# Part II: Improve Large Language Models for Search/Graph-Augmented Scenarios

Bowen Jin, Yu Zhang, Yunyi Zhang, Jiawei Han
Department of Computer Science
University of Illinois at Urbana-Champaign
May 1, 2025

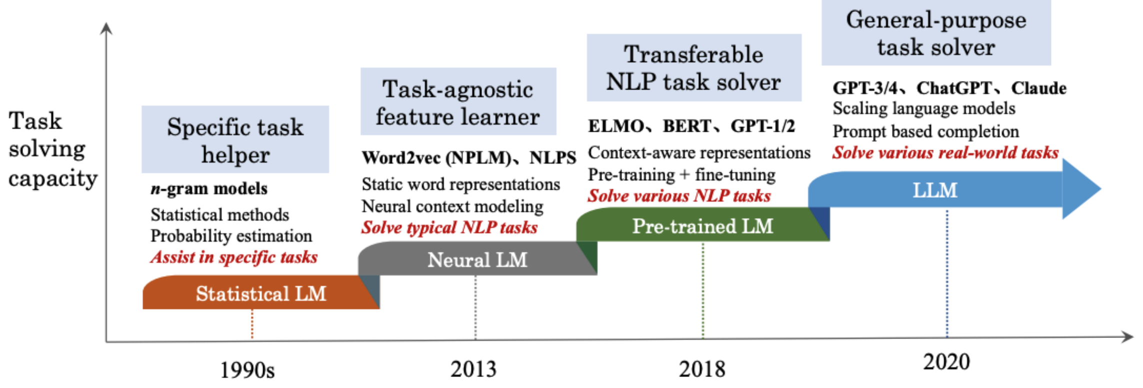
Tutorial Website:

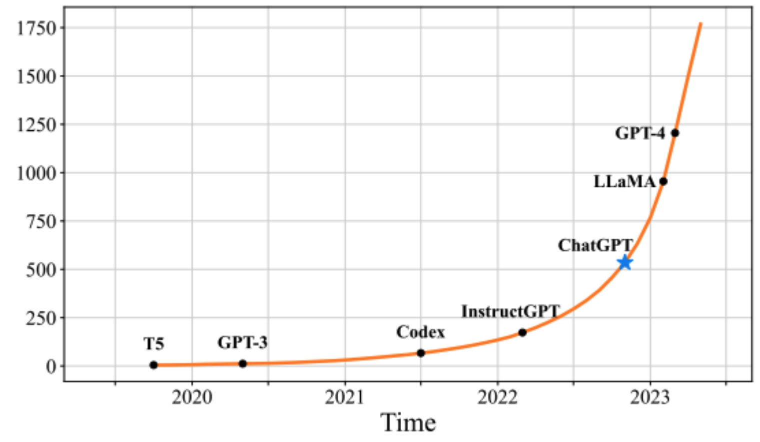
# Outline

- Improving LLMs for search-augmented scenarios
  - From General-Purpose Models to RAG-Optimized LLMs
  - Beyond RAG Pipelines: Towards Search-Enabled LLM Agents
- Improving LLMs for graph-augmented scenarios
  - Pretrained LLM Retrievers for Graph-Augmented Search
  - Reasoning over Graph-Structured Knowledge with LLMs

# Large Language Models (LLMs)

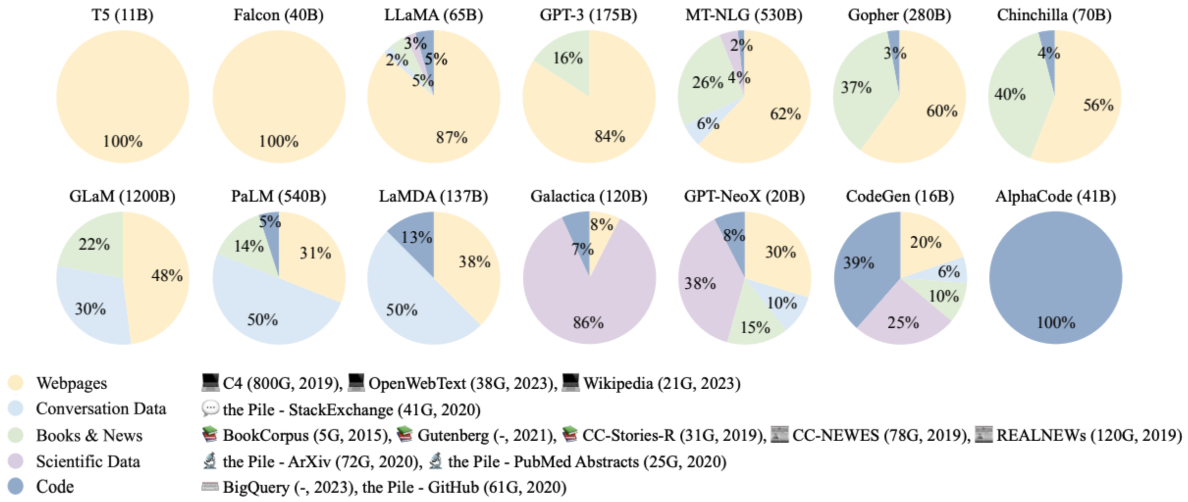
- LLMs have demonstrated their strong **text** encoding/decoding ability.




(b) Query="Large Language Model"

- LLMs have shown newly found emergent ability (e.g., **reasoning**).



# RAG support various applications

- Large language models often struggle with factual inaccuracies and produce hallucinated content when faced with knowledge-intensive questions.
- *Retrieval Augmented Generation (RAG)* incorporates information retrieved from an external knowledge sources into the context to provide up-to-date information and specify obscure facts.

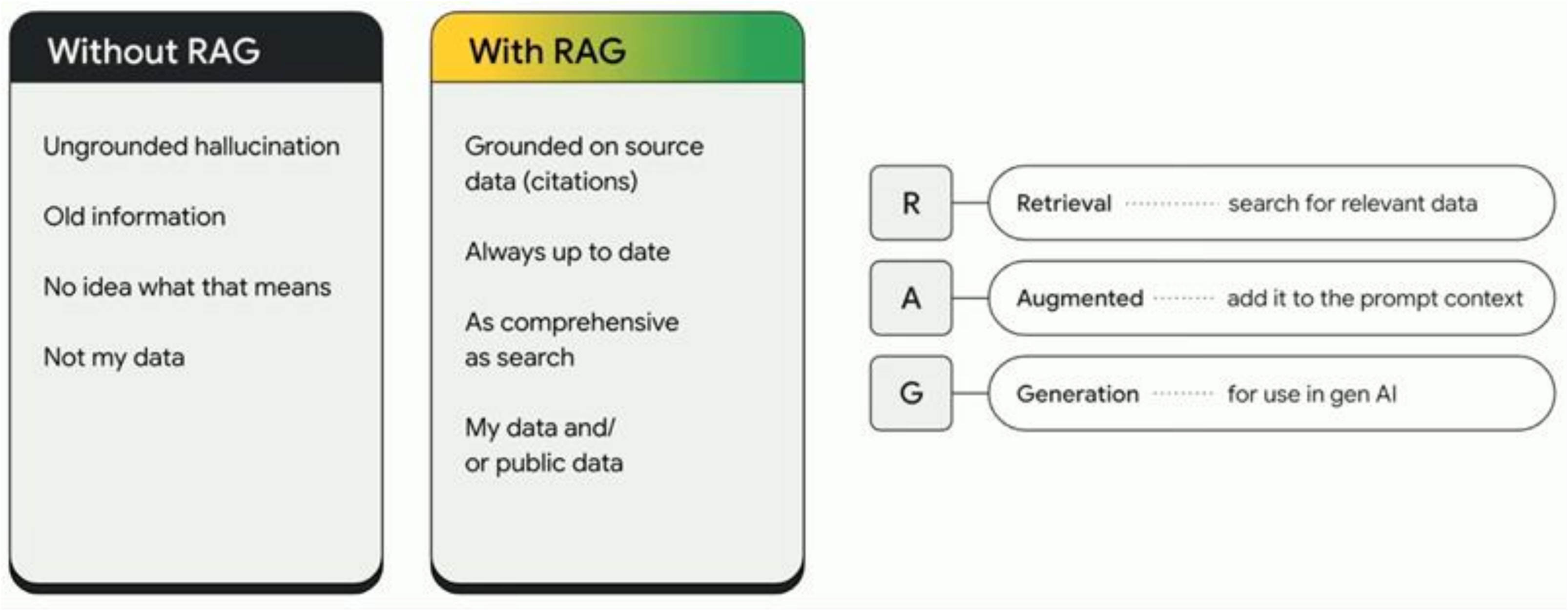
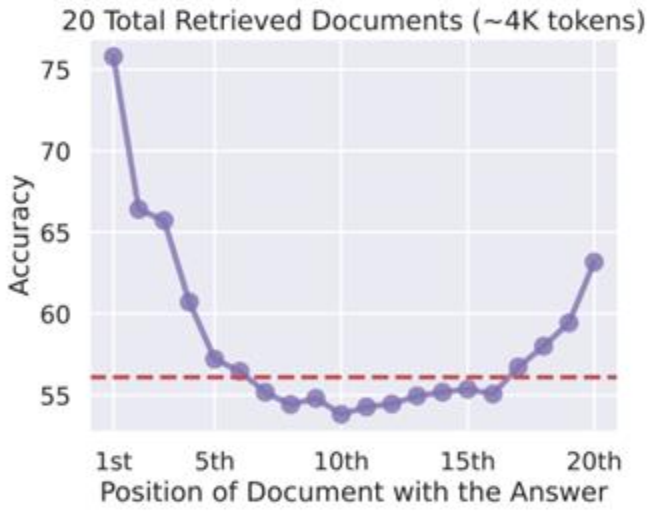



Image from: https://cloud.google.com/use-cases/retrieval-augmented-generation

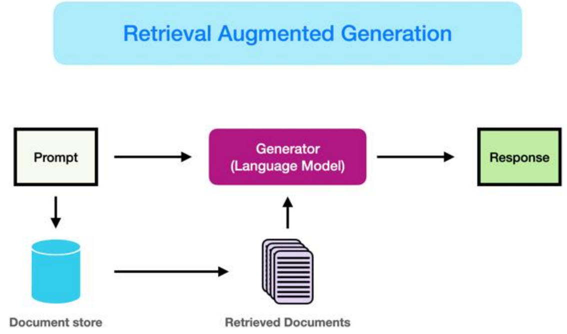
# Challenges in RAG

- In the RAG pipeline, LLMs can be easily **distracted** by the **irrelevant** retrieved information.

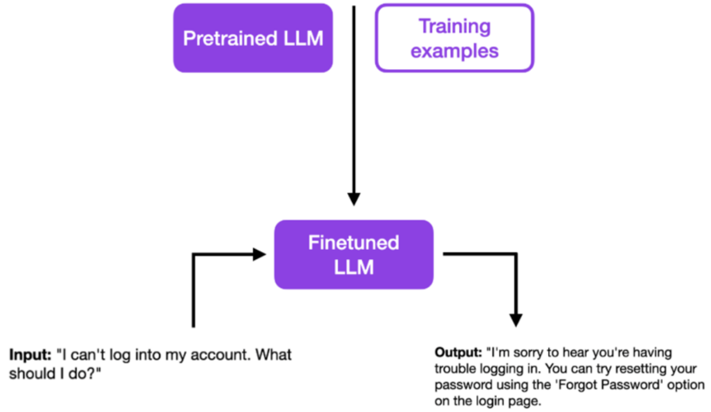



- The retrieval is always not perfect.

- Traditional RAG methods rely on **prompting**, without fundamental LLM improvement for search.




- Supervised finetuning (SFT) for RAG needs labeled trajectories and is **hard to scale**.

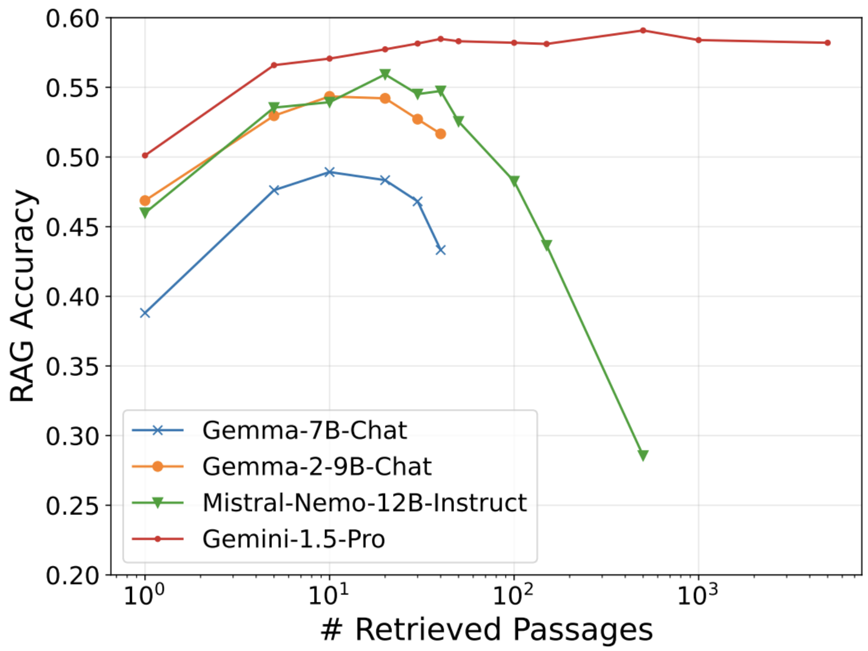


# Outline

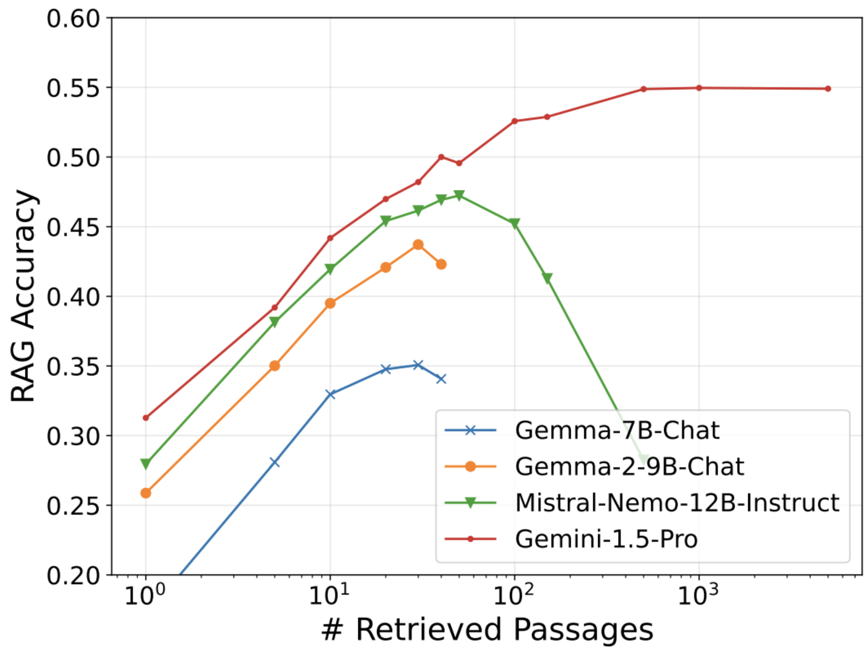
- Improving LLMs for search-augmented scenarios
  - From General-Purpose Models to RAG-Optimized LLMs
  - Beyond RAG Pipelines: Towards Search-Enabled LLM Agents

