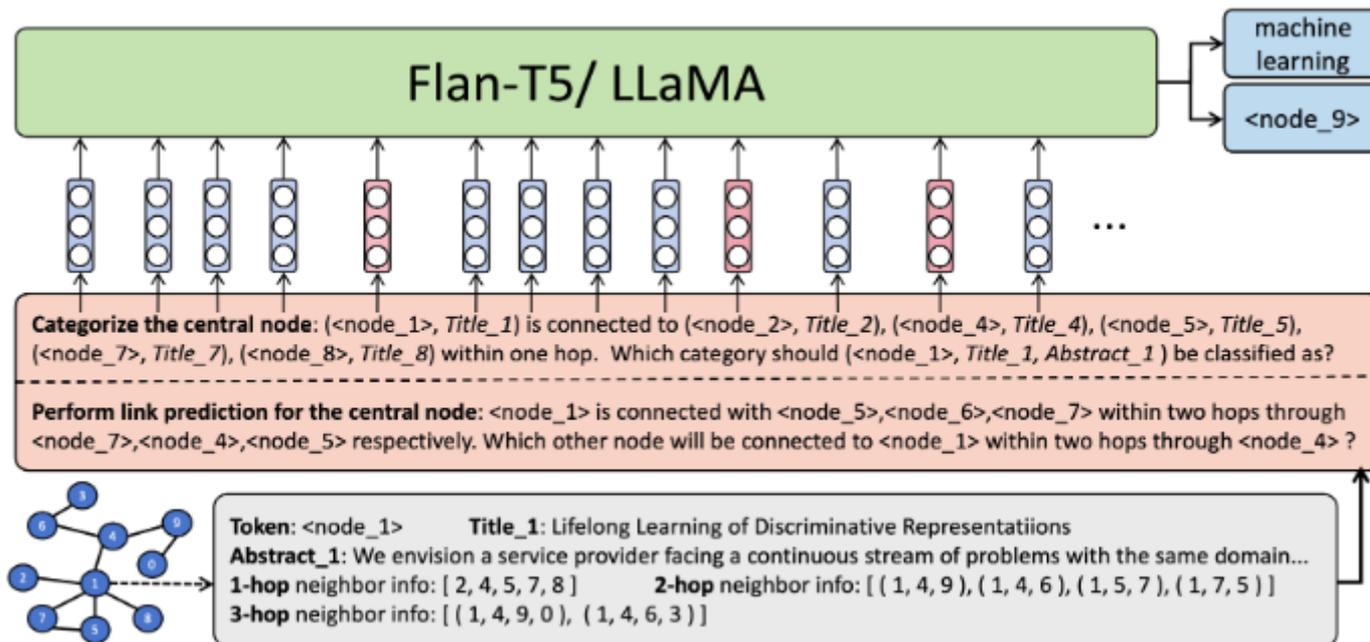
Tuning-required LLMs: InstructGLM

- InstructGLM: Language is All a Graph Needs (EACL '24)
 - ✓ GNN 아키텍처 없이 LLM이 그래프 구조를 처리할 수 있도록 함



Tuning-required LLMs: InstructGLM

- Prompt 설계: 자연어로 그래프 구조를 설명
 - ✓ Multi-hop 구조
 - “Node v is connected to nodes [v1, v2]. Node v1 is also connected to [v3, v4] ...”
 - ✓ 노드의 BoW, TF-IDF 또는 임베딩 정보도 자연어로 함께 포함
 - ✓ 각 노드를 “new token”으로 정의하여 LLM vocabulary에 포함
- Flan-T5, Llama-7B 기반으로 multi-task 학습

