

Beyond Scaling: Frontiers of Retrieval-Augmented Language Models

Akari Asai

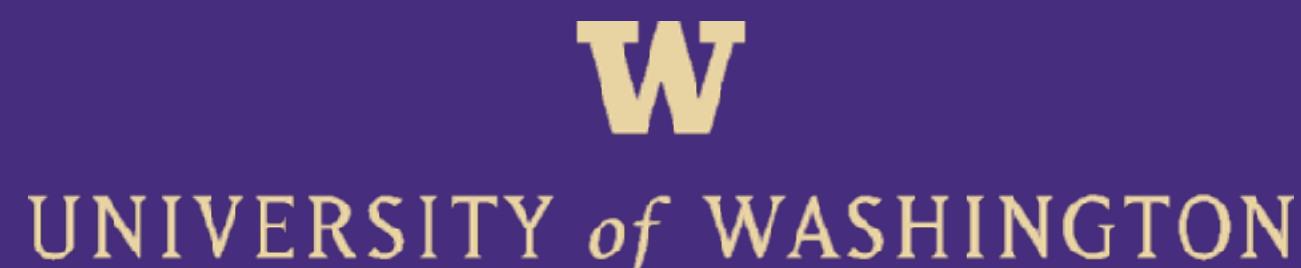
University of Washington



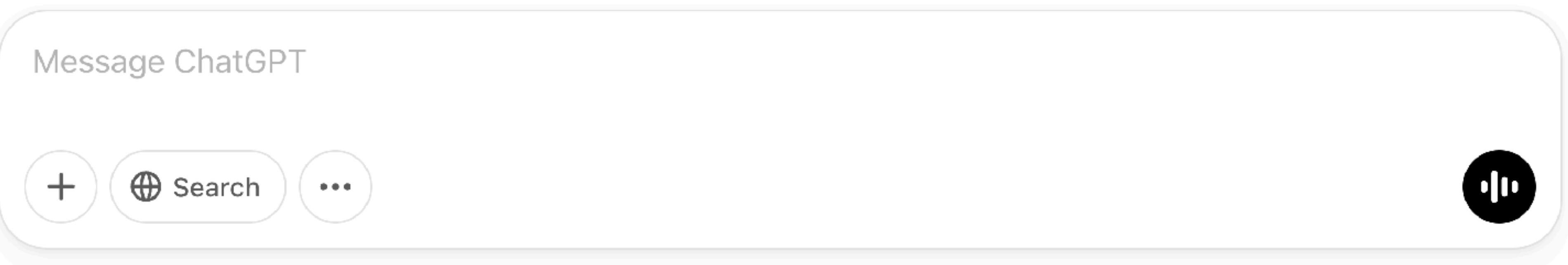
akariasai.github.io/



akari@cs.washington.edu



What can I help with?



What can I help with?

Message ChatGPT



Search

...



Please generate an image of a pug wearing a University of Washington T-shirt, standing in front of the iconic Suzzallo Library and looking ahead.



What can I help with?

Message ChatGPT



Search



Please generate an image of a pug wearing a University of Washington T-shirt, standing in front of the iconic Suzzallo Library and looking ahead.



generate a line chart showing the total number of users, by adding the user counts from each week

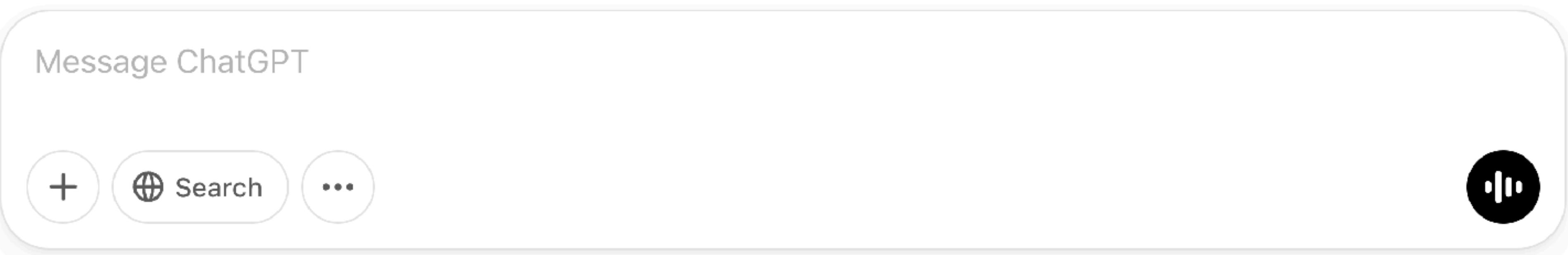
```
import matplotlib.pyplot as plt    Always show details  |  Copy

# Convert the 'date' column to datetime for easier manipulation
data['date'] = pd.to_datetime(data['date'])

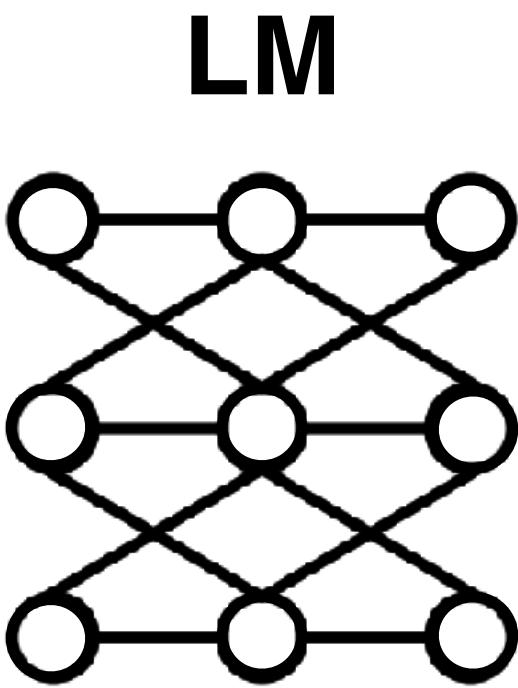
# Resample data to weekly frequency, summing the 'user_count'
weekly_user_counts = data.set_index('date').resample('W')['user_count'].sum()

# Plot the weekly user count as a line chart
plt.figure(figsize=(10, 6))
plt.plot(weekly_user_counts.index, weekly_user_counts.values, marker='o')
plt.title('Total Number of Users per Week')
plt.xlabel('Week')
plt.ylabel('Total Users')
plt.grid(True)
plt.tight_layout()
plt.show()
```

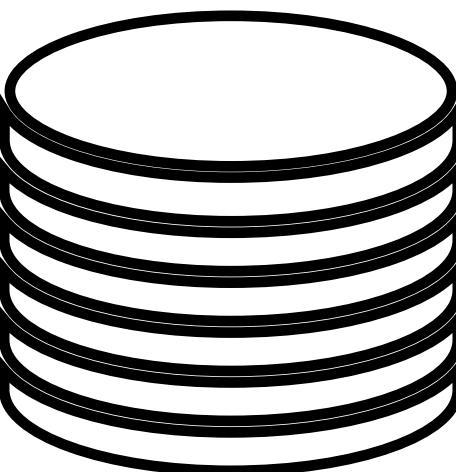
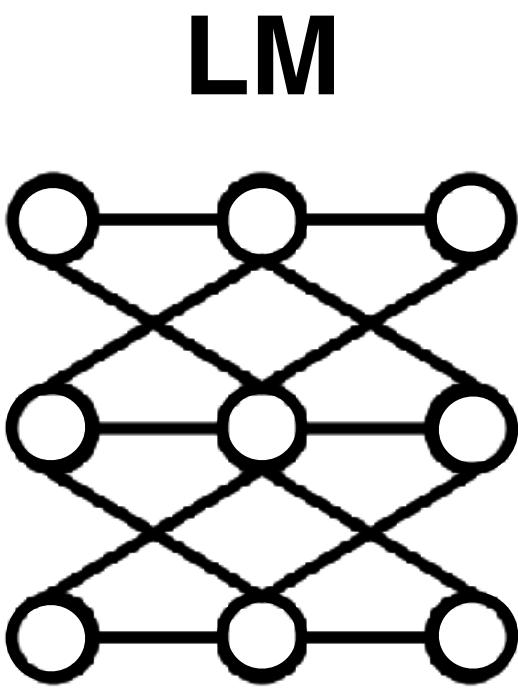
What can I help with?



Monolithic Language Models (LMs)

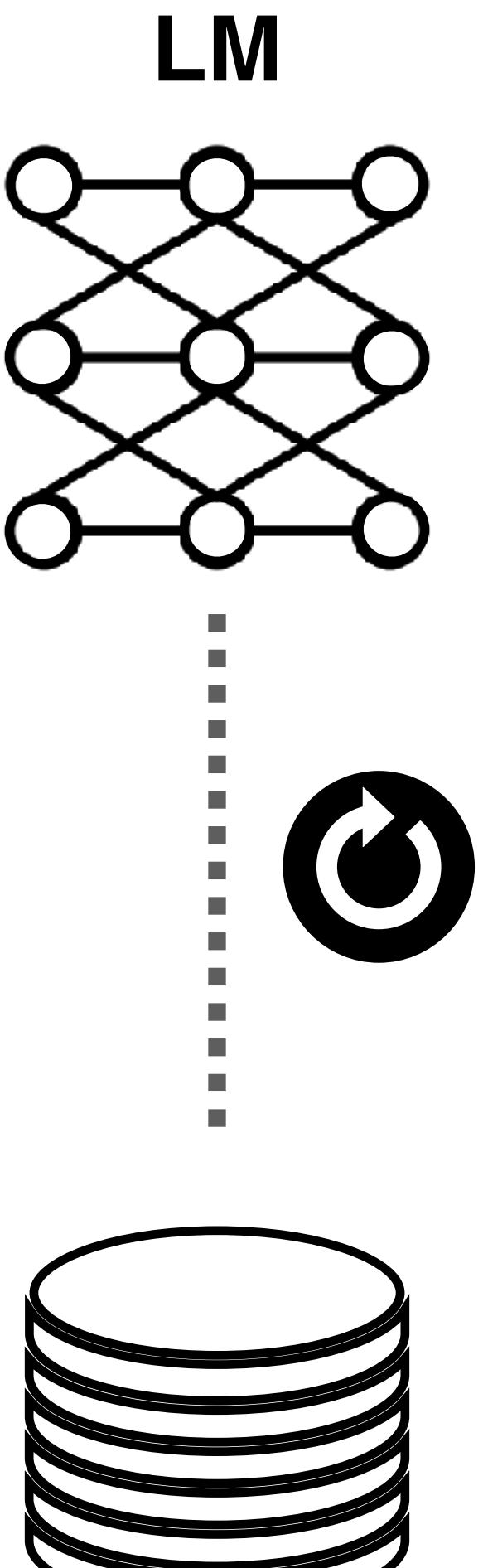


Monolithic Language Models (LMs)

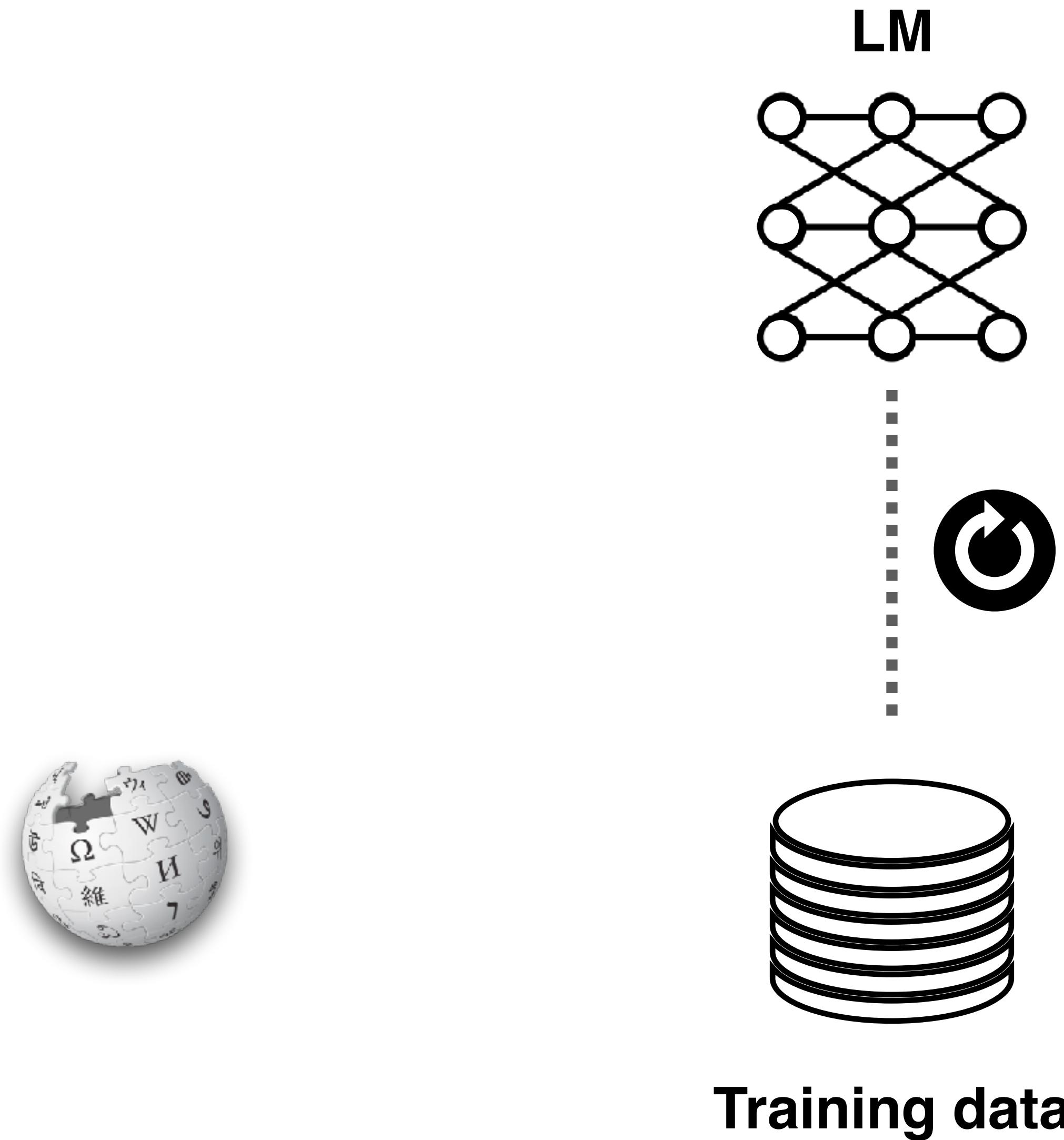


Training data

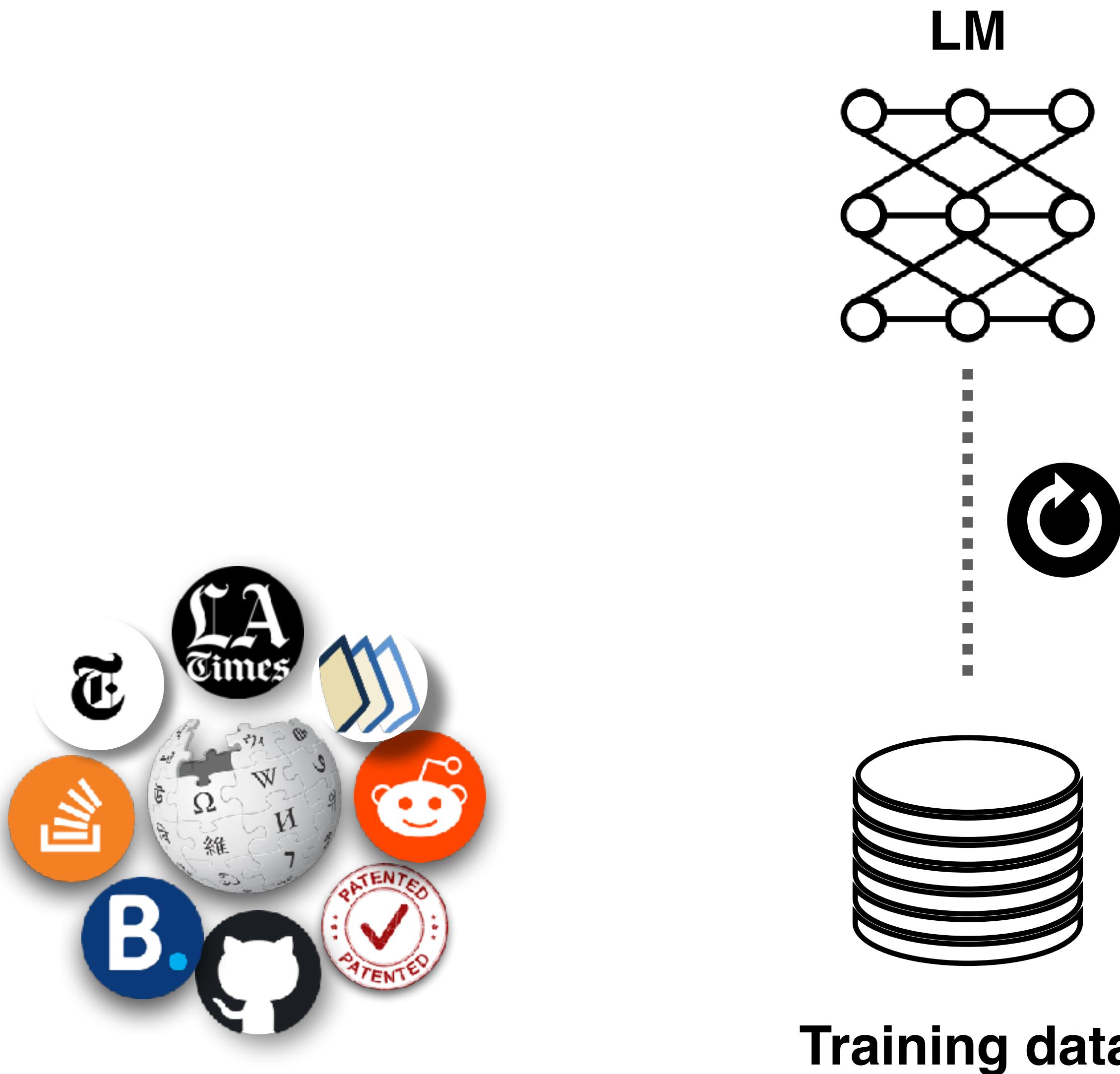
Monolithic Language Models (LMs)



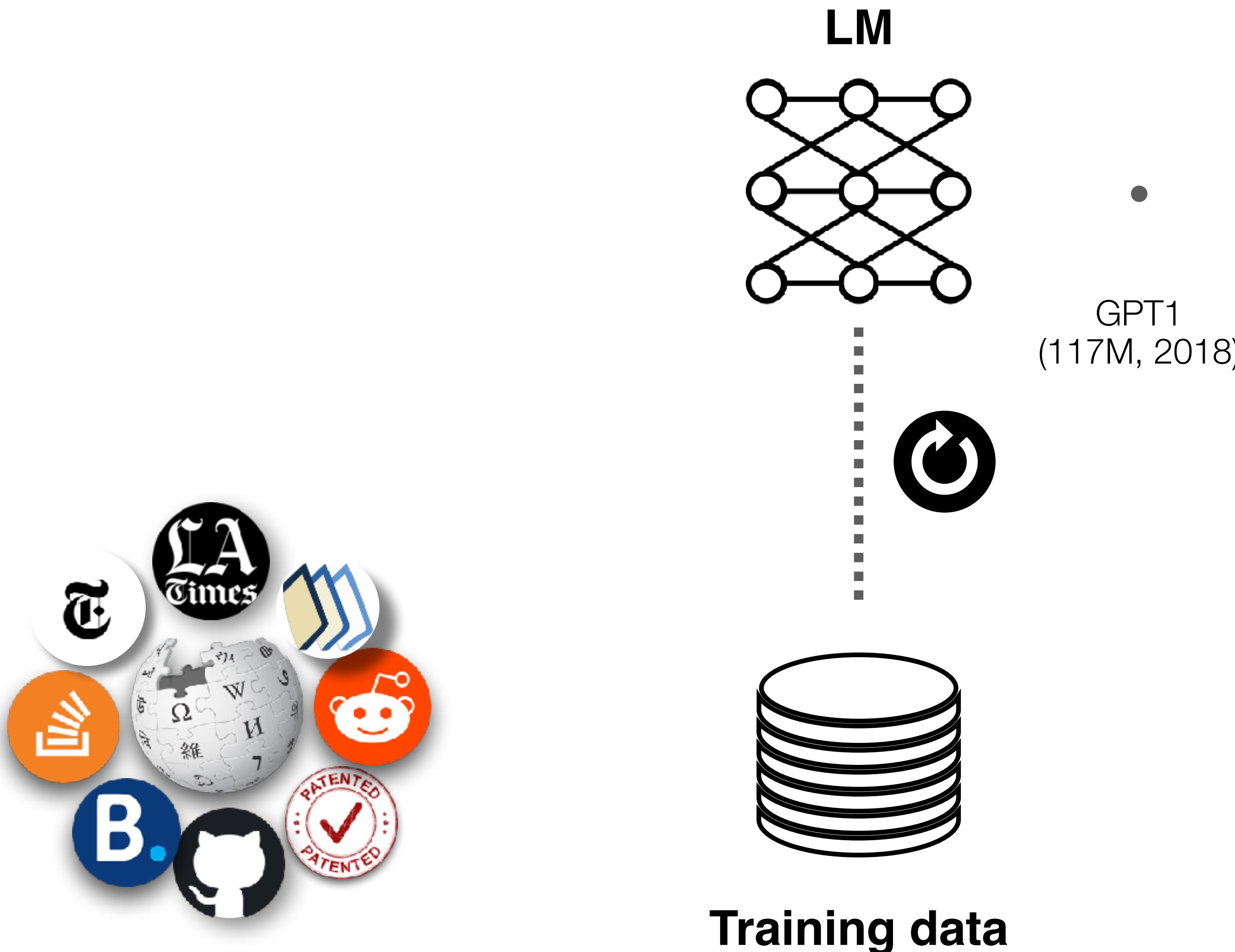
Monolithic Language Models (LMs)



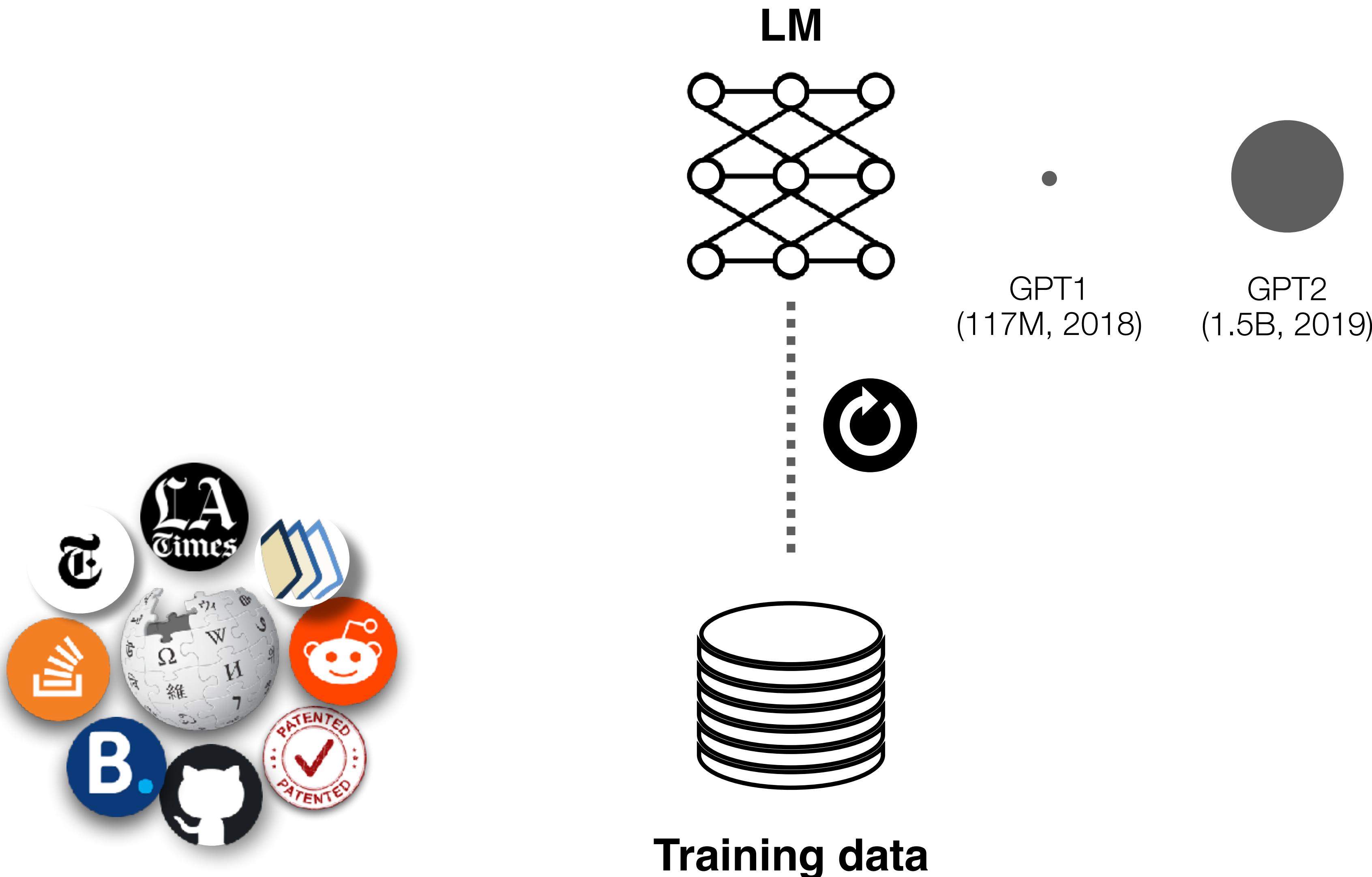
Monolithic Language Models (LMs)



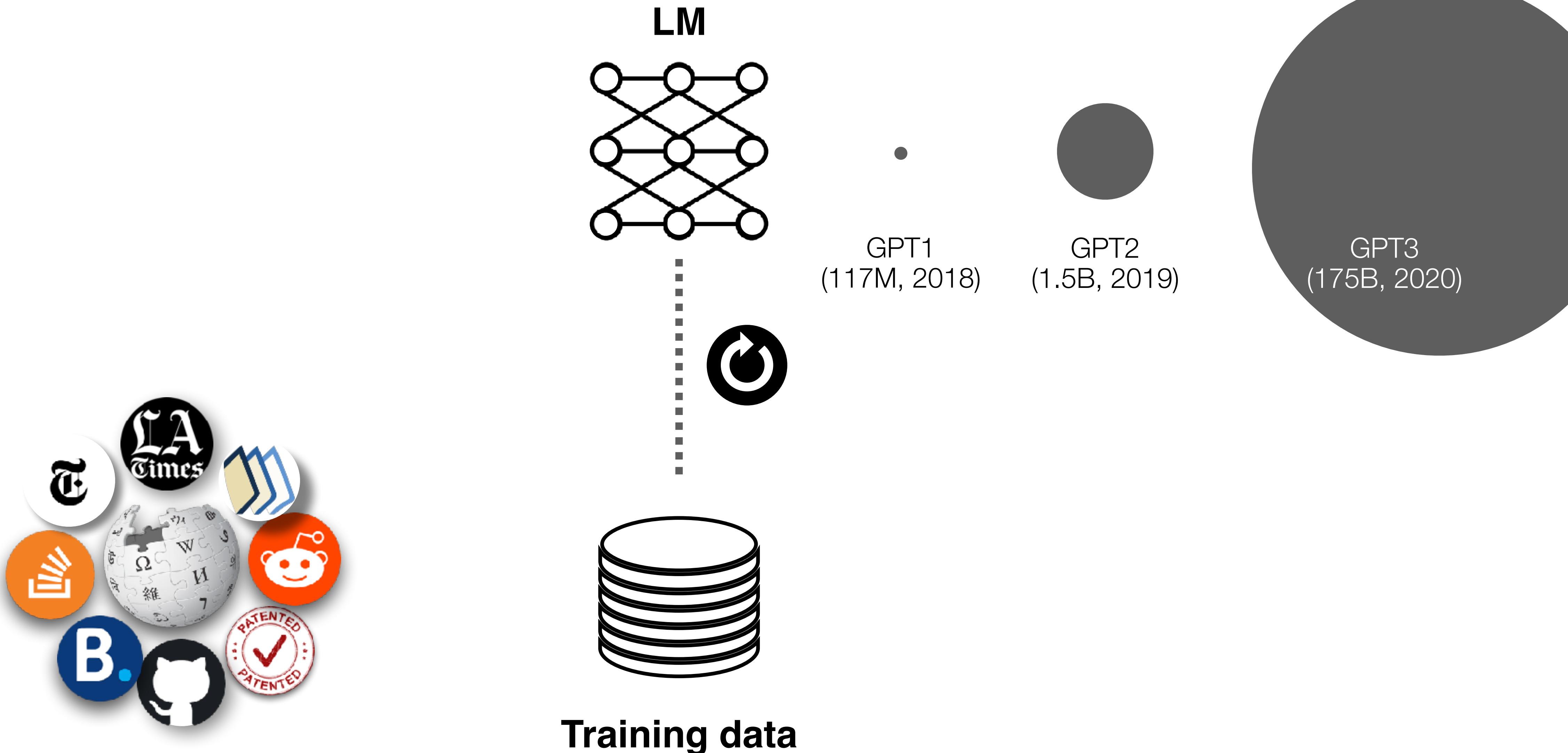
Monolithic Language Models (LMs)



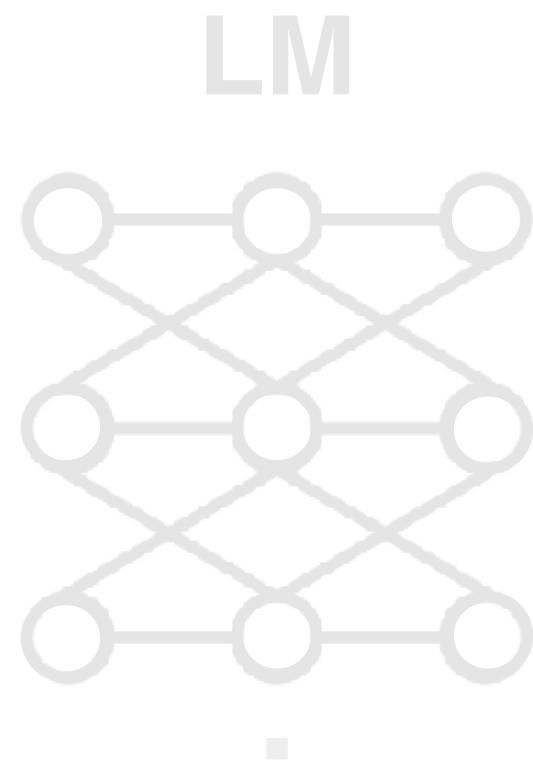
Monolithic Language Models (LMs)



Monolithic Language Models (LMs)



Monolithic Language Models (LMs)



Scaling has driven the rapid progress of **monolithic LMs**



Training data

Are there any studies showing if LMs can help scientists to synthesize scientific literature? Can you suggest some papers?



Are there any studies showing if LMs can help scientists to synthesize scientific literature? Can you suggest some papers?



Are there any studies showing if LMs can help scientists to synthesize scientific literature? Can you suggest some papers?



Yes, there are recent studies exploring how large language models (LLMs) can assist scientists in synthesizing scientific literature.

1. ChatGPT for Scientific Writing: Evaluating its Utility for Synthesizing and Summarizing Research (2023)
2. Automating Systematic Reviews with Pre-trained Language Models
3. Can Large Language Models Enhance Scientific Reasoning?

Are there any studies showing if LMs can help scientists to synthesize scientific literature? Can you suggest some papers?



Yes, there are recent studies exploring how large language models (LLMs) can assist scientists in synthesizing scientific literature.

1. ChatGPT for Scientific Writing: Evaluating its Utility for Synthesizing and Summarizing Research (2023)
2. Automating Systematic Reviews with Pre-trained Language Models
3. Can Large Language Models Enhance Scientific Reasoning?

Are there any studies showing if LMs can help scientists to synthesize scientific literature? Can you suggest some papers?



Yes, there are recent studies exploring how large language models (LLMs) can assist scientists in synthesizing scientific literature.

1. ChatGPT for Scientific Writing: Evaluating its Utility for Synthesizing and Summarizing Research (2023)
2. Automating Systematic Reviews with Pre-trained Language Models
3. Can Large Language Models Enhance Scientific Reasoning?



None of these papers exist!

Are there any studies showing if LMs can help scientists to synthesize scientific literature? Can you suggest some papers?



Yes, there are recent studies exploring how large language models (LLMs) can assist scientists in synthesizing scientific literature.

1. ChatGPT for Scientific Writing: Evaluating its Utility for Synthesizing and Summarizing Research (2023)
2. Automating Systematic Reviews with Pre-trained Language Models
3. Can Large Language Models Enhance Scientific Reasoning?



Factual inaccuracies in LM output



None of these papers exist!

Are there any studies showing if LMs can help scientists to synthesize scientific literature? Can you suggest some papers?



Yes, there are recent studies exploring how large language models (LLMs) can assist scientists in synthesizing scientific literature.

1. ChatGPT for Scientific Writing: Evaluating its Utility for Synthesizing and Summarizing Research (2023)
2. Automating Systematic Reviews with Pre-trained Language Models
3. Can Large Language Models Enhance Scientific Reasoning?



Yes, there are recent studies exploring how large language models (LLMs) can assist scientists in synthesizing scientific literature.

1. ChatGPT for Scientific Writing: Evaluating its Utility for Synthesizing and Summarizing Research (2023)
2. Automating Systematic Reviews with Pre-trained Language Models
3. Can Large Language Models Enhance Scientific Reasoning?



Yes, there are recent studies exploring how large language models (LLMs) can assist scientists in synthesizing scientific literature.

1. ChatGPT for Scientific Writing: Evaluating its Utility for Synthesizing and Summarizing Research (2023)
2. Automating Systematic Reviews with Pre-trained Language Models
3. Can Large Language Models Enhance Scientific Reasoning?

Has anyone tried GPT4o in this area?





Yes, there are recent studies exploring how large language models (LLMs) can assist scientists in synthesizing scientific literature.

1. ChatGPT for Scientific Writing: Evaluating its Utility for Synthesizing and Summarizing Research (2023)
2. Automating Systematic Reviews with Pre-trained Language Models
3. Can Large Language Models Enhance Scientific Reasoning?

Has anyone tried GPT4o in this area?



*Released in May 2024 and default for ChatGPT



Yes, there are recent studies exploring how large language models (LLMs) can assist scientists in synthesizing scientific literature.

1. ChatGPT for Scientific Writing: Evaluating its Utility for Synthesizing and Summarizing Research (2023)
2. Automating Systematic Reviews with Pre-trained Language Models
3. Can Large Language Models Enhance Scientific Reasoning?

Has anyone tried GPT4o in this area?



*Released in May 2024 and default for ChatGPT





Yes, there are recent studies exploring how large language models (LLMs) can assist scientists in synthesizing scientific literature.

1. ChatGPT for Scientific Writing: Evaluating its Utility for Synthesizing and Summarizing Research (2023)
2. Automating Systematic Reviews with Pre-trained Language Models
3. Can Large Language Models Enhance Scientific Reasoning?



Has anyone tried GPT4o in this area?



*Released in May 2024 and default for ChatGPT

Based on general knowledge, there is no widely recognized AI model or term specifically called "GPT4o." If "GPT4o" refers to a recent variant of GPT4 or a new model released in 2024, I cannot comment.



Yes, there are recent studies exploring how large language models (LLMs) can assist scientists in synthesizing scientific literature.

1. ChatGPT for Scientific Writing: Evaluating its Utility for Synthesizing and Summarizing Research (2023)
2. Automating Systematic Reviews with Pre-trained Language Models
3. Can Large Language Models Enhance Scientific Reasoning?

Has anyone tried GPT4o in this area?



*Released in May 2024 and default for ChatGPT



Based on general knowledge, there is no widely recognized AI model or term specifically called "GPT4o." If "GPT4o" refers to a recent variant of GPT4 or a new model released in 2024, I cannot comment.



Outdated knowledge



Yes, there are recent studies exploring how large language models (LLMs) can assist scientists in synthesizing scientific literature.

1. ChatGPT for Scientific Writing: Evaluating its Utility for Synthesizing and Summarizing Research (2023)
2. Automating Systematic Reviews with Pre-trained Language Models
3. Can Large Language Models Enhance Scientific Reasoning?



Factual inaccuracies in LLM output



Based on general knowledge, there is no widely recognized AI model or term specifically called "GPT4o." If "GPT4o" refers to a recent variant of GPT-4 or a new model released in 2024, I cannot comment.



Outdated knowledge



Yes, there are recent studies exploring how large language models (LLMs) can assist scientists in synthesizing scientific literature.

1. ChatGPT for Scientific Writing: Evaluating its Utility for Synthesizing and Summarizing Research (2023)
2. Automating Systematic Reviews with Pre-trained Language Models
3. Can Large Language Models Estimate Scientific Research? (2023)

“GPT4 training costs **more than \$100 million”**

— Sam Altman, OpenAI CEO

<https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/> (4/17/2023)



Based on general knowledge, there is no widely recognized AI model or term specifically called “GPT4o.” If “GPT4o” refers to a recent variant of GPT-4 or a new model released in 2024, I cannot comment.



Outdated knowledge



Yes, there are recent studies exploring how large language models (LLMs) can assist scientists in synthesizing scientific literature.

1. ChatGPT for Scientific Writing: Evaluating its Utility for Synthesizing and Summarizing Research (2023)
2. Automating Systematic Reviews with Pre-trained Language Models
3. Can Large Language Models Estimate Scientific Research? (2023)

“GPT4 training costs **more than \$100 million”**

— Sam Altman, OpenAI CEO

<https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/> (4/17/2023)

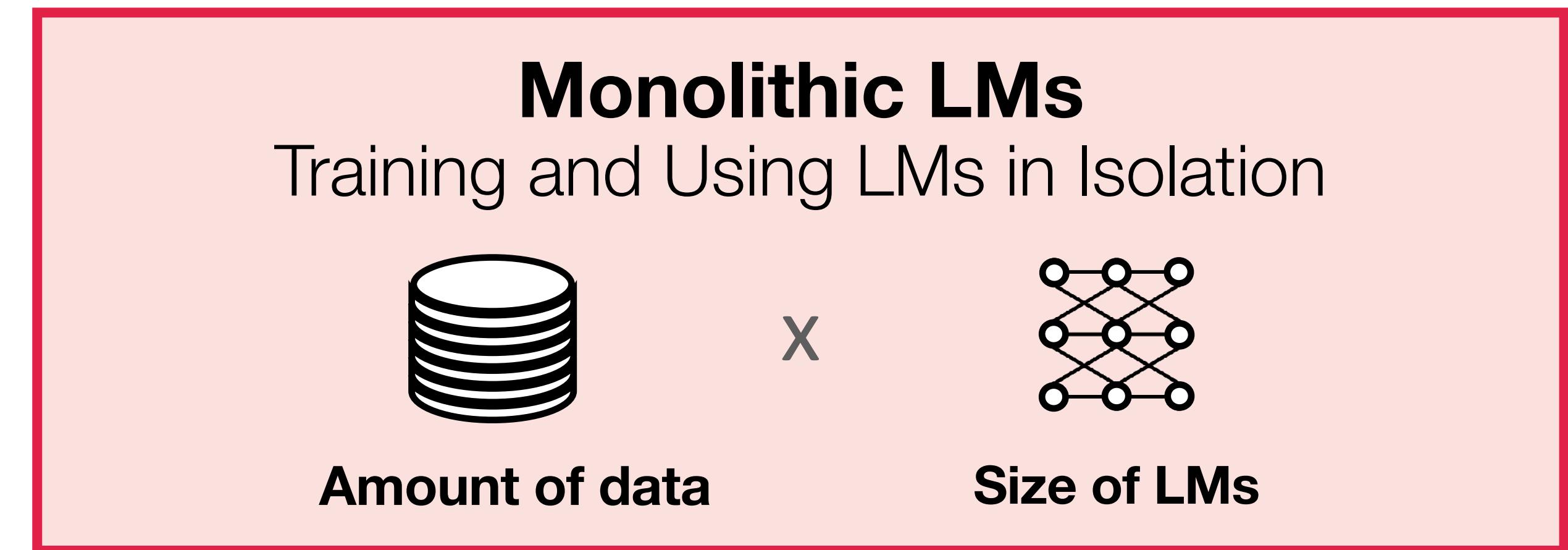
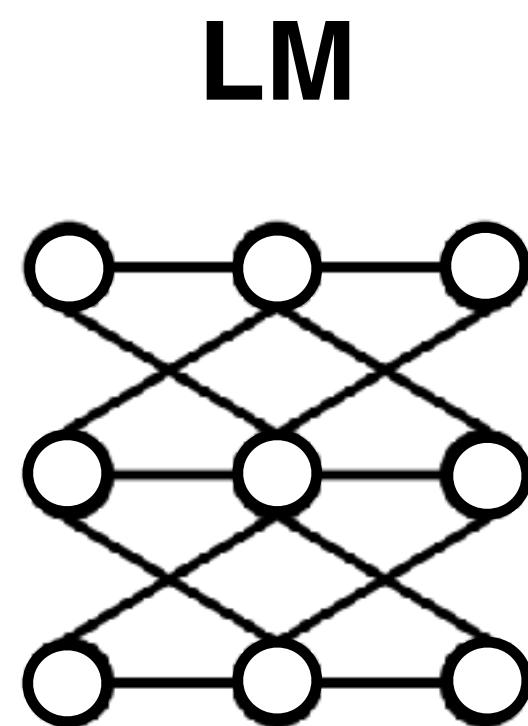


Based on general knowledge, there is no widely recognized AI model or term specifically referring to GPT-4. The “Rapidly increasing cost of training” refers to a recent variant of GPT-4 or a new model released in 2024, I cannot comment.



Outdated knowledge

Today: Scaling Monolithic LMs

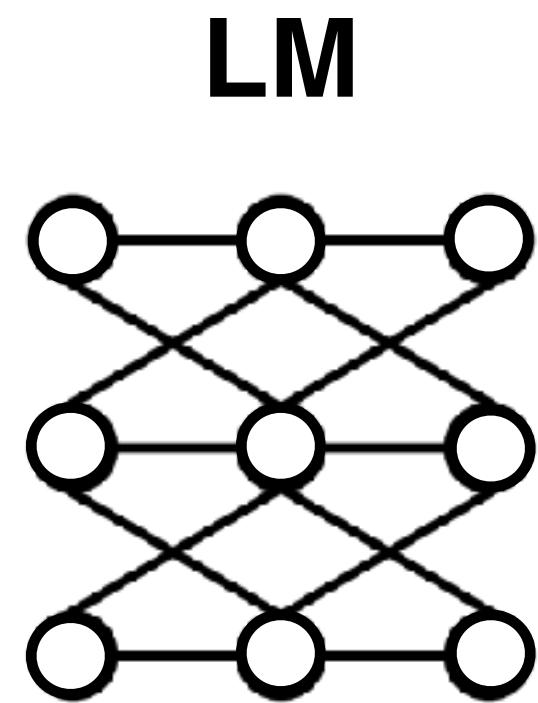


- :(Factual inaccuracies in LM output
- :(Outdated knowledge
- :(Rapidly increasing cost of training

From Scaling Monolithic LMs

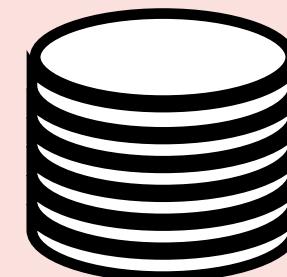
From Scaling **Monolithic** LMs to **Augmented** LMs

From Scaling **Monolithic LMs** to **Augmented LMs**



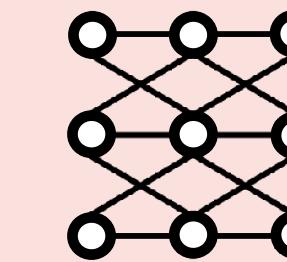
Monolithic LMs

Training and using LMs in isolation



Amount of data

X

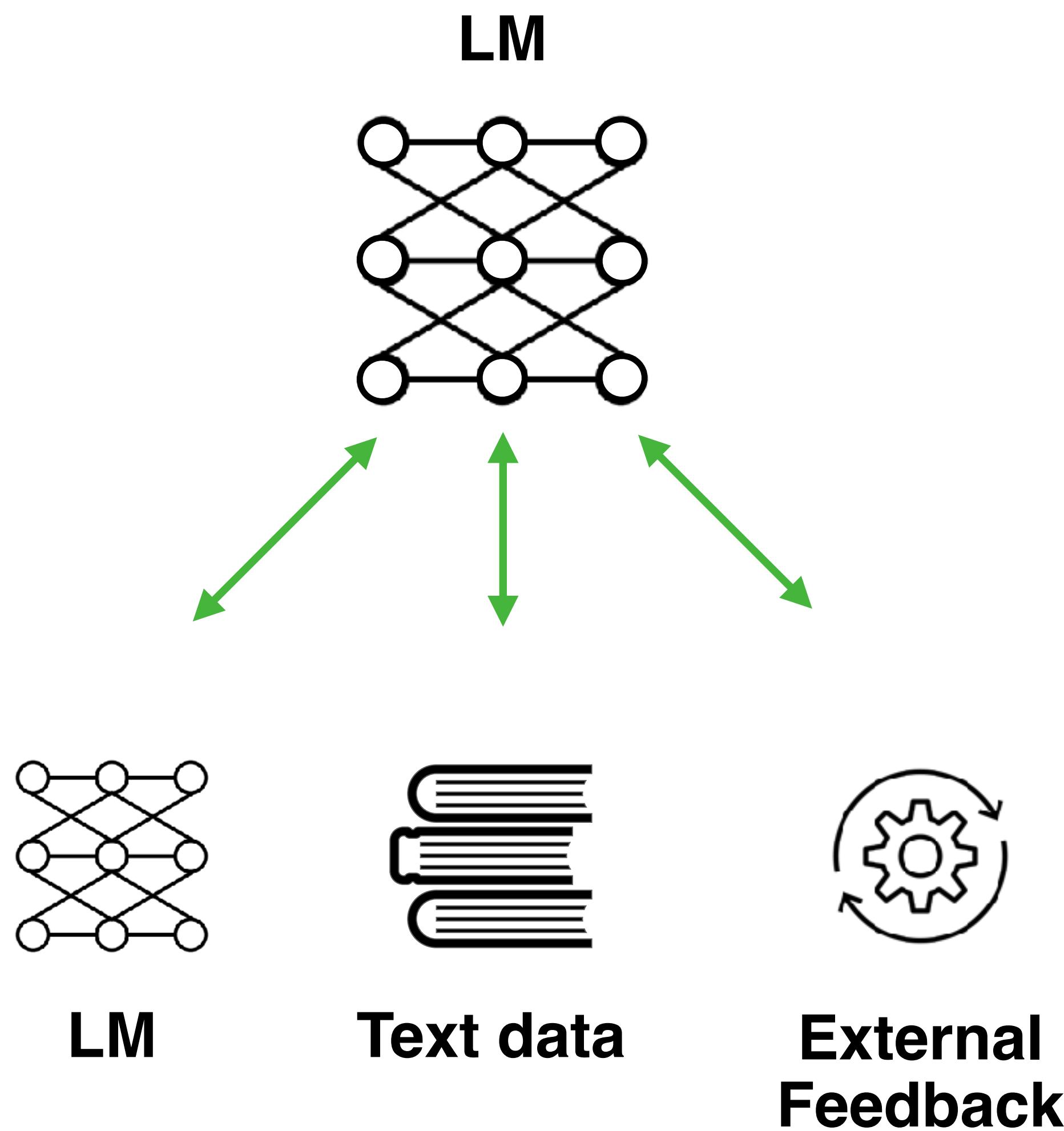


Size of LMs

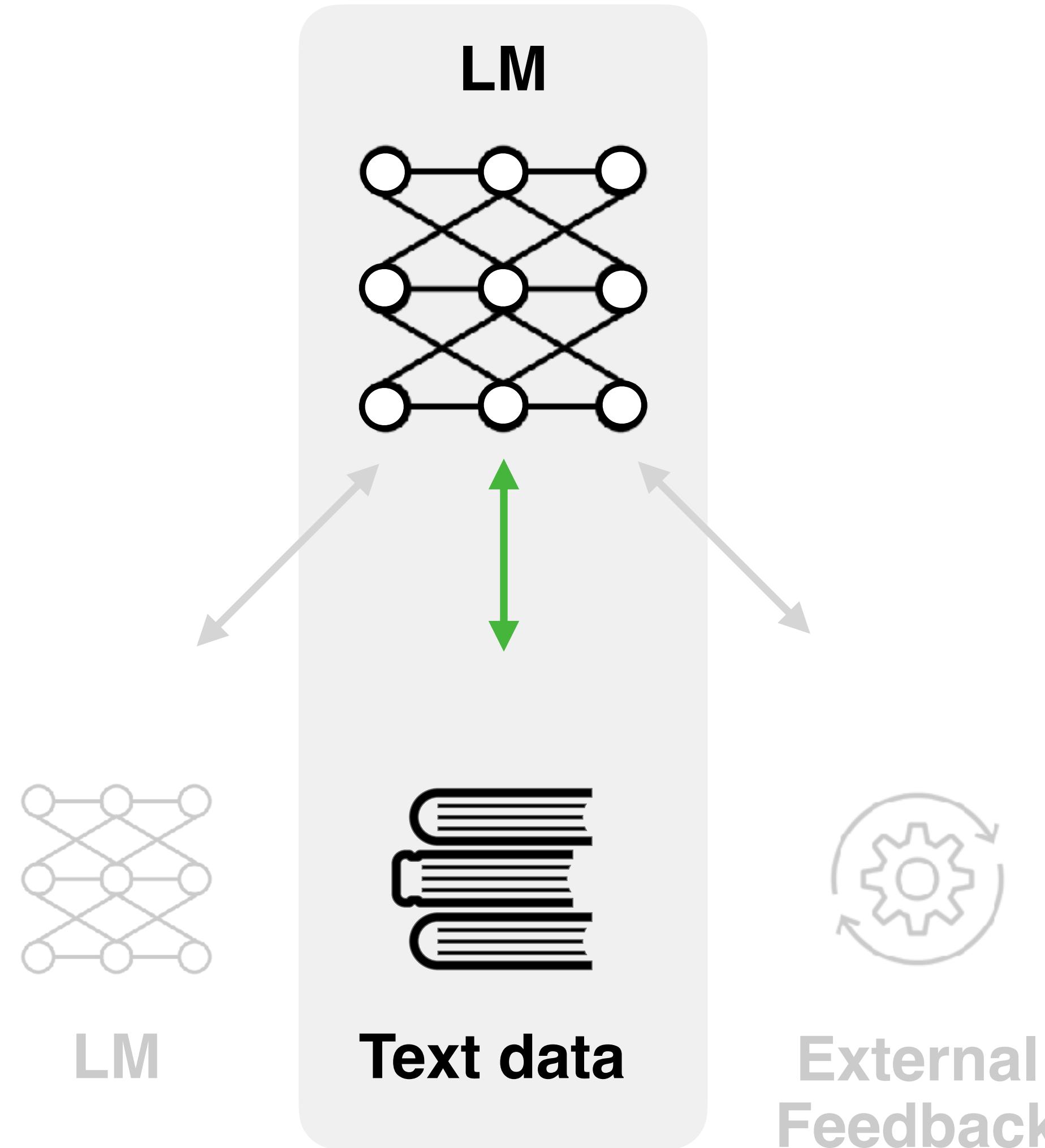
Augmented LMs

Training and using LMs with other modules
to go beyond their parametric capabilities

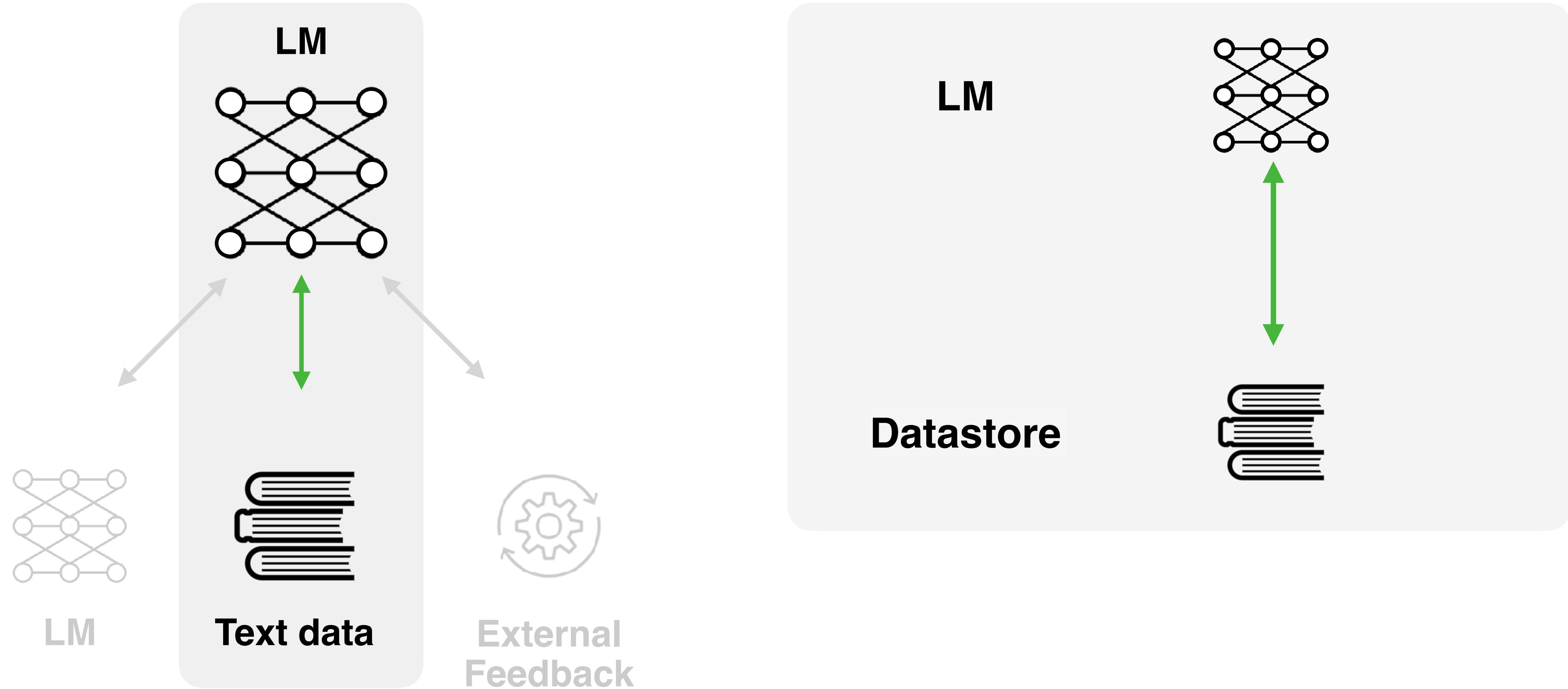
From Scaling **Monolithic LMs** to **Augmented LMs**



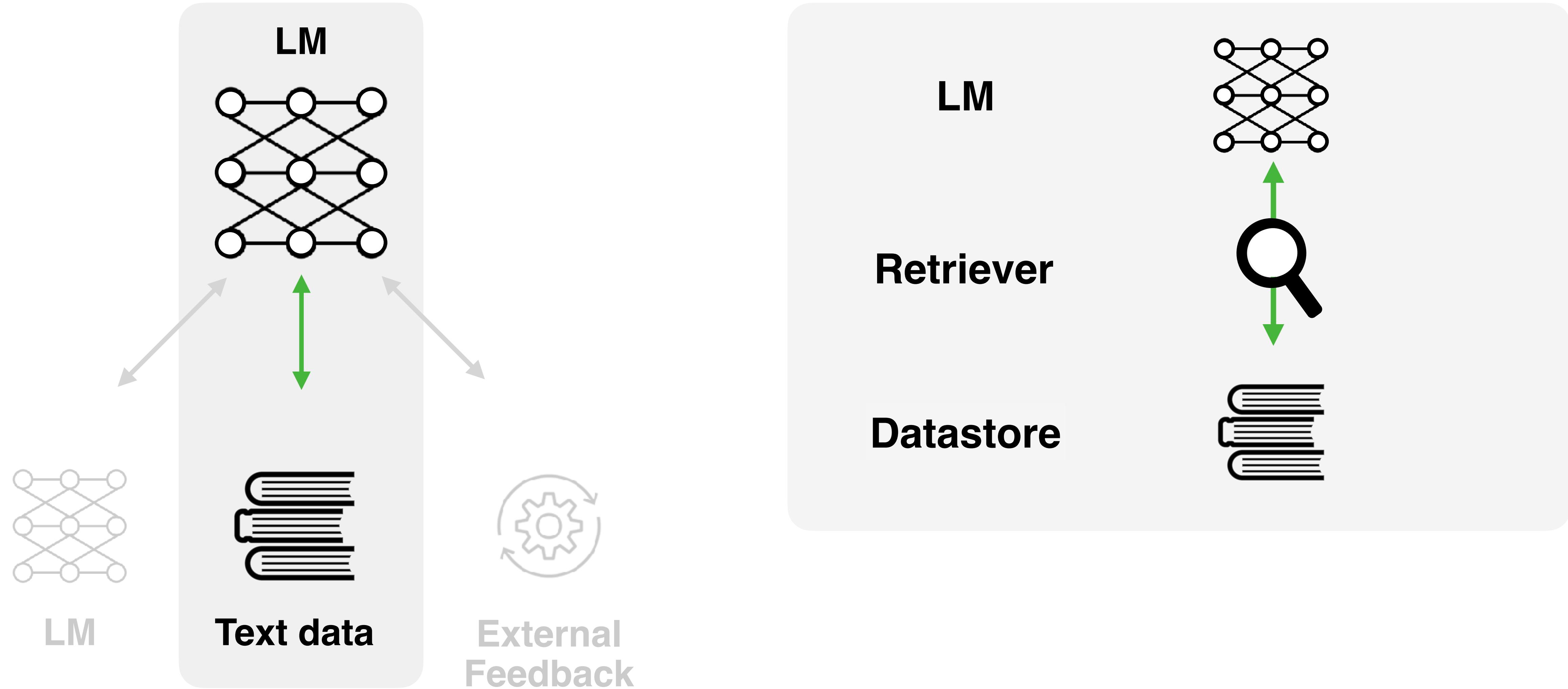
Retrieval-Augmented LMs



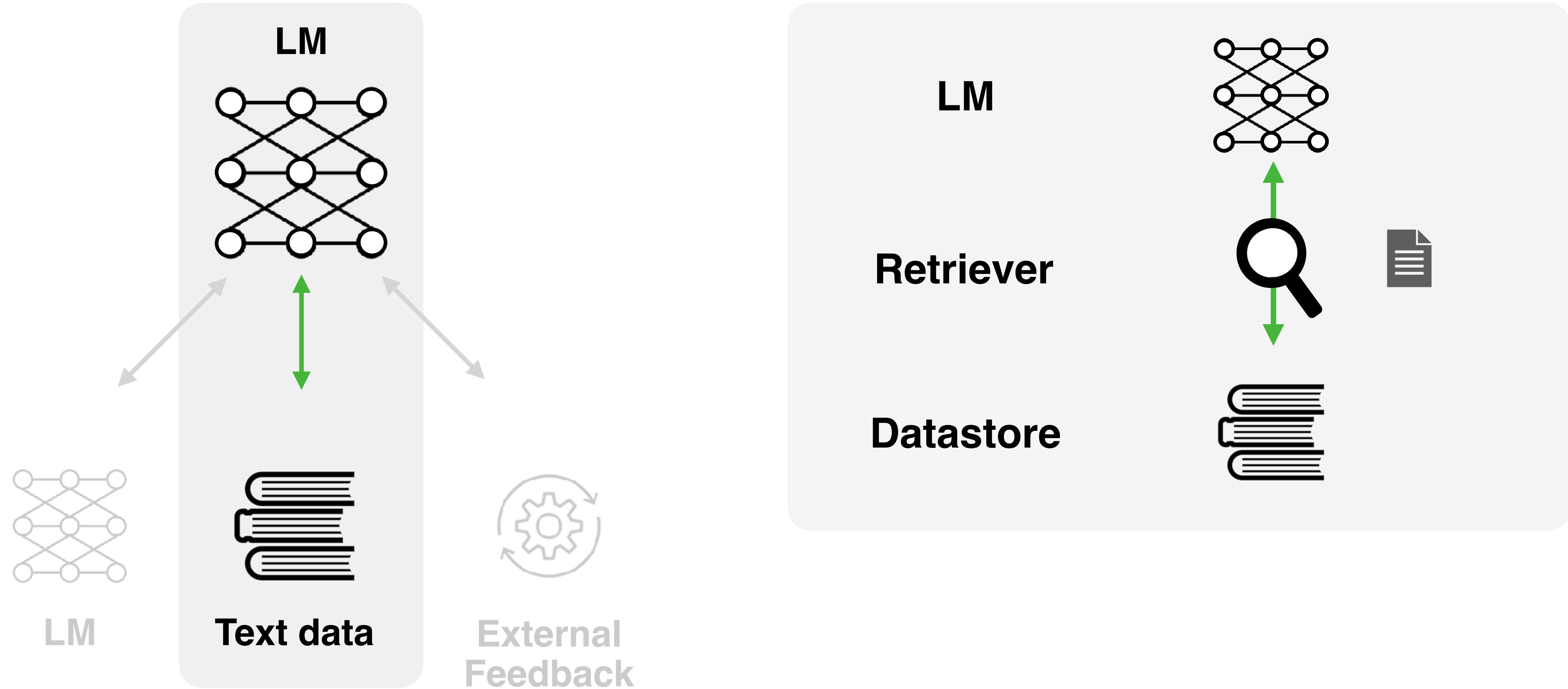
Retrieval-Augmented LMs



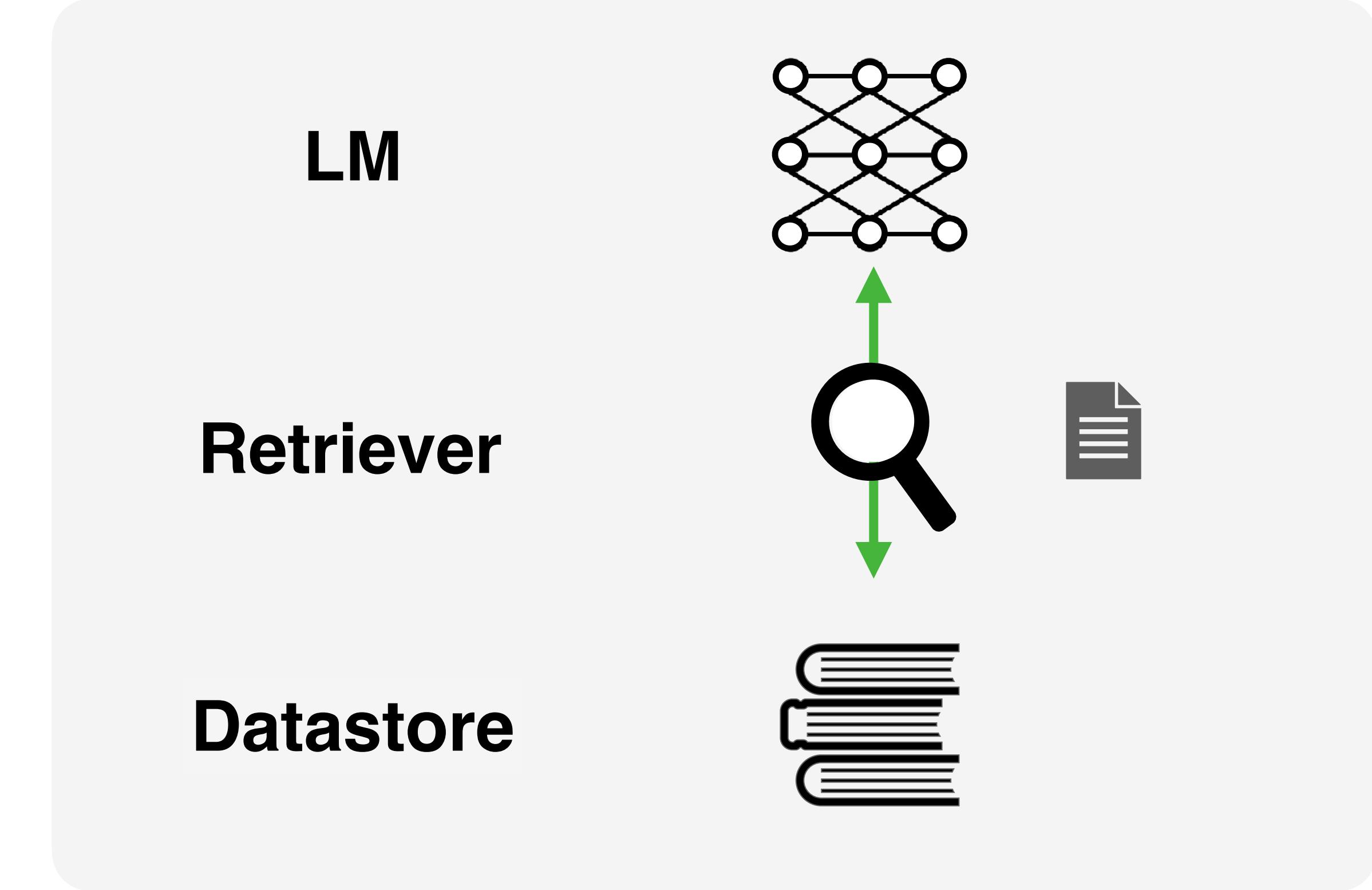
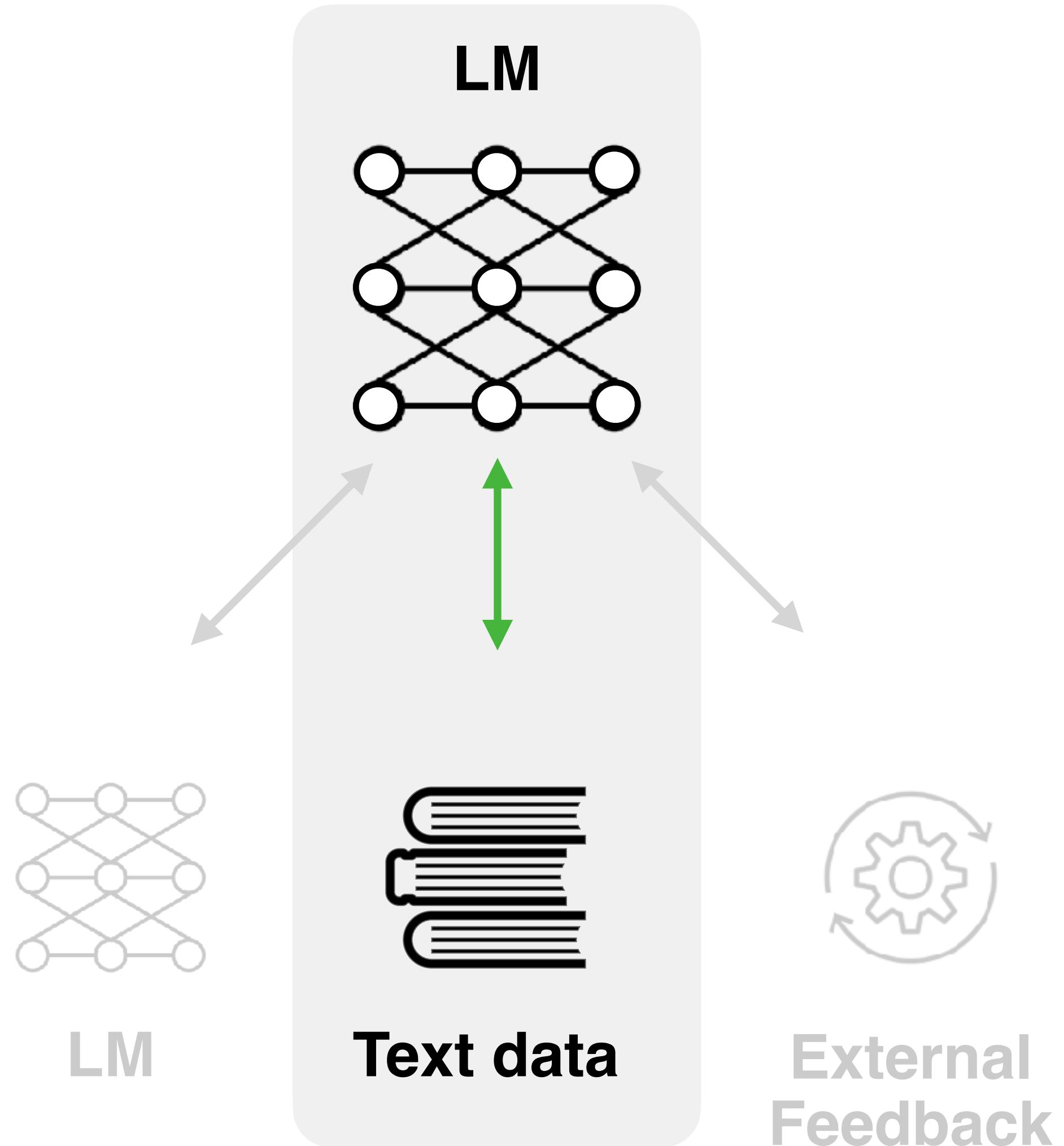
Retrieval-Augmented LMs



Retrieval-Augmented LMs



Retrieval-Augmented LMs



OpenAI



Google DeepMind



Microsoft

Frontiers of Retrieval-Augmented LMs

Frontiers of Retrieval-Augmented LMs



From Monolithic to Retrieval-Augmented LMs

Why?

Frontiers of Retrieval-Augmented LMs



From Monolithic to Retrieval-Augmented LMs

Why?



New Foundations of Retrieval-Augmented LMs

How?

Frontiers of Retrieval-Augmented LMs



From Monolithic to Retrieval-Augmented LMs

Why?



New Foundations of Retrieval-Augmented LMs

How?



Real-World Impacts of Retrieval-Augmented LMs

What?

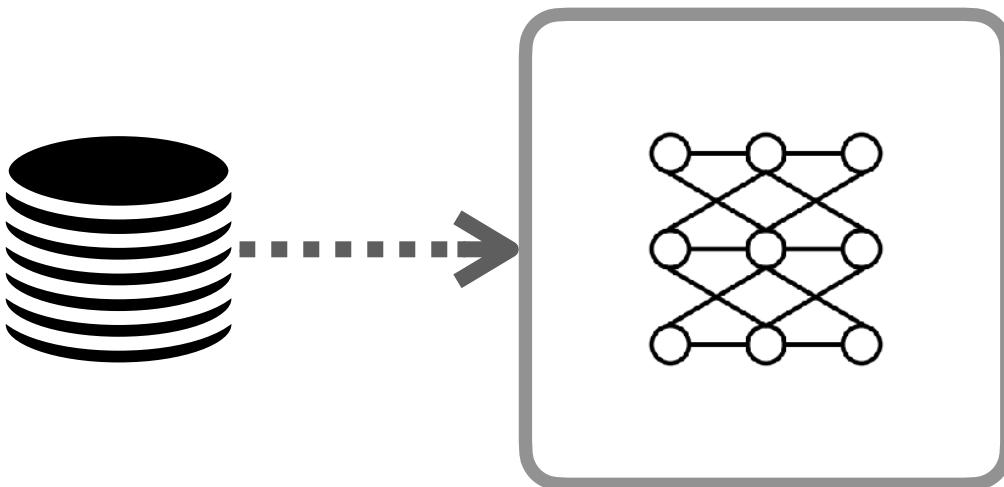


From Monolithic to Retrieval-Augmented LMs



From Monolithic to Retrieval-Augmented LMs

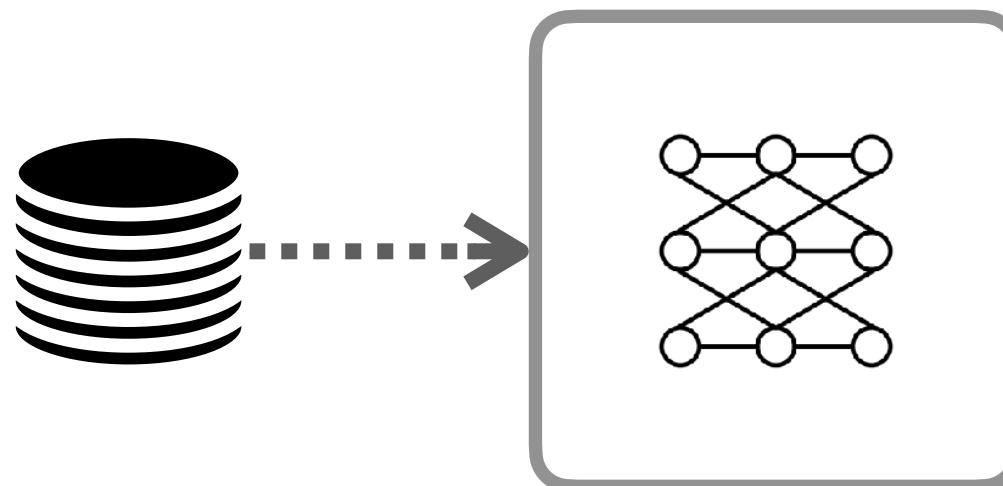
Monolithic LMs





From Monolithic to Retrieval-Augmented LMs

Monolithic LMs

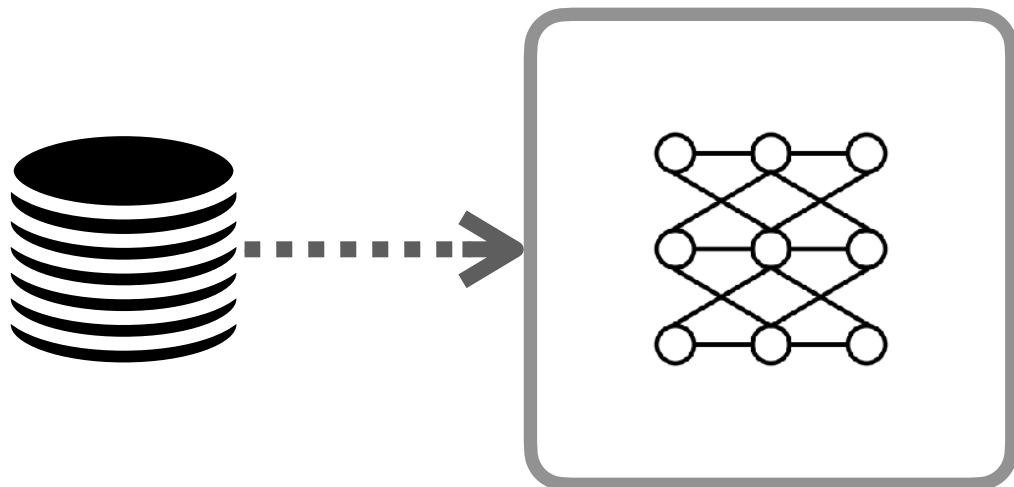


Asai*, Mallen* (contributed equally) et al. ACL 2023 (**Most Influential Paper #5, Most-Viewed Video Awards**) | Shao, He, **Asai** et al. NeurIPS 2024
Kasai, ..., **Asai** et al., NeurIPS D&B 2023 | Mishra, **Asai** et al., COLM 2024 | **Asai** et al., ACL Tutorial 2023 | Kamoda, **Asai** et al., COLING 2025



From Monolithic to Retrieval-Augmented LMs

Monolithic LMs



Challenges remain after scaling



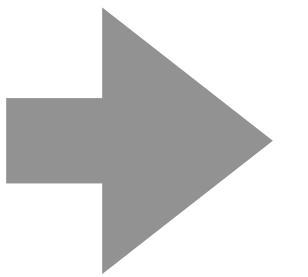
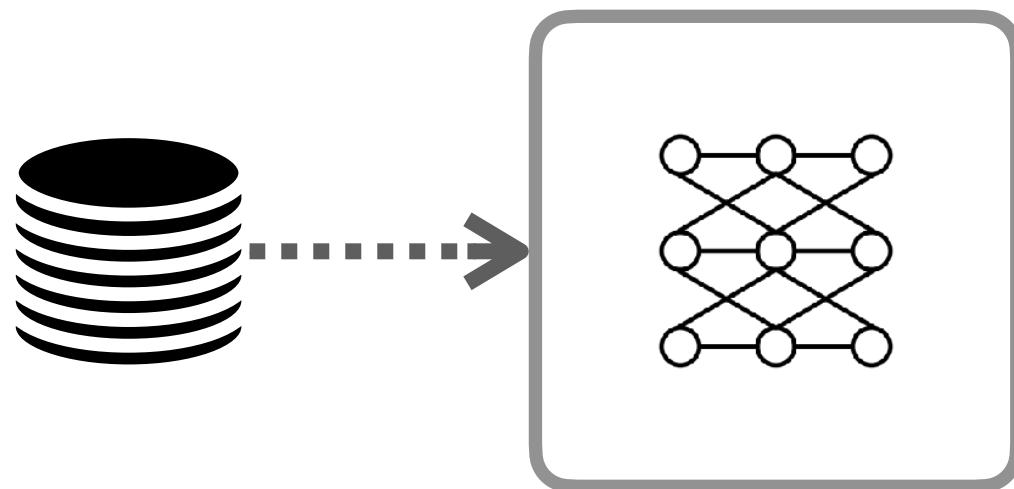
Massive training costs

Asai*, Mallen* (contributed equally) et al. ACL 2023 (**Most Influential Paper #5, Most-Viewed Video Awards**) | Shao, He, **Asai** et al. NeurIPS 2024
Kasai, ..., **Asai** et al., NeurIPS D&B 2023 | Mishra, **Asai** et al., COLM 2024 | **Asai** et al., ACL Tutorial 2023 | Kamoda, **Asai** et al., COLING 2025



From Monolithic to Retrieval-Augmented LMs

Monolithic LMs



Challenges remain after scaling



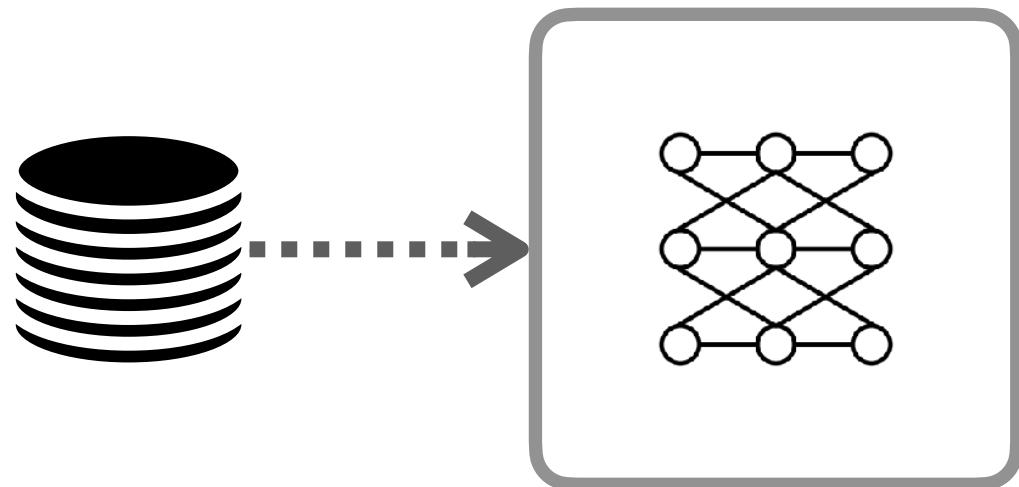
Massive training costs

Asai*, Mallen* (contributed equally) et al. ACL 2023 (**Most Influential Paper #5, Most-Viewed Video Awards**) | Shao, He, **Asai** et al. NeurIPS 2024
Kasai, ..., **Asai** et al., NeurIPS D&B 2023 | Mishra, **Asai** et al., COLM 2024 | **Asai** et al., ACL Tutorial 2023 | Kamoda, **Asai** et al., COLING 2025



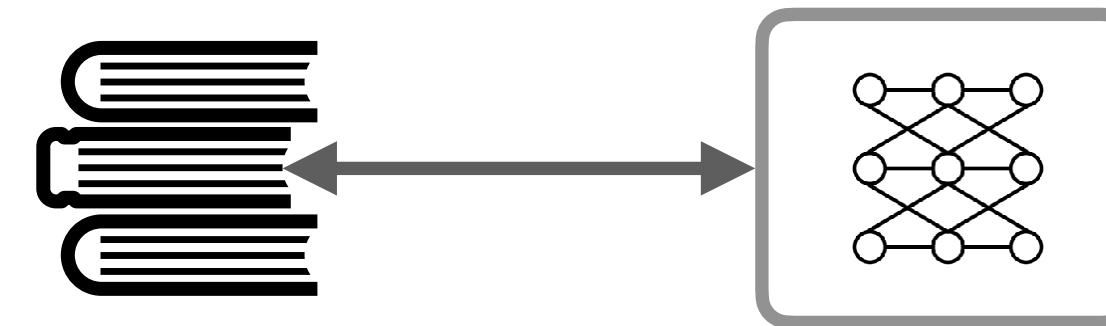
From Monolithic to Retrieval-Augmented LMs

Monolithic LMs



- :(Challenges remain after scaling
- :(Massive training costs

Retrieval-Augmented LMs

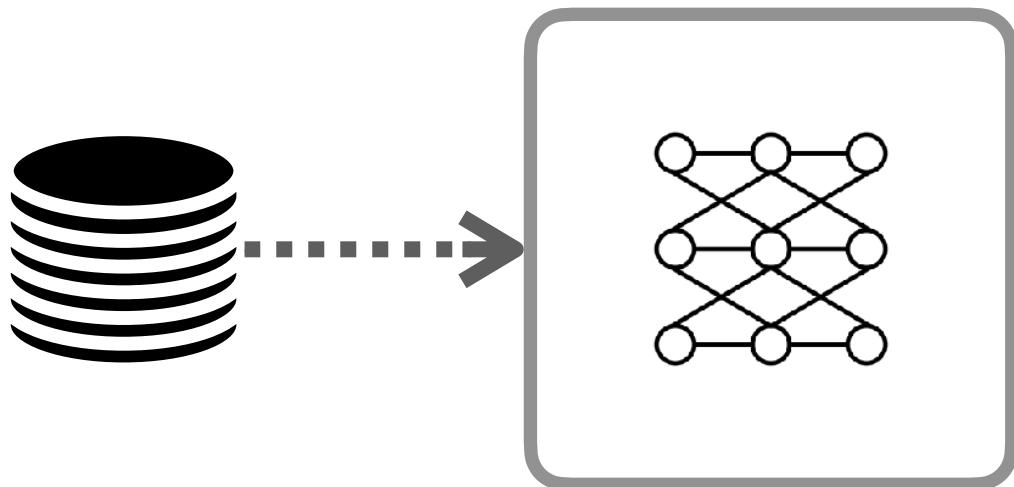


Asai*, Mallen* (contributed equally) et al. ACL 2023 (**Most Influential Paper #5, Most-Viewed Video Awards**) | Shao, He, **Asai** et al. NeurIPS 2024
Kasai, ..., **Asai** et al., NeurIPS D&B 2023 | Mishra, **Asai** et al., COLM 2024 | **Asai** et al., ACL Tutorial 2023 | Kamoda, **Asai** et al., COLING 2025



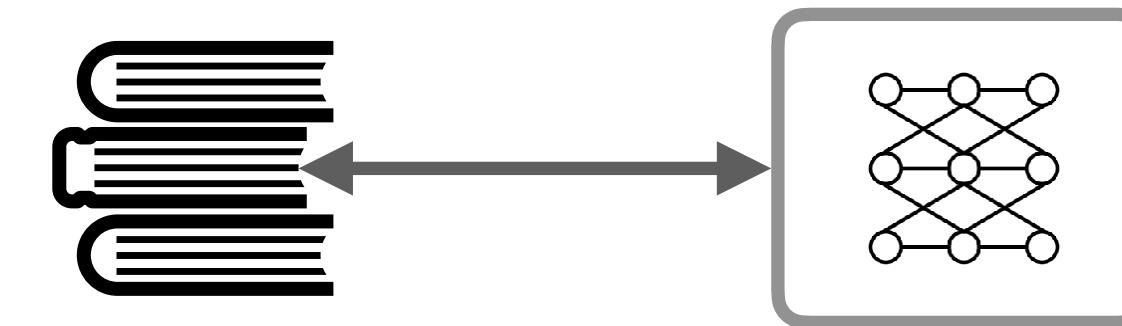
From Monolithic to Retrieval-Augmented LMs

Monolithic LMs



- :(Challenges remain after scaling
- :(Massive training costs

Retrieval-Augmented LMs



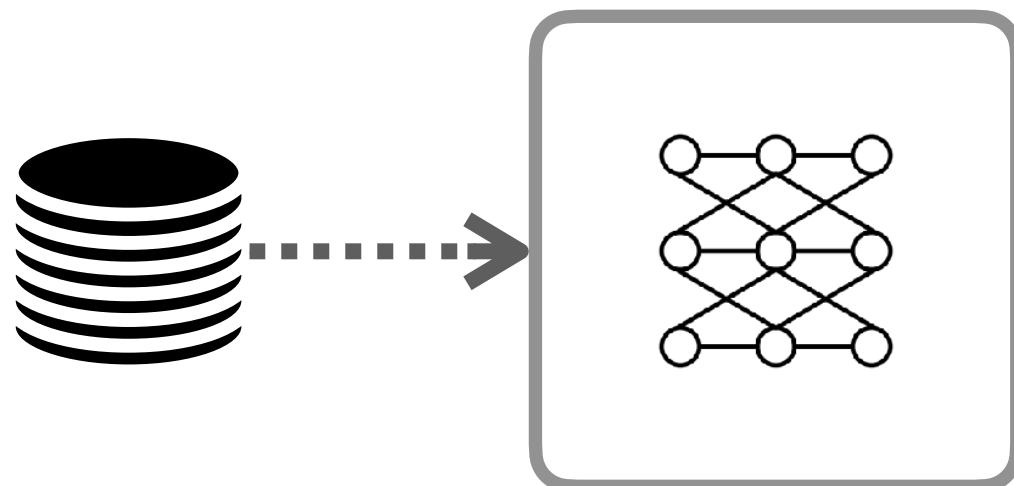
- : Address challenges effectively

Asai*, Mallen* (contributed equally) et al. ACL 2023 (**Most Influential Paper #5, Most-Viewed Video Awards**) | Shao, He, **Asai** et al. NeurIPS 2024
Kasai, ..., **Asai** et al., NeurIPS D&B 2023 | Mishra, **Asai** et al., COLM 2024 | **Asai** et al., ACL Tutorial 2023 | Kamoda, **Asai** et al., COLING 2025



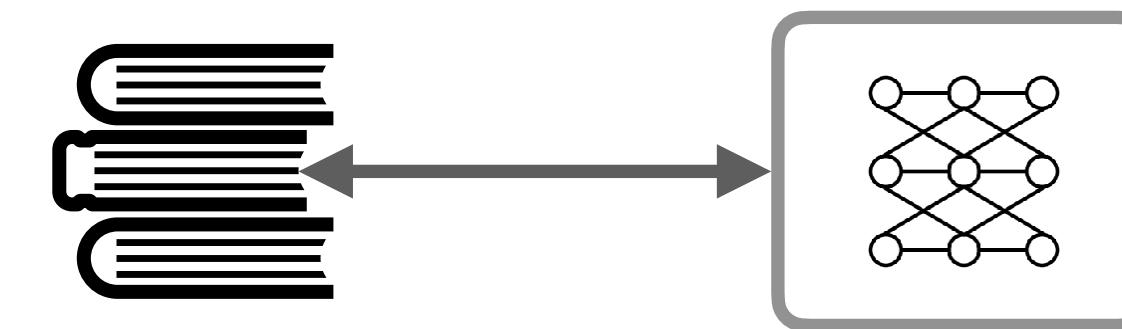
From Monolithic to Retrieval-Augmented LMs

Monolithic LMs



- :(Challenges remain after scaling
- :(Massive training costs

Retrieval-Augmented LMs



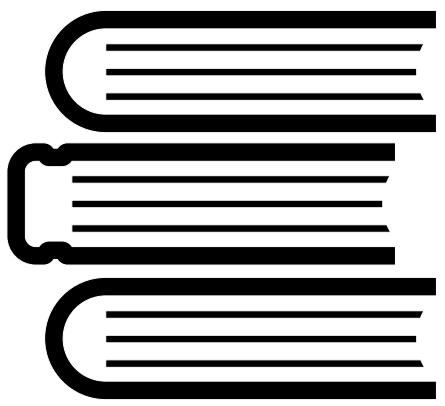
- : Address challenges effectively
- : Reducing training cost



New Foundations of Retrieval-Augmented LMs



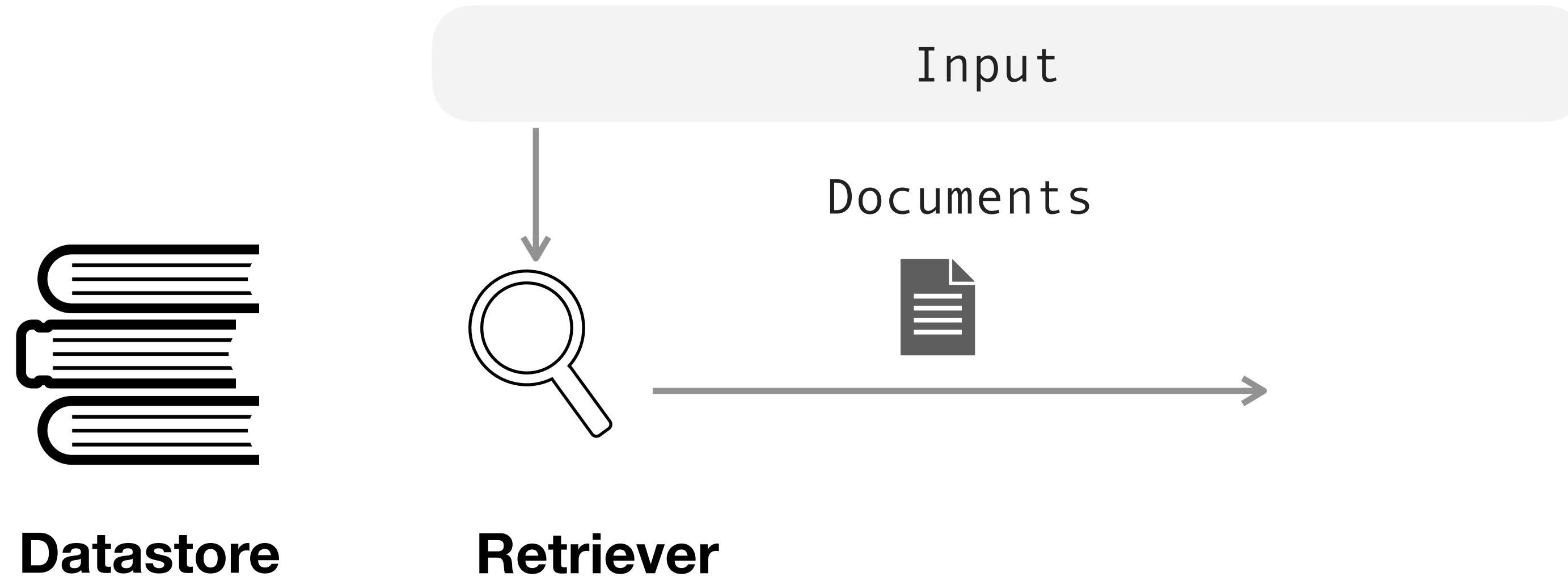
New Foundations of Retrieval-Augmented LMs