- Improving LLMs for graph-augmented scenarios
  - Pretrained LLM Retrievers for Graph-Augmented Search
  - Reasoning over Graph-Structured Knowledge with LLMs

# General LLMs Fails in complex long-context RAG scenarios




(a) RAG performance with e5 retriever




(b) RAG performance with BM25 retriever

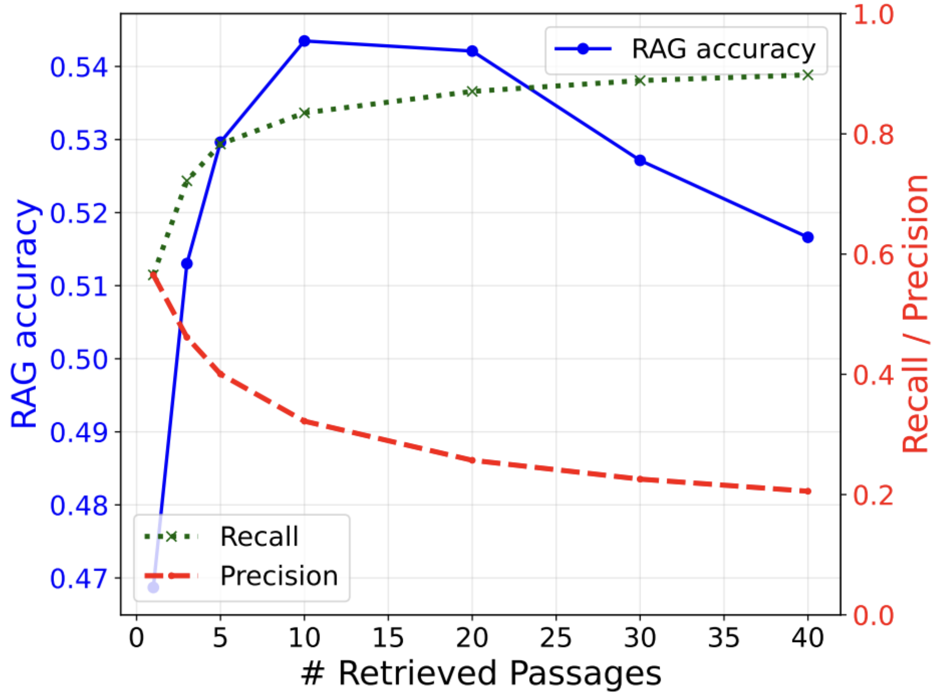
**Insights**

1) The effectiveness of increasing retrieved context size in RAG depends on the strength of the retriever.
2) With a strong retriever, performance exhibits an “inverted-U pattern”, while a weak retriever shows more consistent, albeit potentially limited, improvement.
3) This suggests that factors beyond simply the amount of retrieved information are at play.

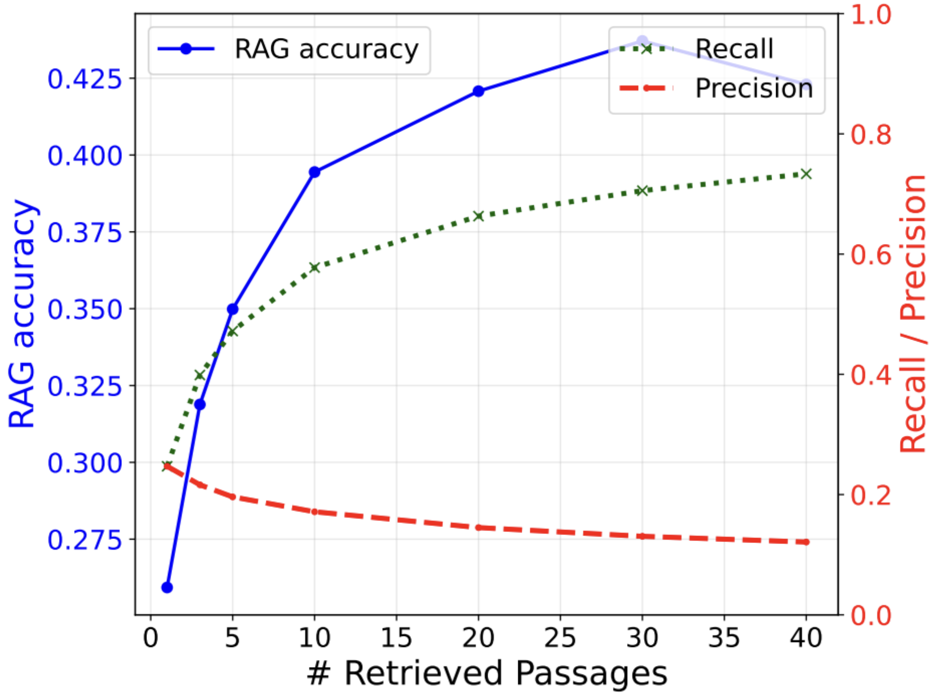


Jin, et al. Long-Context LLMs Meet RAG: Overcoming Challenges for Long Inputs in RAG. ICLR’25.

# Irrelevant information can mislead LLMs in RAG




(a) Retrieval with e5 retriever



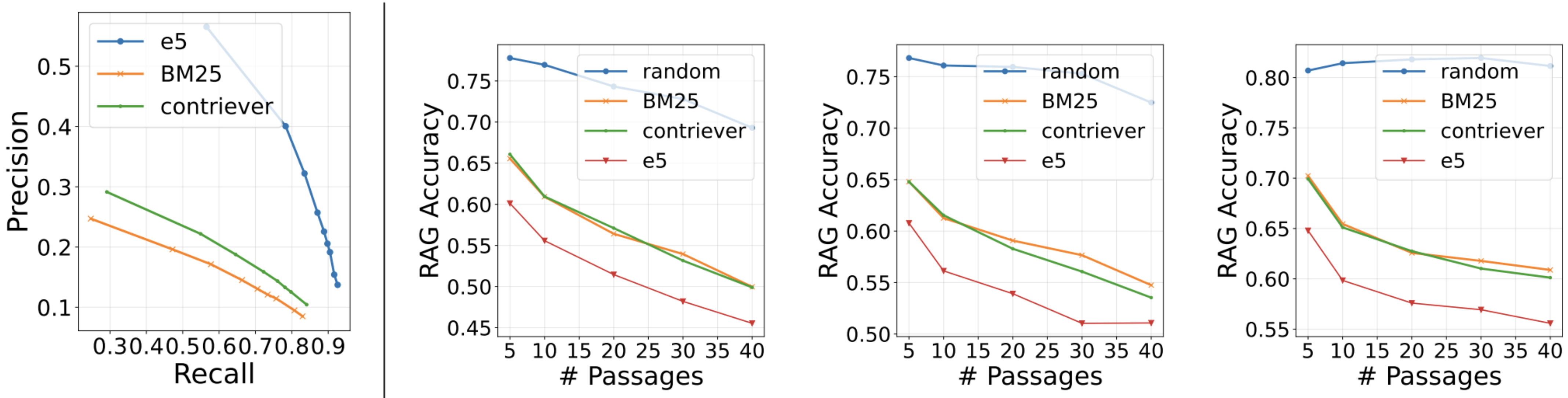

(b) Retrieval with BM25 retriever

**Insights**

- **Influence of irrelevant passages**: The discrepancy between retrieval recall and RAG accuracy underscores the detrimental effect of irrelevant retrieved passages ("hard negatives") on the LLMs' performance.
- **Limitations of precision as a metric**: The contrasting performance trends observed with e5 and BM25, despite the former's higher precision, reveal that precision alone is an inadequate measure of retrieval quality in this context, when the end-to-end performance is considered.

Jin, et al. Long-Context LLMs Meet RAG: Overcoming Challenges for Long Inputs in RAG. ICLR'25.

# Negative hardness correlates with retriever strength




(a) Retrievers (b) Gemma2-9B-Chat (c) Mistral-12B-Instruct (d) Gemini-1.5-Pro

**Observations**

- **Sensitivity to hard negatives**: Across all LLMs, increasing the number of hard negative passages generally leads to a decline in RAG answer accuracy.
- **Retriever strength and hard negative difficulty**: The strength of the retriever directly correlates with the difficulty of the retrieved hard negatives. LLMs struggle more with hard negatives from stronger retrievers.
- **Distinguishing random and hard negatives**: While Gemini-1.5-Pro demonstrates robustness to random negatives, it remains susceptible to the influence of hard negatives.

Jin, et al. Long-Context LLMs Meet RAG: Overcoming Challenges for Long Inputs in RAG. ICLR'25.

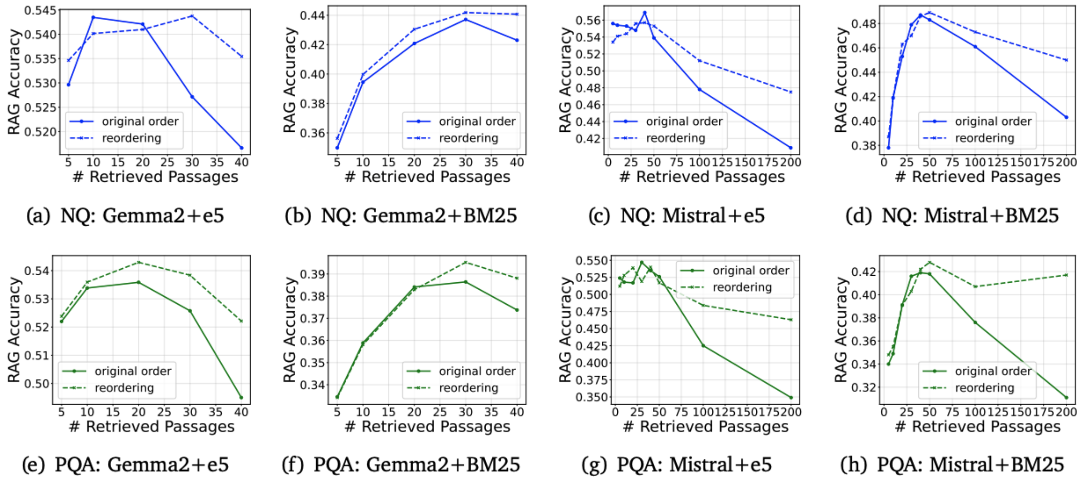
# Simple and effective training-free RAG improvement

## Retrieval reordering

- ”Lost-in-the-middle”: LLMs exhibit a tendency to prioritize information presented at the beginning and end of an input sequence, while paying less attention to the middle.
- Retrieval reordering leverages the inherent "lost-in-the-middle" phenomenon observed in LLMs to mitigate the negative effects of hard negatives.

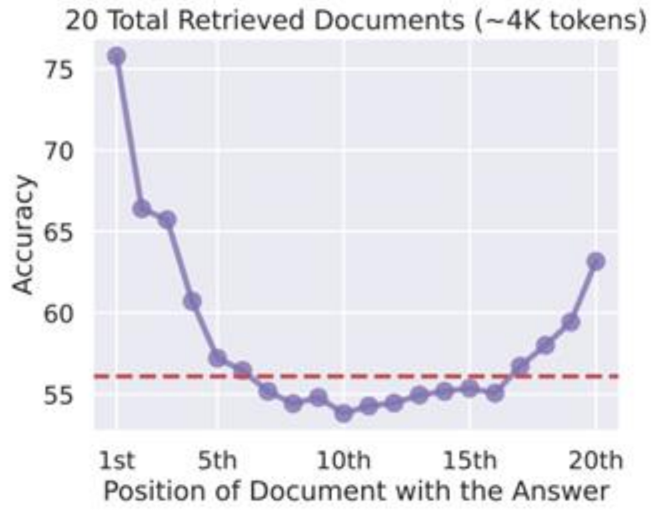
Given a query $q$ and a set of retrieved passages $d_1, d_2, \ldots, d_k$ with decreasing relevance scores:

$$[I, d_1, d_2, \ldots, d_{k-1}, d_k, q] \longrightarrow [I, d_1, d_3, \ldots, d_4, d_2, q]$$




(a) NQ: Gemma2+e5 (b) NQ: Gemma2+BM25 (c) NQ: Mistral+e5 (d) NQ: Mistral+BM25

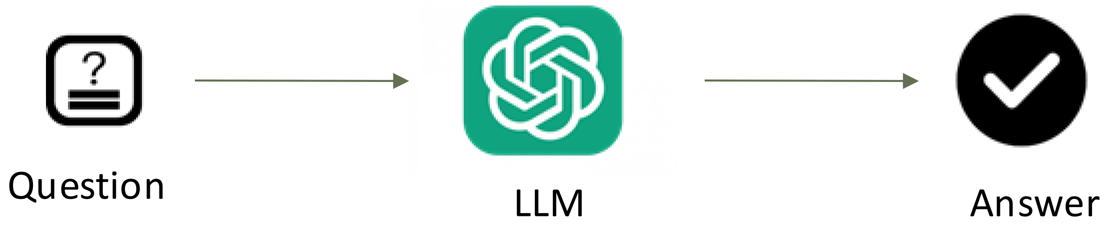
(e) PQA: Gemma2+e5 (f) PQA: Gemma2+BM25 (g) PQA: Mistral+e5 (h) PQA: Mistral+BM25




Jin, et al. Long-Context LLMs Meet RAG: Overcoming Challenges for Long Inputs in RAG. ICLR’25.