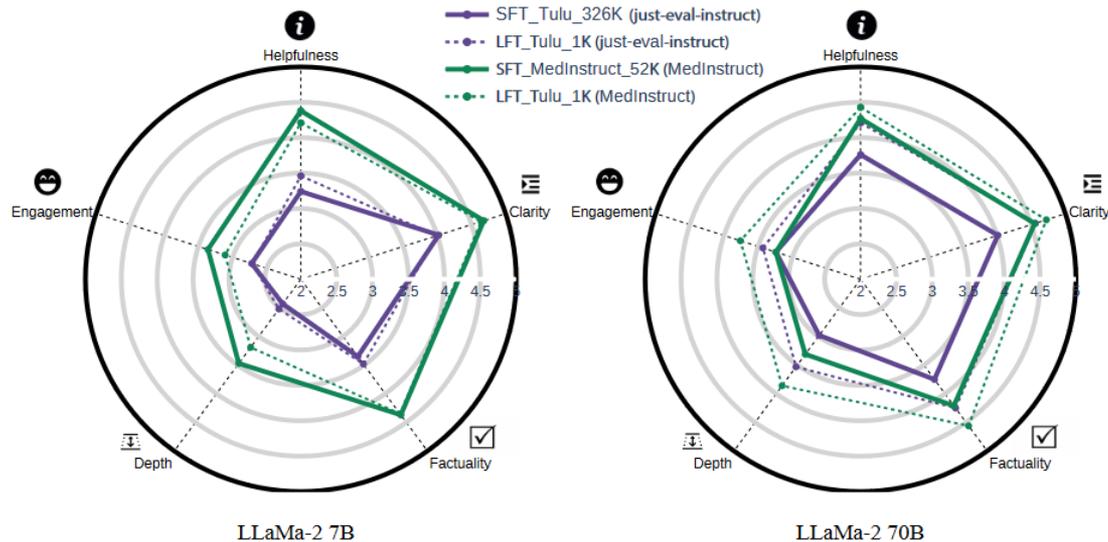


Evaluation

- Automatic Evaluation
 - ✓ Benchmark task나 평가용 dataset에서 LLM이 생성한 응답을 사전 정의된 기준과 비교해 평가
 - 답변의 품질(clarity, helpfulness, correctness, depth 등)을 정량화하여 측정
 - ✓ AlpacaEval, MT-Bench, HELM
- Human Evaluation
 - ✓ 응답의 각 측면에 점수를 매기고 주관적 피드백을 함께 제공
- LLM 기반 평가
 - ✓ GPT-4에서 특정 프롬프트를 주고 “이 응답을 helpfulness, clarity, factuality, depth, engagement 측면에서 평가하라”고 지시
 - ✓ LLM-as-a-judge
- Token 분포 분석
 - ✓ 모델이 생성하는 토큰의 분포 변화(KL-divergence 등)를 통해 성능 평가
 - ✓ Pre-trained model 대비 instruction tuned model이 얼마나 “새로운 지식을 생성”하거나 “기존 지식을 보존”하는지를 분석

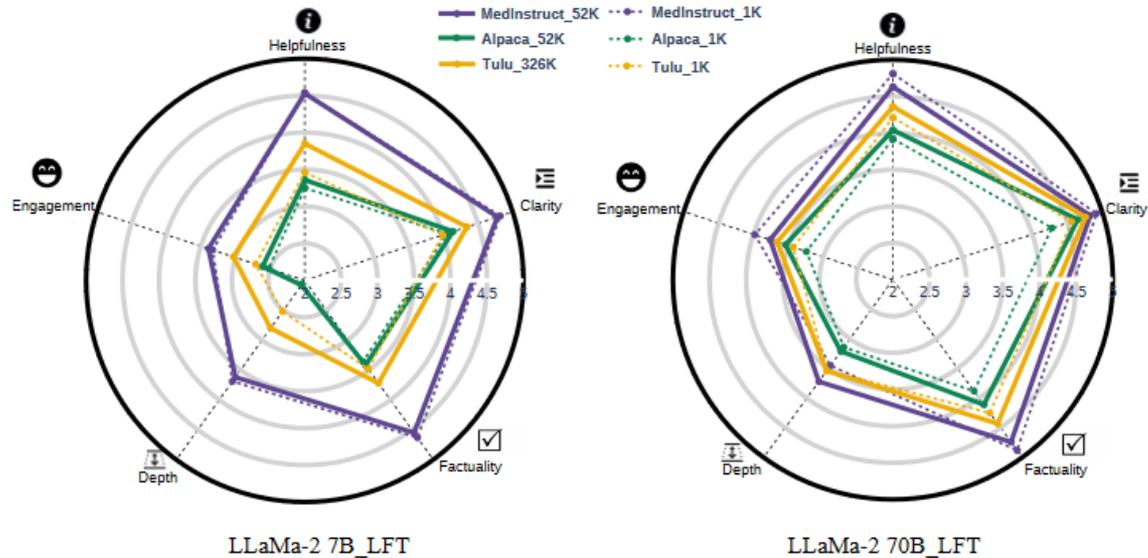
Limitations

- Instruction tuning is not a knowledge enhancer
 - ✓ LoRA fine-tuning은 단순히 응답의 시작 부분만 학습하고, 나머지는 사전학습된 지식을 그대로 출력하는 경향이 있음
 - ✓ Full-parameter fine-tuning은 새로운 지식을 배우지만 기존 지식과 충돌하여 knowledge degradation 문제 발생