Datastore

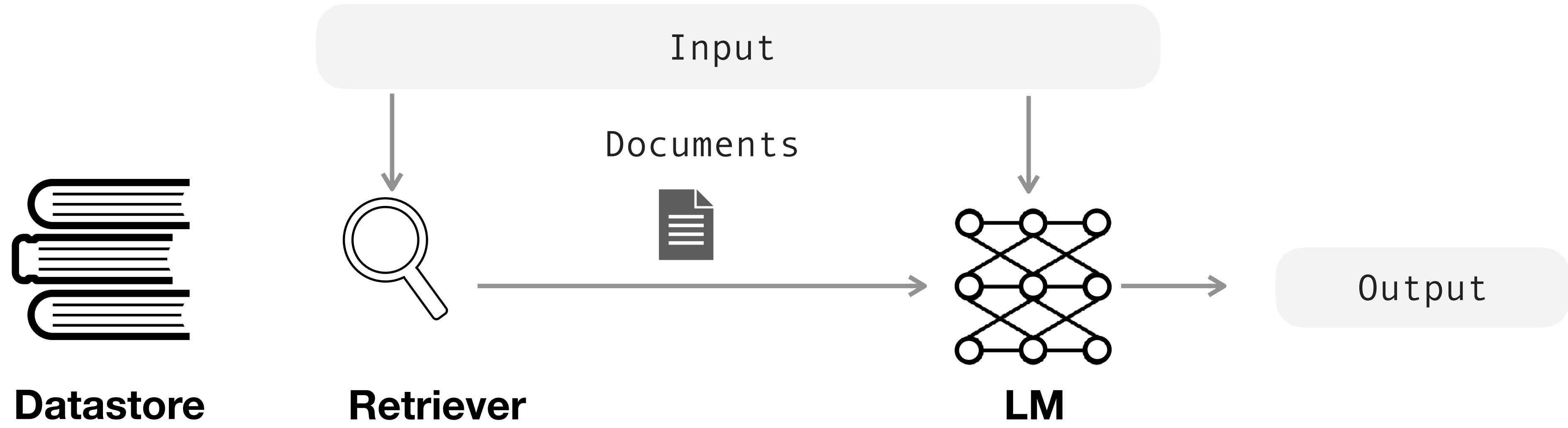


New Foundations of Retrieval-Augmented LMs



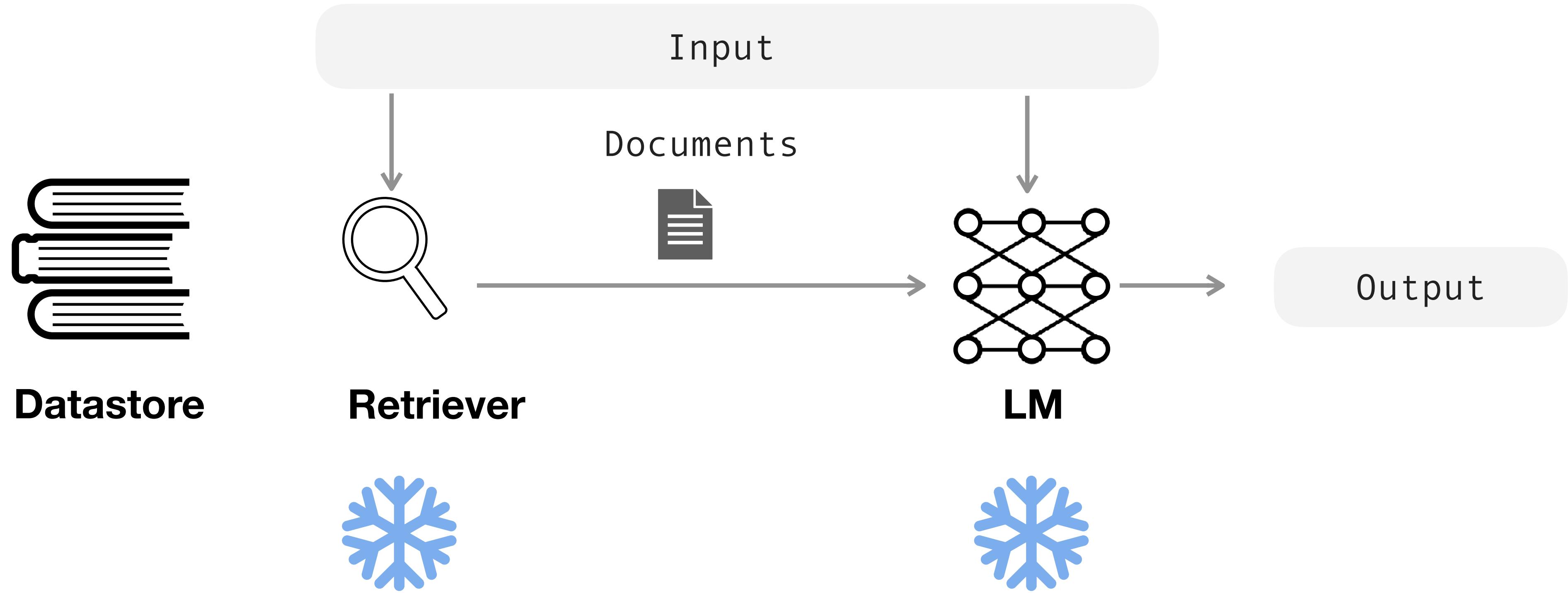


New Foundations of Retrieval-Augmented LMs



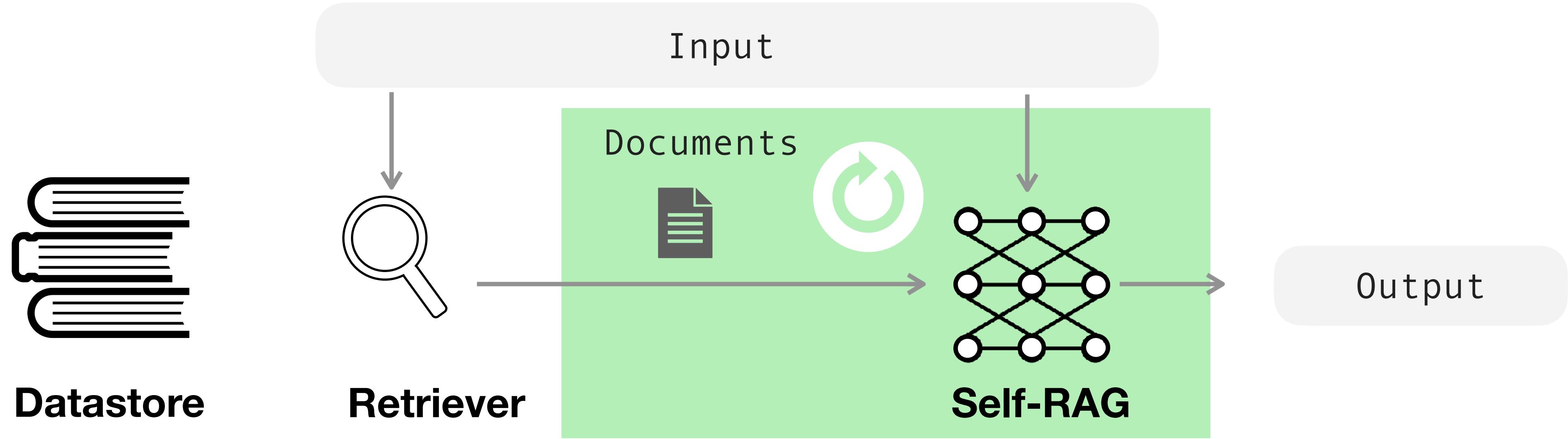


New Foundations of Retrieval-Augmented LMs





New Foundations of Retrieval-Augmented LMs

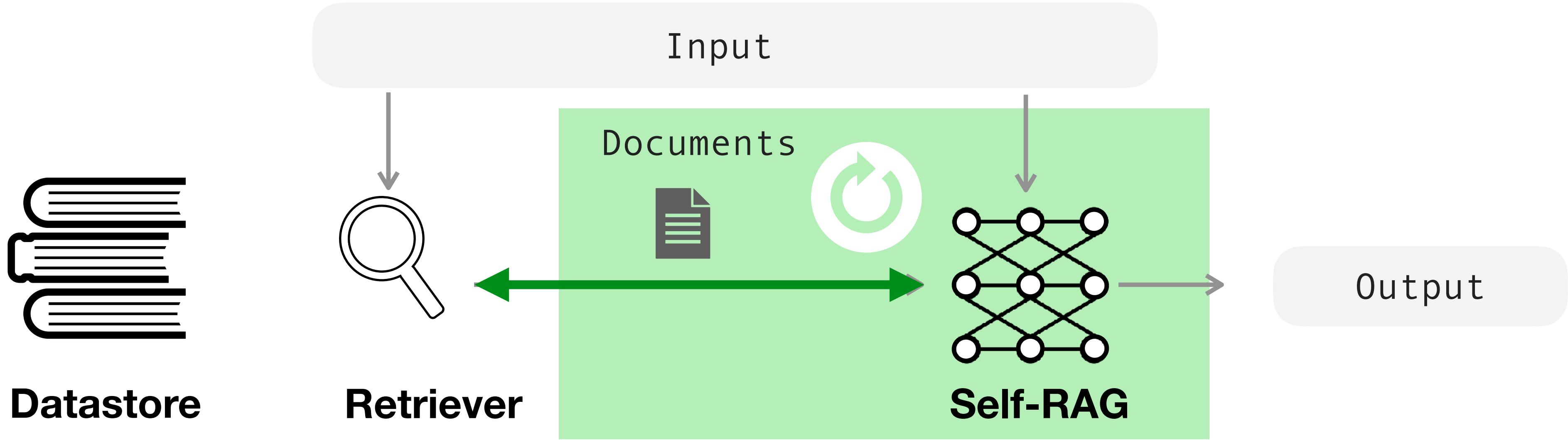


Asai et al., ICLR 2024 (**Oral - Top 1%; Most Influential ICLR 2024 Paper #13; Best Paper Honorable Mention at a NeurIPS Workshop**)

Asai et al., ACL 2020 | Yamada, **Asai** et al., EMNLP 2020 (**Most Influential EMNLP 2020 Paper #8**) | **Asai** and Choi, ACL 2021 | **Asai** et al., NAACL 2022 (**Oral**)



New Foundations of Retrieval-Augmented LMs



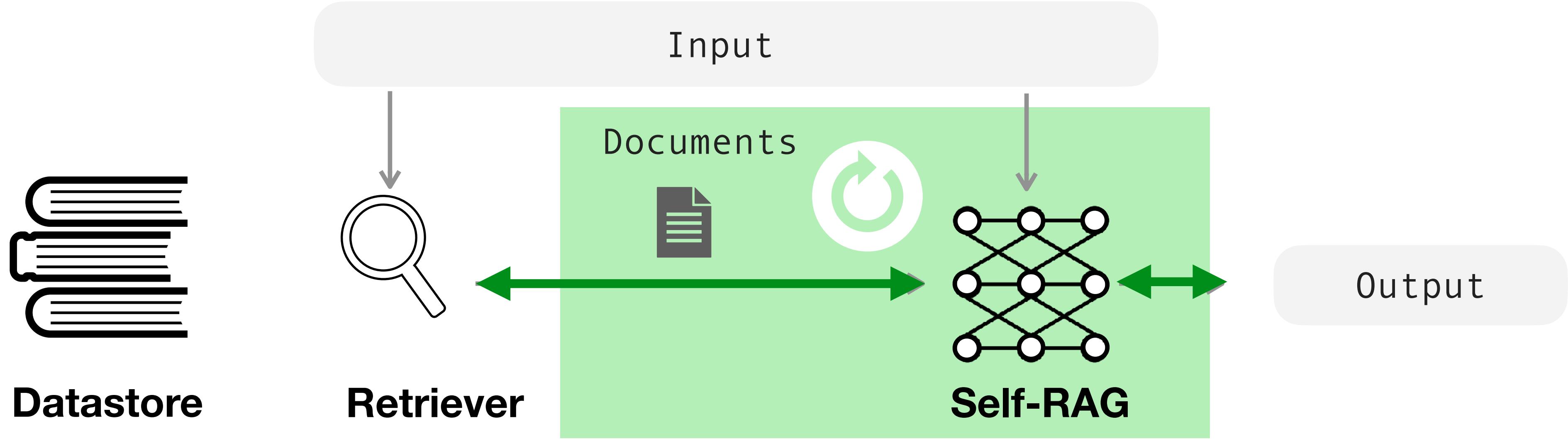
Retrieval when needed

Asai et al., ICLR 2024 (**Oral - Top 1%; Most Influential ICLR 2024 Paper #13; Best Paper Honorable Mention at a NeurIPS Workshop**)

Asai et al., ACL 2020 | Yamada, **Asai** et al., EMNLP 2020 (**Most Influential EMNLP 2020 Paper #8**) | **Asai** and Choi, ACL 2021 | **Asai** et al., NAACL 2022 (**Oral**)



New Foundations of Retrieval-Augmented LMs



Retrieval when needed



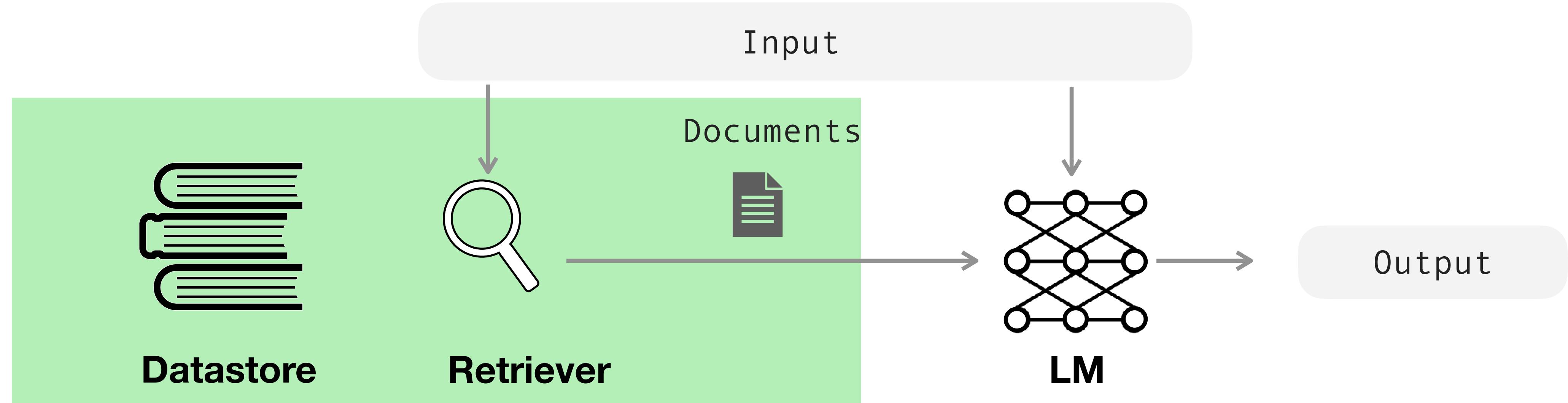
Self-evaluation

Asai et al., ICLR 2024 (**Oral - Top 1%; Most Influential ICLR 2024 Paper #13; Best Paper Honorable Mention at a NeurIPS Workshop**)

Asai et al., ACL 2020 | Yamada, **Asai** et al., EMNLP 2020 (**Most Influential EMNLP 2020 Paper #8**) | **Asai** and Choi, ACL 2021 | **Asai** et al., NAACL 2022 (**Oral**)



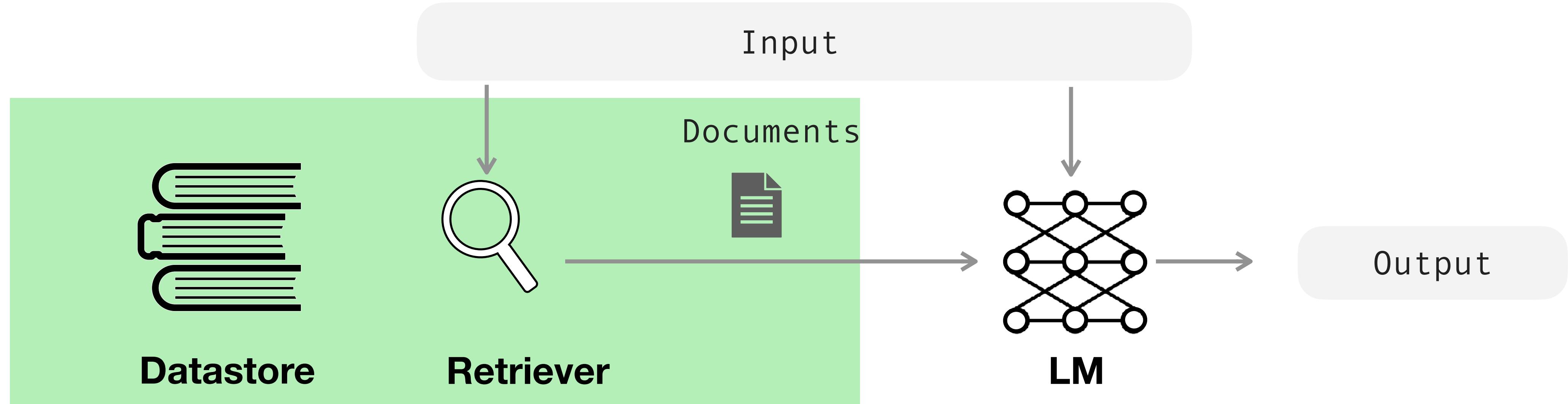
New Foundations of Retrieval-Augmented LMs



Asai et al., ACL Findings 2023 | Shao, He, **Asai** et al., NeurIPS 2024 | Yamada, **Asai** et al., EMNLP 2020 (**Most Influential EMNLP 2020 Paper #8**)
Asai et al. ICLR 2020 | Yamada, **Asai** et al. ACL 2021 | **Asai** et al., NeurIPS 2021 | Lin, **Asai** et al., EMNLP Findings 2023



New Foundations of Retrieval-Augmented LMs

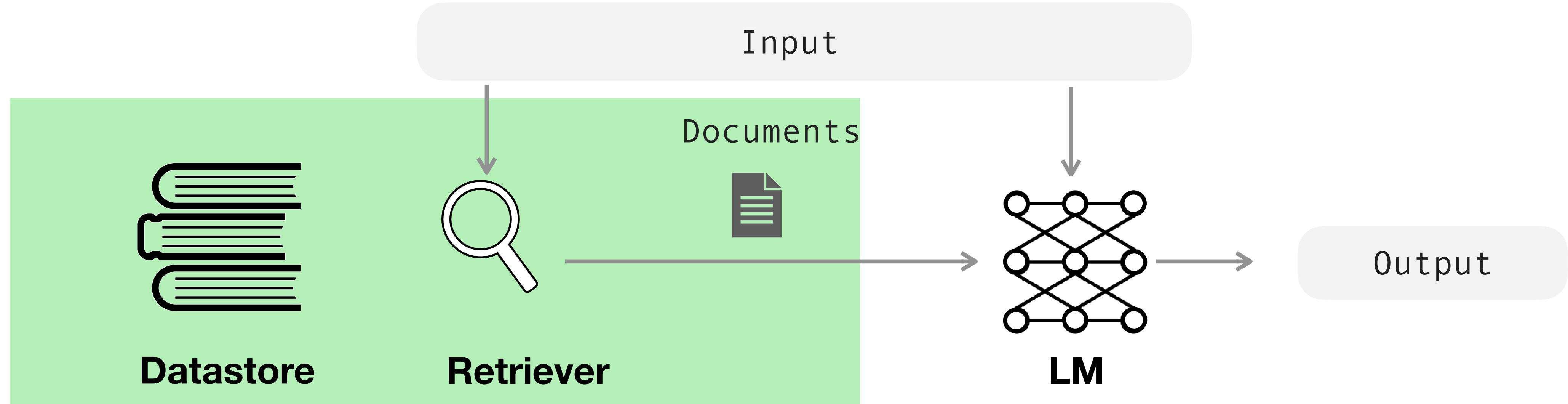


Massively expand datastore

Asai et al., ACL Findings 2023 | Shao, He, **Asai** et al., NeurIPS 2024 | Yamada, **Asai** et al., EMNLP 2020 (**Most Influential EMNLP 2020 Paper #8**)
Asai et al. ICLR 2020 | Yamada, **Asai** et al. ACL 2021 | **Asai** et al., NeurIPS 2021 | Lin, **Asai** et al., EMNLP Findings 2023



New Foundations of Retrieval-Augmented LMs



Massively expand datastore

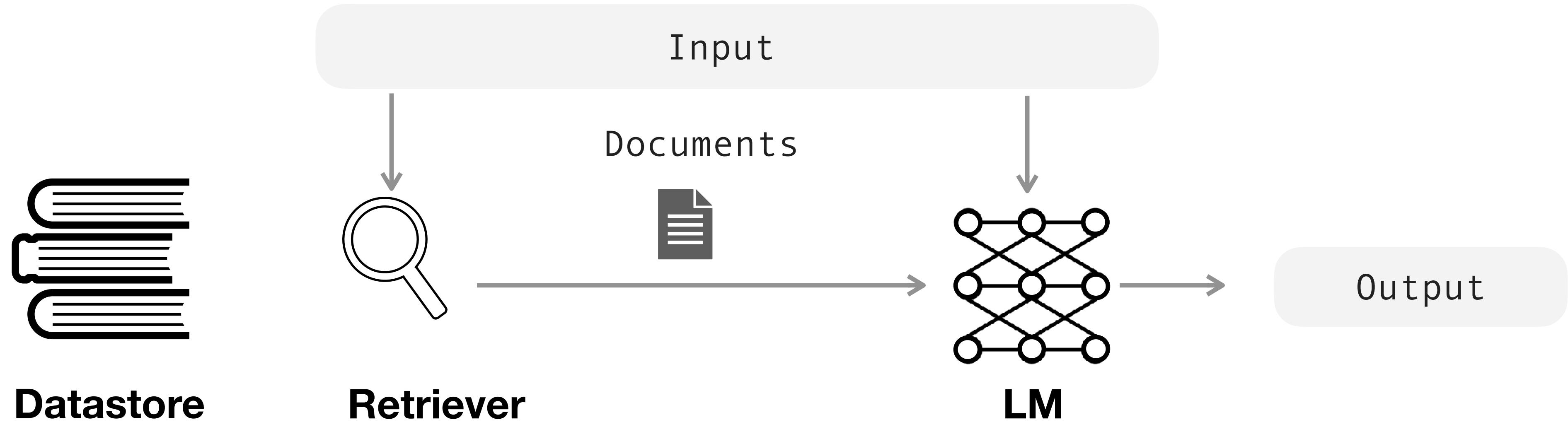


Make retrievers adaptive to instructions

Asai et al., ACL Findings 2023 | Shao, He, **Asai** et al., NeurIPS 2024 | Yamada, **Asai** et al., EMNLP 2020 (**Most Influential EMNLP 2020 Paper #8**)
Asai et al. ICLR 2020 | Yamada, **Asai** et al. ACL 2021 | **Asai** et al., NeurIPS 2021 | Lin, **Asai** et al., EMNLP Findings 2023

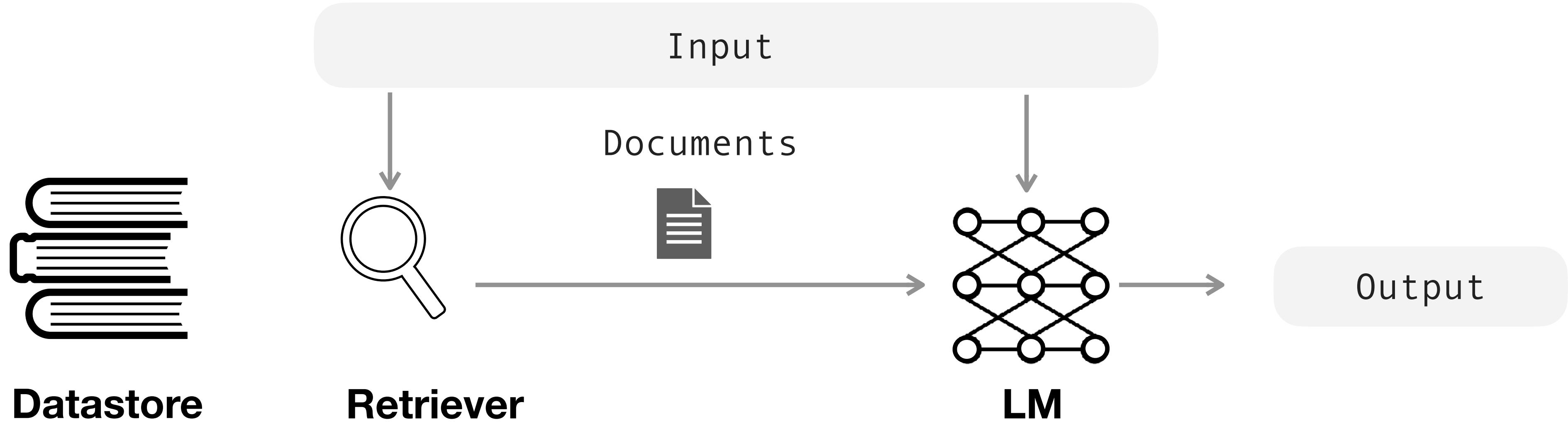


New Foundations of Retrieval-Augmented LMs





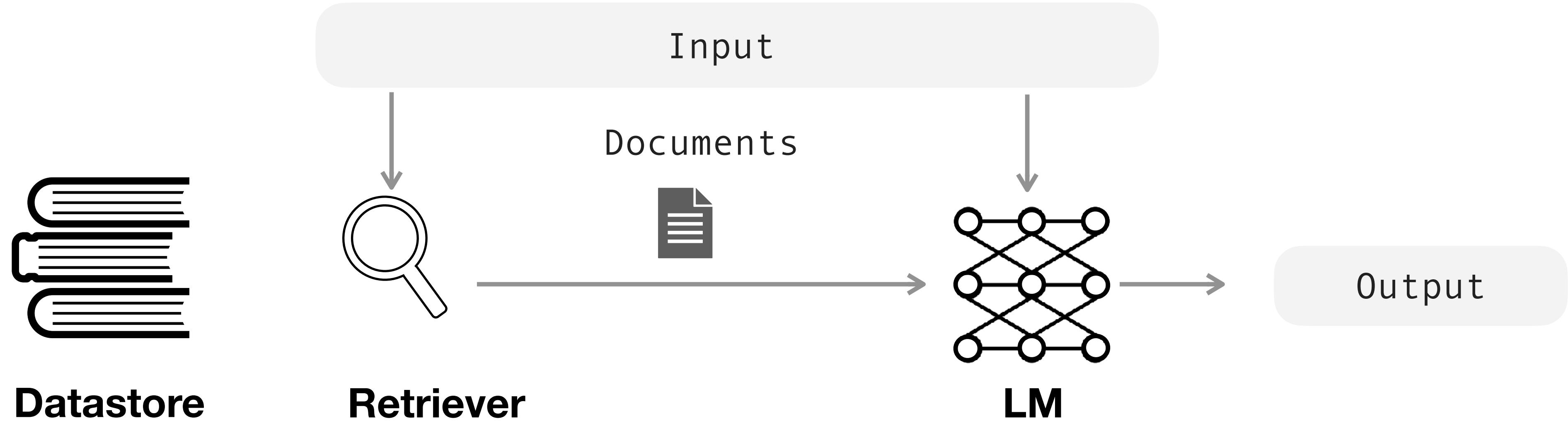
New Foundations of Retrieval-Augmented LMs



Opens up **new research areas**

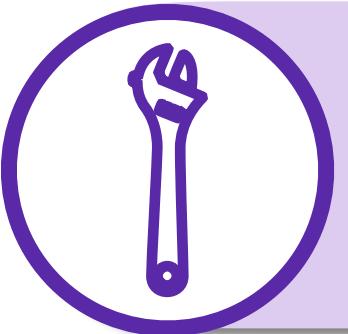


New Foundations of Retrieval-Augmented LMs

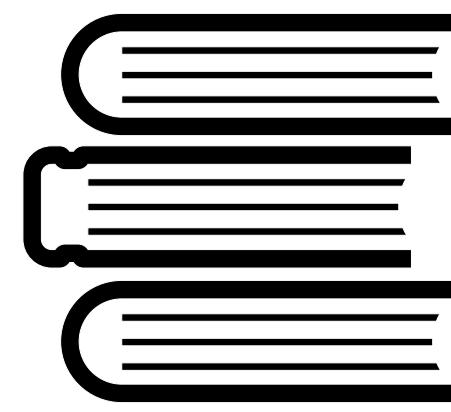


Opens up **new research areas**

Foundations of **state-of-the-art systems**



New Foundations of Retrieval-Augmented LMs



Datastore

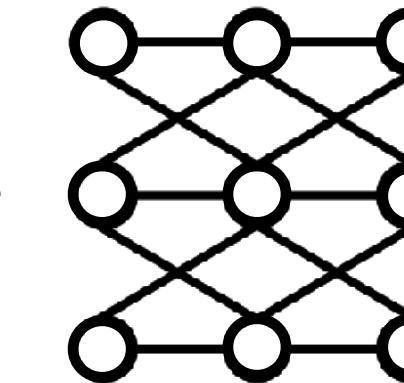


Retriever

Documents



Input



LM

Output

Opens up **new research areas**



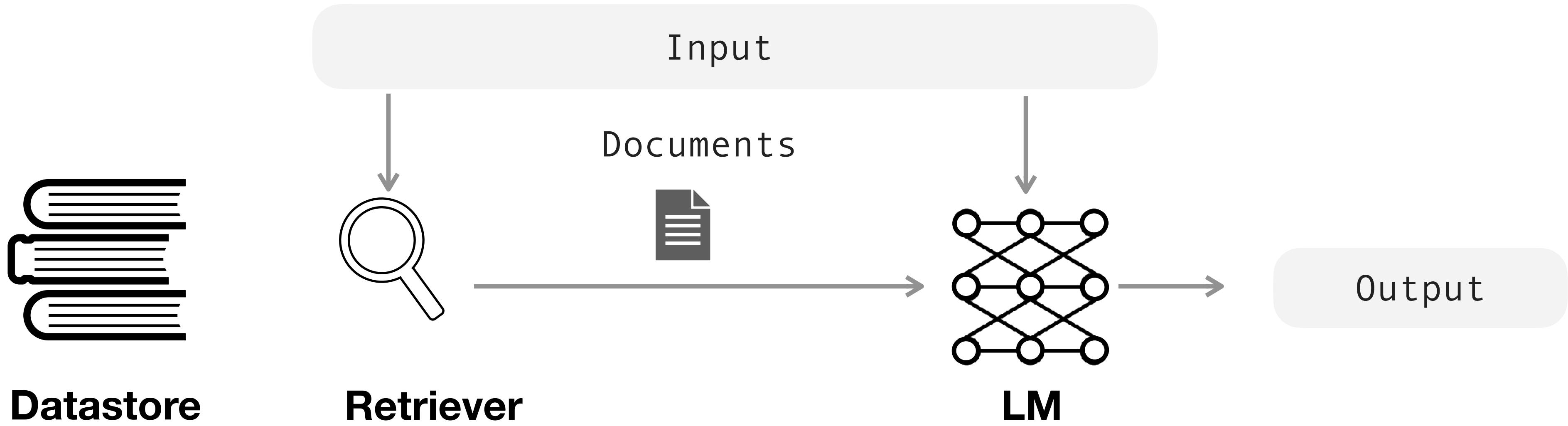
Foundations of **state-of-the-art systems**

Integrations into **major libraries**





New Foundations of Retrieval-Augmented LMs



Opens up **new research areas**

Foundations of **state-of-the-art systems**

Integrations into **major libraries**



Used in **industry applications**





Real-World Impacts of Retrieval-Augmented LMs



Real-World Impacts of Retrieval-Augmented LMs

Are there any studies showing if LMs can help scientists to synthesize scientific literature?
Could you suggest recent papers?

Answer

Share Disclaimer

Yes, there are several studies that demonstrate the potential of Large Language Models (LLMs) to assist scientists in synthesizing scientific literature. Here are some recent papers that showcase the capabilities of LLMs in this area:

1. OpenScholar: Synthesizing Scientific Literature with Retrieval-augmented LMs ([Asai et al. 2024](#))

- This paper introduces OpenScholar, a retrieval-augmented language model that can synthesize scientific literature by identifying relevant passages from 45 million open-access papers and generating citation-backed responses. The authors evaluate OpenScholar on a large-scale multi-domain benchmark, ScholarQABench, and show that it outperforms GPT-4o and PaperQA2 in correctness, with human evaluators preferring OpenScholar's responses over expert-written ones 51% and 70% of the time, respectively.

2. LLMs4Synthesis: Leveraging Large Language Models for Scientific Synthesis ([Giglou et al. 2024¹](#)) ([Giglou et al. 2024²](#))

- This paper proposes the LLMs4Synthesis framework, which aims to enhance the capabilities of LLMs in generating high-quality scientific syntheses. The authors introduce new synthesis types, quality evaluation criteria, and leverage reinforcement learning and AI feedback to optimize synthesis quality. They also provide publicly available resources to further advance research in this domain.

3. SciPIP: An LLM-based Scientific Paper Idea Proposer ([Wang et al. 2024](#))

- This paper introduces SciPIP, an LLM-based system that leverages advanced natural language processing capabilities to generate scientific paper ideas across disciplines, making it a promising tool for researchers.

These studies demonstrate the potential of LLMs in synthesizing scientific literature across various domains ([Clercq et al. 2024](#)). However, it is essential to remember that LLMs are not yet fully reliable and must be used judiciously and in conjunction with human expertise ([Giglou et al. 2024¹](#)).

SciPIP: An LLM-based Scientific Paper Idea Proposer

Wenxiao Wang, Lihui Gu, Liye Zhang, Yunxiang Luo, Yi Dai, Chen Shen et al. arXiv.org, 2024.

*Around 60 years ago, scientists began exploring scientific discoveries based on literature retrieval, known as Literature-Based Discovery (LBD) (Swanson, 1986). This approach concentrated on a specific, narrow type of hypothesis: the connections between pairs of concepts, often involving drugs and diseases. LBD introduced the

<https://openscholar.allen.ai/>

Asai et al., Nature (Under Review) 2024 | Wang*, **Asai*** (equally contributed) NAACL Findings 2025 | **Asai** et al. NAACL 2021 (**Oral**) | **Asai** et al., NeurIPS 2021
Shen, **Asai** et al., ACL (Industry) 2023 | Ogundepo, ..., **Asai**, et al., EMNLP Findings 2023 | **Asai** et al., NAACL 2024 (**Oral**) | Yu*, **Asai*** et al., EMNLP Findings 2022



Real-World Impacts of Retrieval-Augmented LMs

Are there any studies showing if LMs can help scientists to synthesize scientific literature?
Could you suggest recent papers?

Answer

Share Disclaimer

Yes, there are several studies that demonstrate the potential of Large Language Models (LLMs) to assist scientists in synthesizing scientific literature. Here are some recent papers that showcase the capabilities of LLMs in this area:

1. OpenScholar: Synthesizing Scientific Literature with Retrieval-augmented LMs (Asai et al. 2024)

- This paper introduces OpenScholar, a retrieval-augmented language model that can synthesize scientific literature by identifying relevant passages from 45 million open-access papers and generating citation-backed responses. The authors evaluate OpenScholar on a large-scale multi-domain benchmark, ScholarQABench, and show that it outperforms GPT-4o and PaperQA2 in correctness, with human evaluators preferring OpenScholar's responses over expert-written ones 51% and 70% of the time, respectively.

2. LLMs4Synthesis: Leveraging Large Language Models for Scientific Synthesis (Giglou et al. 2024¹) (Giglou et al. 2024²)

- This paper proposes the LLMs4Synthesis framework, which aims to enhance the capabilities of LLMs in generating high-quality scientific syntheses. The authors introduce new synthesis types, quality evaluation criteria, and leverage reinforcement learning and AI feedback to optimize synthesis quality. They also provide publicly available resources to further advance research in this domain.

3. SciPIP: An LLM-based Scientific Paper Idea Proposer (Wang et al. 2024)

- This paper introduces SciPIP, an LLM-based system that leverages advanced natural language processing capabilities to generate ideas for scientific papers across disciplines, making it a promising tool for researchers.

These studies demonstrate the potential of LLMs in synthesizing scientific literature across various domains (Clercq et al. 2024). However, it is essential to use these tools judiciously and in conjunction with human expertise (Giglou et al. 2024¹).

SciPIP: An LLM-based Scientific Paper Idea Proposer

Wenxiao Wang, Lihui Gu, Liye Zhang, Yunxiang Luo, Yi Dai, Chen Shen et al. arXiv.org, 2024.

*Around 60 years ago, scientists began exploring scientific discoveries based on literature retrieval, known as Literature-Based Discovery (LBD) (Swanson, 1986). This approach concentrated on a specific, narrow type of hypothesis: the connections between pairs of concepts, often involving drugs and diseases. LBD introduced the

<https://openscholar.allen.ai/>

Asai et al., Nature (Under Review) 2024 | Wang*, **Asai*** (equally contributed) NAACL Findings 2025 | **Asai** et al. NAACL 2021 (**Oral**) | **Asai** et al., NeurIPS 2021
Shen, **Asai** et al., ACL (Industry) 2023 | Ogundepo, ..., **Asai**, et al., EMNLP Findings 2023 | **Asai** et al., NAACL 2024 (**Oral**) | Yu*, **Asai*** et al., EMNLP Findings 2022



Real-World Impacts of Retrieval-Augmented LMs

Are there any studies showing if LMs can help scientists to synthesize scientific literature?
Could you suggest recent papers?

Answer

Share Disclaimer

Yes, there are several studies that demonstrate the potential of Large Language Models (LLMs) to assist scientists in synthesizing scientific literature. Here are some recent papers that showcase the capabilities of LLMs in this area:

1. OpenScholar: Synthesizing Scientific Literature with Retrieval-augmented LMs (Asai et al. 2024)

- This paper introduces OpenScholar, a retrieval-augmented language model that can synthesize scientific literature by identifying relevant passages from 45 million open-access papers and generating citation-backed responses. The authors evaluate OpenScholar on a large-scale multi-domain benchmark, ScholarQABench, and show that it outperforms GPT-4o and PaperQA2 in correctness, with human evaluators preferring OpenScholar's responses over expert-written ones 51% and 70% of the time, respectively.

2. LLMs4Synthesis: Leveraging Large Language Models for Scientific Synthesis (Giglou et al. 2024¹) (Giglou et al. 2024²)

- This paper proposes the LLMs4Synthesis framework, which aims to enhance the capabilities of LLMs in generating high-quality scientific syntheses. The authors introduce new synthesis types, quality evaluation criteria, and leverage reinforcement learning and AI feedback to optimize synthesis quality. They also provide publicly available resources to further advance research in this domain.

3. SciPIP: An LLM-based Scientific Paper Idea Proposer (Wang et al. 2024)

- This paper introduces SciPIP, an LLM-based system that leverages advanced natural language processing capabilities to generate ideas for scientific papers across disciplines, making it a promising tool for researchers.

These studies demonstrate the potential of LLMs in synthesizing scientific literature across various domains (Clercq et al. 2024). However, it is essential to remember that LLMs are not yet fully reliable and must be used judiciously and in conjunction with human expertise (Giglou et al. 2024¹).

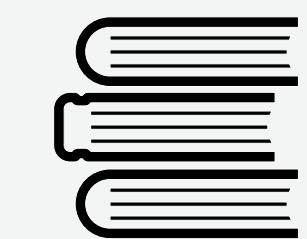
SciPIP: An LLM-based Scientific Paper Idea Proposer

Wenxiao Wang, Lihui Gu, Liye Zhang, Yunxiang Luo, Yi Dai, Chen Shen et al. arXiv.org, 2024.

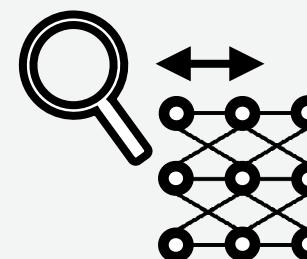
*Around 60 years ago, scientists began exploring scientific discoveries based on literature retrieval, known as Literature-Based Discovery (LBD) (Swanson, 1986). This approach concentrated on a specific, narrow type of hypothesis: the connections between pairs of concepts, often involving drugs and diseases. LBD introduced the

<https://openscholar.allen.ai/>

Asai et al., Nature (Under Review) 2024 | Wang*, **Asai*** (equally contributed) NAACL Findings 2025 | **Asai** et al. NAACL 2021 (**Oral**) | **Asai** et al., NeurIPS 2021
Shen, **Asai** et al., ACL (Industry) 2023 | Ogundepo, ..., **Asai**, et al., EMNLP Findings 2023 | **Asai** et al., NAACL 2024 (**Oral**) | Yu*, **Asai*** et al., EMNLP Findings 2022



Scientific datastore



Improve algorithms



New expert evaluation



Open and public demo

8B fully open LM **outperforms GPT4o**



Real-World Impacts of Retrieval-Augmented LMs

Are there any studies showing if LMs can help scientists to synthesize scientific literature?
Could you suggest recent papers?

Answer

Share Disclaimer

Yes, there are several studies that demonstrate the potential of Large Language Models (LLMs) to assist scientists in synthesizing scientific literature. Here are some recent papers that showcase the capabilities of LLMs in this area:

1. OpenScholar: Synthesizing Scientific Literature with Retrieval-augmented LMs (Asai et al. 2024)

- This paper introduces OpenScholar, a retrieval-augmented language model that can synthesize scientific literature by identifying relevant passages from 45 million open-access papers and generating citation-backed responses. The authors evaluate OpenScholar on a large-scale multi-domain benchmark, ScholarQABench, and show that it outperforms GPT-4o and PaperQA2 in correctness, with human evaluators preferring OpenScholar's responses over expert-written ones 51% and 70% of the time, respectively.

2. LLMs4Synthesis: Leveraging Large Language Models for Scientific Synthesis (Giglou et al. 2024¹) (Giglou et al. 2024²)

- This paper proposes the LLMs4Synthesis framework, which aims to enhance the capabilities of LLMs in generating high-quality scientific syntheses. The authors introduce new synthesis types, quality evaluation criteria, and leverage reinforcement learning and AI feedback to optimize synthesis quality. They also provide publicly available resources to further advance research in this domain.

3. SciPIP: An LLM-based Scientific Paper Idea Proposer (Wang et al. 2024)

- This paper introduces SciPIP, an LLM-based system that leverages advanced natural language processing capabilities to generate ideas for scientific papers across disciplines, making it a promising tool for researchers.

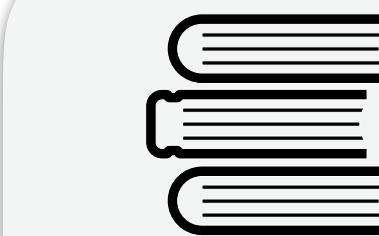
These studies demonstrate the potential of LLMs in synthesizing scientific literature across various domains (Clercq et al. 2024). However, it is essential to remember that LLMs are not perfect and must be used judiciously and in conjunction with human expertise (Giglou et al. 2024¹).

SciPIP: An LLM-based Scientific Paper Idea Proposer

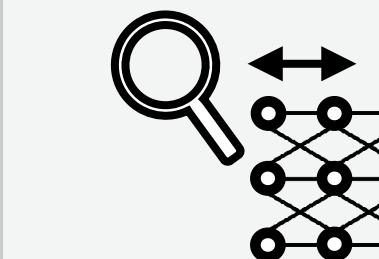
Wenxiao Wang, Lihui Gu, Liye Zhang, Yunxiang Luo, Yi Dai, Chen Shen et al. arXiv.org, 2024.

¹Around 60 years ago, scientists began exploring scientific discoveries based on literature retrieval, known as Literature-Based Discovery (LBD) (Swanson, 1986). This approach concentrated on a specific, narrow type of hypothesis: the connections between pairs of concepts, often involving drugs and diseases. LBD introduced the

<https://openscholar.allen.ai/>



Scientific datastore



Improve algorithms



New expert evaluation



Open and public demo

8B fully open LM **outperforms GPT4o**

30k researchers have used OpenScholar

Asai et al., Nature (Under Review) 2024 | Wang*, **Asai*** (equally contributed) NAACL Findings 2025 | **Asai** et al. NAACL 2021 (**Oral**) | **Asai** et al., NeurIPS 2021
Shen, **Asai** et al., ACL (Industry) 2023 | Ogundepo, ..., **Asai**, et al., EMNLP Findings 2023 | **Asai** et al., NAACL 2024 (**Oral**) | Yu*, **Asai*** et al., EMNLP Findings 2022

Frontiers of Retrieval-Augmented LMs



From Monolithic to Retrieval-Augmented LMs

Why?



New Foundations of Retrieval-Augmented LMs

How?



Real-World Impacts of Retrieval-Augmented LMs

What?

When Do LMs Hallucinate?

When Do LMs **Hallucinate**?

LMs output contains factual inaccuracies

When Do LMs Hallucinate?

When Do LMs Hallucinate?



Who is Sam Altman?

(2022/07)

When Do LMs Hallucinate?



Who is Sam Altman?

(2022/07)



(GPT3)

When Do LMs Hallucinate?



Who is Sam Altman?

(2022/07)



(GPT3)

Sam Altman is an American entrepreneur, investor, and prominent figure in the technology industry. He is best known as the CEO of OpenAI, an organization dedicated to developing and ensuring the safe use of artificial general intelligence (AGI). Before joining OpenAI, Altman was the president of Y Combinator, a renowned startup accelerator that has supported companies like Airbnb, Dropbox, and Stripe. He also co-founded Loopt, a location-based social networking app, which was acquired in 2012, raising more than \$30 million in venture capital.

When Do LMs Hallucinate?



Who is Sam Altman?

(2022/07)



(GPT3)

Sam Altman is an American entrepreneur, investor, and prominent figure in the technology industry. He is best known as the CEO of OpenAI, an organization dedicated to developing and ensuring the safe use of artificial general intelligence (AGI). Before joining OpenAI, Altman was the president of Y Combinator, a renowned startup accelerator that has supported companies like Airbnb, Dropbox, and Stripe. He also co-founded Loopt, a location-based social networking app, which was acquired in 2012, raising more than \$30 million in venture capital.



When Do LMs Hallucinate?



Who is Akari Asai?

(2022/07)

When Do LMs Hallucinate?



Who is Akari Asai?

(2022/07)



(GPT3)

When Do LMs Hallucinate?



Who is Akari Asai?

(2022/07)



(GPT3)

Akari Asai is a Japanese voice actress known for her versatile vocal range and her ability to bring animated characters to life. Having graduated from a prestigious voice acting academy in Tokyo, she quickly made her mark in the industry. Akari is best known for her roles in popular anime series where she often voices lead characters that appeal to a wide audience. Her performances are noted for their emotional depth and realism, making her a favorite among fans. Aside from anime, Akari has also lent her voice to numerous video games and dubbed international films, showcasing her adaptability and skill across different media.

When Do LMs Hallucinate?



Who is Akari Asai?

(2022/07)



(GPT3)

Akari Asai is a Japanese voice actress known for her versatile vocal range and her ability to bring animated characters to life. Having graduated from a prestigious voice acting academy in Tokyo, she quickly made her mark in the industry. Akari is best known for her roles in popular anime series where she often voices lead characters that appeal to a wide audience. Her performances are noted for their emotional depth and realism, making her a favorite among fans. Aside from anime, Akari has also lent her voice to numerous video games and dubbed international films, showcasing her adaptability and skill across different media.



When Do LMs Hallucinate?



Who is Akari Asai?

(2022/07)



(GPT3)

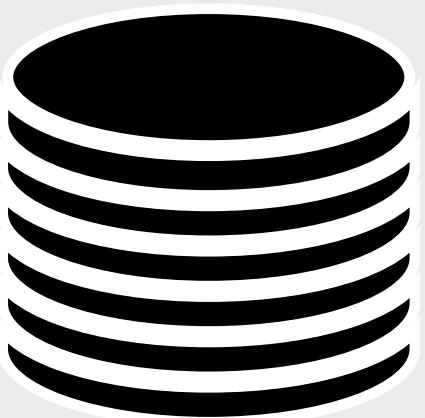
Akari Asai is a Japanese voice actress known for her versatile vocal range and her ability to bring animated characters to life. Having graduated from a prestigious voice acting academy in Tokyo, she quickly made her mark in the industry. Akari is best known for her roles in popular anime series where she often voices lead characters that appeal to a wide audience. Her performances are noted for their emotional depth and realism, making her a favorite among fans. Aside from anime, Akari has also lent her voice to numerous video games and dubbed international films, showcasing her adaptability and skill across different media.



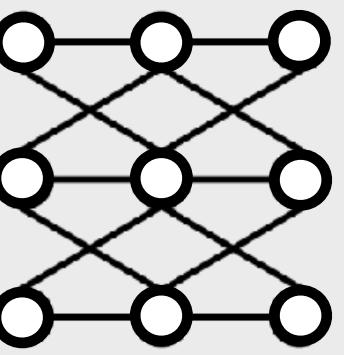
Back in 2022, we didn't know **why** and **when** LMs hallucinate

How LM Training Works

Pre-training Data

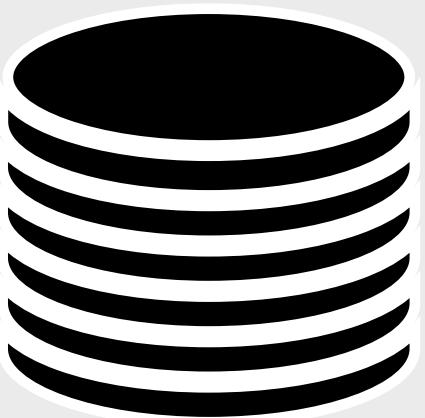


$\text{LM } \theta$



How LM Training Works

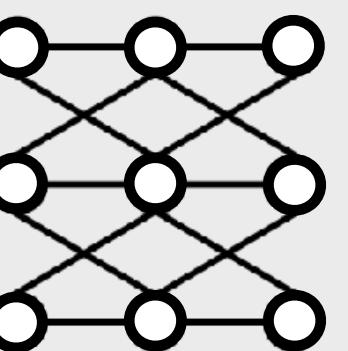
Pre-training Data



ChatGPT is developed
by OpenAI

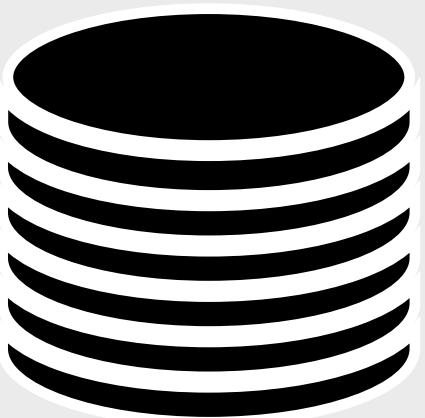


$\text{LM } \theta$



How LM Training Works

Pre-training Data



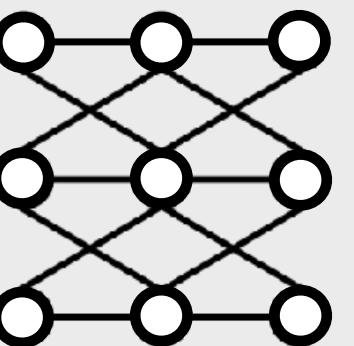
ChatGPT is developed
by OpenAI



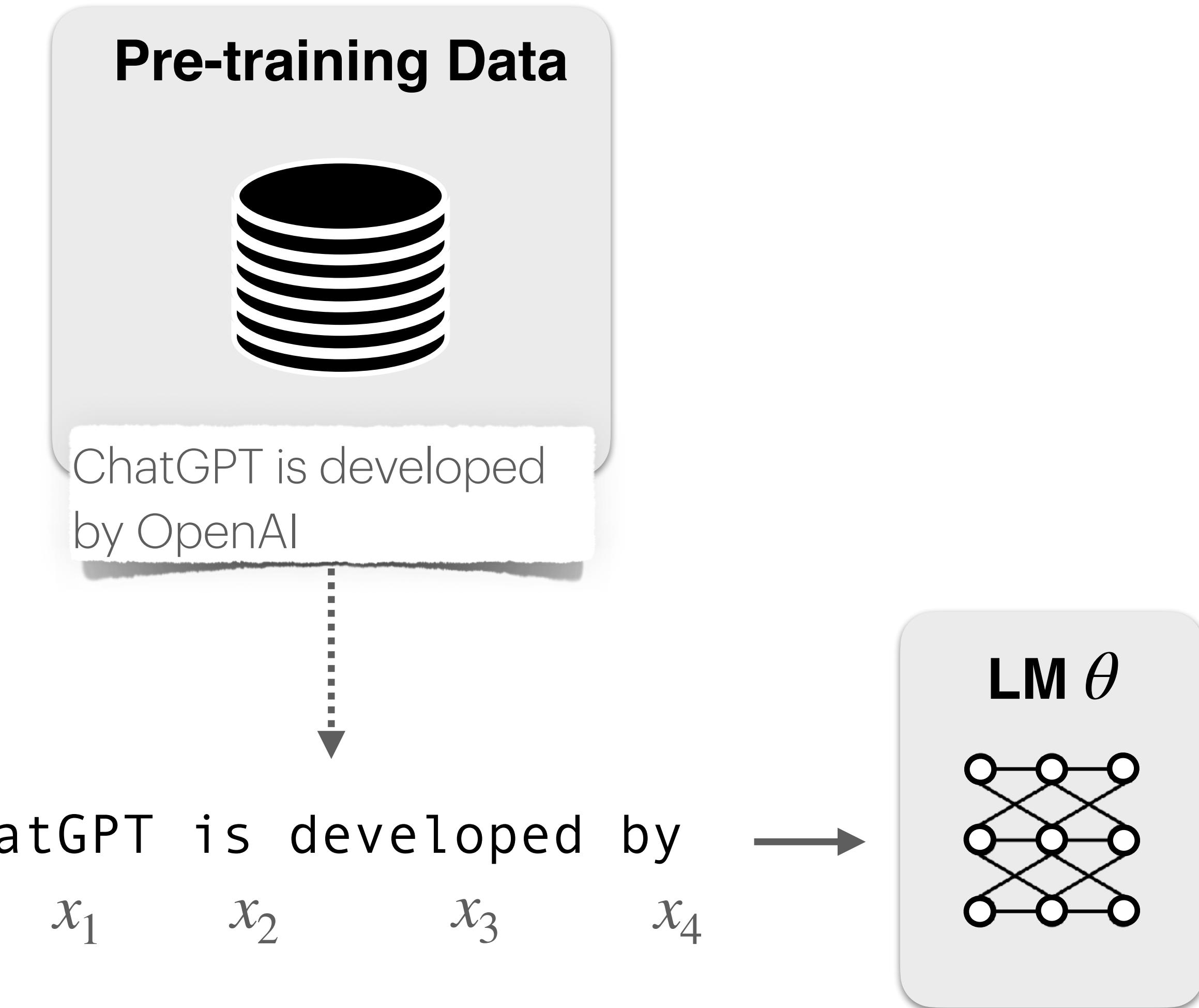
ChatGPT is developed by

$x_1 \quad x_2 \quad x_3 \quad x_4$

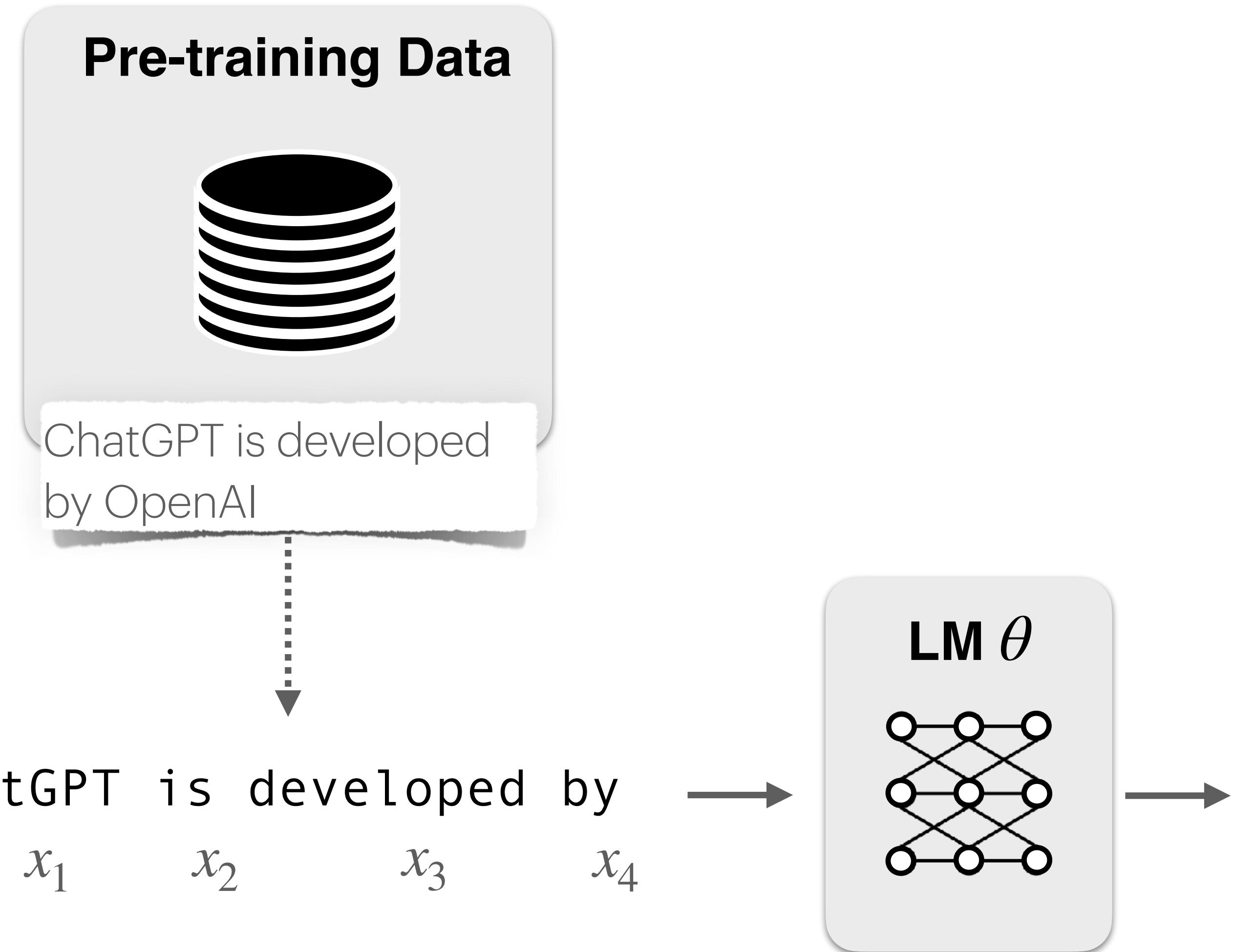
LM θ



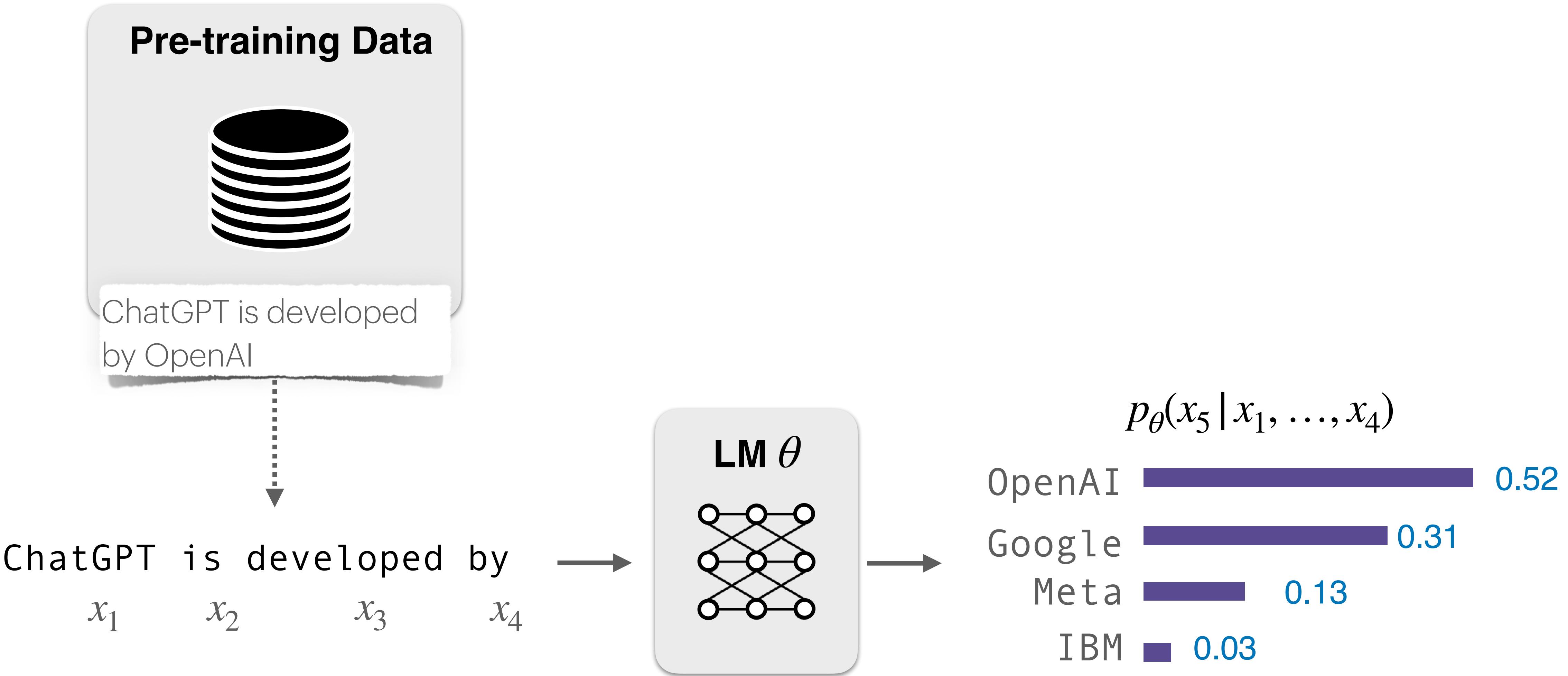
How LM Training Works



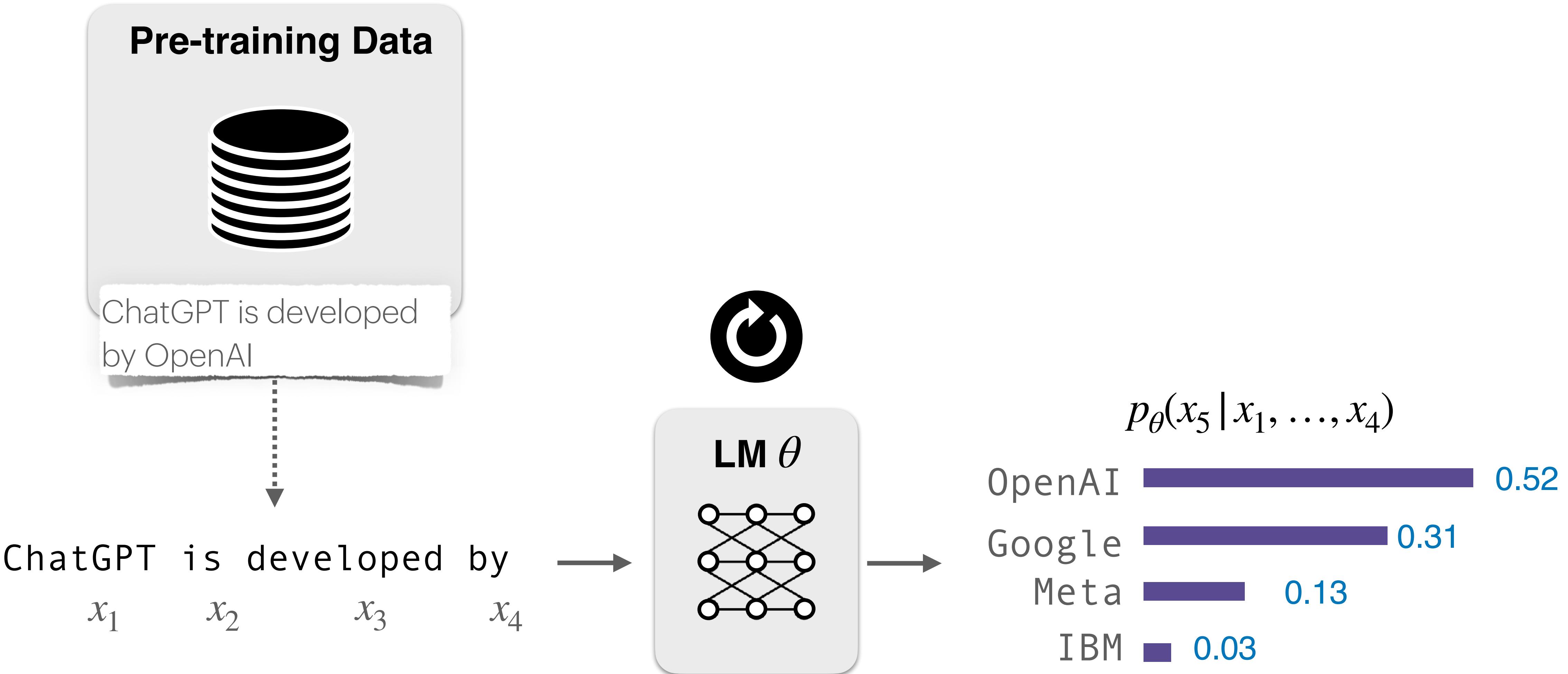
How LM Training Works



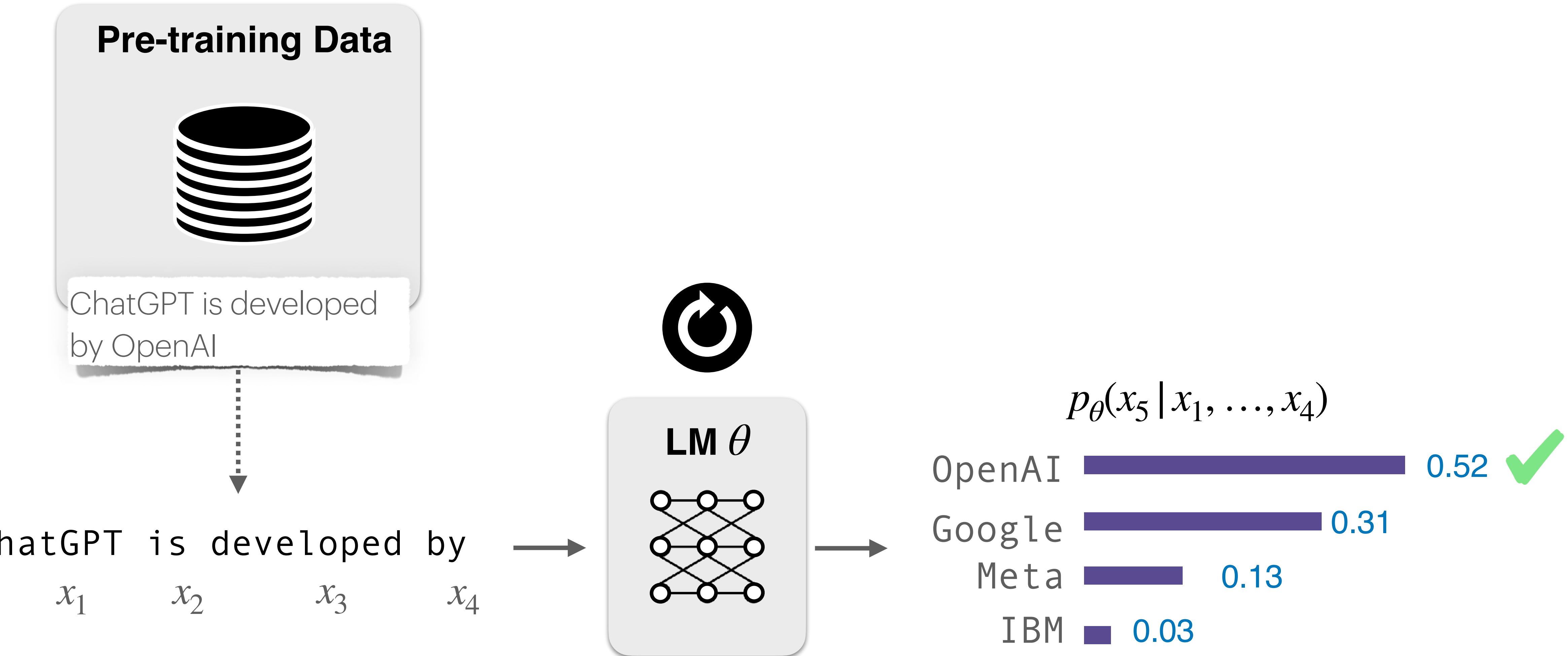
How LM Training Works



How LM Training Works



How LM Training Works



Popularity of Knowledge

Who is Sam Altman?



Who is Akari Asai?



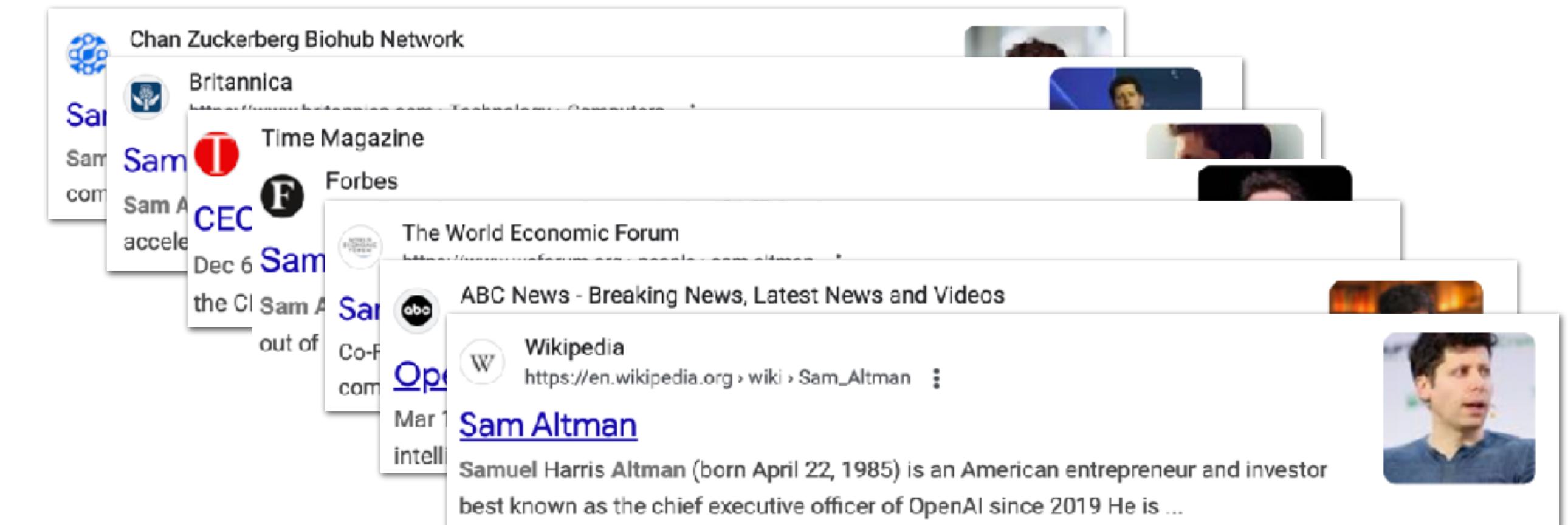
Asai* and Mallen* (contributed equally) et al. ACL 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories.

ACL Best Video Award - Most Viewed

Most Influential ACL 2023 Paper #5 (from Paper Digest)

Popularity of Knowledge

Who is Sam Altman?



Who is Akari Asai?



Asai* and Mallen* (contributed equally) et al. ACL 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories.
ACL Best Video Award - Most Viewed Most Influential ACL 2023 Paper #5 (from Paper Digest)

Popularity of Knowledge

Who is Sam Altman?



Chan Zuckerberg Biohub Network

Britannica

Time Magazine

Forbes

The World Economic Forum

ABC News - Breaking News, Latest News and Videos

Wikipedia

https://en.wikipedia.org/wiki/Sam_Altman

[Sam Altman](#)

Samuel Harris Altman (born April 22, 1985) is an American entrepreneur and investor best known as the chief executive officer of OpenAI since 2019 He is ...

Who is Akari Asai?



 Akari Asai

 EECS Rising Stars 2022
<https://risingstars.utexas.edu> :

Akari Asai - EECS Rising Stars 2022

Akari Asai is a Ph.D. student in the Paul G. Allen School of Computer Science & Engineering at the University of Washington, advised by Prof.

Asai* and Mallen* (contributed equally) et al. ACL 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories.

ACL Best Video Award - Most Viewed

Most Influential ACL 2023 Paper #5 (from Paper Digest) 

Popularity of Knowledge

Who is Sam Altman?



Chan Zuckerberg Biohub Network

Britannica

Time Magazine

Forbes

The World Economic Forum

ABC News - Breaking News, Latest News and Videos

Wikipedia

https://en.wikipedia.org/wiki/Sam_Altman

[Sam Altman](#)

Samuel Harris Altman (born April 22, 1985) is an American entrepreneur and investor best known as the chief executive officer of OpenAI since 2019. He is ...

Who is Akari Asai?



Akari Asai

EECS Rising Stars 2022

<https://risingstars.utexas.edu>

Akari Asai - EECS Rising Stars 2022

Akari Asai is a Ph.D. student in the Paul G. Allen School of Computer Science & Engineering at the University of Washington, advised by Prof.

Asai* and Mallen* (contributed equally) et al. ACL 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories.

ACL Best Video Award - Most Viewed

Most Influential ACL 2023 Paper #5 (from Paper Digest) 

Popularity of Knowledge

Who is Sam Altman?



Subject popularity
(~Wikipedia page views)



Chan Zuckerberg Biohub Network

Britannica

Sam

Sam com

Sam T

Time Magazine

Forbes

CEC

Dec 6 Sam

the CI Sam A

out of

Sam

ABC News - Breaking News, Latest News and Videos

Wikipedia

https://en.wikipedia.org/wiki/Sam_Altman

Sam Altman

Samuel Harris Altman (born April 22, 1985) is an American entrepreneur and investor best known as the chief executive officer of OpenAI since 2019 He is ...

A photo of Sam Altman is visible on the right.

Who is Akari Asai?



Akari Asai

EECS Rising Stars 2022

https://risingstars.utexas.edu

Akari Asai - EECS Rising Stars 2022

Akari Asai is a Ph.D. student in the Paul G. Allen School of Computer Science & Engineering at the University of Washington, advised by Prof.

Asai* and Mallen* (contributed equally) et al. ACL 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories.

ACL Best Video Award - Most Viewed

Most Influential ACL 2023 Paper #5 (from Paper Digest)

Effect of Scaling on Hallucination

Accuracy on
PopQA

100

75

50

25

0

1

10

100

Size of LMs (Billion Parameters)

- Popular Questions
 - *Monthly page views > 100k
- Rare Questions
 - *Monthly page views < 1k

Effect of Scaling on Hallucination

Accuracy on
PopQA

100

75

50

25

0

1

10

100

Size of LMs (Billion Parameters)

- Popular Questions
 - *Monthly page views > 100k
- Rare Questions
 - *Monthly page views < 1k

Effect of Scaling on Hallucination

Accuracy on
PopQA

100

75

50

25

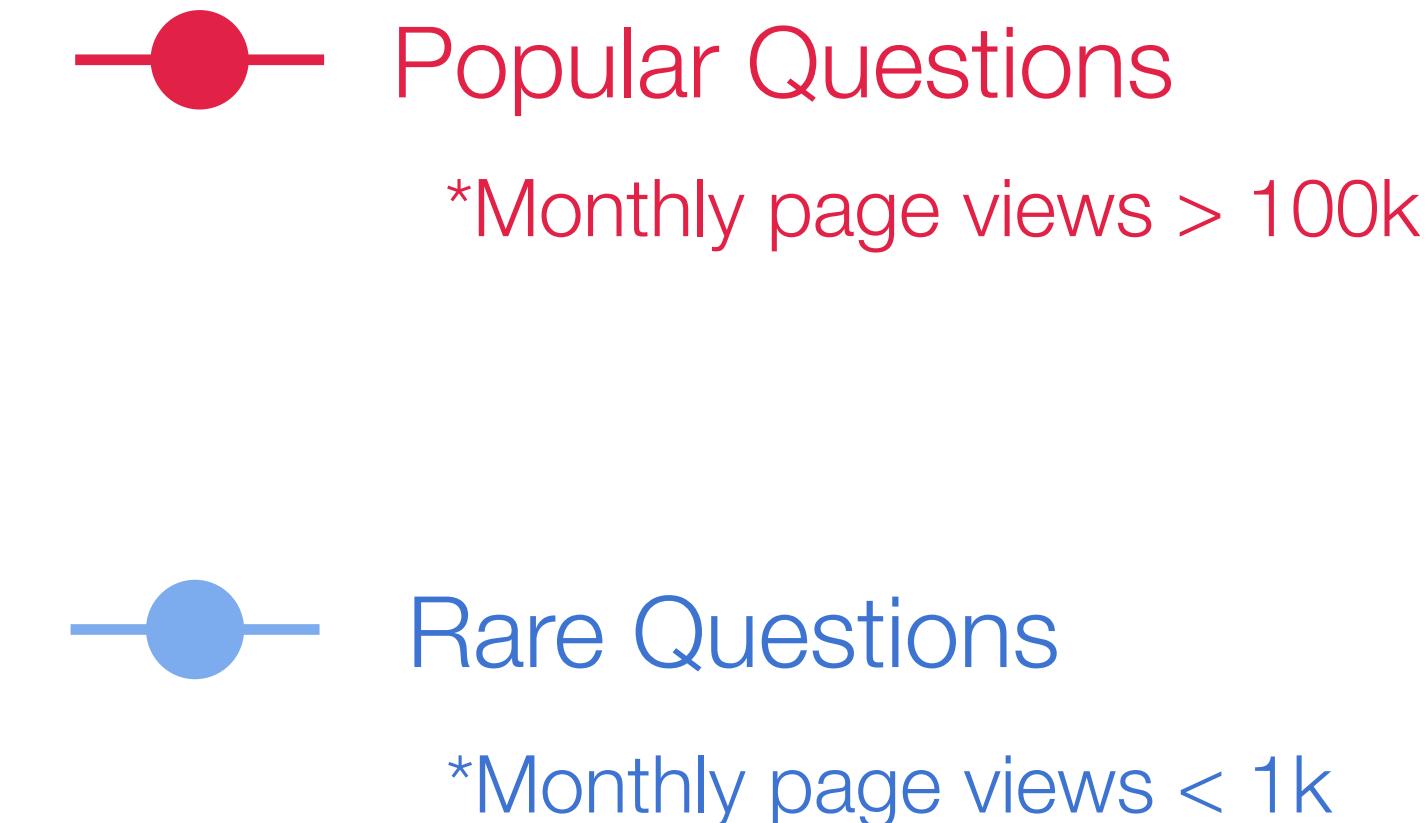
0

1

10

100

Size of LMs (Billion Parameters)



Effect of Scaling on Hallucination

Accuracy on
PopQA

100

75

50

25

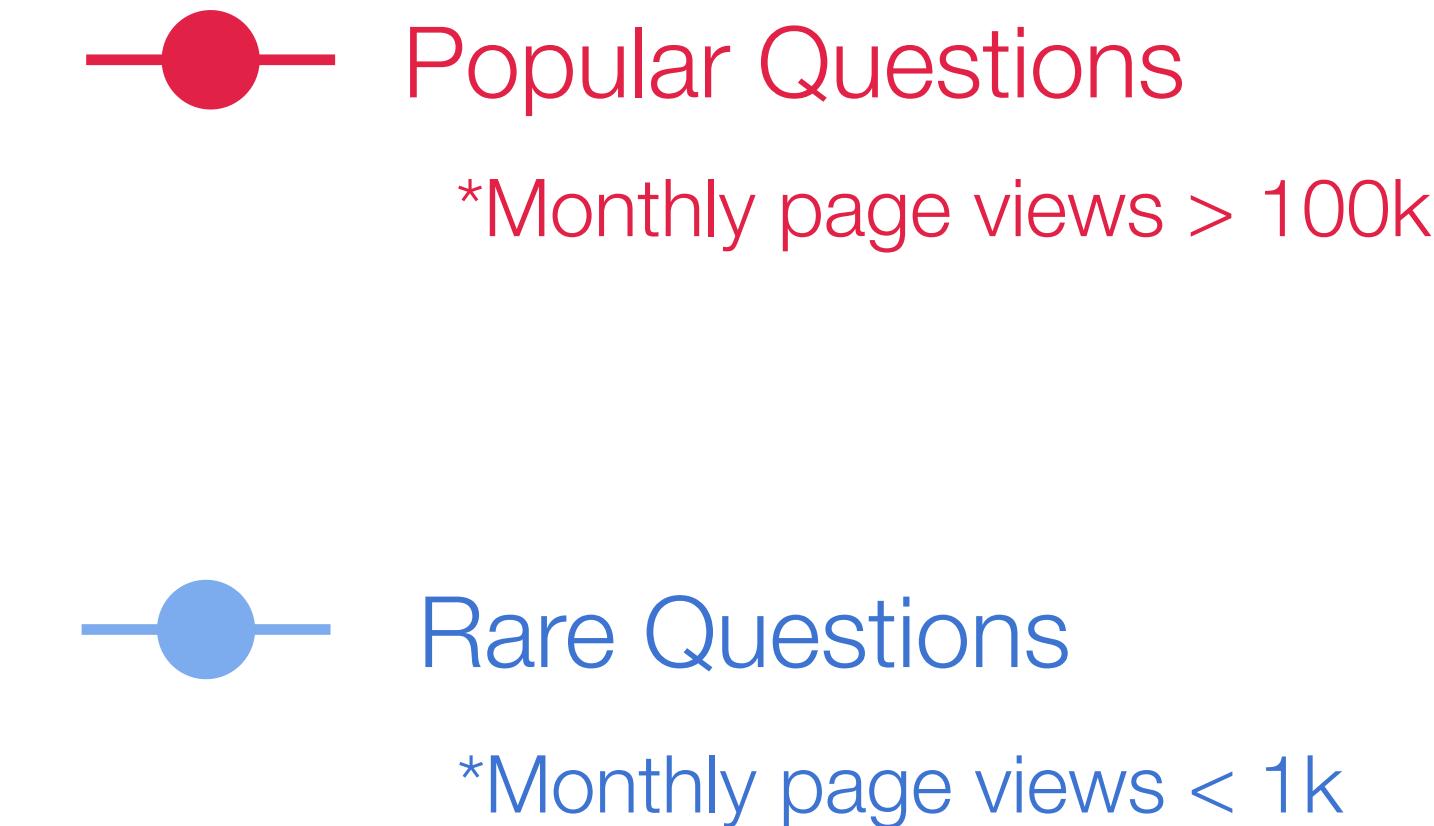
0

1

10

100

Size of LMs (Billion Parameters)



Effect of Scaling on Hallucination

Accuracy on
PopQA

100

75

50

25

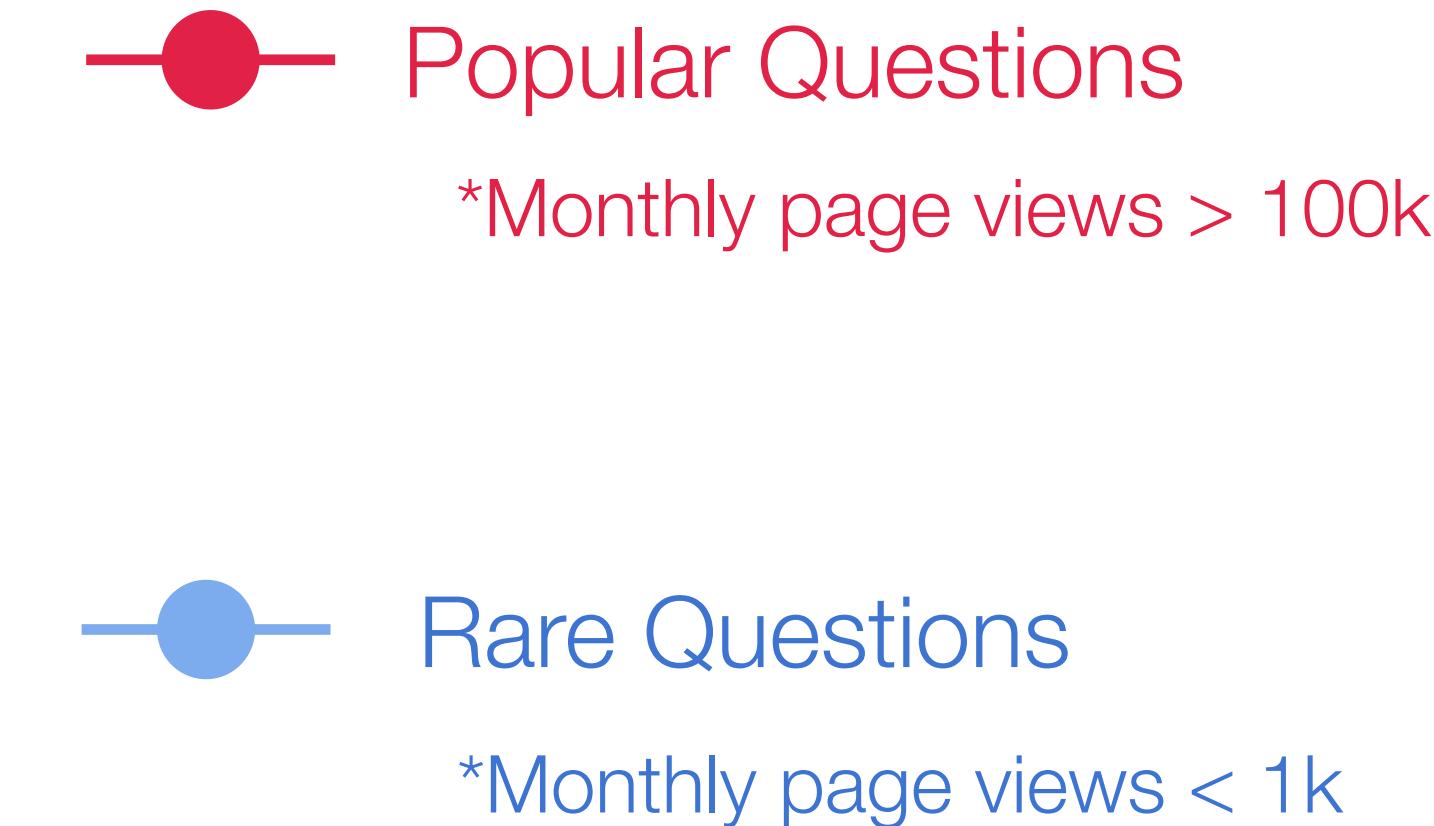
0

1

10

100

Size of LMs (Billion Parameters)



Popular Questions

*Monthly page views > 100k

Rare Questions

*Monthly page views < 1k

Effect of Scaling on Hallucination

Accuracy on
PopQA

100

75

50

25

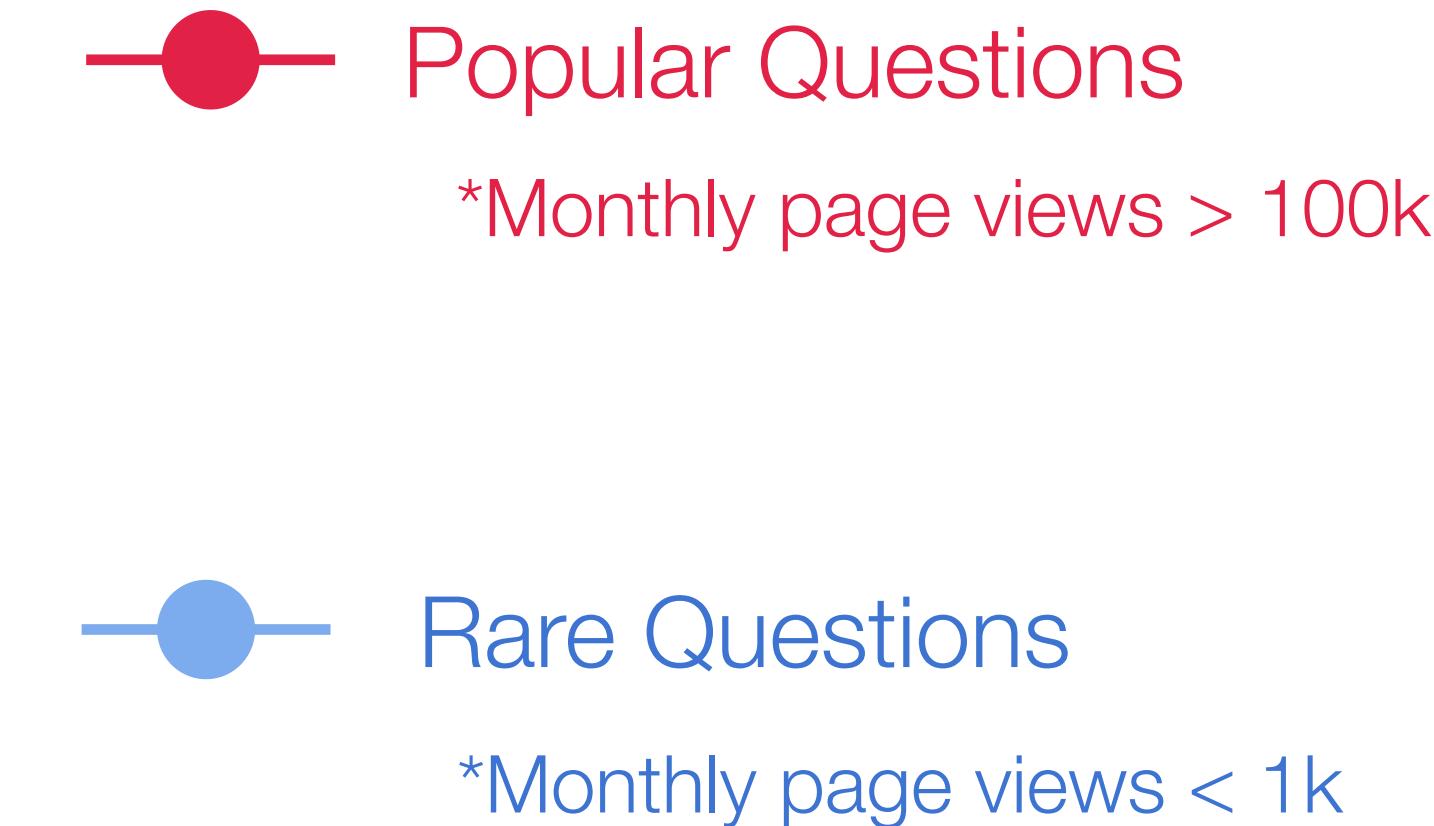
0

1

10

100

Size of LMs (Billion Parameters)



Effect of Scaling on Hallucination

Accuracy on
PopQA

100

75

50

25

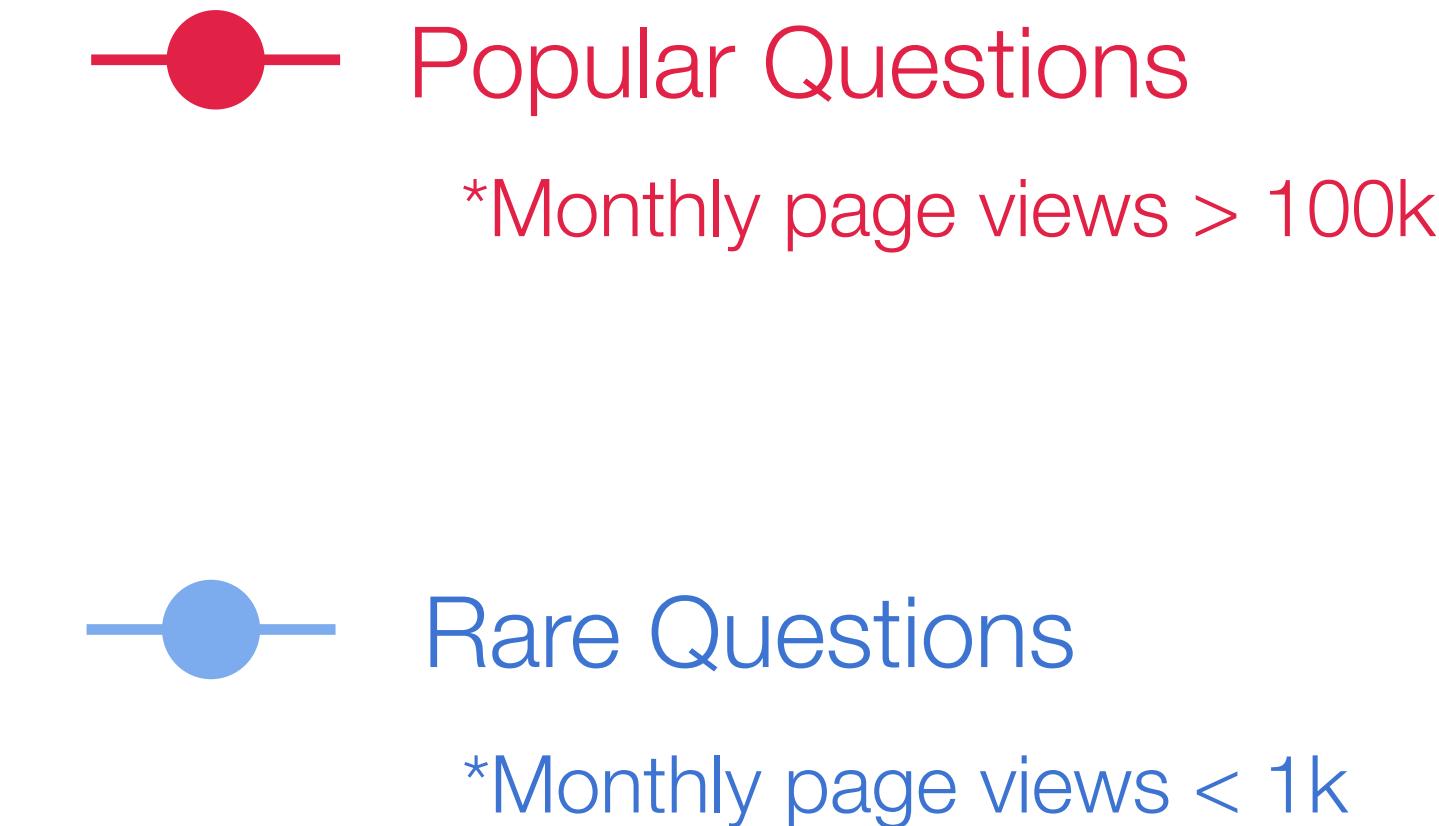
0

1

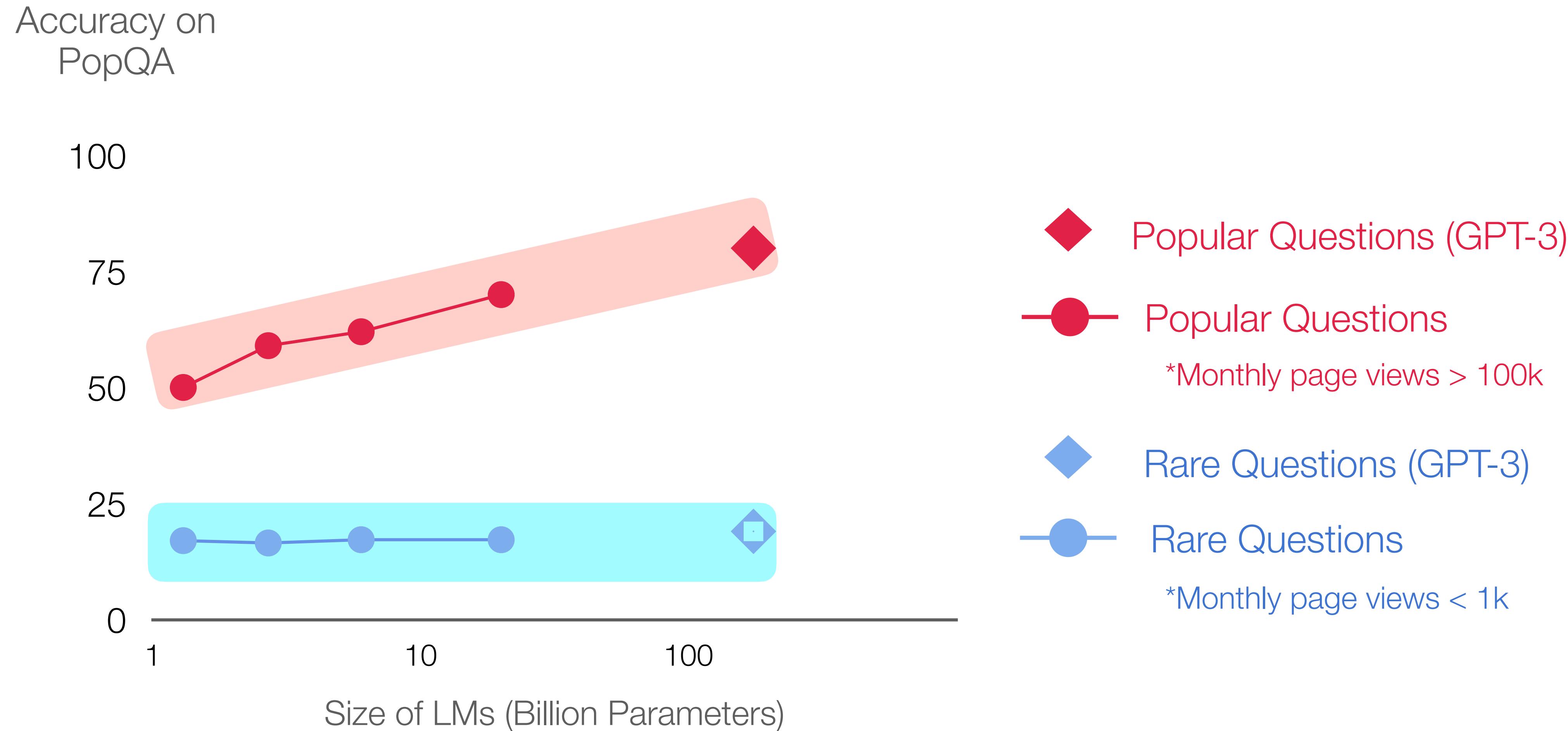
10

100

Size of LMs (Billion Parameters)



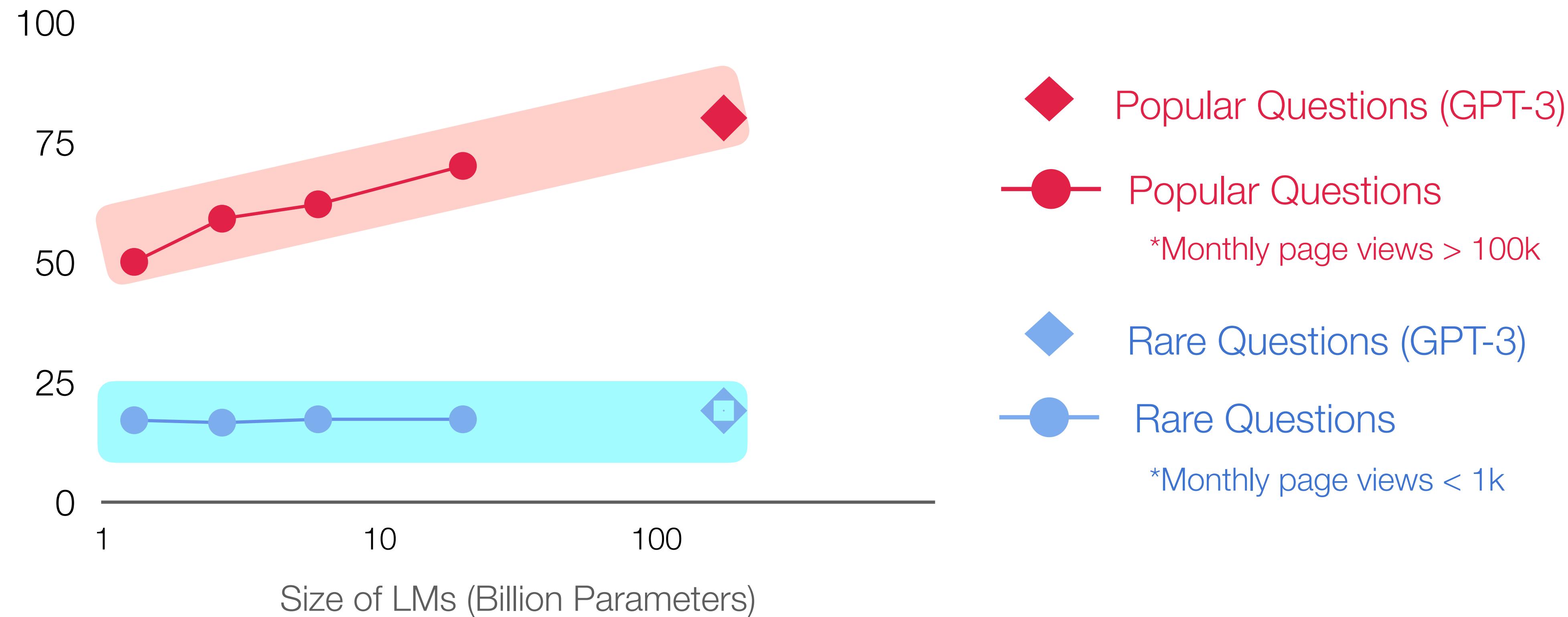
Effect of Scaling on Hallucination



Effect of Scaling on Hallucination

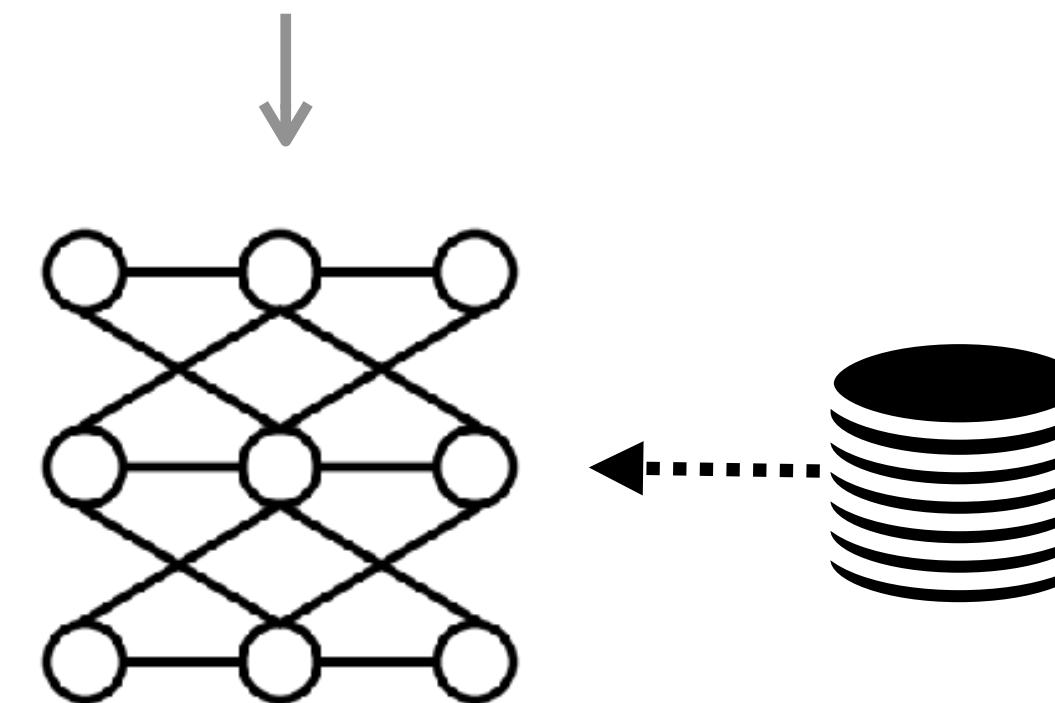
Accuracy on
PopQA

LMs struggle on less popular questions (= **long-tail knowledge**)



How Can We Overcome Long-Tail Hallucination?