# Improving Robustness for RAG via Data-Augmented Fine-Tuning

## Vanilla LLM finetuning




**Finetune the LLM to obtain knowledge**

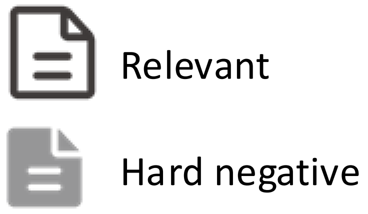
## Implicitly improving LLM for RAG through fine-tuning


**Finetune the long context LLM to be robust to hard negatives**

# Improving Robustness for RAG via Data-Augmented Fine-Tuning

## Enhancing relevance identification through reasoning augmentation


**Finetune the long context LLM to reason to identify hard negatives**

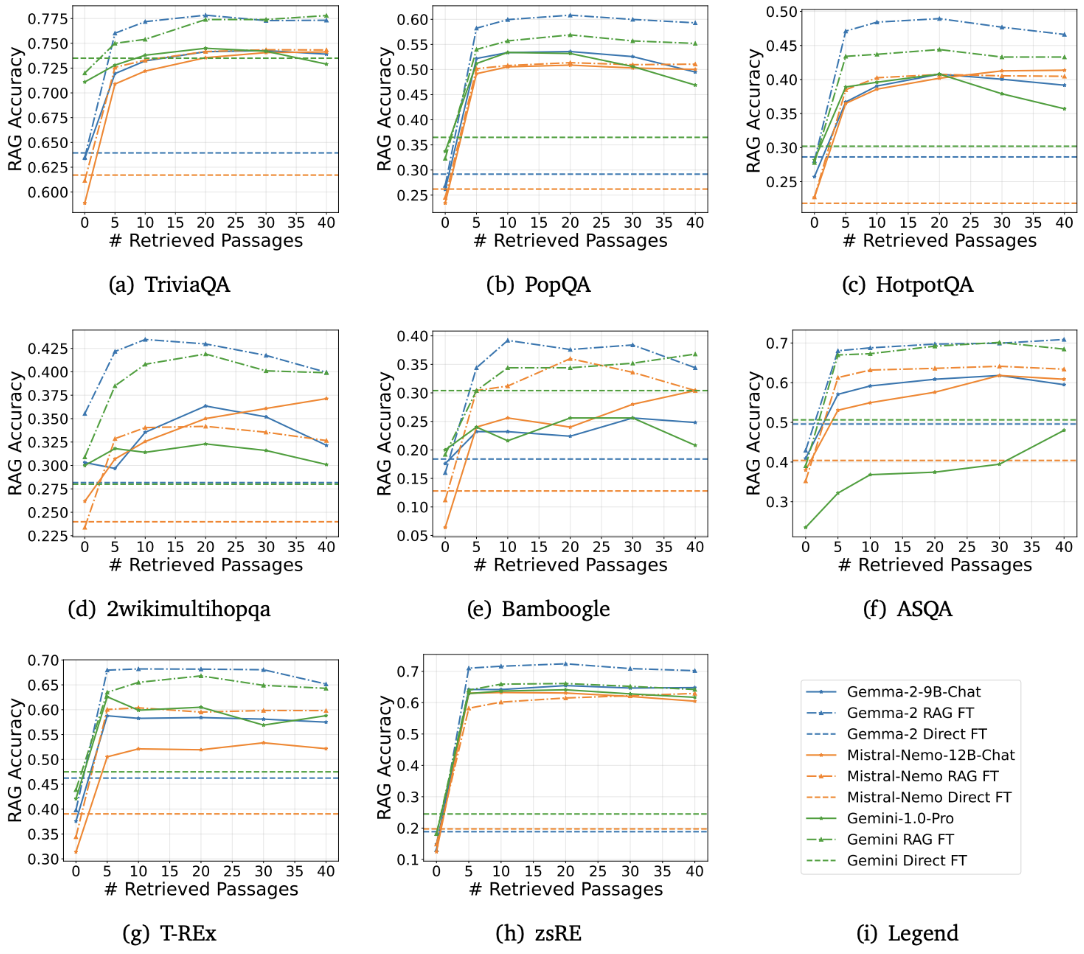




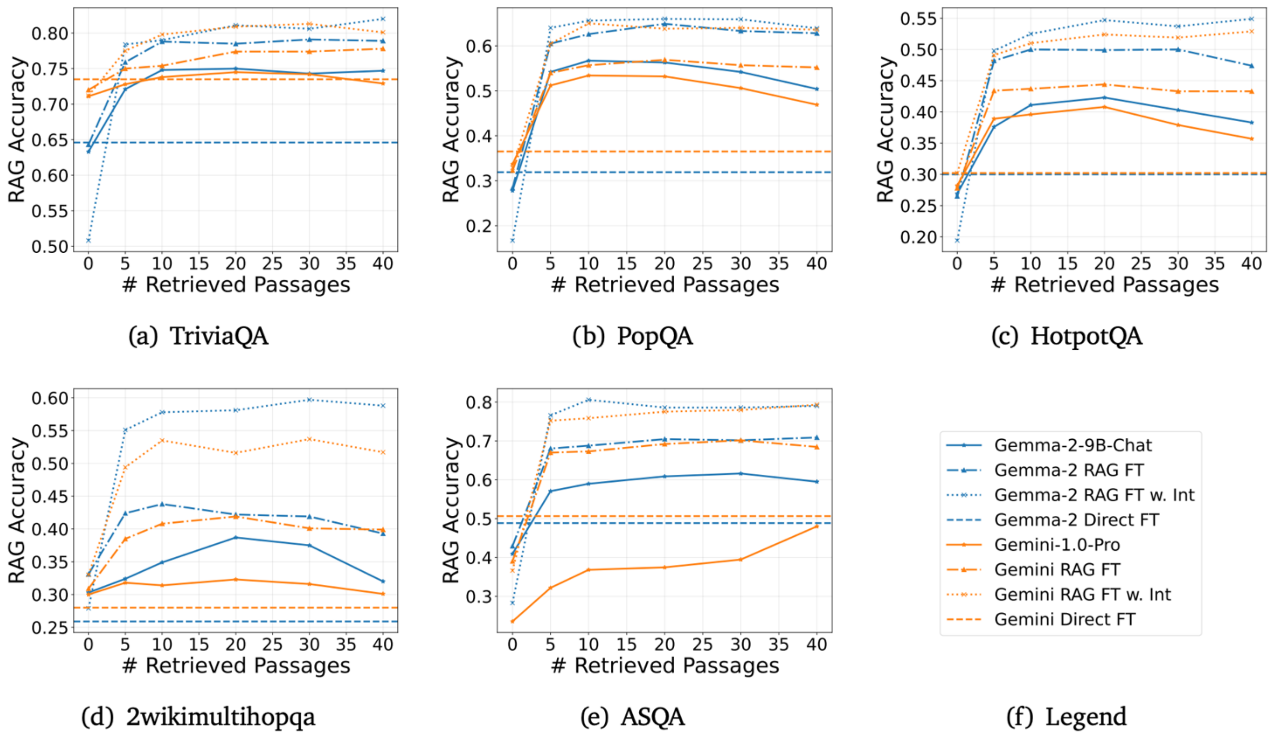

Jin, et al. Long-Context LLMs Meet RAG: Overcoming Challenges for Long Inputs in RAG. ICLR’25.

# Improving Robustness for RAG via Data-Augmented Fine-Tuning

## Implicitly improving LLM robustness through fine-tuning




## Enhancing relevance identification through reasoning augmentation




Jin, et al. Long-Context LLMs Meet RAG: Overcoming Challenges for Long Inputs in RAG. ICLR'25.

# Outline

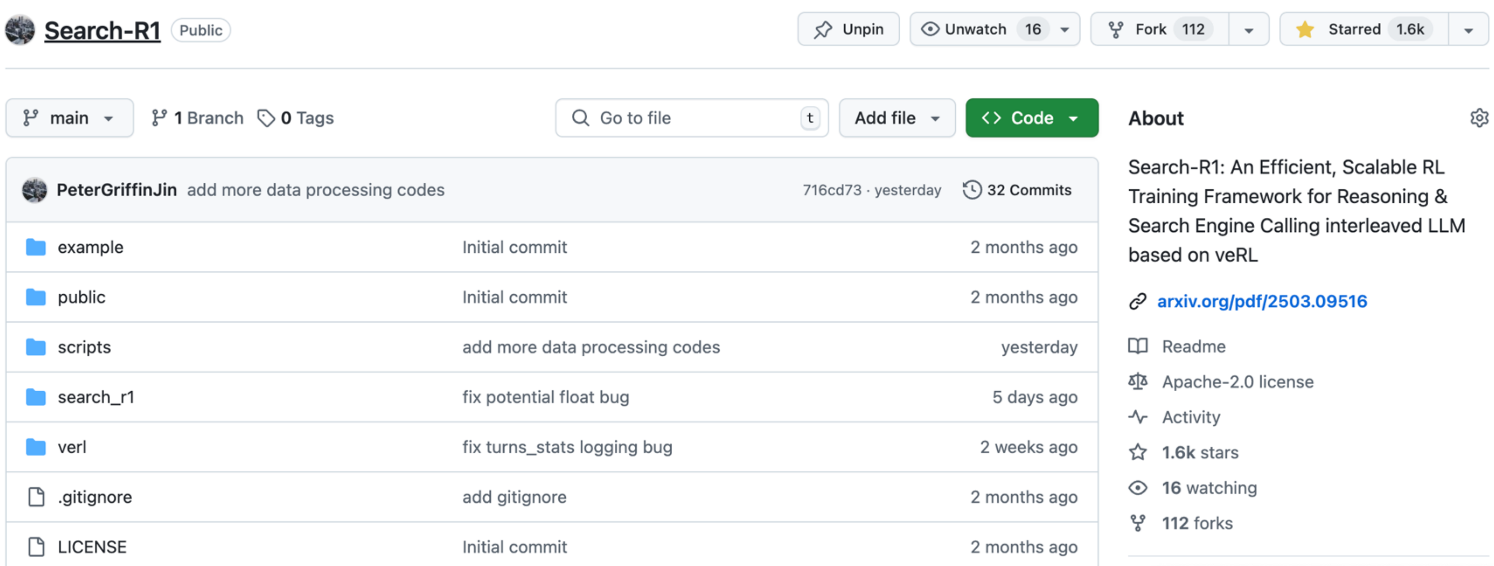
- Improving LLMs for search-augmented scenarios
  - From General-Purpose Models to RAG-Optimized LLMs
  - Beyond RAG Pipelines: Towards Search-Enabled LLM Agents

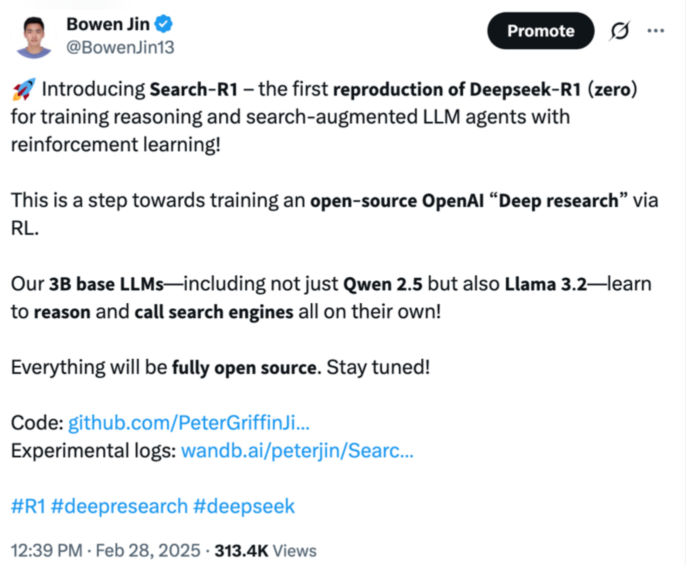
- Improving LLMs for graph-augmented scenarios
  - Pretrained LLM Retrievers for Graph-Augmented Search
  - Reasoning over Graph-Structured Knowledge with LLMs

# Evolving from RAG Pipelines to Search-Enabled Agents: Search-R1

- It is an efficient, scalable **reinforcement learning (RL)** training framework which can teach LLMs to **reason** and **call search engines** in an interleaved fashion.

- We show that deepseek-R1 (zero) style RL training can be extended and let the LLM learn to call a search engine and do reasoning simultaneously.



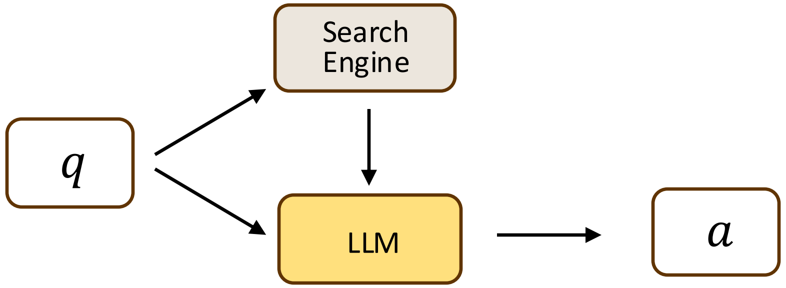




Jin, et al. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. arxiv'25.


15

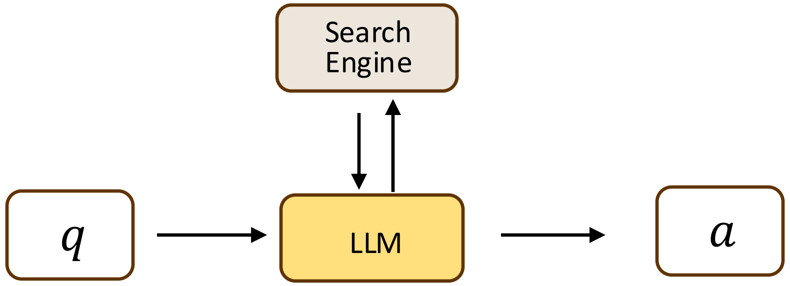
# Evolving from RAG Pipelines to Search-Enabled Agents: Search-R1

- Traditional RAG is adopted as an inference time strategy and usually focus on single turn retrieval based on the input question.




<query> q </query>
<info> I </info>
<answer> a <answer>

- Search-R1 improves LLM for search with RL training and enables multi-turn agentic interaction with the search engine.

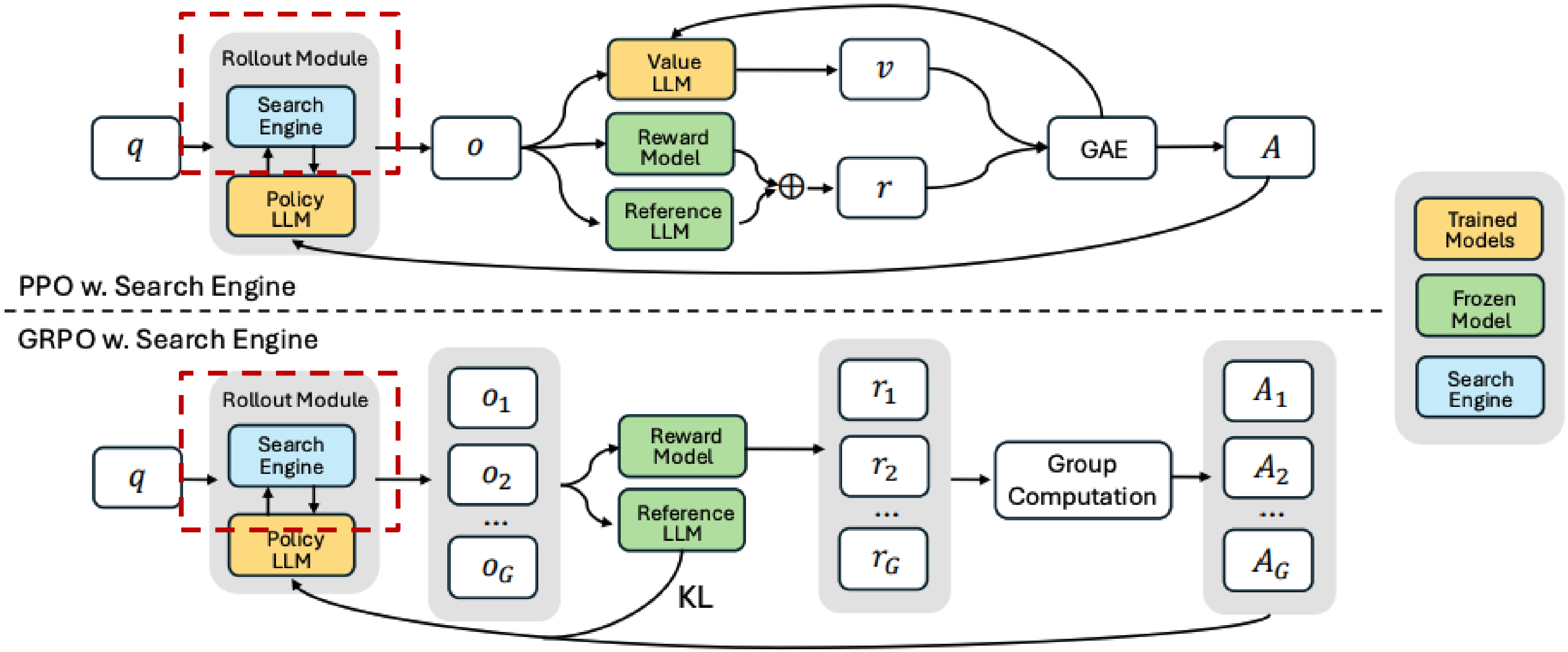



<reason> $r_1$ </reason>
<search> $q_1$ </search>
<info> $I_1$ </info>
…
<reason> $r_t$ </reason>
<answer> a <answer>


16

# Evolving from RAG Pipelines to Search-Enabled Agents: Search-R1

- Reinforcement Learning with a Search Engine




Jin, et al. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. arxiv'25.