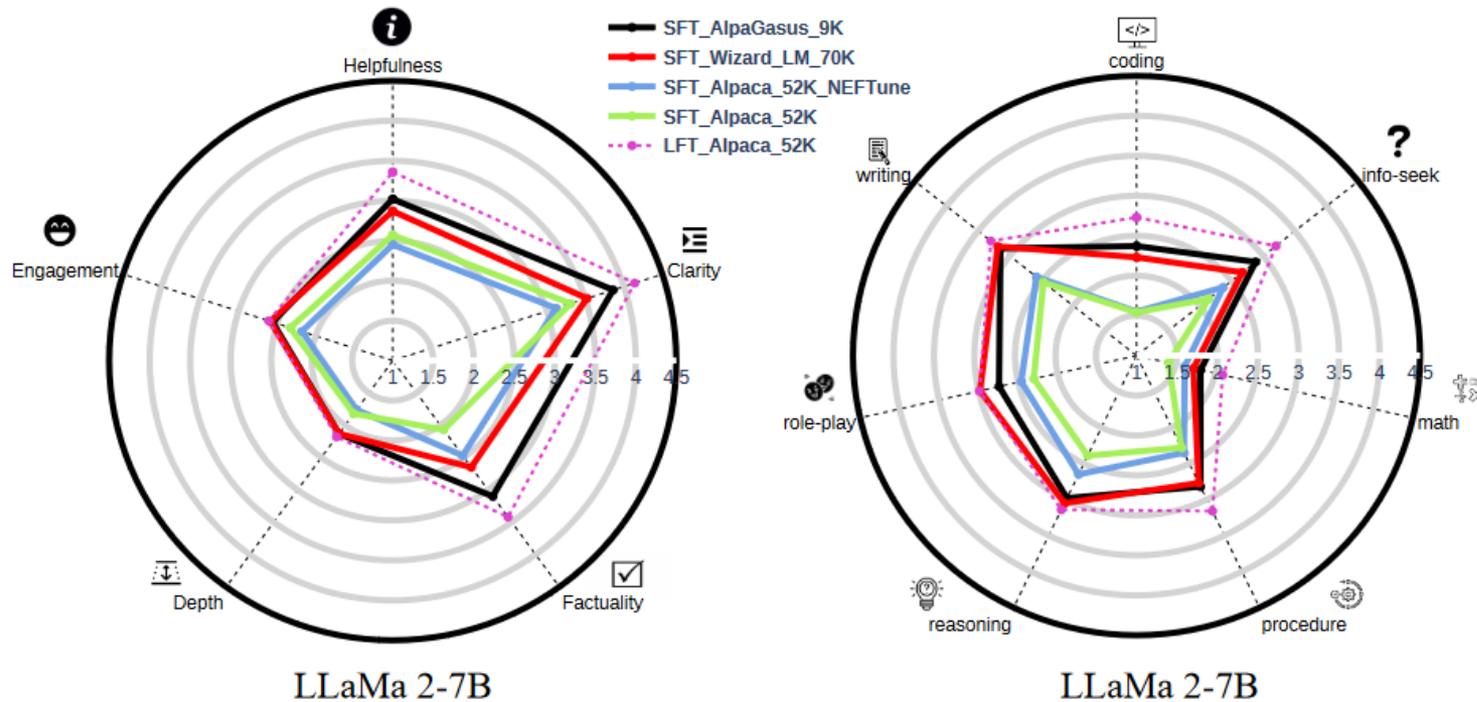
Limitations

- Pattern-copying often hurts performance
 - ✓ 모델이 instruction dataset의 답변 패턴을 과도하게 복사하여 응답의 정확성이 저하
- Instruction tuning dataset may cause hallucinations
 - ✓ Instruction dataset에서 비슷한 개념을 다루는 다른 샘플로부터 잘못된 토큰을 가져옴



Limitations

- Various methods to improve instruction tuning do not improve model performance
 - ✓ 여러 개선법(NEFTune, Dataset filtering 등)이 실제로는 지식 향상에 기여하지 못함



Future works

- Hallucination의 탐지 및 완화
- Multi-modal로의 확장
 - ✓ Text뿐만 아니라 image, video 등 여러 모달리티 데이터와 통합
- Alignment methods
 - ✓ DPO, RLHF과 같은 alignment 방식을 추가로 적용
- Retrieval-Augmented Generation (RAG) 도입
 - ✓ 모델의 내부 지식에만 의존하지 않고 외부 지식 검색을 통해 강화
- Task-specific Instruction tuning
 - ✓ 특정 task에 맞게 구성된 instruction dataset을 통해 특정 능력 향상을 추구