Who is Akari Asai?



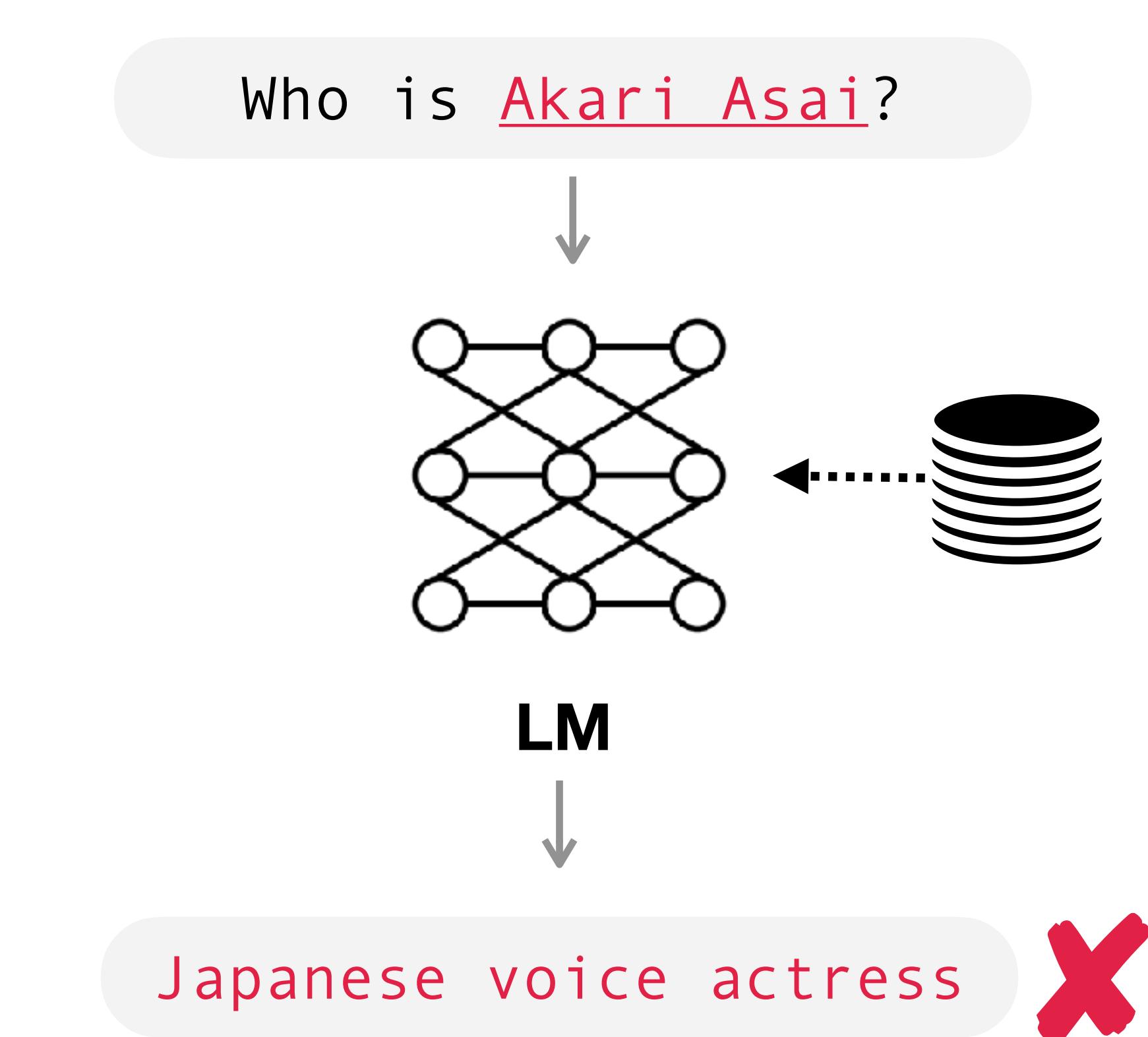
LM

Japanese voice actress



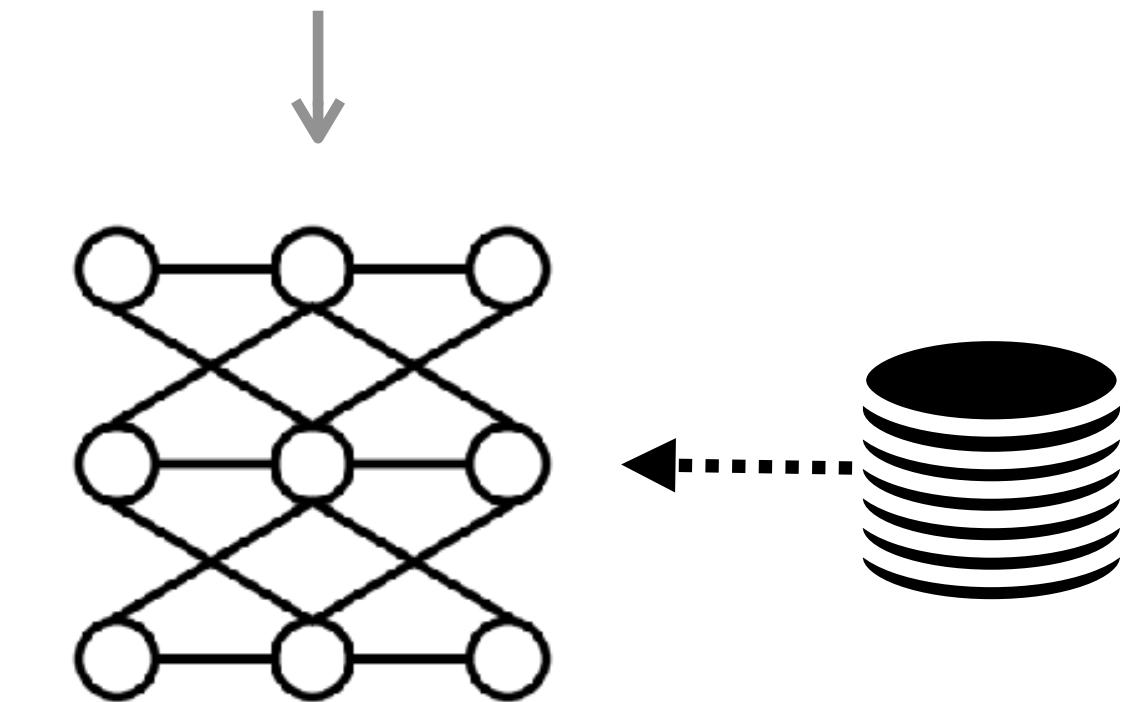
Akari Asai is a Ph.D. student in the Paul G. Allen School of Computer Science at University of Washington

How Can We Overcome Long-Tail Hallucination?



How Can We Overcome Long-Tail Hallucination?

Who is Akari Asai?



LM

Japanese voice actress

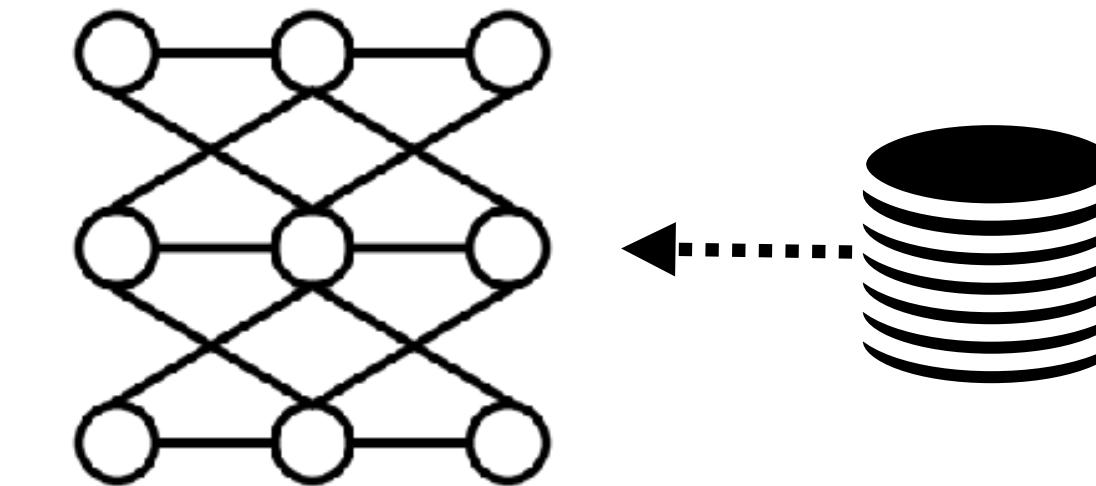


How Can We Overcome Long-Tail Hallucination?

Who is Akari Asai?



Who is Akari Asai?



LM

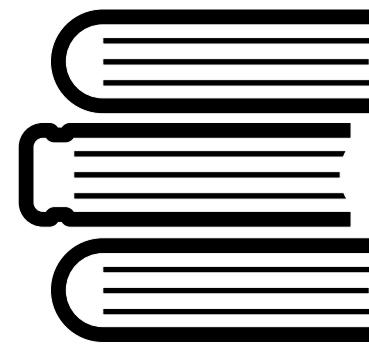


Japanese voice actress



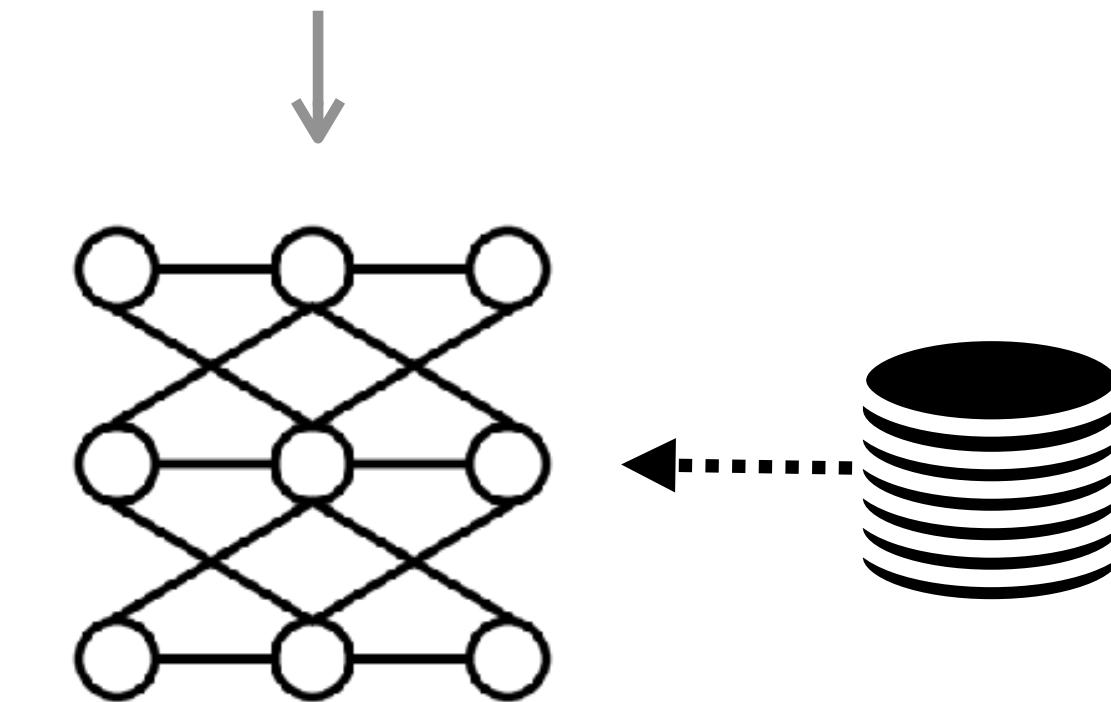
How Can We Overcome Long-Tail Hallucination?

Who is Akari Asai?



Akari Asai is a Ph.D. student in the Paul G. Allen School of Computer Science at University of Washington

Who is Akari Asai?



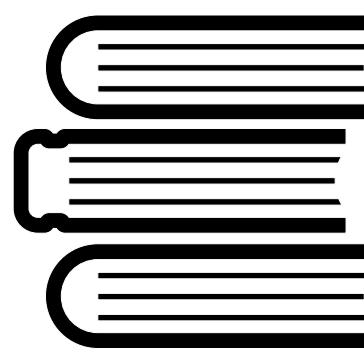
LM

Japanese voice actress

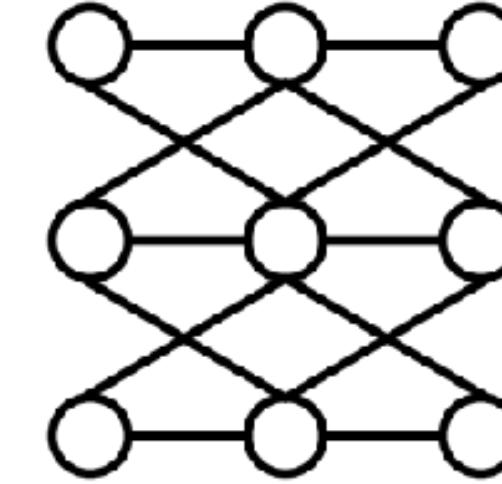


How Can We Overcome Long-Tail Hallucination?

Who is Akari Asai?



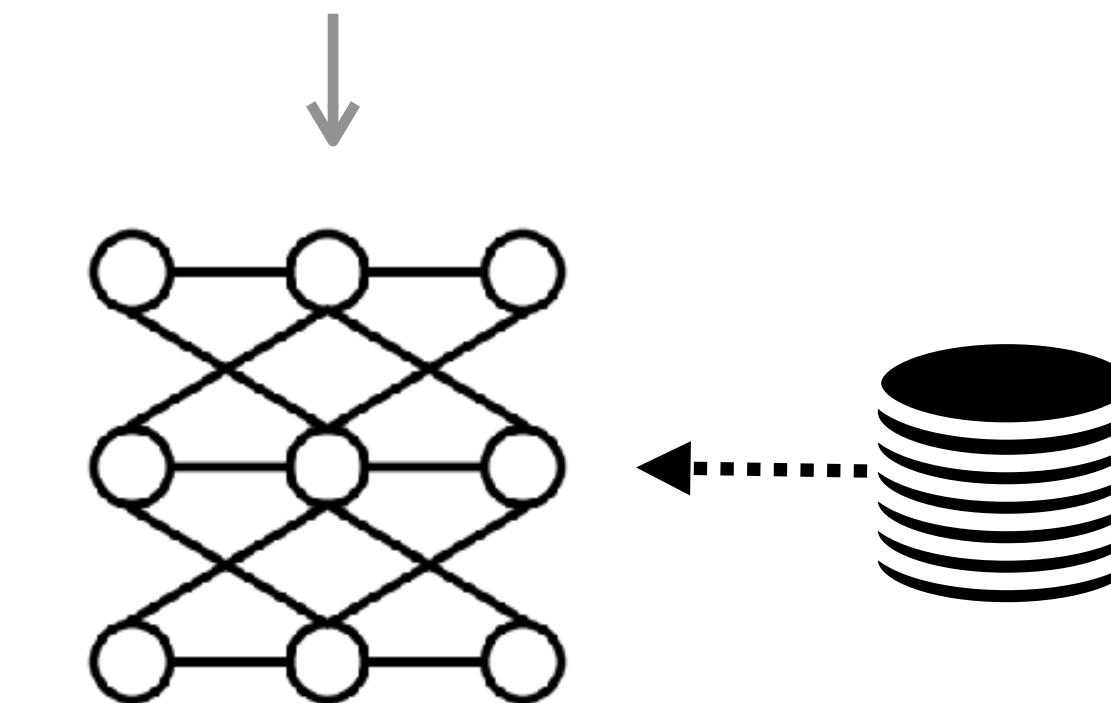
Akari Asai is a Ph.D. student in the Paul G. Allen School of Computer Science at University of Washington



LM

Ph.D. student at UW

Who is Akari Asai?

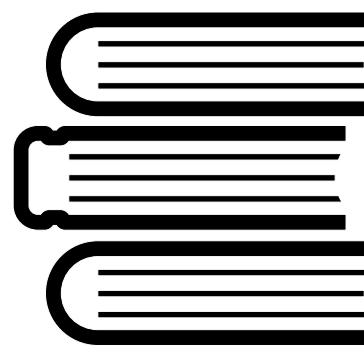


Japanese voice actress

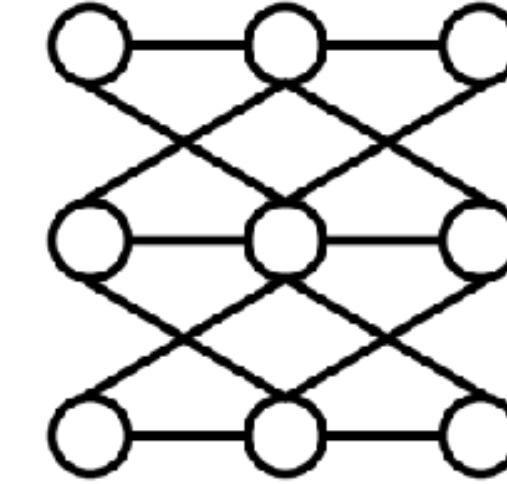


How Can We Overcome Long-Tail Hallucination?

Who is Akari Asai?



Akari Asai is a Ph.D. student in the Paul G. Allen School of Computer Science at University of Washington

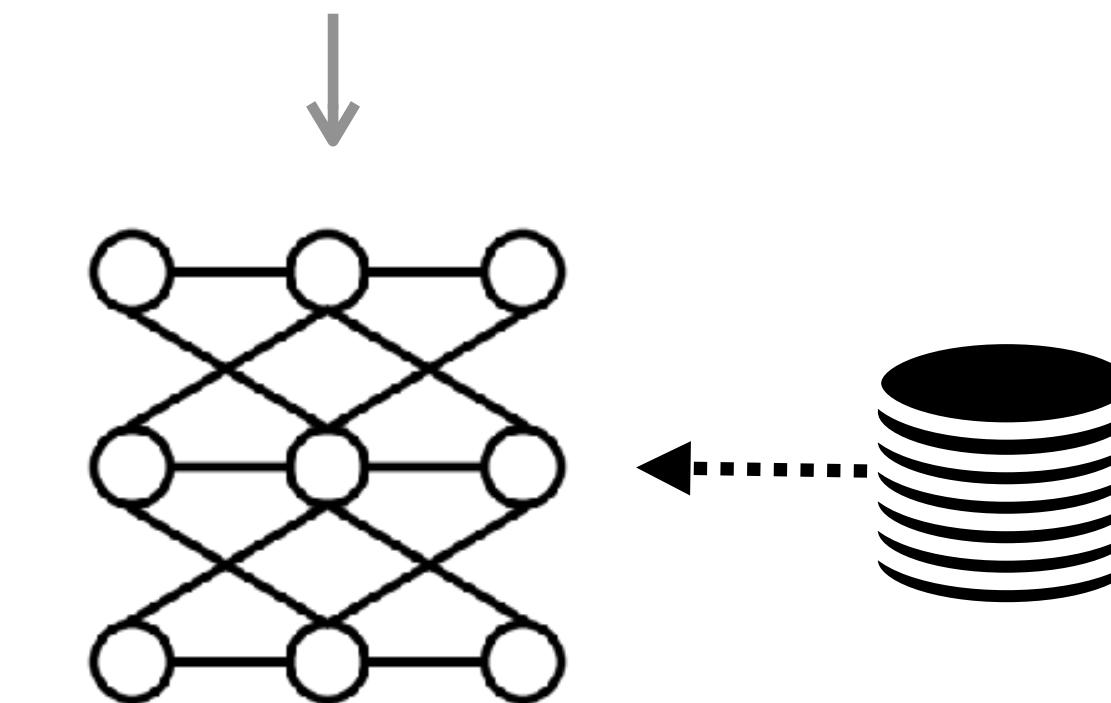


LM

Ph.D. student at UW



Who is Akari Asai?

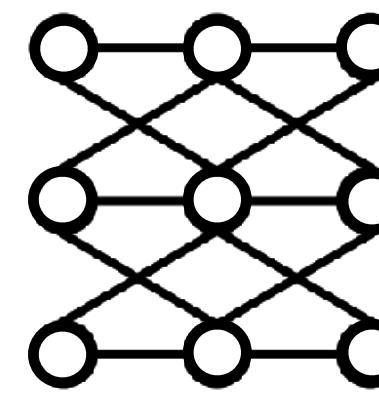


Japanese voice actress



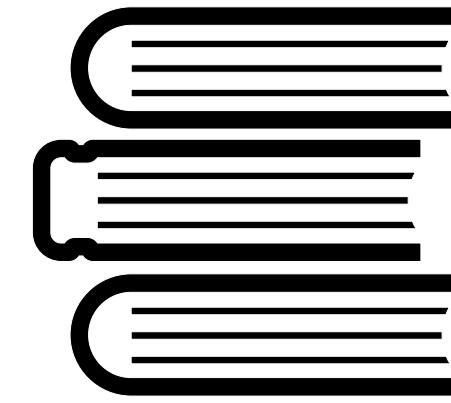
How Retrieval-Augmented LMs Work

How Retrieval-Augmented LMs Work



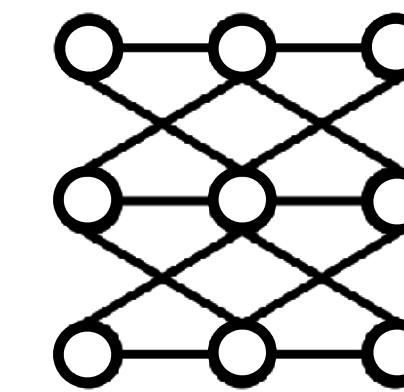
LM

How Retrieval-Augmented LMs Work



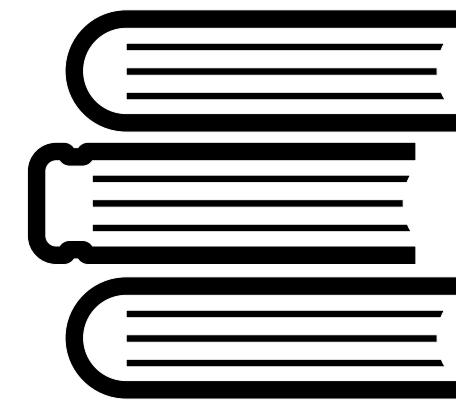
Datastore

Collections of a large number of documents



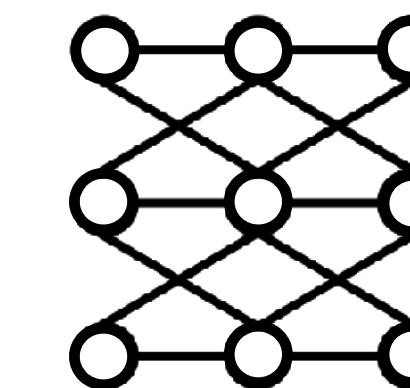
LM

How Retrieval-Augmented LMs Work



Datastore

Collections of a large number of documents



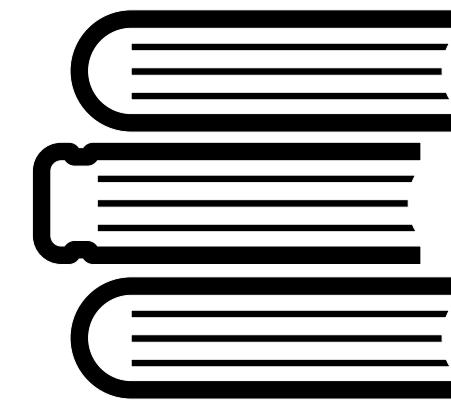
LM

GPT-4o is a pre-trained transformer developed by OpenAI.

Transformers is a series of science fiction action films based on the Transformers franchise.

GPT4o was released by OpenAI in May 2024.

How Retrieval-Augmented LMs Work



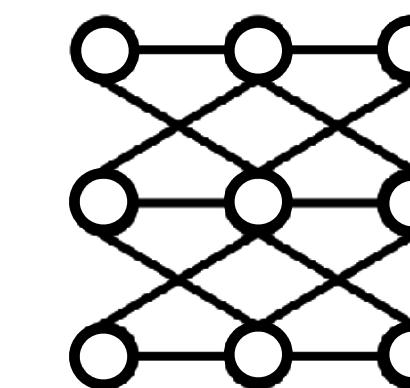
Datastore

Collections of a large number of documents



Retriever

Retrieve top k documents in datastore



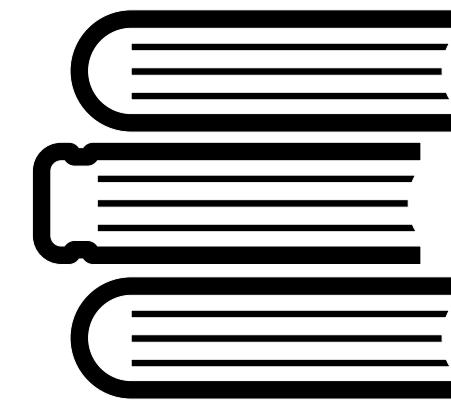
LM

GPT-4o is a pre-trained transformer developed by OpenAI.

Transformers is a series of science fiction action films based on the Transformers franchise.

GPT4o was released by OpenAI in May 2024.

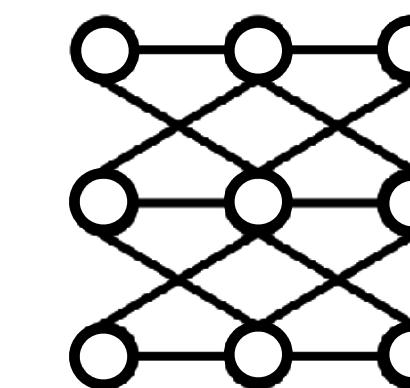
How Retrieval-Augmented LMs Work



Datastore



Retriever



LM

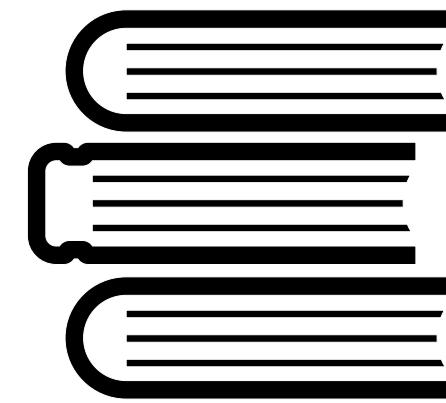
GPT-4o is a pre-trained transformer developed by OpenAI.

Transformers is a series of science fiction action films based on the Transformers franchise.

GPT4o was released by OpenAI in May 2024.

How Retrieval-Augmented LMs Work

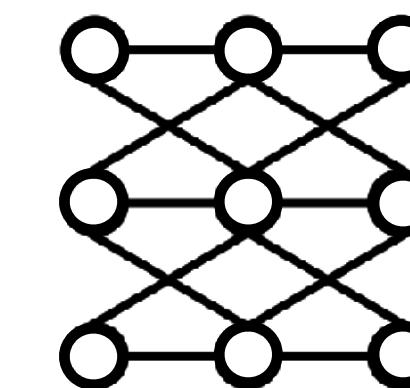
x : Which company developed GPT4o?



Datastore



Retriever



LM

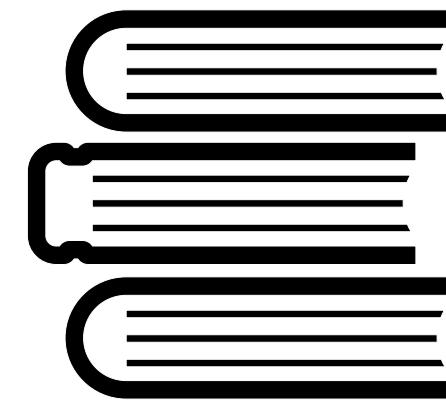
GPT-4o is a pre-trained transformer developed by OpenAI.

Transformers is a series of science fiction action films based on the Transformers franchise.

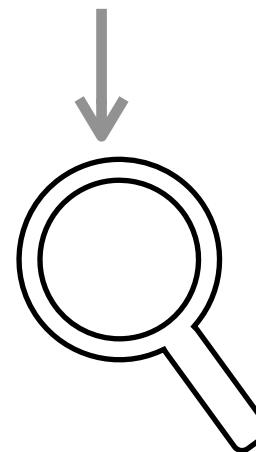
GPT4o was released by OpenAI in May 2024.

How Retrieval-Augmented LMs Work

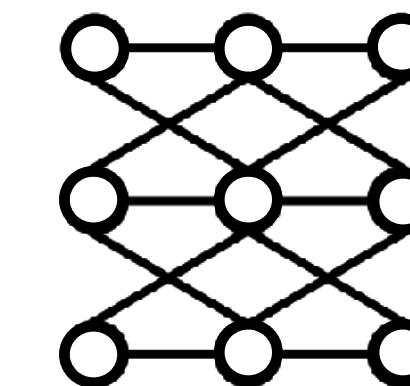
x : Which company developed GPT4o?



Datastore



Retriever



LM

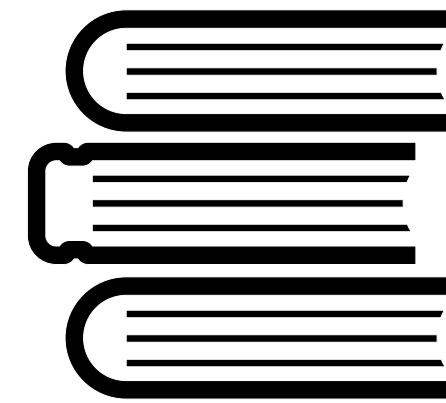
GPT-4o is a pre-trained transformer developed by OpenAI.

Transformers is a series of science fiction action films based on the Transformers franchise.

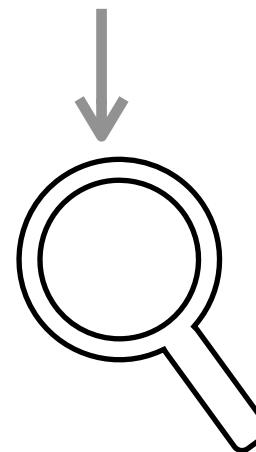
GPT4o was released by OpenAI in May 2024.

How Retrieval-Augmented LMs Work

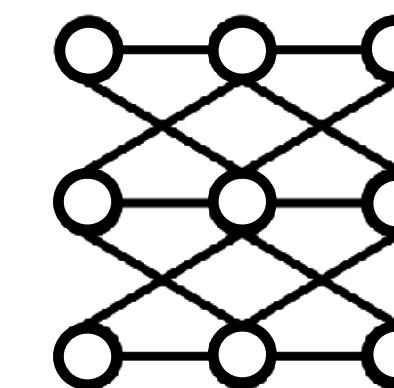
x : Which company developed GPT4o?



Datastore



Retriever



LM

$\text{Sim}(\cdot | x)$

GPT-4o is a pre-trained transformer developed by OpenAI.

0.9

Transformers is a series of science fiction action films based on the Transformers franchise.

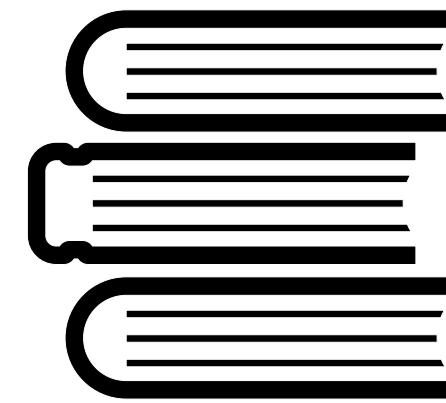
0.1

GPT4o was released by OpenAI in May 2024.

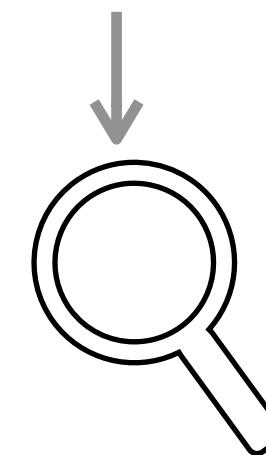
0.8

How Retrieval-Augmented LMs Work

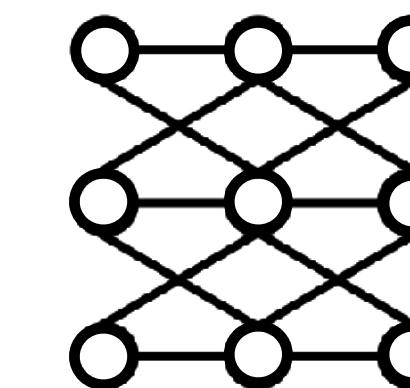
x : Which company developed GPT4o?



Datastore



Retriever



LM

$$D \in \text{Top}_k \text{Sim}(\cdot | x)$$

GPT-4o is a pre-trained transformer developed by OpenAI.

Transformers is a series of science fiction action films based on the Transformers franchise.

GPT4o was released by OpenAI in May 2024.

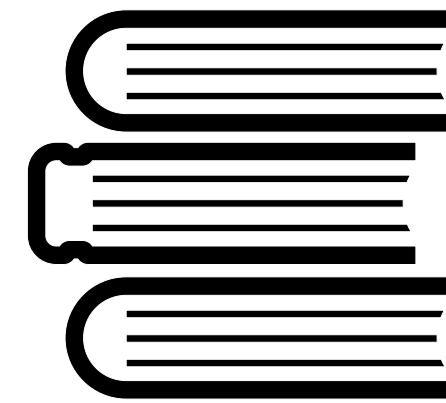
0.9

0.1

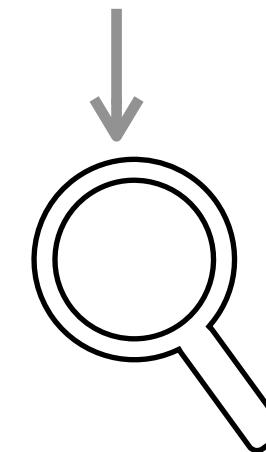
0.8

How Retrieval-Augmented LMs Work

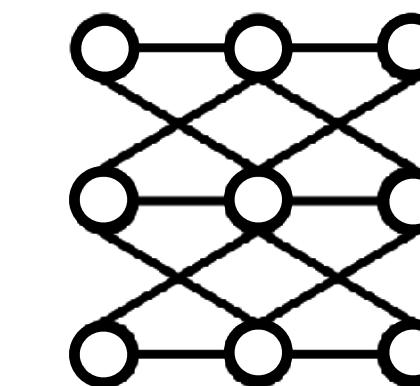
x : Which company developed GPT4o?



Datastore



Retriever



LM

$$D \in \text{Top}_k \text{Sim}(\cdot | x)$$

GPT-4o is a pre-trained transformer developed by OpenAI.

0.9

Transformers is a series of science fiction action films based on the Transformers franchise.

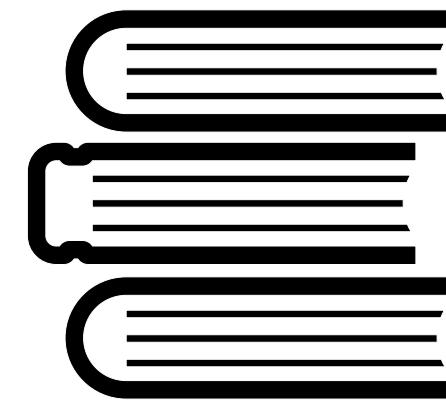
0.1

GPT4o was released by OpenAI in May 2024.

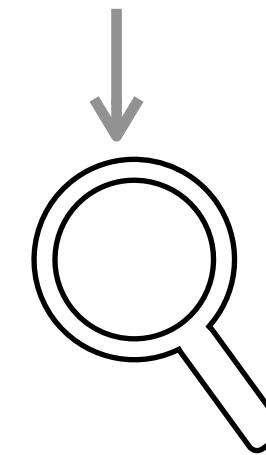
0.8

How Retrieval-Augmented LMs Work

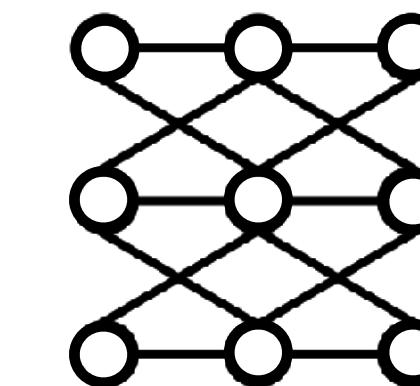
x : Which company developed GPT4o?



Datastore



Retriever



LM

$$D \in \text{Top}_k \text{Sim}(\cdot | x)$$

GPT-4o is a pre-trained transformer developed by OpenAI.

0.9

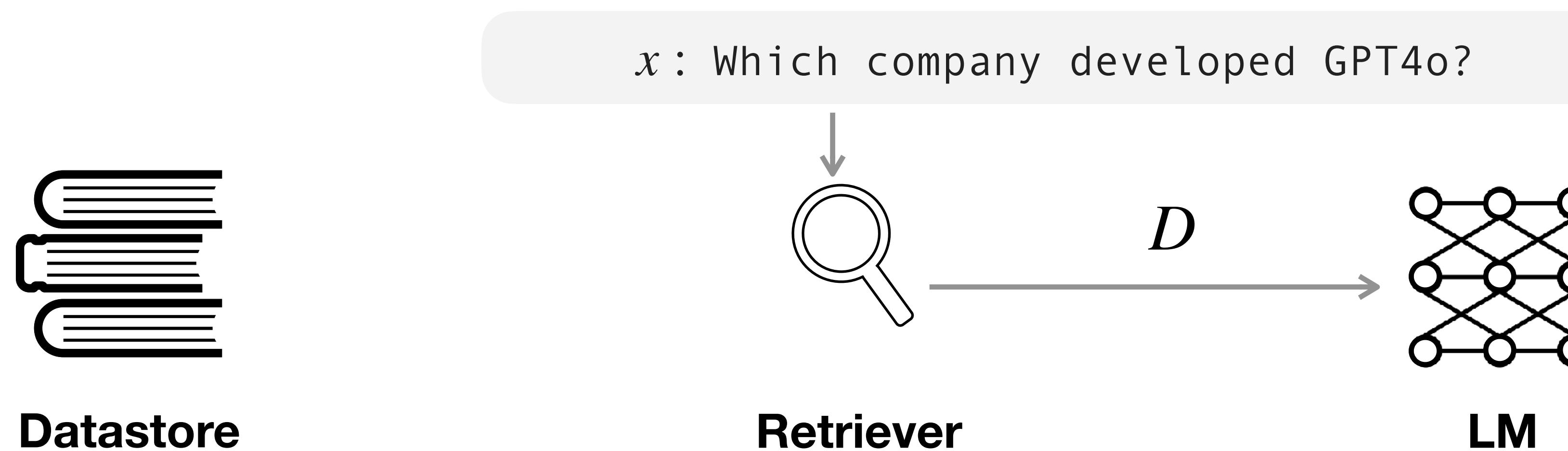
Transformers is a series of science fiction action films based on the Transformers franchise.

0.1

GPT4o was released by OpenAI in May 2024.

0.8

How Retrieval-Augmented LMs Work



$$D \in \text{Top}_k \text{Sim}(\cdot | x)$$

Transformers is a series of science fiction action films based on the Transformers franchise.

GPT-4o is a pre-trained transformer developed by OpenAI.

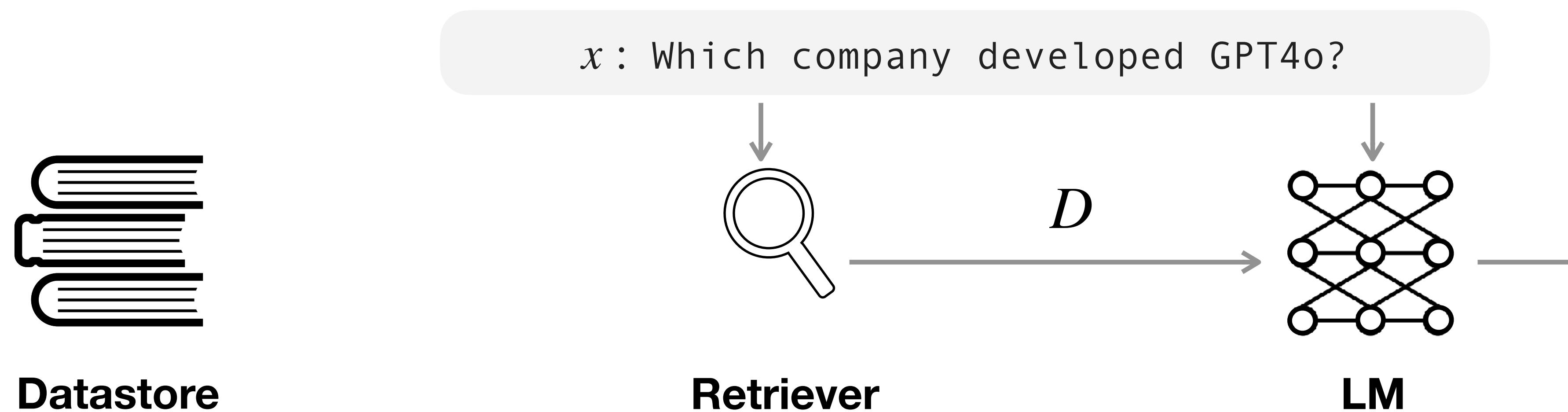
0.9

0.1

GPT4o was released by OpenAI in May 2024.

0.8

How Retrieval-Augmented LMs Work



$$D \in \text{Top}_k \text{Sim}(\cdot | x)$$

Transformers is a series of science fiction action films based on the Transformers franchise.

GPT-4o is a pre-trained transformer developed by OpenAI.

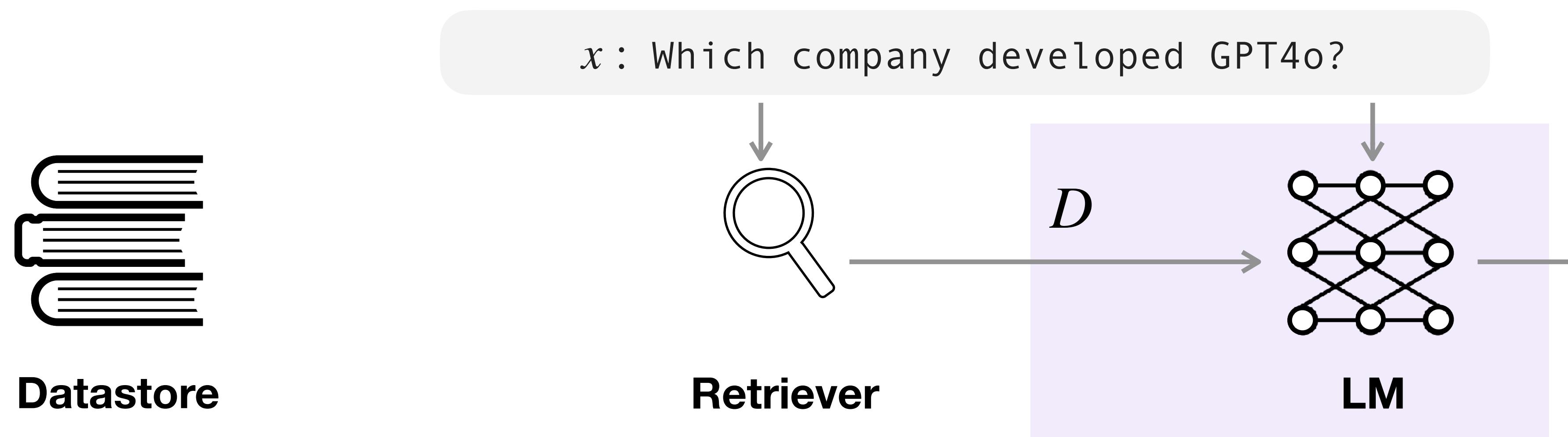
0.1

0.9

GPT4o was released by OpenAI in May 2024.

0.8

How Retrieval-Augmented LMs Work



$$D \in \text{Top}_k \text{Sim}(\cdot | x)$$

Transformers is a series of science fiction action films based on the Transformers franchise.

GPT-4o is a pre-trained transformer developed by OpenAI.

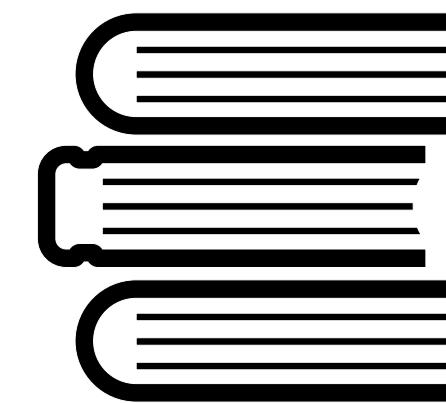
0.1

0.9

GPT4o was released by OpenAI in May 2024.

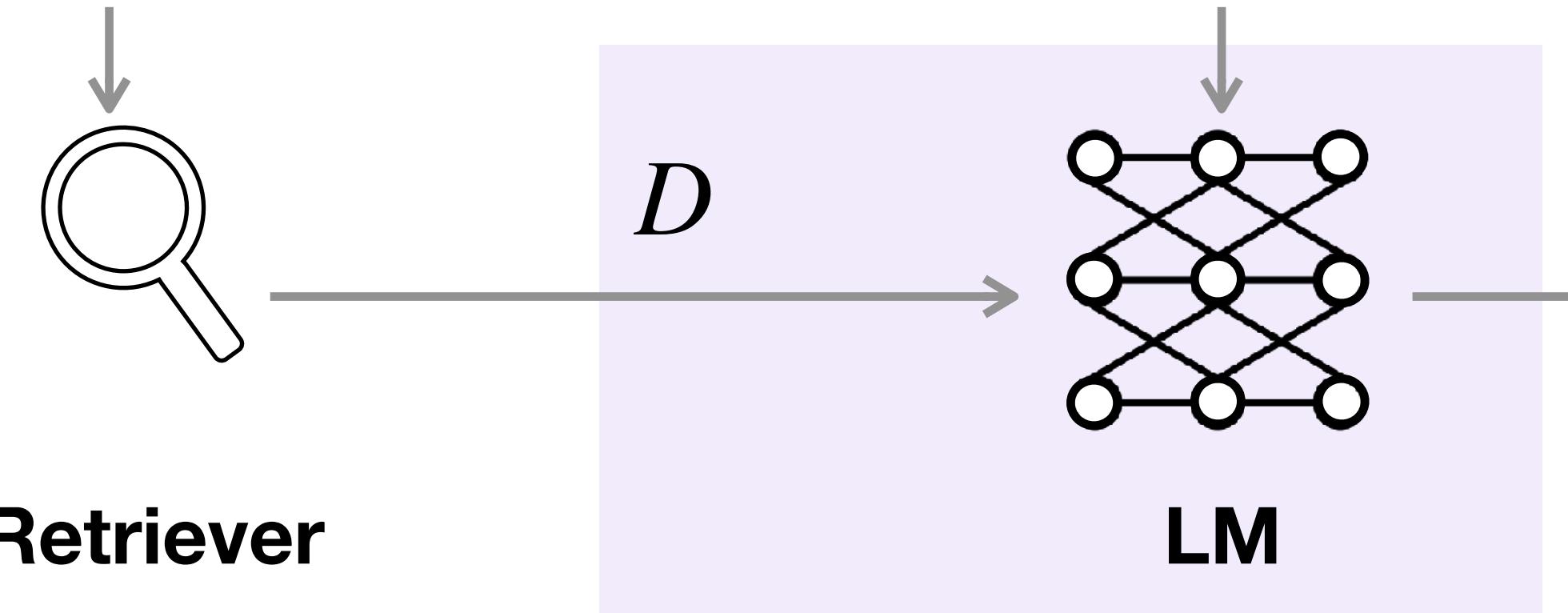
0.8

In-context Retrieval-Augmented Generation (RAG)



Datastore

x : Which company developed GPT4o?



$$D \in \text{Top}_k \text{Sim}(\cdot | x)$$

Transformers is a series of science fiction action films based on the Transformers franchise.

GPT-4o is a pre-trained transformer developed by OpenAI.

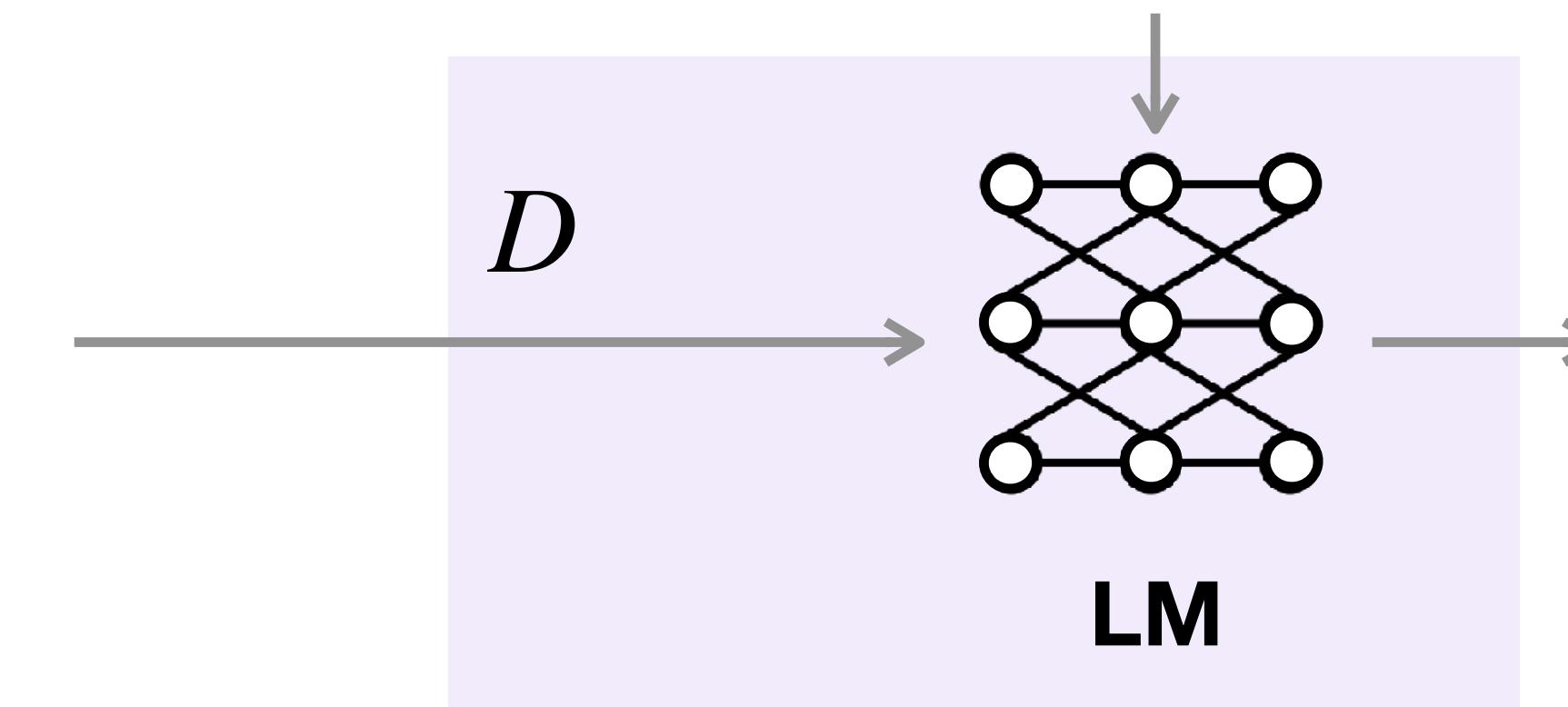
0.9

0.1

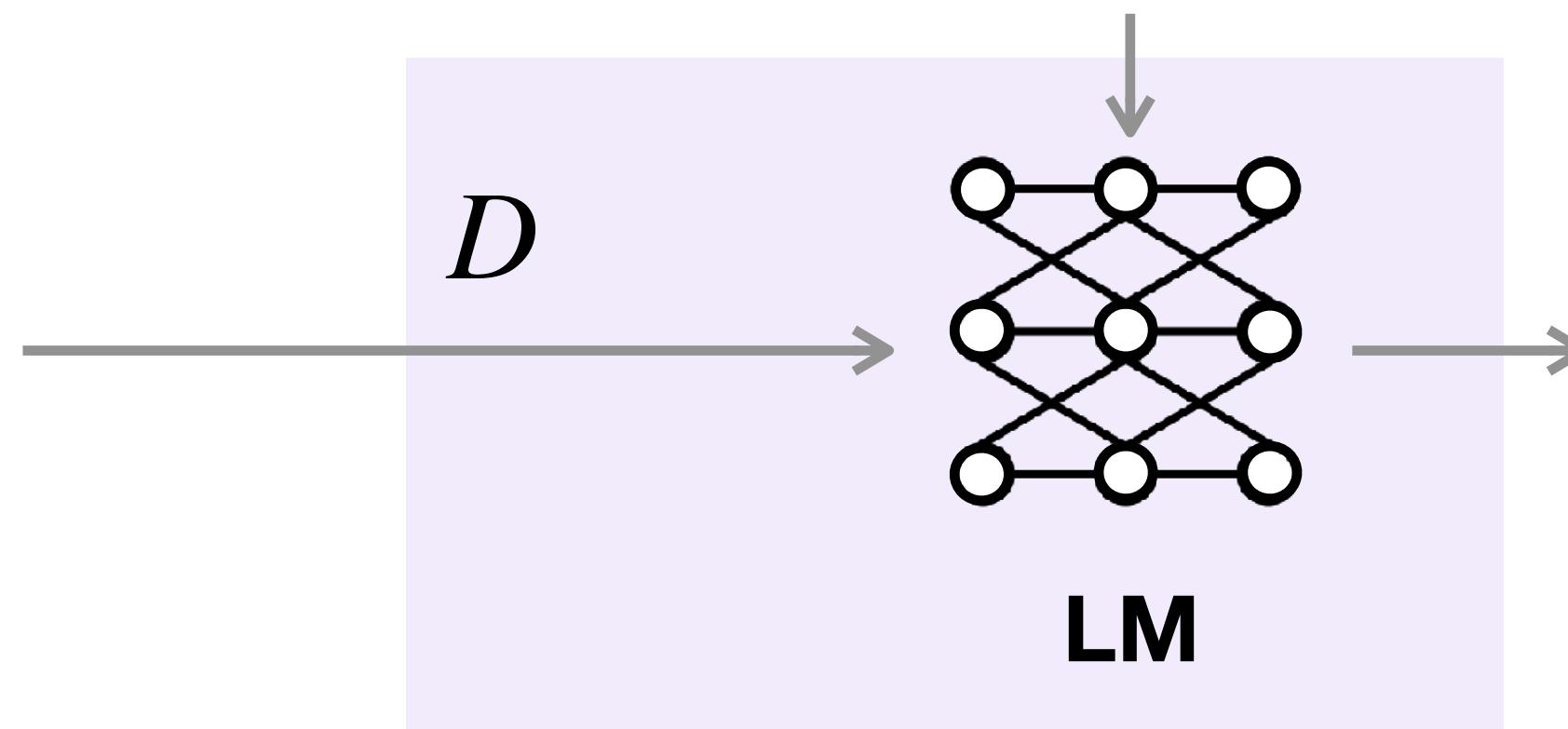
GPT4o was released by OpenAI in May 2024.

0.8

In-context Retrieval-Augmented Generation (RAG)

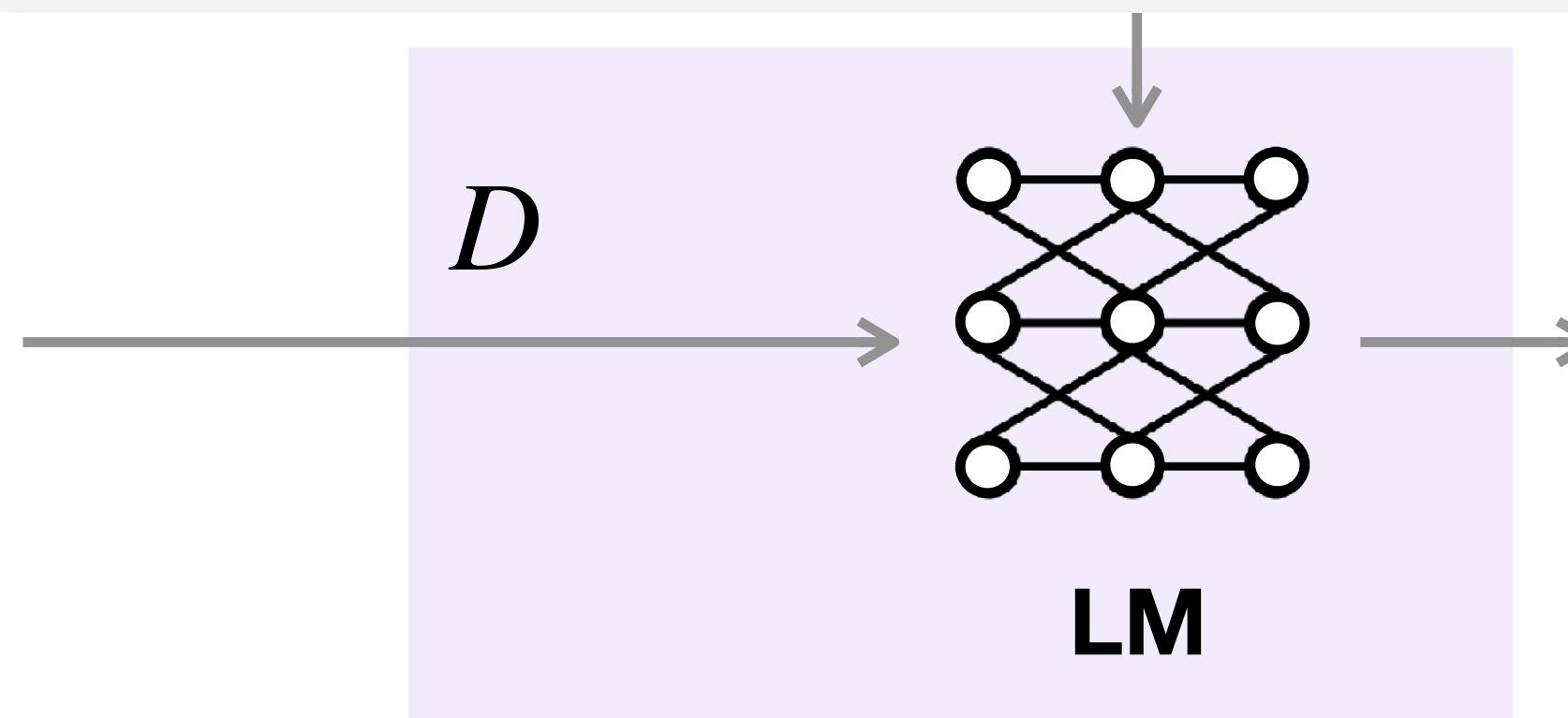


In-context Retrieval-Augmented Generation (RAG)



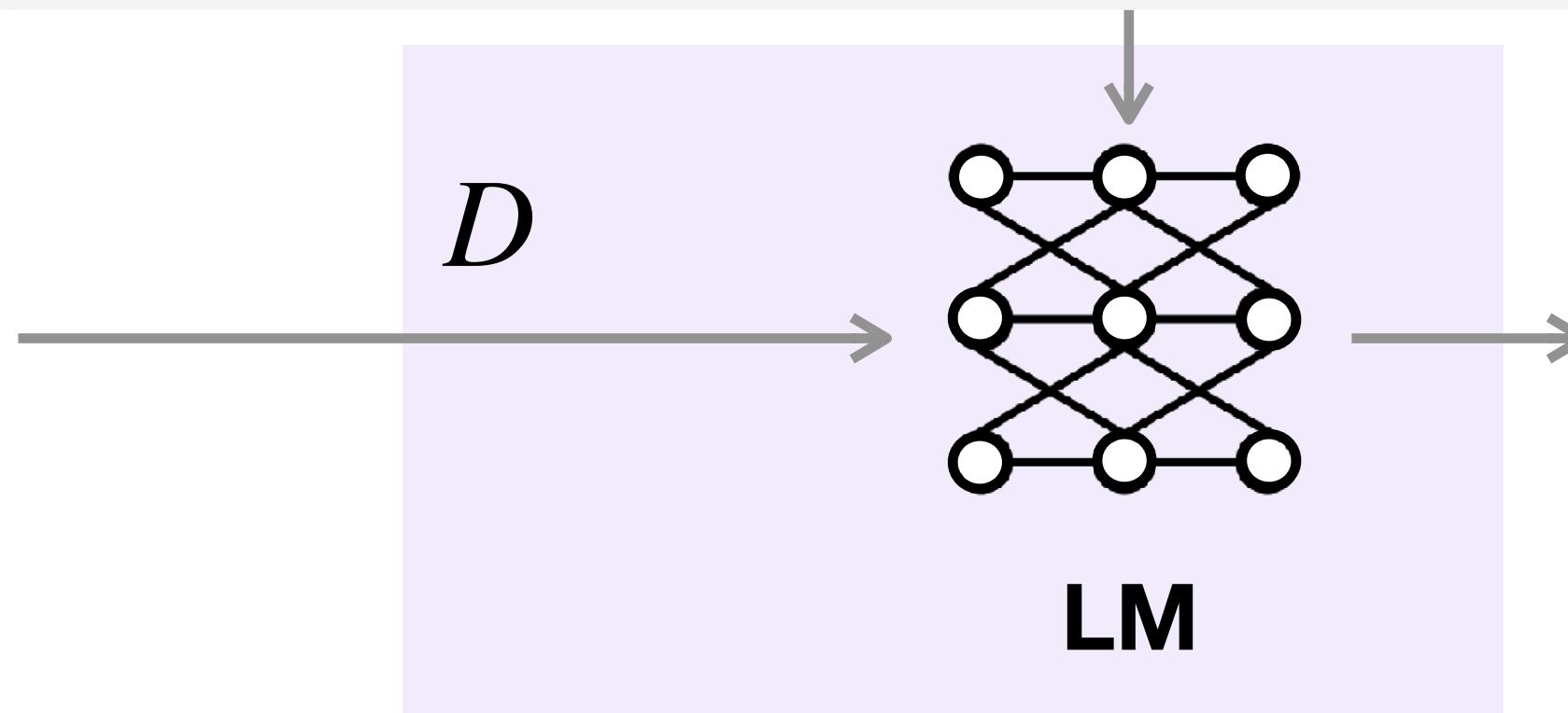
In-context Retrieval-Augmented Generation (RAG)

x : Which company developed GPT4o?



In-context Retrieval-Augmented Generation (RAG)

x : Which company developed GPT4o?

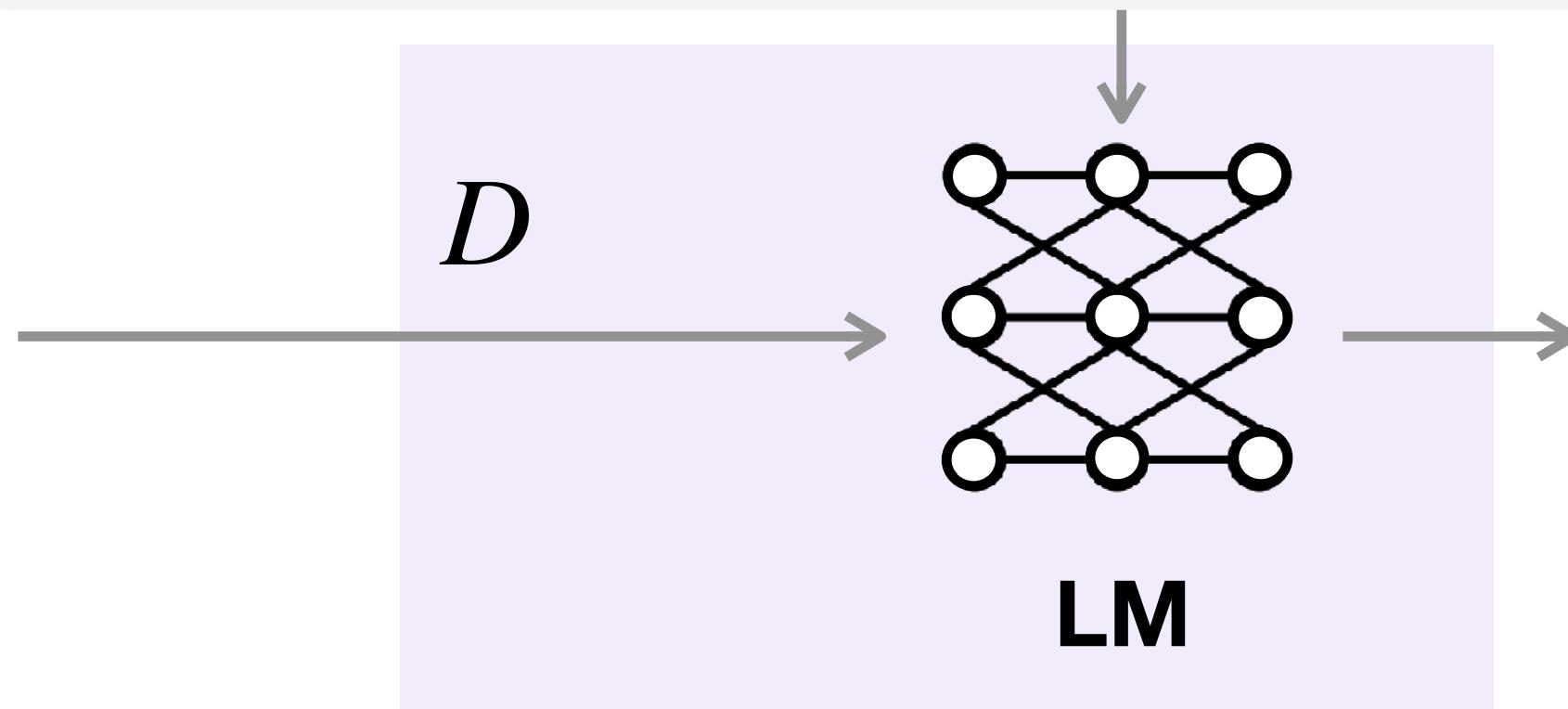


GPT-4o is a pre-trained transformer developed by OpenAI.

GPT4o was released by OpenAI in May 2024.

In-context Retrieval-Augmented Generation (RAG)

x : Which company developed GPT4o?



Question: Which company developed GPT4o?

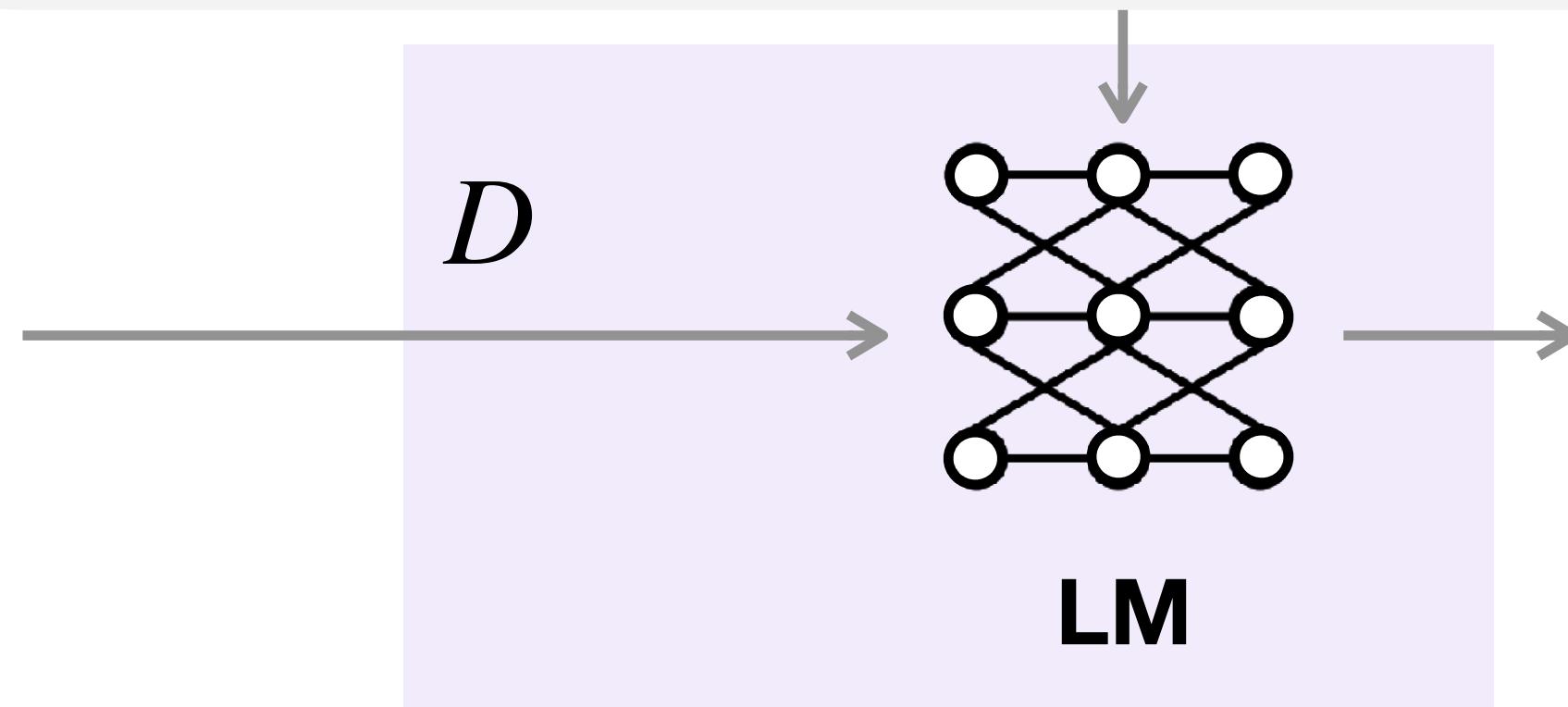
References:

GPT-4o is a pre-trained transformer developed by OpenAI.

GPT4o was released by OpenAI in May 2024.

In-context Retrieval-Augmented Generation (RAG)

x : Which company developed GPT4o?



Question: Which company developed GPT4o?

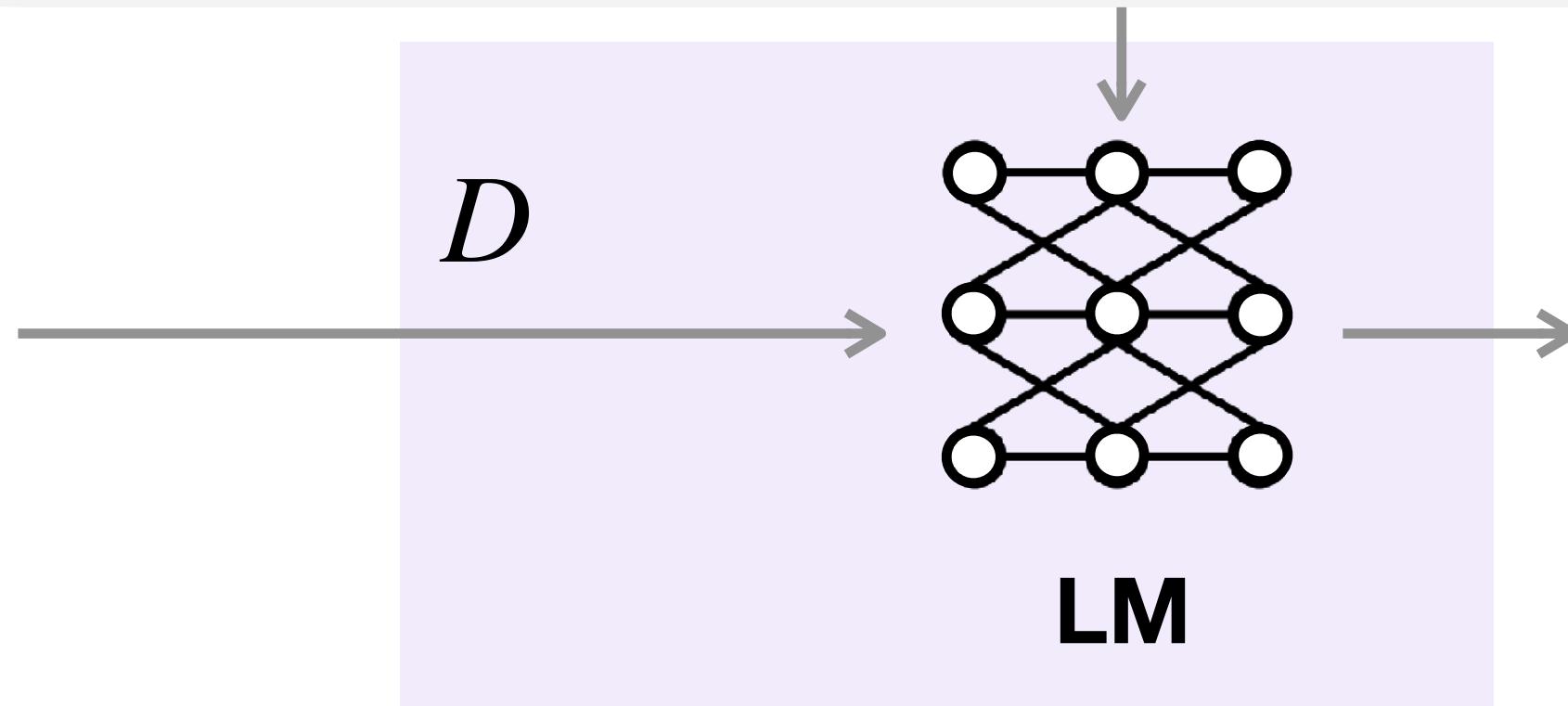
References:

GPT-4o is a pre-trained transformer developed by OpenAI.

GPT4o was released by OpenAI in May 2024.

In-context Retrieval-Augmented Generation (RAG)

x : Which company developed GPT4o?



Question: Which company developed GPT4o?

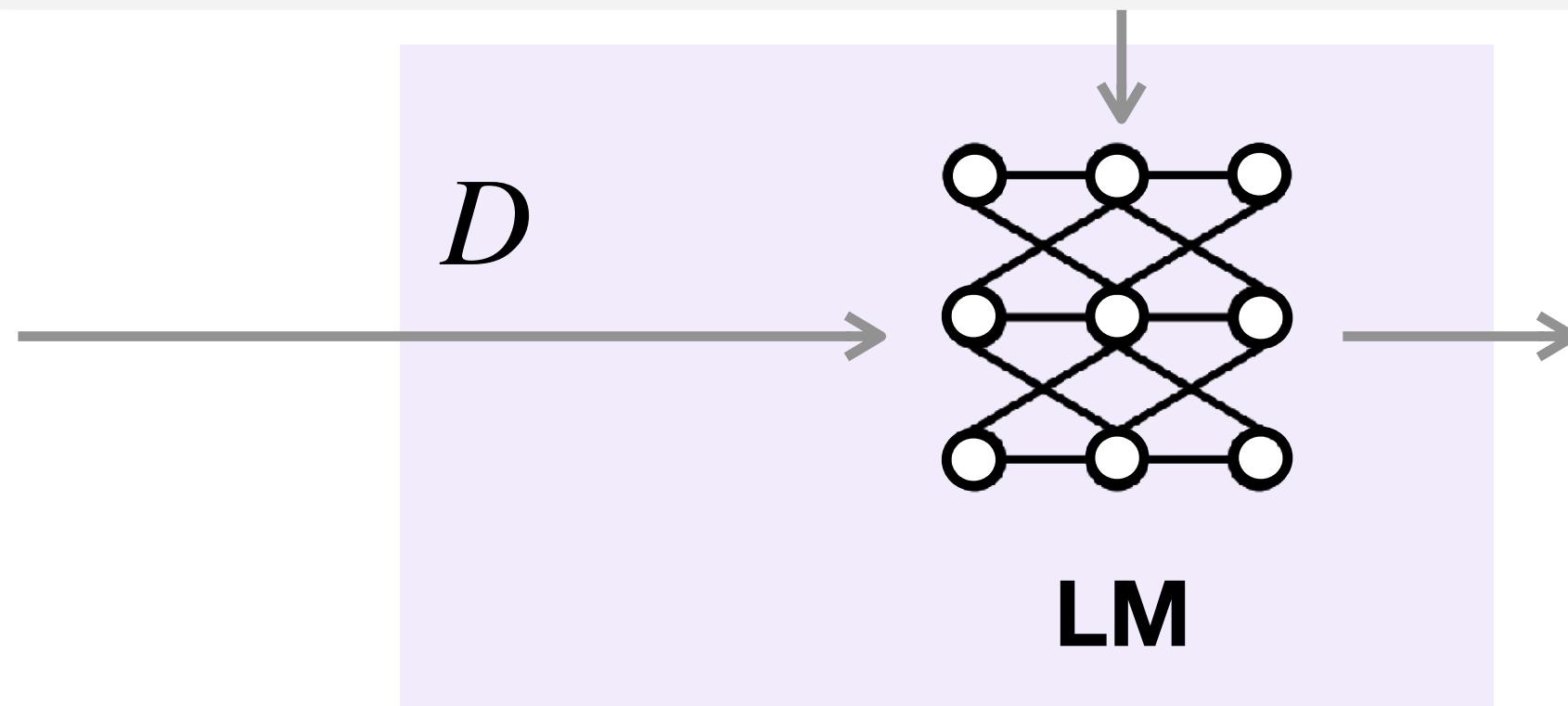
References:

GPT-4o is a pre-trained transformer developed by OpenAI.

GPT4o was released by OpenAI in May 2024.

In-context Retrieval-Augmented Generation (RAG)

x : Which company developed GPT4o?

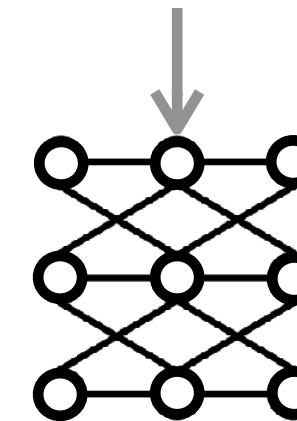


Question: Which company developed GPT4o?