17

# Experiments

- Search-R1 consistently outperforms strong baseline methods.
- Search-R1 surpasses RL-based training for LLM reasoning without retrieval (R1).

| Method | NQ | TriviaQA | PopQA | HotpotQA | 2wiki | Musique | Bamboogle | Avg. |
|---|---|---|---|---|---|---|---|---|
| **Qwen2.5-7b-Base/Instruct** | | | | | | | | |
| Direct Inference | 0.134 | 0.408 | 0.140 | 0.183 | 0.250 | 0.031 | 0.120 | 0.181 |
| CoT | 0.048 | 0.185 | 0.054 | 0.092 | 0.111 | 0.022 | 0.232 | 0.106 |
| IRCoT | 0.224 | 0.478 | 0.301 | 0.133 | 0.149 | 0.072 | 0.224 | 0.239 |
| Search-o1 | 0.151 | 0.443 | 0.131 | 0.187 | 0.176 | 0.058 | 0.296 | 0.206 |
| RAG | 0.349 | 0.585 | 0.392 | 0.299 | 0.235 | 0.058 | 0.208 | 0.304 |
| SFT | 0.318 | 0.354 | 0.121 | 0.217 | 0.259 | 0.066 | 0.112 | 0.207 |
| R1-base | 0.297 | 0.539 | 0.202 | 0.242 | 0.273 | 0.083 | 0.296 | 0.276 |
| R1-instruct | 0.270 | 0.537 | 0.199 | 0.237 | 0.292 | 0.072 | 0.293 | 0.271 |
| Search-R1-base | **0.412** | 0.568 | **0.428** | 0.356 | 0.322 | 0.142 | 0.384 | 0.373 |
| Search-R1-instruct | 0.397 | **0.606** | 0.404 | **0.380** | **0.326** | **0.168** | **0.408** | **0.384** |
| **Qwen2.5-3b-Base/Instruct** | | | | | | | | |
| Direct Inference | 0.106 | 0.288 | 0.108 | 0.149 | 0.244 | 0.020 | 0.024 | 0.134 |
| CoT | 0.023 | 0.032 | 0.005 | 0.021 | 0.021 | 0.002 | 0.000 | 0.015 |
| IRCoT | 0.111 | 0.312 | 0.200 | 0.164 | 0.171 | 0.067 | 0.240 | 0.181 |
| Search-o1 | 0.238 | 0.472 | 0.262 | 0.221 | 0.218 | 0.054 | **0.320** | 0.255 |
| RAG | **0.348** | **0.544** | **0.387** | 0.255 | 0.226 | 0.047 | 0.080 | 0.270 |
| SFT | 0.249 | 0.292 | 0.104 | 0.186 | 0.248 | 0.044 | 0.112 | 0.176 |
| R1-base | 0.226 | 0.455 | 0.173 | 0.201 | 0.268 | 0.055 | 0.224 | 0.229 |
| R1-instruct | 0.210 | 0.449 | 0.171 | 0.208 | 0.275 | 0.060 | 0.192 | 0.224 |
| Search-R1-base | 0.341 | 0.513 | 0.362 | 0.263 | 0.273 | 0.076 | 0.211 | 0.292 |
| Search-R1-instruct | 0.323 | 0.537 | 0.364 | **0.308** | **0.336** | **0.105** | 0.315 | **0.327** |
| **LLaMA3.2-3b-Base/Instruct** | | | | | | | | |
| Direct Inference | 0.139 | 0.368 | 0.124 | 0.122 | 0.107 | 0.015 | 0.064 | 0.134 |
| CoT | 0.246 | 0.487 | 0.166 | 0.051 | 0.083 | 0.006 | 0.024 | 0.152 |
| IRCoT | 0.363 | 0.566 | 0.428 | 0.238 | 0.236 | 0.072 | 0.208 | 0.301 |
| Search-o1 | 0.107 | 0.203 | 0.093 | 0.132 | 0.117 | 0.035 | 0.176 | 0.123 |
| RAG | 0.317 | 0.551 | 0.337 | 0.234 | 0.118 | 0.034 | 0.064 | 0.237 |
| SFT | 0.320 | 0.341 | 0.122 | 0.206 | 0.257 | 0.064 | 0.120 | 0.204 |
| R1-base | 0.290 | 0.514 | 0.237 | 0.234 | **0.279** | 0.055 | 0.146 | 0.251 |
| R1-instruct | 0.384 | 0.549 | 0.228 | 0.238 | 0.269 | 0.074 | **0.315** | 0.294 |
| Search-R1-base | **0.394** | **0.596** | **0.437** | 0.280 | 0.264 | 0.056 | 0.105 | 0.305 |
| Search-R1-instruct | 0.357 | 0.578 | 0.378 | **0.314** | 0.233 | **0.090** | 0.306 | **0.322** |

Jin, et al. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. arxiv'25.


18

# Case studies

- SEARCH-R1 conduct multi-step reasoning, search, with self-verification and finally answer the question.
- Interleaved Reasoning and Retrieval Enhances Problem Analysis.
- Self-Verification though Iterative Retrieval.

**Question**: What type of profession does Chris Jericho and Gary Barlow have in common?
**Ground Truth**: musician
**SEARCH-R1**:
<think> I need to find the type of profession that Chris Jericho and Gary Barlow have in common. I'll search for it. </think>
<search> Chris Jericho and Gary Barlow profession </search>
<information> Doc 2(Title: "Chris Jericho") ""Do You Wanna Start a War"" (2014) and ""Judas"" (2017) are comprised entirely or predominantly of original compositions. He has also appeared on numerous TV shows over the years, having competed in the 2011 season of ""Dancing With the Stars"", lasting until the sixth week. Prior to that, Jericho hosted the ABC game show ""Downfall"" and the Revolver Golden Gods Awards. Christopher Keith Irvine was born on November 9, 1970, in Manhasset, New York, but grew up in Winnipeg. His father, ice hockey player Ted Irvine, had been playing for the New York Rangers at the time of his birth. ... </information>
<think> I found out that Chris Jericho is a professional wrestler, musician, author, and actor. Now I need to find if Gary Barlow has any of these professions in common. </think>
<search> Gary Barlow profession </search>
<information> Doc 1(Title: "Gary Barlow") hits, twelve number one singles and seven number one albums with Take That. Barlow has received six Ivor Novello Awards from the British Academy of Songwriters, Composers and Authors, including the award for Outstanding Services to British Music. He has sold over 50 million records worldwide. Barlow was born on 20 January 1971 in Frodsham, Cheshire, the second son of Colin (died 15 October 2009) and Marjorie Barlow. In Barlow's autobiography, he relates that his love of music began at an early age, and calls himself ""one of those kids that's forever dancing in front of the TV looking at ... </information>
<think> I found out that Gary Barlow is a singer, songwriter, musician, and record producer. Now I need to find if there is a common profession between Chris Jericho and Gary Barlow. </think>
<search> common profession of Chris Jericho and Gary Barlow </search>
<information> Doc 1(Title: "Gary Barlow") Gary Barlow Gary Barlow (born 20 January 1971) is an English singer, songwriter, musician and record producer. He is best known as the lead singer of the British pop group Take That. Barlow also served as head judge of ""The X Factor UK"" in 2011, 2012 and 2013 and ""Let It Shine"" in 2017. Barlow is one of Britain's most successful songwriters, having written fourteen number one singles and twenty-four top 10 hits. As a solo artist he has had three number one singles, six top 10 singles and two number one albums, and has additionally had seventeen top 5 ... </information>
<think> I found out that Gary Barlow is a singer, songwriter, musician, and record producer. Now I need to find if Chris Jericho has any of these professions in common. </think>
<search> Chris Jericho profession </search>
<information> Doc 1(Title: "Chris Jericho") Chris Jericho Christopher Keith Irvine (born November 9, 1970), better known by the ring name Chris Jericho, is an American-born professional wrestler, musician, author and actor. He currently wrestles for New Japan Pro-Wrestling (NJPW) where he is the current IWGP Intercontinental Champion in his first reign. Jericho also makes sporadic appearances for WWE on the SmackDown brand, with whom he has worked since 1999. Within professional wrestling, he is known for his over-the-top, rock star persona. Jericho is regarded as one of the greatest professional wrestlers of all time. During the 1990s, Jericho performed for the American organizations Extreme Championship ... </information>
<think> I found out that Chris Jericho is a professional wrestler, musician, author, and actor. Now I can compare it to Gary Barlow's professions. I found out that Gary Barlow is a singer, songwriter, musician, and record producer. The common profession between Chris Jericho and Gary Barlow is musician. </think>
<answer> musician </answer>

**Question**: Curious is a women's fragrance by a singer born in what city and state?
**Ground Truth**: McComb, Mississippi
**R1**:
<think> The singer is Beyoncé, who was born in Houston, Texas. </think>
<answer> Houston </answer>
**SEARCH-R1**:
<think> I need to find out which city and state a singer, by whom the women's fragrance "Curious" was created, was born in. </think>
<search> Curious fragrance information </search>
<information> Doc 1(Title: "Curious (fragrance)") Curious (fragrance) Curious is a women's fragrance by Britney Spears for Elizabeth Arden, and is the first perfume to be endorsed by Britney Spears, preceding ""Fantasy."" Spears reportedly earned $52 million from the endorsement. ""Curious"" was released in September 2004, and was very successful internationally. It was the number one fragrance of 2004 in department stores, and in 2005, ""Curious"" was honored by the Fragrance Foundation as Best Women's Fragrance. The fragrance is a white floral scent. Its notes are Louisiana magnolia, golden Anjou pear, lotus flower, tuberose, star jasmine, pink cyclamen, vanilla-infused musk, sandalwood and blonde woods. Elizabeth Arden ... </information>
<think> I need to find out which city and state Britney Spears, the singer behind "Curious" fragrance, was born in. </think>
<search> Britney Spears birthplace </search>
<information> Doc 1(Title: "Britney Spears") the best-selling female albums artist of the 2000s. In 2004, she launched a perfume brand with Elizabeth Arden, Inc., from which sales exceeded US$1.5 billion, . Spears was born in McComb, Mississippi, the second child of Lynne Irene Bridges and James Parnell Spears. Her maternal grandmother, Lillian Portell, was English (born in London), and one of Spears's maternal great-great-grandfathers was Maltese. Her siblings are Bryan James and Jamie Lynn. Britney was born in the Bible Belt, where socially conservative evangelical Protestantism is a particularly strong religious influence. Spears was baptized into the Southern Baptist Convention, but later in life studied </information>
<think> Since Britney Spears was born in McComb, Mississippi, I need to find out which city and state that is. </think>
<search> McComb, Mississippi location </search>
<information> Doc 1(Title: "McComb, Mississippi") McComb, Mississippi McComb is a city in Pike County, Mississippi, United States, approximately south of Jackson. As of the 2010 census, the city had a total population of 12,790. It is the principal city of the McComb, Mississippi Micropolitan Statistical Area. McComb was founded in 1872 after Henry Simpson McComb of the New Orleans, Jackson and Great Northern Railroad, a predecessor of the Illinois Central Railroad (now part of the Canadian National Railway), decided to move the railroad's maintenance shops away from New Orleans, Louisiana, to avoid the attractions of that city's saloons. The railroad purchased land in Pike County, </information>
<think> I have found that Britney Spears was born in McComb, Mississippi. Since the question asks for the city and state, I just need to confirm the state, which I already know from the previous search. </think>
<answer> McComb, Mississippi </answer>

Jin, et al. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. arxiv'25.


19

# Search-R1 Resources

Code

X

Paper

Data / Model

Logs



20

# Outline

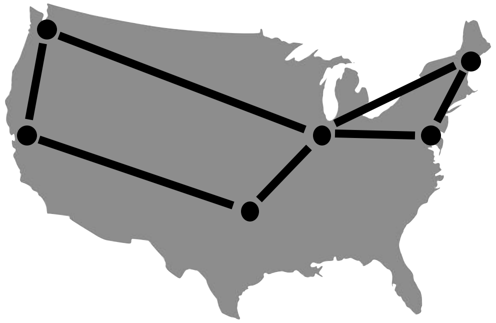
- Improving LLMs for search-augmented scenarios
  - From General-Purpose Models to RAG-Optimized LLMs
  - Beyond RAG Pipelines: Towards Search-Enabled LLM Agents

- Improving LLMs for graph-augmented scenarios
  - Pretrained LLM Retrievers for Graph-Augmented Search
  - Reasoning over Graph-Structured Knowledge with LLMs

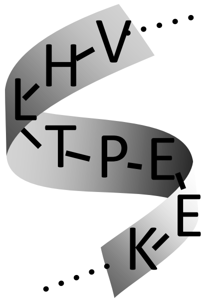
# Graphs

- Graph data is ubiquitous in real world.

Traffic Graphs

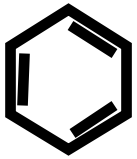


Protein Graphs

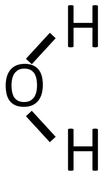



"Myoglobin holds oxygen in muscles."