Instruction tuned LLMs

Type	Dataset Name	# of Instances	# of Lang	Construction	Open-source
Human-Crafted	UnifiedQA (Khashabi et al., 2020) ¹	750K	En	human-crafted	Yes
	UnifiedSKG (Xie et al., 2022) ³	0.8M	En	human-crafted	Yes
	Natural Instructions (Honovich et al., 2022) ⁴	193K	En	human-crafted	Yes
	Super-Natural Instructions (Wang et al., 2022f) ⁵	5M	55 Lang	human-crafted	Yes
	P3 (Sanh et al., 2021) ⁶	12M	En	human-crafted	Yes
	xP3 (Muennighoff et al., 2022) ⁷	81M	46 Lang	human-crafted	Yes
	Flan 2021 (Longpre et al., 2023) ⁸	4.4M	En	human-crafted	Yes
	COIG (Zhang et al., 2023a) ⁹	-	-	-	Yes
	InstructGPT (Ouyang et al., 2022)	13K	Multi	human-crafted	No
	Dolly (Conover et al., 2023a) ¹⁶	15K	En	human-crafted	Yes
	LIMA (Zhou et al., 2023a) ¹⁸	1K	En	human-crafted	Yes
ChatGPT (OpenAI, 2022)	-	Multi	human-crafted	No	
OpenAssistant (Köpf et al., 2023) ²⁰	161,443	Multi	human-crafted	Yes	
Synthetic Data (Distillation)	OIG (LAION.ai, 2023) ²	43M	En	ChatGPT (No technique reports)	Yes
	Unnatural Instructions (Honovich et al., 2022) ¹⁰	240K	En	InstructGPT-Generated	Yes
	InstructWild (Xue et al., 2023) ¹²	104K	-	ChatGPT-Generated	Yes
	Evol-Instruct / WizardLM (Xu et al., 2023a) ¹³	52K	En	ChatGPT-generated	Yes
	Alpaca (Taori et al., 2023a) ¹⁴	52K	En	InstructGPT-generated	Yes
	LogiCoT (Liu et al., 2023a) ¹⁵	-	En	GPT-4-Generated	Yes
	GPT-4-LLM (Peng et al., 2023) ¹⁷	52K	En&Zh	GPT-4-Generated	Yes
	Vicuna (Chiang et al., 2023)	70K	En	Real User-ChatGPT Conversations	No
	Baize v1 (Conover et al., 2023b) ²¹	111.5K	En	ChatGPT-Generated	Yes
	UltraChat (Ding et al., 2023a) ²²	675K	En&Zh	GPT 3/4-Generated	Yes
	Guanaco (JosephusCheung, 2021) ¹⁹	534,530	Multi	GPT (Unknown Version)-Generated	Yes
	Orca (Mukherjee et al., 2023) ²³	1.5M	En	GPT 3.5/4-Generated	Yes
	ShareGPT ²⁴	90K	Multi	Real User-ChatGPT Conversations	Yes
	WildChat ²⁵	150K	Multi	Real User-ChatGPT Conversations	Yes
	WizardCoder (Luo et al., 2023)	-	Code	LLaMa 2-Generated	No
	MagiCoder (Wei et al., 2023b) ²⁶	75K/110K	Code	GPT-3.5-Generated	Yes
	WaveCoder (Yu et al., 2023)	-	Code	GPT 4-Generated	No
	Phi-1 (Gunasekar et al., 2023) ²⁷	6B Tokens	Code Q and A	GPT-3.5-Generated	Yes
Phi-1.5 (Li et al., 2023i)	-	Code Q and A	GPT-3.5-Generated	No	
Nectar (Zhu et al., 2023a) ²⁸	183K	En	GPT 4-Generated	Yes	
Synthetic Data (Self-Improvement)	Self-Instruct (Wang et al., 2022e) ¹¹	52K	En	InstructGPT-Generated	Yes
	Instruction Backtranslation (Li et al., 2023g)	502K	En	LLaMa-Generated	No
	SPIN (Chen et al., 2024b) ²⁹	49.8K	En	Zephyr-Generated	Yes

References

- https://www.ibm.com/think/topics/instruction-tuning?utm_source=chatgpt.com
- <https://www.superannotate.com/blog/llm-fine-tuning>
- <https://jins-sw.tistory.com/entry/LLM-Pretrained-vs-Instruction-Tuned-Model>
- <https://learn-ai.tistory.com/entry/Paper-Review-FLAN-Finetuned-Language-Models-Are-Zero-Shot-Learners>
- <https://velog.io/@hellemon/SELF-INSTRUCT-Aligning-Language-Models-with-Self-Generated-Instructions>
- https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/
- <https://huggingface.co/datasets/tatsu-lab/alpaca/viewer/default/train?row=47&views%5B%5D=train>
- <https://huggingface.co/meta-llama/Meta-Llama-3-8B?text=%EA%B0%90%EA%B8%B0%EC%95%BD%EA%B3%BC+%ED%95%B4%EC%97%B4%EC%A0%9C%EB%A5%BC+%EA%B0%99%EC%9D%B4+%EB%A8%B9%EC%96%B4%EB%8F%84+%EB%90%98%EB%82%98%EC%9A%94%3F>
- <https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct>
- <https://llm4graph-tutorial.github.io/>