References:

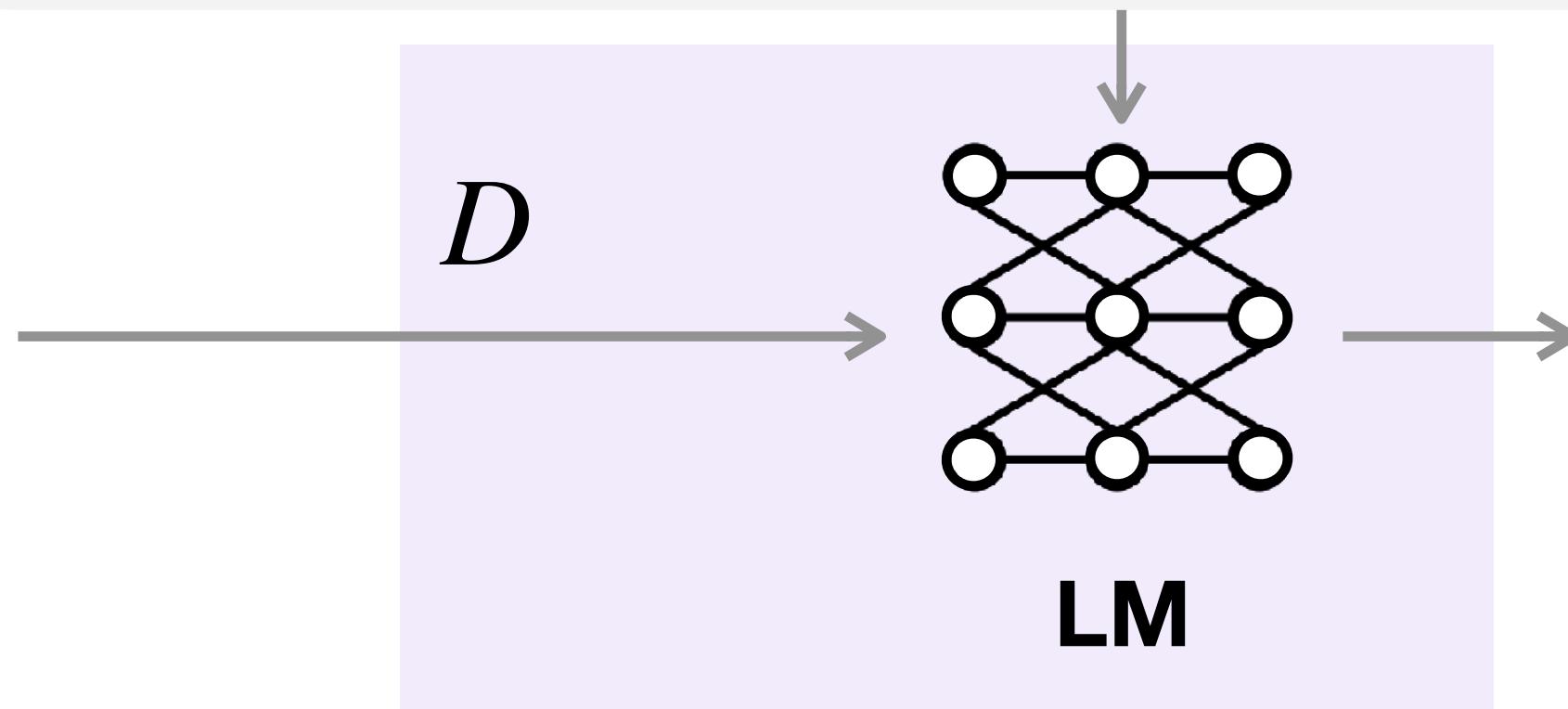
GPT-4o is a pre-trained transformer developed by OpenAI.

GPT4o was released by OpenAI in May 2024.



In-context Retrieval-Augmented Generation (RAG)

x : Which company developed GPT4o?

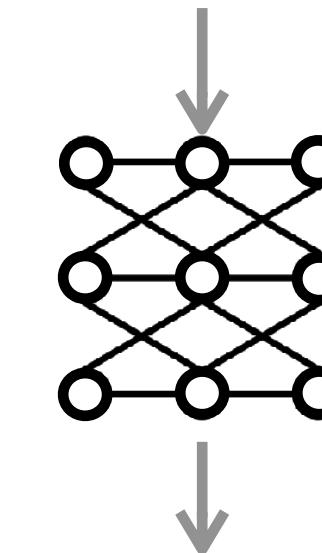


Question: Which company developed GPT4o?

References:

GPT-4o is a pre-trained transformer developed by OpenAI.

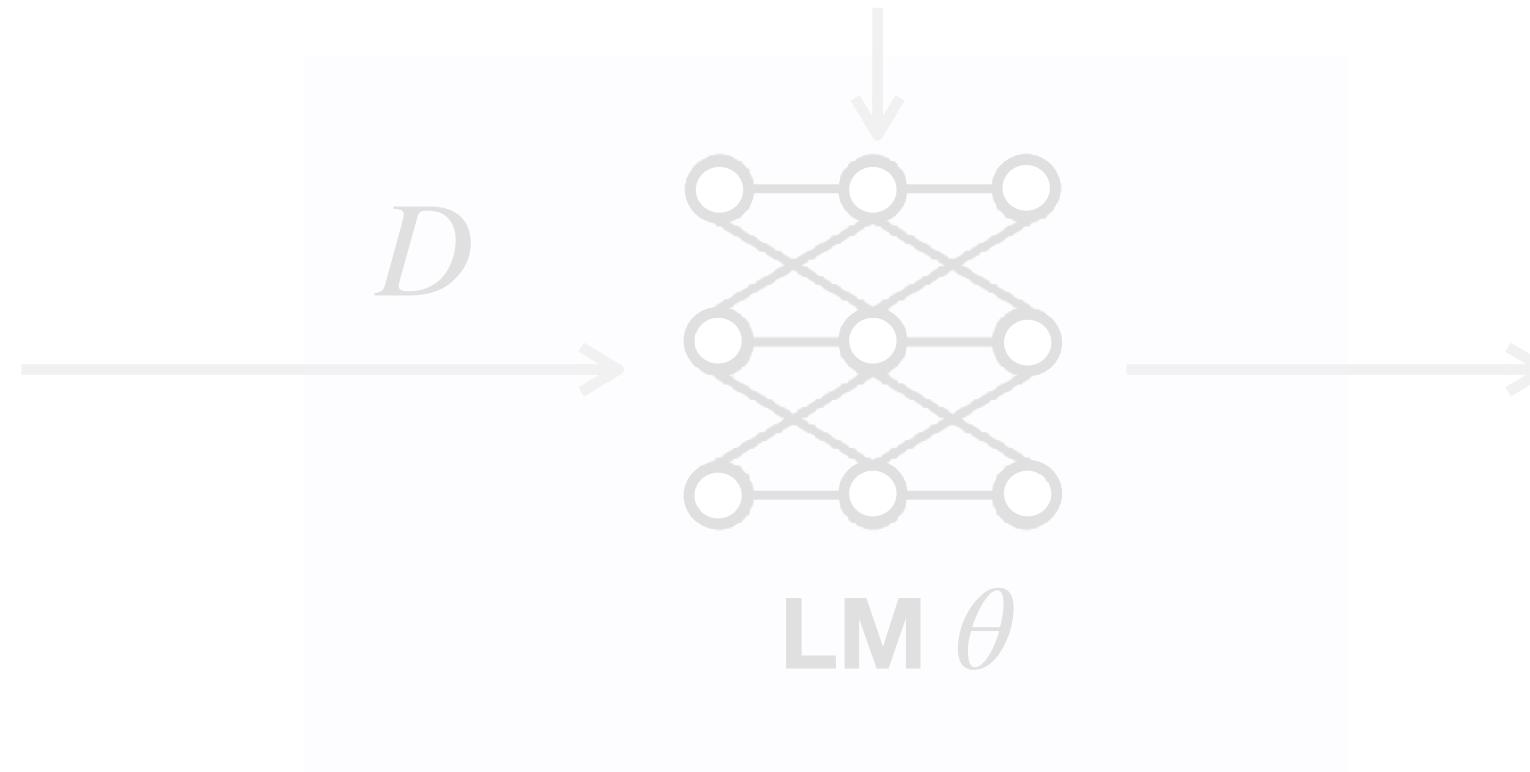
GPT4o was released by OpenAI in May 2024.



y : OpenAI

In-context Retrieval-Augmented Generation (RAG)

x : Which company developed GPT4o?



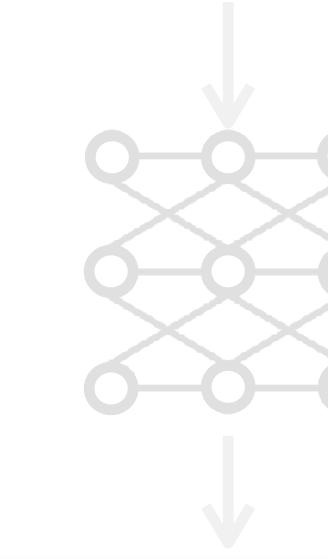
Question: Who developed GPT4o?

References:

D

GPT-4o is a pre-trained transformer developed by OpenAI.

GPT4o is released by OpenAI in May 2024.



In-context Retrieval-Augmented Generation (RAG)

x : Which company developed GPT4o?



Question: Who developed GPT4o?

References:

Previous work **fine-tuned LMs** for specific downstream tasks

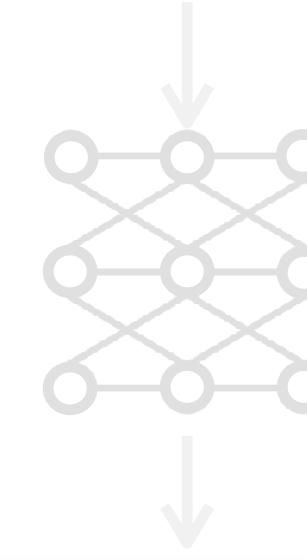
(Chen et al., 2017; Guu et al., 2020; Lewis et al., 2020)

LM θ

D

GPT-4o is a pre-trained transformer developed by OpenAI.

GPT4o is released by OpenAI in May 2024.



y : OpenAI

In-context Retrieval-Augmented Generation (RAG)

x : Which company developed GPT4o?



Question: Who developed GPT4o?

References:

Previous work **fine-tuned LMs** for specific downstream tasks

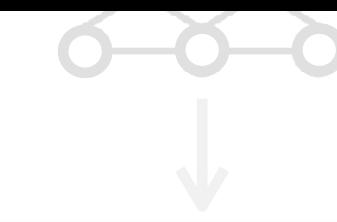
(Chen et al., 2017; Guu et al., 2020; Lewis et al., 2020)

LM θ

Our approach proposes **combining retrieval with off-the-shelf LMs**

GPT-4o is a pre-trained transformer
developed by OpenAI.

GPT4o is released by OpenAI in May 2024.



y : OpenAI

In-context Retrieval-Augmented Generation (RAG)

x : Which company developed GPT4o?



Question: Who developed GPT4o?

References:

Previous work **fine-tuned LMs** for specific downstream tasks

(Chen et al., 2017; Guu et al., 2020; Lewis et al., 2020)

LM θ

Our approach proposes **combining retrieval with off-the-shelf LMs**

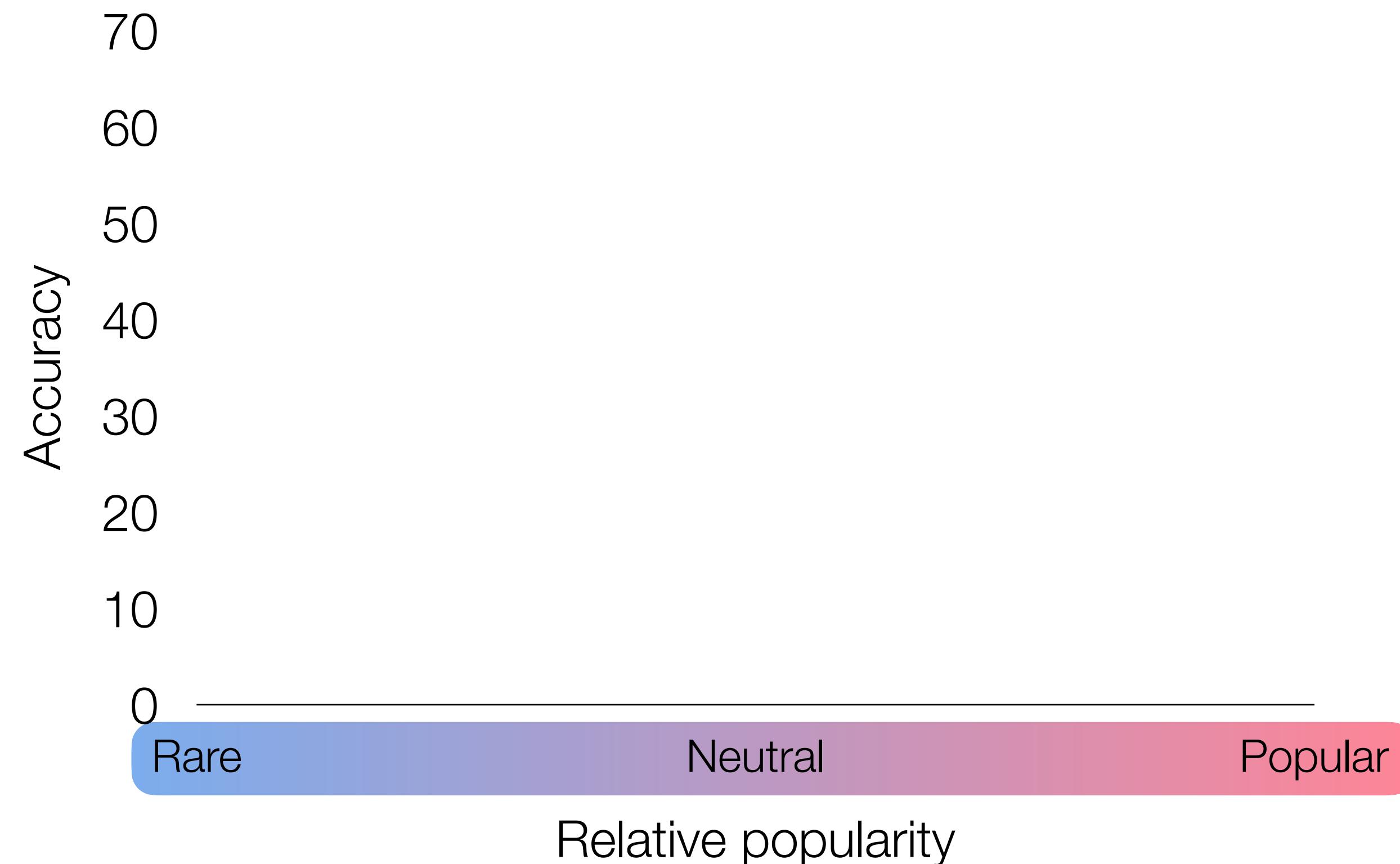
GPT-4o is a pre-trained transformer developed by OpenAI.

GPT4o is released by OpenAI in May 2024.

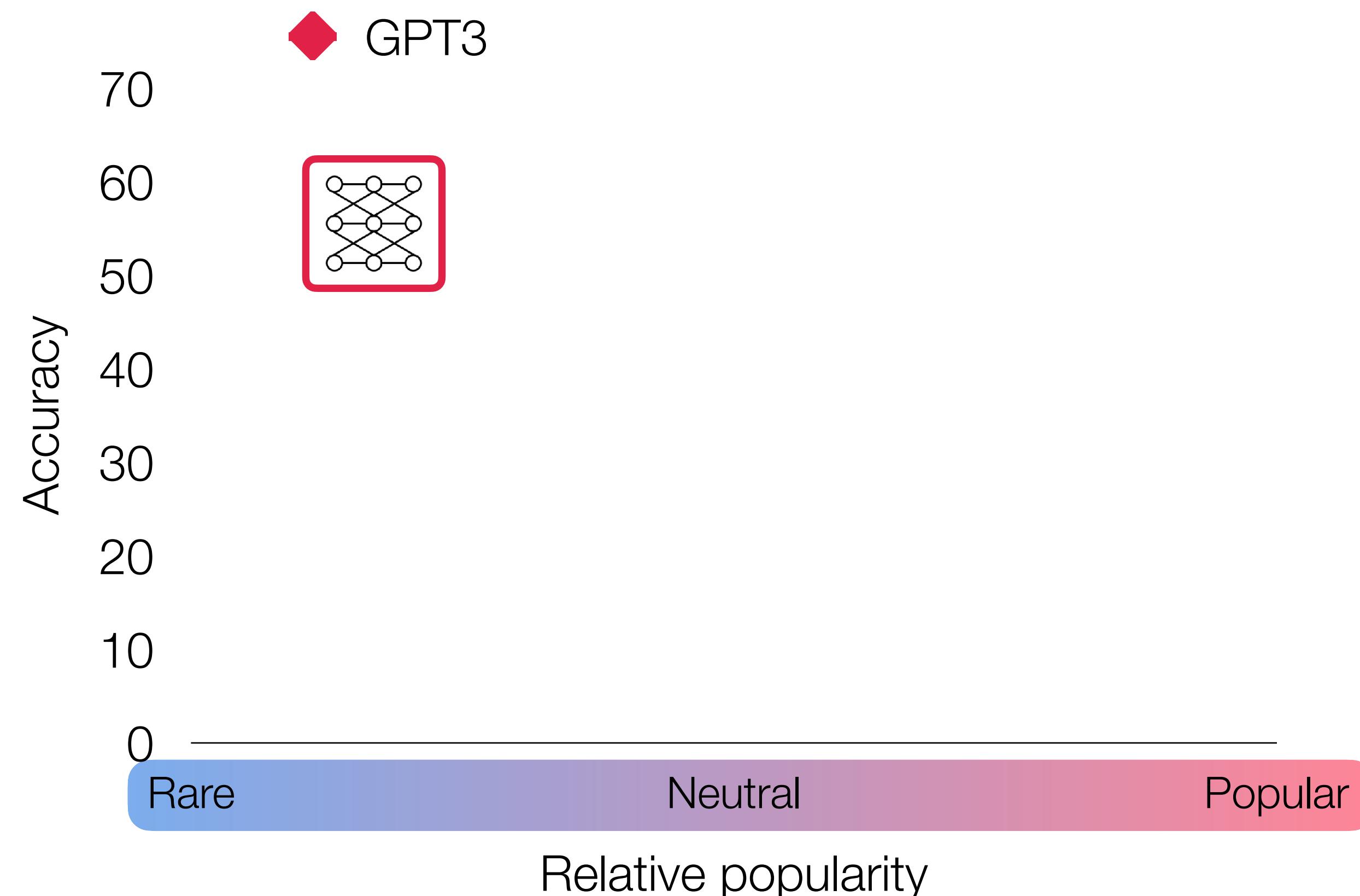


Now widely referred to as **RAG**

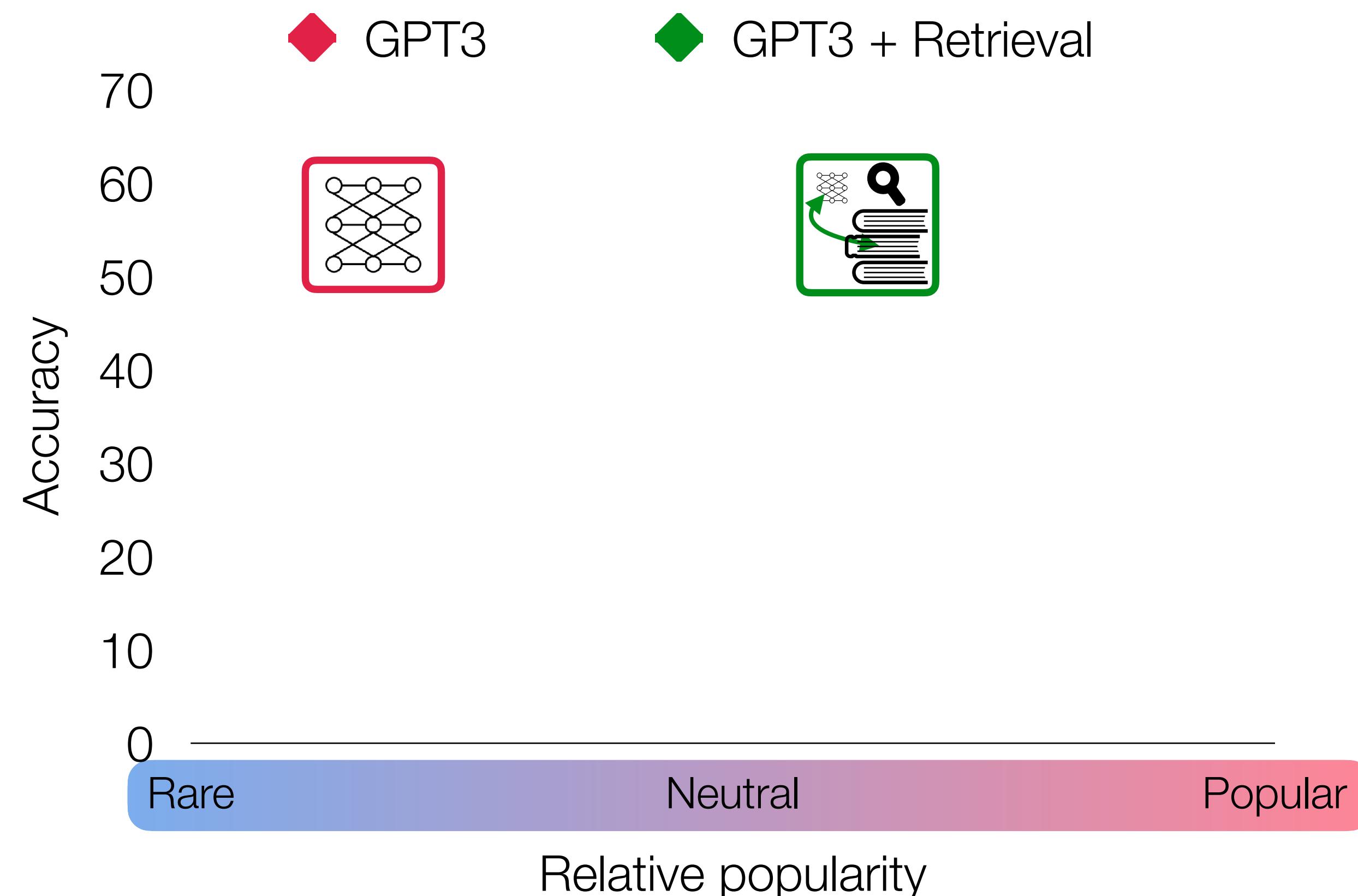
Effectiveness on Long-Tail Questions



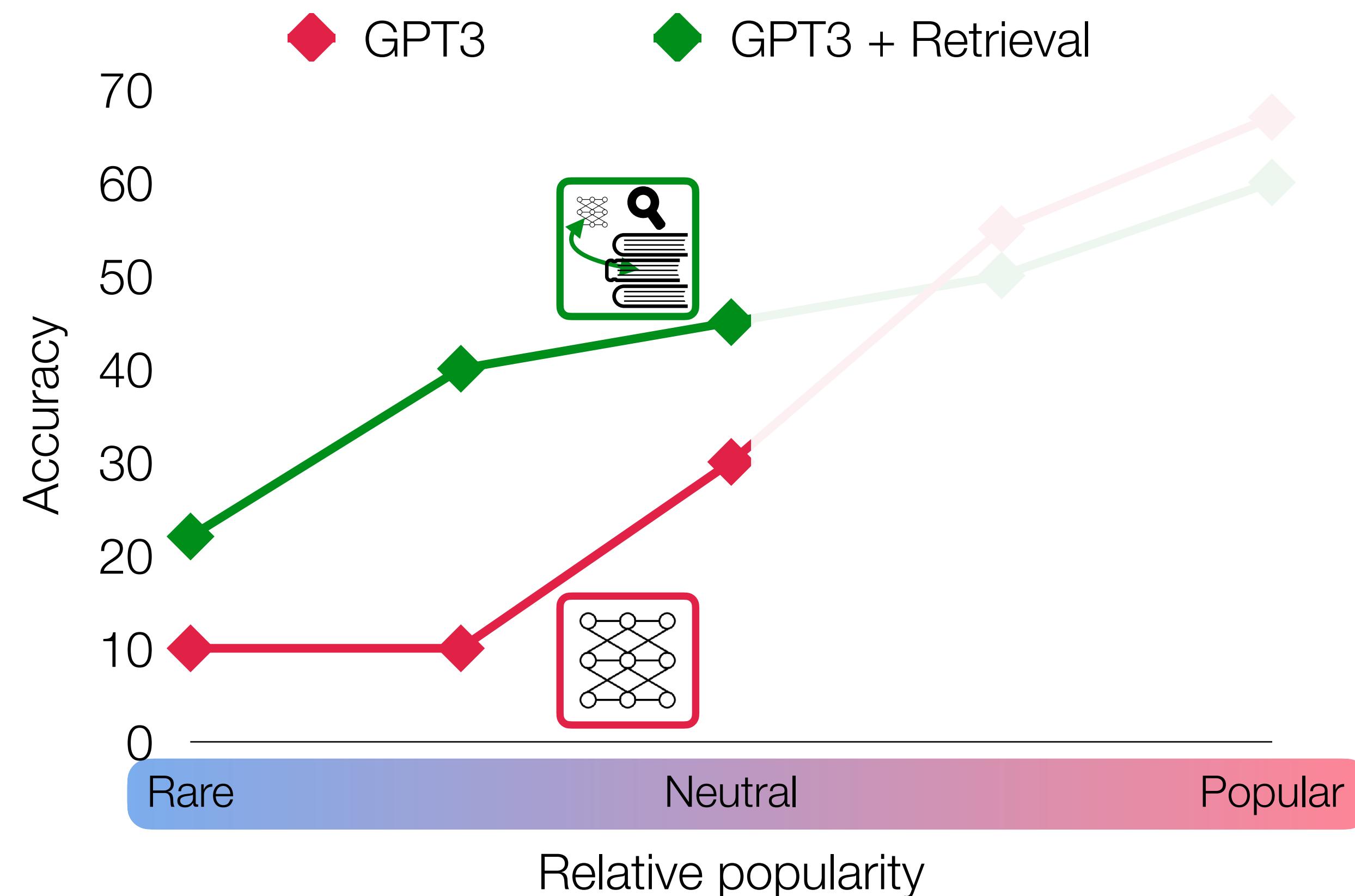
Effectiveness on Long-Tail Questions



Effectiveness on Long-Tail Questions

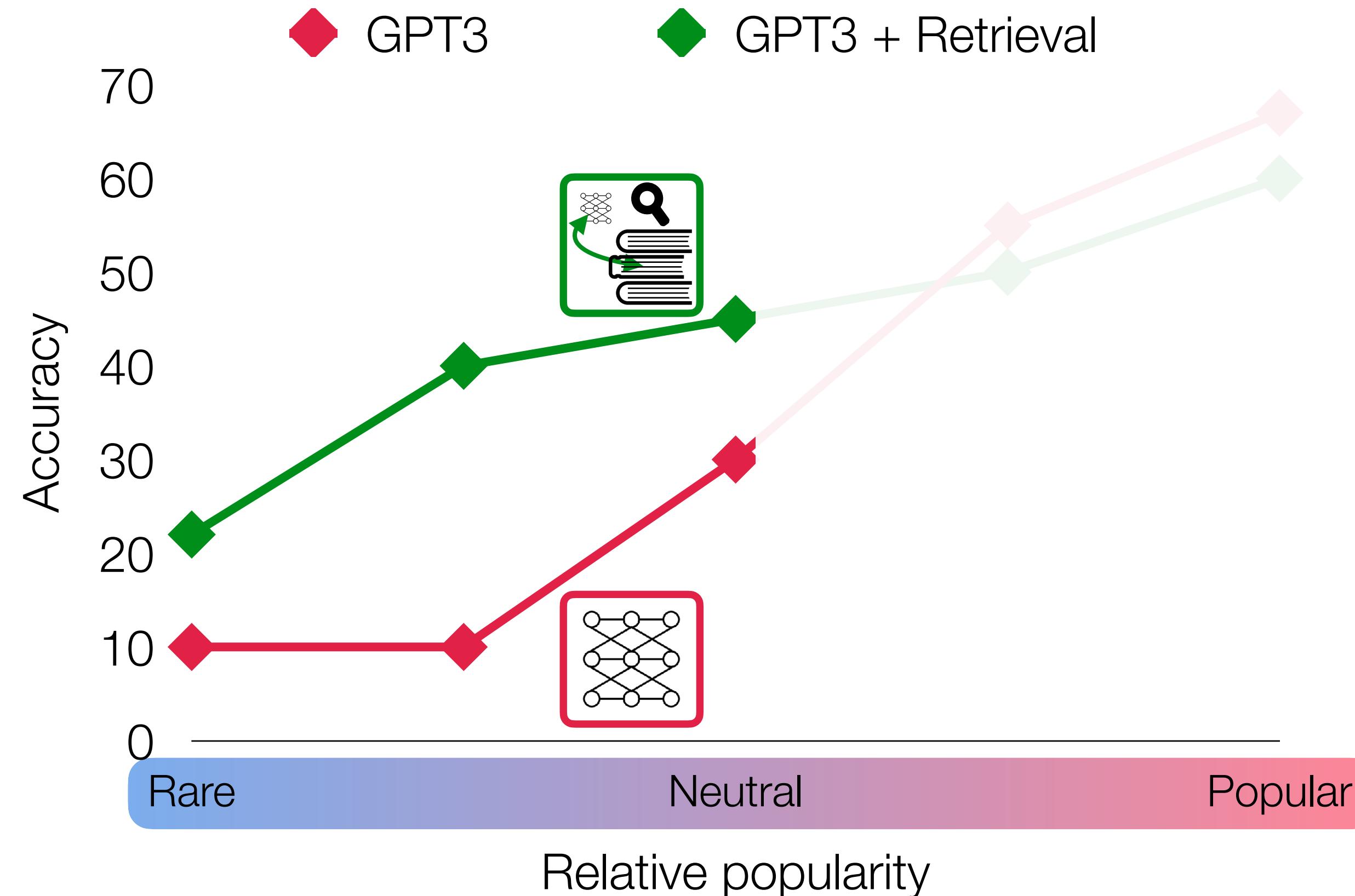


Effectiveness on Long-Tail Questions



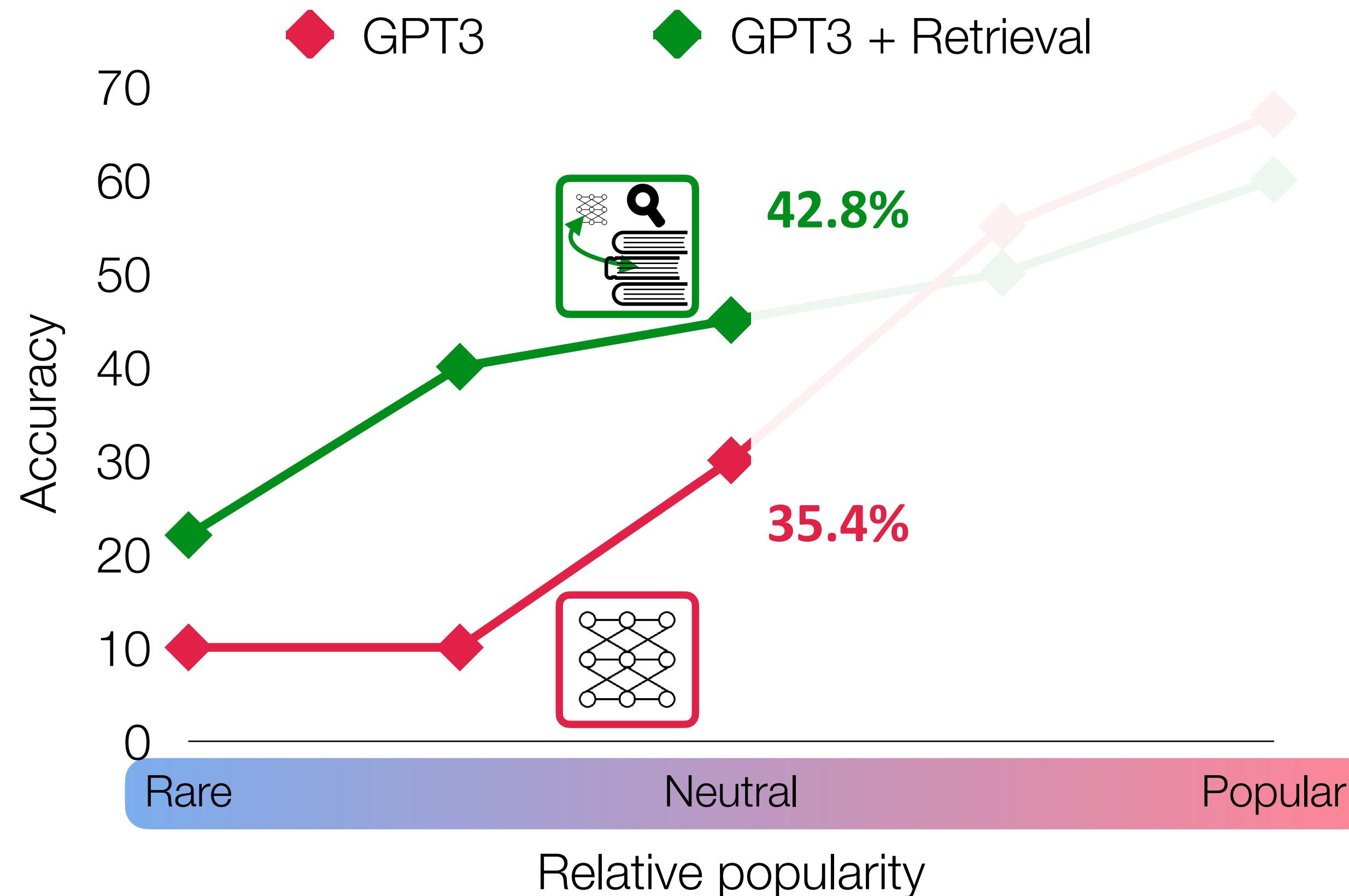
Effectiveness on Long-Tail Questions

Retrieval-Augmented LMs are highly effective for long-tail questions

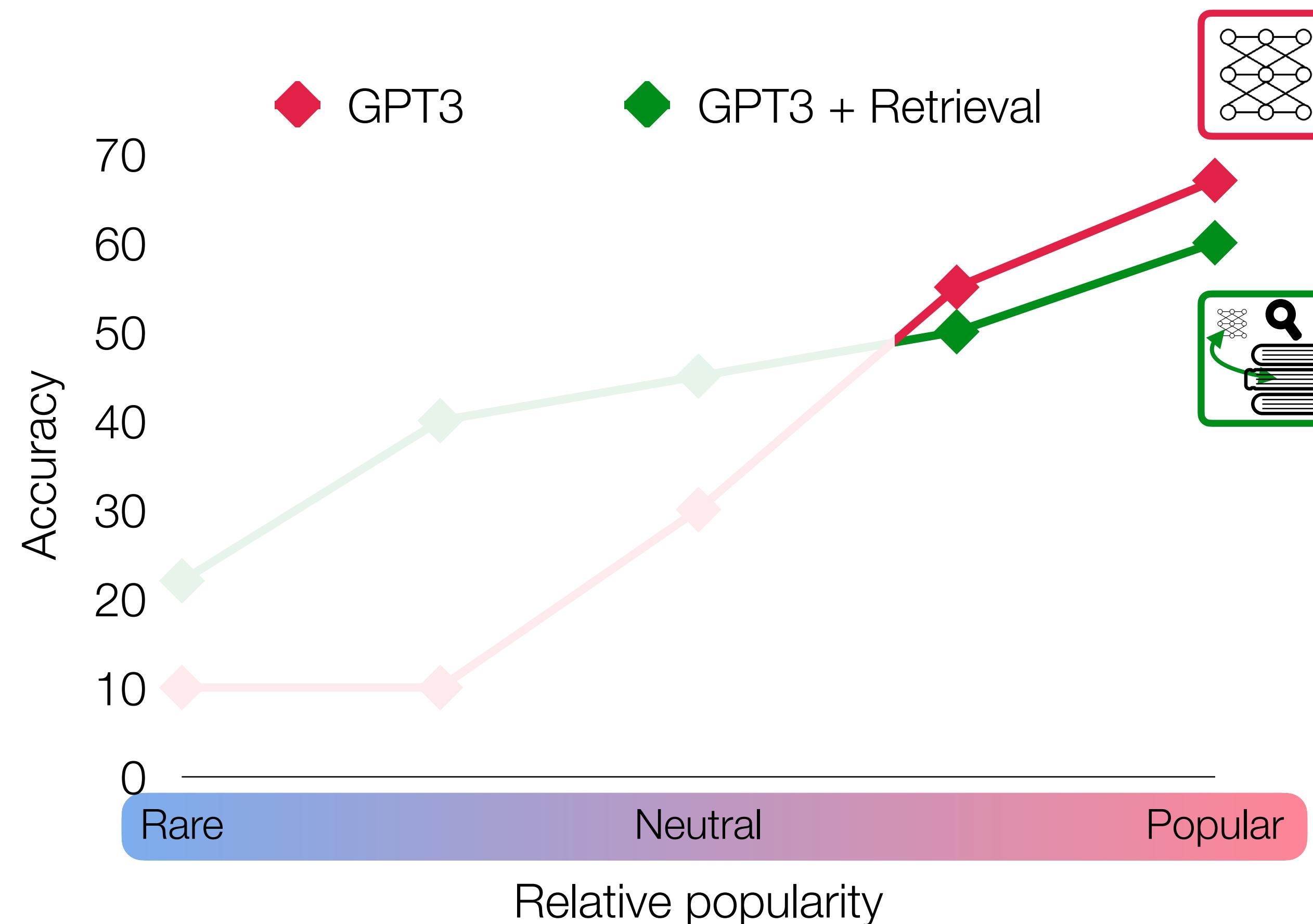


Effectiveness on Long-Tail Questions

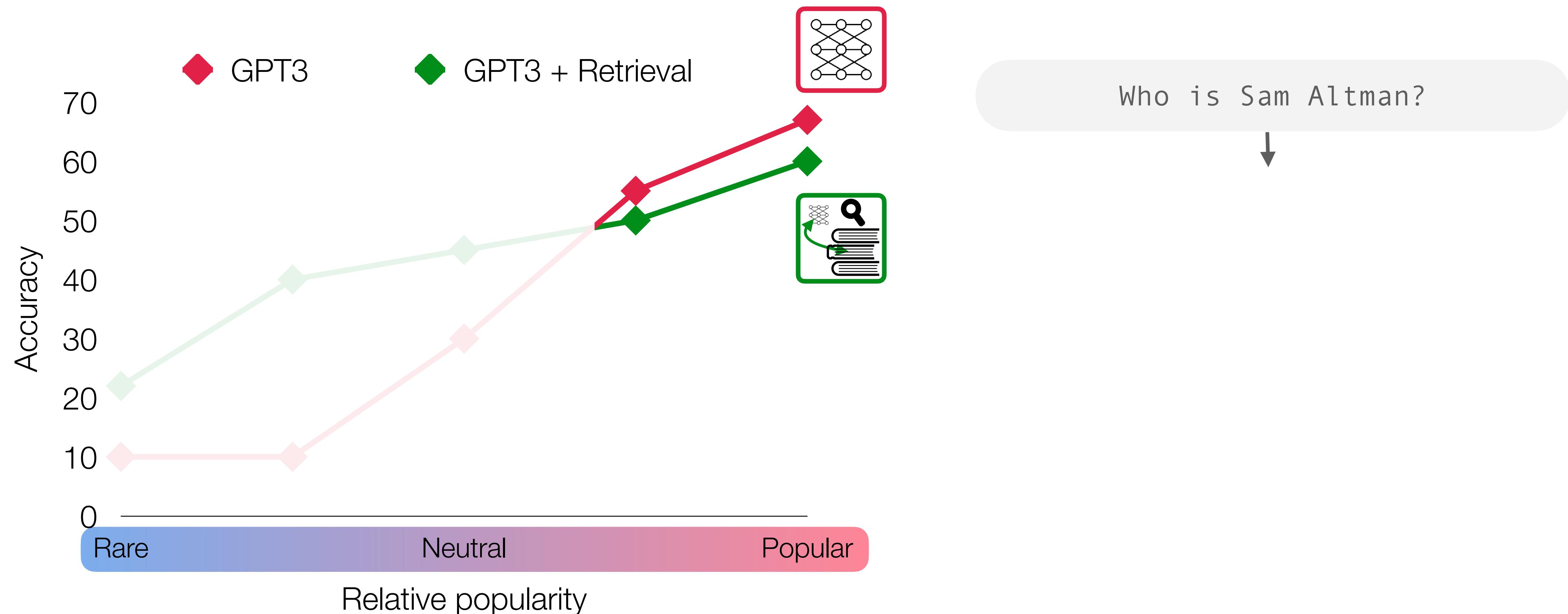
Retrieval-Augmented LMs are highly effective for long-tail questions



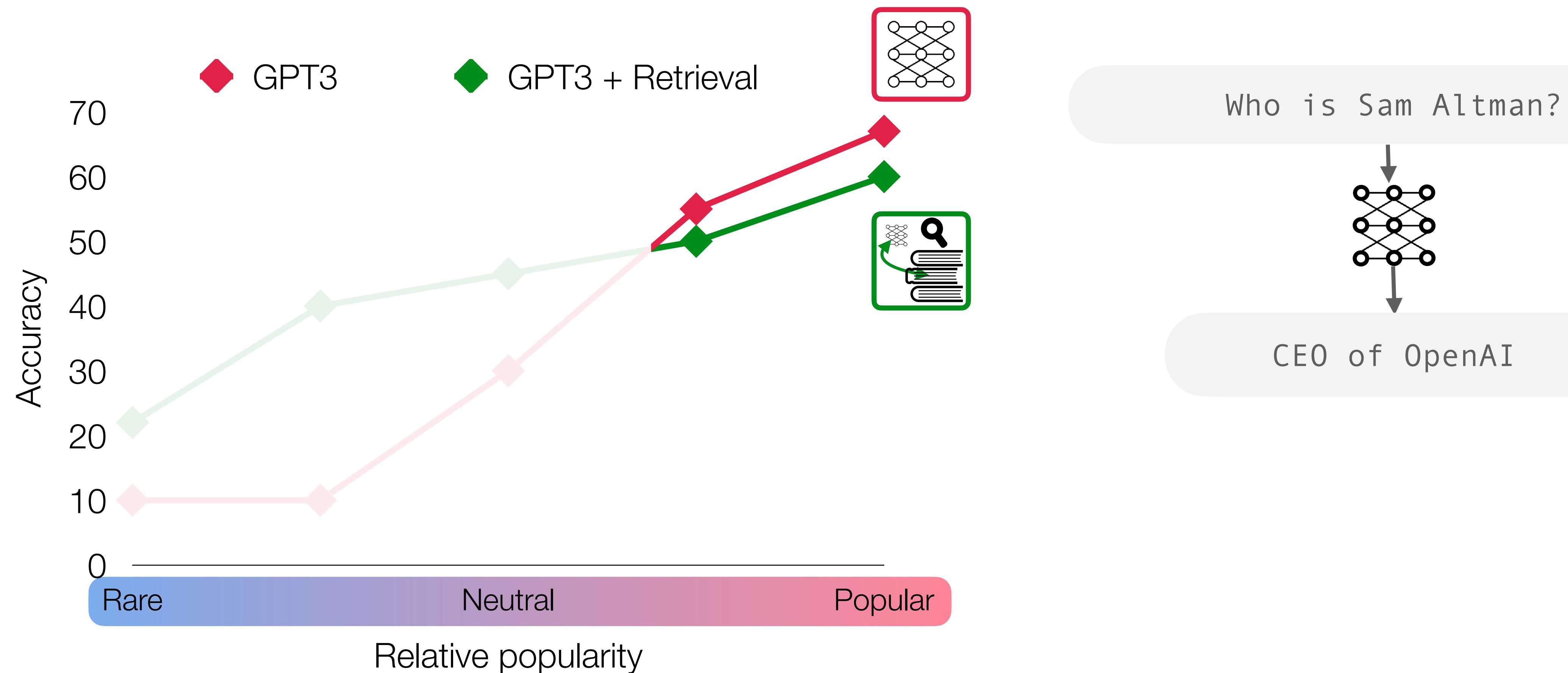
Effectiveness on Long-Tail Questions



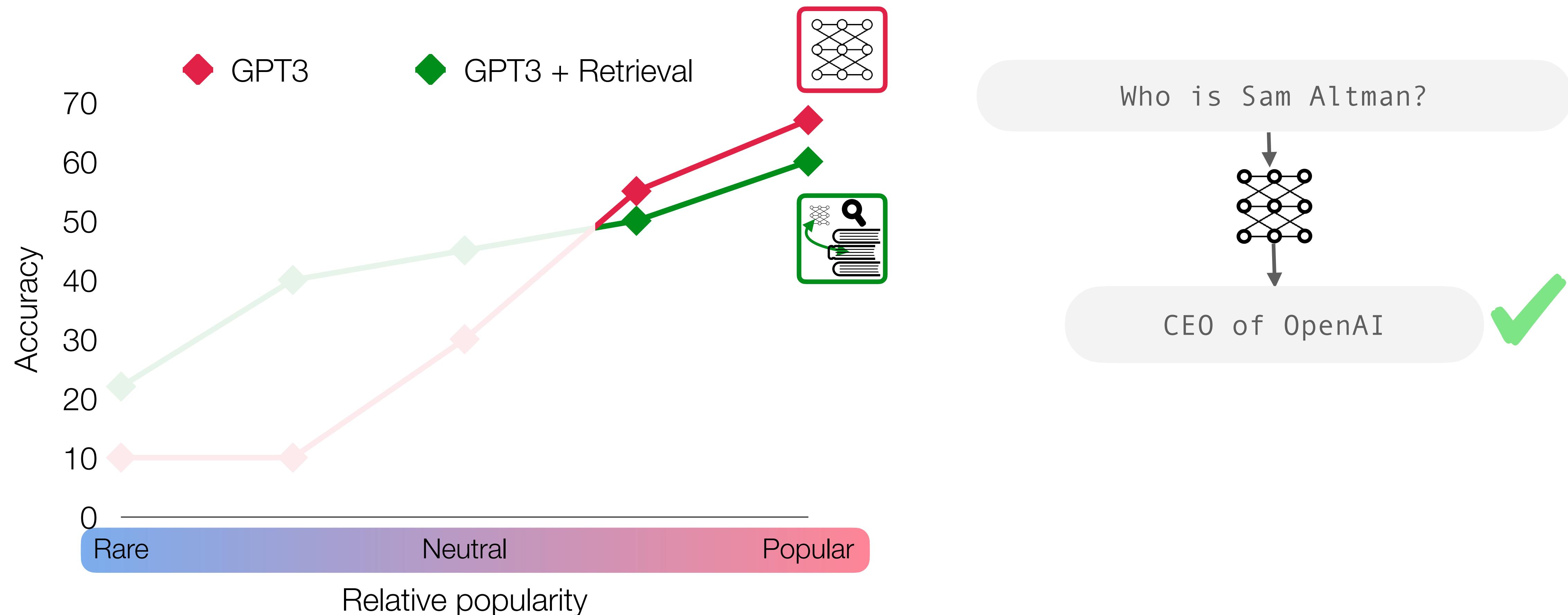
Effectiveness on Long-Tail Questions



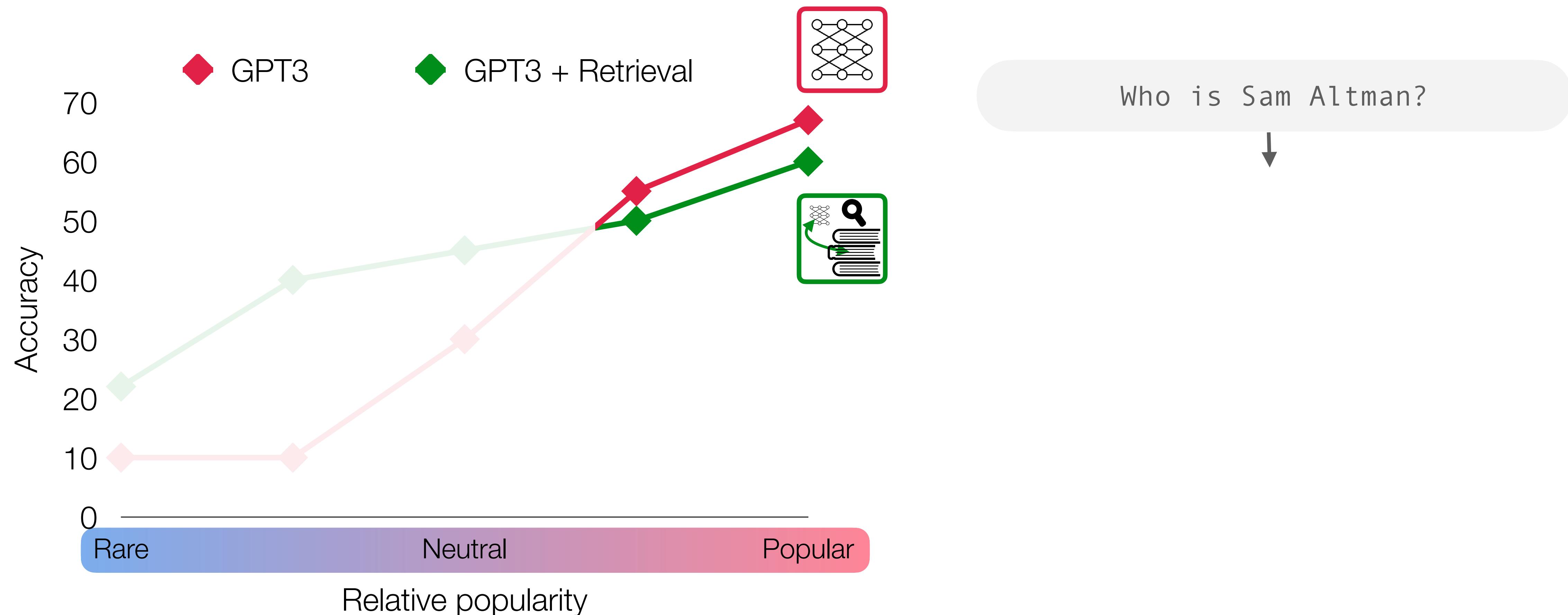
Effectiveness on Long-Tail Questions



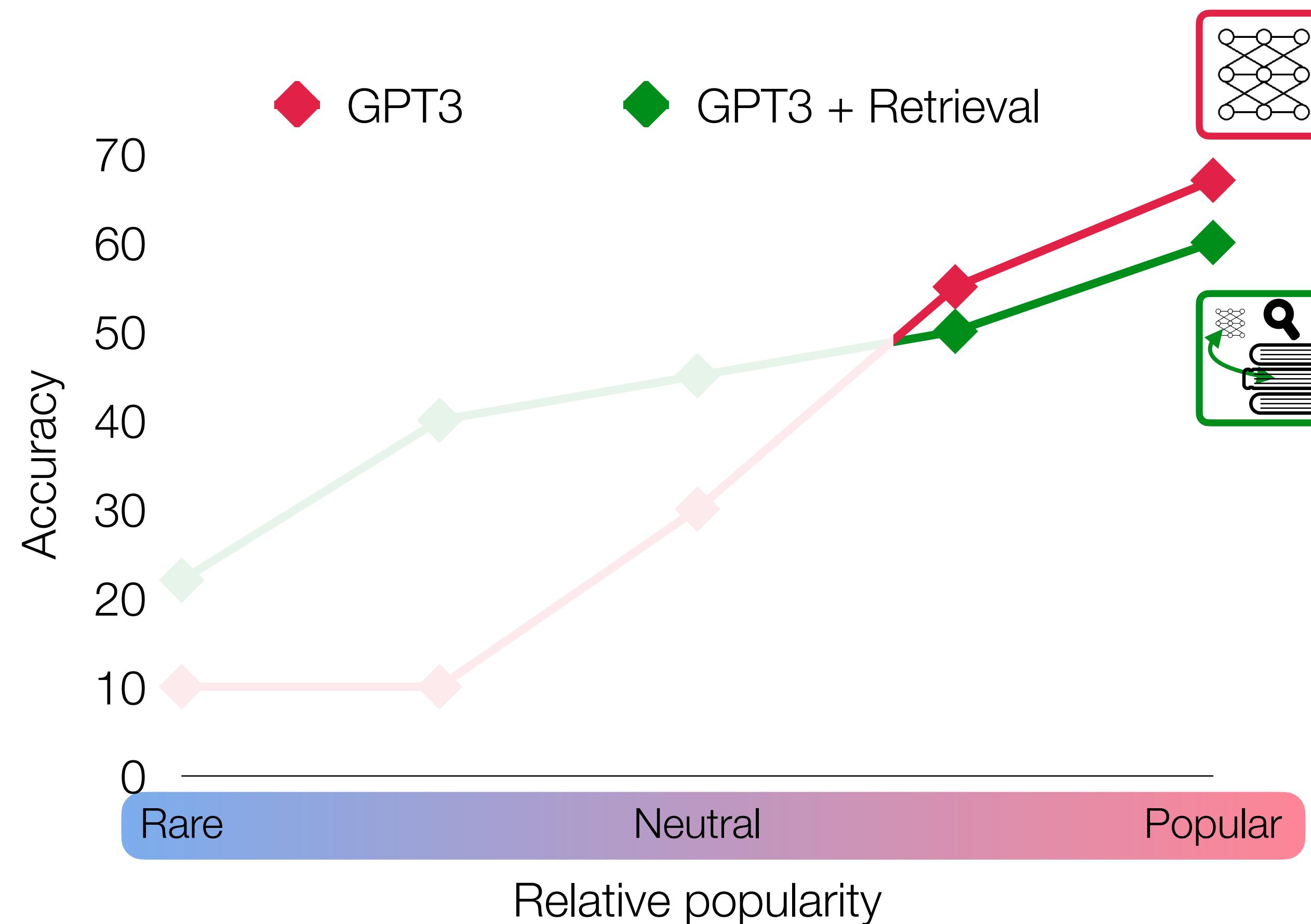
Effectiveness on Long-Tail Questions



Effectiveness on Long-Tail Questions



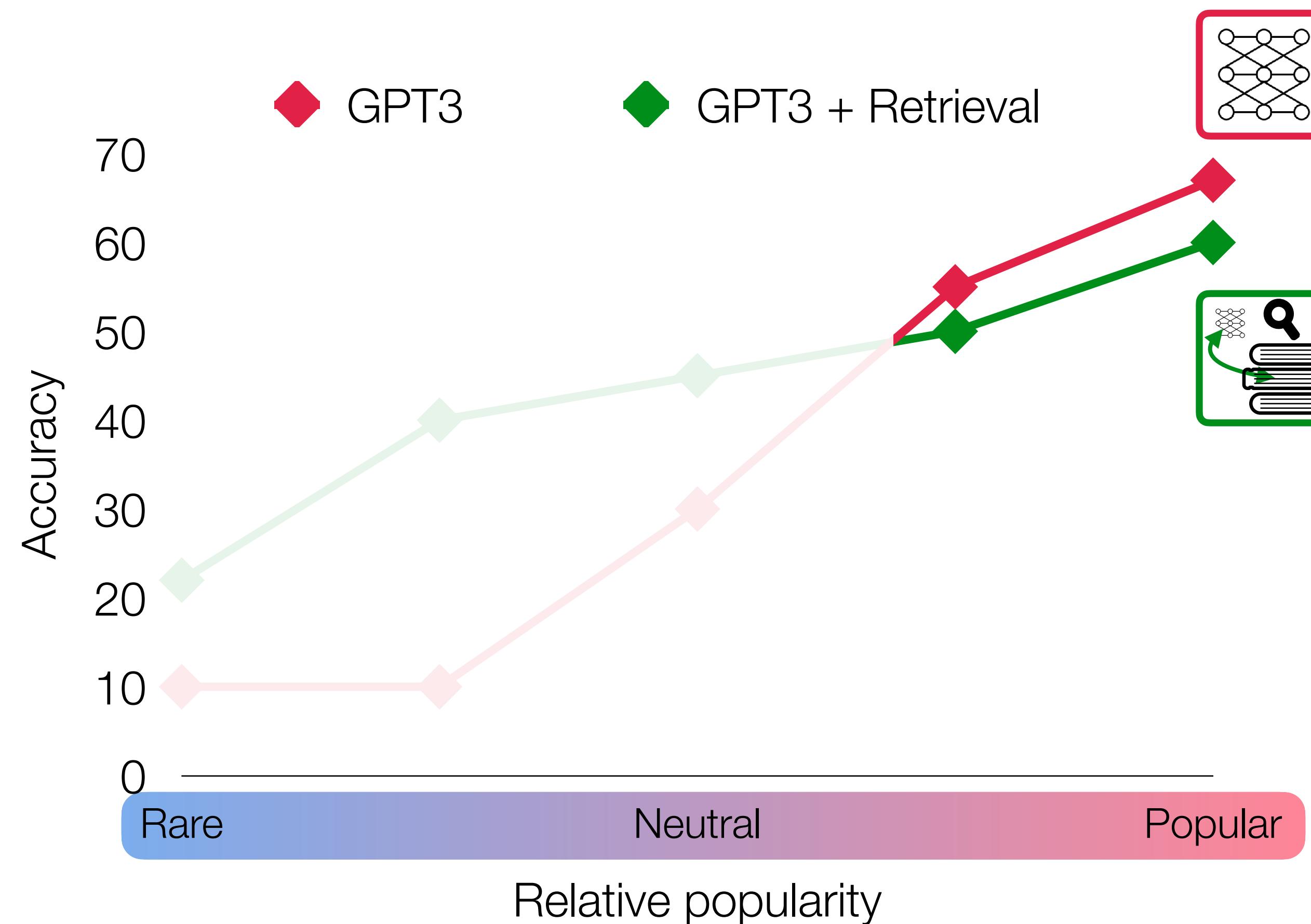
Effectiveness on Long-Tail Questions



Who is Sam Altman?

Samuel Adams (Sam Adams) was an American statesman, political philosopher and a founding father of US.

Effectiveness on Long-Tail Questions



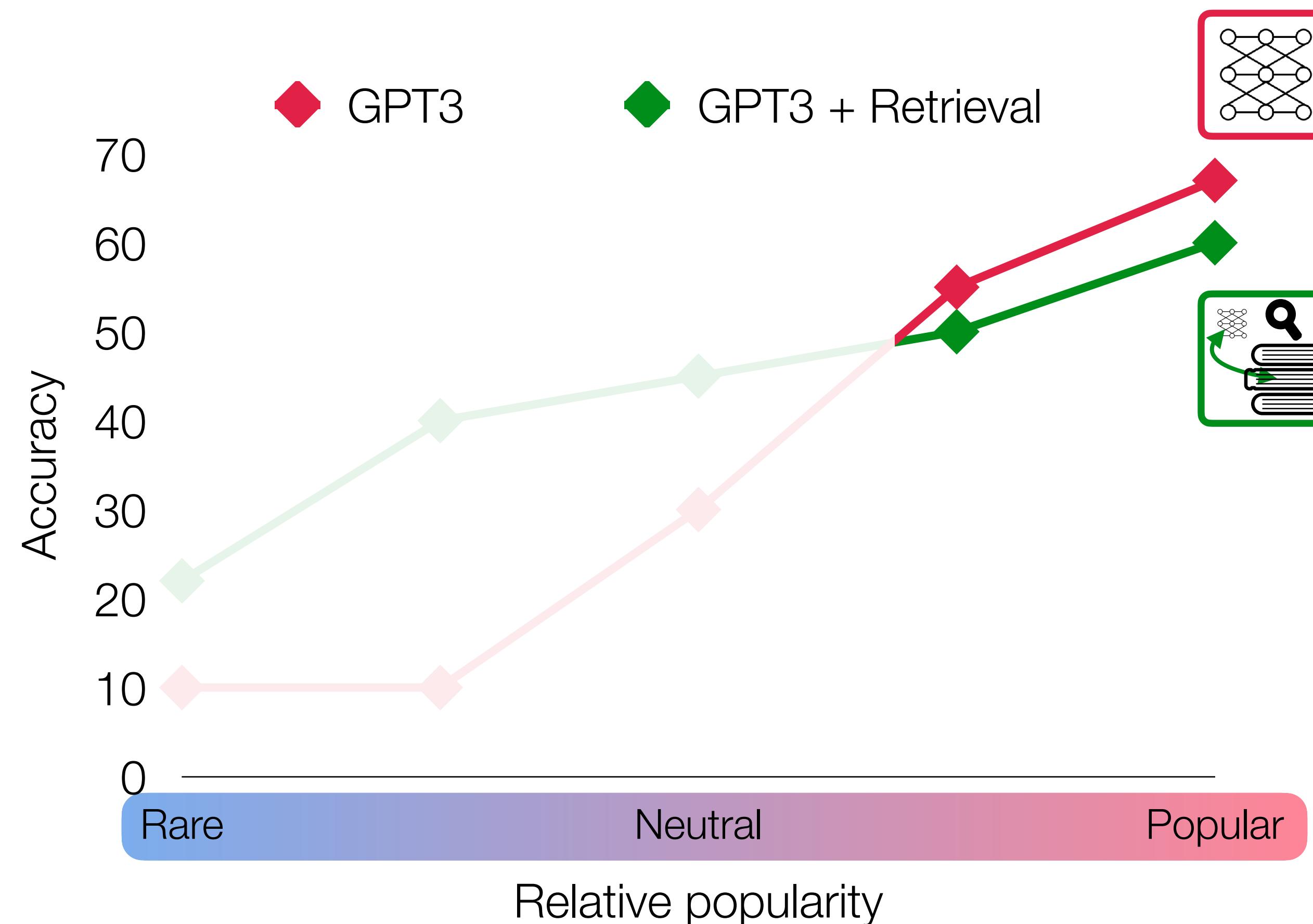
Who is Sam Altman?

Samuel Adams (Sam Adams) was an American statesman, political philosopher and a founding father of US.

He is a founding father of the US

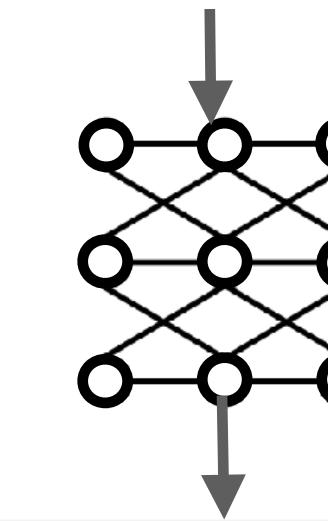
Effectiveness on Long-Tail Questions

LLMs are easily distracted by incorrect retrieval results (even when they already know the answer!)



Who is Sam Altman?

Samuel Adams (Sam Adams) was an American statesman, political philosopher and a founding father of US.



X

He is a founding father of the US

Impact

Understanding hallucinations

FActScore
(Min et al., 2023)

DYNAMICQA
(Marjanović et al., 2023)

Thread of Thought
(Zhou et al., 2023)

MiniCheck
(Tang et al., 2024)

KGQuiz
(Bai et al., 2024)

MoMe
(Li et al., 2024)

Lynx
(Ravi et al., 2024)

ANAH
(Ji et al., 2024)

FAVA
(Mishra et al., 2024)

New RAG Methods

WITQA
(Maekawa et al., 2024)

Adaptive RAG
(Jeong et al., 2024)

ActiveRAG
(Xu et al., 2024)

RQ-RAG
(Chan et al., 2024)

FILCO
(Wang et al., 2024)

AstuteRAG
(Wang et al., 2024)

RetroLLM
(Li et al., 2024)

CRAG
(Yan et al., 2024)

PopQA in benchmarks

(PopQA total downloads: 30k>)

The Hallucination Leaderboard
(Hong et al., 2024)

FLASK
(Ye et al., 2023)

MINTQA
(He et al., 2024)

Hallucinations Leaderboard

The Hallucinations Leaderboard aims to track, rank and evaluate hallucinations in LLMs.

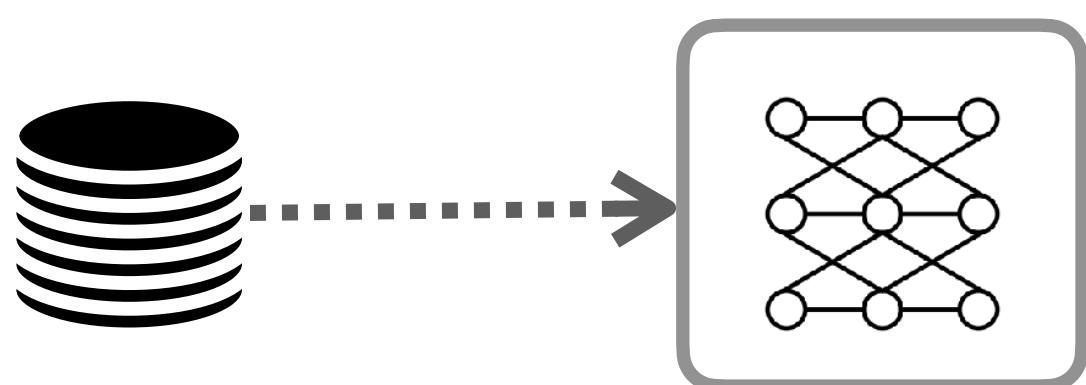
It evaluates the propensity for hallucination in Large Language Models (LLMs) across a diverse array of tasks, including Closed-book Open-domain QA, Summarization, Reading Comprehension, Instruction Following, Fact-Checking, and Hallucination Detection. The evaluation encompasses a wide range of datasets such as NQ Open, TriviaQA, TruthfulQA, XSum, CNN/DM, RACE, SQuADv2, MemoTrap, IFEval, FEVER, FaithDial, True-False, HatefulEval, NQ-Swap, and PopQA, offering a comprehensive assessment of each model's performance in generating accurate and contextually relevant content.

<https://huggingface.co/spaces/hallucinations-leaderboard/leaderboard>



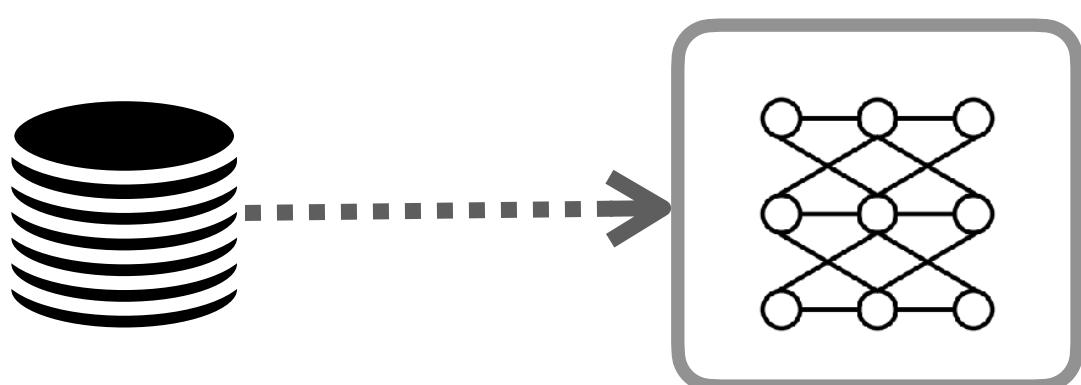
Summary

Monolithic LMs



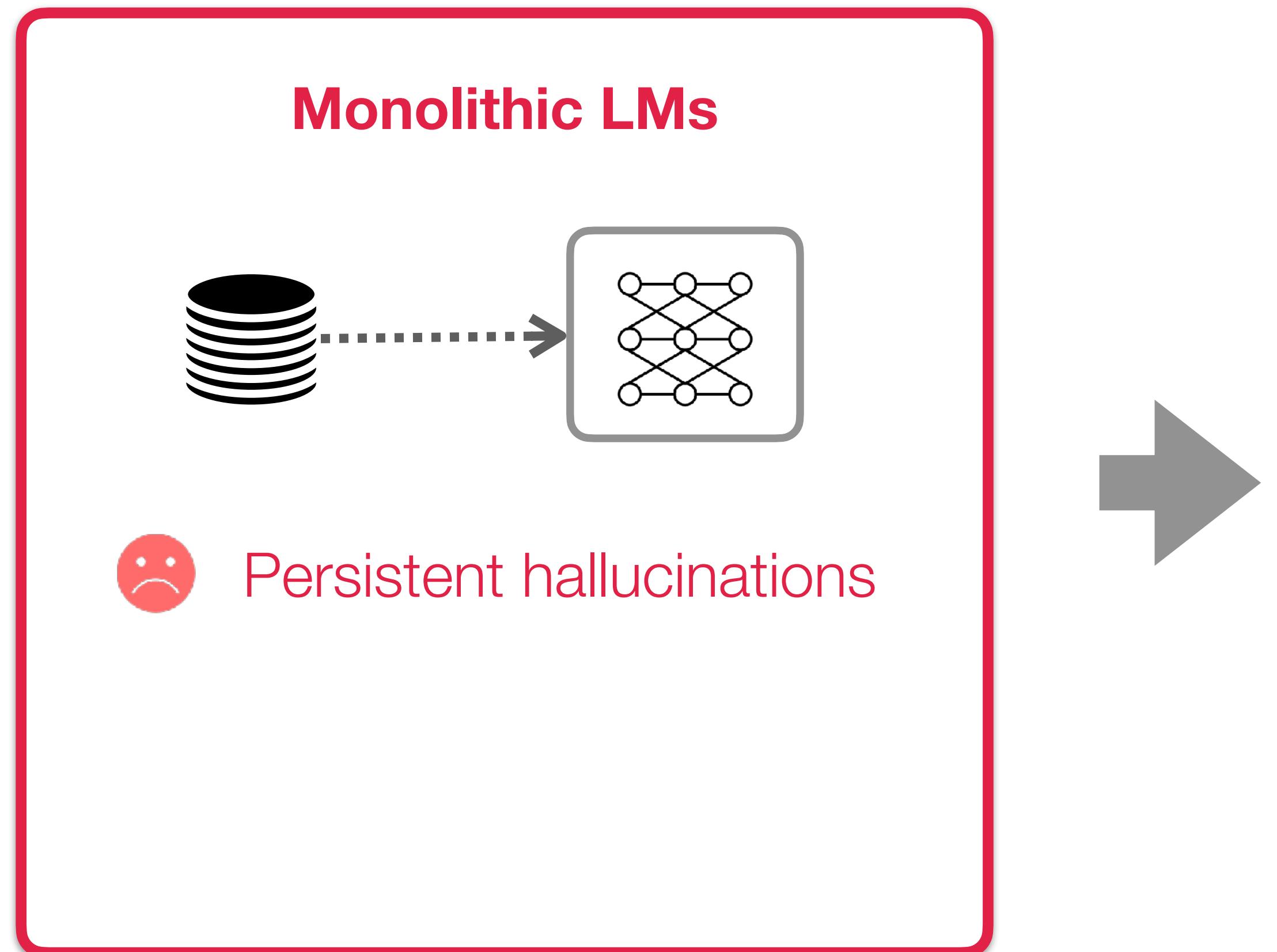
Summary

Monolithic LMs

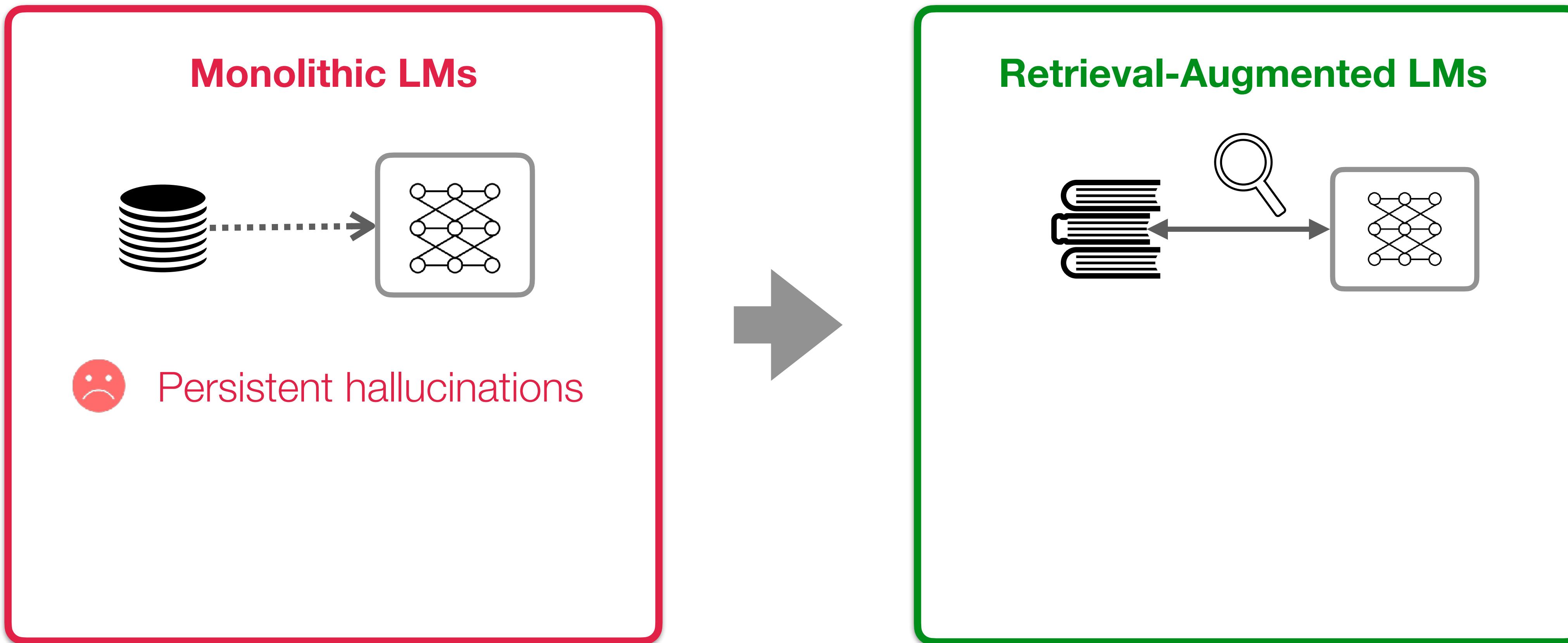


Persistent hallucinations

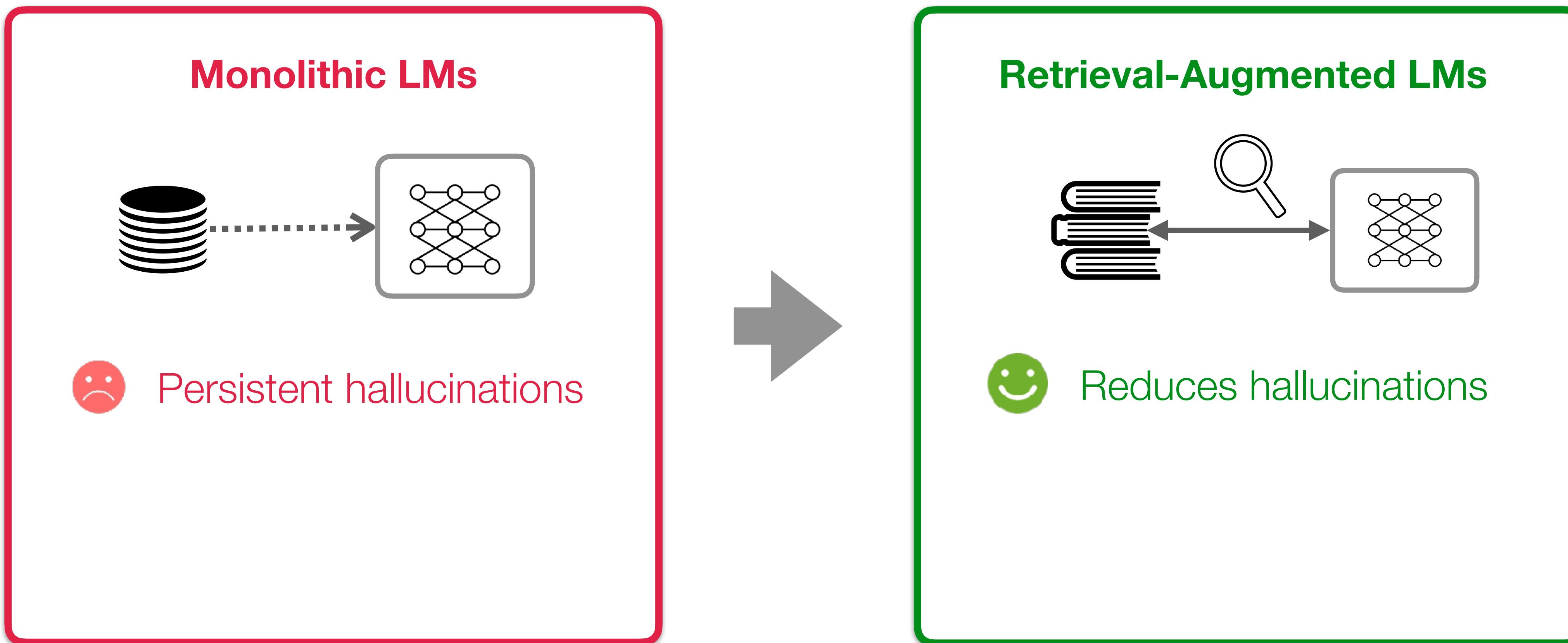
Summary



Summary

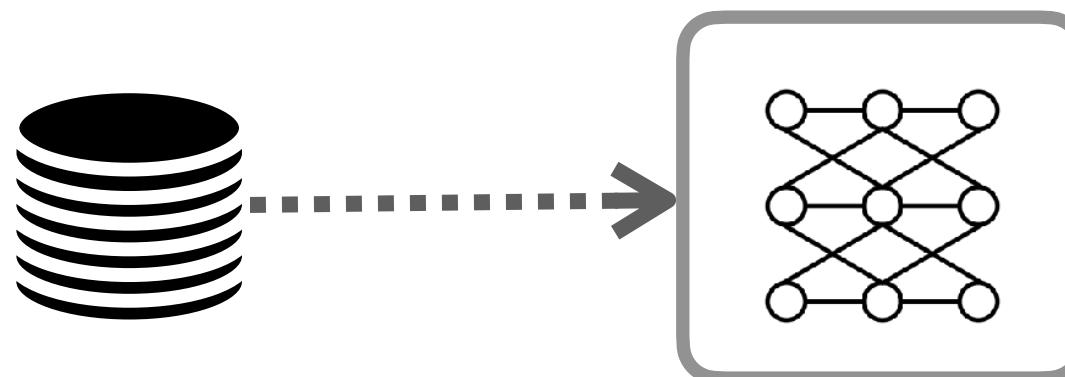


Summary



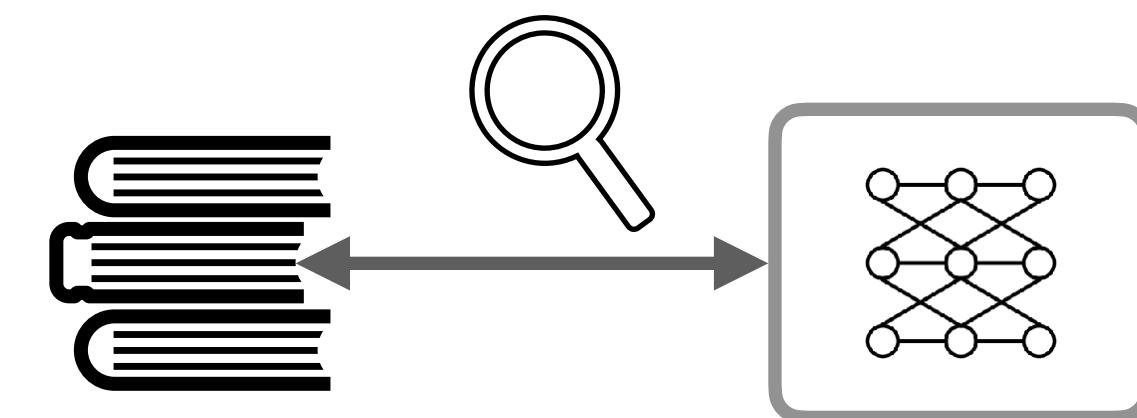
Summary

Monolithic LMs



- :(Persistent hallucinations
- :(Obsolete knowledge

Retrieval-Augmented LMs

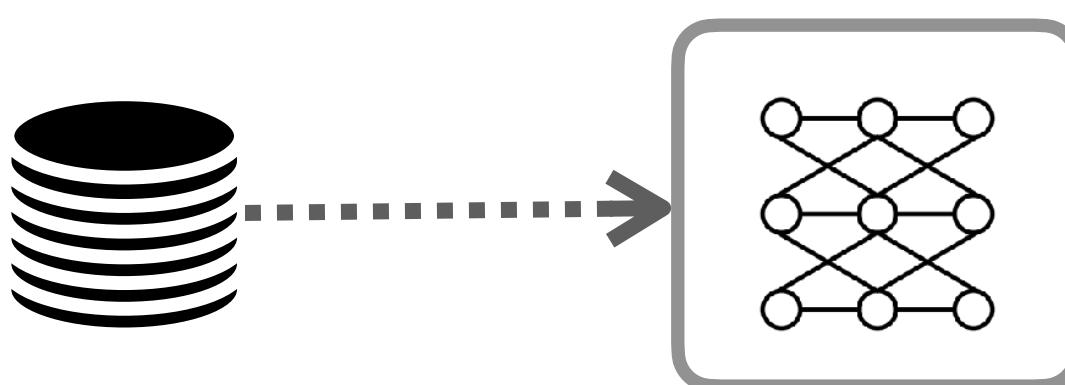


- : Smiley face Reduces hallucinations
- : Smiley face Updates knowledge w/o re-training

Kasai, ..., Asai et al. NeurIPS D&B 2023. RealTimeQA: What's the Answer Right Now?

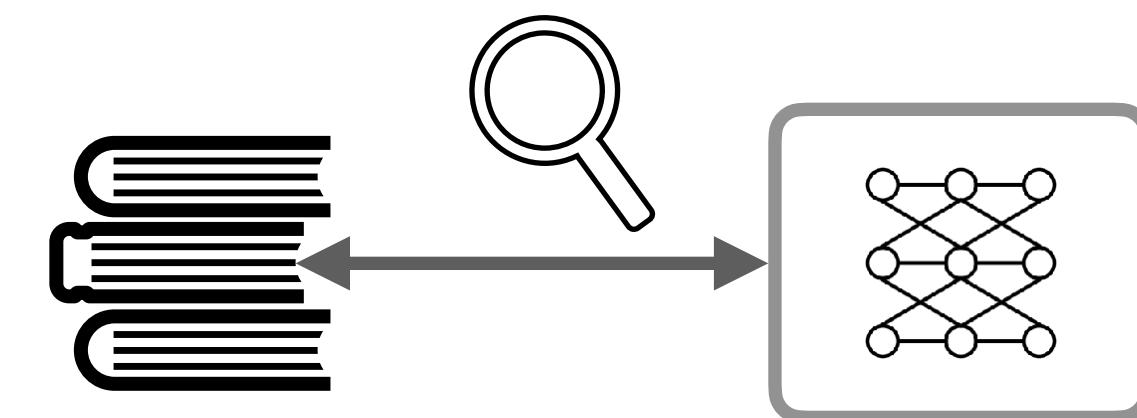
Summary

Monolithic LMs



- :(Persistent hallucinations
- :(Obsolete knowledge
- :(High training costs

Retrieval-Augmented LMs



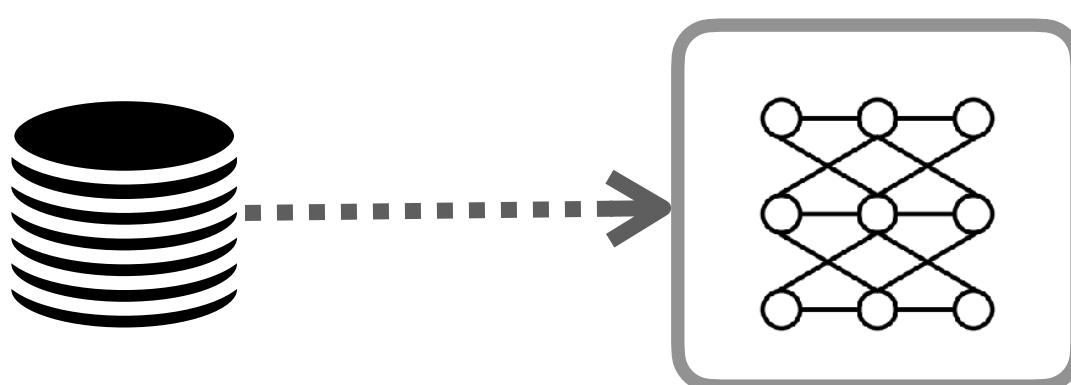
- : Smiley Reduces hallucinations
- : Smiley Updates knowledge w/o re-training
- : Smiley Enables efficient training

Kasai, ..., Asai et al. NeurIPS D&B 2023. RealTimeQA: What's the Answer Right Now?

Shao, He, Asai et al. NeurIPS 2024. Scaling Retrieval-Based Language Models with a Trillion-Token Datastore.

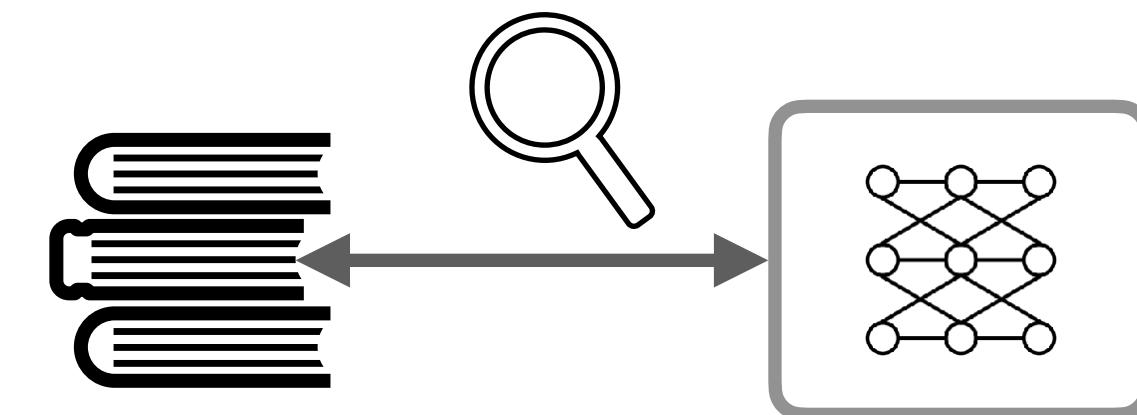
Summary

Monolithic LMs



- 😢 Persistent hallucinations
- 😢 Obsolete knowledge
- 😢 High training costs

Retrieval-Augmented LMs



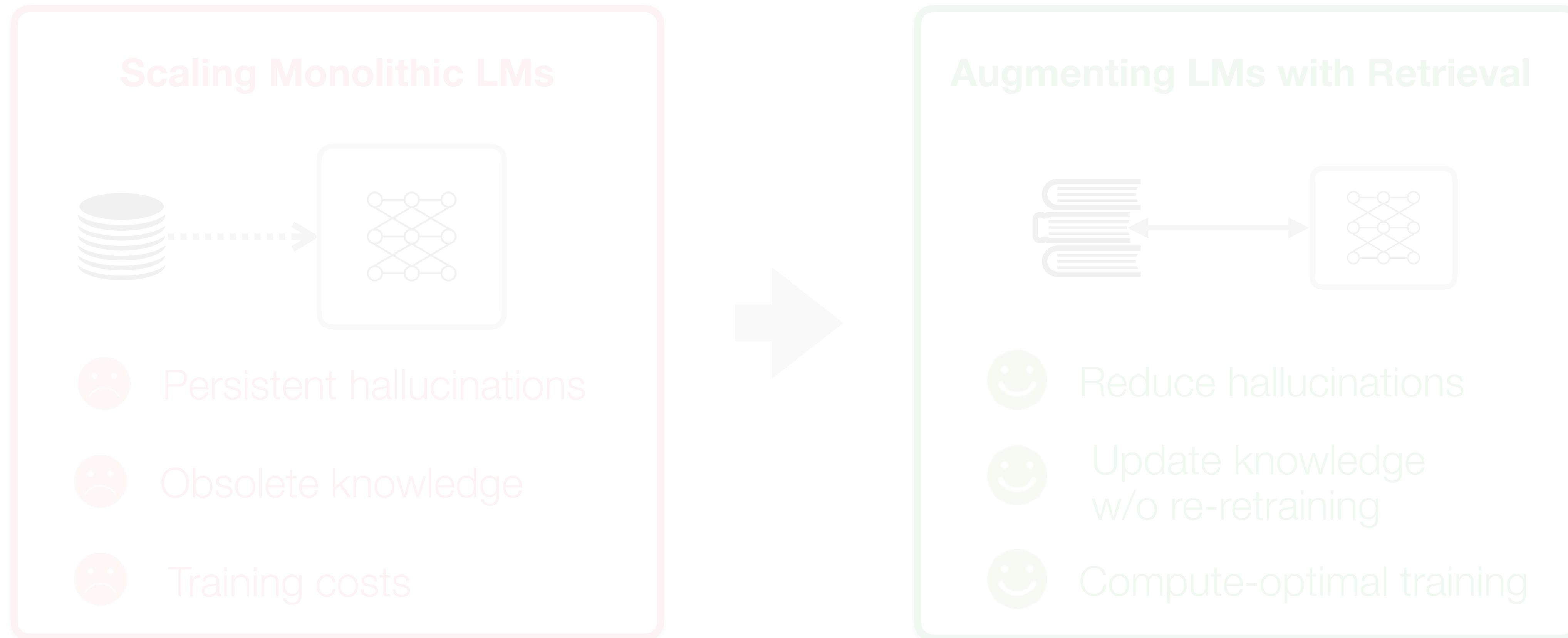
- 😊 Reduces hallucinations
- 😊 Updates knowledge w/o re-training
- 😊 Enables efficient training

Kasai, ..., Asai et al. NeurIPS D&B 2023. RealTimeQA: What's the Answer Right Now?

Shao, He, Asai et al. NeurIPS 2024. Scaling Retrieval-Based Language Models with a Trillion-Token Datastore.

Asai et al. ACL Tutorial 2023. ACL Tutorial on Retrieval-based LMs and Applications.

Summary



Kasai, ..., [Asai et al.](#) NeurIPS D&B 2023. RealTimeQA: What's the Answer Right Now?

Shao, He, [Asai et al.](#) NeurIPS 2024. Scaling Retrieval-Based Language Models with a Trillion-Token Datastore.

Summary

October 31, 2024

Introducing ChatGPT search

Get fast, timely answers with links to relevant web sources.

Plus and Team users can try it now ➔ Download Chrome extension ➔



- Persistent hallucinations
- Obsolete knowledge
- Training costs



Augmenting LMs with Retrieval



- Reduce hallucinations
- Update knowledge w/o re-training
- Compute-optimal training

Kasai, ..., Asai et al. NeurIPS D&B 2023. RealTimeQA: What's the Answer Right Now?

Shao, He, Asai et al. NeurIPS 2024. Scaling Retrieval-Based Language Models with a Trillion-Token Datastore.

Summary

October 31, 2024

Introducing ChatGPT search

Get fast, timely answers with links to relevant web sources.

Plus and Team users can try it now ↗ Download Chrome extension ↗

- Persistent hallucinations
- Obsolete knowledge
- Training costs



Who is Sam Altman?

Samuel Adams (Sam Adams) was an American statesman, political philosopher and a founding father of US.

Reduce hallucinations

Update knowledge w/o re-training

He is a founding father of the US

Kasai, ..., Asai et al. NeurIPS D&B 2023. RealTimeQA: What's the Answer Right Now?