Molecule Graphs

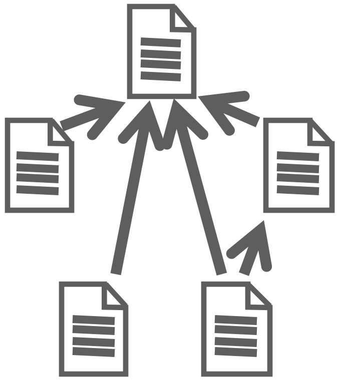


"Benzene is toxic"

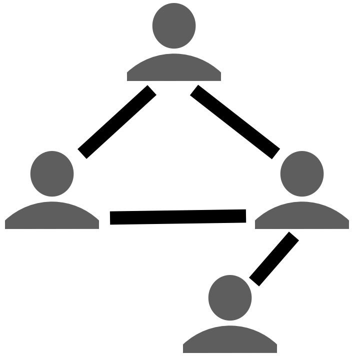



"Water is less toxic"

Academic Graphs



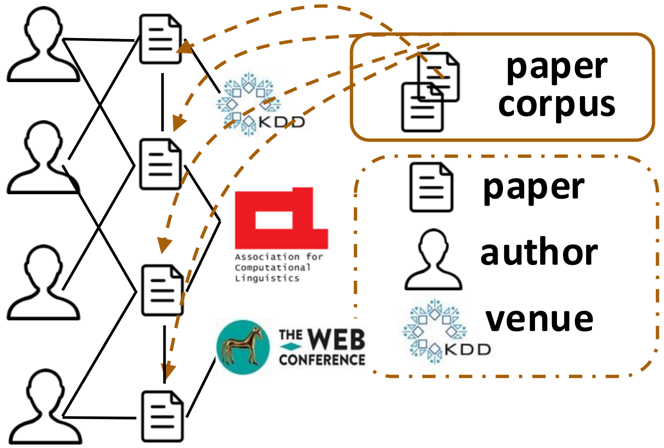
Social Graphs

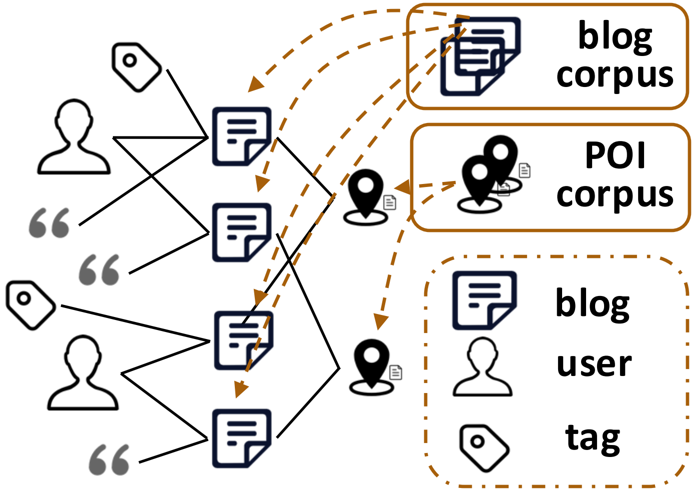


# Why LLM on Graphs?

- In real world, text and graph usually appears simultaneously.
  - Text data are associated with rich structure information in the form of graphs.

Academic Network

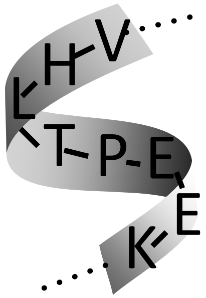


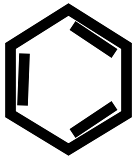

Social Media Network

  - Graph data are captioned with rich textual information.

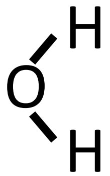
Protein Graphs




"Myoglobin holds oxygen in muscles."



"Benzene is toxic"

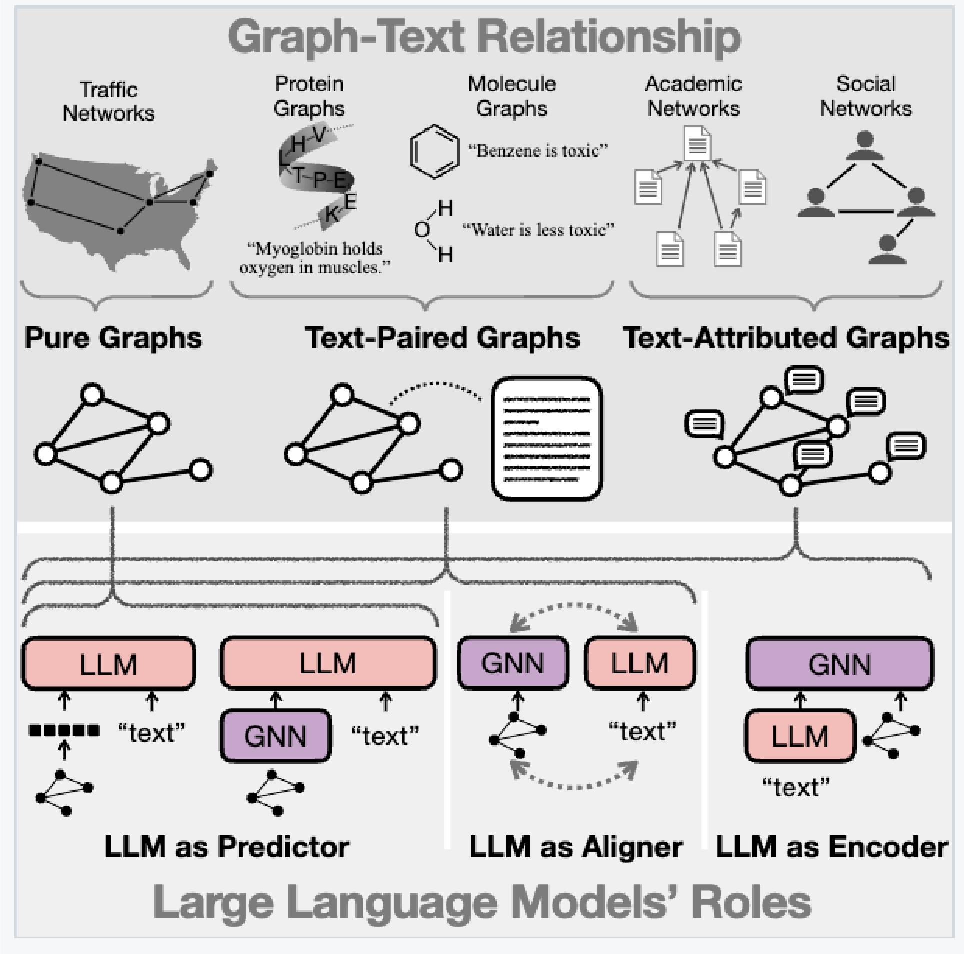



"Water is less toxic"

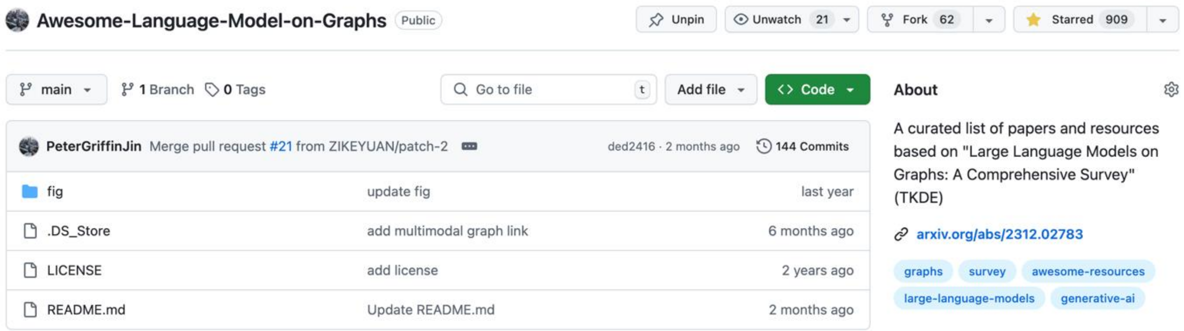
Molecule Graphs

# Survey

- A survey paper of LLM & graphs




- A resource repo of LLM & graphs




paper

repo



Jin, et al. Large Language Model on Graphs: A Comprehensive Survey. TKDE. 2024.

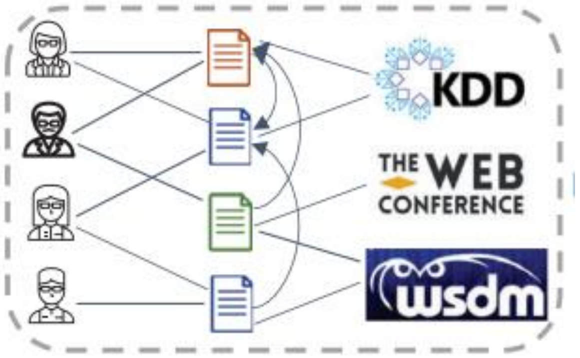
# Outline

- Improving LLMs for search-augmented scenarios
  - From General-Purpose Models to RAG-Optimized LLMs
  - Beyond RAG Pipelines: Towards Search-Enabled LLM Agents
- Improving LLMs for graph-augmented scenarios
  - Pretrained LLM Retrievers for Graph-Augmented Search
  - Reasoning over Graph-Structured Knowledge with LLMs

# Representation learning on text-attributed graphs

- Given a text-attributed network, people are interested in various tasks.
  - Node classification, link prediction, and node clustering.

- Learn representations for nodes/edges which can be utilized in various tasks.
  - Textual information & structure information




→ $f(x)$ Model → $Z$ Embeddings $g(z)$ →

$R$

Clustering

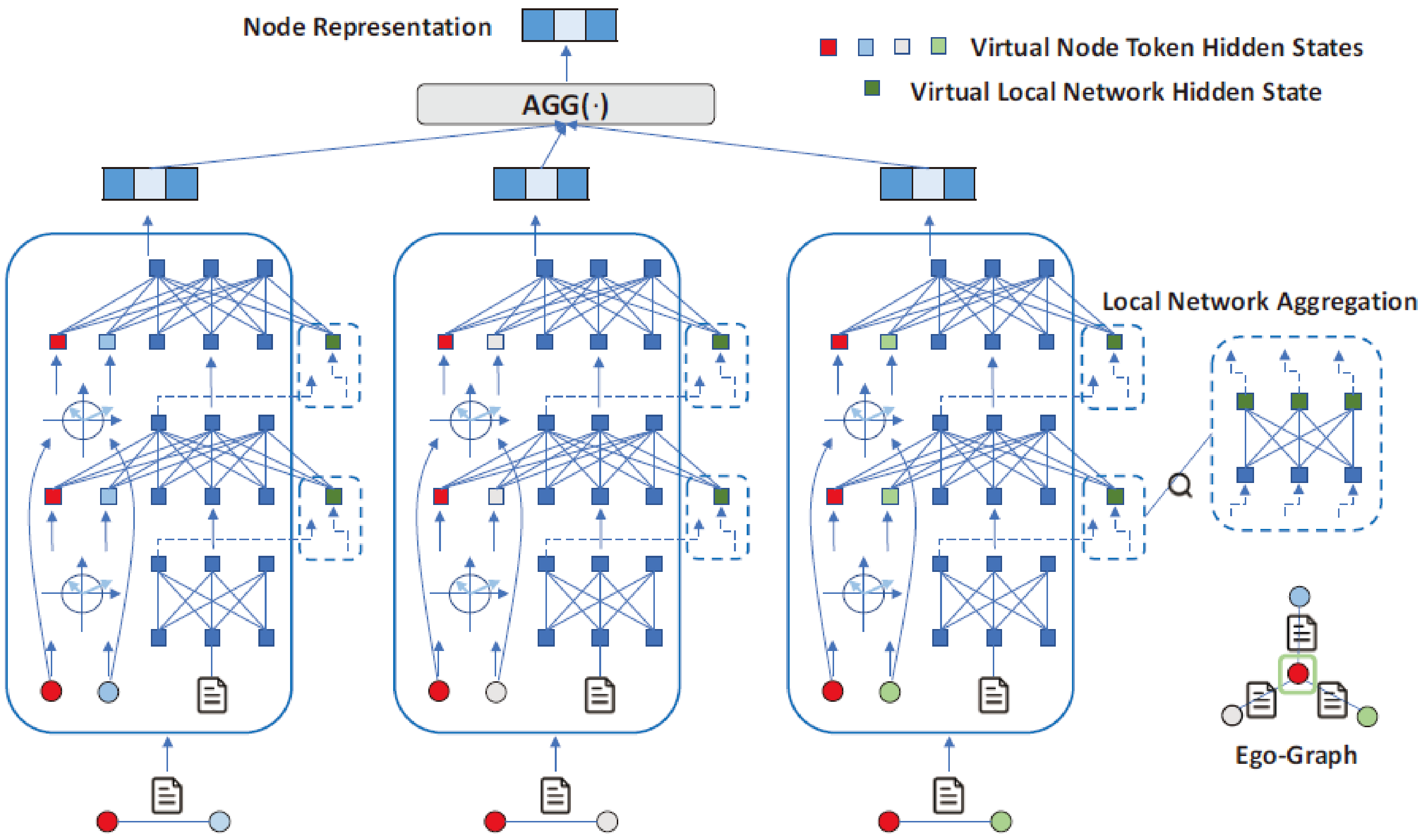
Classification

Link prediction

$$X \xrightarrow{f(x)} Z \xrightarrow{g(z)} R$$

# Graph-Empowered LLM: Edgeformers

- Text-aware node representation learning (Edgeformer-N)
  - Aggregate edge representations
  - Enhance edge representations with node's local network structure




Jin, et al. Edgeformers: Graph-Empowered Transformers for Representation Learning on Textual-Edge Networks. ICLR'23.

# Graph-Empowered LLM: Edgeformers

## ❑ Edge classification

Table 7: Edge classification performance on Amazon-Movie, Amazon-App, Goodreads-Crime, and Goodreads-Children.

| | Amazon-Movie | | Amazon-Apps | | Goodreads-Crime | | Goodreads-Children | |
|---|---|---|---|---|---|---|---|---|
| Model | Macro-F1 | Micro-F1 | Macro-F1 | Micro-F1 | Macro-F1 | Micro-F1 | Macro-F1 | Micro-F1 |
| TF-IDF | 50.01 | 64.22 | 48.30 | 62.88 | 43.07 | 51.72 | 39.42 | 49.90 |
| TF-IDF+nodes | 53.59 | 66.34 | 50.56 | 65.08 | 49.35 | 57.50 | 47.32 | 56.78 |
| EHGNN | 49.90 | 64.04 | 48.20 | 63.63 | 44.49 | 52.30 | 40.01 | 50.23 |
| BERT | 61.38 | 71.36 | 59.11 | 69.27 | 56.41 | 61.29 | 51.57 | 57.72 |
| BERT+nodes | *63.00* | *72.45* | *59.72* | *70.82* | *58.64* | *65.02* | *54.42* | *60.46* |
| BERT+EHGNN | 61.45 | 70.73 | 58.86 | 70.79 | 56.92 | 61.66 | 52.46 | 57.97 |
| BERT+MaxSAGE | 61.57 | 70.79 | 58.95 | 70.45 | 57.20 | 61.98 | 52.75 | 58.53 |
| GraphFormers | 61.73 | 71.52 | 59.67 | 70.19 | 57.49 | 62.37 | 52.93 | 58.34 |
| Edgeformer-E | **64.18** | **73.59** | **60.67** | **71.28** | **61.03** | **65.86** | **57.45** | **61.71** |

## ❑ Link prediction

| | Amazon-Movie | | Amazon-Apps | | Goodreads-Crime | | Goodreads-Children | | StackOverflow | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | MRR | NDCG | MRR | NDCG | MRR | NDCG | MRR | NDCG | MRR | NDCG |
| MF | 0.2032 | 0.3546 | 0.1482 | 0.3052 | 0.1923 | 0.3443 | 0.1137 | 0.2716 | 0.1040 | 0.2642 |
| MeanSAGE | 0.2138 | 0.3657 | 0.1766 | 0.3343 | 0.1832 | 0.3368 | 0.1066 | 0.2647 | 0.1174 | 0.2768 |
| MaxSAGE | 0.2178 | 0.3694 | 0.1674 | 0.3258 | 0.1846 | 0.3387 | 0.1066 | 0.2647 | 0.1173 | 0.2769 |
| GIN | 0.2140 | 0.3648 | 0.1797 | 0.3362 | 0.1846 | 0.3374 | 0.1128 | 0.2700 | 0.1189 | 0.2778 |
| CensNet | 0.2048 | 0.3568 | 0.1894 | 0.3457 | 0.1880 | 0.3398 | 0.1157 | 0.2726 | 0.1235 | 0.2806 |
| NENN | 0.2565 | 0.4032 | 0.1996 | 0.3552 | 0.2173 | 0.3670 | 0.1297 | 0.2854 | 0.1257 | 0.2854 |
| BERT | 0.2391 | 0.3864 | 0.1790 | 0.3350 | 0.1986 | 0.3498 | 0.1274 | 0.2836 | 0.1666 | 0.3252 |
| BERT+MaxSAGE | 0.2780 | 0.4224 | 0.2055 | 0.3602 | 0.2193 | 0.3694 | 0.1312 | 0.2872 | 0.1681 | 0.3264 |
| BERT+MeanSAGE | 0.2491 | 0.3972 | 0.1983 | 0.3540 | 0.1952 | 0.3477 | 0.1223 | 0.2791 | 0.1678 | 0.3264 |
| BERT+GIN | 0.2573 | 0.4037 | 0.2000 | 0.3552 | 0.2007 | 0.3522 | 0.1238 | 0.2801 | *0.1708* | *0.3279* |
| GraphFormers | 0.2756 | 0.4198 | 0.2066 | 0.3607 | 0.2176 | 0.3684 | 0.1323 | 0.2887 | 0.1693 | 0.3278 |
| BERT+CensNet | 0.1919 | 0.3462 | 0.1544 | 0.3132 | 0.1437 | 0.3000 | 0.0847 | 0.2436 | 0.1173 | 0.2789 |
| BERT+NENN | *0.2821* | *0.4256* | *0.2127* | *0.3666* | *0.2262* | *0.3756* | *0.1365* | *0.2925* | 0.1619 | 0.3215 |
| Edgeformer-N | **0.2919** | **0.4344** | **0.2239** | **0.3771** | **0.2395** | **0.3875** | **0.1446** | **0.3000** | **0.1754** | **0.3339** |
| + Δ % | 3.5% | 2.1% | 5.3% | 2.9% | 5.9% | 3.2% | 5.9% | 2.6% | 2.7% | 1.8% |

## ❑ Node classification

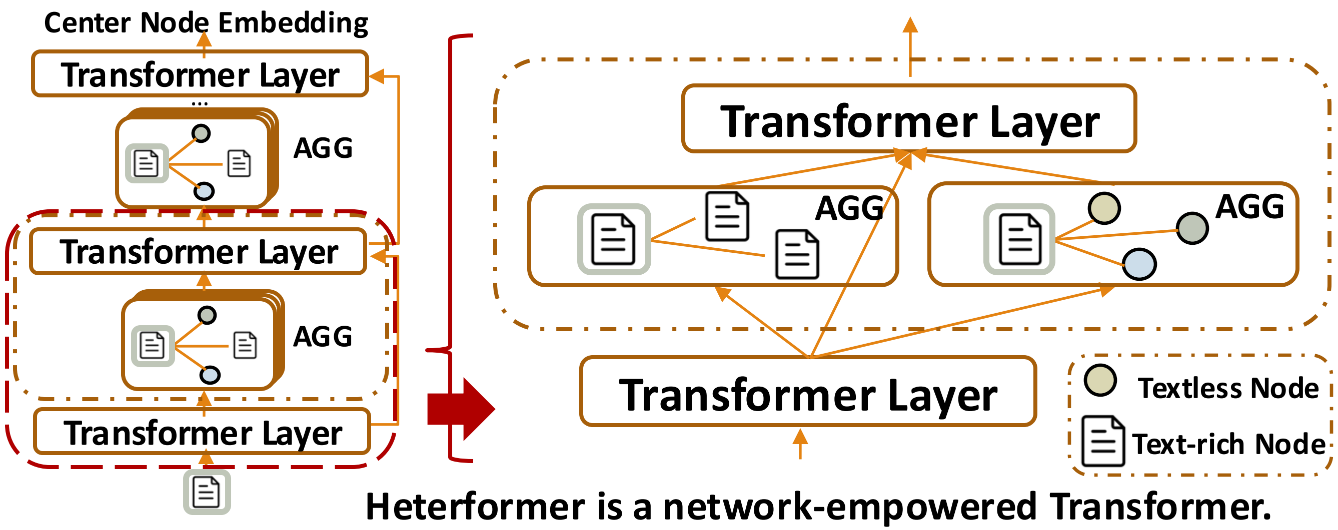
| | Amazon-Movie | | | Amazon-Apps | | |
|---|---|---|---|---|---|---|
| Model | Macro-F1 | Micro-F1 | PREC | Macro-F1 | Micro-F1 | PREC |
| MF | 0.7566±0.0017 | 0.8234±0.0013 | 0.8241±0.0013 | 0.4647±0.0151 | 0.8393±0.0012 | 0.8462±0.0006 |
| CensNet | 0.8528±0.0010 | 0.8839±0.0008 | 0.8845±0.0007 | 0.2782±0.0168 | 0.8279±0.0006 | 0.8331±0.0005 |
| NENN | 0.9186±0.0008 | 0.9341±0.0008 | 0.9347±0.0007 | 0.3408±0.0082 | 0.8789±0.0019 | 0.8819±0.0017 |
| BERT | 0.9209±0.0005 | 0.9361±0.0003 | 0.9367±0.0003 | 0.7608±0.0175 | 0.9283±0.0015 | 0.9337±0.0015 |
| BERT+CensNet | 0.9032±0.0006 | 0.9221±0.0004 | 0.9227±0.0004 | 0.5750±0.0277 | 0.8692±0.0034 | 0.8731±0.0028 |
| BERT+NENN | 0.9247±0.0005 | 0.9387±0.0004 | 0.9393±0.0005 | 0.7556±0.0092 | 0.9306±0.0008 | 0.9382±0.0006 |
| Edgeformer-N | **0.9276±0.0007** | **0.9411±0.0006** | **0.9417±0.0005** | **0.7758±0.0100** | **0.9339±0.0007** | **0.9431±0.0005** |



Jin, et al. Edgeformers: Graph-Empowered Transformers for Representation Learning on Textual-Edge Networks. ICLR'23.

# Graph-Empowered LLM: Heterformer

- Overall framework
  - Heterformer: a graph-empowered Transformer.
  - Unifying text semantic encoding and network signal capturing.




Heterformer is a network-empowered Transformer.



Jin, et al. Heterformer: Transformer-based Deep Node Representation Learning on Heterogeneous Text-Rich Networks. KDD'23.

# Graph-Empowered LLM: Heterformer

## ❑ Link prediction

| | Method | DBLP PREC | DBLP MRR | DBLP NDCG | Twitter PREC | Twitter MRR | Twitter NDCG | Goodreads PREC | Goodreads MRR | Goodreads NDCG |
|---|---|---|---|---|---|---|---|---|---|---|
| | MeanSAGE | 0.7019 | 0.7964 | 0.8437 | 0.6489 | 0.7450 | 0.7991 | 0.6302 | 0.7409 | 0.8001 |
| | BERT | 0.7569 | 0.8340 | 0.8726 | 0.7179 | 0.7833 | 0.8265 | 0.5571 | 0.6668 | 0.7395 |
| Homo GNN | BERT+MeanSAGE | 0.8131 | 0.8779 | 0.9070 | 0.7201 | 0.7845 | 0.8275 | 0.7301 | 0.8167 | 0.8594 |
| | BERT+MAXSAGE | 0.8193 | 0.8825 | 0.9105 | 0.7198 | 0.7845 | 0.8276 | 0.7280 | 0.8164 | 0.8593 |
| | BERT+GAT | 0.8119 | 0.8771 | 0.9063 | 0.7231 | 0.7873 | 0.8300 | 0.7333 | 0.8170 | 0.8593 |
| | GraphFormers | 0.8324 | 0.8916 | 0.9175 | 0.7258 | 0.7891 | 0.8312 | 0.7444 | 0.8260 | 0.8665 |
| Hetero GNN | BERT+RGCN | 0.7979 | 0.8633 | 0.8945 | 0.7111 | 0.7764 | 0.8209 | 0.7488 | 0.8303 | 0.8699 |
| | BERT+HAN | 0.8136 | 0.8782 | 0.9072 | 0.7237 | 0.7880 | 0.8306 | 0.7329 | 0.8174 | 0.8597 |
| | BERT+HGT | 0.8170 | 0.8814 | 0.9098 | 0.7153 | 0.7800 | 0.8237 | 0.7224 | 0.8112 | 0.8552 |
| | BERT+SHGN | 0.8149 | 0.8785 | 0.9074 | 0.7218 | 0.7866 | 0.8295 | 0.7362 | 0.8195 | 0.8613 |
| | GraphFormers++ | 0.8233 | 0.8856 | 0.9130 | 0.7159 | 0.7799 | 0.8236 | 0.7536 | 0.8328 | 0.8717 |
| | Heterformer | **0.8474*** | **0.9019*** | **0.9255*** | **0.7272*** | **0.7908*** | **0.8328*** | **0.7633*** | **0.8400*** | **0.8773*** |

## ❑ Node classification

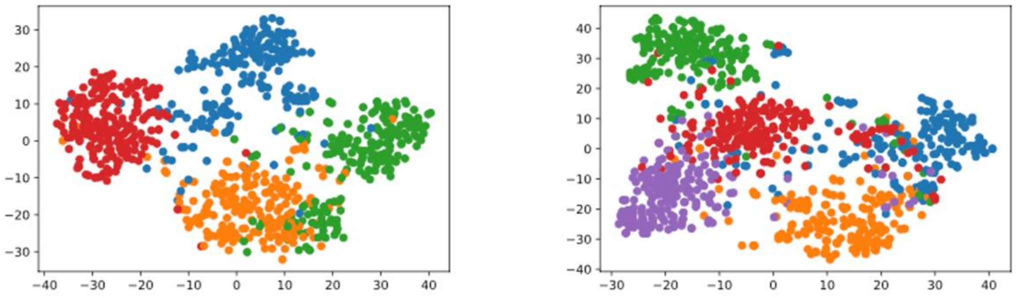
**Table 3: Transductive text-rich node classification.**

| Method | DBLP Micro-F1 | DBLP Macro-F1 | Goodreads Micro-F1 | Goodreads Macro-F1 |
|---|---|---|---|---|
| BERT | 0.6119 | 0.5476 | 0.8364 | 0.7713 |
| BERT+MaxSAGE | 0.6179 | 0.5511 | 0.8447 | 0.7866 |
| BERT+MeanSAGE | 0.6198 | 0.5522 | 0.8420 | 0.7826 |
| BERT+GAT | 0.5943 | 0.5175 | 0.8328 | 0.7713 |
| GraphFormers | 0.6256 | 0.5616 | 0.8388 | 0.7786 |
| BERT+HAN | 0.5965 | 0.5211 | 0.8351 | 0.7747 |
| BERT+HGT | 0.6575 | 0.5951 | 0.8474 | 0.7928 |
| BERT+SHGN | 0.5982 | 0.5214 | 0.8345 | 0.7737 |
| GraphFormers++ | 0.6474 | 0.5790 | 0.8516 | 0.7993 |
| Heterformer | **0.6695*** | **0.6062*** | **0.8578*** | **0.8076*** |

## ❑ Node clustering

| Method | DBLP NMI | DBLP ARI | Goodreads NMI | Goodreads ARI |
|---|---|---|---|---|
| BERT | 0.2570 | 0.3349 | 0.2325 | 0.4013 |
| BERT+MaxSAGE | 0.2615 | 0.3490 | 0.2205 | 0.4173 |
| BERT+MeanSAGE | 0.2628 | 0.3488 | **0.2449** | **0.4329** |
| BERT+GAT | 0.2598 | 0.3419 | 0.2408 | 0.4185 |
| GraphFormers | 0.2633 | 0.3455 | 0.2362 | 0.4139 |
| BERT+HAN | 0.2568 | 0.3401 | 0.2391 | 0.4266 |
| BERT+HGT | 0.2469 | 0.3392 | 0.2427 | 0.4296 |
| BERT+SHGN | 0.2589 | 0.3431 | 0.2373 | 0.4171 |
| GraphFormers++ | 0.2566 | 0.3432 | 0.2372 | 0.4211 |
| Heterformer | **0.2707*** | **0.3639*** | 0.2429 | 0.4199 |

## ❑ Embedding visualization



(a) DBLP (b) Goodreads

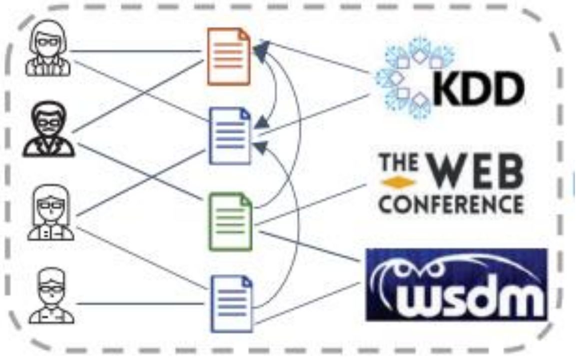
Jin, et al. Heterformer: Transformer-based Deep Node Representation Learning on Heterogeneous Text-Rich Networks. KDD’23.

# Why do we need language model pretraining on network?

- Given a text-rich network, people are interested in various downstream tasks
  - Document/node classification, document retrieval and link prediction

- Text-attributed network contains rich unsupervised semantic information
  - Alleviate human labeling burden for downstream tasks

PLM → 




→ G-Adapted PLM → Retrieval, Reranking, Classification, Link prediction

**Pretraining** on a Text-attributed Network G

**Finetuning** on downstream tasks

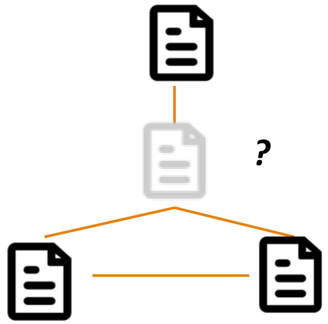
Jin, et al. Patton: Language Model Pretraining on Text-rich Networks. ACL'23.