Shao, He, Asai et al. NeurIPS 2024. Scaling Retrieval-Based Language Models with a Trillion-Token Datastore.

Summary

October 31, 2024

Introducing ChatGPT search

Get fast, timely answers with links to relevant web sources.

Plus and Team users can try it now ↗ Download Chrome extension ↗

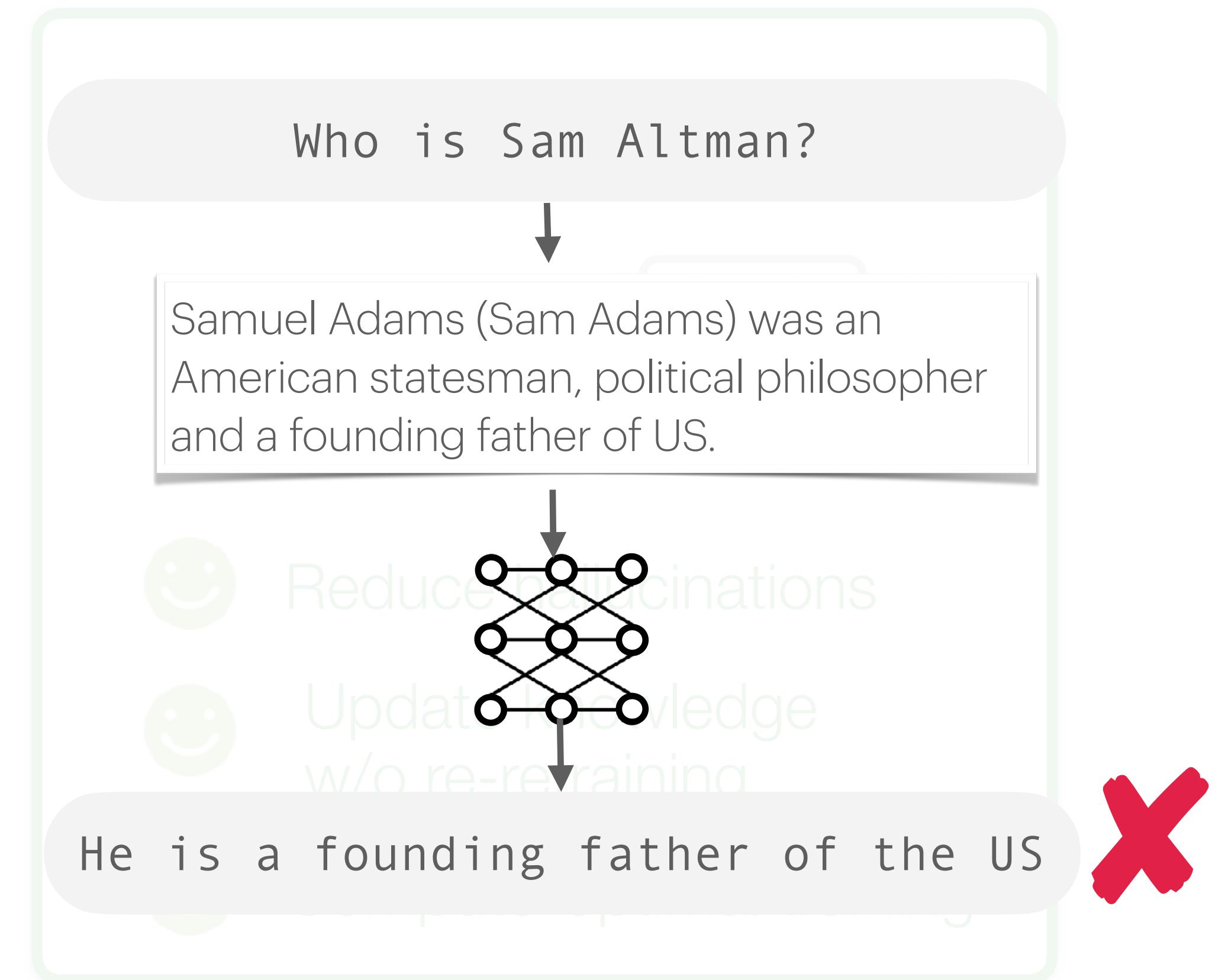
Persistent hallucinations

G Gizmodo

Google Search Is Now a Giant Hallucination

Google rolled out AI overviews across the United States this month, exposing its flagship product to the hallucinations of large language...

2024/05/24



Kasai, ..., Asai et al. NeurIPS D&B 2023. RealTimeQA: What's the Answer Right Now?

Shao, He, Asai et al. NeurIPS 2024. Scaling Retrieval-Based Language Models with a Trillion-Token Datastore.

Frontiers of Retrieval-Augmented LMs



From Monolithic to Retrieval-Augmented LMs

Why?



New Foundations of Retrieval-Augmented LMs

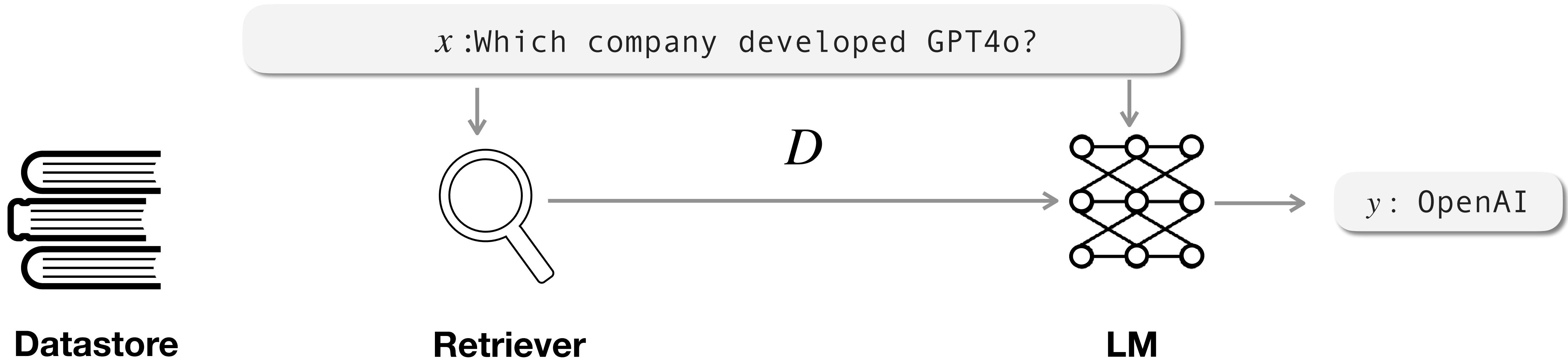
How?



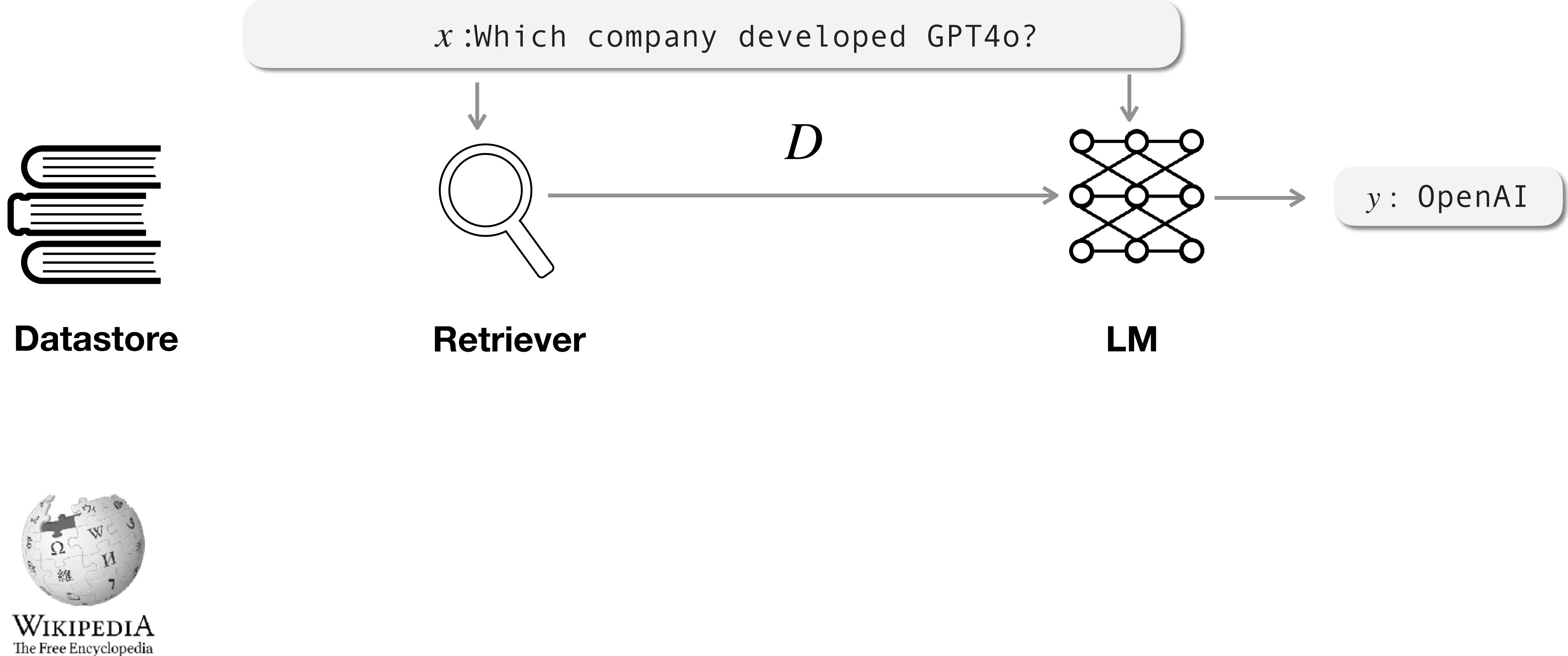
Real-World Impacts of Retrieval-Augmented LMs

What?

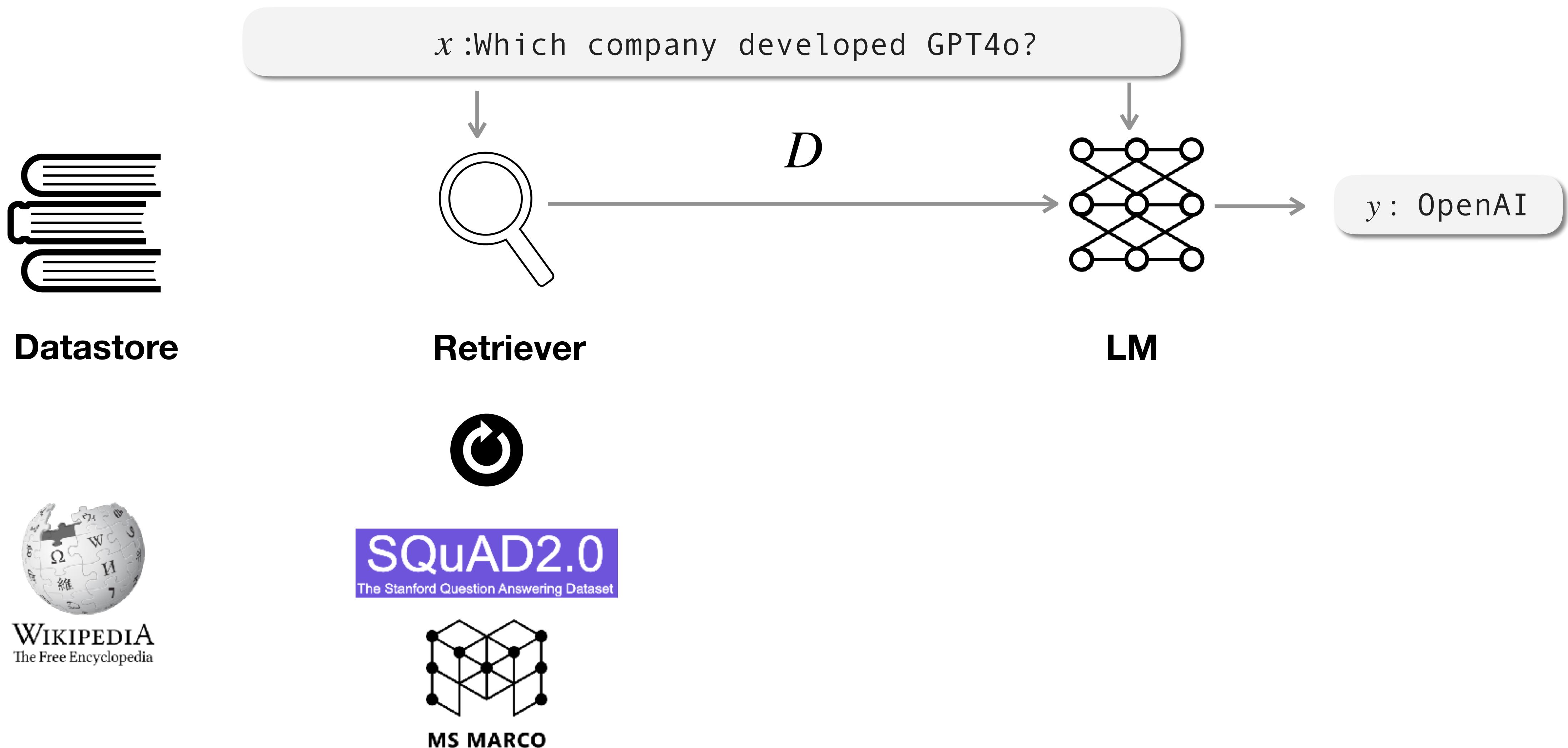
Common RAG Pipelines



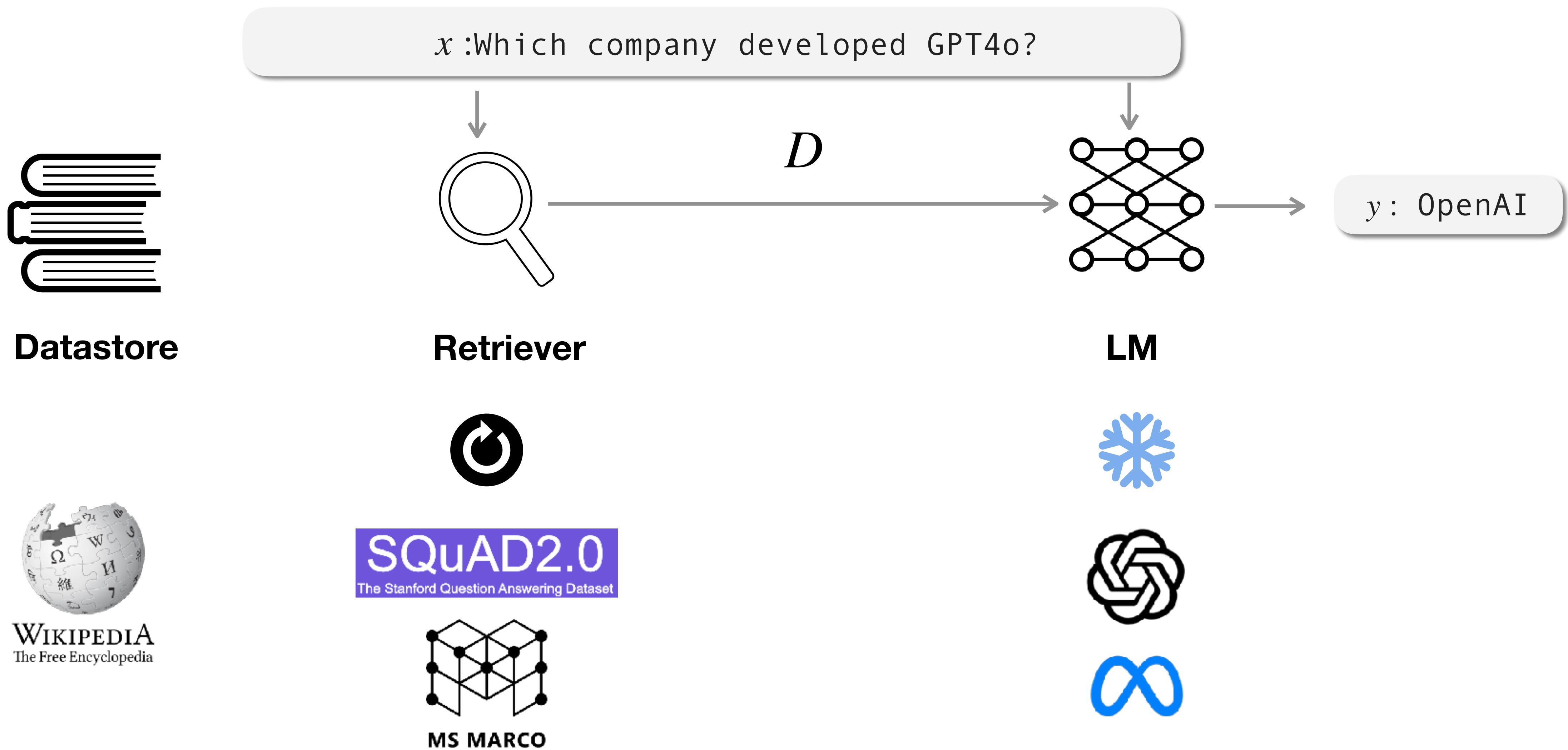
Common RAG Pipelines



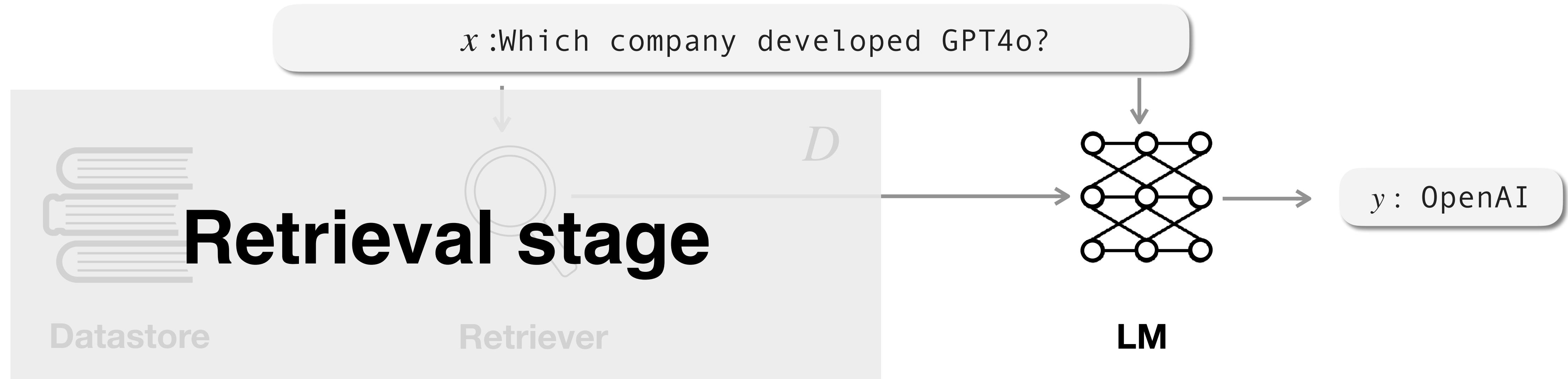
Common RAG Pipelines



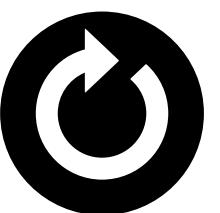
Common RAG Pipelines



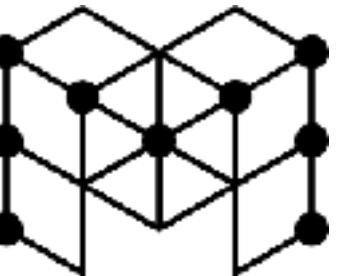
Common RAG Pipelines



WIKIPEDIA
The Free Encyclopedia



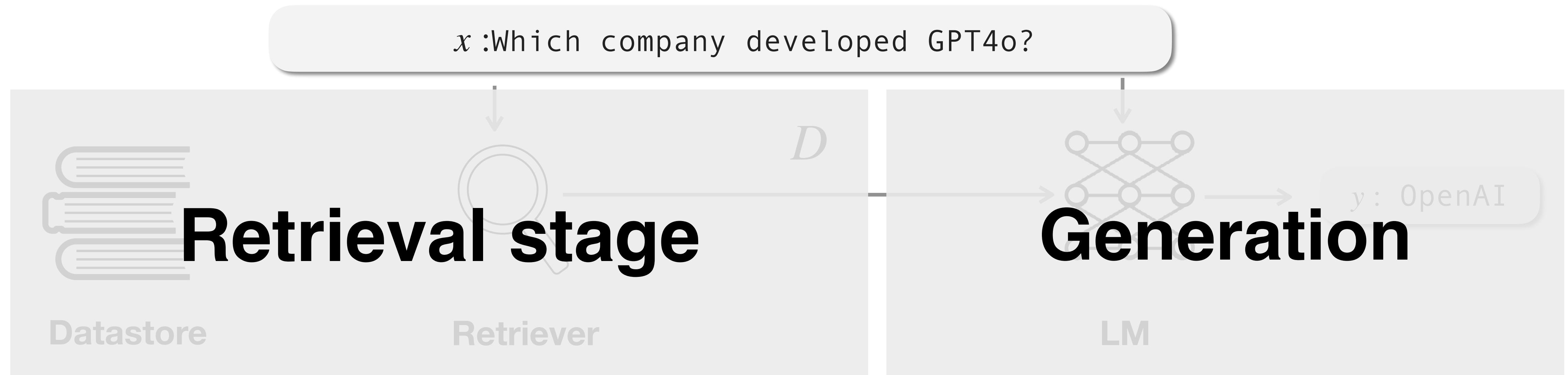
SQuAD2.0
The Stanford Question Answering Dataset



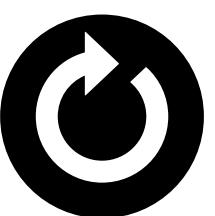
MS MARCO



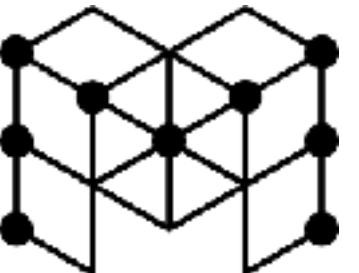
Common RAG Pipelines



WIKIPEDIA
The Free Encyclopedia



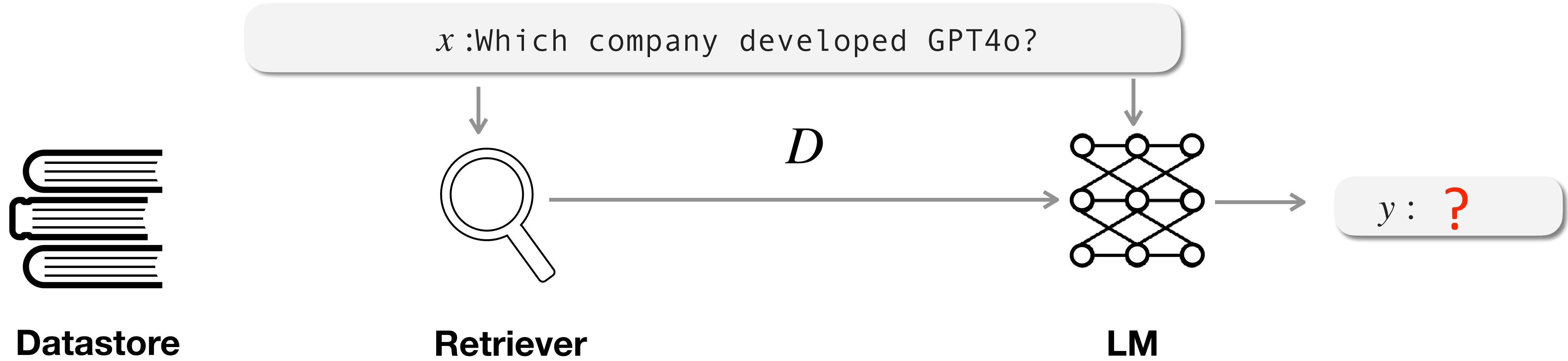
SQuAD2.0
The Stanford Question Answering Dataset



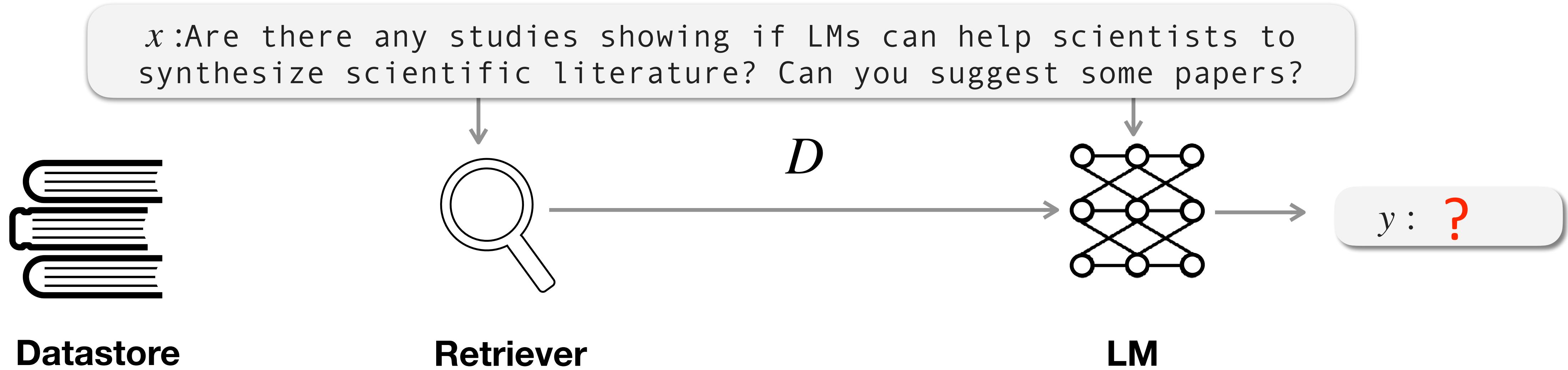
MS MARCO



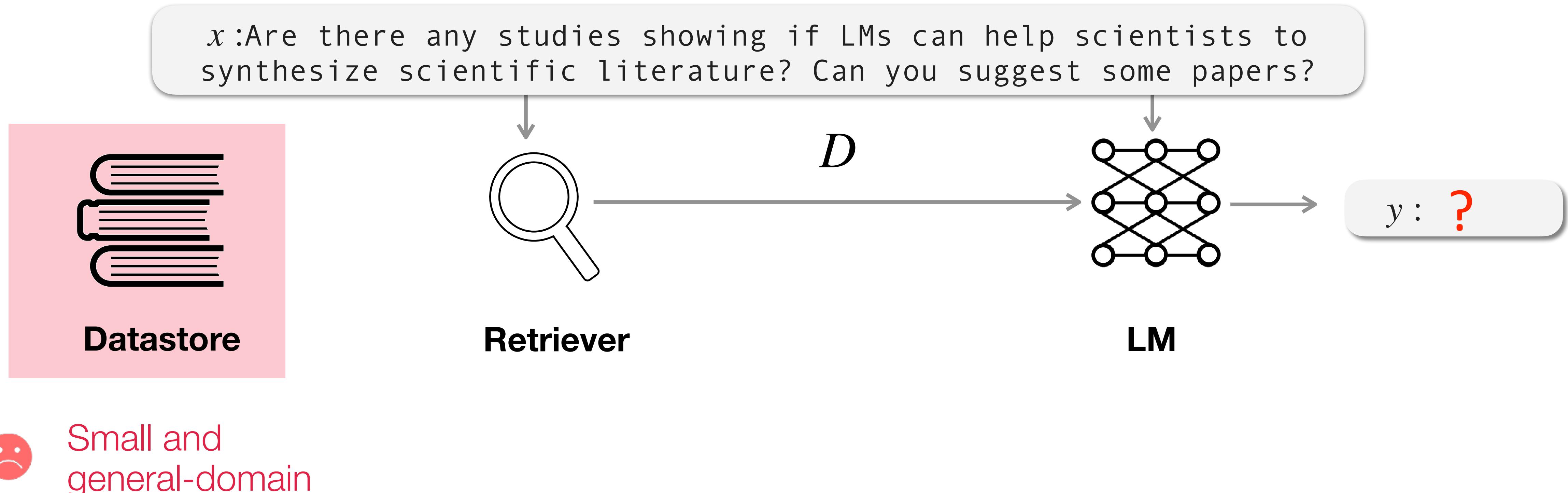
Challenges in Common RAG Pipelines



Challenges in Common RAG Pipelines

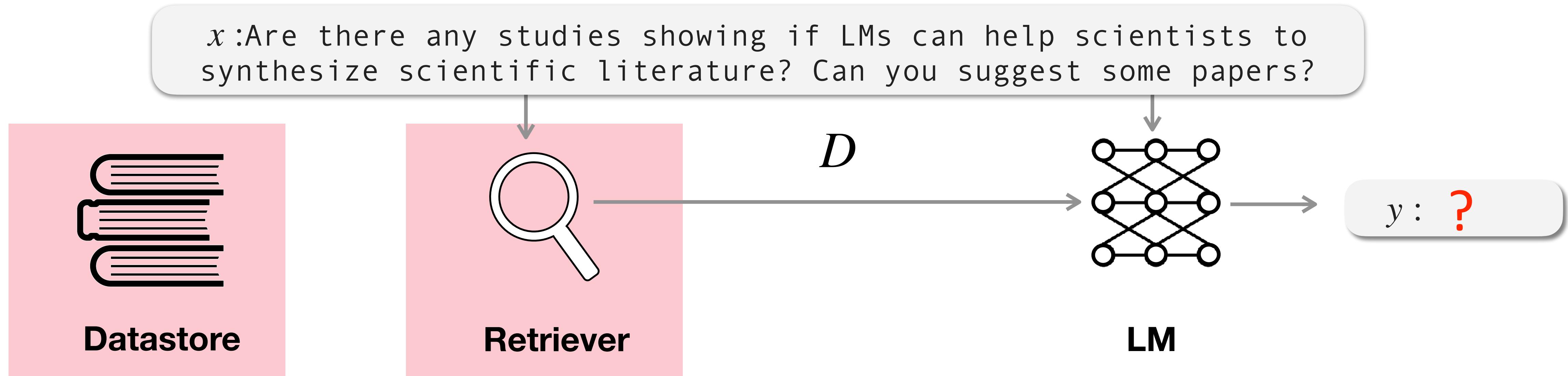


Challenges in Common RAG Pipelines



Small and general-domain

Challenges in Common RAG Pipelines

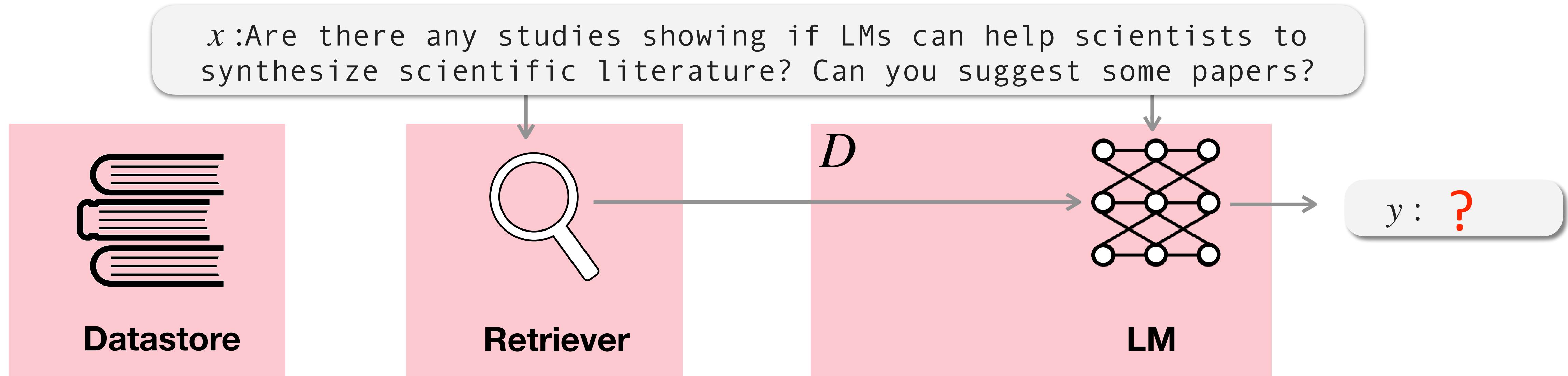


Small and general-domain



Limited generalization ability

Challenges in Common RAG Pipelines

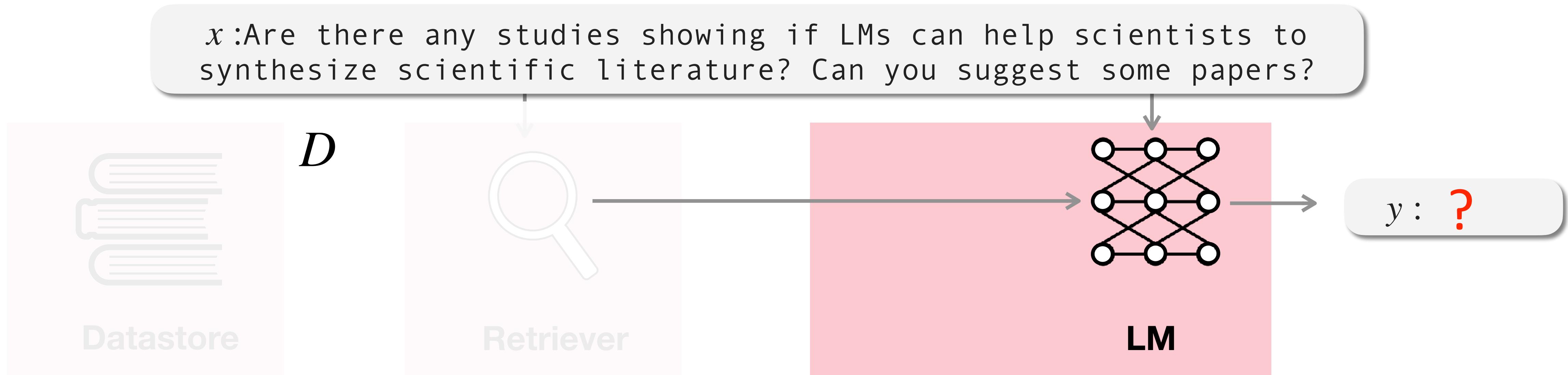


Small and general-domain



Limited generalization ability

Challenges in Common RAG Pipelines

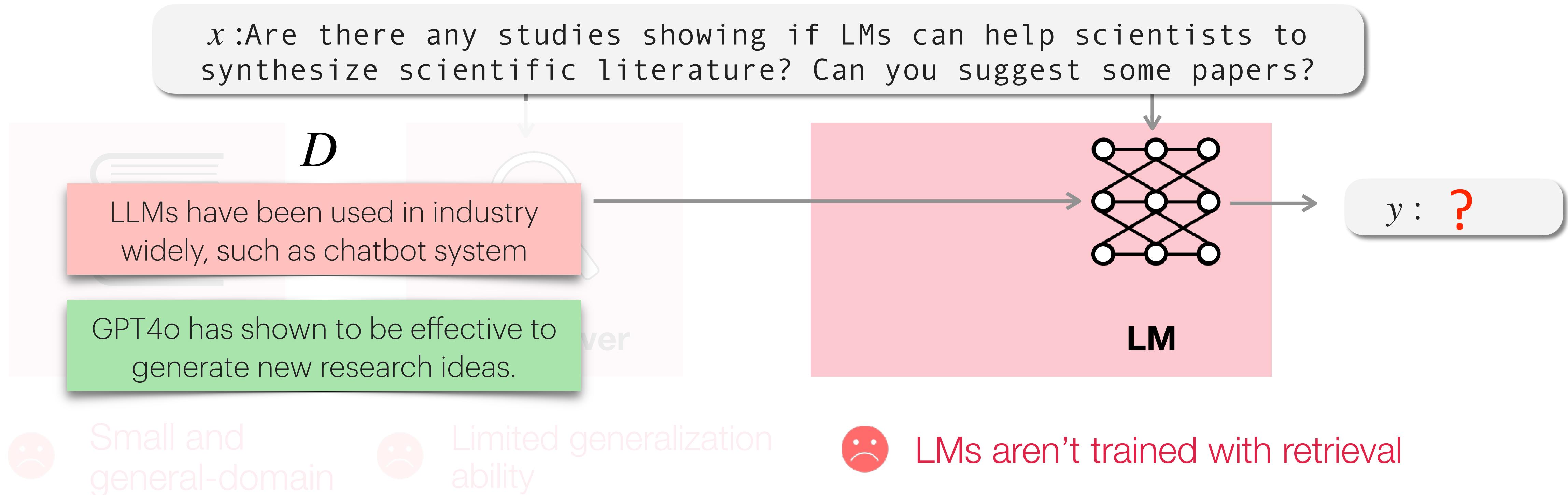


:(Small and general-domain

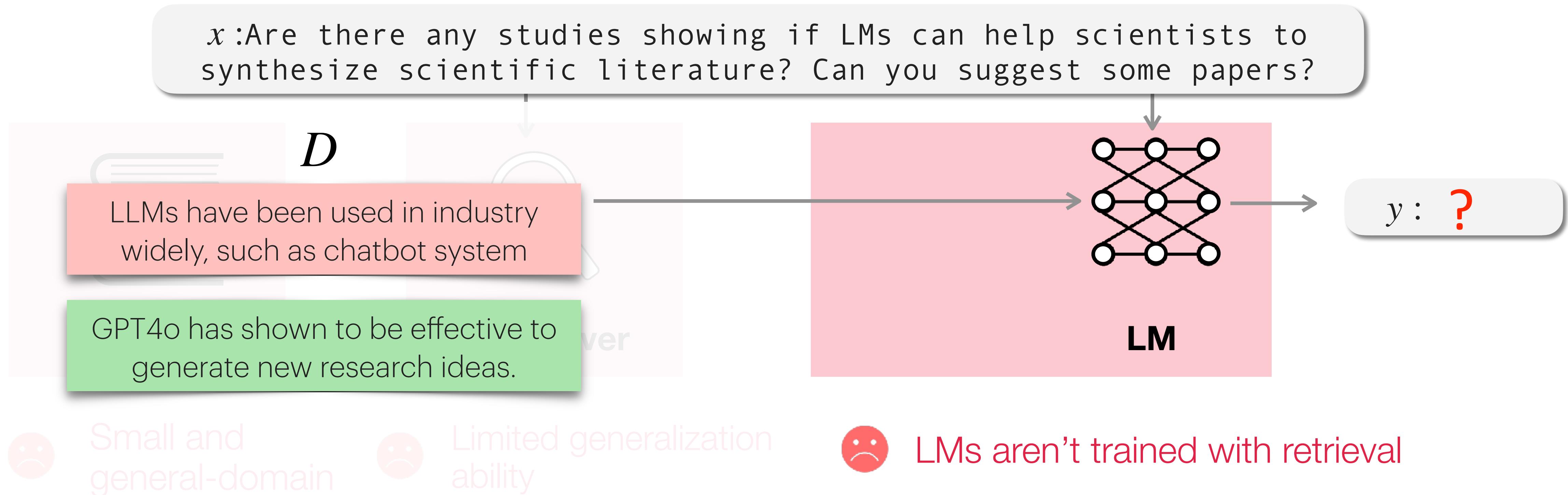
:(Limited generalization ability

:(LMs aren't trained with retrieval

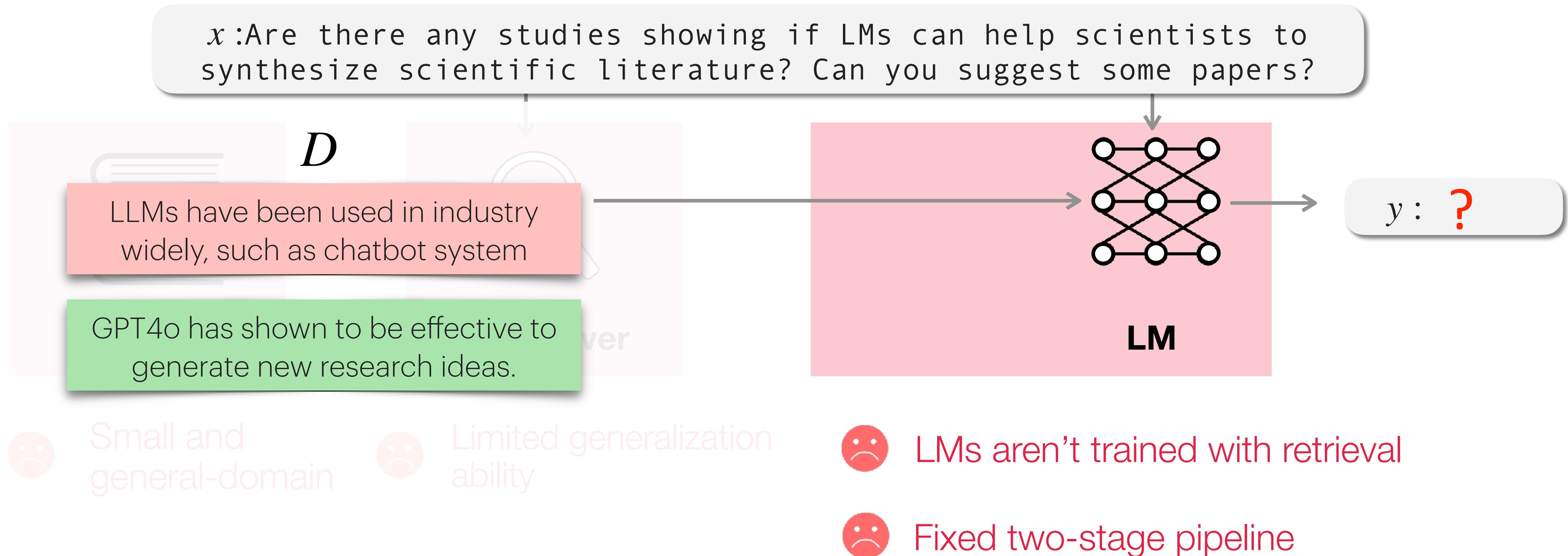
Challenges in Common RAG Pipelines



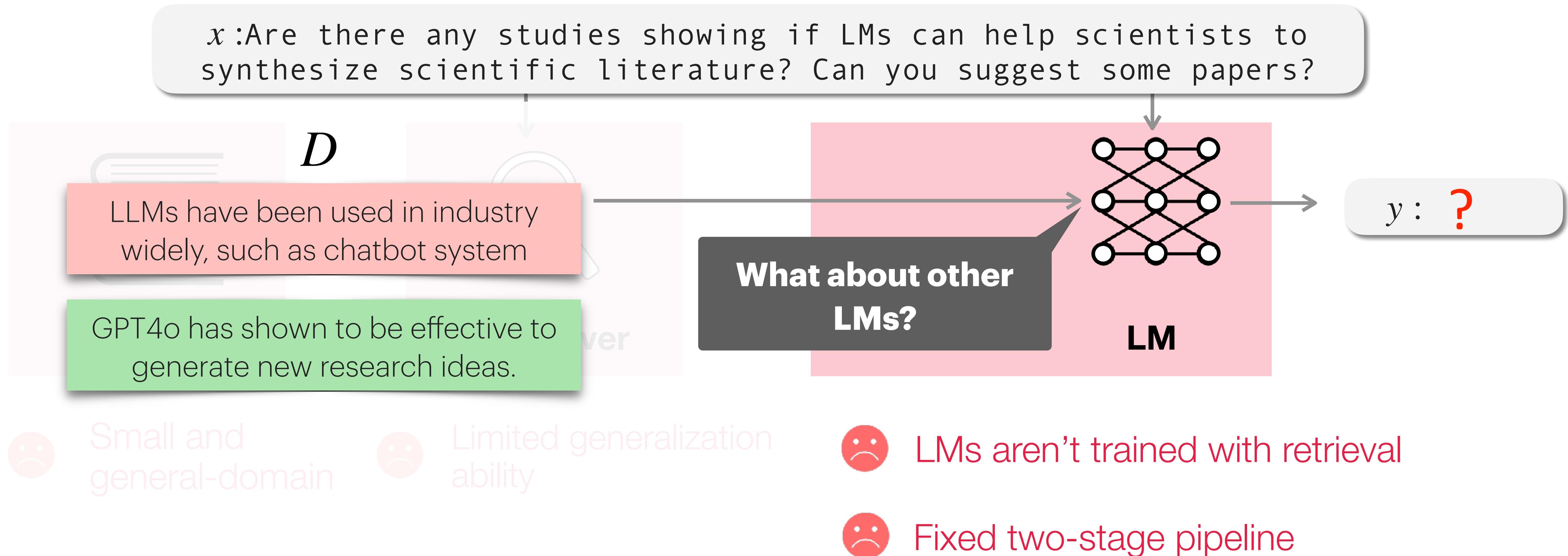
Challenges in Common RAG Pipelines



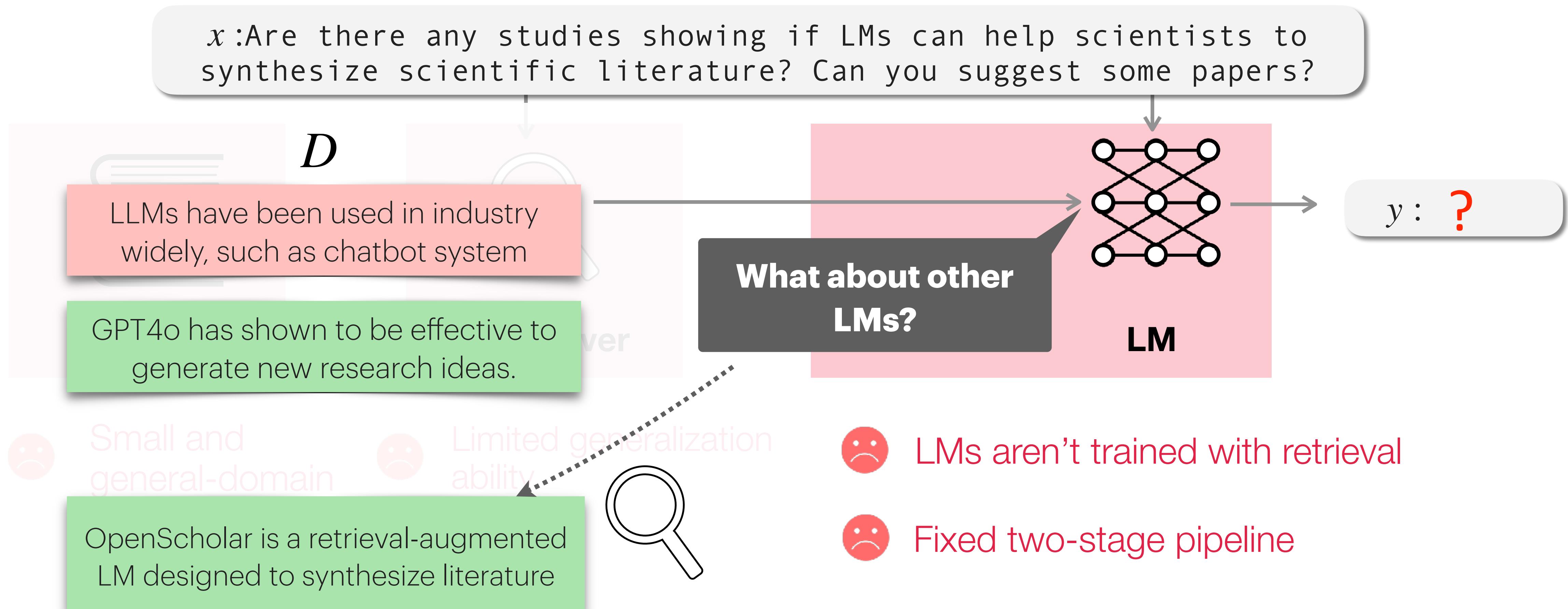
Challenges in Common RAG Pipelines



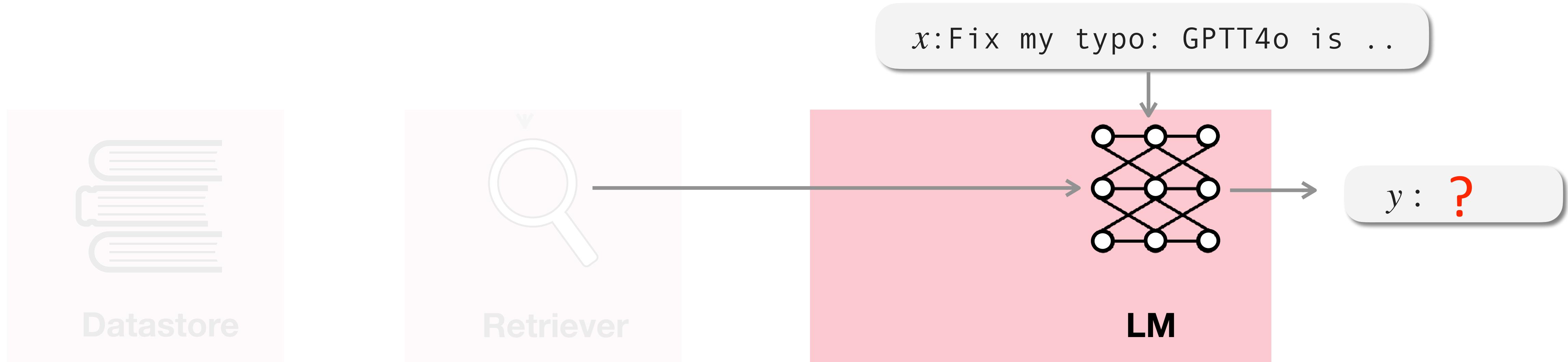
Challenges in Common RAG Pipelines



Challenges in Common RAG Pipelines

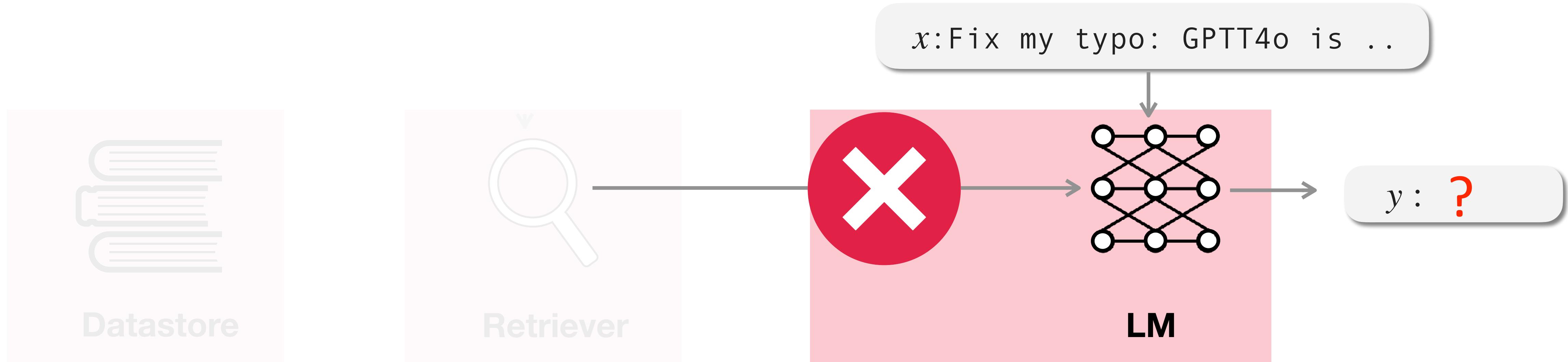


Challenges in Common RAG Pipelines



- :(Small, general-domain
- :(Limited generalization ability
- :(LMs aren't trained with retrieval
- :(Fixed two-stage pipeline

Challenges in Common RAG Pipelines



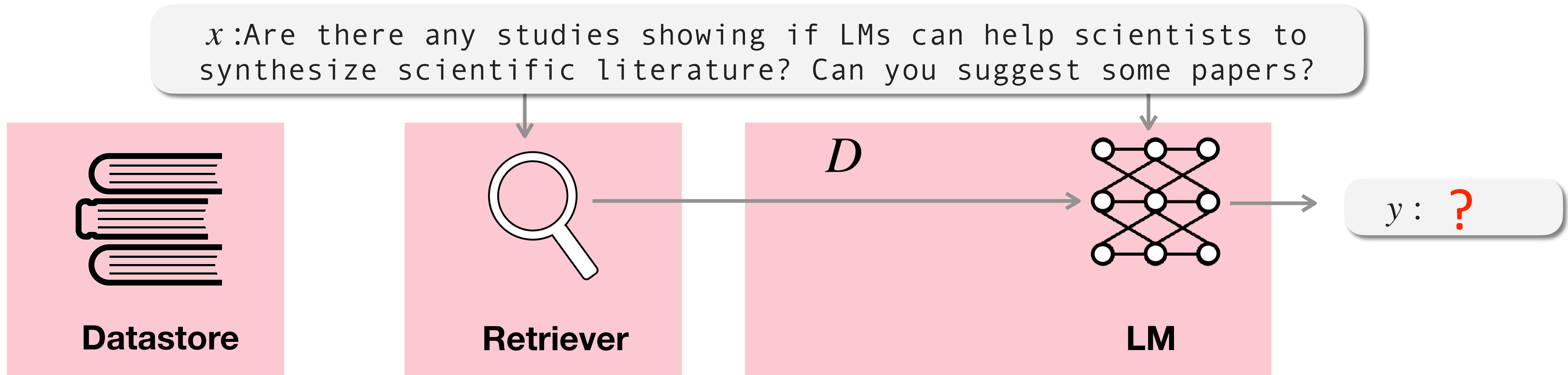
:(Small,
general-domain

:(Limited generalization
ability

:(LMs aren't trained with retrieval

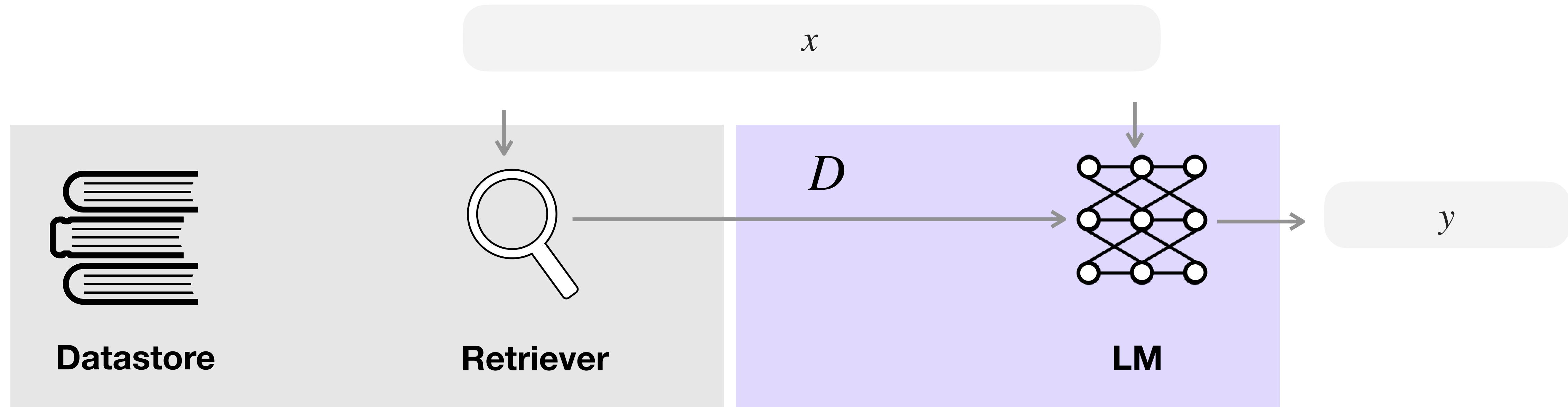
:(Fixed two-stage pipeline

Challenges in Common RAG Pipelines



- :(Small and general-domain
- :(Limited generalization ability
- :(LMs aren't trained with retrieval
- :(Fixed two-stage pipeline

Challenges in Common RAG Pipelines



Goal

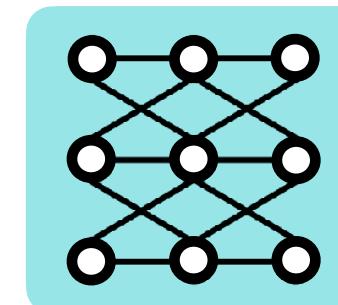
Develop a versatile and adaptive retrieval mechanism to enhance D

Goal

Design and train LMs to effectively and efficiently incorporate D

Self-RAG: Self-Reflective Retrieval-Augmented Generation

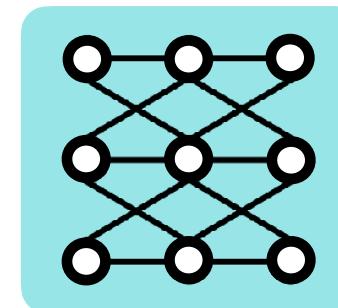
- :(LMs aren't trained with retrieval
- :(Fixed two-stage pipeline



Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop Most Influential ICLR 2024 Paper #13

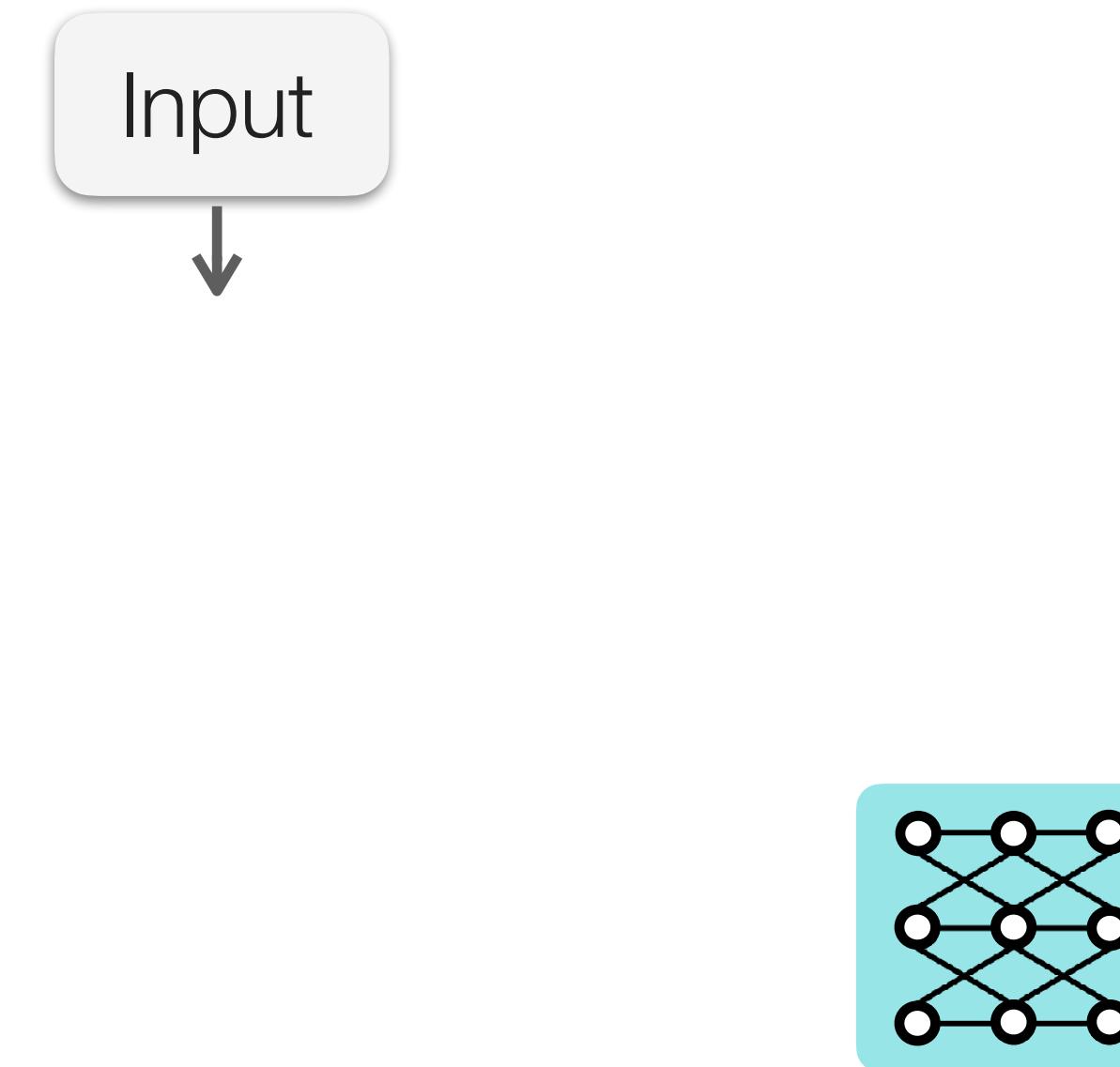
Self-RAG: Self-Reflective Retrieval-Augmented Generation



Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop Most Influential ICLR 2024 Paper #13

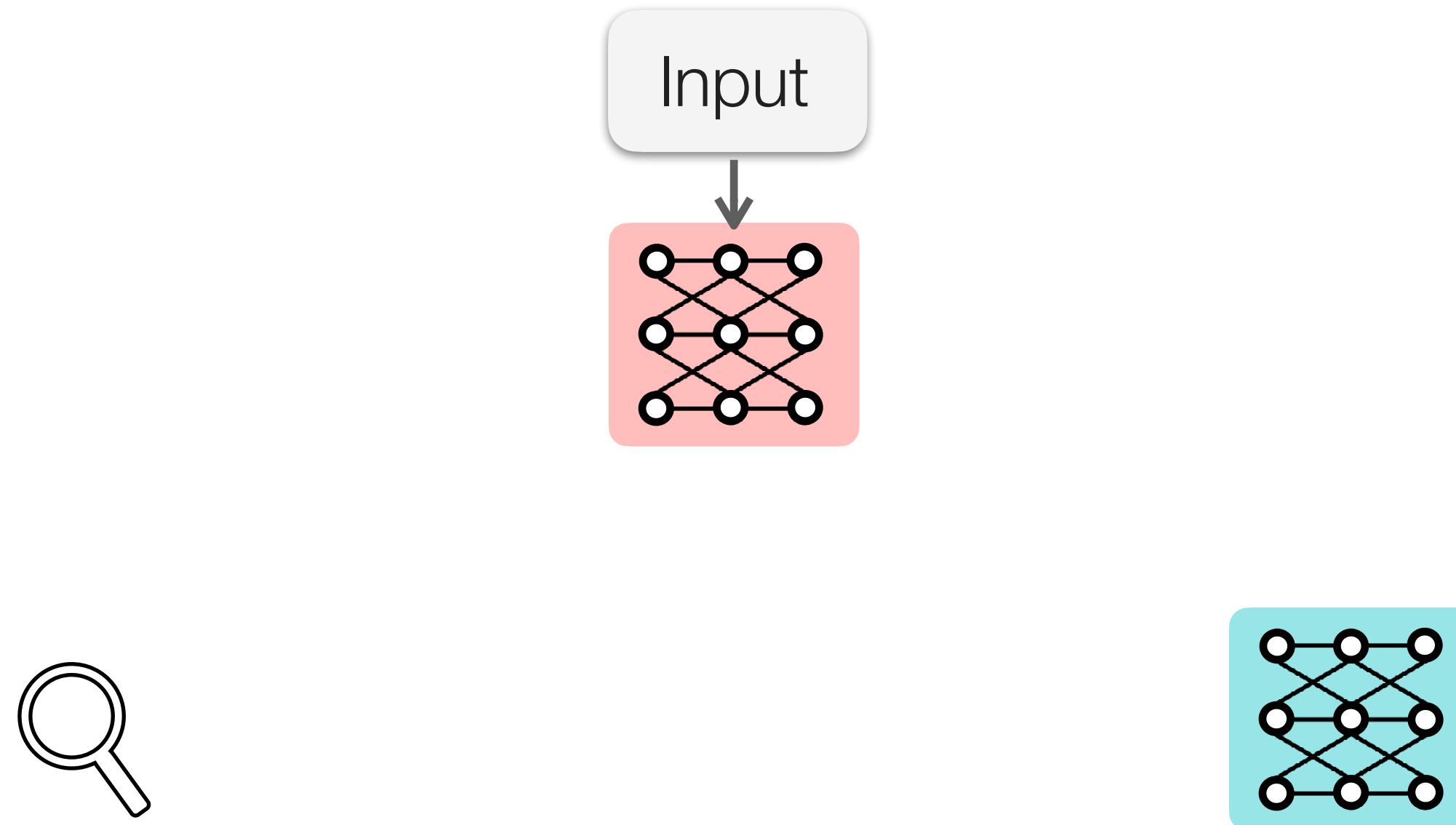
Self-RAG: Self-Reflective Retrieval-Augmented Generation



Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop Most Influential ICLR 2024 Paper #13

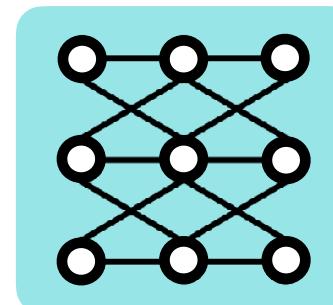
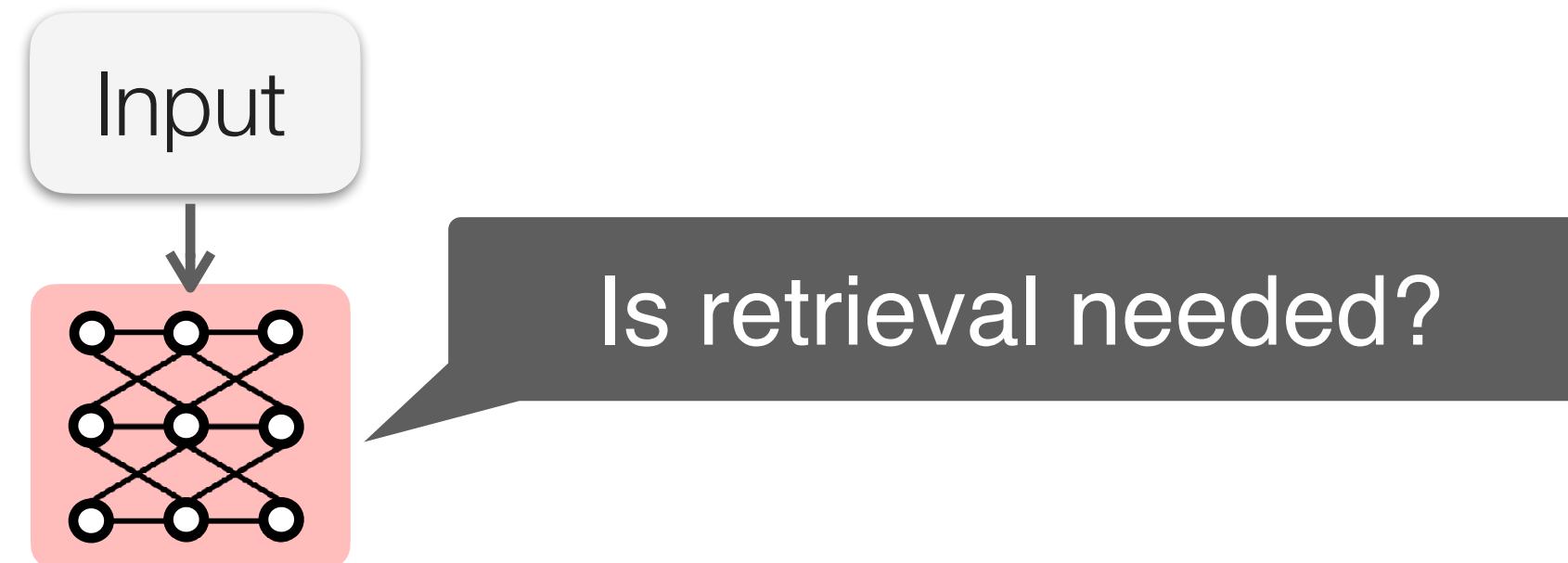
Self-RAG: Self-Reflective Retrieval-Augmented Generation



Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop Most Influential ICLR 2024 Paper #13

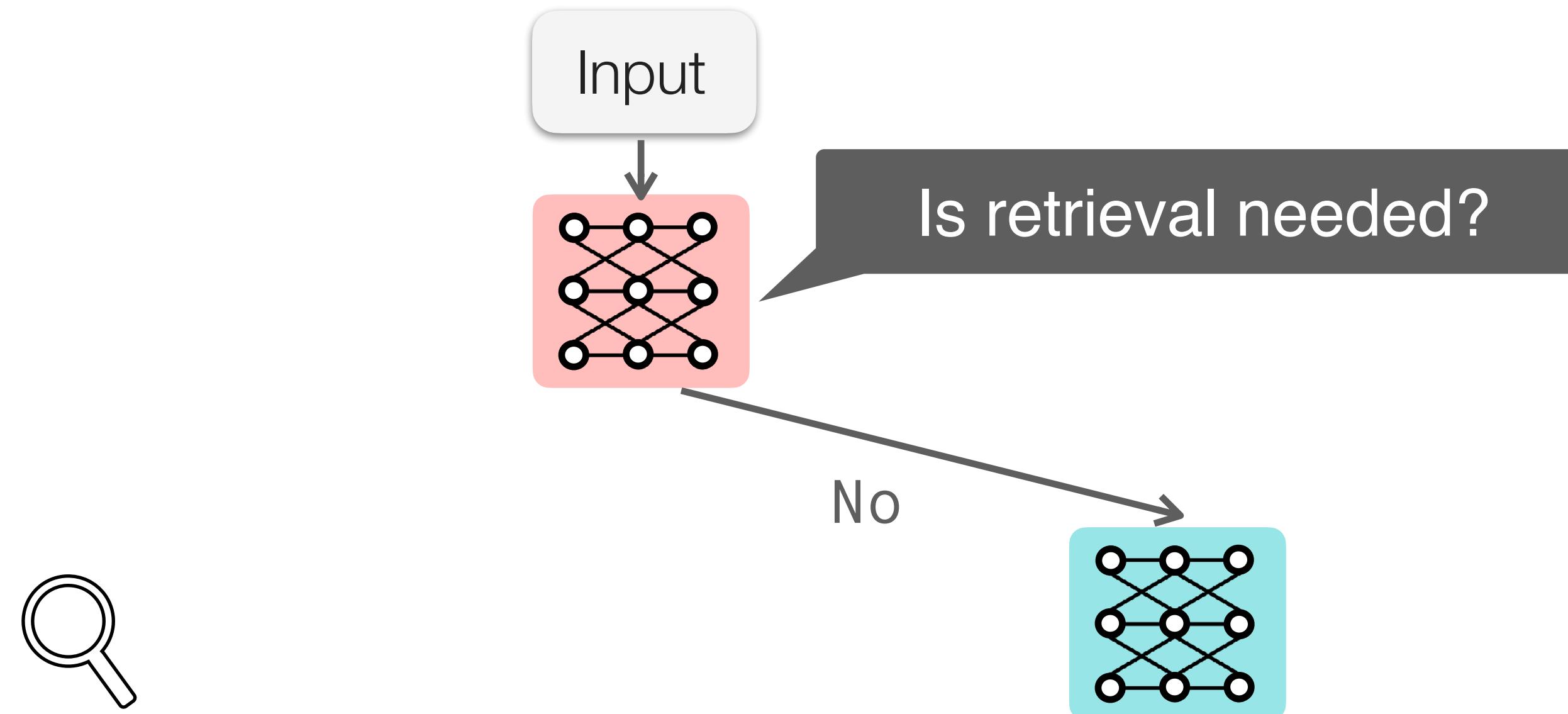
Self-RAG: Self-Reflective Retrieval-Augmented Generation



Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop Most Influential ICLR 2024 Paper #13

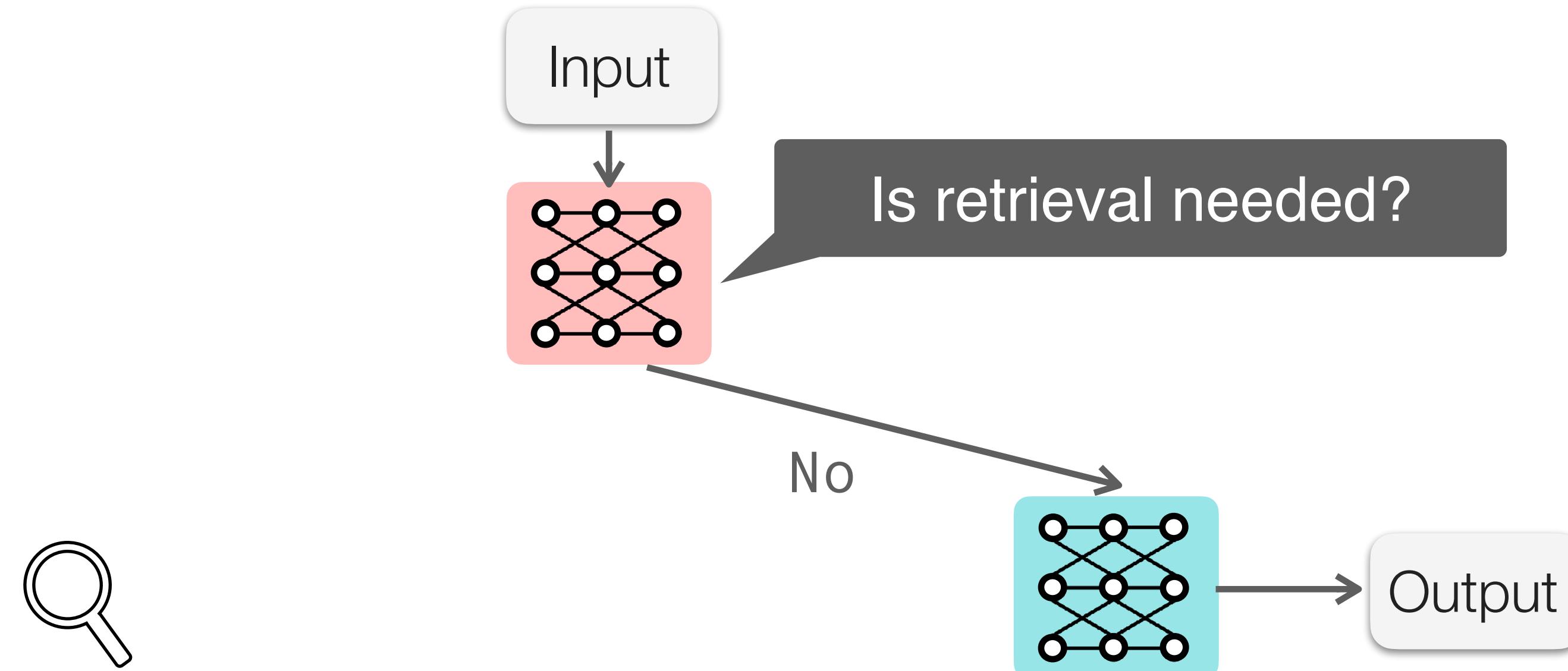
Self-RAG: Self-Reflective Retrieval-Augmented Generation



Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop Most Influential ICLR 2024 Paper #13

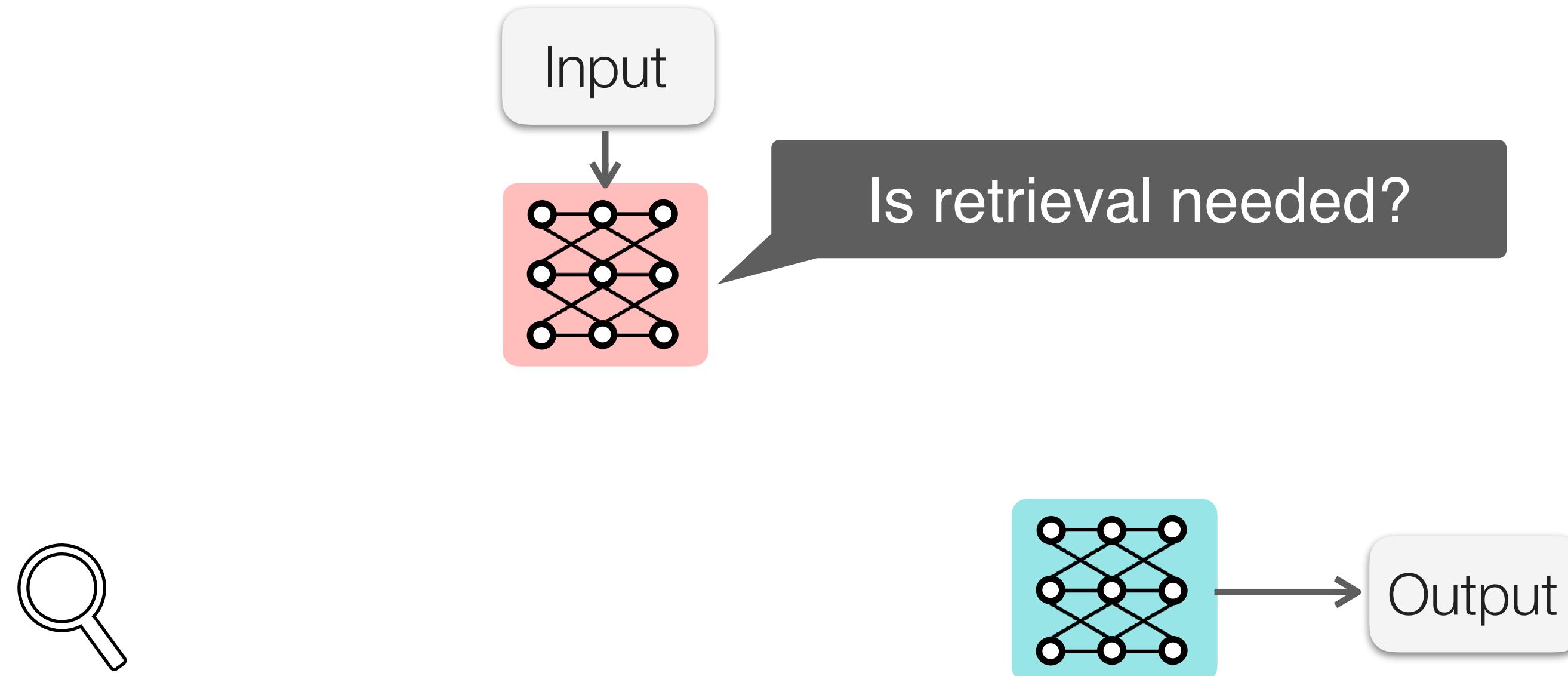
Self-RAG: Self-Reflective Retrieval-Augmented Generation



Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop Most Influential ICLR 2024 Paper #13

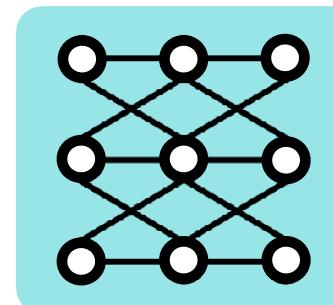
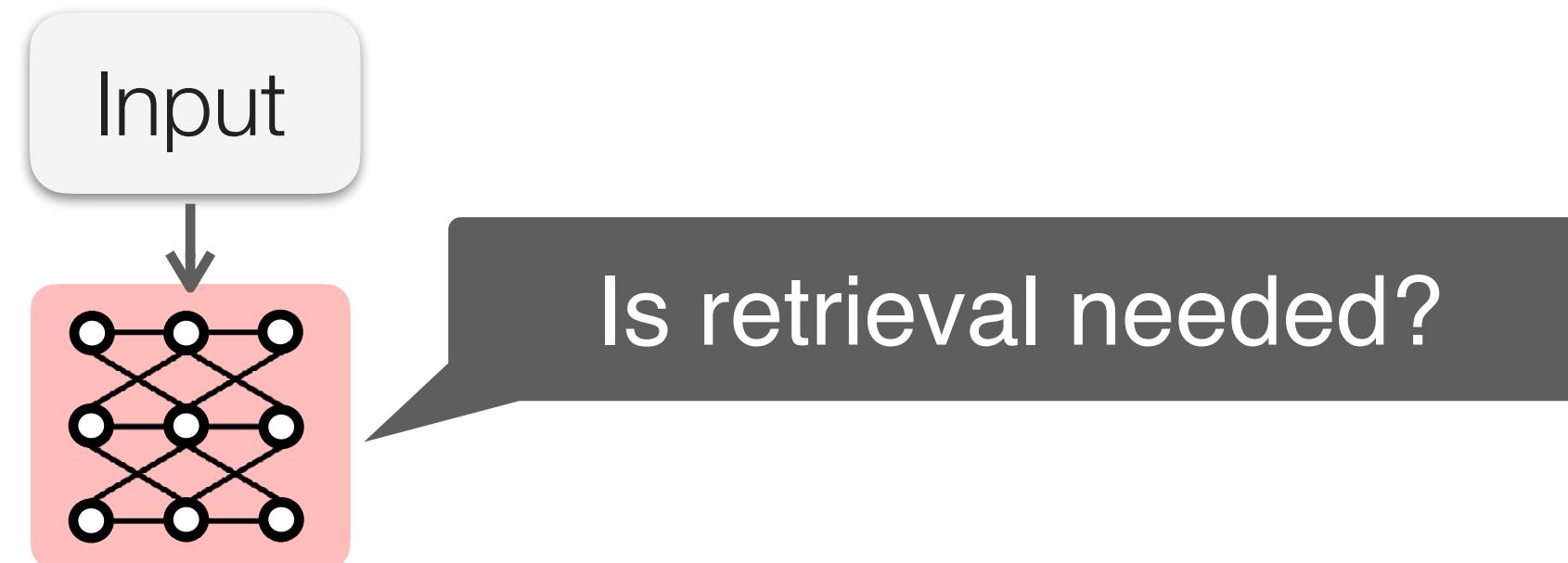
Self-RAG: Self-Reflective Retrieval-Augmented Generation



Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop Most Influential ICLR 2024 Paper #13

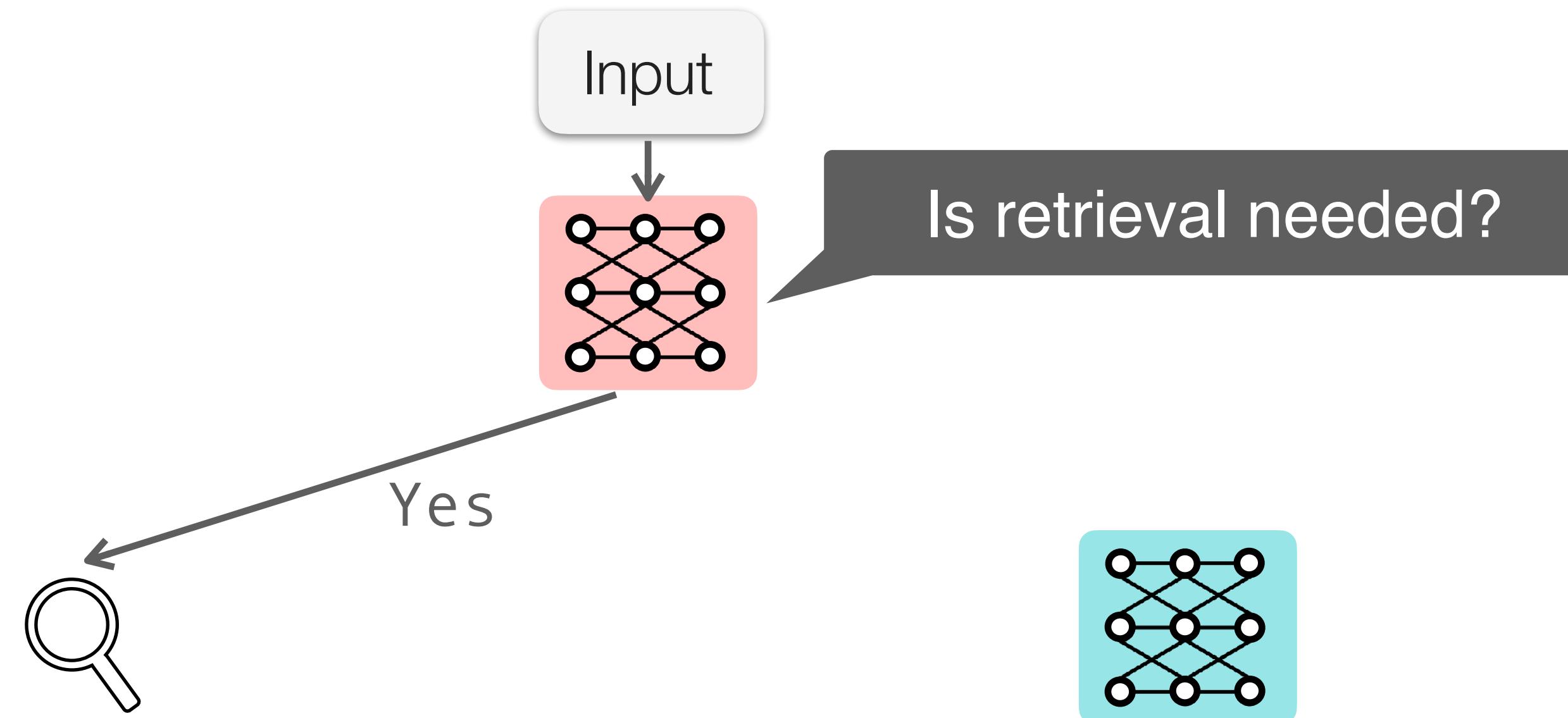
Self-RAG: Self-Reflective Retrieval-Augmented Generation



Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop Most Influential ICLR 2024 Paper #13

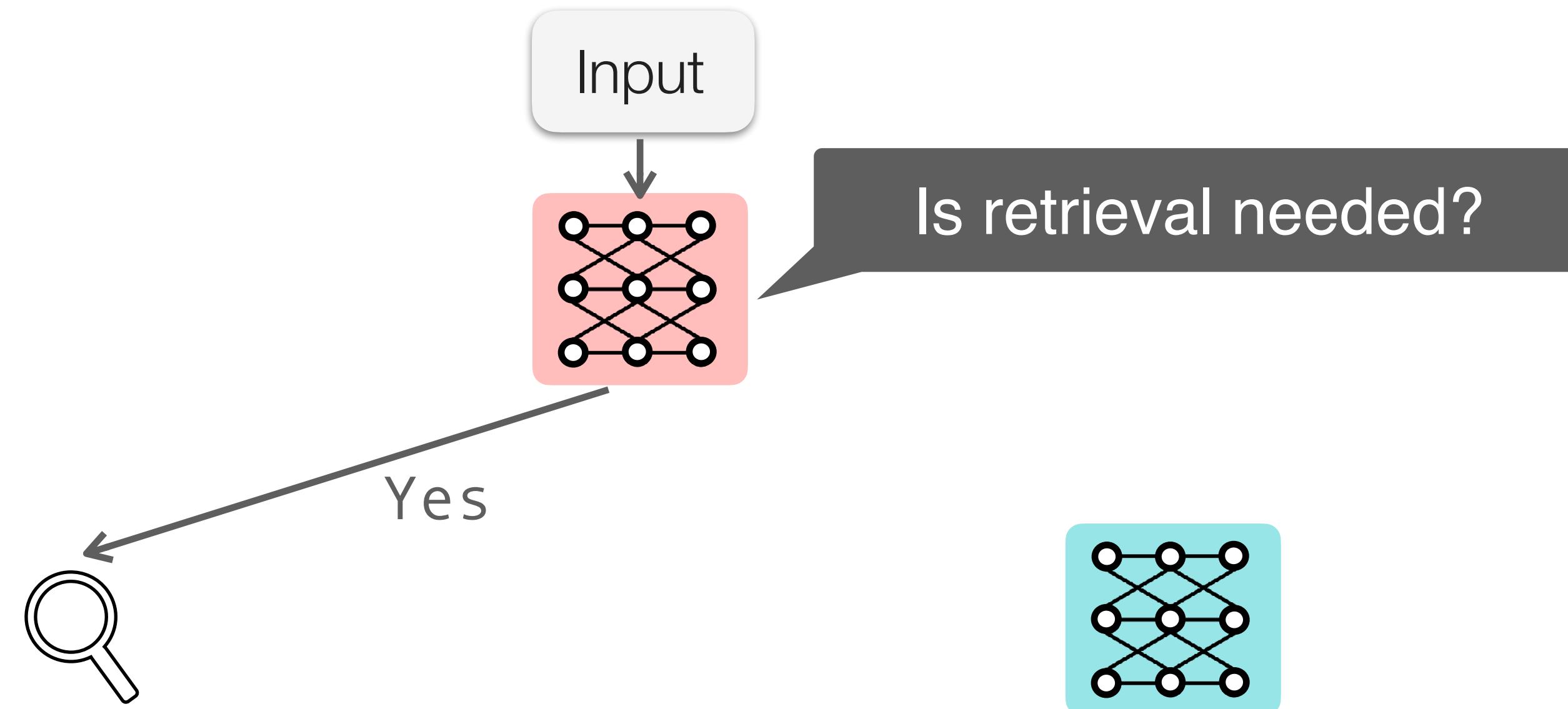
Self-RAG: Self-Reflective Retrieval-Augmented Generation



Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop Most Influential ICLR 2024 Paper #13

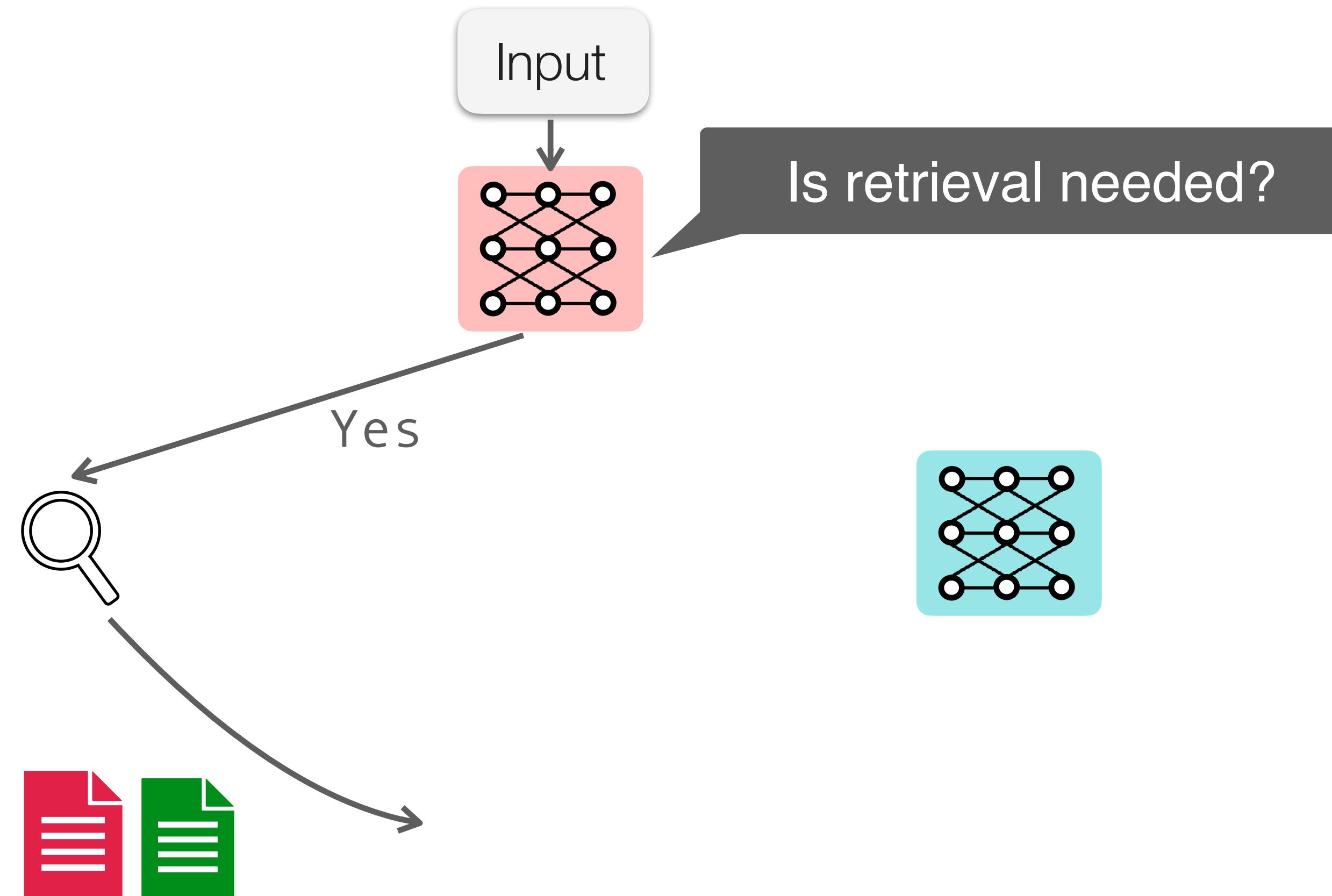
Self-RAG: Self-Reflective Retrieval-Augmented Generation



Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop Most Influential ICLR 2024 Paper #13

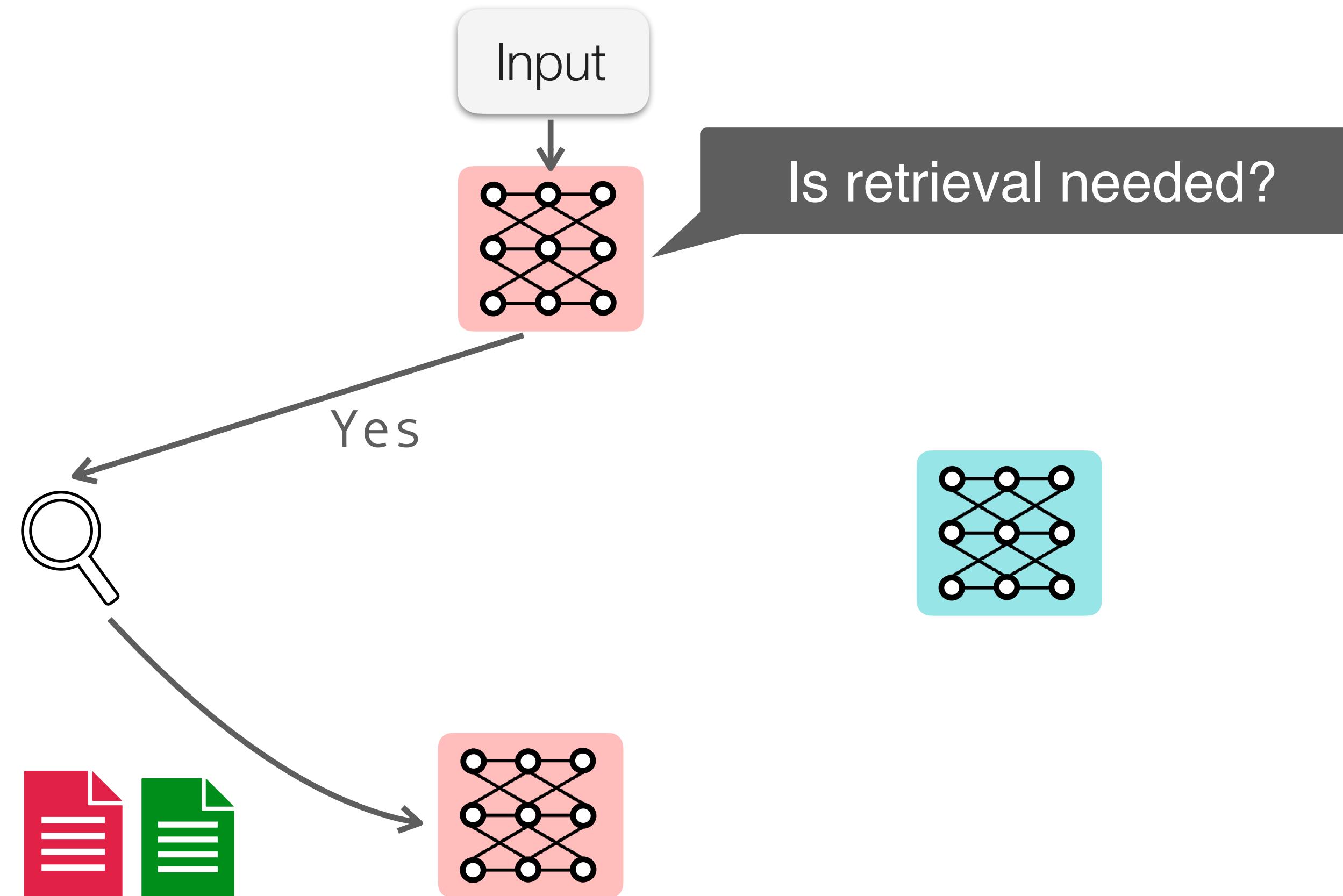
Self-RAG: Self-Reflective Retrieval-Augmented Generation



Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop Most Influential ICLR 2024 Paper #13

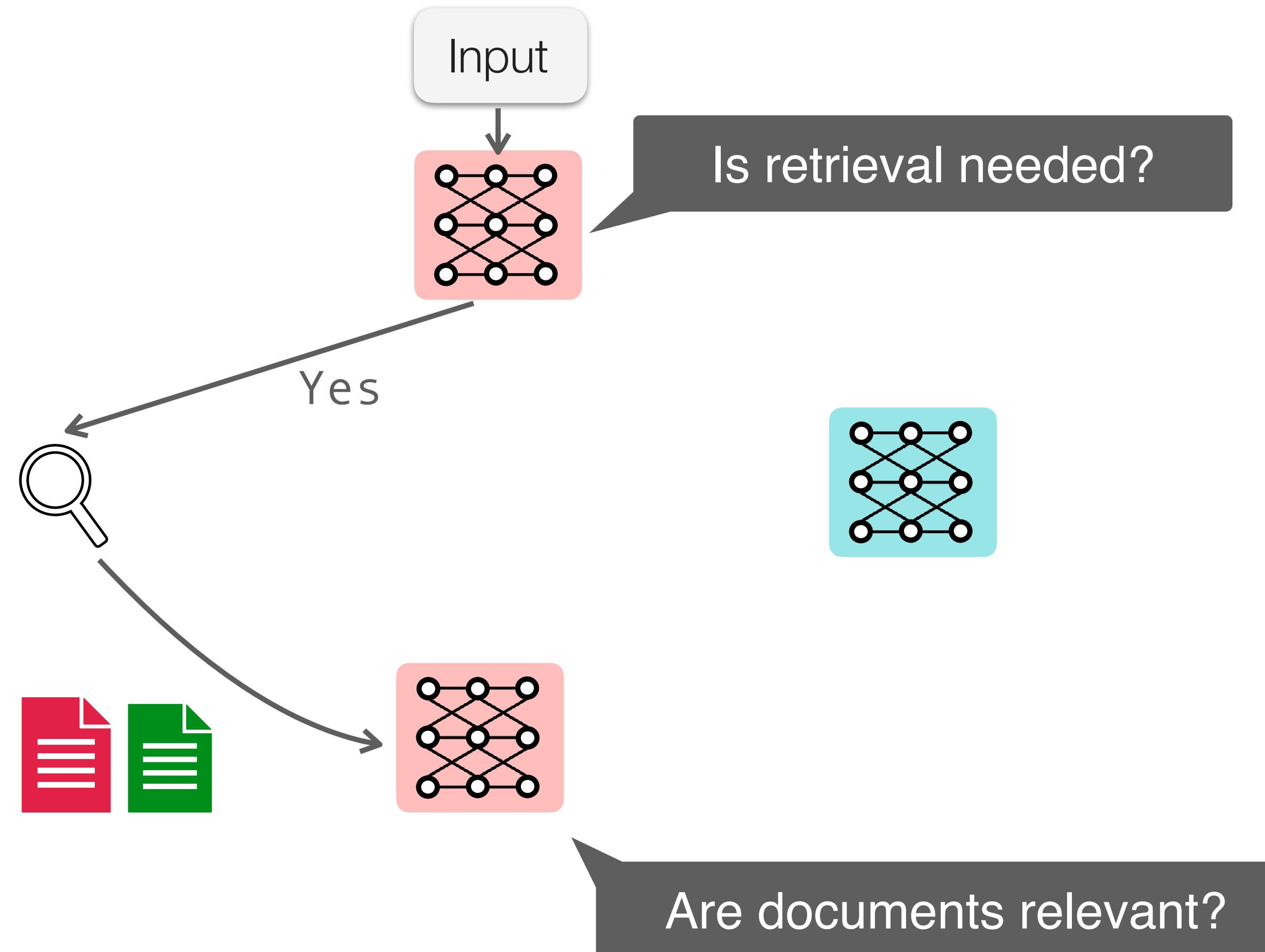
Self-RAG: Self-Reflective Retrieval-Augmented Generation



Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop Most Influential ICLR 2024 Paper #13

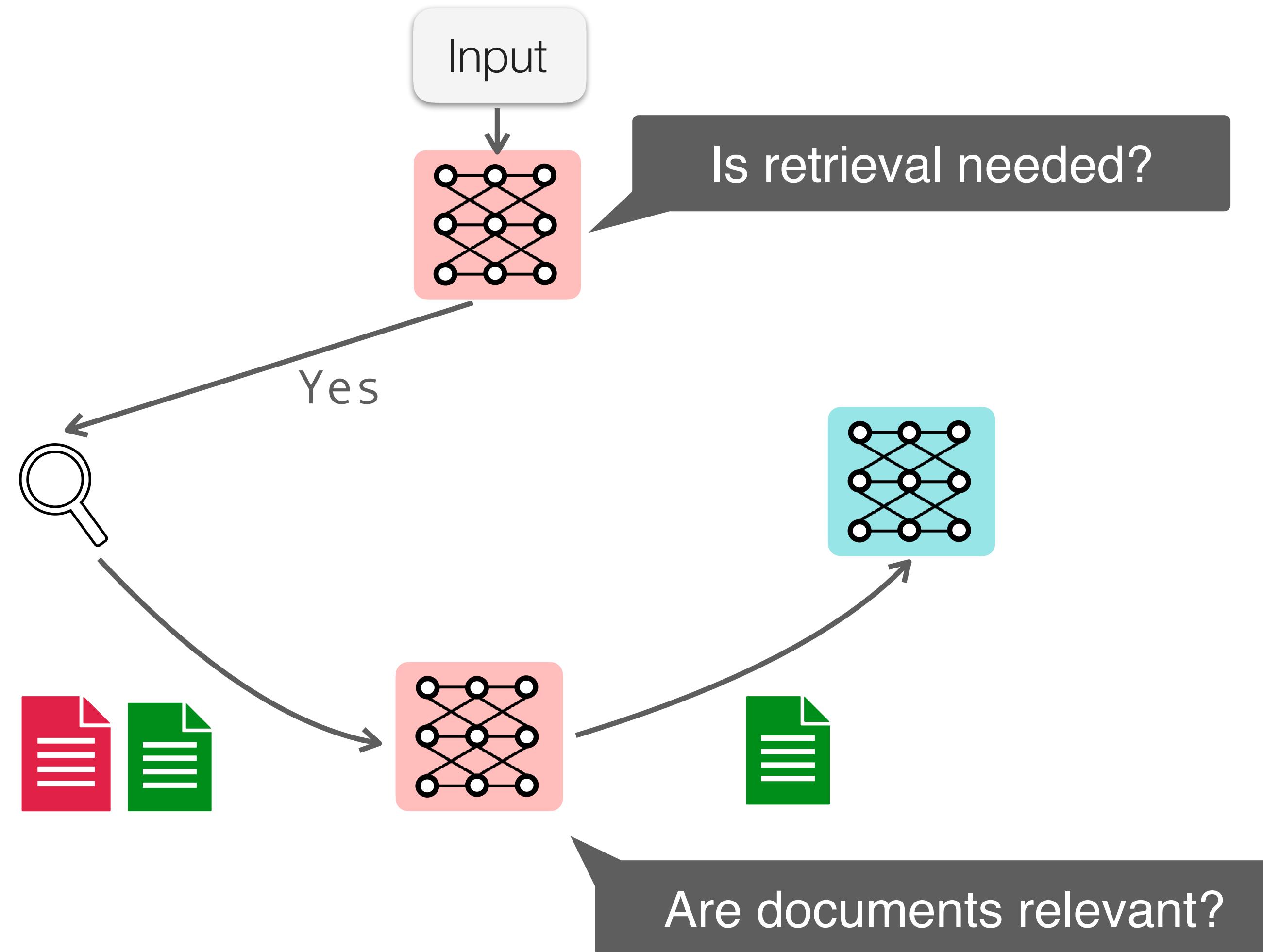
Self-RAG: Self-Reflective Retrieval-Augmented Generation



Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop Most Influential ICLR 2024 Paper #13

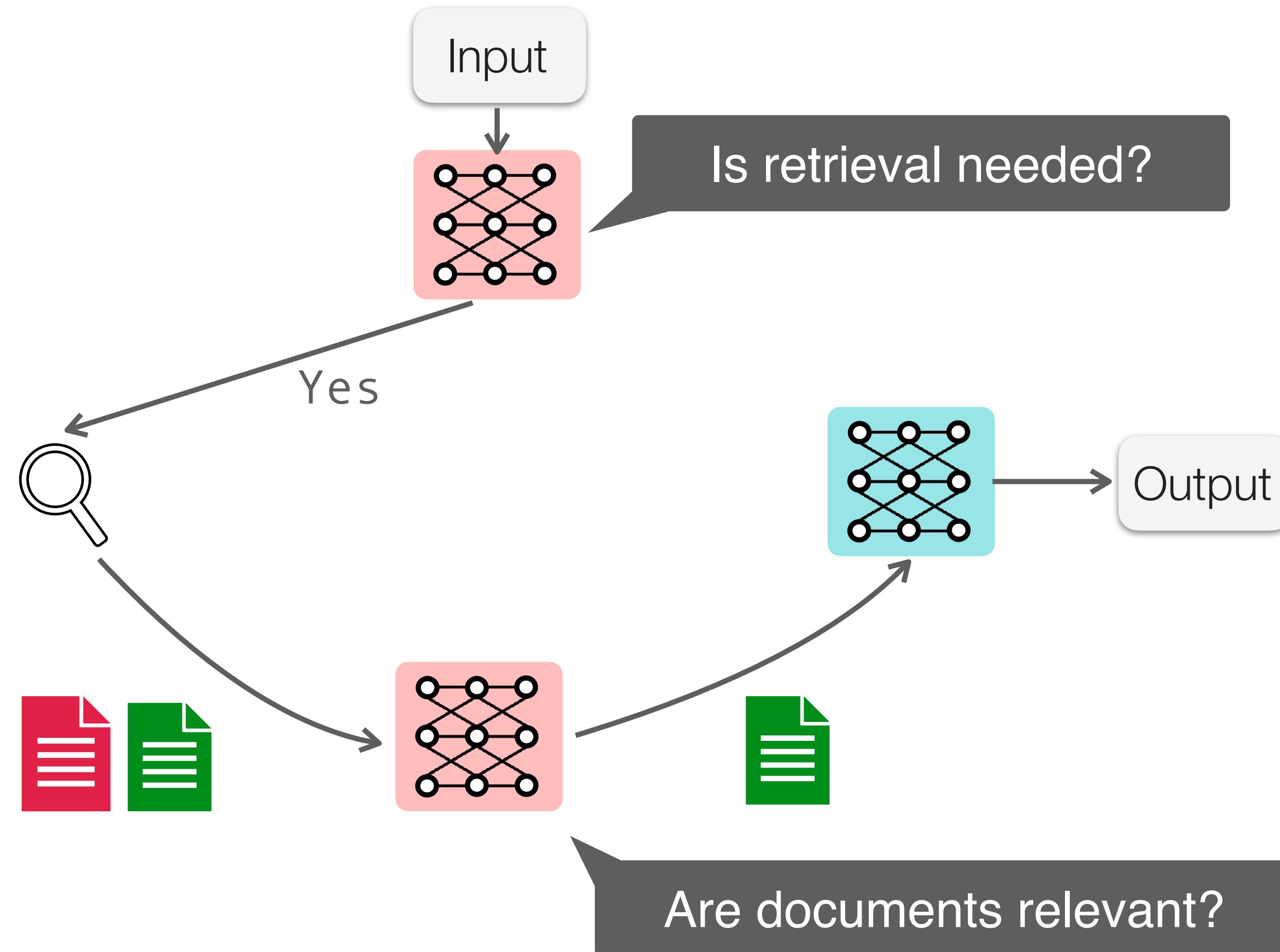
Self-RAG: Self-Reflective Retrieval-Augmented Generation



Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop Most Influential ICLR 2024 Paper #13

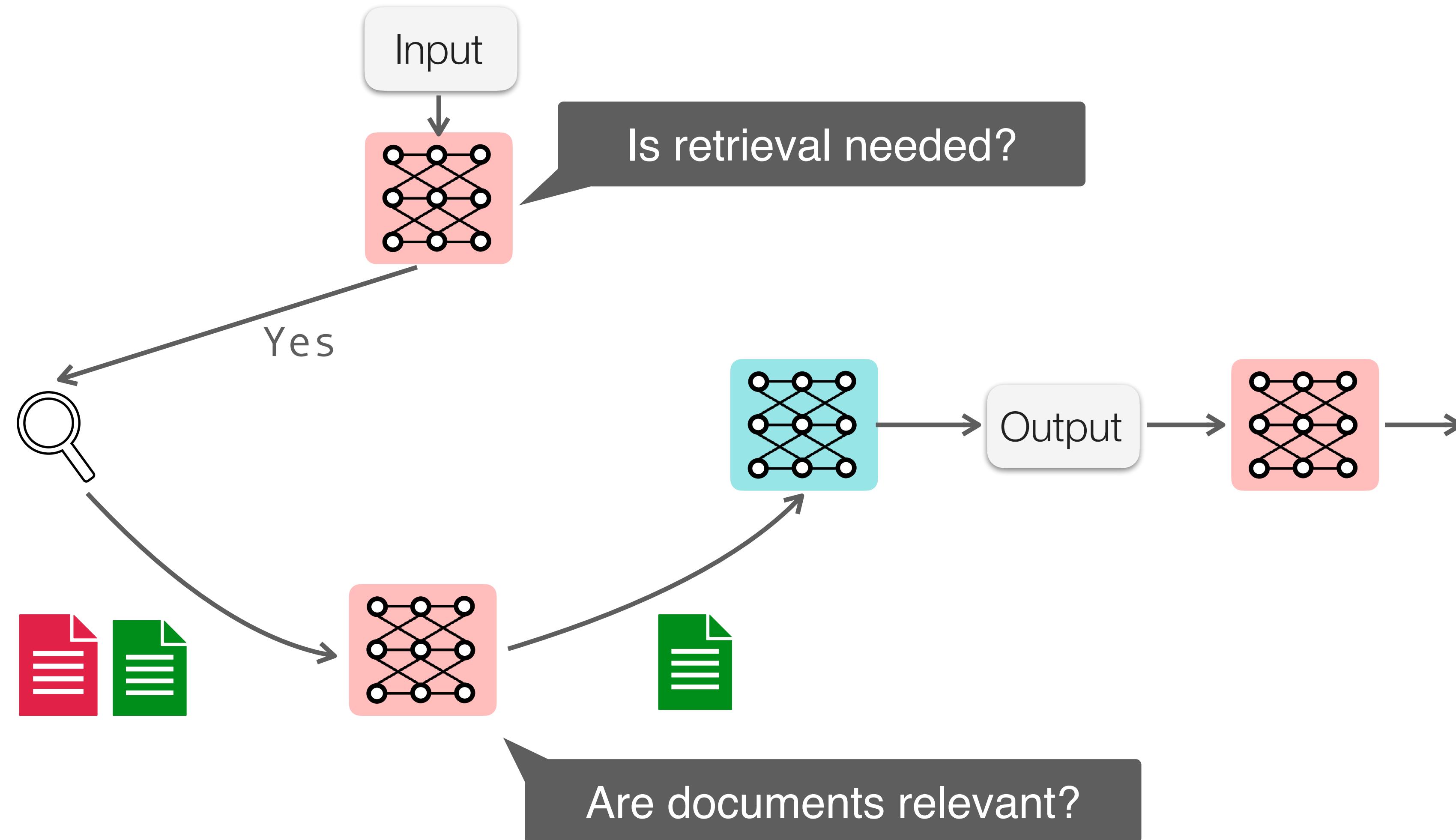
Self-RAG: Self-Reflective Retrieval-Augmented Generation



Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop Most Influential ICLR 2024 Paper #13

Self-RAG: Self-Reflective Retrieval-Augmented Generation

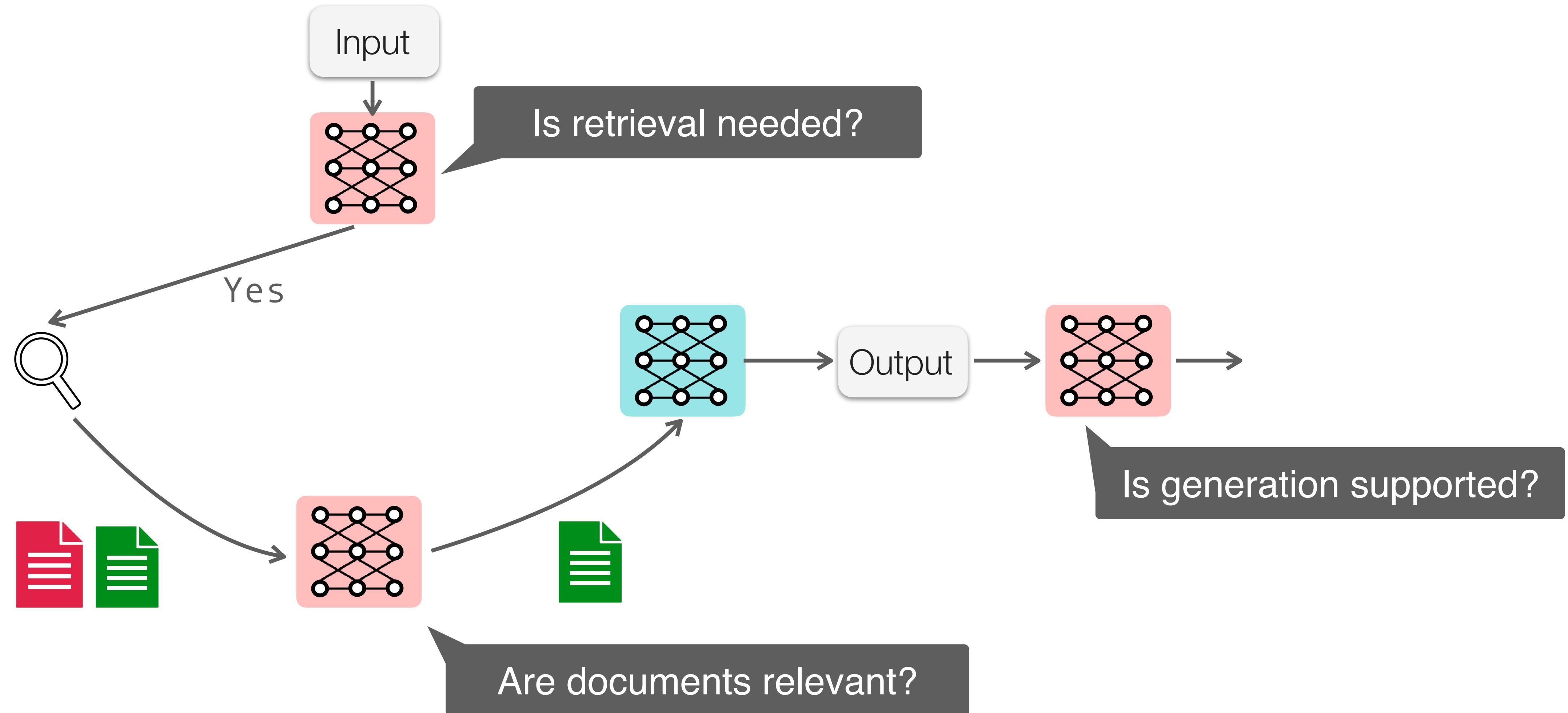


Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop

Most Influential ICLR 2024 Paper #13

Self-RAG: Self-Reflective Retrieval-Augmented Generation

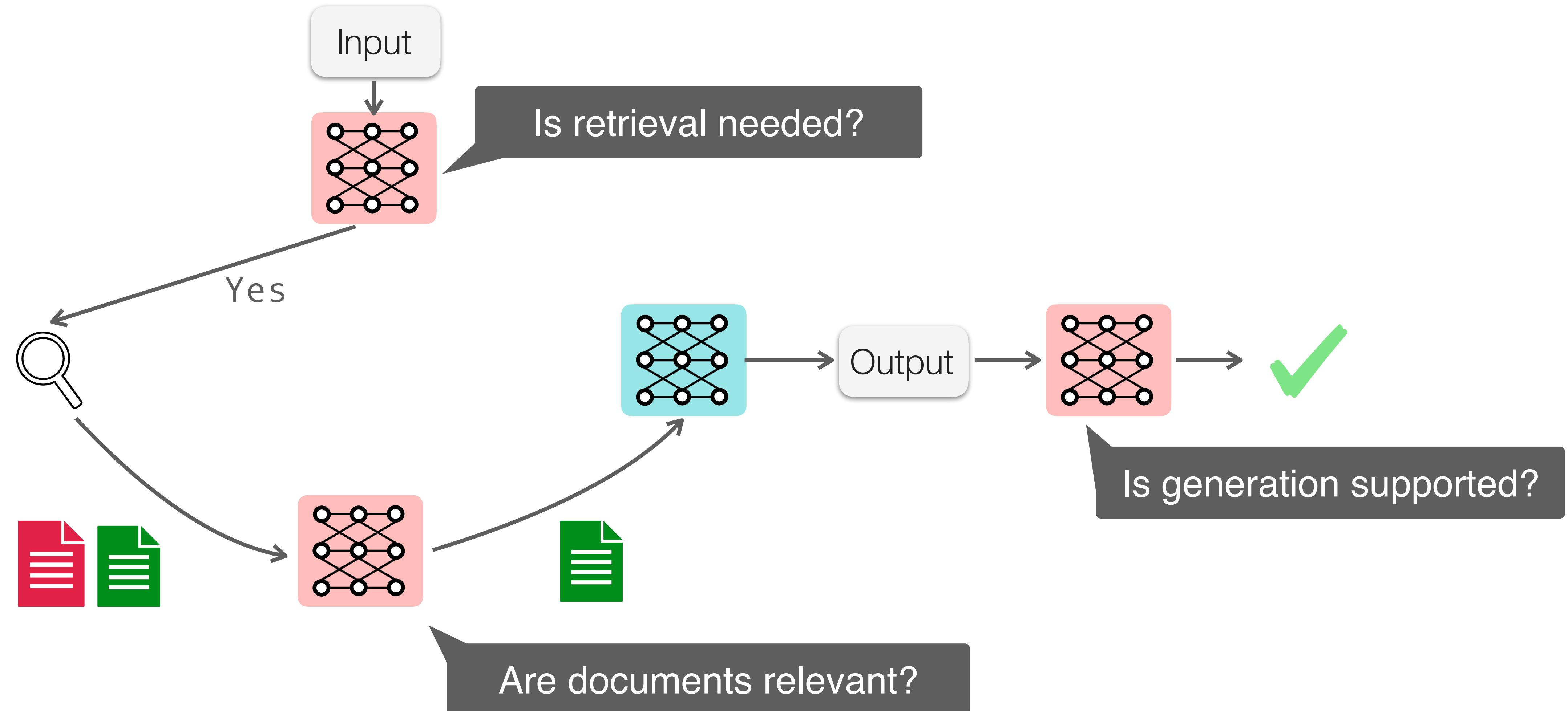


Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop

Most Influential ICLR 2024 Paper #13

Self-RAG: Self-Reflective Retrieval-Augmented Generation



Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

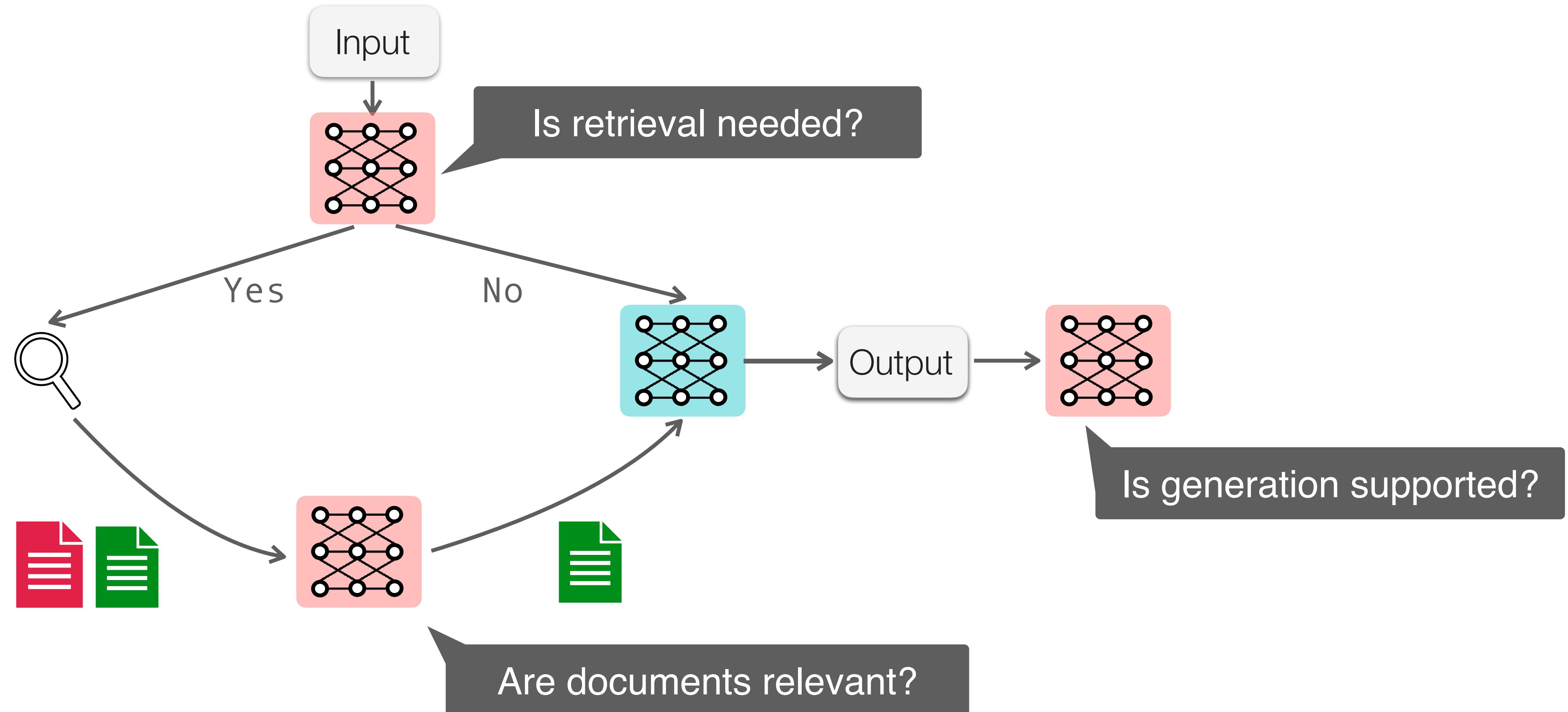
Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop



Most Influential ICLR 2024 Paper #13



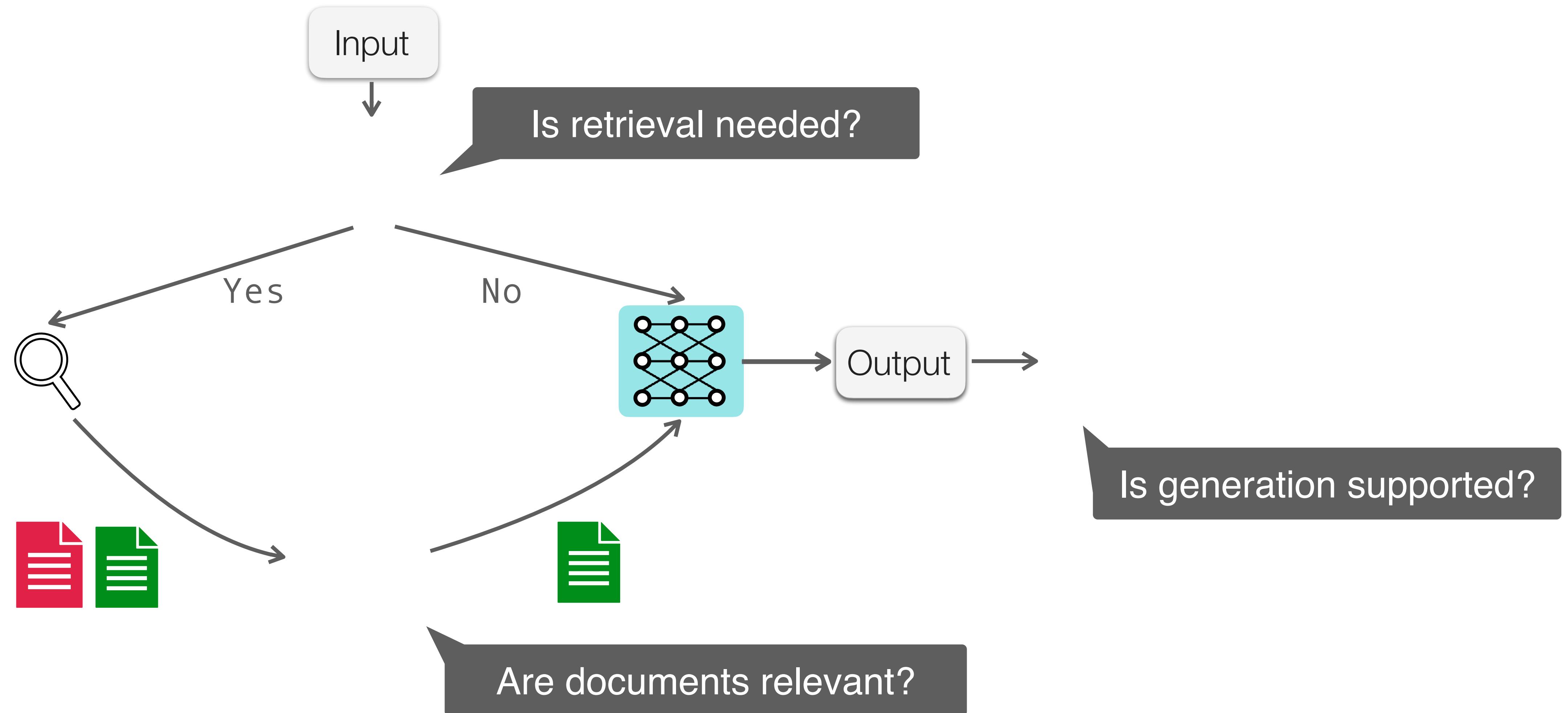
Self-RAG: Self-Reflective Retrieval-Augmented Generation



Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop Most Influential ICLR 2024 Paper #13

Self-RAG: Self-Reflective Retrieval-Augmented Generation

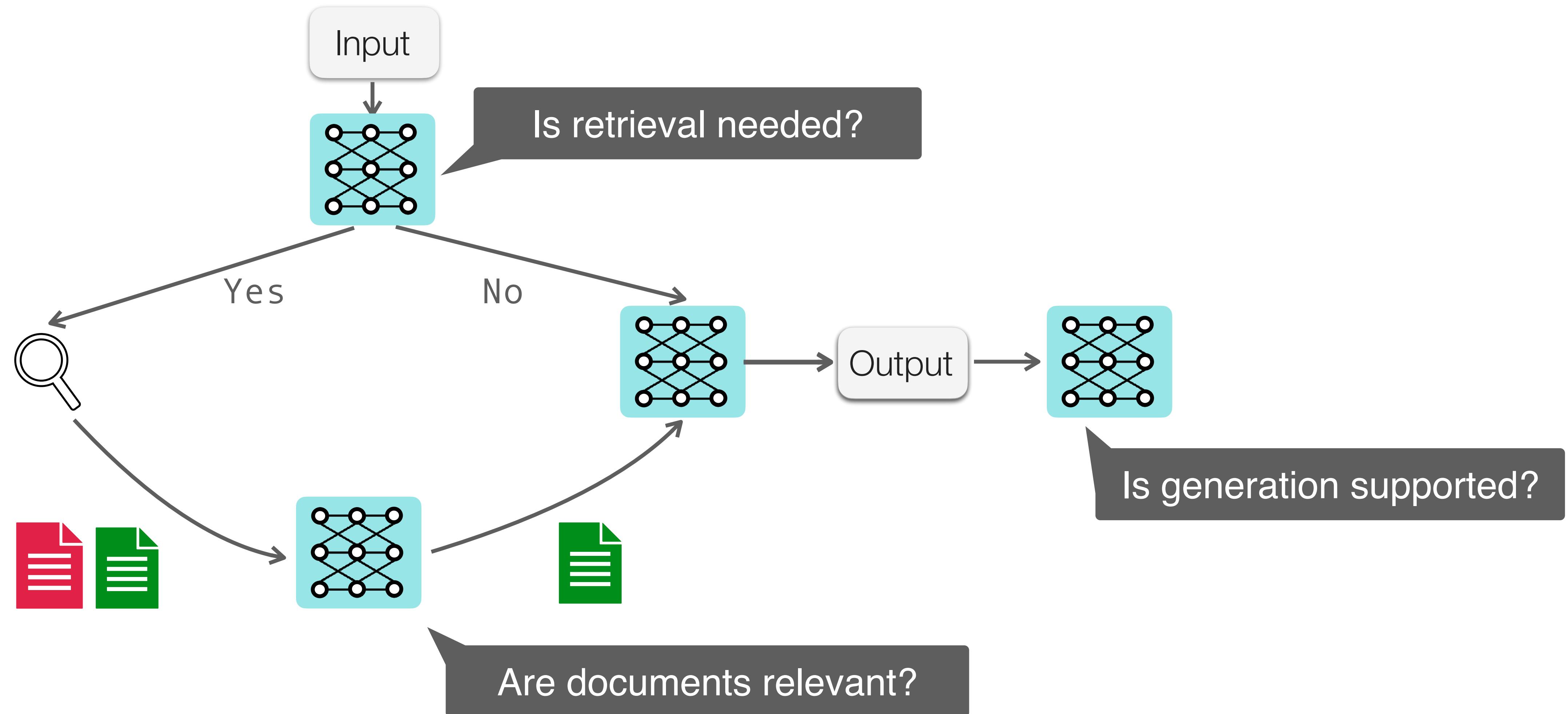


Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop

Most Influential ICLR 2024 Paper #13

Self-RAG: Self-Reflective Retrieval-Augmented Generation

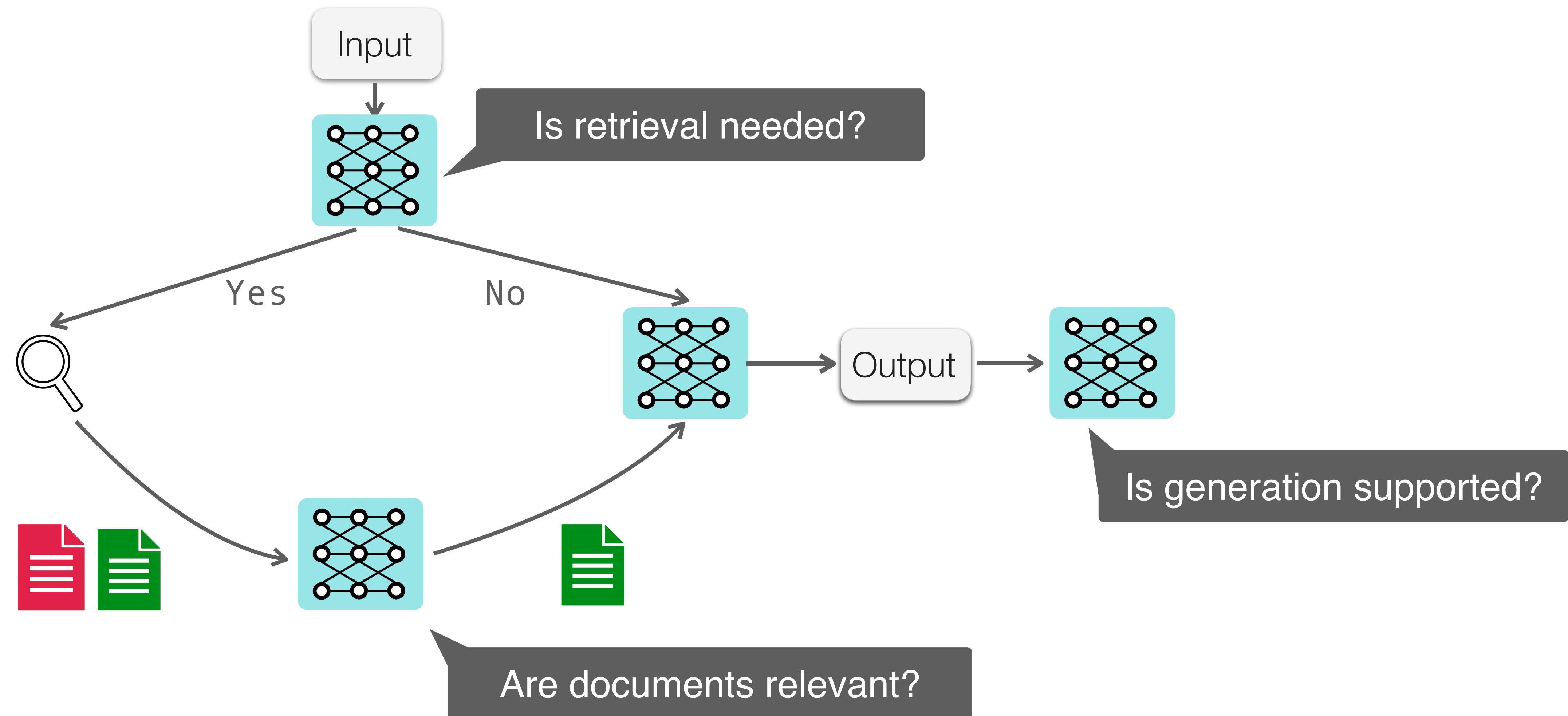


Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop

Honorable mention at NeurIPS Instruction Tuning Workshop  Most Influential ICLR 2024 Paper #13 

Self-RAG: Self-Reflective Retrieval-Augmented Generation



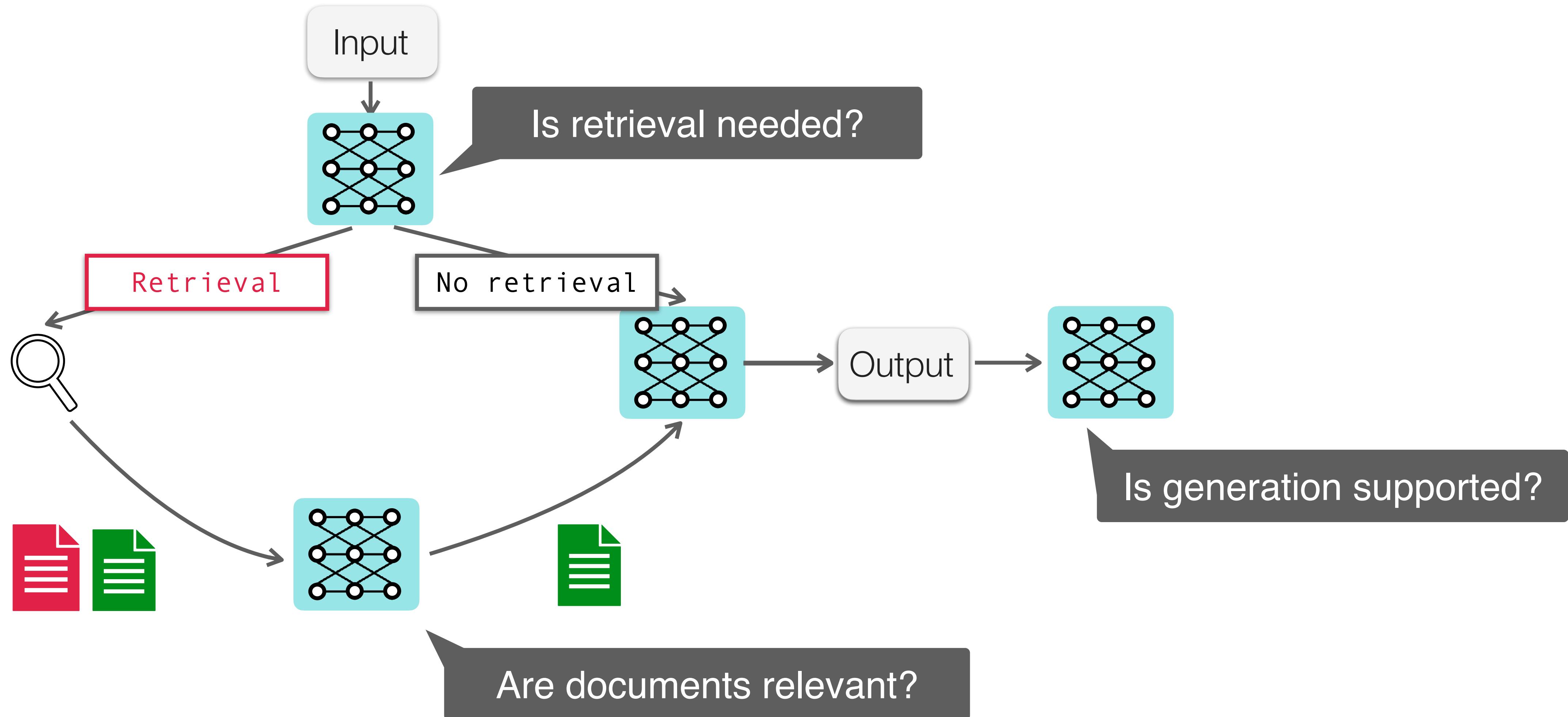
Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop 



Most Influential ICLR 2024 Paper #13

Self-RAG: Self-Reflective Retrieval-Augmented Generation



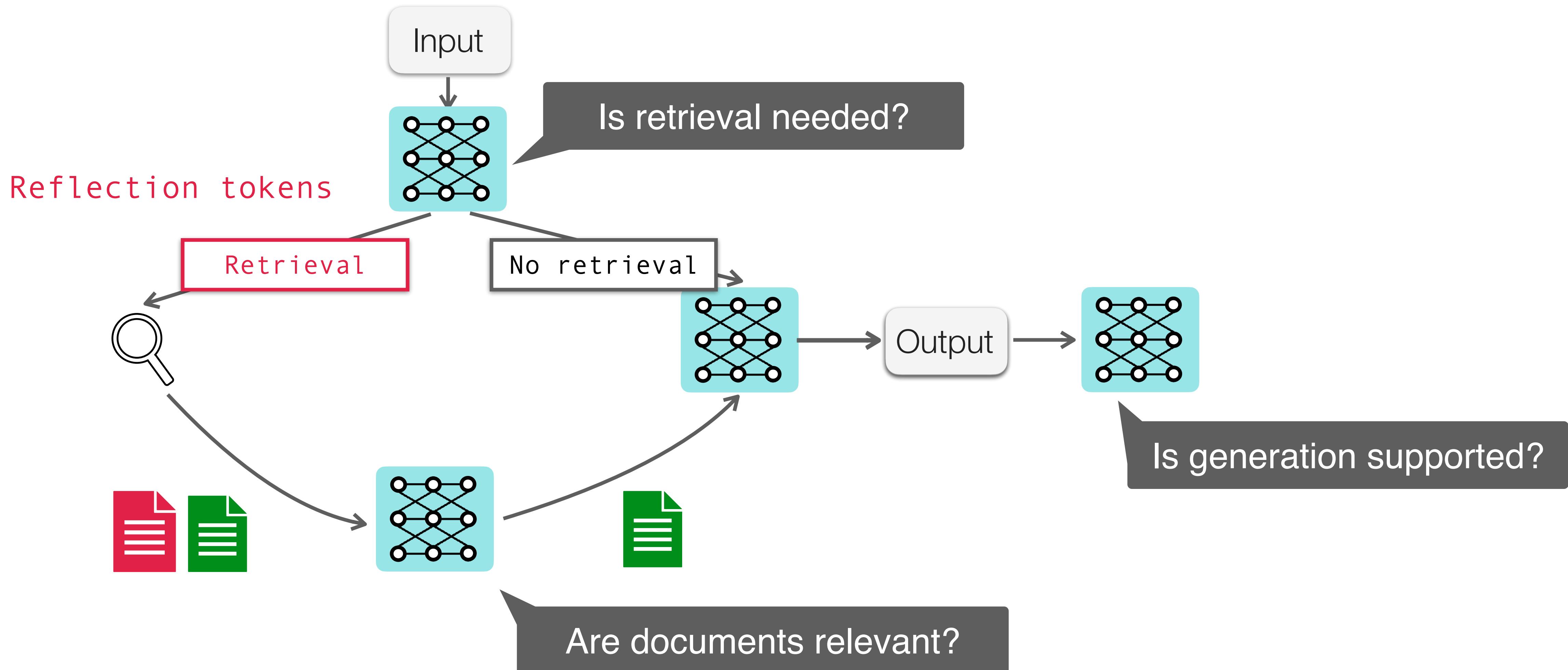
Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop 



Most Influential ICLR 2024 Paper #13 

Self-RAG: Self-Reflective Retrieval-Augmented Generation

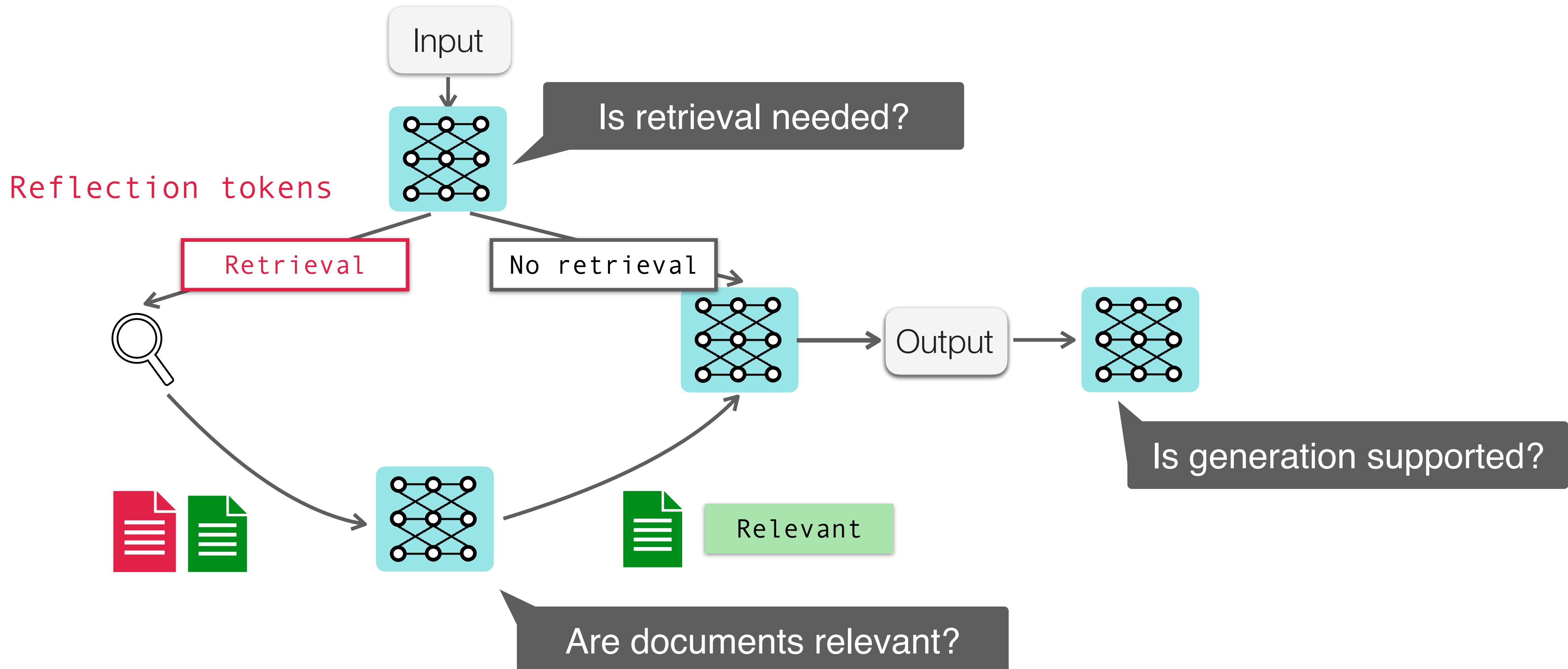


Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop

Most Influential ICLR 2024 Paper #13

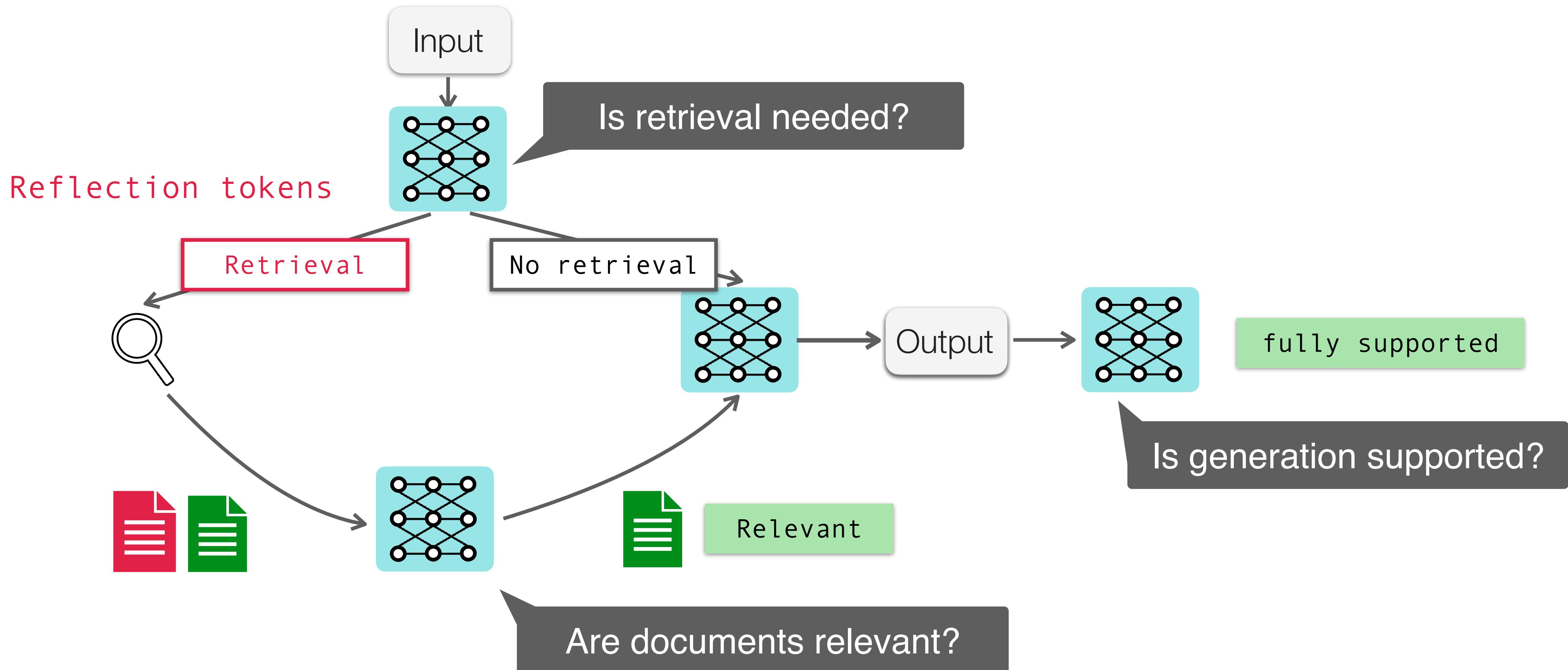
Self-RAG: Self-Reflective Retrieval-Augmented Generation



Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop Most Influential ICLR 2024 Paper #13

Self-RAG: Self-Reflective Retrieval-Augmented Generation



Asai et al. ICLR 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

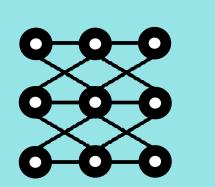
Oral - ICLR Top 1% & Best Paper Honorable mention at NeurIPS Instruction Tuning Workshop

Most Influential ICLR 2024 Paper #13

Self-RAG Inference

Input

Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature

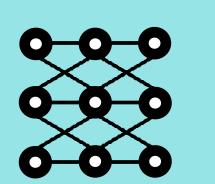


Sentence 1

Self-RAG Inference

Input

Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature

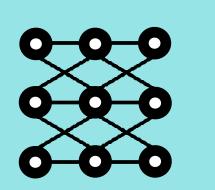


Sentence 1 Certainly!

Self-RAG Inference

Input

Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature



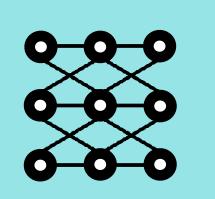
Sentence 1 Certainly!

Retrieval

Self-RAG Inference

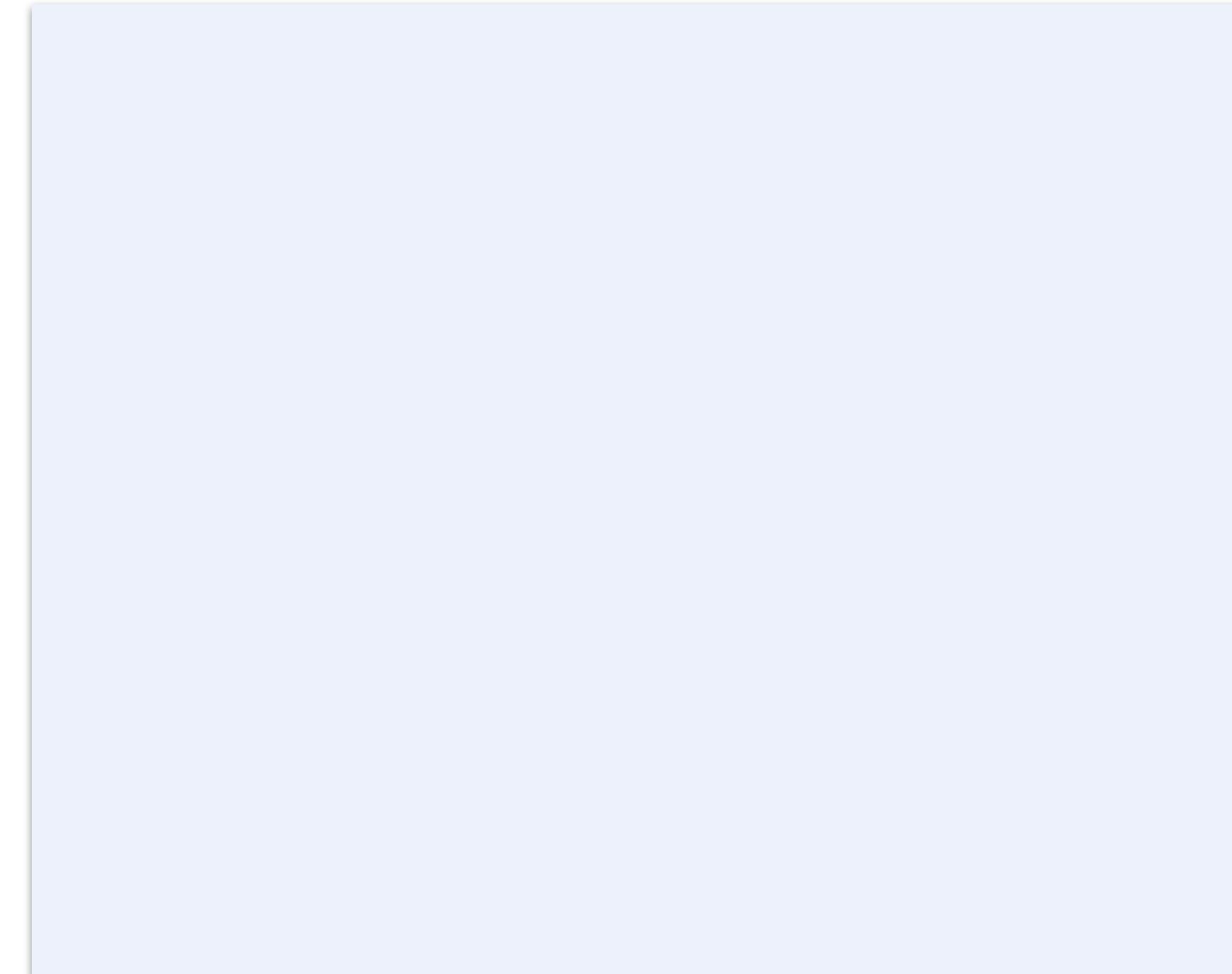
Input

Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature



Sentence 1 Certainly!

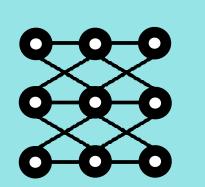
Retrieval



Self-RAG Inference

Input

Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature



Sentence 1 Certainly!

Retrieval



LLMs have been used in industry widely, such as chatbot system

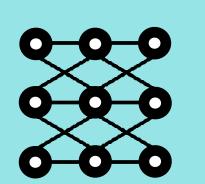
OpenScholar is a retrieval-augmented LM designed to synthesize literature

GPT4o has shown to be effective to generate new research ideas.

Self-RAG Inference

Input

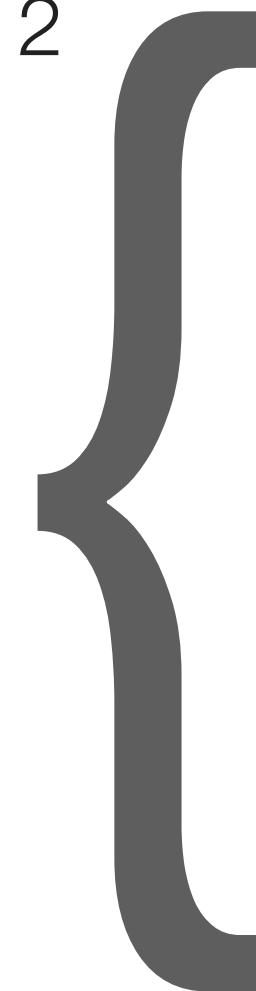
Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature



Sentence 1 Certainly! Retrieval



Sentence 2



LLMs have been used in industry widely, such as chatbot system

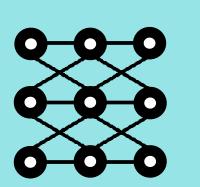
OpenScholar is a retrieval-augmented LM designed to synthesize literature

GPT4o has shown to be effective to generate new research ideas.

Self-RAG Inference

Input

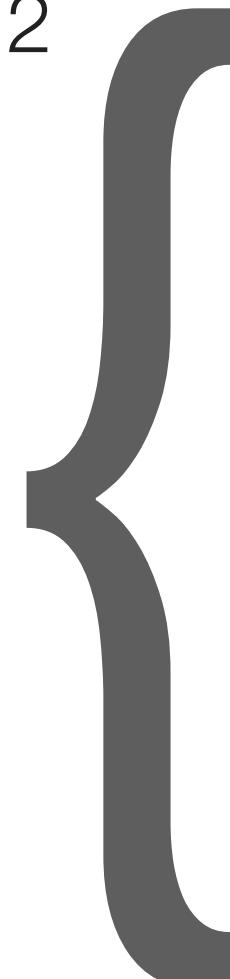
Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature



Sentence 1 Certainly! Retrieval



Sentence 2



Irrelevant

Relevant

Relevant

LLMs have been used in industry widely, such as chatbot system

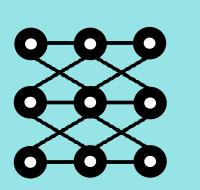
OpenScholar is a retrieval-augmented LM designed to synthesize literature

GPT4o has shown to be effective to generate new research ideas.

Self-RAG Inference

Input

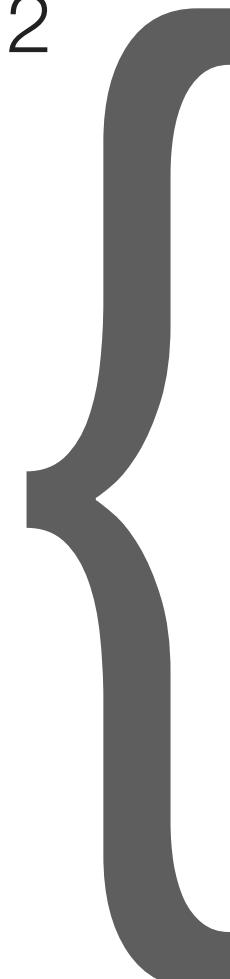
Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature



Sentence 1 Certainly! Retrieval



Sentence 2



Irrelevant

~~LLMs have been widely used in science.~~

Relevant

Relevant

LLMs have been used in industry widely, such as chatbot system

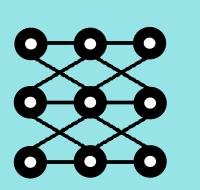
OpenScholar is a retrieval-augmented LM designed to synthesize literature

GPT4o has shown to be effective to generate new research ideas.

Self-RAG Inference

Input

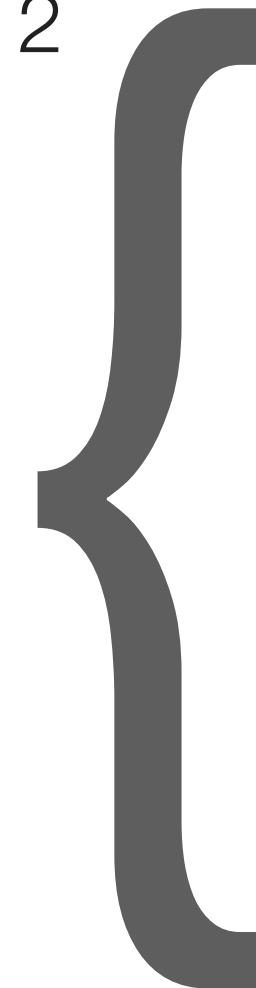
Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature



Sentence 1 Certainly! Retrieval



Sentence 2 { Irrelevant ~~LLMs have been widely used in science.~~



Relevant

Relevant

LLMs have been used in industry widely, such as chatbot system

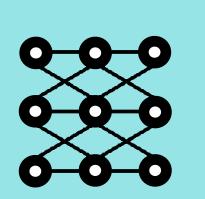
OpenScholar is a retrieval-augmented LM designed to synthesize literature

GPT4o has shown to be effective to generate new research ideas.

Self-RAG Inference

Input

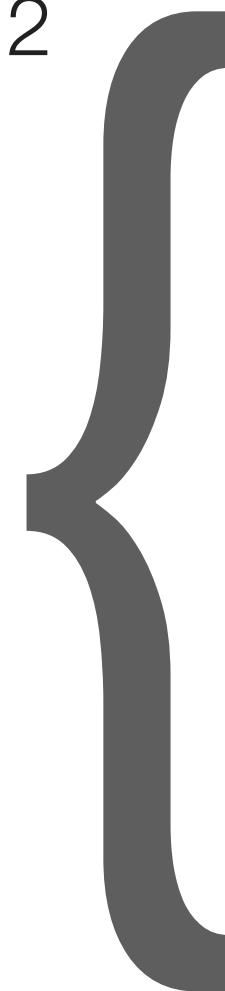
Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature



Sentence 1 Certainly! Retrieval



Sentence 2



Relevant

Relevant

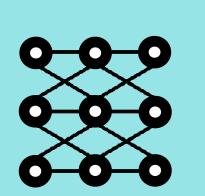
OpenScholar is a retrieval-augmented LM designed to synthesize literature

GPT4o has shown to be effective to generate new research ideas.

Self-RAG Inference

Input

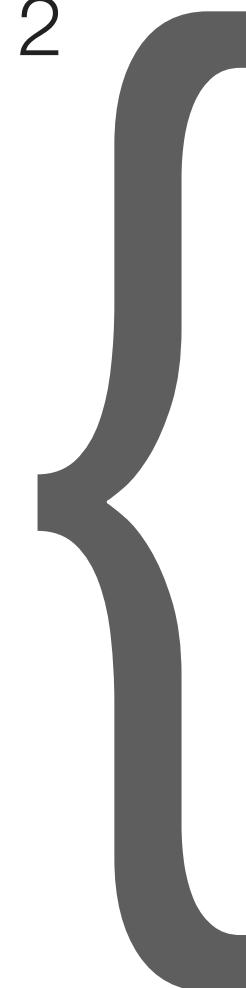
Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature



Sentence 1 Certainly! Retrieval



Sentence 2



Relevant

Relevant

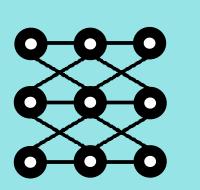
OpenScholar is a retrieval-augmented LM designed to synthesize literature

GPT4o has shown to be effective to generate new research ideas.

Self-RAG Inference

Input

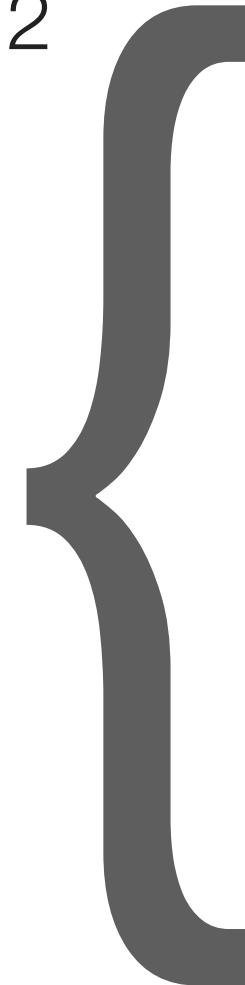
Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature



Sentence 1 Certainly! Retrieval



Sentence 2



Relevant

Relevant

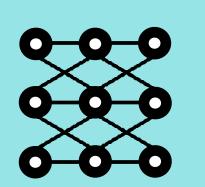
OpenScholar is a retrieval-augmented LM designed to synthesize literature

GPT4o has shown to be effective to generate new research ideas.

Self-RAG Inference

Input

Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature



Sentence 1 Certainly! Retrieval



Sentence 2

{ Relevant OpenScholar is an LM for literature synthesis.

OpenScholar is a retrieval-augmented LM designed to synthesize literature

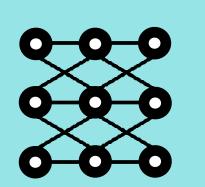
Relevant

GPT4o has shown to be effective to generate new research ideas.

Self-RAG Inference

Input

Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature



Sentence 1 Certainly! Retrieval



Sentence 2

{ Relevant OpenScholar is an LM for literature synthesis.

OpenScholar is a retrieval-augmented LM designed to synthesize literature

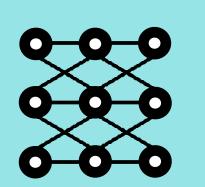
Relevant Studies show GPT4o can help scientists for idea generations and literature synthesis.

GPT4o has shown to be effective to generate new research ideas.

Self-RAG Inference

Input

Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature



Sentence 1 Certainly! Retrieval



Sentence 2

{

Relevant OpenScholar is an LM for
literature synthesis. fully supported

OpenScholar is a retrieval-augmented
LM designed to synthesize literature

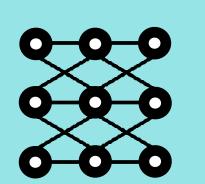
Relevant Studies show GPT4o can help
scientists for idea generations and
literature synthesis. Partially supported

GPT4o has shown to be effective to
generate new research ideas.

Self-RAG Inference

Input

Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature



Sentence 1 Certainly! Retrieval

Sentence 2

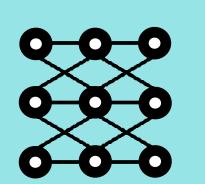
{ Relevant OpenScholar is an LM for
literature synthesis. fully supported

Relevant Studies show GPT4o can help
scientists for idea generations and
~~literature synthesis.~~ Partially supported

Self-RAG Inference

Input

Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature



Sentence 1 Certainly! Retrieval

0.9

Sentence 2

{

Relevant OpenScholar is an LM for
literature synthesis. fully supported

0.7

0.6

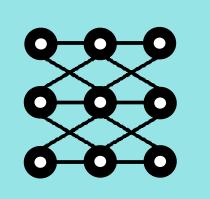
Relevant Studies show GPT4o can help
scientists for idea generations and
~~literature synthesis.~~ Partially supported

0.3

Self-RAG Inference

Input

Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature



Sentence 1 Certainly! Retrieval
0.9

Sentence 2 { Relevant OpenScholar is an LM for
literature synthesis. fully supported
0.7

{

0.6
Relevant Studies show GPT4o can help
scientists for idea generations and
~~literature synthesis.~~ Partially supported

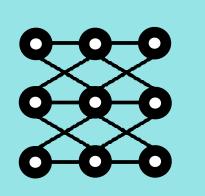
0.3

$$S_i = \sum w_i (\begin{array}{l} \text{Relevant} \\ \text{Supported} \end{array})$$

Self-RAG Inference

Input

Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature



Sentence 1 Certainly! Retrieval

0.9

Sentence 2 { Relevant OpenScholar is an LM for
literature synthesis. fully supported

0.7

0.6
Relevant Studies show GPT4o can help
scientists for idea generations and
~~literature synthesis.~~ Partially supported

0.3

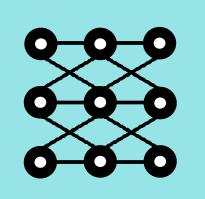
$$S_i = \sum w_i (\begin{array}{l} \text{Relevant} \\ \text{Supported} \end{array})$$

0.9

Self-RAG Inference

Input

Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature



Sentence 1 Certainly! Retrieval

0.9

Sentence 2 { Relevant OpenScholar is an LM for
literature synthesis. fully supported

0.7

0.6
Relevant Studies show GPT4o can help
scientists for idea generations and
~~literature synthesis.~~ Partially supported

0.3

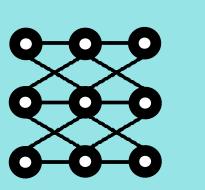
$$S_i = \sum w_i (\begin{array}{l} \text{Relevant} \\ \text{Supported} \end{array})$$

0.9

Self-RAG Inference

Input

Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature



Sentence 1 Certainly! Retrieval

Sentence 2

{

Relevant OpenScholar is an LM for
literature synthesis. fully supported

0.9

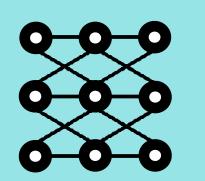
Relevant Studies show GPT4o can help
scientists for idea generations and
~~literature synthesis.~~ Partially supported

0.5

Self-RAG Inference

Input

Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature



Sentence 1 Certainly! Retrieval

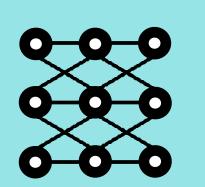
Sentence 2

Relevant OpenScholar is an LM for
literature synthesis. fully supported

Self-RAG Inference

Input

Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature



Sentence 1 Certainly! Retrieval

Sentence 2

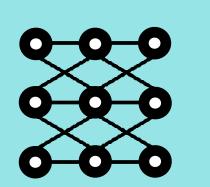
Relevant OpenScholar is an LM for
literature synthesis. fully supported

Sentence 3

Self-RAG Inference

Input

Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature



Sentence 1 Certainly! Retrieval

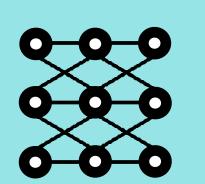
Sentence 2 Relevant OpenScholar is an LM for
literature synthesis. fully supported

Sentence 3 OpenScholar uses retrieval-augmented
approaches.

Self-RAG Inference

Input

Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature



Sentence 1 Certainly! Retrieval

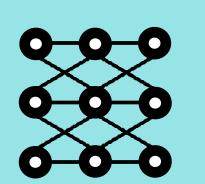
Sentence 2 Relevant OpenScholar is an LM for literature synthesis. fully supported

Sentence 3 OpenScholar uses retrieval-augmented approaches. fully supported

Self-RAG Inference

Input

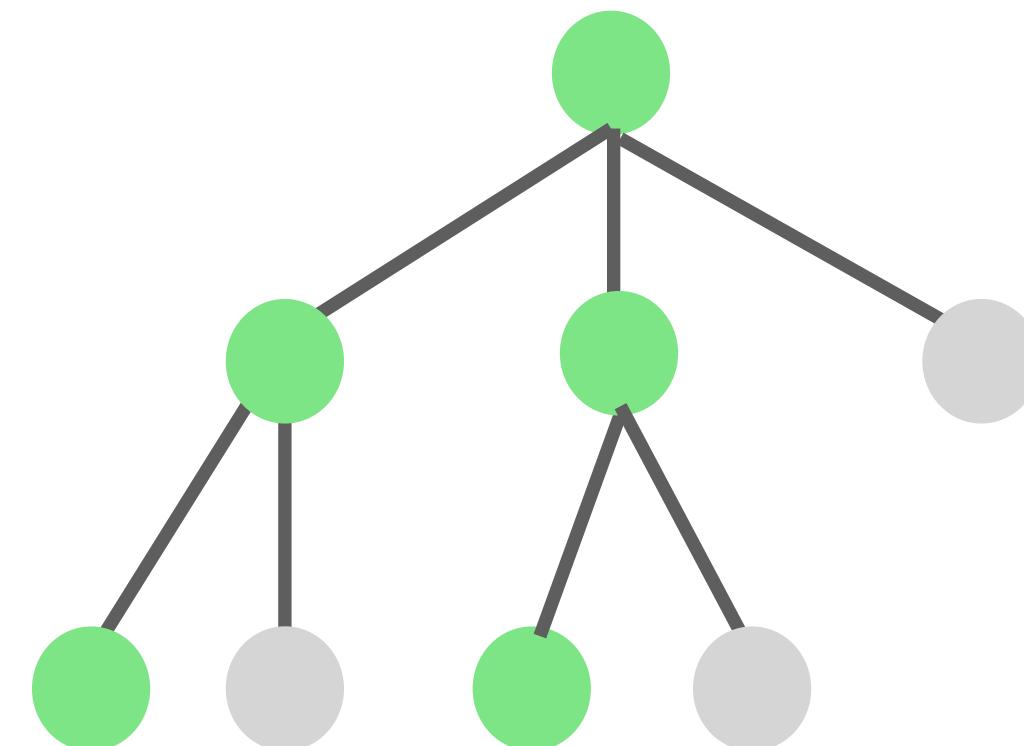
Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature



Sentence 1 Certainly! Retrieval

Sentence 2 Relevant OpenScholar is an LM for literature synthesis. fully supported

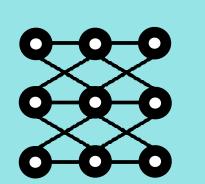
Sentence 3 OpenScholar uses retrieval-augmented approaches. fully supported



Self-RAG Inference

Input

Suggest papers showing LLMs' effectiveness helping scientist to synthesize scientific literature

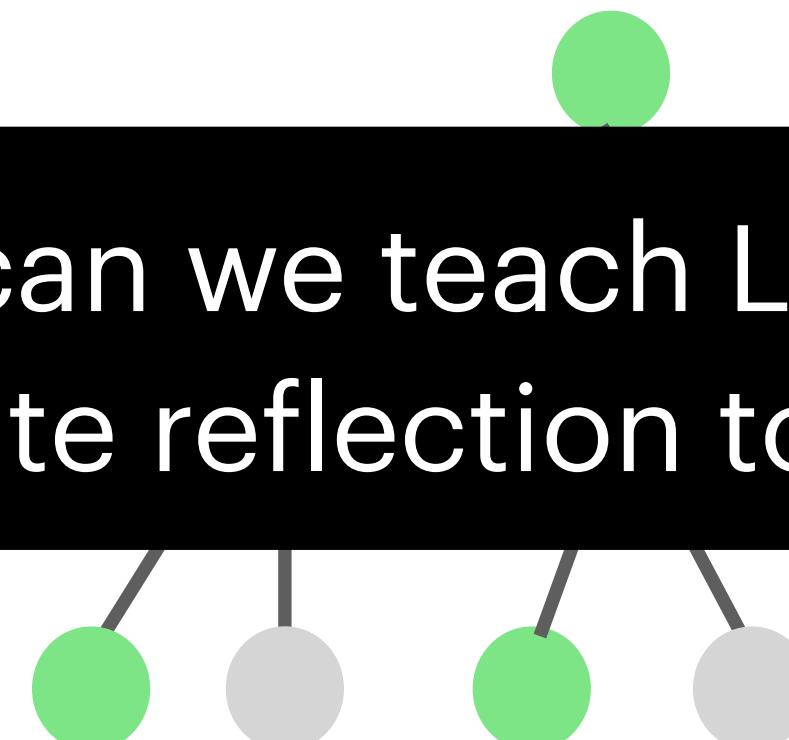


Sentence 1 Certainly! Retrieval

Sentence 2 Relevant OpenScholar is an LM for literature synthesis. fully supported

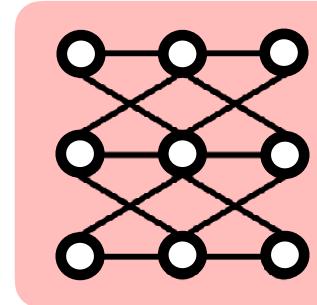
Sentence 3 OpenScholar uses retrieval-augmented approaches. fully supported

How can we teach LMs to generate reflection tokens?