# Language Model Pretraining: Patton


- Pretraining strategy 1: Network-contextualized masked language modeling
  - In node MLM -> Network contextualized MLM
  - Use both in-node text context and neighbor node context to conduct masked token prediction
  - Facilitate the LM to understand both in-node token correlation and network-contextualized text semantic relatedness

$$\mathcal{L}_{\text{NMLM}} = -\sum_{i \in M_t} \log p(w_i | \boldsymbol{H}_x, \boldsymbol{z}_x),$$

- Pretraining strategy 2: Masked Node Prediction
  - We dynamically hold out a subset of nodes from the network ($M_v \subseteq V$), mask them, and train the LM to predict the masked nodes based on the adjacent network structure.
  - LM will absorb document semantic hints hidden inside the network structure.




$$\mathcal{L}_{\text{MNP}} = -\sum_{v_j \in M_v} \log p(v_j | \boldsymbol{G}_{v_j})$$

Jin, et al. Patton: Language Model Pretraining on Text-rich Networks. ACL'23.

# Language Model Pretraining: Patton

## Retrieval

Table 3: Experiment results on Retrieval. We show the $\text{mean}_{\text{std}}$ of three runs for all the methods.

| Method | Mathematics R@50 | Mathematics R@100 | Geology R@50 | Geology R@100 | Economy R@50 | Economy R@100 | Clothes R@50 | Clothes R@100 | Sports R@50 | Sports R@100 |
|---|---|---|---|---|---|---|---|---|---|---|
| BM25 | 20.76 | 24.55 | 19.02 | 20.92 | 19.14 | 22.49 | 15.76 | 15.88 | 22.00 | 23.96 |
| BERT | $16.73_{0.17}$ | $22.66_{0.18}$ | $18.82_{0.39}$ | $25.94_{0.39}$ | $23.95_{0.25}$ | $31.54_{0.21}$ | $40.77_{1.68}$ | $50.40_{1.41}$ | $32.37_{1.09}$ | $43.32_{0.96}$ |
| GraphFormers | $16.65_{0.12}$ | $22.41_{0.10}$ | $18,92_{0.60}$ | $25.94_{0.39}$ | $24.48_{0.36}$ | $32.16_{0.40}$ | $41.77_{2.05}$ | $51.26_{2.27}$ | $32.39_{0.89}$ | $43.29_{1.12}$ |
| SciBERT | $24.70_{0.17}$ | $33.55_{0.31}$ | $23.71_{0.89}$ | $30.94_{0.95}$ | $29.80_{0.66}$ | $38.66_{0.52}$ | - | - | - | - |
| SPECTER | $23.86_{0.25}$ | $31.11_{0.31}$ | $26.56_{1.05}$ | $34.04_{1.32}$ | $31.26_{0.15}$ | $40.79_{0.11}$ | - | - | - | - |
| SimCSE (unsup) | $17.91_{0.26}$ | $23.19_{0.29}$ | $20.45_{0.20}$ | $26.82_{0.26}$ | $25.83_{0.23}$ | $33.42_{0.28}$ | $44.90_{0.35}$ | $54.76_{0.38}$ | $38.81_{0.35}$ | $49.30_{0.44}$ |
| SimCSE (sup) | $20.29_{0.41}$ | $26.23_{0.51}$ | $22.34_{0.49}$ | $29.63_{0.55}$ | $28.07_{0.38}$ | $36.51_{0.37}$ | $44.69_{0.59}$ | $54.70_{0.77}$ | $40.31_{0.43}$ | $50.55_{0.41}$ |
| LinkBERT | $17.25_{0.30}$ | $23.21_{0.47}$ | $17.14_{0.75}$ | $23.05_{0.74}$ | $22.69_{0.30}$ | $30.77_{0.36}$ | $28.66_{2.97}$ | $37.79_{3.82}$ | $31.97_{0.54}$ | $41.77_{0.67}$ |
| BERT.MLM | $20.69_{0.21}$ | $27.17_{0.25}$ | $32.13_{0.36}$ | $41.74_{0.42}$ | $27.13_{0.04}$ | $36.00_{0.14}$ | $52.41_{1.71}$ | $63.72_{1.79}$ | $54.10_{0.81}$ | $63.14_{0.83}$ |
| SciBERT.MLM | $20.65_{0.21}$ | $27.67_{0.32}$ | $31.65_{0.71}$ | $40.52_{0.76}$ | $29.23_{0.67}$ | $39.18_{0.73}$ | - | - | - | - |
| SimCSE.in-domain | $24.54_{0.05}$ | $31.66_{0.09}$ | $33.97_{0.07}$ | $44.09_{0.19}$ | $28.44_{0.31}$ | $37.81_{0.27}$ | $61.42_{0.84}$ | $72.25_{0.86}$ | $53.77_{0.22}$ | $63.73_{0.30}$ |
| PATTON | $27.44_{0.15}$ | $34.97_{0.21}$ | $34.94_{0.23}$ | $45.01_{0.28}$ | $32.10_{0.51}$ | $42.19_{0.62}$ | $\mathbf{68.62}_{0.38}$ | $\mathbf{77.54}_{0.19}$ | $\mathbf{58.63}_{0.31}$ | $\mathbf{68.53}_{0.55}$ |
| SciPATTON | $\mathbf{31.40}_{0.52}$ | $\mathbf{40.38}_{0.66}$ | $\mathbf{40.69}_{0.52}$ | $\mathbf{51.31}_{0.48}$ | $\mathbf{35.82}_{0.69}$ | $46.05_{0.69}$ | - | - | - | - |
| w/o NMLM | $30.85_{0.14}$ | $39.89_{0.23}$ | $39.29_{0.07}$ | $49.59_{0.11}$ | $35.17_{0.31}$ | $\mathbf{46.07}_{0.20}$ | $65.60_{0.26}$ | $75.19_{0.32}$ | $57.05_{0.14}$ | $67.22_{0.12}$ |
| w/o MNP | $22.47_{0.07}$ | $30.20_{0.15}$ | $31.28_{0.89}$ | $40.54_{0.97}$ | $29.54_{0.36}$ | $39.57_{0.57}$ | $60.20_{0.73}$ | $69.85_{0.52}$ | $51.73_{0.41}$ | $60.35_{0.78}$ |

## Classification

Table 2: Experiment results on Classification. We show the $\text{mean}_{\text{std}}$ of three runs for all the methods.

| Method | Mathematics Macro-F1 | Mathematics Micro-F1 | Geology Macro-F1 | Geology Micro-F1 | Economy Macro-F1 | Economy Micro-F1 | Clothes Macro-F1 | Clothes Micro-F1 | Sports Macro-F1 | Sports Micro-F1 |
|---|---|---|---|---|---|---|---|---|---|---|
| BERT | $18.14_{0.07}$ | $22.04_{0.32}$ | $21.97_{0.87}$ | $29.63_{0.36}$ | $14.17_{0.08}$ | $19.77_{0.12}$ | $45.10_{1.47}$ | $68.54_{2.25}$ | $31.88_{0.23}$ | $34.58_{0.56}$ |
| GraphFormers | $18.69_{0.52}$ | $23.24_{0.46}$ | $22.64_{0.92}$ | $31.02_{1.16}$ | $13.68_{1.03}$ | $19.00_{1.44}$ | $46.27_{1.92}$ | $68.97_{2.46}$ | $43.77_{0.63}$ | $50.47_{0.78}$ |
| SciBERT | $23.50_{0.64}$ | $23.10_{2.23}$ | $29.49_{1.25}$ | $37.82_{1.89}$ | $15.91_{0.48}$ | $21.32_{0.66}$ | - | - | - | - |
| SPECTER | $23.37_{0.07}$ | $29.83_{0.96}$ | $30.40_{0.48}$ | $38.54_{0.77}$ | $16.16_{0.17}$ | $19.84_{0.47}$ | - | - | - | - |
| SimCSE (unsup) | $20.12_{0.08}$ | $26.11_{0.39}$ | $38.78_{0.19}$ | $38.55_{0.17}$ | $14.54_{0.26}$ | $19.07_{0.43}$ | $42.70_{2.32}$ | $58.72_{0.34}$ | $41.91_{0.85}$ | $59.19_{0.55}$ |
| SimCSE (sup) | $20.39_{0.07}$ | $25.56_{0.00}$ | $25.66_{0.28}$ | $33.89_{0.40}$ | $15.03_{0.53}$ | $18.64_{1.32}$ | $52.82_{0.87}$ | $75.54_{0.98}$ | $46.69_{0.10}$ | $59.19_{0.55}$ |
| LinkBERT | $15.78_{0.91}$ | $19.75_{1.19}$ | $24.08_{0.58}$ | $31.32_{0.04}$ | $12.71_{0.12}$ | $16.39_{0.22}$ | $44.94_{2.52}$ | $65.33_{4.34}$ | $35.60_{0.33}$ | $38.30_{0.09}$ |
| BERT.MLM | $23.44_{0.39}$ | $31.75_{0.58}$ | $36.31_{0.36}$ | $48.04_{0.69}$ | $16.60_{0.21}$ | $22.71_{1.16}$ | $46.98_{0.84}$ | $68.00_{0.84}$ | $62.21_{0.13}$ | $75.43_{0.74}$ |
| SciBERT.MLM | $23.34_{0.42}$ | $30.11_{0.97}$ | $36.94_{0.28}$ | $46.54_{0.40}$ | $16.28_{0.38}$ | $21.41_{0.81}$ | - | - | - | - |
| SimCSE.in-domain | $25.15_{0.09}$ | $29.85_{0.20}$ | $38.91_{0.08}$ | $48.93_{0.14}$ | $18.08_{0.22}$ | $23.79_{0.44}$ | $57.03_{0.20}$ | $80.16_{0.31}$ | $65.57_{0.35}$ | $75.22_{0.18}$ |
| PATTON | $\mathbf{27.58}_{0.03}$ | $\mathbf{32.82}_{0.01}$ | $39.35_{0.06}$ | $48.19_{0.15}$ | $19.32_{0.05}$ | $25.12_{0.05}$ | $\mathbf{60.14}_{0.28}$ | $\mathbf{84.88}_{0.09}$ | $\mathbf{67.57}_{0.08}$ | $\mathbf{78.60}_{0.15}$ |
| SciPATTON | $27.35_{0.04}$ | $31.70_{0.01}$ | $\mathbf{39.65}_{0.10}$ | $\mathbf{48.93}_{0.06}$ | $\mathbf{19.91}_{0.08}$ | $\mathbf{25.68}_{0.32}$ | - | - | - | - |
| w/o NMLM | $25.91_{0.45}$ | $27.79_{2.07}$ | $38.78_{0.19}$ | $48.48_{0.17}$ | $18.86_{0.23}$ | $24.25_{0.26}$ | $56.68_{0.24}$ | $80.27_{0.17}$ | $65.83_{0.28}$ | $76.24_{0.54}$ |
| w/o MNP | $24.79_{0.65}$ | $29.44_{1.50}$ | $38.00_{0.73}$ | $47.82_{1.06}$ | $18.69_{0.59}$ | $25.63_{1.44}$ | $47.35_{1.20}$ | $68.50_{2.60}$ | $64.23_{1.53}$ | $76.03_{1.67}$ |

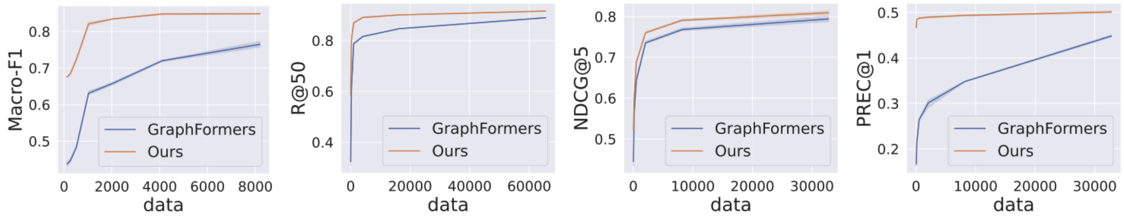
## Link prediction

Table 5: Experiment results on Link Prediction. We show the $\text{mean}_{\text{std}}$ of three runs for all the methods.

| Method | Mathematics PREC@1 | Mathematics MRR | Geology PREC@1 | Geology MRR | Economy PREC@1 | Economy MRR | Clothes PREC@1 | Clothes MRR | Sports PREC@1 | Sports MRR |
|---|---|---|---|---|---|---|---|---|---|---|
| BERT | $6.60_{0.16}$ | $12.96_{0.34}$ | $6.24_{0.76}$ | $12.96_{1.34}$ | $4.12_{0.08}$ | $9.23_{0.15}$ | $24.17_{0.41}$ | $34.20_{0.45}$ | $16.48_{0.45}$ | $25.35_{0.52}$ |
| GraphFormers | $6.91_{0.29}$ | $13.42_{0.34}$ | $6.52_{1.17}$ | $13.34_{1.81}$ | $4.16_{0.21}$ | $9.28_{0.28}$ | $23.79_{0.69}$ | $33.79_{0.66}$ | $16.69_{0.36}$ | $25.74_{0.48}$ |
| SciBERT | $14.08_{0.11}$ | $23.62_{0.10}$ | $7.15_{0.26}$ | $14.11_{0.39}$ | $5.01_{1.04}$ | $10.48_{1.79}$ | - | - | - | - |
| SPECTER | $13.44_{0.5}$ | $21.73_{0.65}$ | $6.85_{0.22}$ | $13.37_{0.34}$ | $6.33_{0.29}$ | $12.41_{0.33}$ | - | - | - | - |
| SimCSE (unsup) | $9.85_{0.10}$ | $16.28_{0.12}$ | $7.47_{0.55}$ | $14.24_{0.89}$ | $5.72_{0.26}$ | $11.02_{0.34}$ | $30.51_{0.09}$ | $40.40_{0.10}$ | $22.99_{0.07}$ | $32.47_{0.06}$ |
| SimCSE (sup) | $10.35_{0.52}$ | $17.01_{0.72}$ | $10.10_{0.04}$ | $17.80_{0.07}$ | $5.72_{0.26}$ | $11.02_{0.34}$ | $35.42_{0.06}$ | $46.07_{0.06}$ | $27.07_{0.15}$ | $37.44_{0.16}$ |
| LinkBERT | $8.05_{0.14}$ | $13.91_{0.09}$ | $6.40_{0.14}$ | $12.99_{0.17}$ | $2.97_{0.08}$ | $6.79_{0.15}$ | $30.33_{0.56}$ | $39.59_{0.64}$ | $19.83_{0.09}$ | $28.32_{0.04}$ |
| BERT.MLM | $17.55_{0.25}$ | $29.22_{0.26}$ | $14.13_{0.19}$ | $25.36_{0.20}$ | $9.02_{0.09}$ | $16.72_{0.15}$ | $42.71_{0.31}$ | $54.54_{0.35}$ | $29.36_{0.09}$ | $41.60_{0.05}$ |
| SciBERT.MLM | $22.44_{0.08}$ | $34.22_{0.05}$ | $16.22_{0.03}$ | $27.02_{0.07}$ | $9.80_{0.00}$ | $17.72_{0.01}$ | - | - | - | - |
| SimCSE.in-domain | $33.55_{0.05}$ | $46.07_{0.07}$ | $24.56_{0.06}$ | $36.89_{0.11}$ | $16.77_{0.10}$ | $26.93_{0.01}$ | $\mathbf{60.41}_{0.03}$ | $\mathbf{71.86}_{0.06}$ | $49.17_{0.04}$ | $63.48_{0.03}$ |
| PATTON | $70.41_{0.11}$ | $80.21_{0.04}$ | $44.76_{0.05}$ | $57.71_{0.04}$ | $57.04_{0.05}$ | $68.35_{0.04}$ | $58.59_{0.12}$ | $70.12_{0.12}$ | $46.68_{0.09}$ | $60.96_{0.23}$ |
| SciPATTON | $\mathbf{71.22}_{0.17}$ | $\mathbf{80.79}_{0.10}$ | $\mathbf{44.95}_{0.24}$ | $\mathbf{57.84}_{0.25}$ | $\mathbf{57.36}_{0.26}$ | $\mathbf{68.71}_{0.31}$ | - | - | - | - |
| w/o NMLM | $71.04_{0.13}$ | $80.60_{0.07}$ | $44.33_{0.23}$ | $57.29_{0.22}$ | $56.64_{0.25}$ | $68.12_{0.16}$ | $60.30_{0.03}$ | $71.67_{0.07}$ | $\mathbf{49.72}_{0.06}$ | $\mathbf{63.76}_{0.04}$ |
| w/o MNP | $63.06_{0.23}$ | $74.26_{0.11}$ | $33.84_{0.60}$ | $47.02_{0.65}$ | $44.46_{0.03}$ | $57.05_{0.04}$ | $49.62_{0.06}$ | $61.61_{0.01}$ | $36.05_{0.20}$ | $49.78_{0.25}$ |

## How pretraining help the model?

### Finetune data size study




(a) Classification (b) Retrieval (c) Reranking (d) Link Prediction

Jin, et al. Patton: Language Model Pretraining on Text-rich Networks. ACL'23.