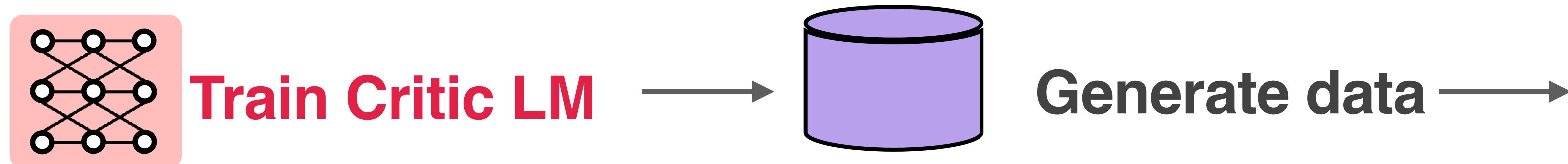
Self-RAG Training

Self-RAG Training



Train Critic LM →

Self-RAG Training



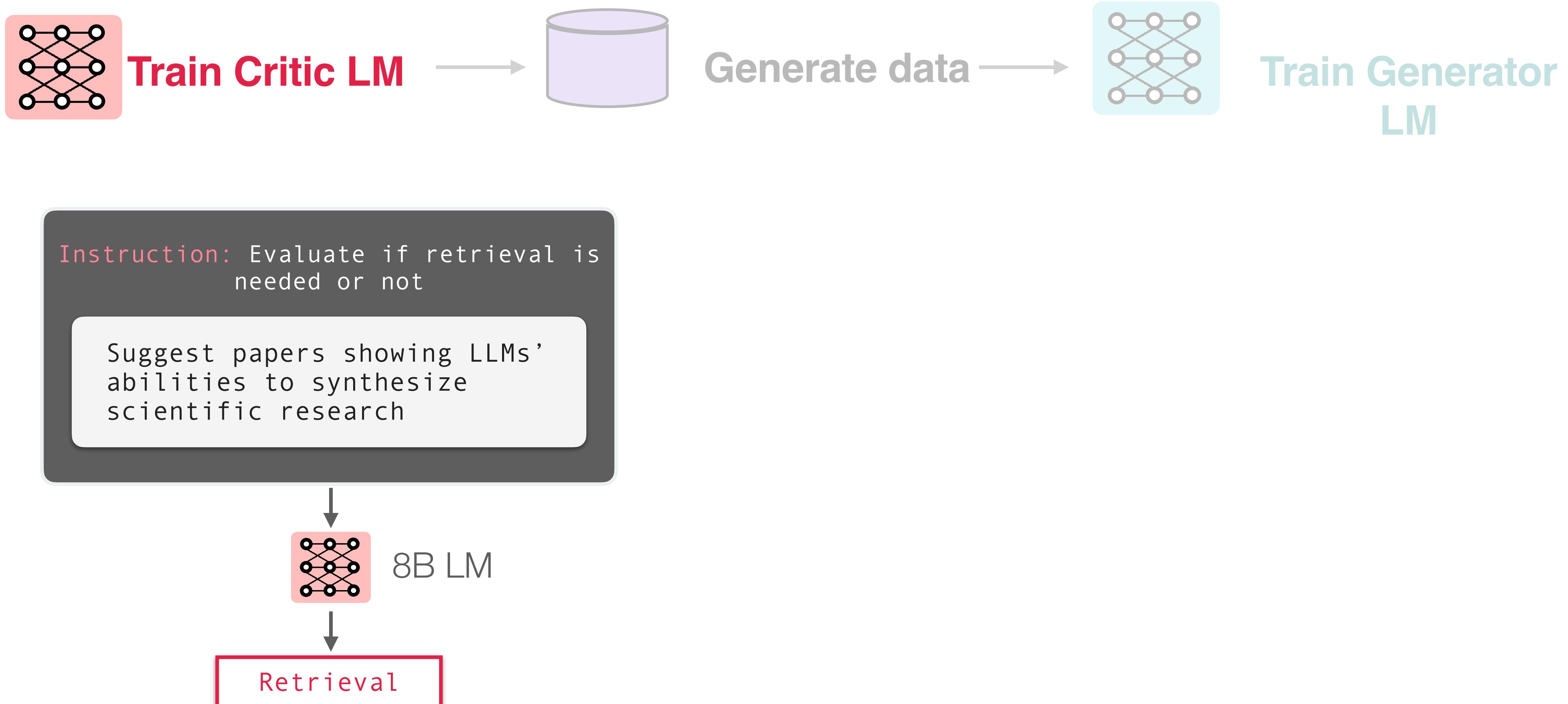
Self-RAG Training



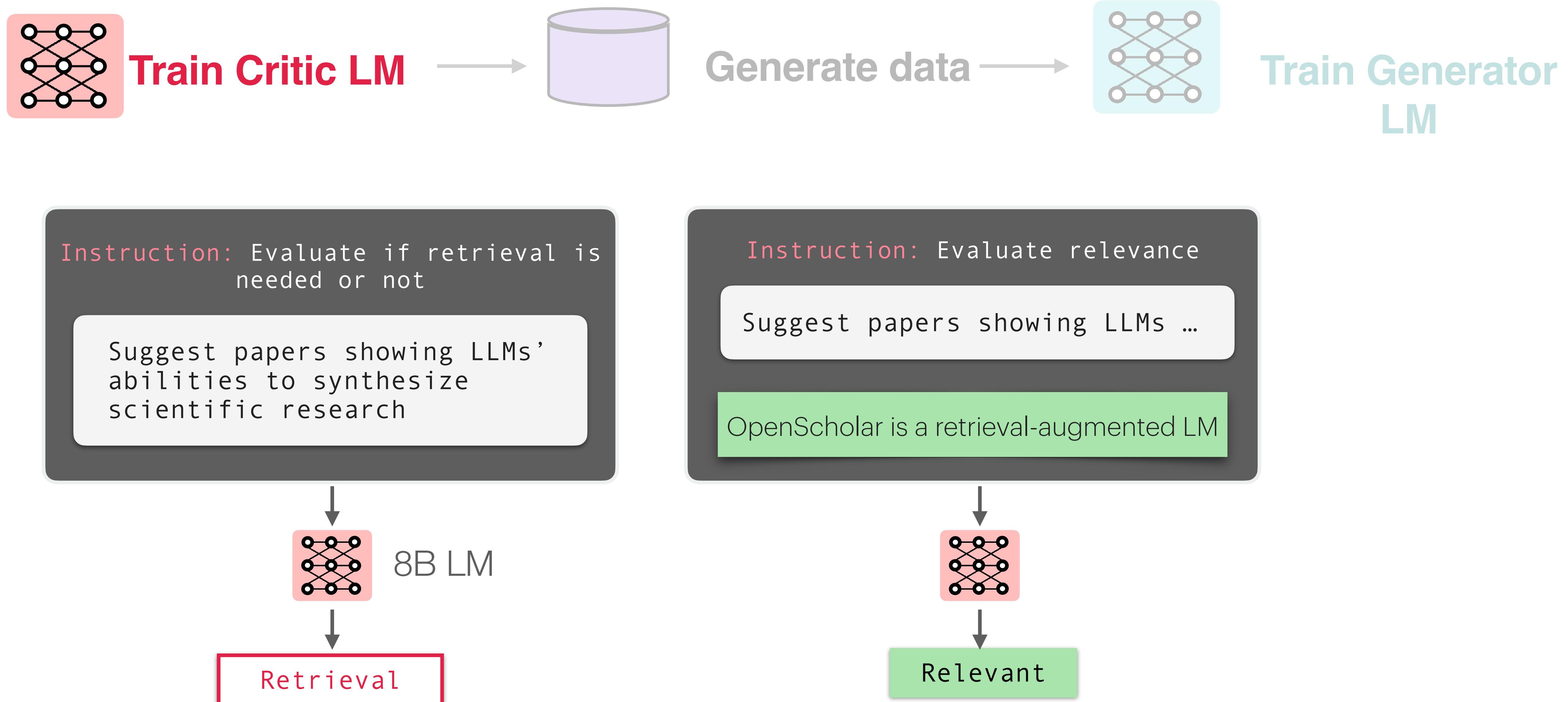
Self-RAG Training: Critic LM



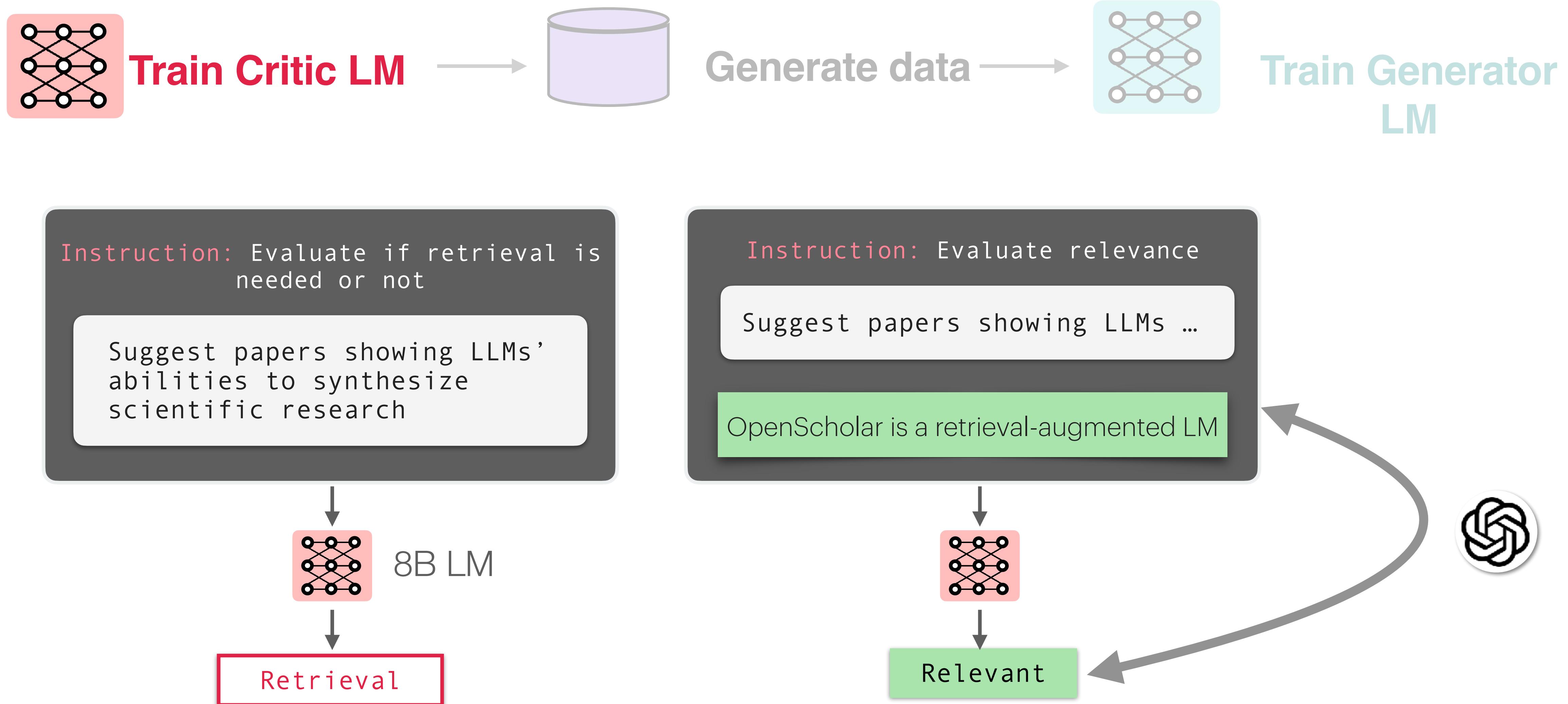
Self-RAG Training: Critic LM



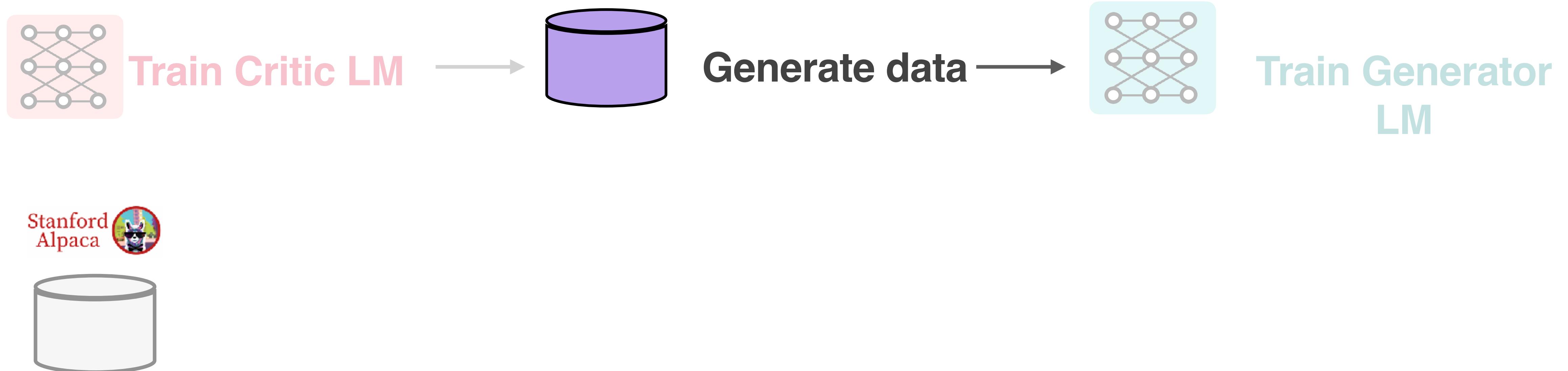
Self-RAG Training: Critic LM



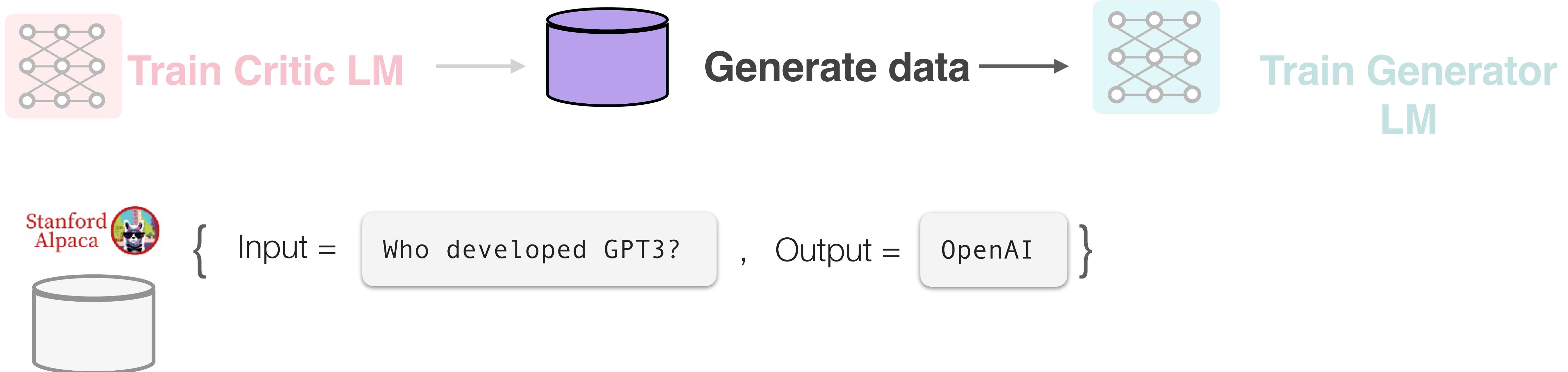
Self-RAG Training: Critic LM



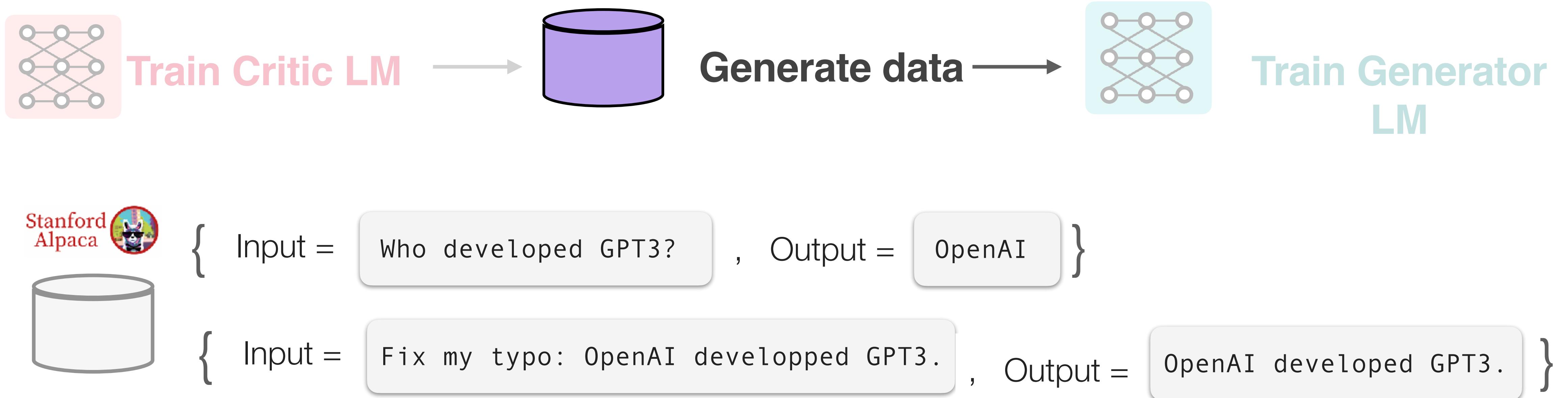
Self-RAG Training: Data Generation



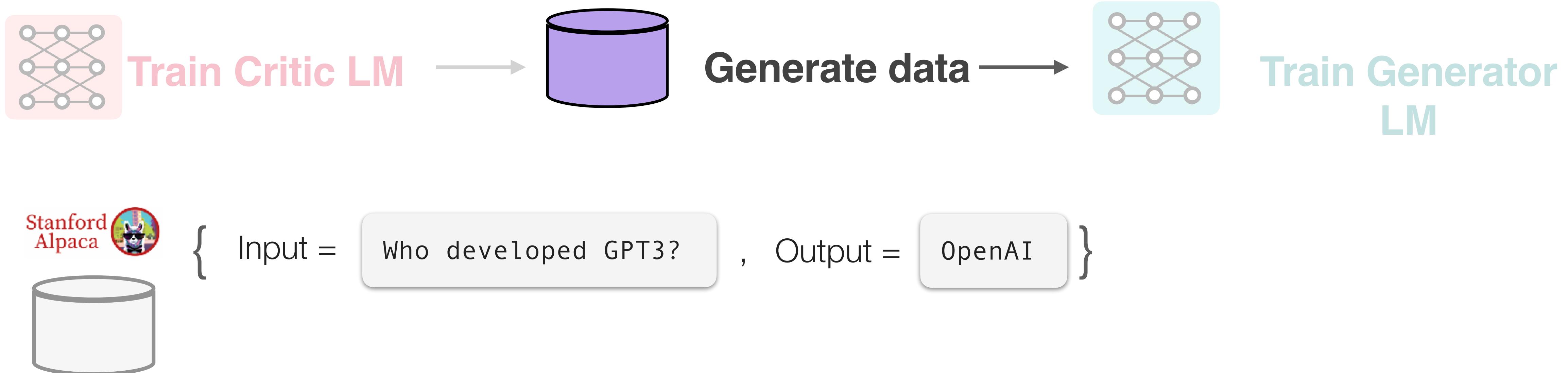
Self-RAG Training: Data Generation



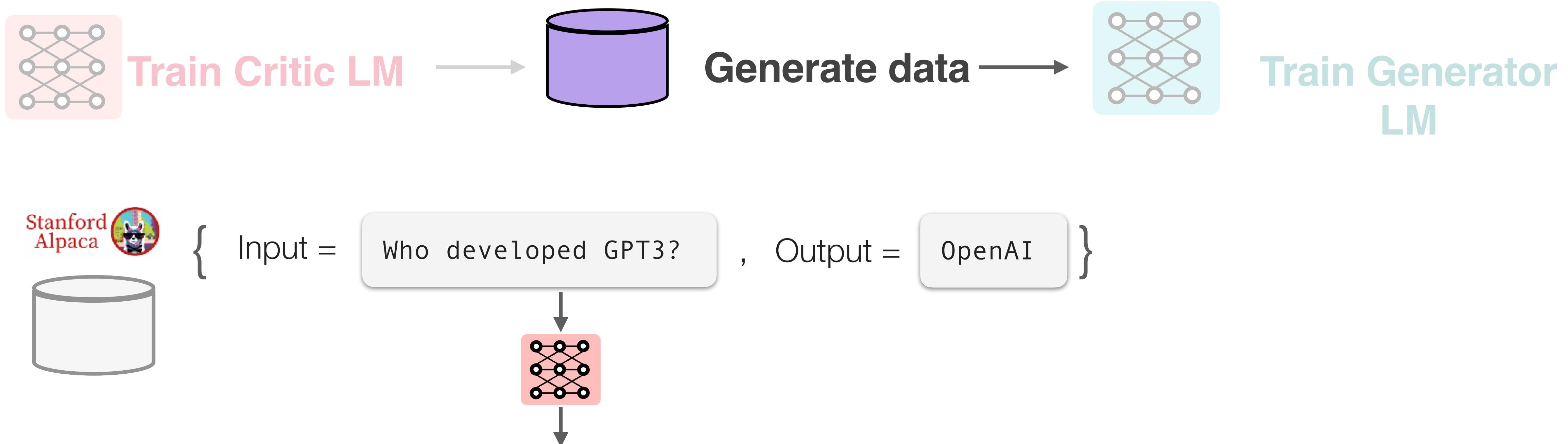
Self-RAG Training: Data Generation



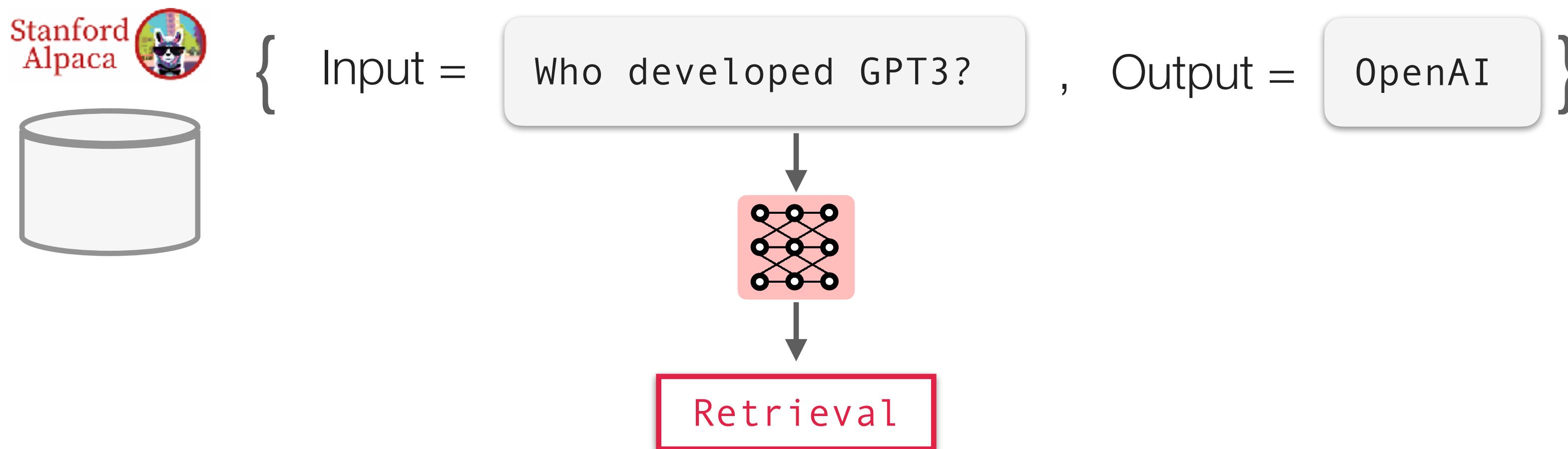
Self-RAG Training: Data Generation



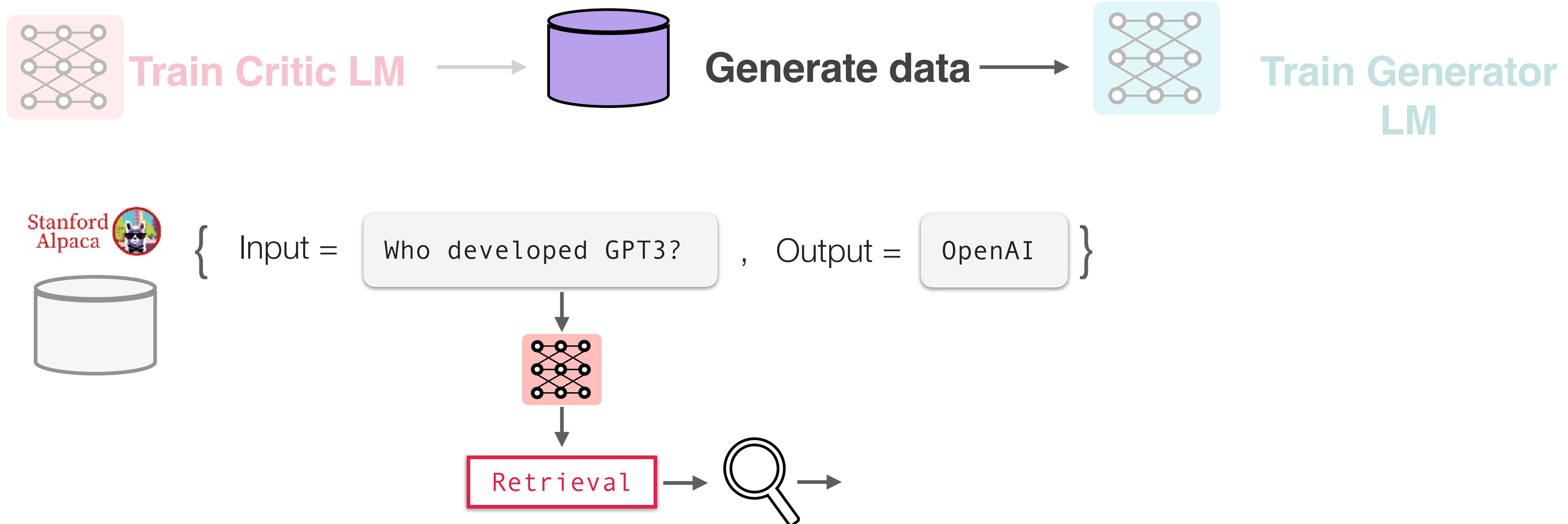
Self-RAG Training: Data Generation



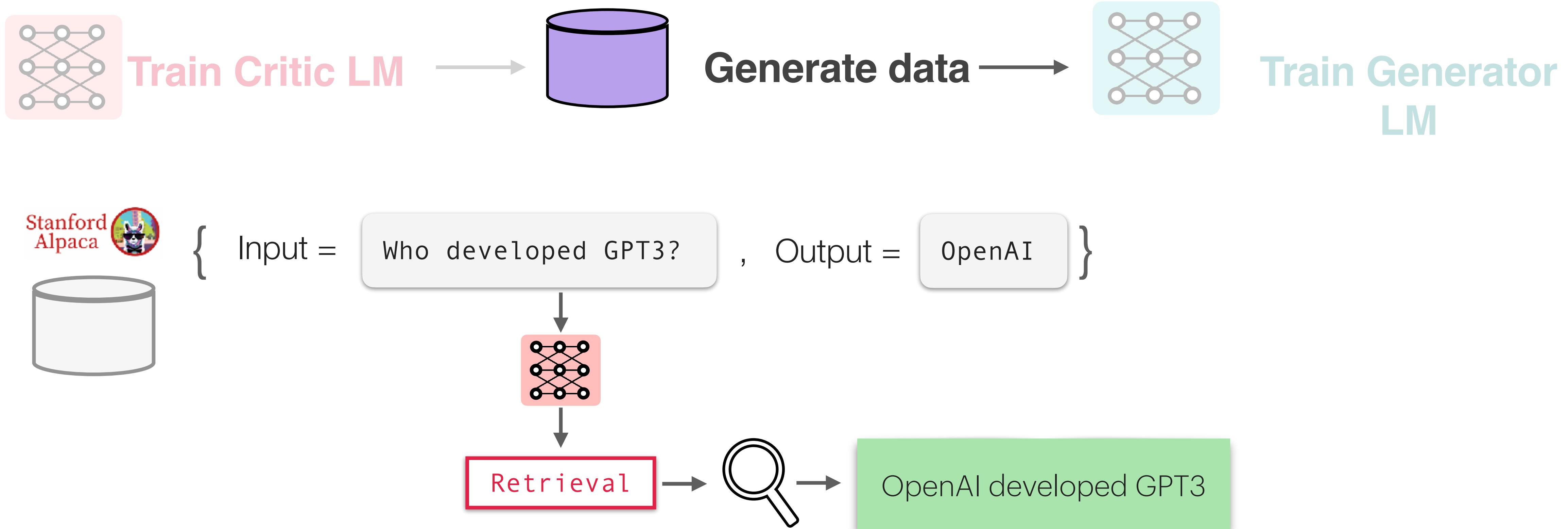
Self-RAG Training: Data Generation



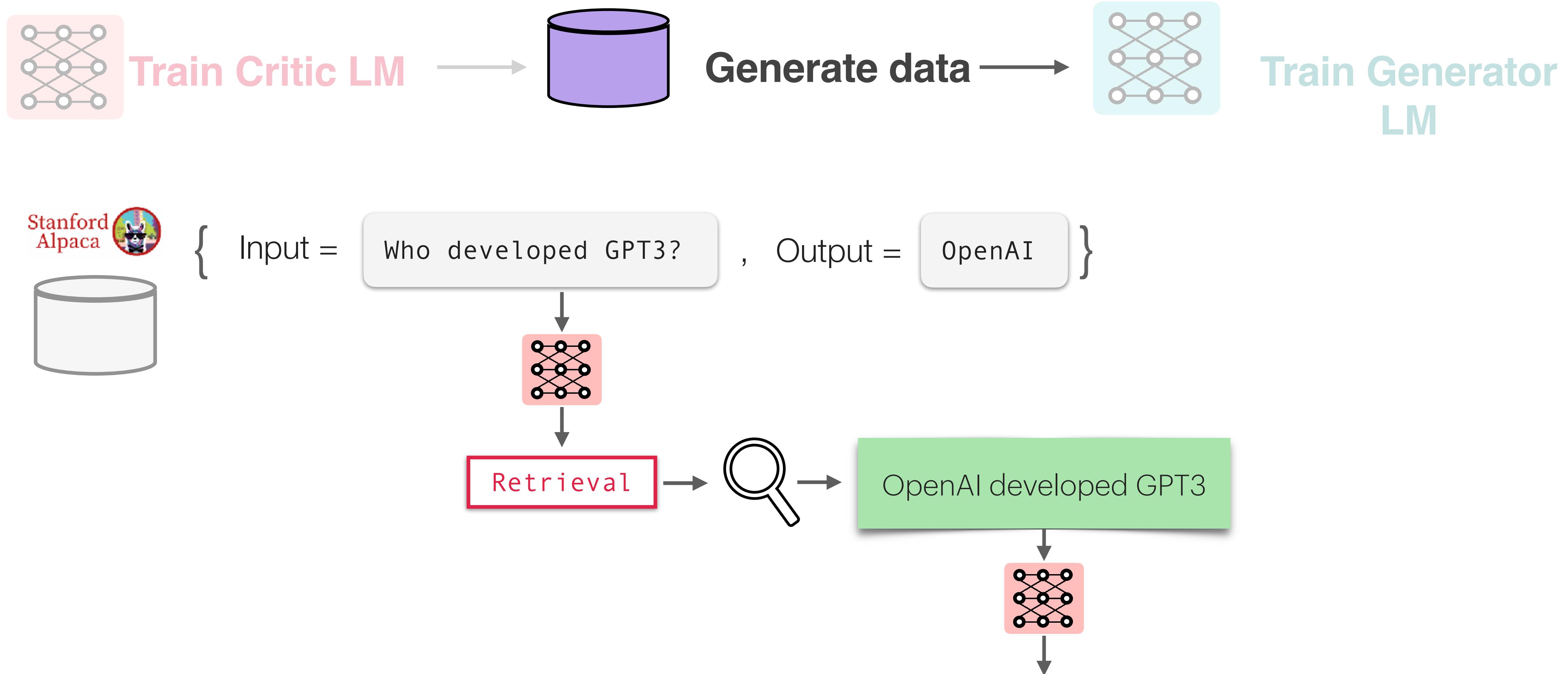
Self-RAG Training: Data Generation



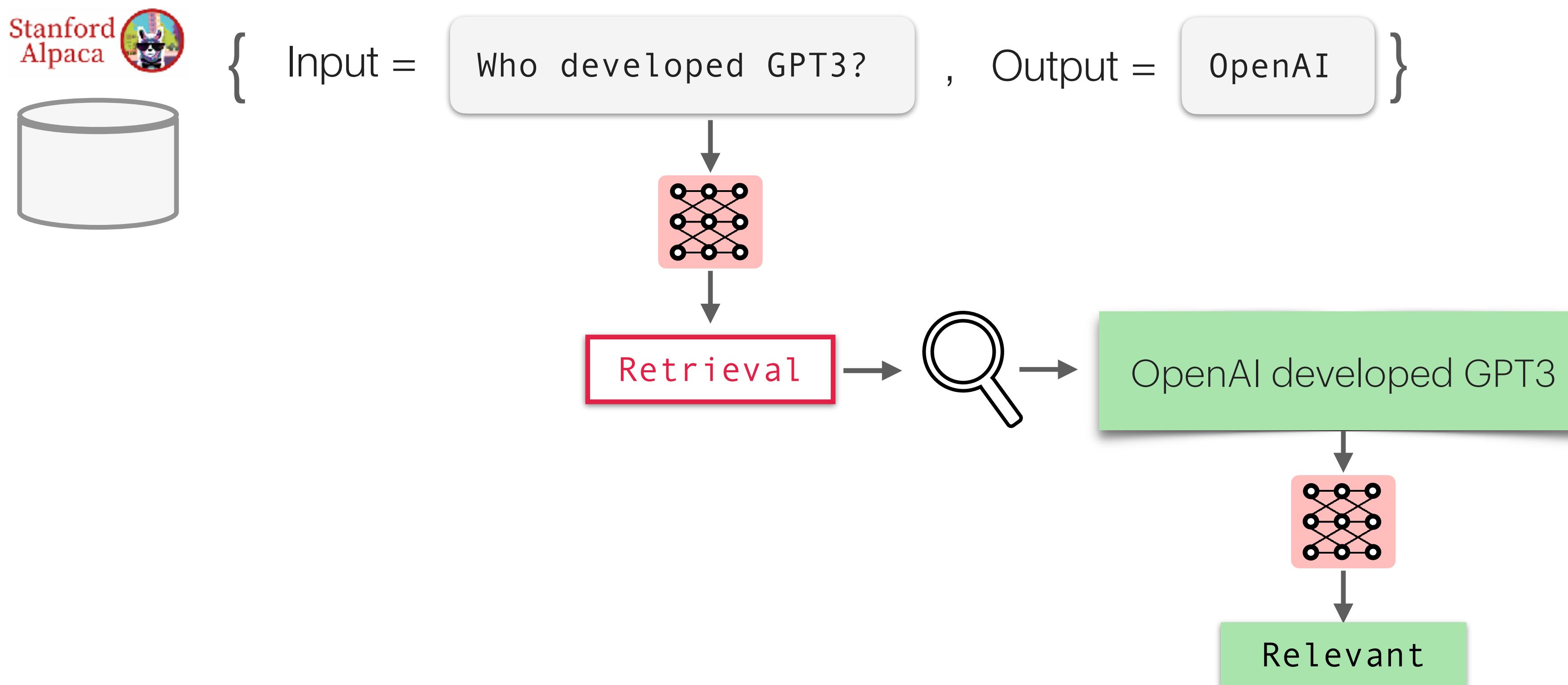
Self-RAG Training: Data Generation



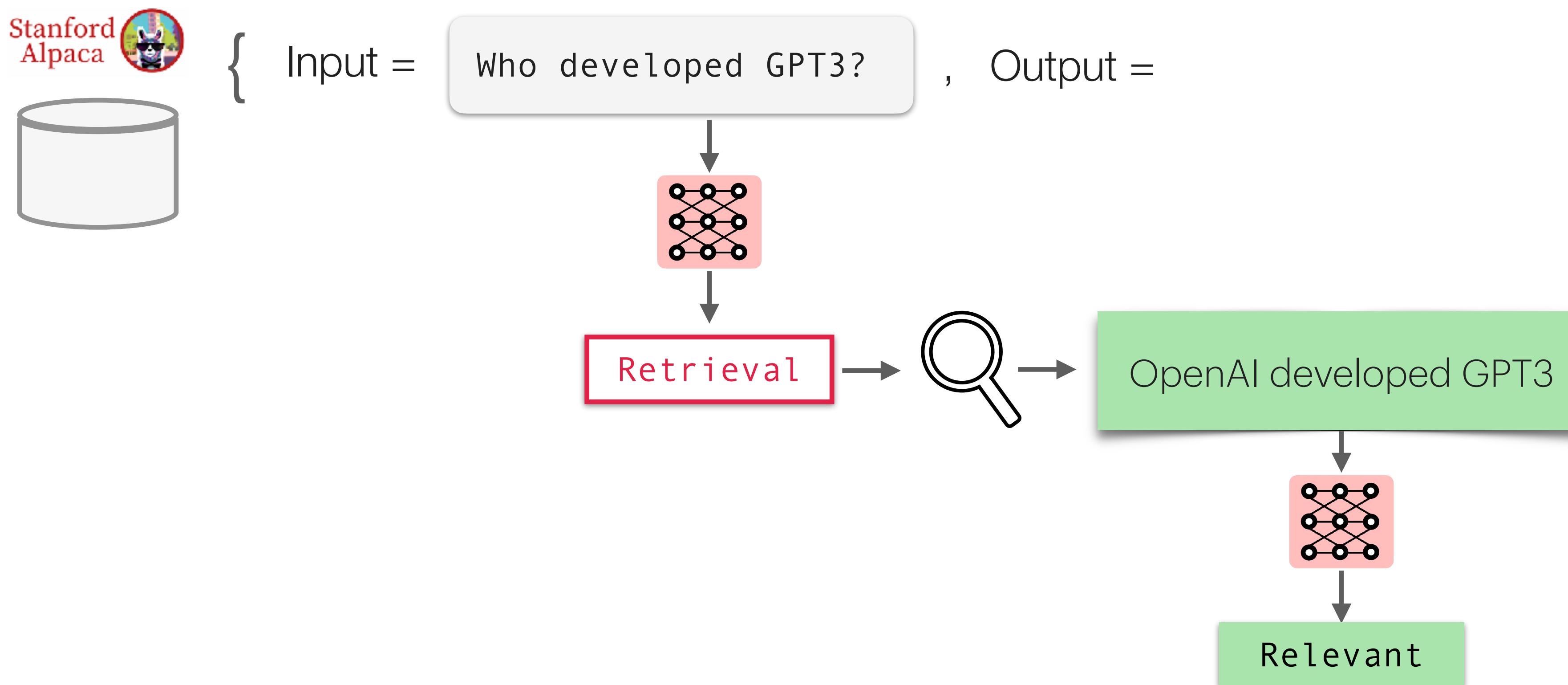
Self-RAG Training: Data Generation



Self-RAG Training: Data Generation



Self-RAG Training: Data Generation



Self-RAG Training: Data Generation



Stanford Alpaca { Input = Who developed GPT3? , Output =



Retrieval

OpenAI developed GPT3

Relevant

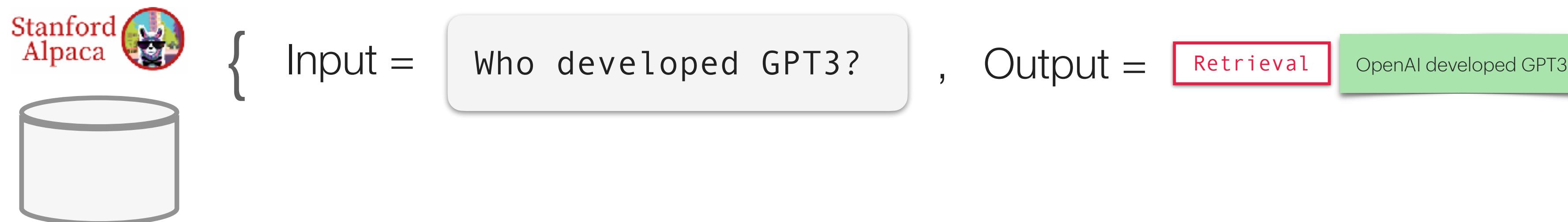
Self-RAG Training: Data Generation



OpenAI developed GPT3

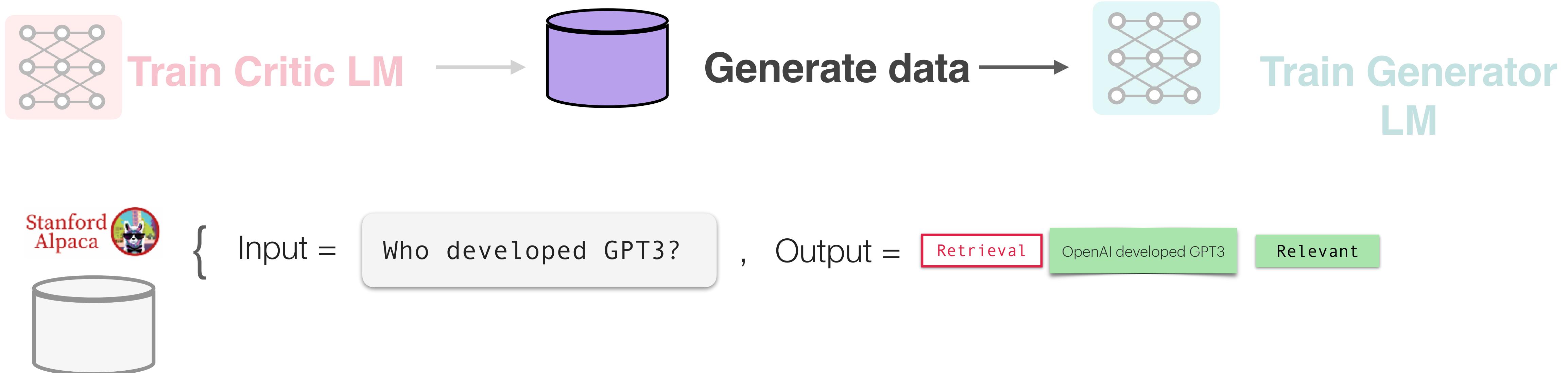
Relevant

Self-RAG Training: Data Generation

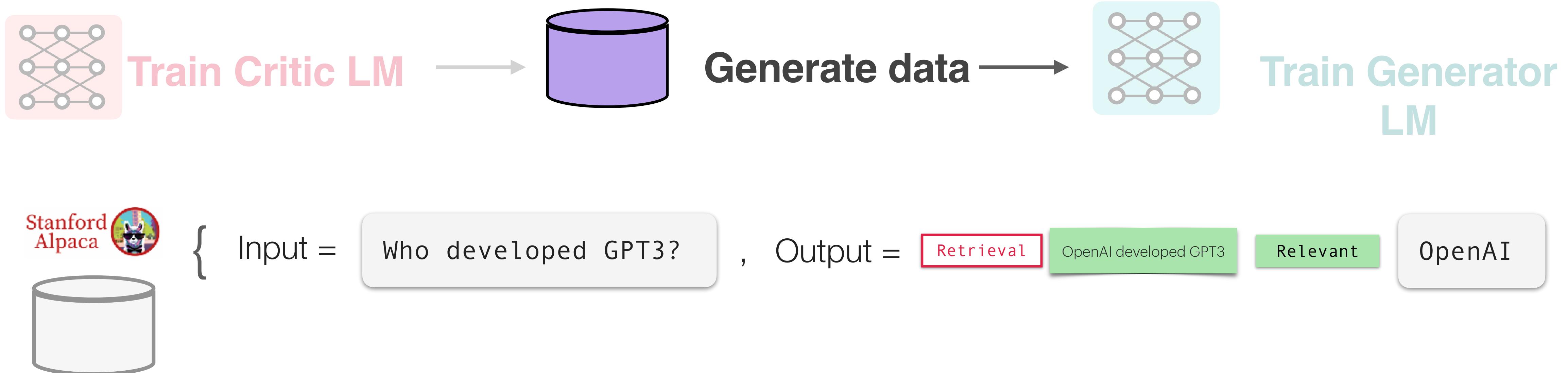


Relevant

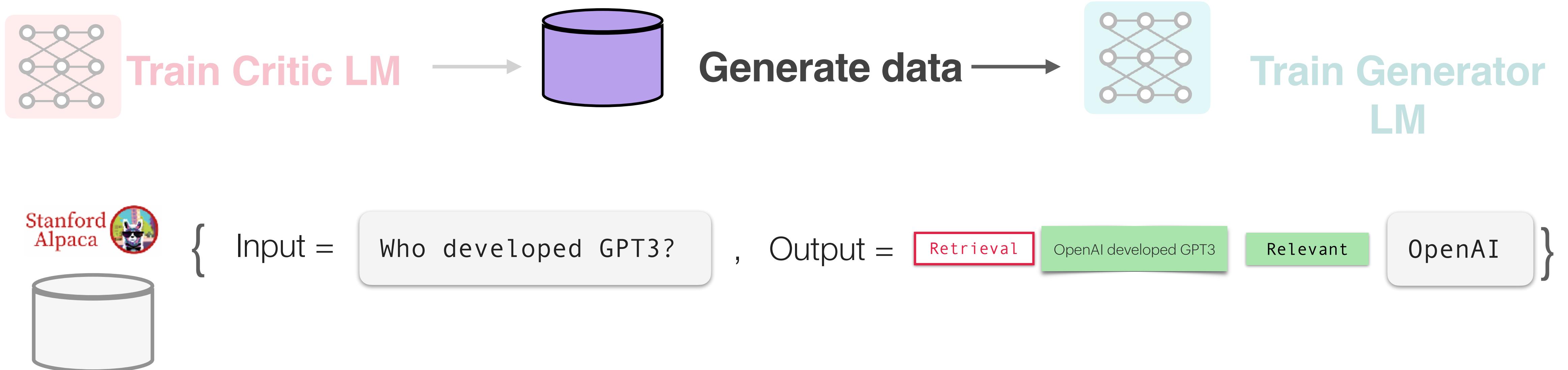
Self-RAG Training: Data Generation



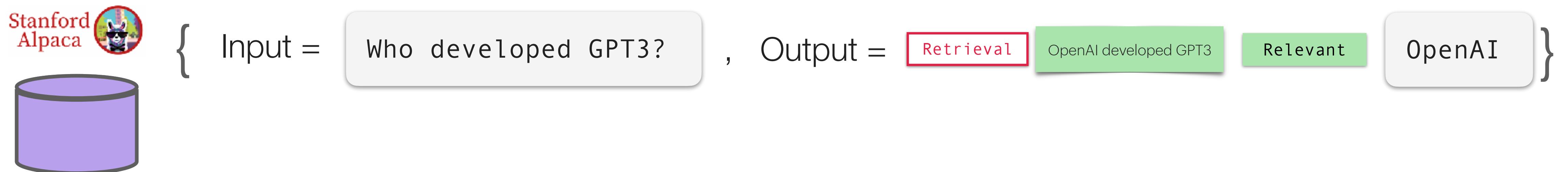
Self-RAG Training: Data Generation



Self-RAG Training: Data Generation

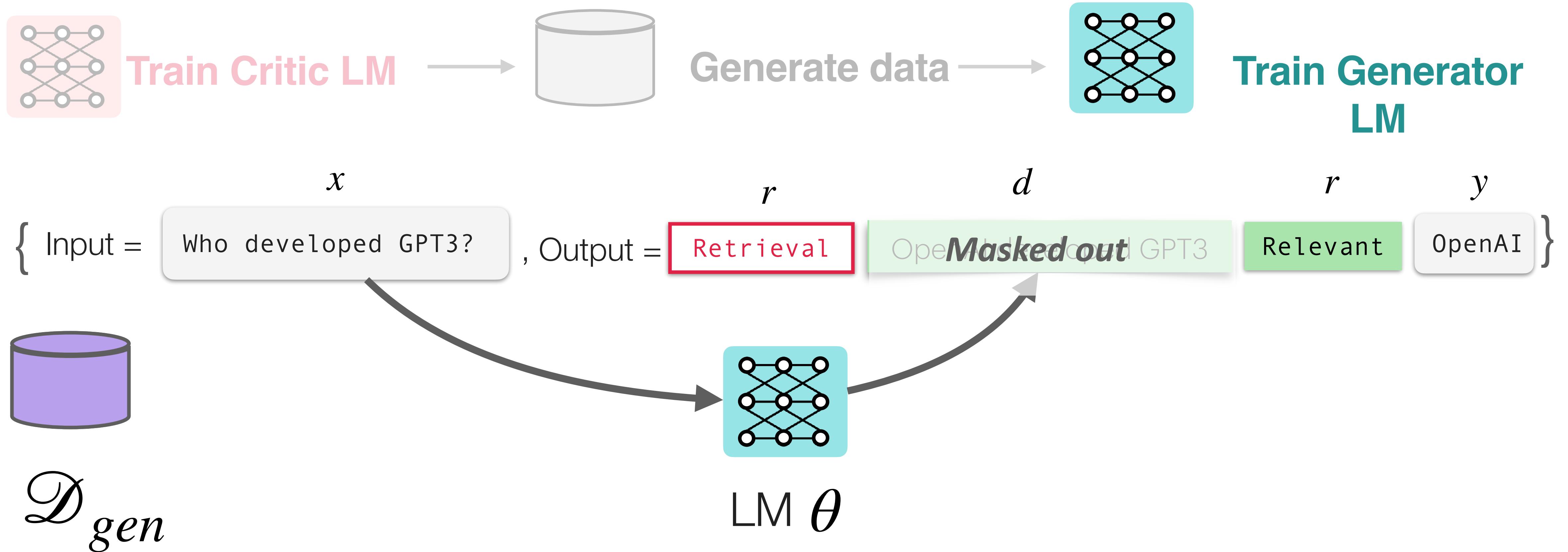


Self-RAG Training: Data Generation

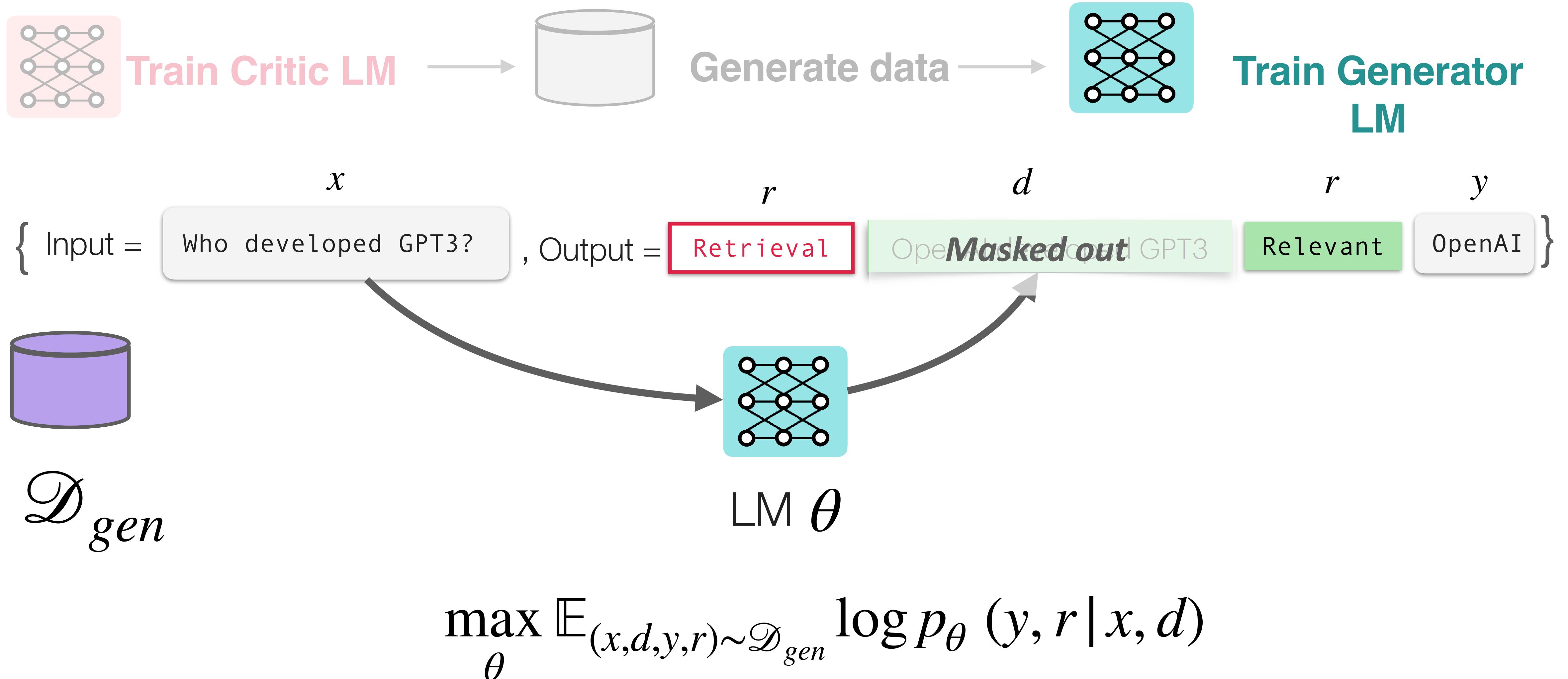


\mathcal{D}_{gen}

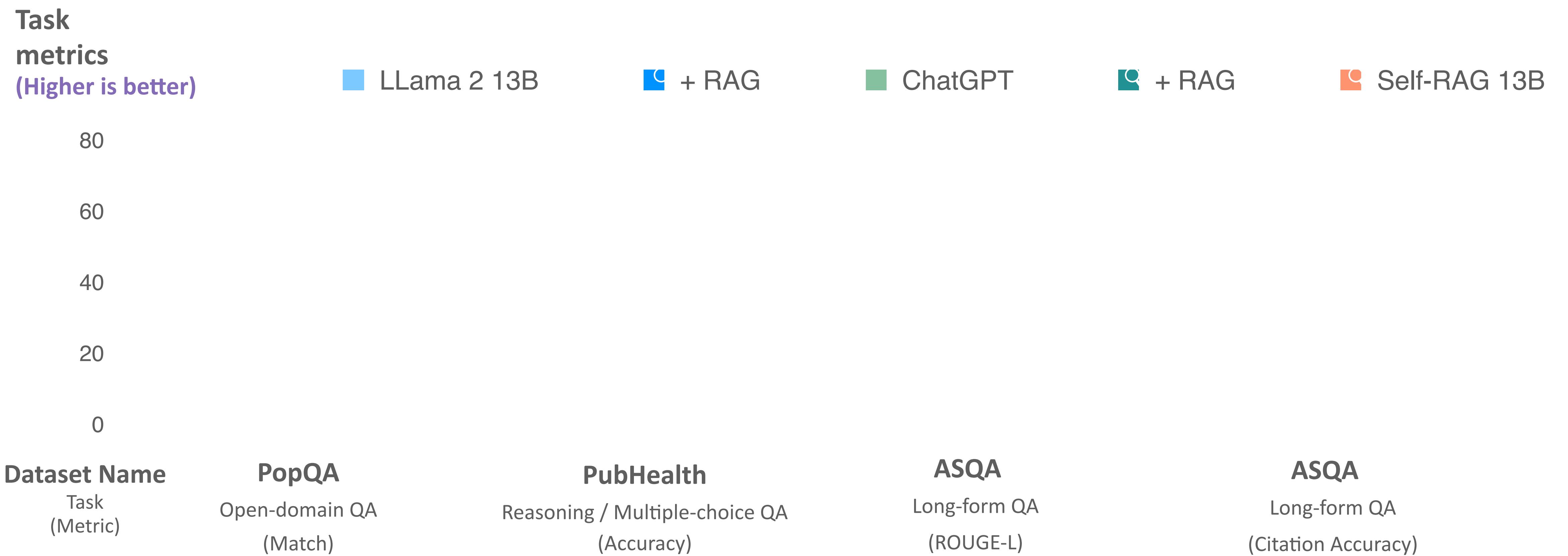
Self-RAG Training: Generator LM



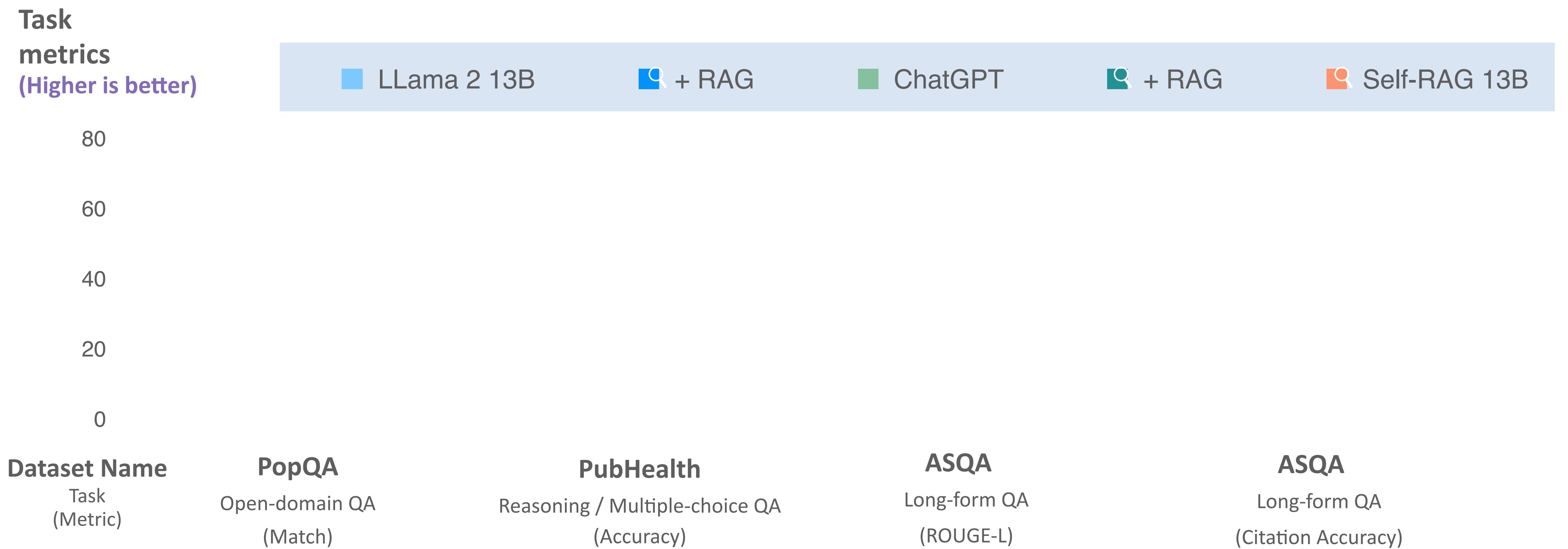
Self-RAG Training: Generator LM



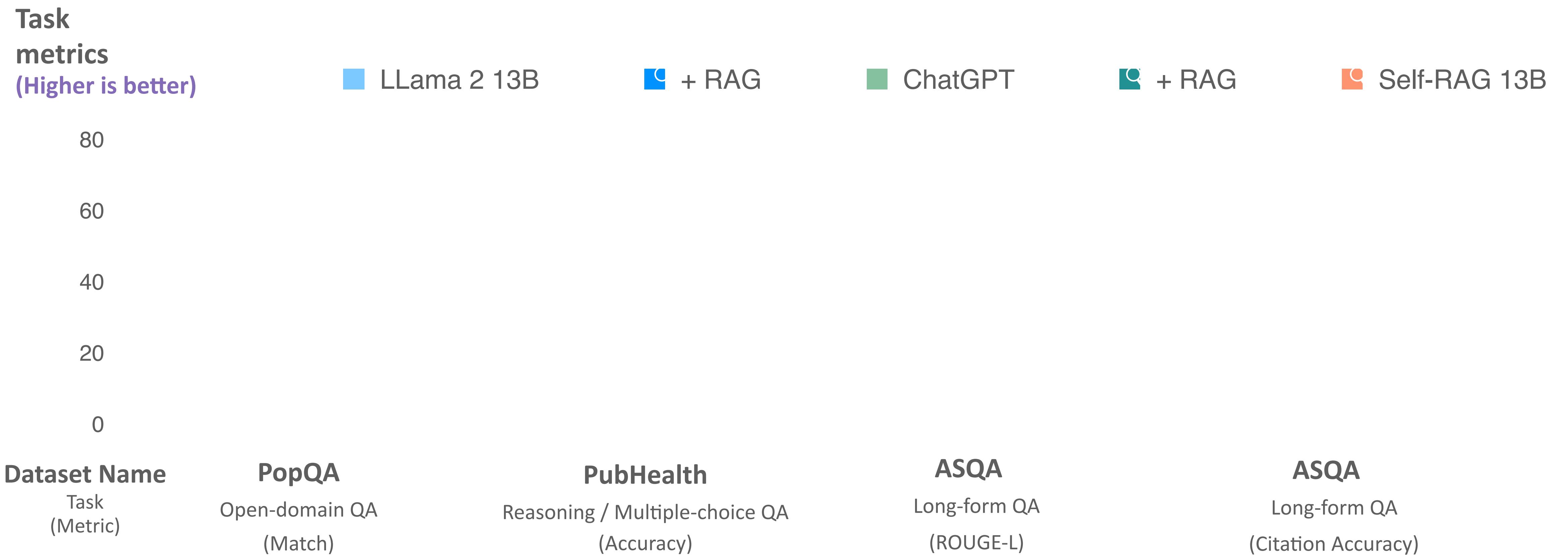
Results



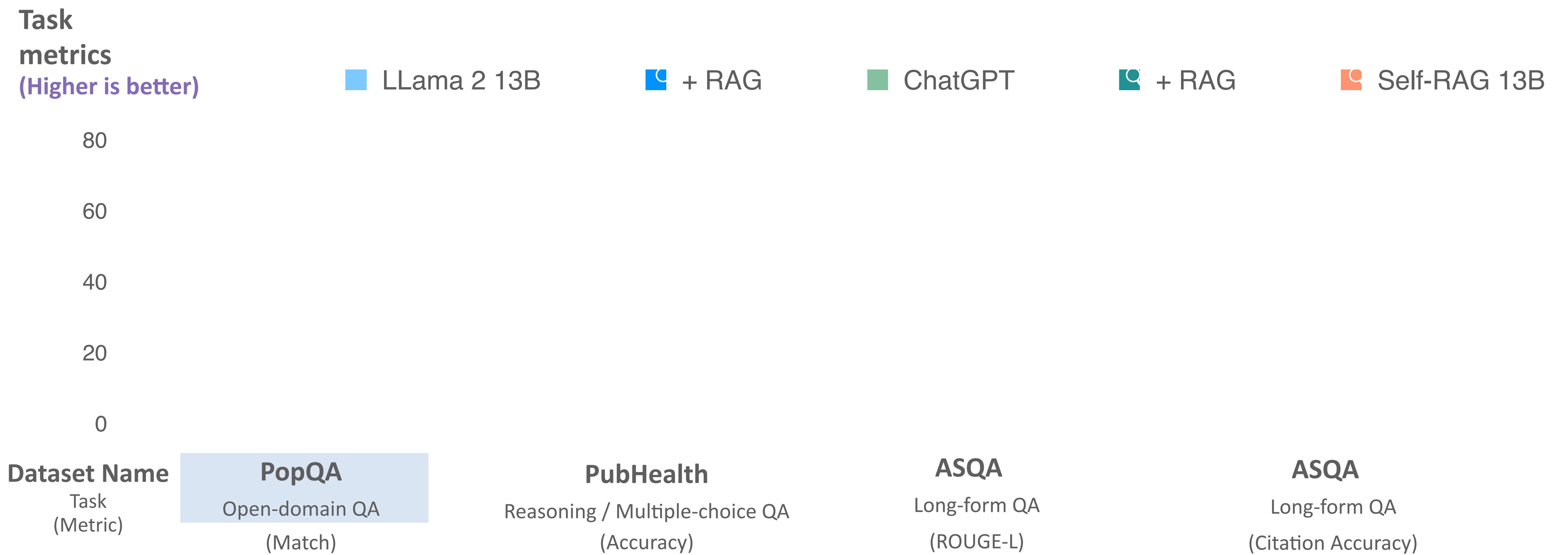
Results



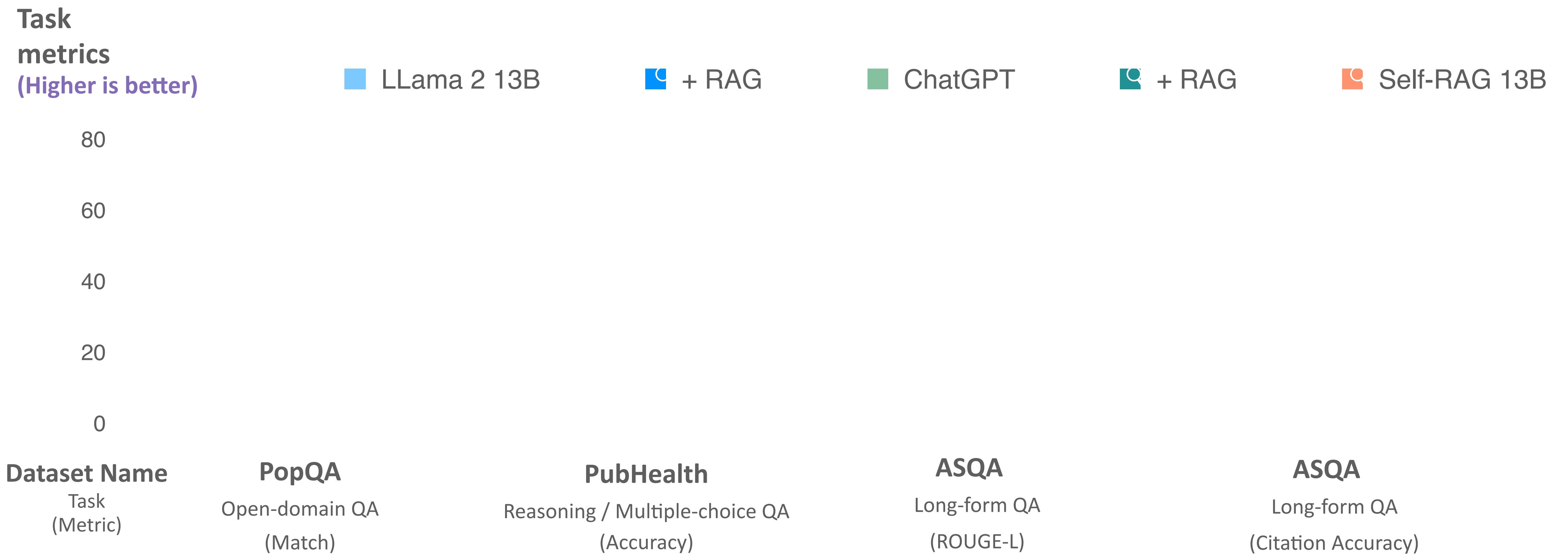
Results



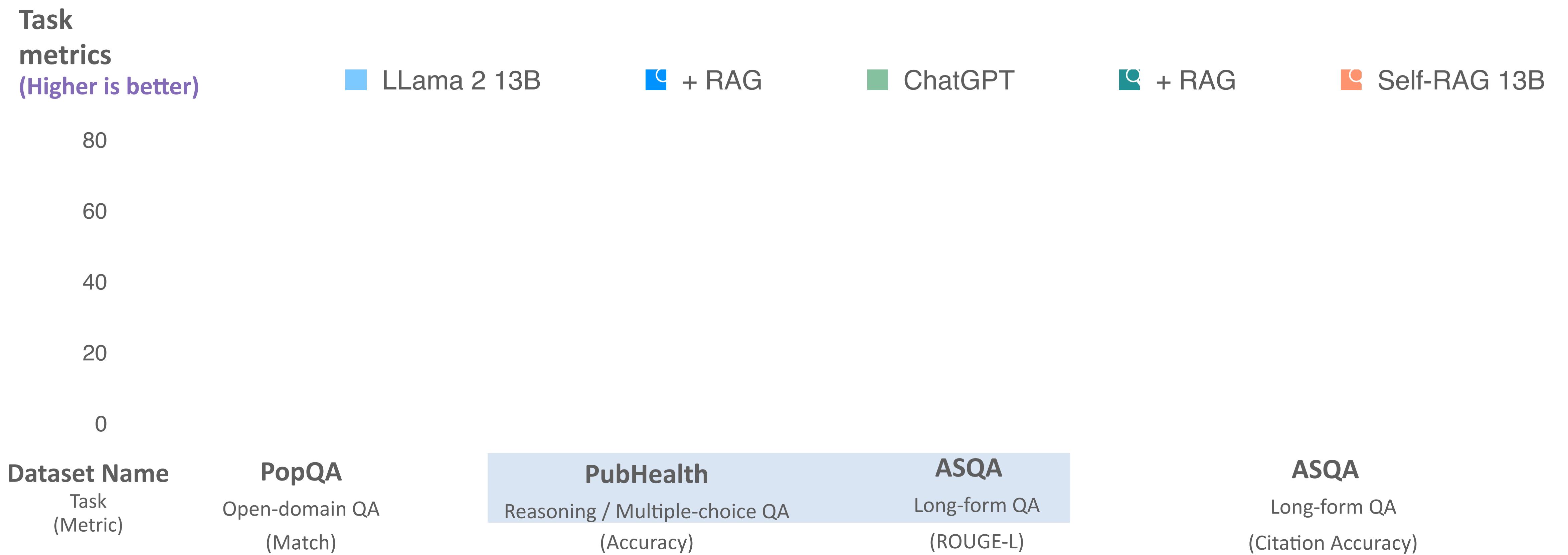
Results



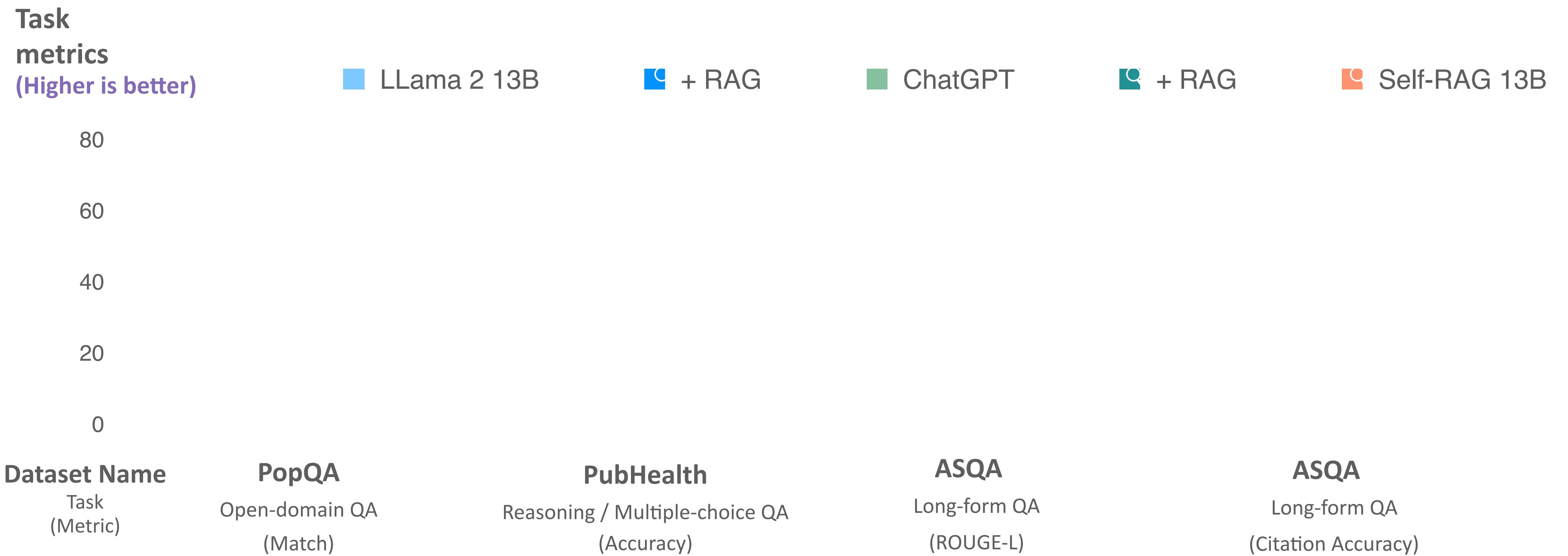
Results



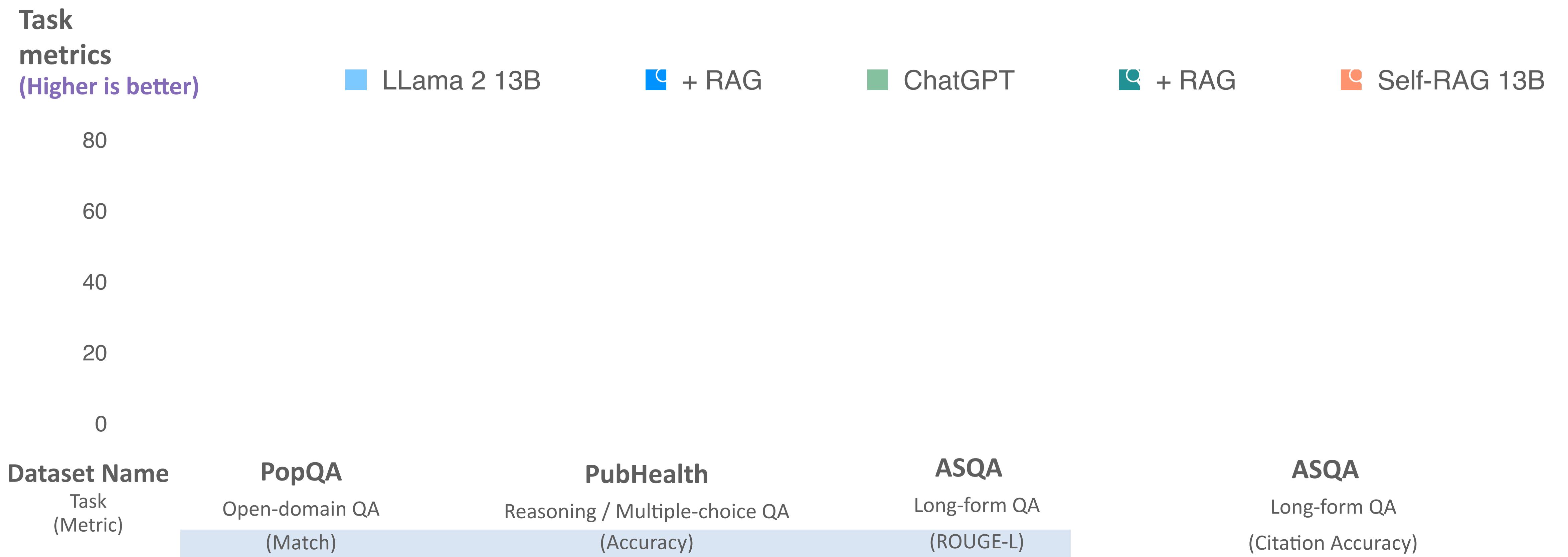
Results



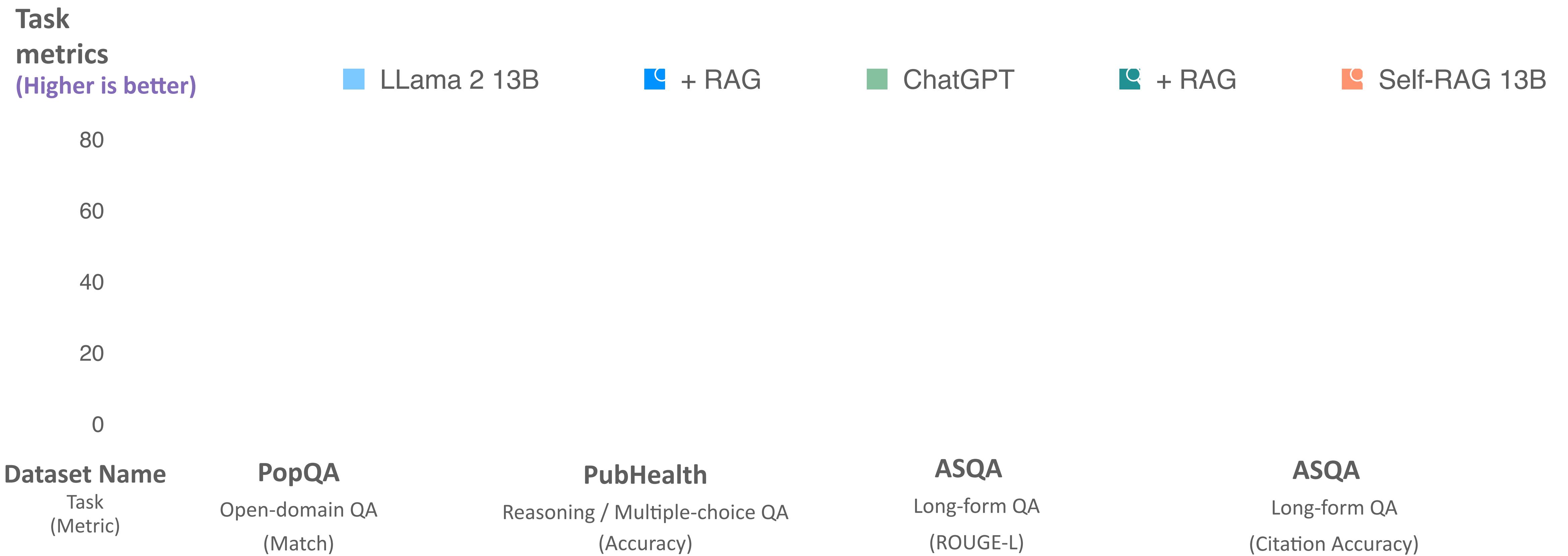
Results



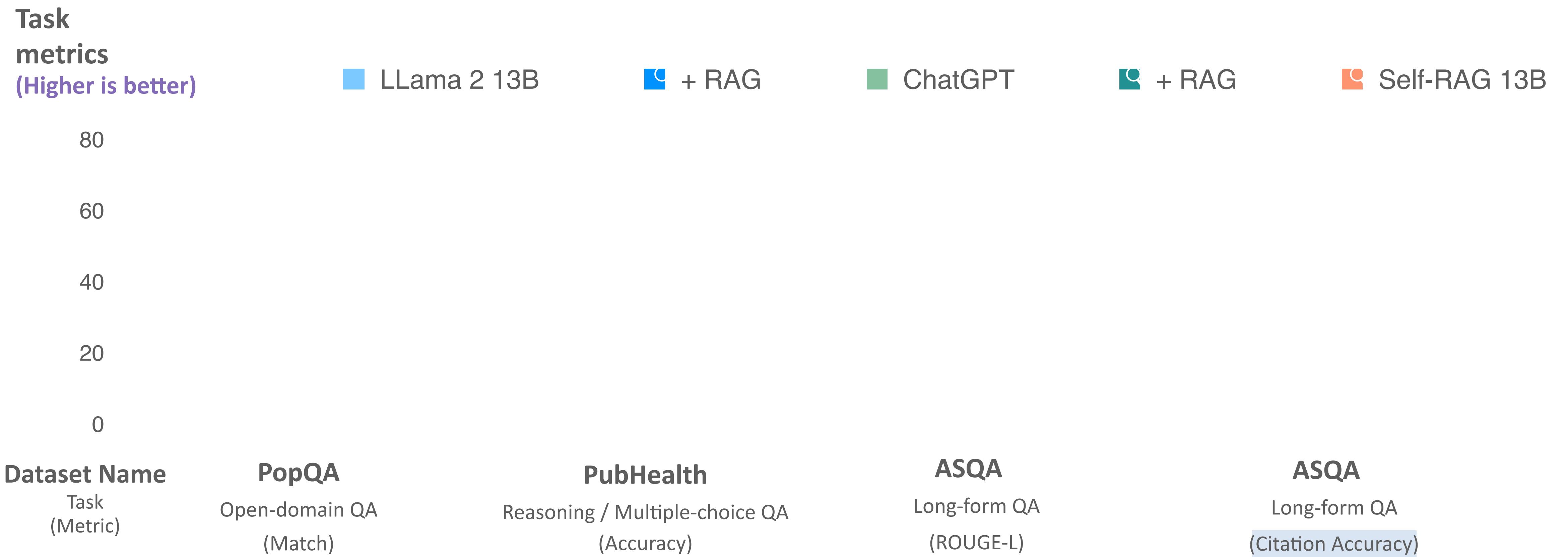
Results



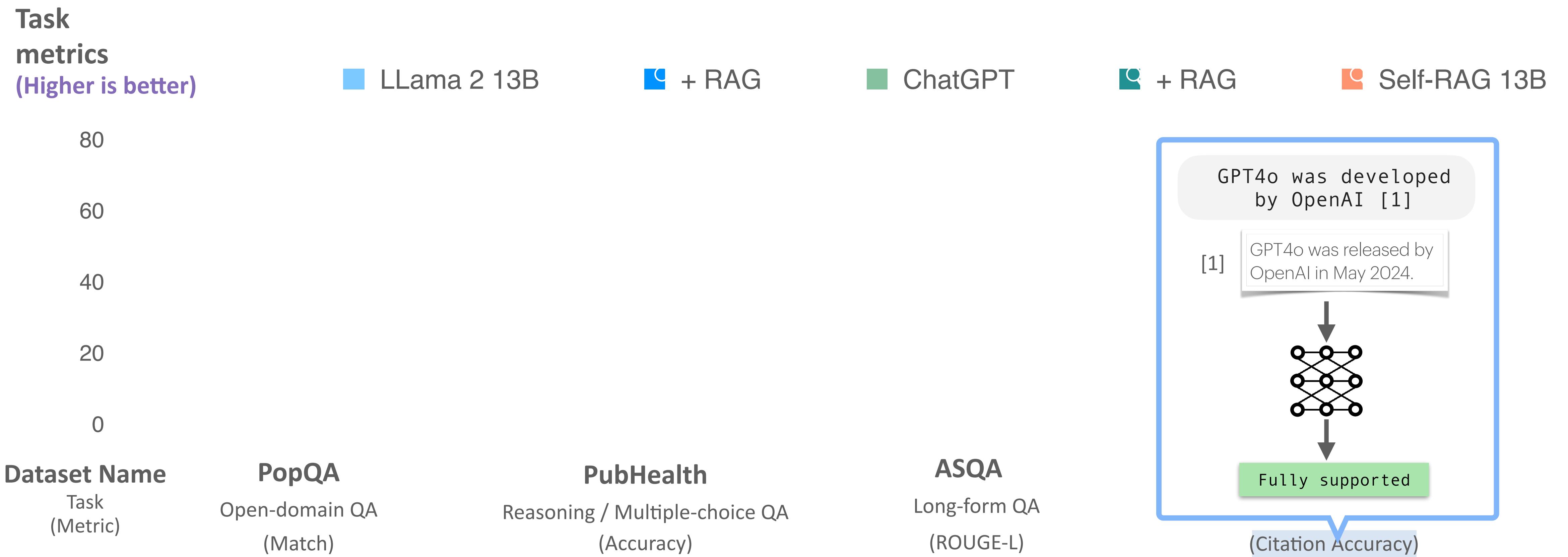
Results



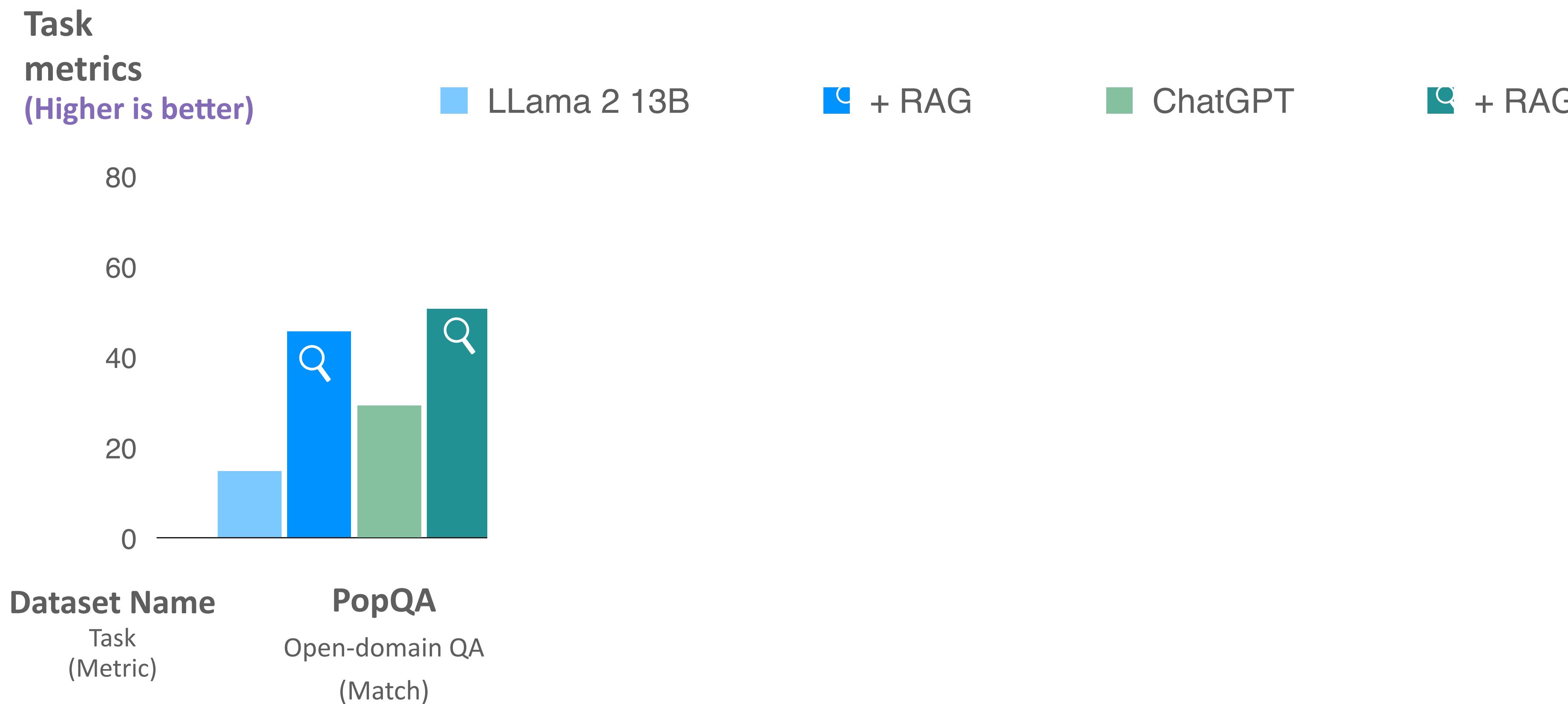
Results



Results



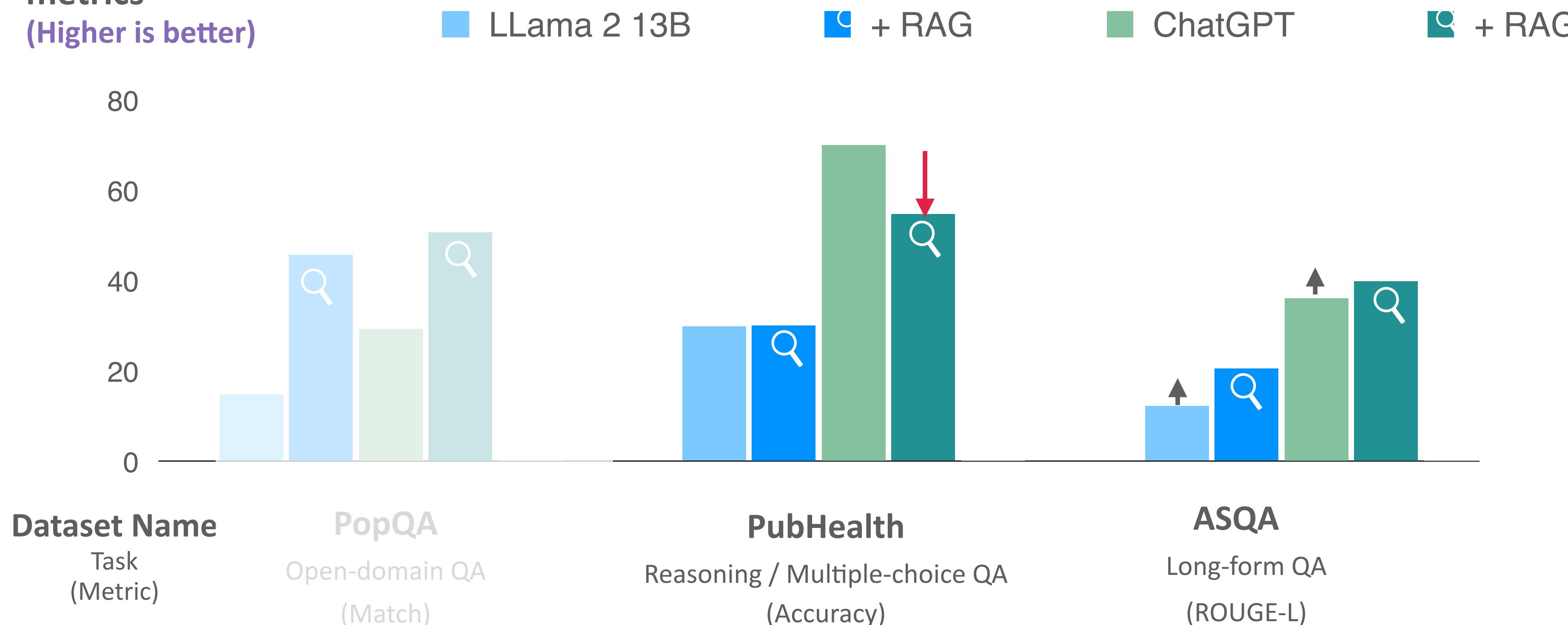
Results



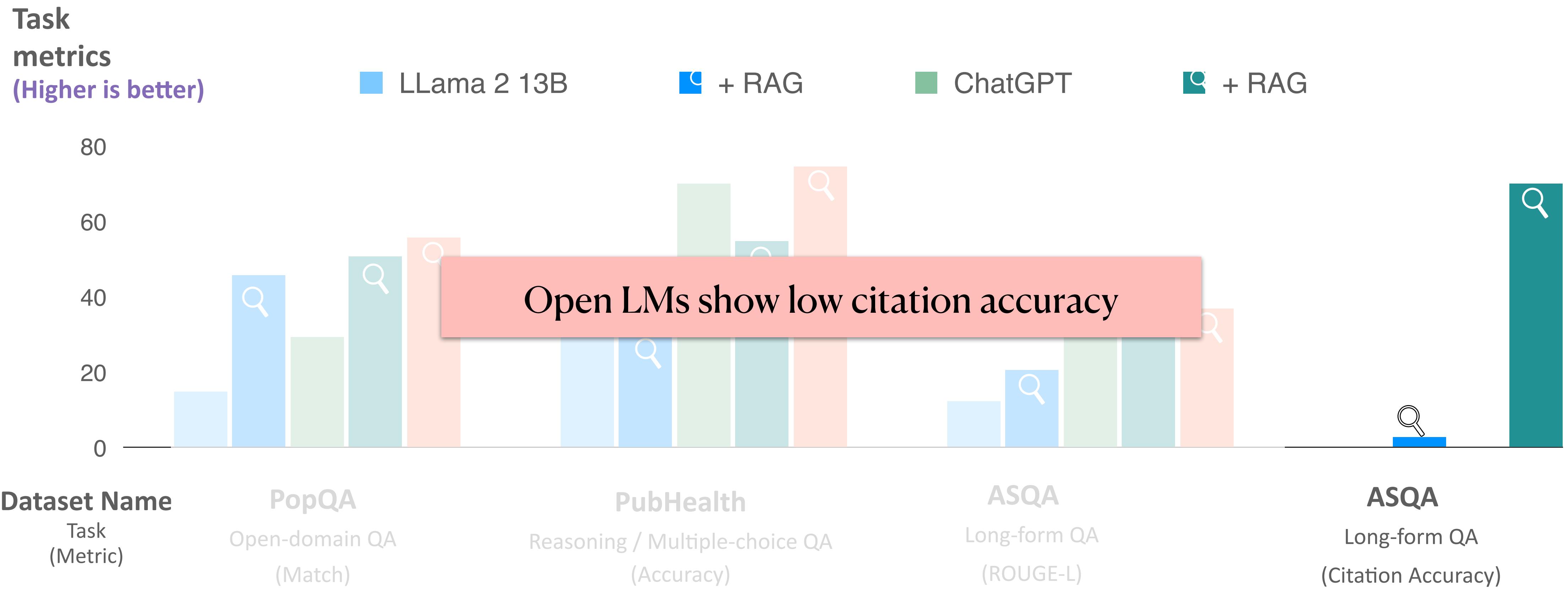
Results

Standard RAG struggles to incorporate knowledge in complex scenarios

Task
metrics
(Higher is better)



Results

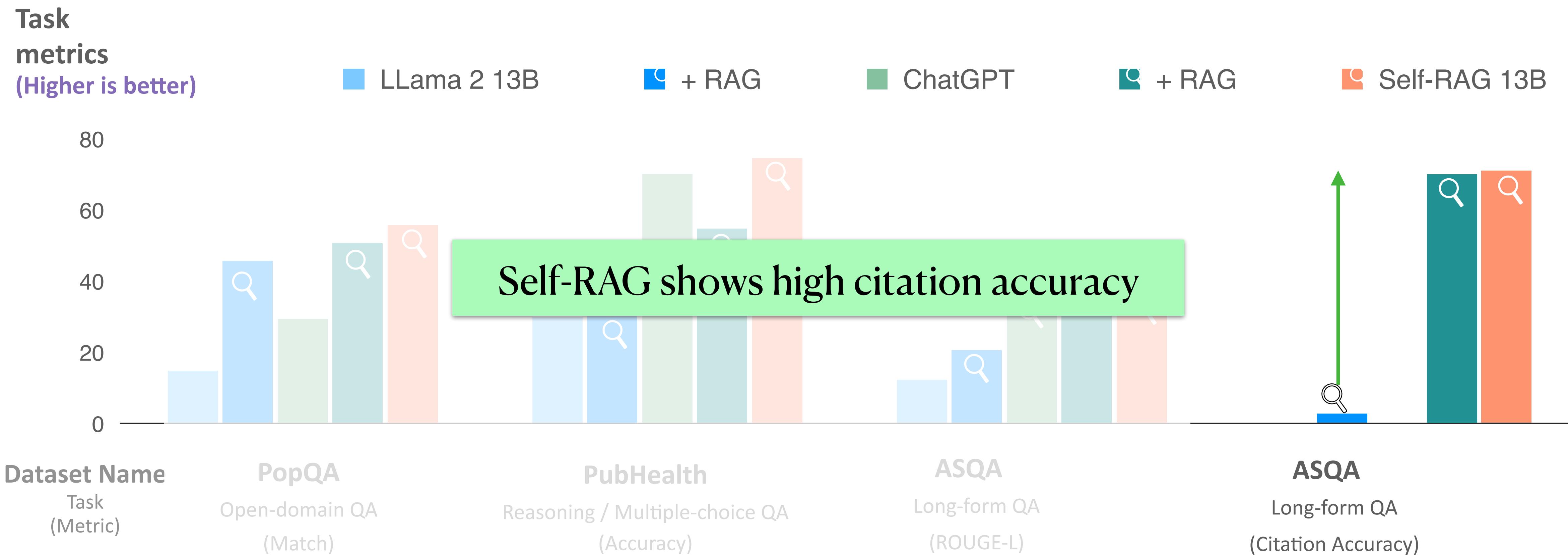


Results

Self-RAG outperforms or matches other models



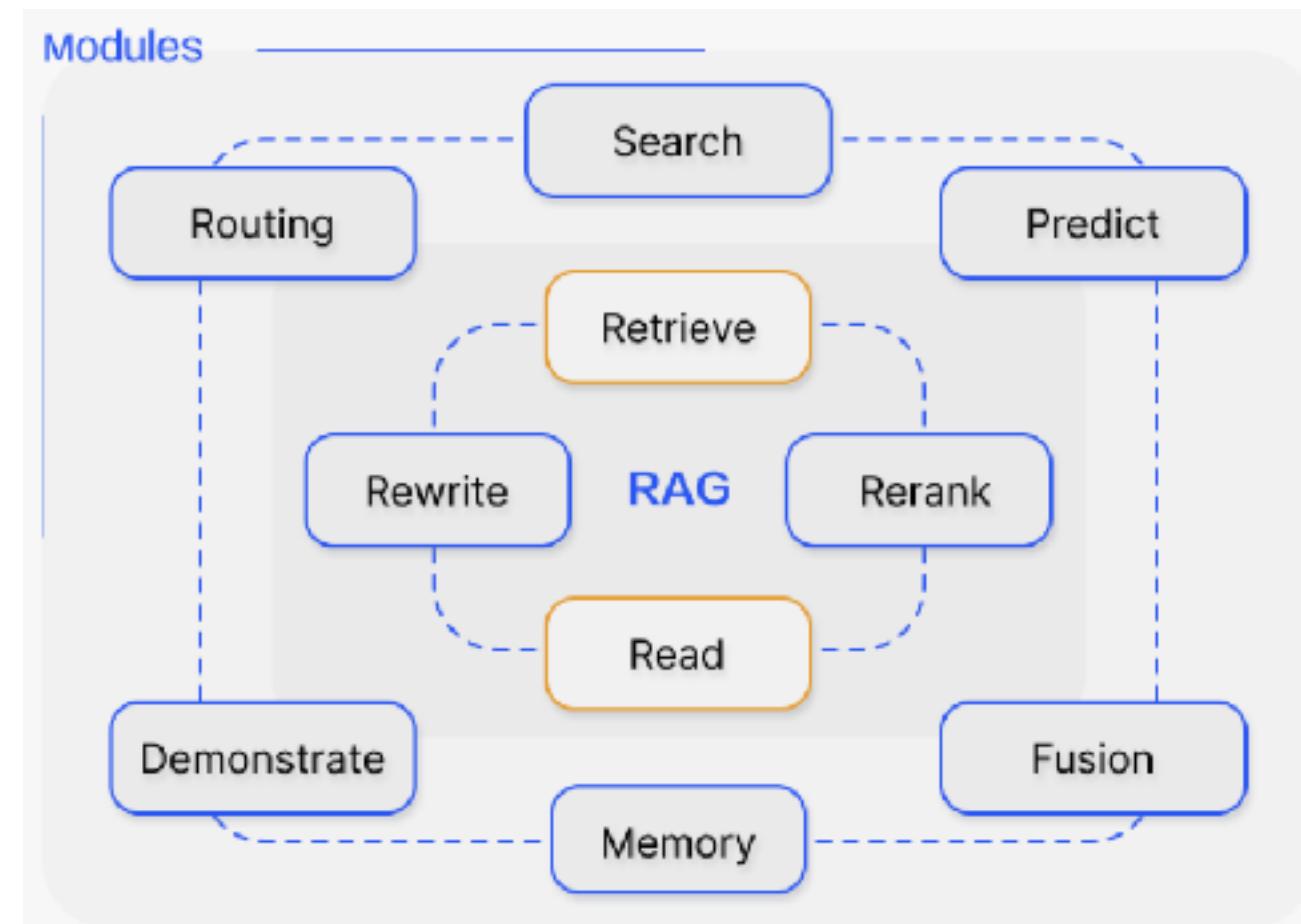
Results



Impact

Impact

New paradigms for advanced RAG

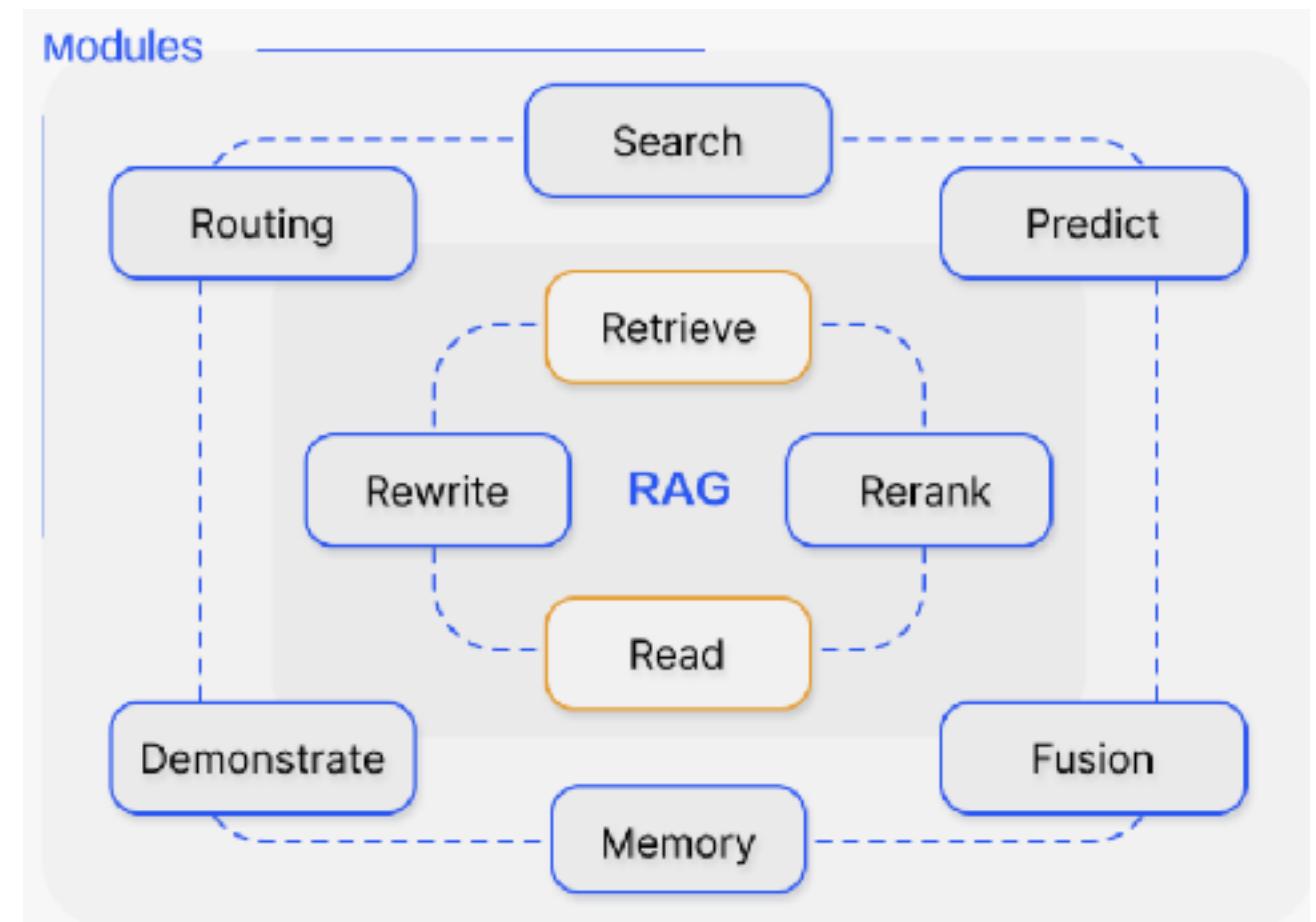


(Gao et al., 2024)

Fork 179 ▾ Starred 2k ▾

Impact

New paradigms for advanced RAG



(Gao et al., 2024)



Applications in expert domains

Self-BioRAG

(Jeong et al., Bioinformatics 2024)

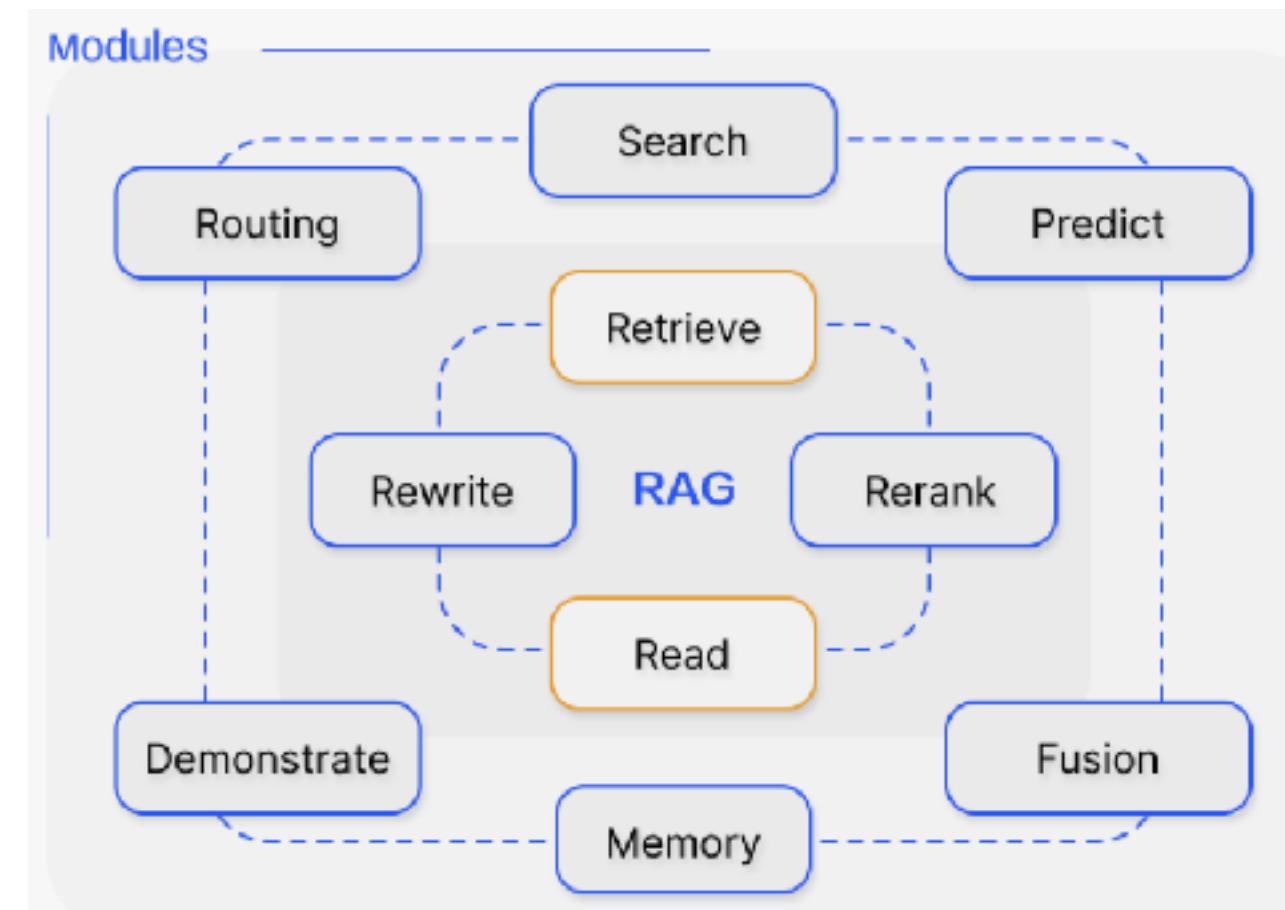
ScholarCopilot
(TIGERLab, 2024)

SciPhi-Self-RAG
(Colegrove, 2023)



Impact

New paradigms for advanced RAG



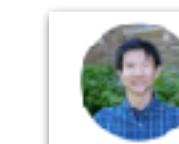
(Gao et al., 2024)

Fork 179 Starred 2k

Rapid industry and community adoptions



[4 million /M]



Jerry Liu @jerryjliu0

Self-RAG in [@llama_index](#)

We're excited to feature **Self-RAG**, a special RAG technique where an LLM can do self-reflection for dynamic retrieval, critique, and generation (@AkariAsai et al.).