# Outline

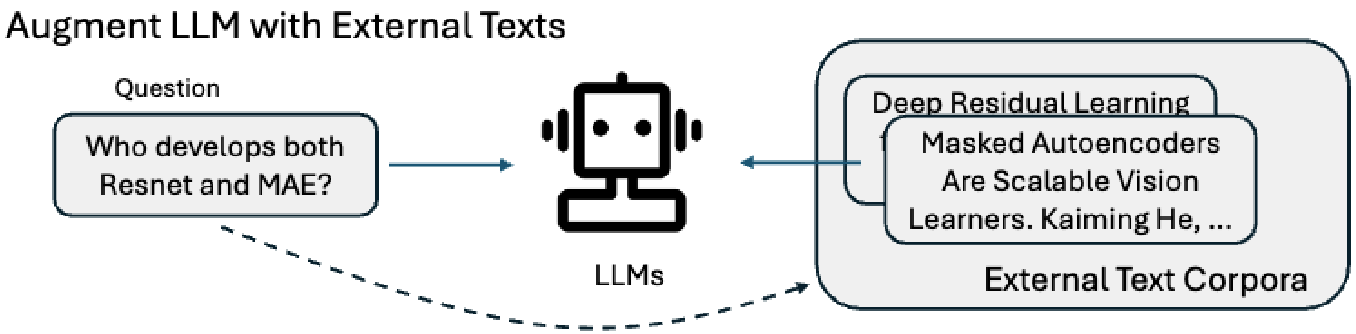
- Improving LLMs for search-augmented scenarios
  - From General-Purpose Models to RAG-Optimized LLMs
  - Beyond RAG Pipelines: Towards Search-Enabled LLM Agents

- Improving LLMs for graph-augmented scenarios
  - Pretrained LLM Retrievers for Graph-Augmented Search
  - Reasoning over Graph-Structured Knowledge with LLMs

# Augment LLM by reasoning on Graphs

- Retrieval-augmented generation (RAG)
  - Motivation
    - LLMs suffer from hallucination
    - External corpus can provide knowledge to mitigate hallucination
  - Pipeline
    - Retriever: fetch knowledge from corpus
    - LLM: inference

**What if the text units in the corpora is linked?**

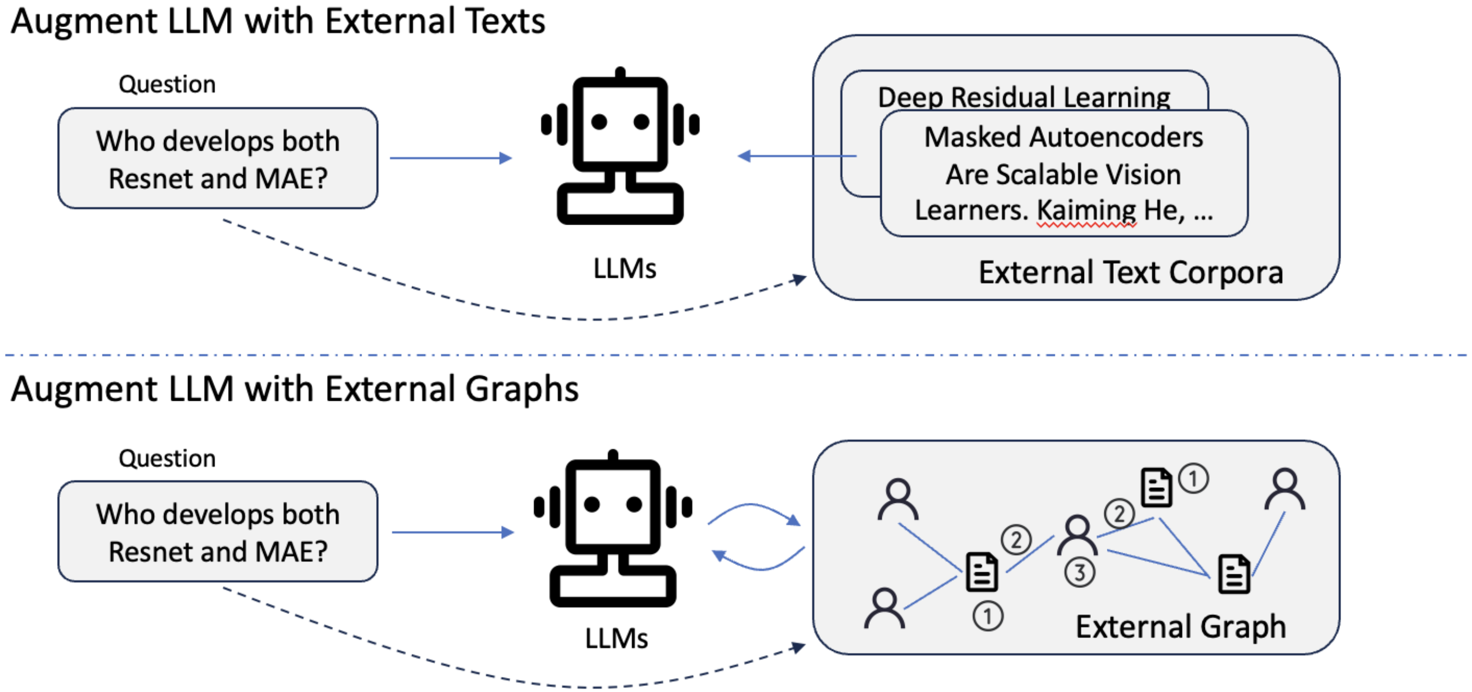



Jin, et al. Graph Chain-of-Thought: Augmenting Large Language Models by Reasoning on Graphs. ACL'24.

# Augment LLM by reasoning on Graphs

## ❑ Motivation

- ❑ This motivates us to explore the problem of augmenting LLMs with external graphs.

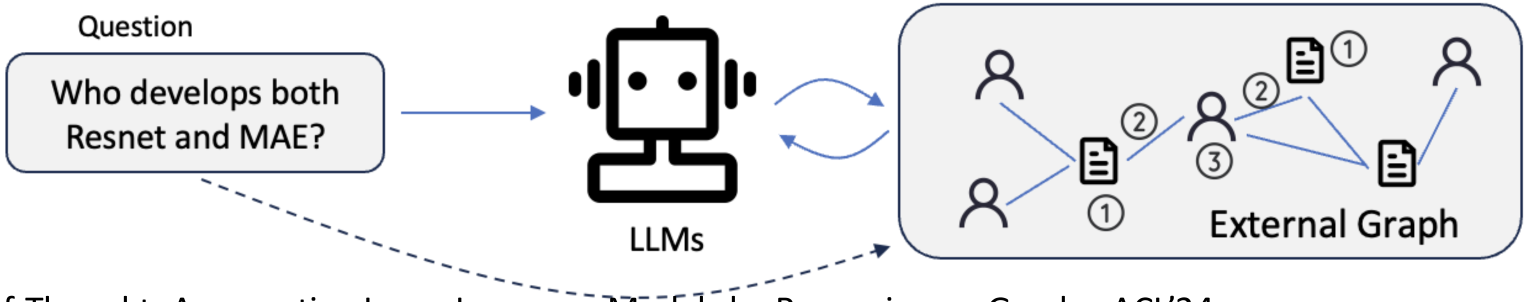



Jin, et al. Graph Chain-of-Thought: Augmenting Large Language Models by Reasoning on Graphs. ACL'24.


36

# Augment LLM by reasoning on Graphs

- **Can RAG be directly adopted for LLMs on graphs?**
  - Structure context:
    - Retrieval augmentation can find individual nodes/texts from the graphs.
    - However, knowledge on the graph also lies in the structure which cannot be captured by single nodes.
  - Graph size explosion:
    - It is feasible to convert local subgraph structure into text descriptions as the input contexts to LLMs.
    - However, the size of the local subgraph increases exponentially as the hop number increases.
    - It will result in an excessively long context sequence and cause LLM to be lost in the middle.

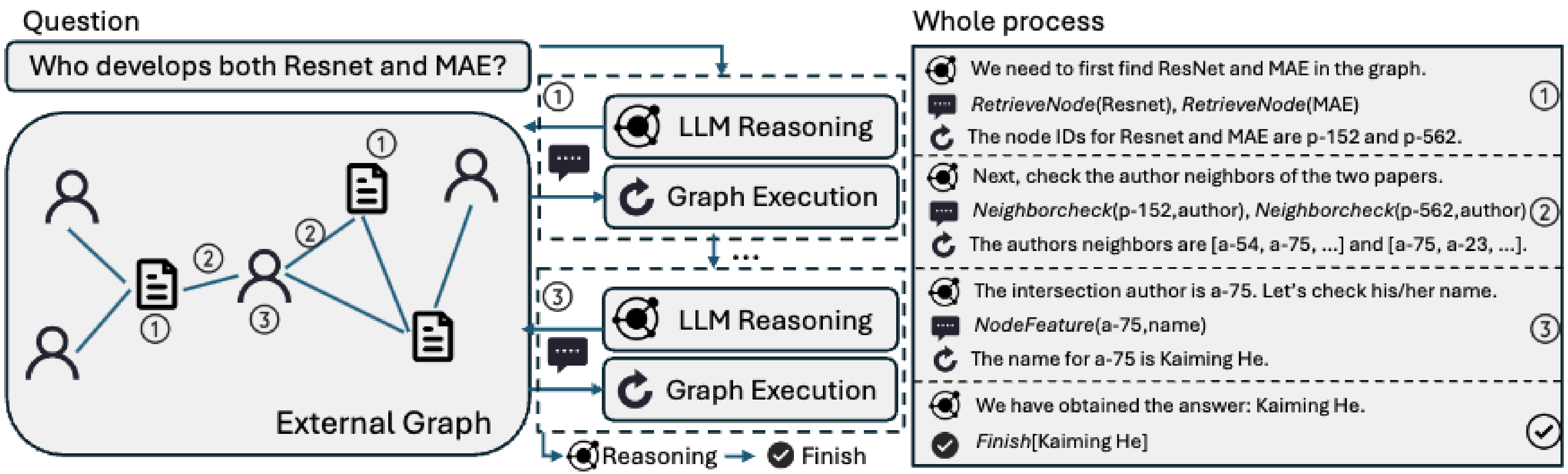



Jin, et al. Graph Chain-of-Thought: Augmenting Large Language Models by Reasoning on Graphs. ACL'24.


37

# Augment LLM by reasoning on Graphs

- Graph Chain-of-Thought
  - Iteratively traverse on graph & reasoning with LLM






Jin, et al. Graph Chain-of-Thought: Augmenting Large Language Models by Reasoning on Graphs. ACL'24.

# Experiments

## Overall performance

| | Model | Academic | | E-commerce | | Literature | | Healthcare | | Legal | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R-L | GPT4score | R-L | GPT4score | R-L | GPT4score | R-L | GPT4score | R-L | GPT4score |
| Base | LLaMA-2-13b-chat | 8.13 | 8.03 | 7.01 | 12.00 | 5.32 | 20.83 | 5.25 | 13.70 | 15.97 | 16.11 |
| | Mixtral-8x7b | 9.02 | 8.14 | 12.54 | 18.00 | 7.50 | 22.50 | 3.88 | 20.00 | 12.74 | 16.11 |
| | GPT-3.5-turbo | 6.05 | 12.80 | 9.18 | 23.50 | 10.43 | 26.67 | 5.83 | 14.44 | 10.51 | 20.00 |
| Text RAG | LLaMA-2-13b-chat | 8.69 | 8.52 | 9.23 | 12.50 | 7.61 | 20.00 | 1.44 | 5.93 | 15.37 | 16.67 |
| | Mixtral-8x7b | 8.44 | 8.02 | 23.14 | 29.50 | 13.35 | 27.92 | 3.22 | 16.67 | 19.69 | 25.00 |
| | GPT-3.5-turbo | 5.83 | 9.91 | 14.06 | 20.00 | 10.04 | 20.83 | 4.57 | 8.52 | 18.14 | 23.89 |
| Graph RAG | LLaMA-2-13b | 22.01 | 22.97 | 12.48 | 20.00 | 9.25 | 20.00 | 2.97 | 4.81 | 17.98 | 17.22 |
| | Mixtral-8x7b | 27.77 | 31.20 | 32.87 | 37.00 | 20.08 | 33.33 | 8.66 | 15.19 | 23.48 | 25.56 |
| | GPT-3.5-turbo | 18.45 | 26.98 | 17.52 | 28.00 | 14.94 | 24.17 | 8.69 | 14.07 | 18.66 | 22.22 |
| | GRAPH-COT | **31.89** | **33.48** | **42.40** | **44.50** | **41.59** | **46.25** | **22.33** | **28.89** | **30.52** | **28.33** |

- Graph-CoT outperforms all the baselines consistently and significantly.
- Base LLMs are exhibiting fairly poor performance, typically because the LLMs may not contain the knowledge needed to answer those questions.
- Graph RAG LLMs outperform text RAG LLMs in most cases since the former can provide more structure-aware context.

Jin, et al. Graph Chain-of-Thought: Augmenting Large Language Models by Reasoning on Graphs. ACL'24.

# Q&A