[24 million /M]



LangChain @LangChainAI · Feb 13, 2024

Self RAG

Self-reflection is one of the most interesting ideas in RAG, giving an LLM the ability to self-correct problems in retrieval and / or generation.

Applications in expert domains

Self-BioRAG

(Jeong et al., Bioinformatics 2024)

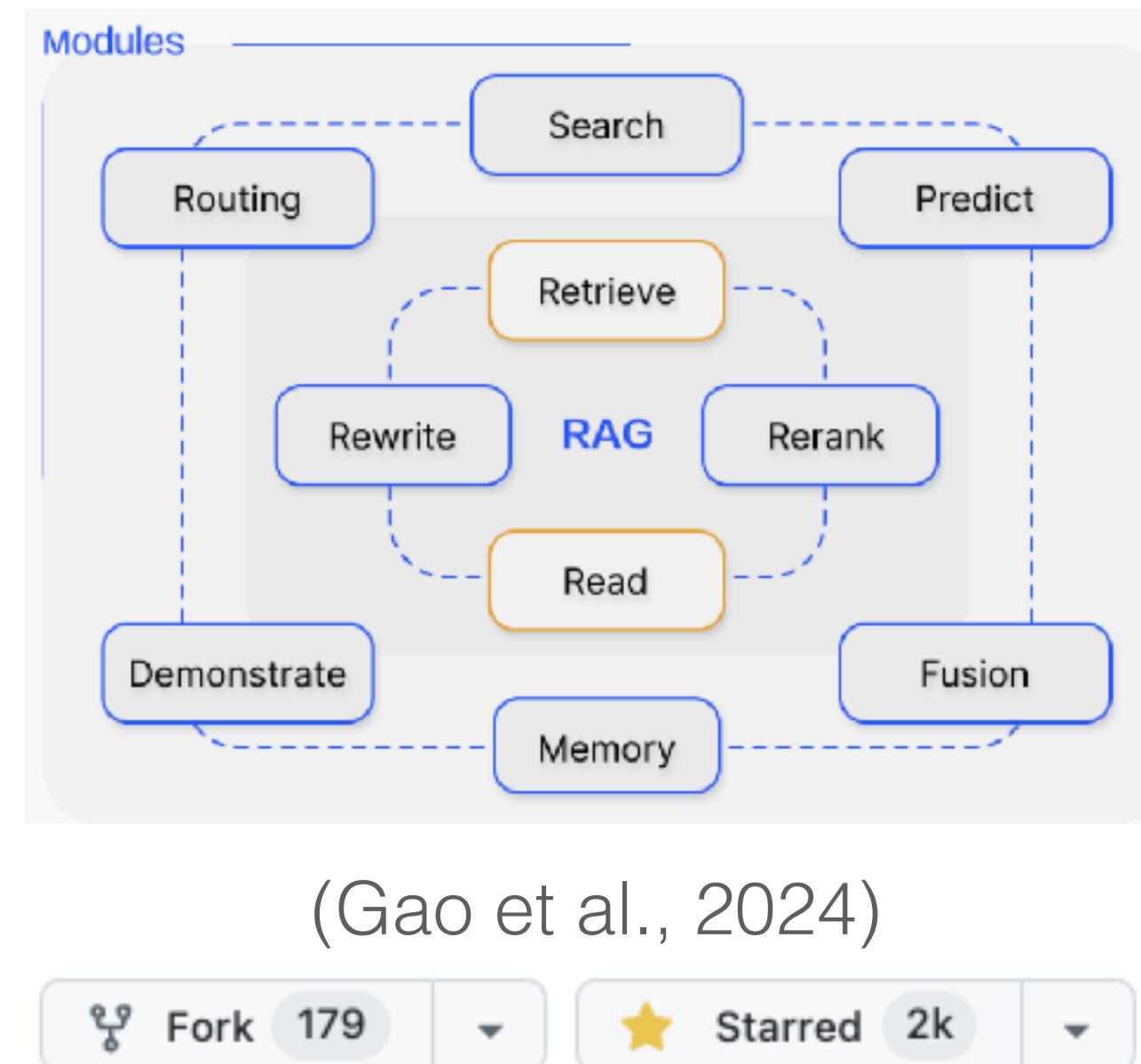
ScholarCopilot
(TIGERLab, 2024)

SciPhi-Self-RAG
(Colegrove, 2023)



Impact

New paradigms for advanced RAG



Rapid industry and community adoptions



[4 million /M]



Self-RAG in @llama_index

We're excited to feature **Self-RAG**, a special RAG technique where an LLM can do self-reflection for dynamic retrieval, critique, and generation (@AkariAsai et al.).



[24 million /M]



Self RAG

Self-reflection is one of the most interesting ideas in **RAG**, giving an LLM the ability to **self-correct** problems in retrieval and / or generation.

Applications in expert domains

Self-BioRAG

(Jeong et al., Bioinformatics 2024)

ScholarCopilot

(TIGERLab, 2024)

SciPhi-Self-RAG

(Colegrove, 2023)



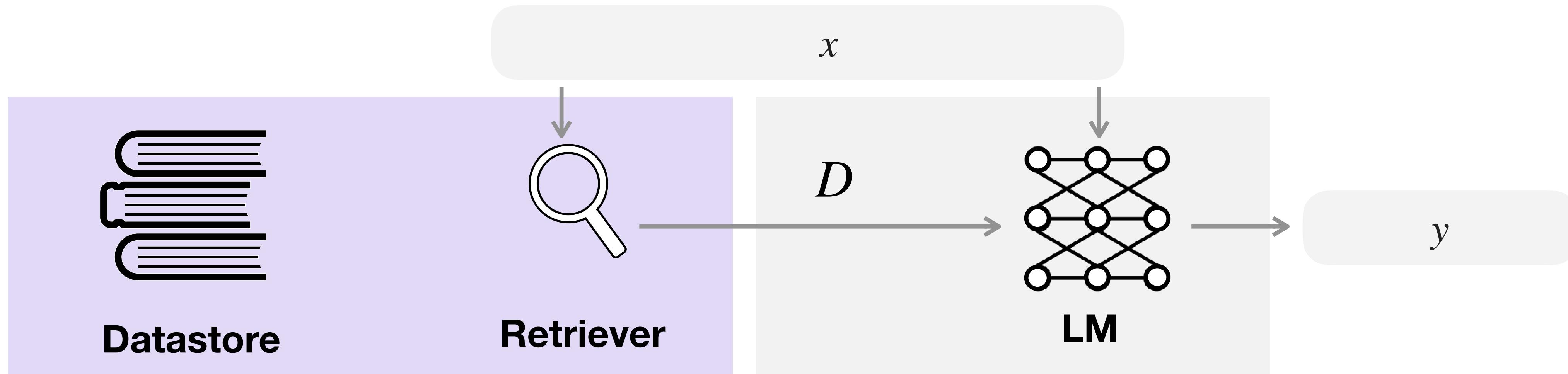
Council Post: How RAG-Powered AI Applications Have A Positive Impact On Businesses

How RAG-Powered AI Applications Have A Positive Impact On Businesses · 1. Extra processing load and increased latency due to the retrieval...



<https://www.forbes.com/councils/forbestechcouncil/2024/07/30/how-rag-powered-ai-applications-have-a-positive-impact-on-businesses/>

New Foundations of Retrieval-Augmented LMs



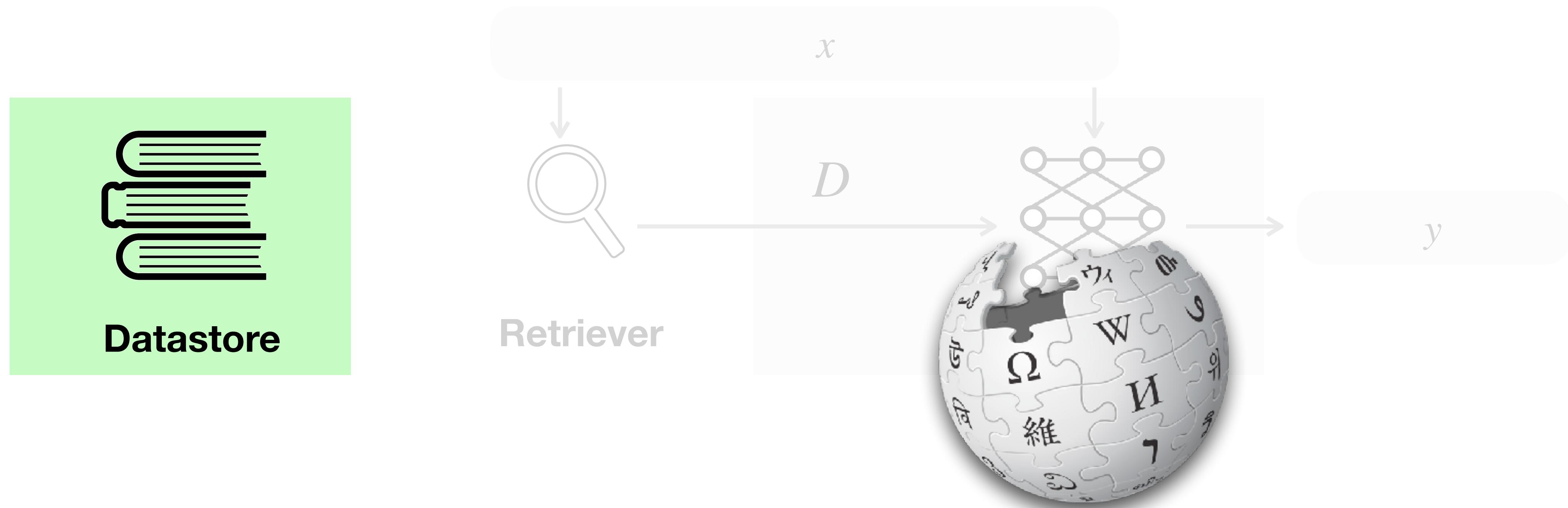
Goal

Develop a versatile and adaptive retrieval mechanism to enhance D

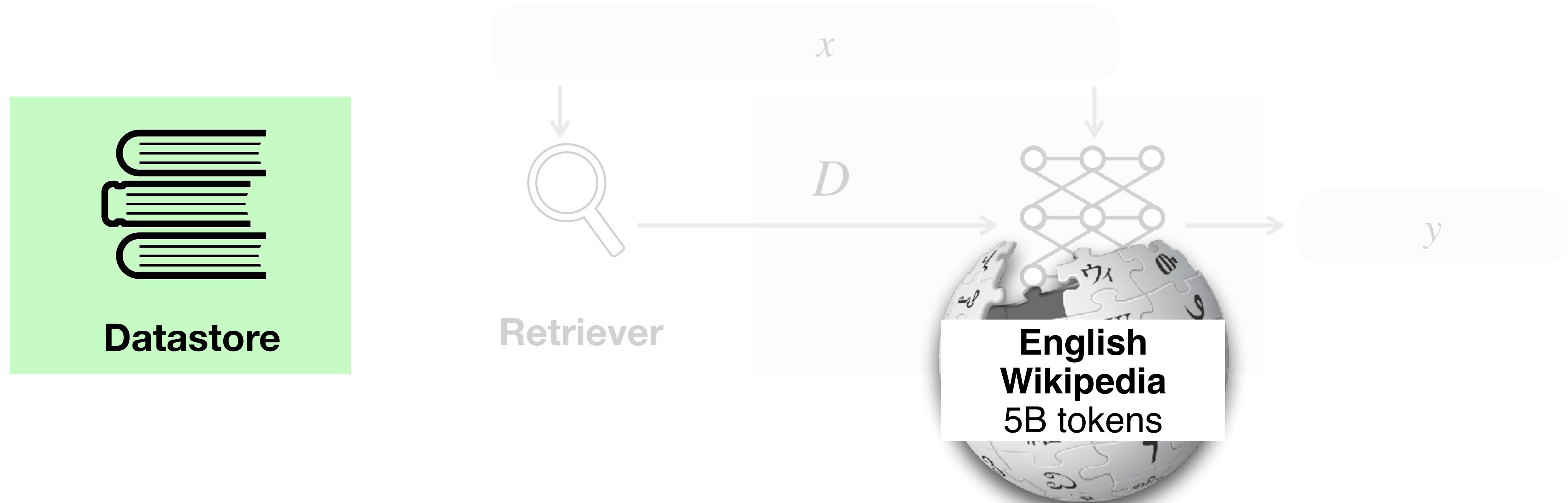
Goal

Design and train LMs to effectively and efficiently incorporate D

Scaling Datastore *Massively*



Scaling Datastore *Massively*



Scaling Datastore *Massively*

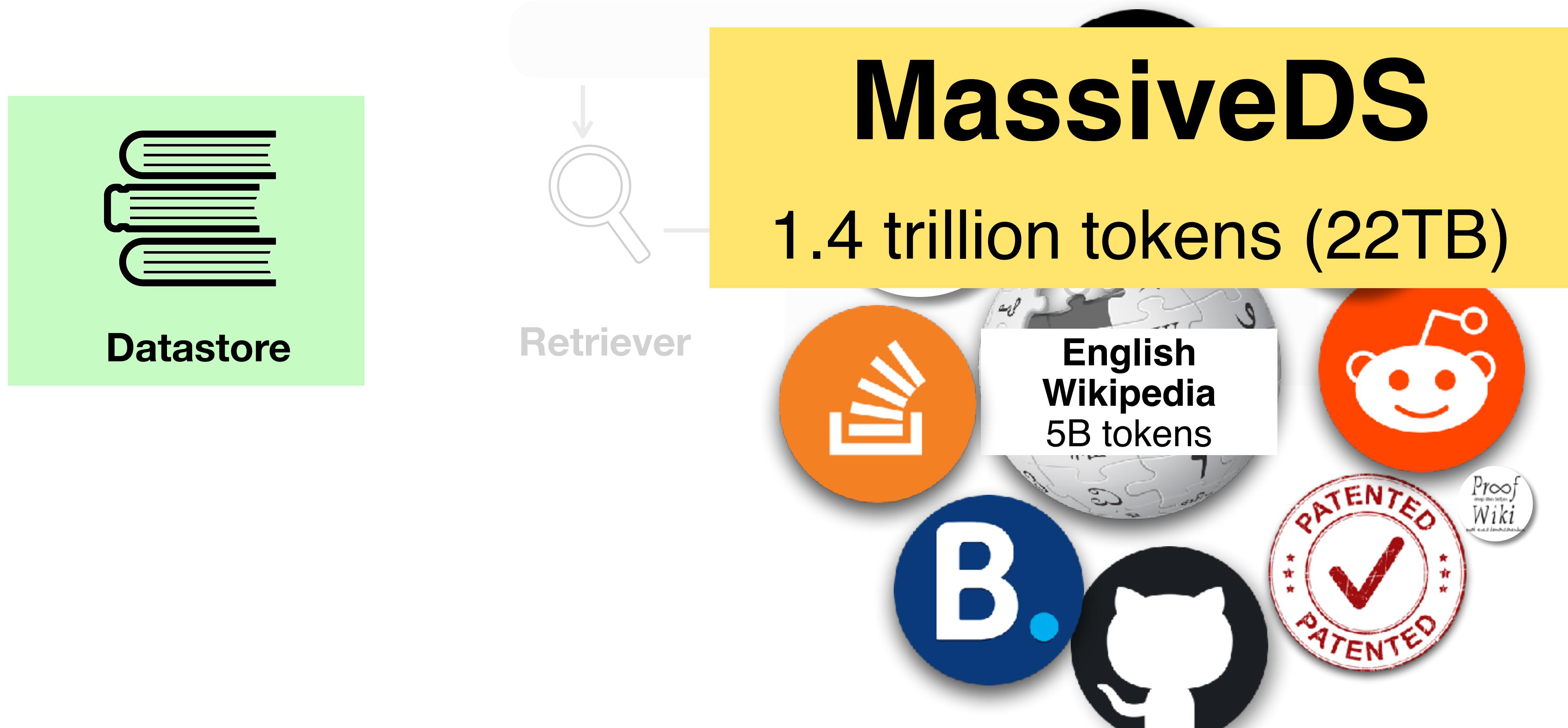


Scaling Datastore *Massively*



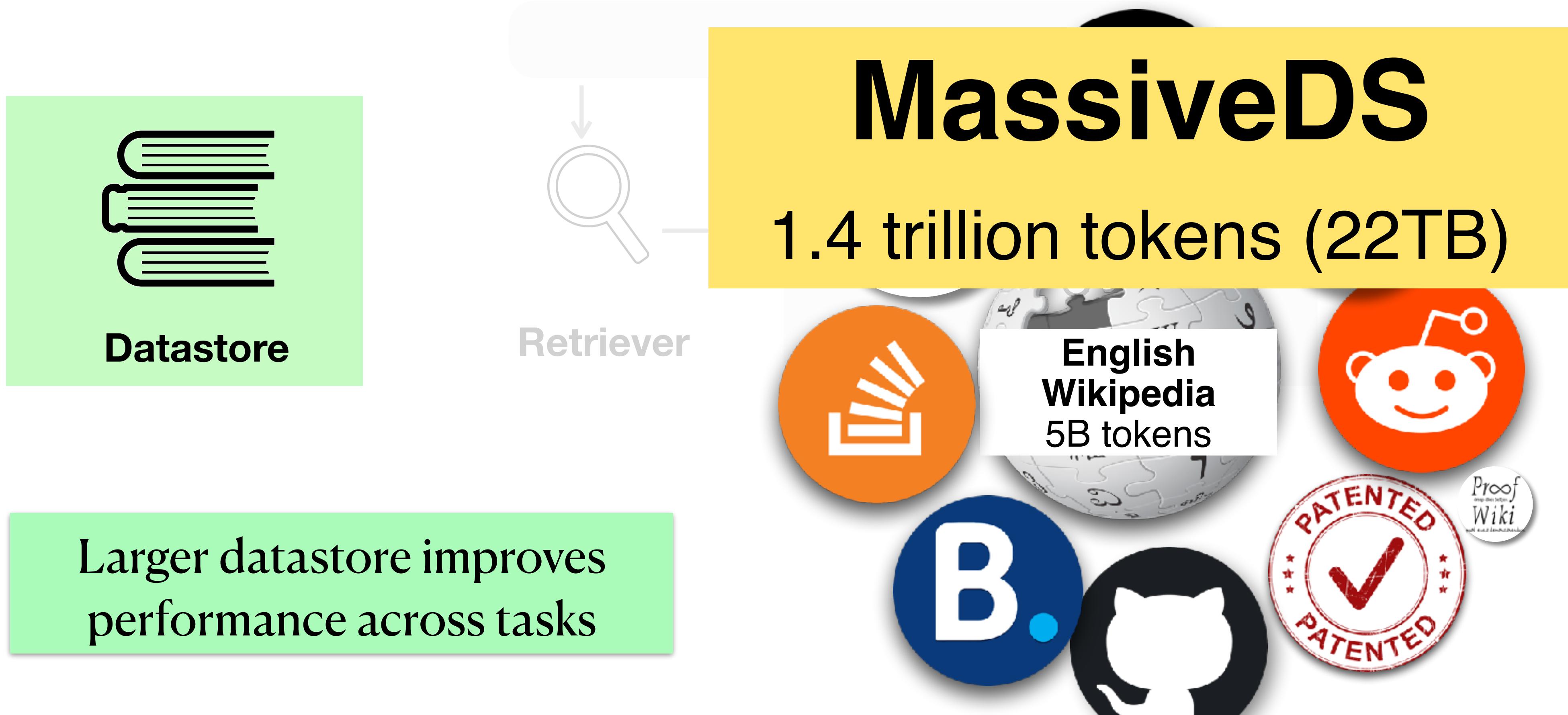
Shao, He, Asai et al. NeurIPS 2024. Scaling Retrieval-Based Language Models with a Trillion-Token Datastore.

Scaling Datastore *Massively*



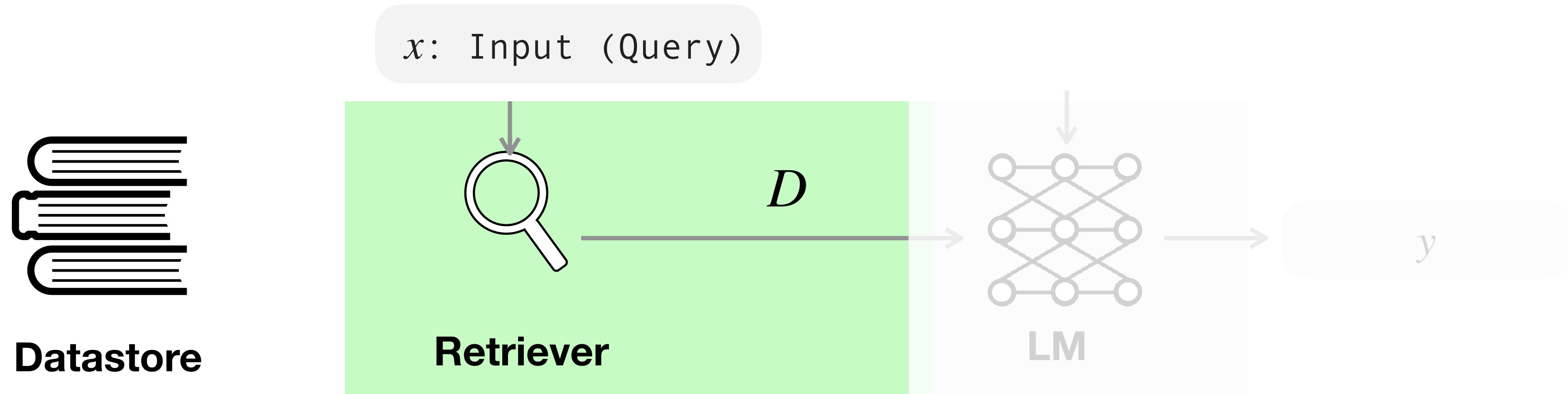
Shao, He, Asai et al. NeurIPS 2024. Scaling Retrieval-Based Language Models with a Trillion-Token Datastore.

Scaling Datastore *Massively*



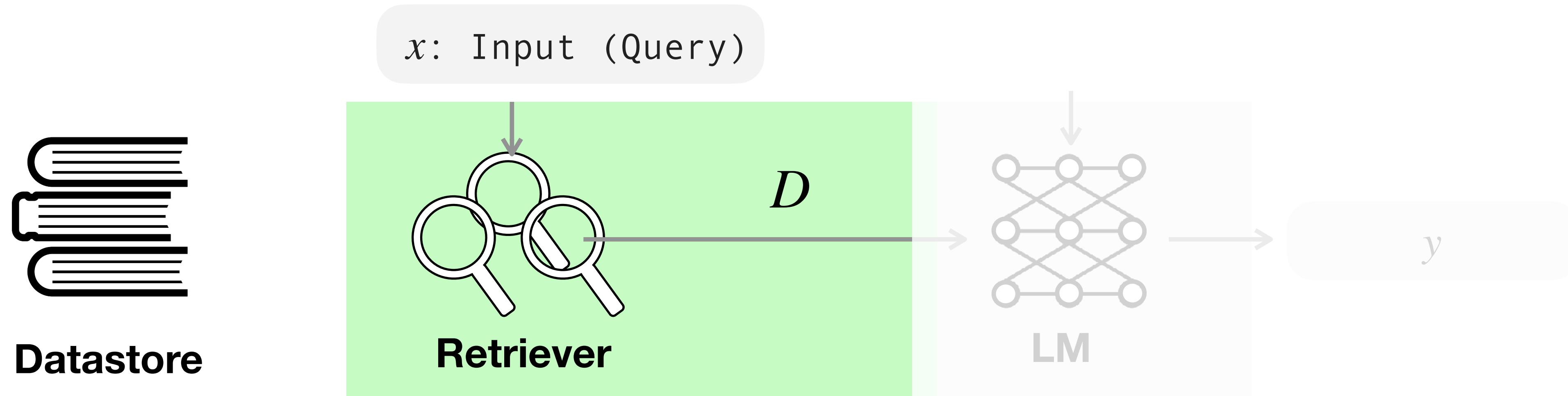
Shao, He, Asai et al. NeurIPS 2024. Scaling Retrieval-Based Language Models with a Trillion-Token Datastore.

Adaptive Instruction-Following Retrievers



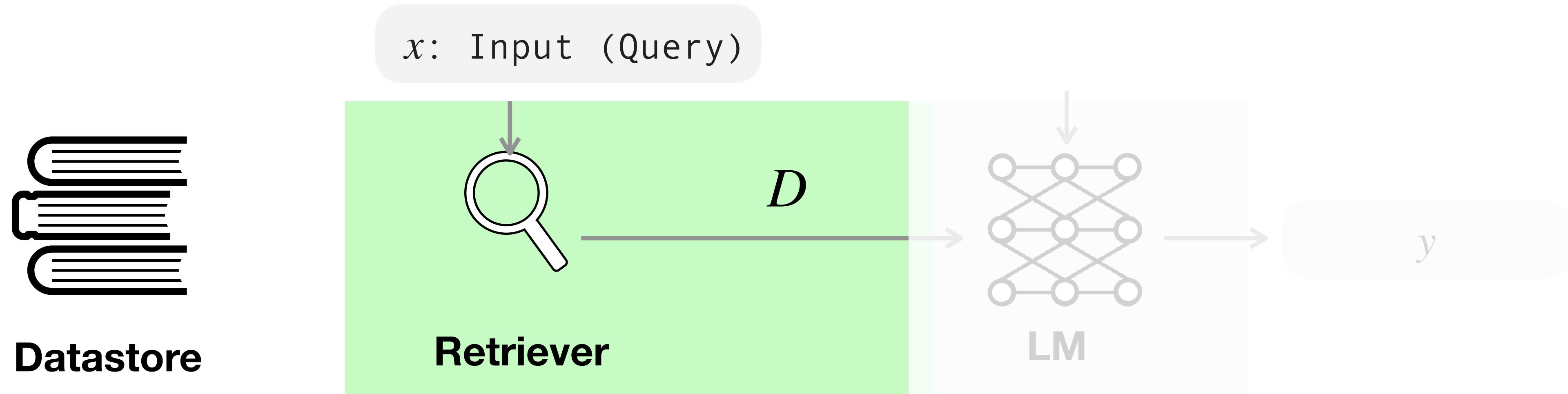
[Asai](#) et al., Findings of ACL 2023. Task-aware Retrieval with Instructions.

Adaptive Instruction-Following Retrievers



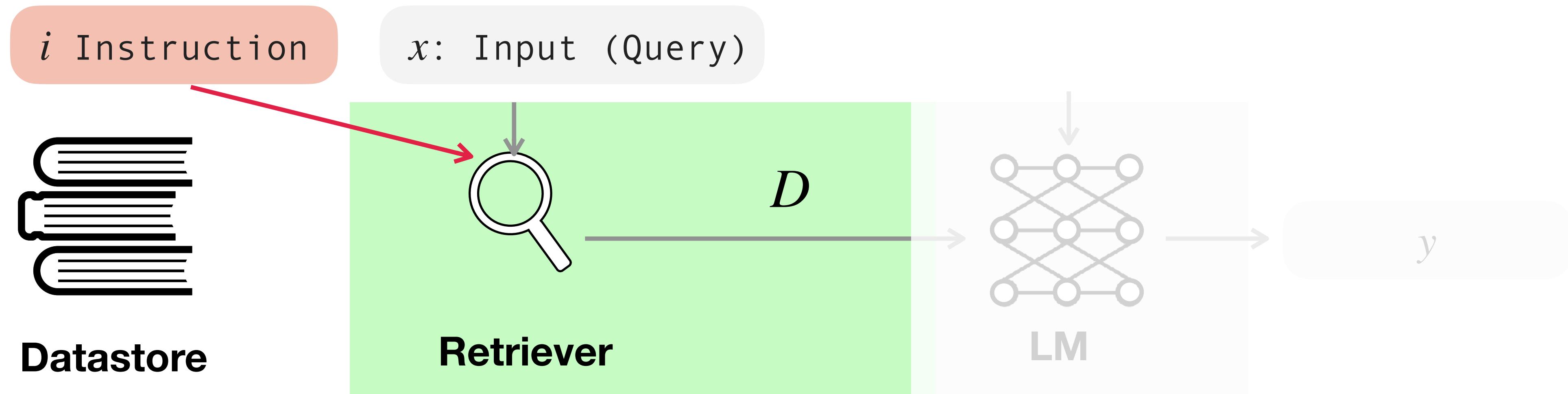
Asai et al., Findings of ACL 2023. Task-aware Retrieval with Instructions.

Adaptive Instruction-Following Retrievers



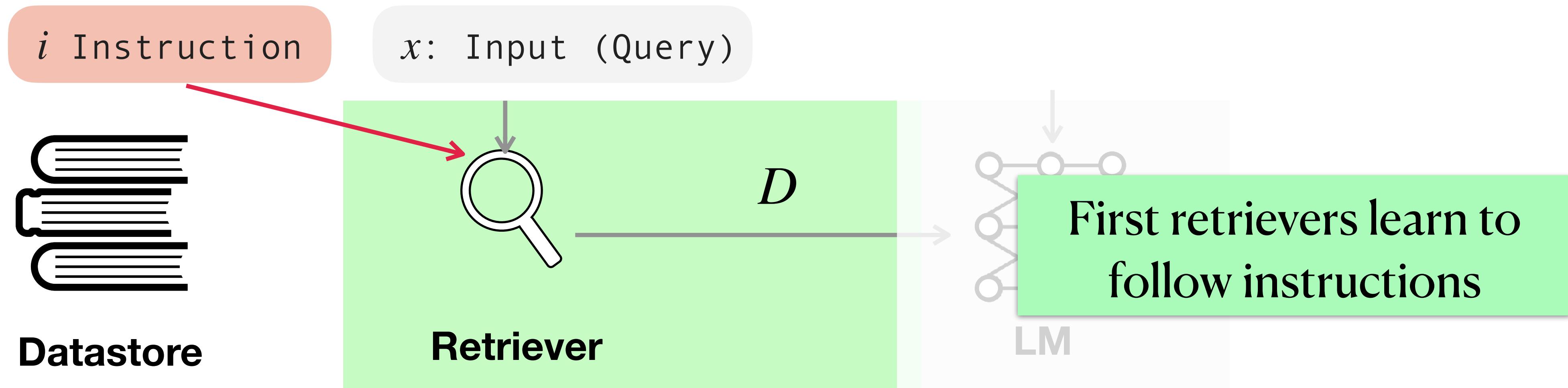
[Asai](#) et al., Findings of ACL 2023. Task-aware Retrieval with Instructions.

Adaptive Instruction-Following Retrievers



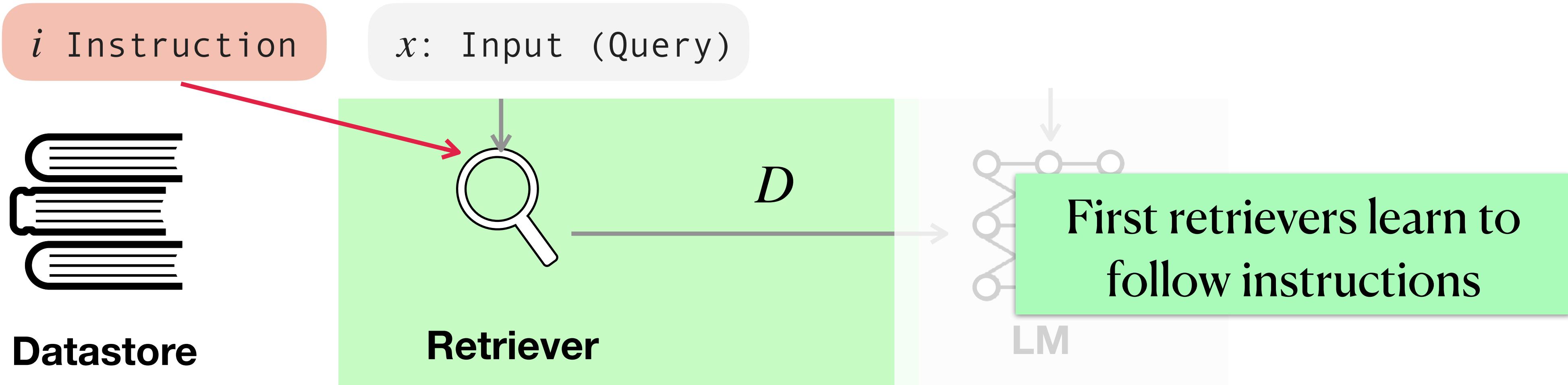
[Asai](#) et al., Findings of ACL 2023. Task-aware Retrieval with Instructions.

Adaptive Instruction-Following Retrievers



[Asai et al., Findings of ACL 2023. Task-aware Retrieval with Instructions.](#)

Adaptive Instruction-Following Retrievers



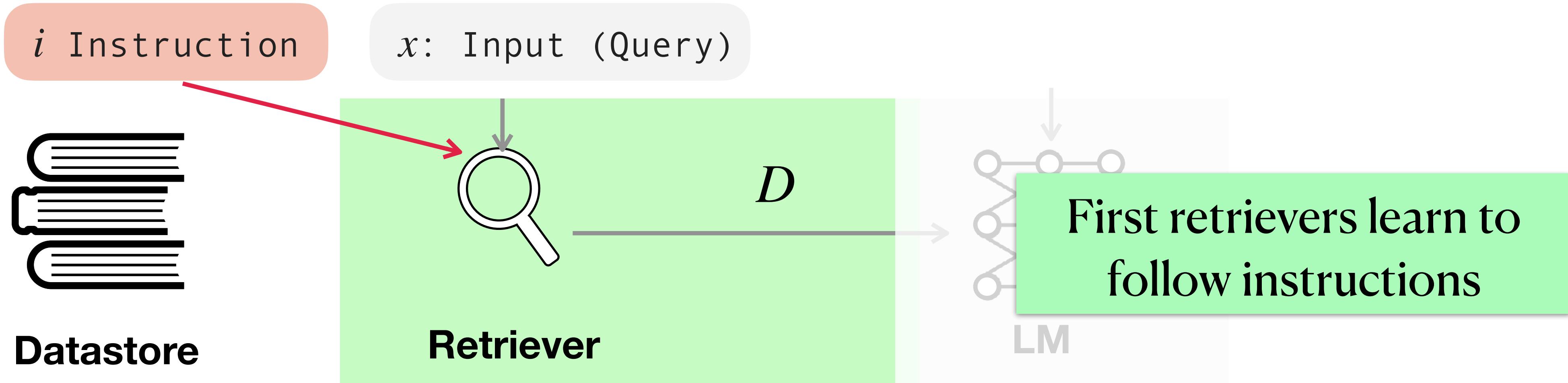
Rank	Model	Model Size (Million Parameters)	Memory Usage (GB, fp32)	Embedding Dimensions	Max Tokens	Average (56 datasets)
1	voyage-3-m-exp					74.03
2	NV-Embed-v2	7851	29.25	4096	32768	72.31

https://huggingface.co/spaces/mteb/leaderboard_legacy

2024). It has also been recently applied for training retrievers and general-purpose embedding models that can adapt their output embeddings with different instructions and task types (Asai et al., 2022).

Asai et al., Findings of ACL 2023. Task-aware Retrieval with Instructions.

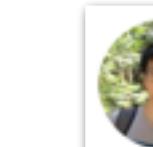
Adaptive Instruction-Following Retrievers



Rank	Model	Model Size (Million Parameters)	Memory Usage (GB, fp32)	Embedding Dimensions	Max Tokens	Average (56 datasets)
1	voyage-3-m-exp					74.03
2	NV-Embed-v2	7851	29.25	4096	32768	72.31

https://huggingface.co/spaces/mteb/leaderboard_legacy

UniIR ☀️: Training and Benchmarking Universal Multimodal Information Retrievers



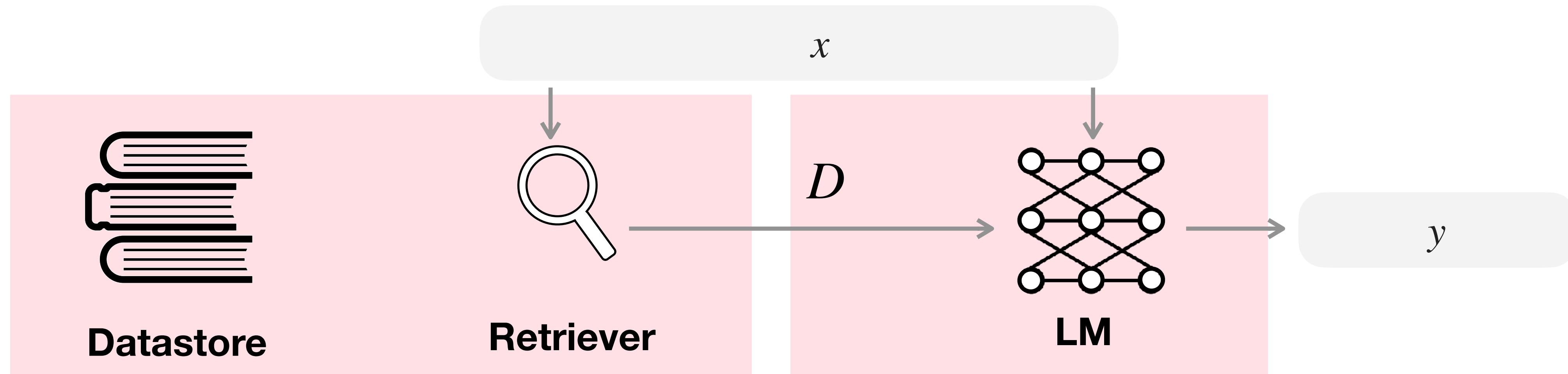
Wenhua Chen @WenhuaChen · Nov 30, 2023 [XL ...](#)
We found that the instruction tuning is crucial for not only boosting models' performance but also enabling the model to generalize to unseen information retrieval tasks. The gain is dramatic in most of the tasks. This project was partly inspired by [@AkariAsai's TART paper](#).

2024). It has also been recently applied for training retrievers and general-purpose embedding models that can adapt their output embeddings with different instructions and task types (Asai et al., 2022).

<https://x.com/WenhuaChen/status/1730326900533244150>

Asai et al., Findings of ACL 2023. Task-aware Retrieval with Instructions.

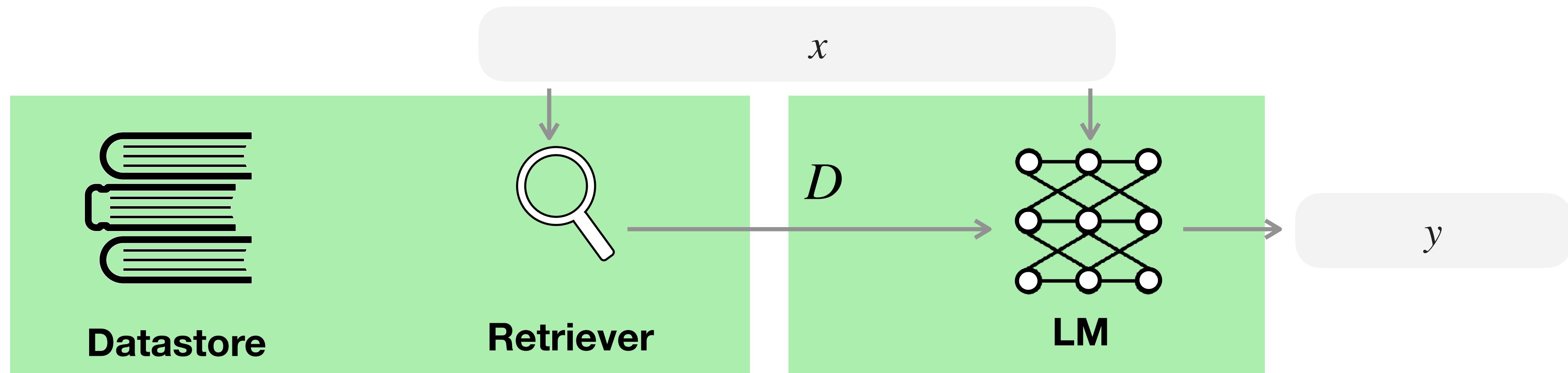
Summary



Limitations of common RAG pipelines

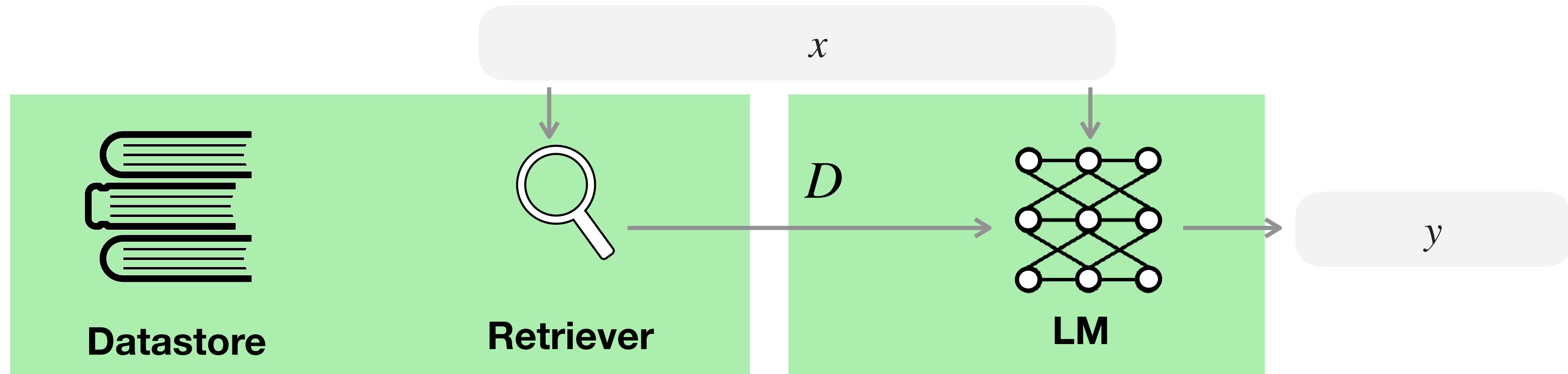
- :(Fails to adapt diverse tasks
- :(Not optimized for retrieval augmentation

Summary



New foundations of Retrieval-Augmented LMs

Summary



New foundations of Retrieval-Augmented LMs

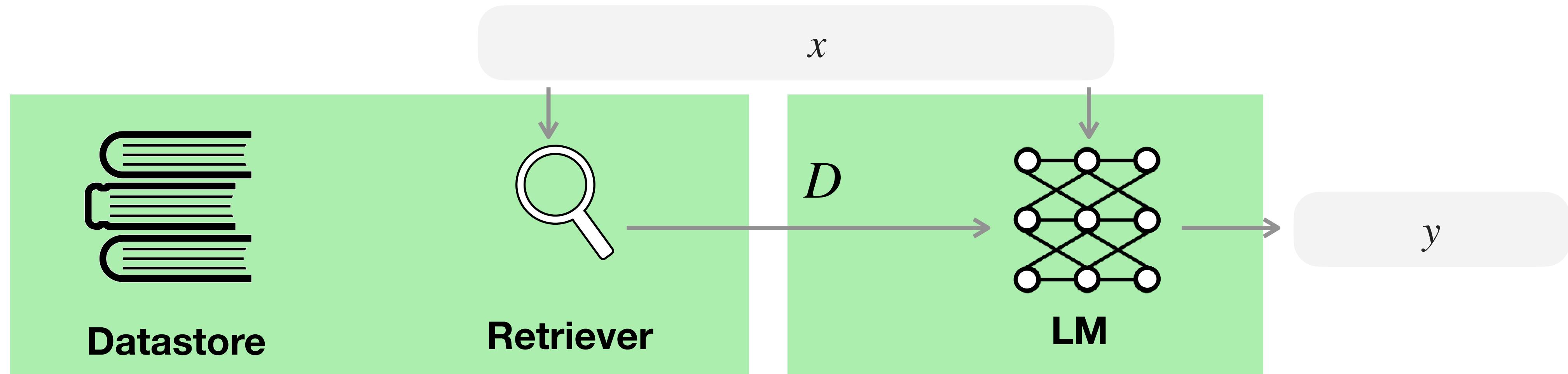


Adaptive inference



Training with retrieval

Summary



New foundations of Retrieval-Augmented LMs



Adaptive inference

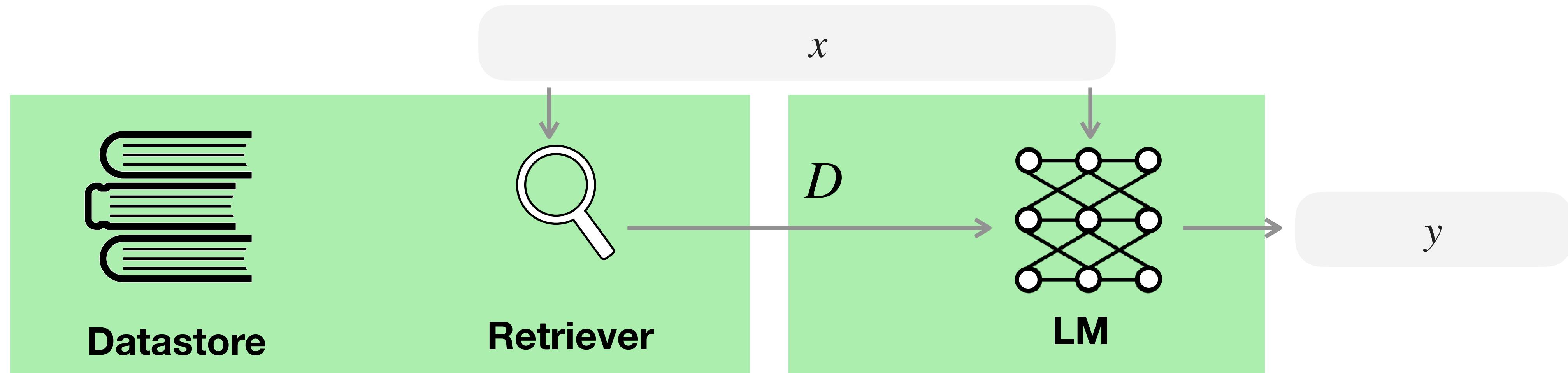


Training with retrieval



Scaling datastore

Summary



New foundations of Retrieval-Augmented LMs



Adaptive inference



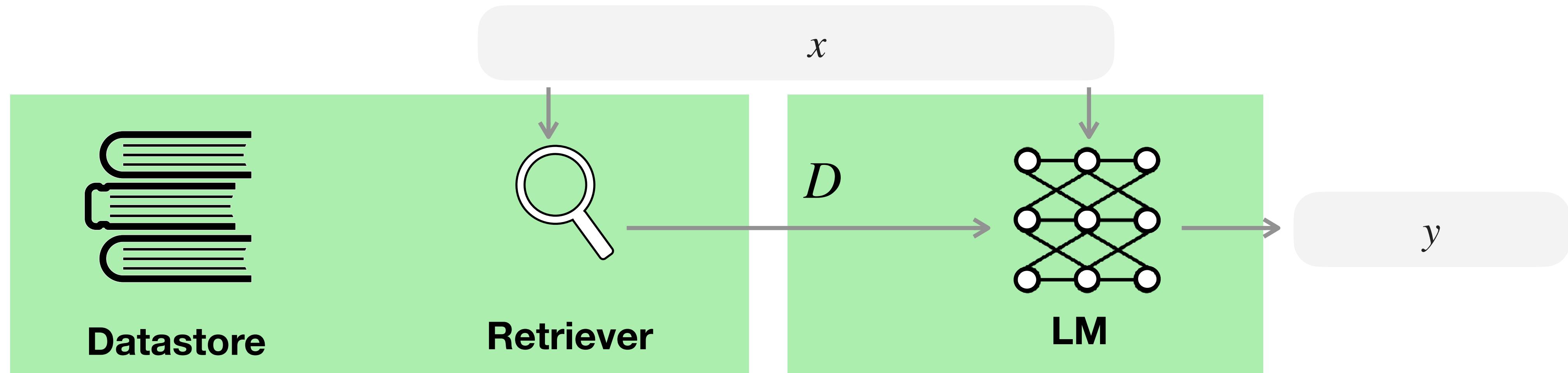
Training with retrieval



Scaling datastore

Foundations of SOTA
systems

Summary



New foundations of Retrieval-Augmented LMs



Adaptive inference



Training with retrieval

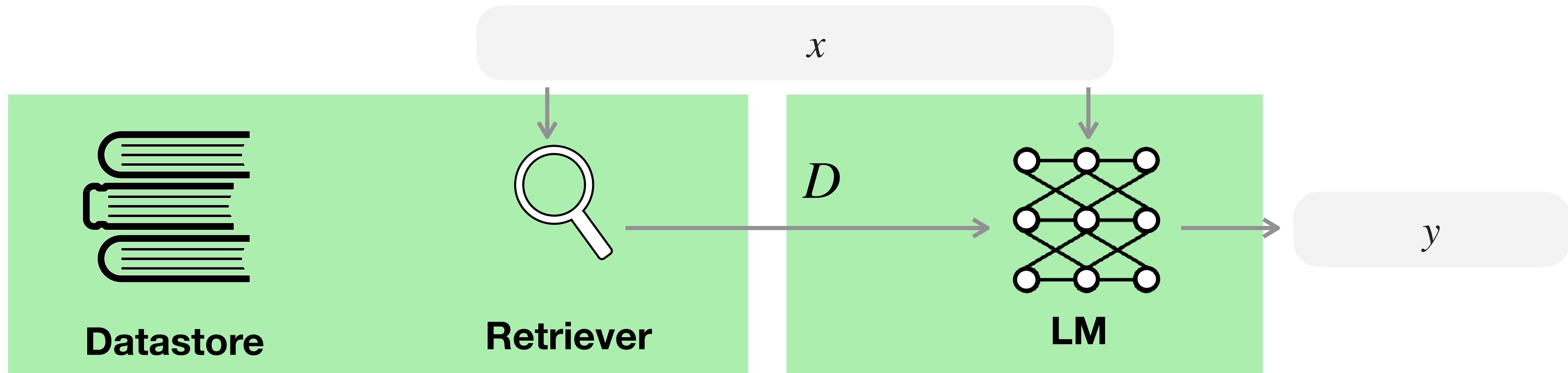


Scaling datastore

Foundations of SOTA
systems

Official support in
major libraries

Summary



New foundations of Retrieval-Augmented LMs



Adaptive inference



Training with retrieval



Scaling datastore

Foundations of SOTA systems

Official support in major libraries

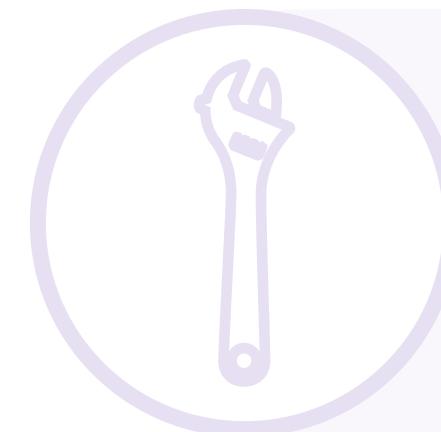
Applications to industrial systems

Frontiers of Retrieval-Augmented LMs



From Monolithic to Retrieval-Augmented LMs

Why?



New Foundations of Retrieval-Augmented LMs

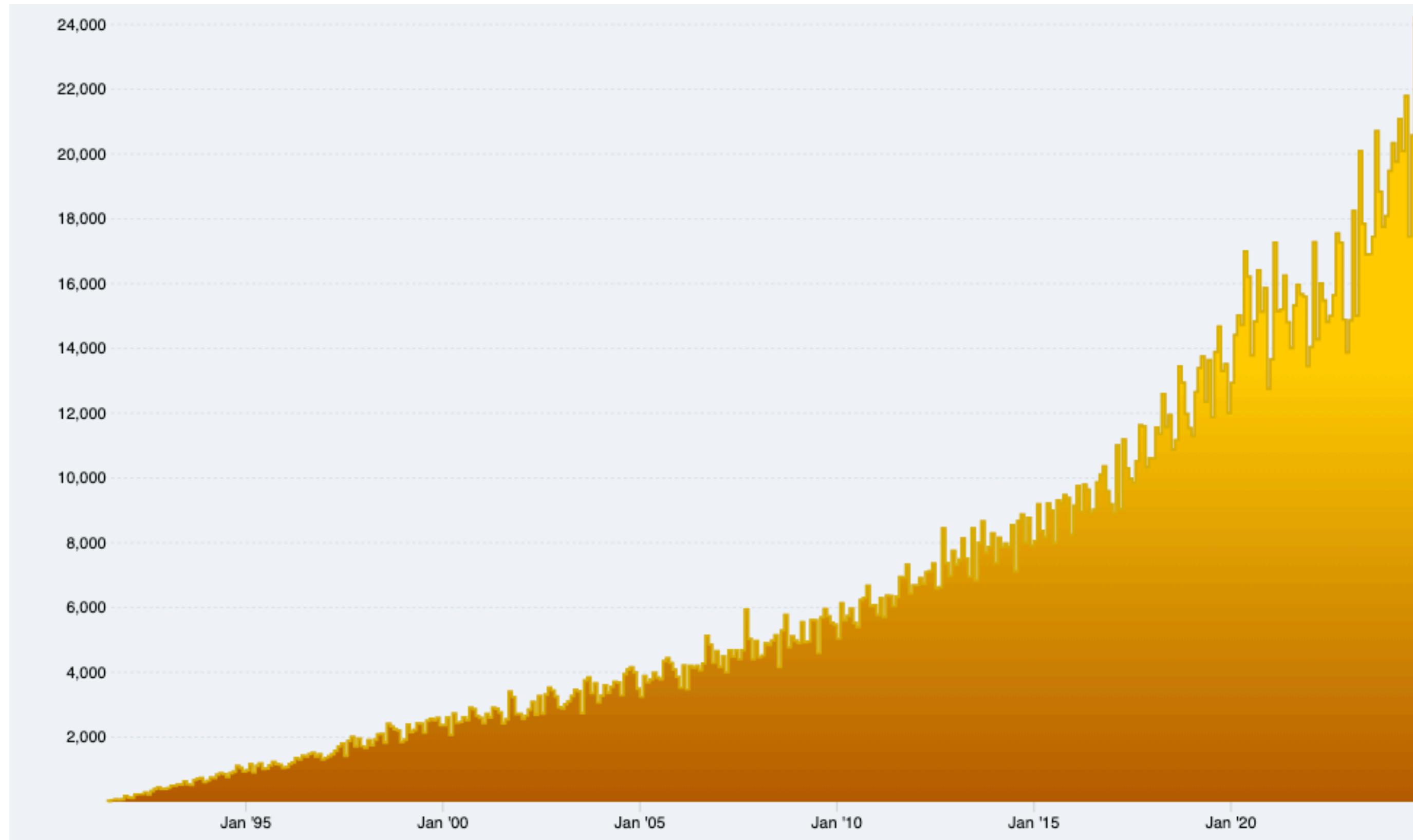
How?



Real-World Impacts of Retrieval-Augmented LMs

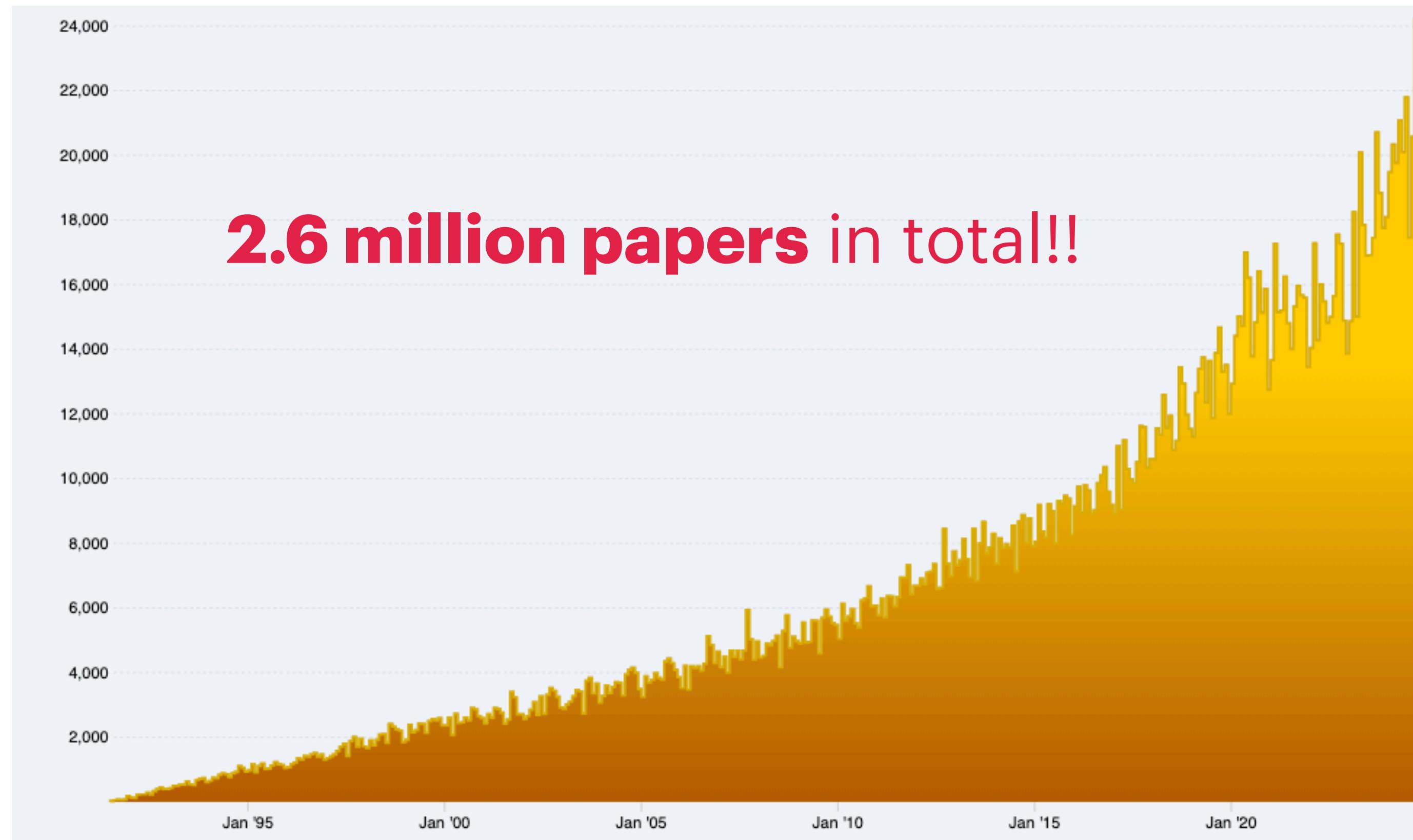
What?

Rapidly Growing Scientific Literature



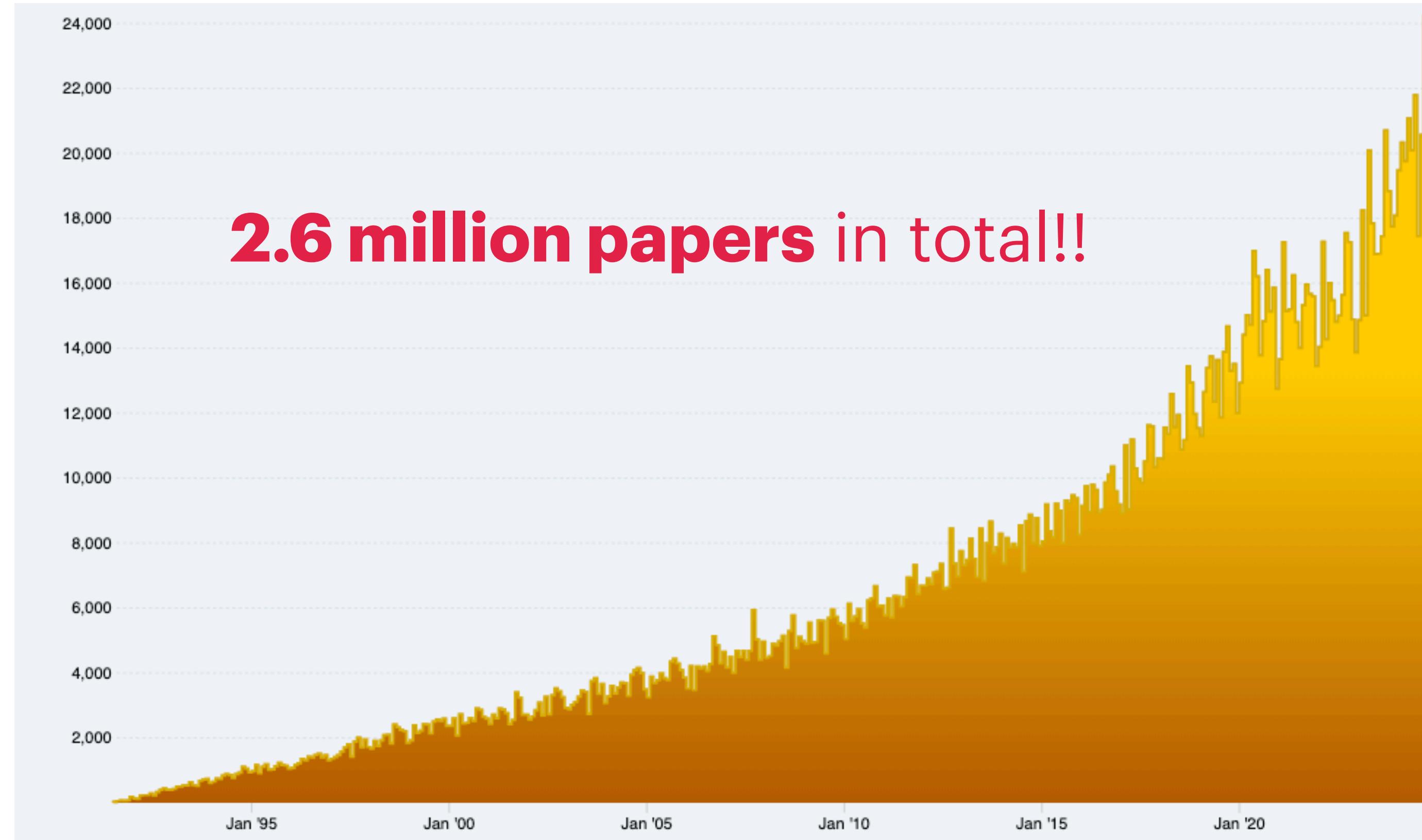
https://arxiv.org/stats/monthly_submissions (January 29, 2025)

Rapidly Growing Scientific Literature



https://arxiv.org/stats/monthly_submissions (January 29, 2025)

Rapidly Growing Scientific Literature



https://arxiv.org/stats/monthly_submissions (January 29, 2025)



Are there any studies showing if LMs can help scientists to synthesize scientific literature? Can you suggest some papers?



Yes, there are recent studies exploring how large language models (LLMs) can assist scientists in synthesizing scientific literature.

1. ChatGPT for Scientific Writing: Evaluating its Utility for Synthesizing and Summarizing Research (2023)
2. Automating Systematic Reviews with Pre-trained Language Models
3. Can Large Language Models Enhance Scientific Reasoning?

Are there any studies showing if LMs can help scientists to synthesize scientific literature? Can you suggest some papers?



Yes, there are recent studies exploring how large language models (LLMs) can assist scientists in synthesizing scientific literature.

1. ChatGPT for Scientific Writing: Evaluating its Utility for Synthesizing and Summarizing Research (2023)
2. Automating Systematic Reviews with Pre-trained Language Models
3. Can Large Language Models Enhance Scientific Reasoning?

78-92% of the model-generated citations don't exist

Are there any studies showing if LMs can help scientists to synthesize scientific literature? Can you suggest some papers?



Yes, there are recent studies exploring how large language models (LLMs) can assist scientists in synthesizing scientific literature.

1. ChatGPT for Scientific Writing: Evaluating its Utility for Synthesizing and Summarizing Research (2023)
2. Automating Systematic Reviews with Pre-trained Language Models
3. Can Large Language Models Enhance Scientific Reasoning?



Yes, there are recent studies exploring how large language models (LLMs) can assist scientists in synthesizing scientific literature.

1. ChatGPT for Scientific Writing: Evaluating its Utility for Synthesizing and Summarizing Research (2023)
2. Automating Systematic Reviews with Pre-trained Language Models
3. Can Large Language Models Enhance Scientific Reasoning?



Yes, there are recent studies exploring how large language models (LLMs) can assist scientists in synthesizing scientific literature.

1. ChatGPT for Scientific Writing: Evaluating its Utility for Synthesizing and Summarizing Research (2023)
2. Automating Systematic Reviews with Pre-trained Language Models
3. Can Large Language Models Enhance Scientific Reasoning?

Has anyone tried GPT4o in this area?





Yes, there are recent studies exploring how large language models (LLMs) can assist scientists in synthesizing scientific literature.

1. ChatGPT for Scientific Writing: Evaluating its Utility for Synthesizing and Summarizing Research (2023)
2. Automating Systematic Reviews with Pre-trained Language Models
3. Can Large Language Models Enhance Scientific Reasoning?



Has anyone tried GPT4o in this area?



Based on general knowledge, there is no widely recognized AI model or term specifically called "GPT4o." If "GPT4o" refers to a recent variant of GPT-4 or a new model released in 2024, I cannot comment.



Can language models synthesize scientific literature?

In a joint project between [Semantic Scholar](#) and the [University of Washington](#), we train and release a fully open, retrieval-augmented language model that can synthesize 8M+ open access research papers to answer scientific questions.

- Download the [full collection](#)--including model weights, training data and retrieval index.
- To learn more about the project, check out our [paper](#).

Type a question...



[Find papers on a topic](#)

[Learn about a concept](#)

[Summarize a paper](#)

[Study an algorithm](#)

[Check for prior work](#)

<https://openscholar.allen.ai/>

Asai et al. 2024. Under Review for Nature. OpenScholar: Synthesizing Scientific Literature with Retrieval-augmented LMs.



Can language models synthesize scientific literature?

In a joint project between [Semantic Scholar](#) and the [University of Washington](#), we train and release a fully open, retrieval-augmented language model that can synthesize 8M+ open access research papers to answer scientific questions.

- Download the [full collection](#)--including model weights, training data and retrieval index.
- To learn more about the project, check out our [paper](#).

Type a question...



[Find papers on a topic](#)

[Learn about a concept](#)

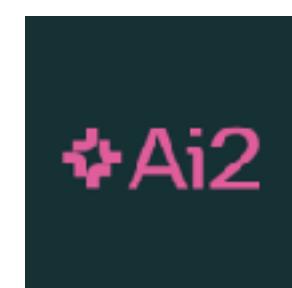
[Summarize a paper](#)

[Study an algorithm](#)

[Check for prior work](#)

<https://openscholar.allen.ai/>

Asai et al. 2024. Under Review for Nature. OpenScholar: Synthesizing Scientific Literature with Retrieval-augmented LMs.



Are there any studies showing if LMs can help scientists to synthesize scientific literature? Could you suggest recent papers?

Generating the initial draft
started 4s ago / estimated: 1 minute



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

<https://openscholar.allen.ai/>

Asai et al. 2024. Under Review for Nature. OpenScholar: Synthesizing Scientific Literature with Retrieval-augmented LMs.



Are there any studies showing if LMs can help scientists to synthesize scientific literature? Could you suggest recent papers?

Generating the initial draft
started 4s ago / estimated: 1 minute



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

<https://openscholar.allen.ai/>

Asai et al. 2024. Under Review for Nature. OpenScholar: Synthesizing Scientific Literature with Retrieval-augmented LMs.

OpenScholar for Scientific Literature Synthesis

OpenScholar for Scientific Literature Synthesis

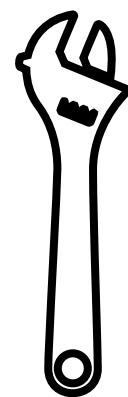


Evaluation: How can we rigorously **evaluate** expert-domain systems?

OpenScholar for Scientific Literature Synthesis



Evaluation: How can we rigorously **evaluate** expert-domain systems?

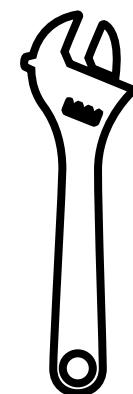


Modeling: How can we **enhance model** in resource-scarce settings?

OpenScholar for Scientific Literature Synthesis



Evaluation: How can we rigorously **evaluate** expert-domain systems?

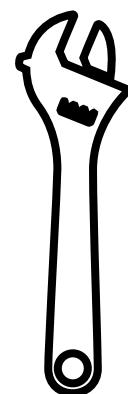


Modeling: How can we **enhance model** in resource-scarce settings?

OpenScholar for Scientific Literature Synthesis



Evaluation: How can we rigorously **evaluate** expert-domain systems?



Modeling: How can we **enhance model** in resource-scarce settings?

- Use off-the-shelf, private LMs & APIs for all components (Lu et al., 2024; Skarlinski et al., 2024)
- Rely on artificial evaluation (e.g., multiple-choice QA) to evaluate systems (Skarlinski et al., 2024)

ScholarQABench: A Comprehensive Evaluation

ScholarQABench: A Comprehensive Evaluation

SciFact

PubMedQA

QASA

SQA - CS

SQA - Bio

SQA - Neuro

SQA - Multi

ScholarQABench: A Comprehensive Evaluation

Understanding single paper

SciFact

PubMedQA

QASA

SQA - CS

SQA - Bio

SQA - Neuro

SQA - Multi

ScholarQABench: A Comprehensive Evaluation

Understanding single paper

SciFact

PubMedQA

QASA

Synthesizing multiple papers (ScholarQA; SQA)

SQA - CS

SQA - Bio

SQA - Neuro

SQA - Multi

ScholarQABench: A Comprehensive Evaluation

Understanding single paper				Synthesizing multiple papers (ScholarQA; SQA)			
Adapted or created	SciFact	PubMedQA	QASA	SQA - CS	SQA - Bio	SQA - Neuro	SQA - Multi
Wadden et al. 2020	Jin et al. 2019	Lee et al. 2023					

ScholarQABench: A Comprehensive Evaluation

Understanding single paper

SciFact

PubMedQA

QASA

Synthesizing multiple papers (ScholarQA; SQA)

SQA - CS

SQA - Bio

SQA - Neuro

SQA - Multi

**Adapted
or created**

Wadden
et al. 2020

Jin
et al. 2019

Lee
et al. 2023



Newly Created 2.2k questions + 220 answers

ScholarQABench: A Comprehensive Evaluation

Understanding single paper

SciFact

PubMedQA

QASA

Synthesizing multiple papers (ScholarQA; SQA)

SQA - CS

SQA - Bio

SQA - Neuro

SQA - Multi

**Adapted
or created**

Wadden
et al. 2020

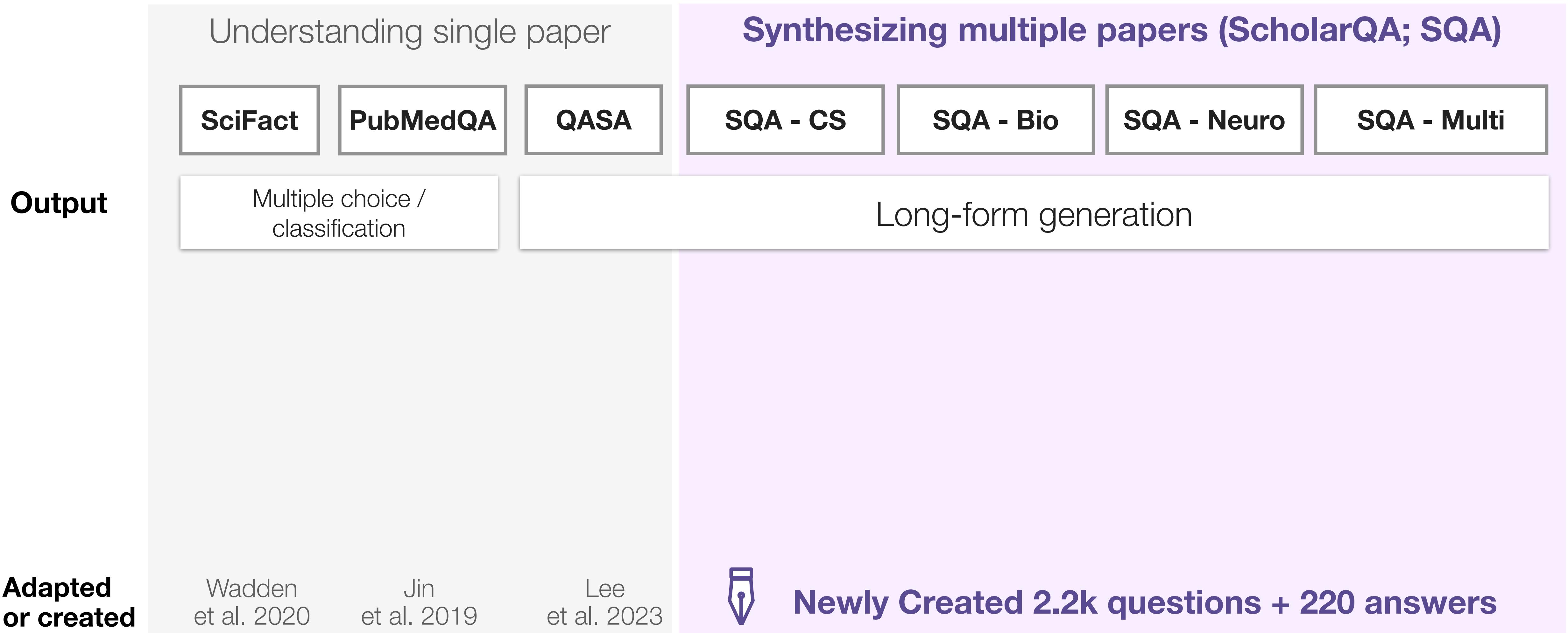
Jin
et al. 2019

Lee
et al. 2023



Newly Created 2.2k questions + 220 answers

ScholarQABench: A Comprehensive Evaluation



ScholarQABench: A Comprehensive Evaluation

	Understanding single paper			Synthesizing multiple papers (ScholarQA; SQA)				
	SciFact	PubMedQA	QASA	SQA - CS	SQA - Bio	SQA - Neuro	SQA - Multi	
Output	Multiple choice / classification				Long-form generation			
Domain	Biomedicine (Biomed)			Computer Science (CS)	Biomed	Neuroscience	CS	
Adapted or created	Wadden et al. 2020	Jin et al. 2019	Lee et al. 2023		Newly Created 2.2k questions + 220 answers			

ScholarQABench: A Comprehensive Evaluation

	Understanding single paper			Synthesizing multiple papers (ScholarQA; SQA)				
	SciFact	PubMedQA	QASA	SQA - CS	SQA - Bio	SQA - Neuro	SQA - Multi	
Output	Multiple choice / classification				Long-form generation			
Domain	Biomedicine (Biomed)			Computer Science (CS)	Biomed	Neuroscience	CS	
Adapted or created	Wadden et al. 2020	Jin et al. 2019	Lee et al. 2023		Newly Created 2.2k questions + 220 answers			

ScholarQABench: ScholarQA-Multi

Physics

Input: What are ways to cool the center-of-mass (CoM) motion of levitated nanoparticles?

Long-form answer:

Currently, the most commonly used cooling method is feedback cooling [1]. By measuring the real-time position ... An alternative method involves levitating the nanosphere in an optical cavity and cooling via coherent scattering [2]. The trapping laser is red-detuned from ...

Citations:

[1] Millikelvin cooling of an optically trapped microsphere in vacuum

[2] Cavity Cooling of a Levitated Nanosphere by Coherent Scattering

ScholarQABench: ScholarQA-Multi

Physics

Input: What are ways to cool the center-of-mass (CoM) motion of levitated nanoparticles?

Long-form answer:

Currently, the most commonly used cooling method is feedback cooling [1]. By measuring the real-time position ... An alternative method involves levitating the nanosphere in an optical cavity and cooling via coherent scattering [2]. The trapping laser is red-detuned from ...

Citations:

- [1] Millikelvin cooling of an optically trapped microsphere in vacuum
- [2] Cavity Cooling of a Levitated Nanosphere by Coherent Scattering

Annotators with Ph.D. spent 58 mins per query on average

ScholarQABench: ScholarQA-CS

CS

Input: What are the best practices to protect a software against vulnerabilities from third party libraries?

Answer rubrics:

ScholarQABench: ScholarQA-CS

CS

Input: What are the best practices to protect a software against vulnerabilities from third party libraries?

Answer rubrics:

Must Have Item-1: The answer should discuss best practices that can be used to prevent these implications, such as reliable source, update monitoring, code.

ScholarQABench: ScholarQA-CS

CS

Input: What are the best practices to protect a software against vulnerabilities from third party libraries?

Answer rubrics:

Must Have Item-1: The answer should discuss best practices that can be used to prevent these implications, such as reliable source, update monitoring, code.

Nice to Have Item-1: The answer could provide some examples of famous third party libraries that can be used in different programming languages.

ScholarQABench: Rubric Accuracy

Input: What are the best practices to protect a software against vulnerabilities from third party libraries?

Answer rubrics:

Must Have Item-1: The answer should discuss best practices that can be used to prevent these implications, such as reliable source, update monitoring, code.

Nice to Have Item-1: The answer could provide some examples of famous third party libraries that can be used in different programming languages.

ScholarQABench: Rubric Accuracy

Input: What are the best practices to protect a software against vulnerabilities from third party libraries?

Must Have Item-1: The answer should discuss best practices that can be used to prevent these implications, such as reliable source, update monitoring, code.

ScholarQABench: Rubric Accuracy

Input: What are the best practices to protect a software against vulnerabilities from third party libraries?

Must Have Item-1: The answer should discuss best practices that can be used to prevent these implications, such as reliable source, update monitoring, code.

ScholarQABench: Rubric Accuracy

Input: What are the best practices to protect a software against vulnerabilities from third party libraries?

Must Have Item-1: The answer should discuss best practices that can be used to prevent these implications, such as reliable source, update monitoring, code.

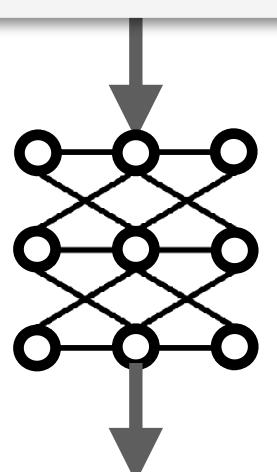
Output: Protecting software against vulnerabilities stemming from third-party libraries is a crucial aspect of software security. Below are some of the best practices based on the existing literature: (1) Developing intelligent tools to automatically and repair vulnerabilities ...

ScholarQABench: Rubric Accuracy

Input: What are the best practices to protect a software against vulnerabilities from third party libraries?

Must Have Item-1: The answer should discuss best practices that can be used to prevent these implications, such as reliable source, update monitoring, code.

Output: Protecting software against vulnerabilities stemming from third-party libraries is a crucial aspect of software security. Below are some of the best practices based on the existing literature: (1) Developing intelligent tools to automatically and repair vulnerabilities ...

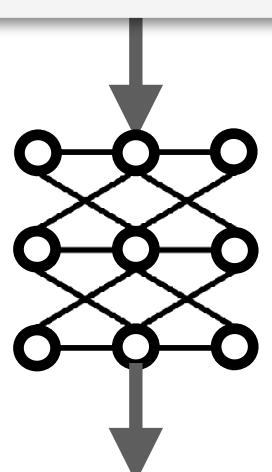


ScholarQABench: Rubric Accuracy

Input: What are the best practices to protect a software against vulnerabilities from third party libraries?

Must Have Item-1: The answer should discuss best practices that can be used to prevent these implications, such as reliable source, update monitoring, code.

Output: Protecting software against vulnerabilities stemming from third-party libraries is a crucial aspect of software security. Below are some of the best practices based on the existing literature: (1) Developing intelligent tools to automatically and repair vulnerabilities ...

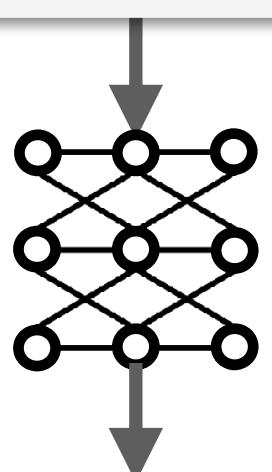


ScholarQABench: Rubric Accuracy

Input: What are the best practices to protect a software against vulnerabilities from third party libraries?

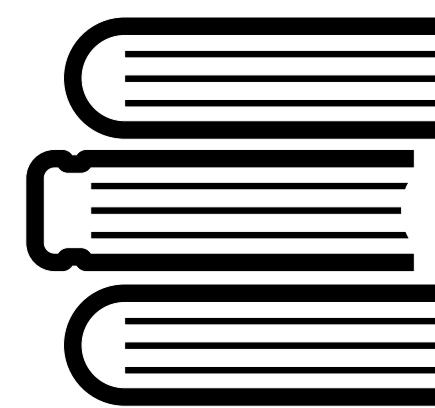
Must Have Item-1: The answer should discuss best practices that can be used to prevent these implications, such as reliable source, update monitoring, code.

Output: Protecting software against vulnerabilities stemming from third-party libraries is a crucial aspect of software security. Below are some of the best practices based on the existing literature: (1) Developing intelligent tools to automatically and repair vulnerabilities ...

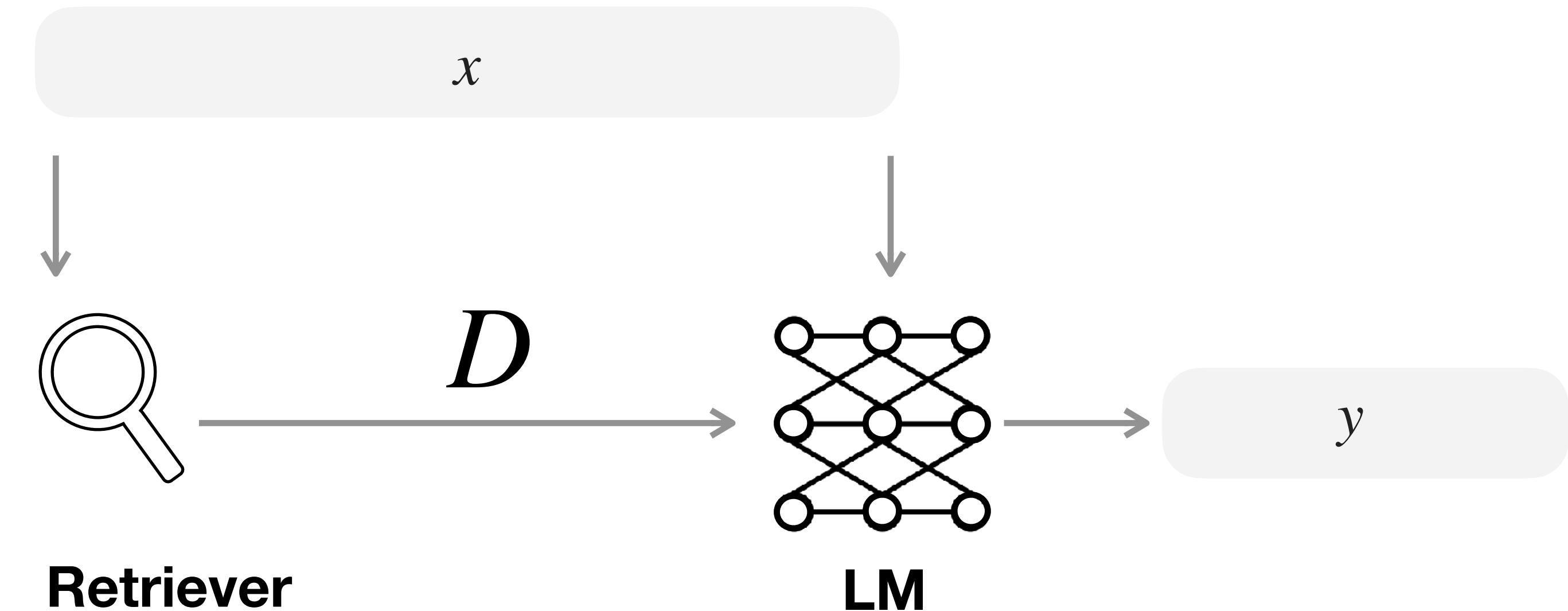


Must have is satisfied

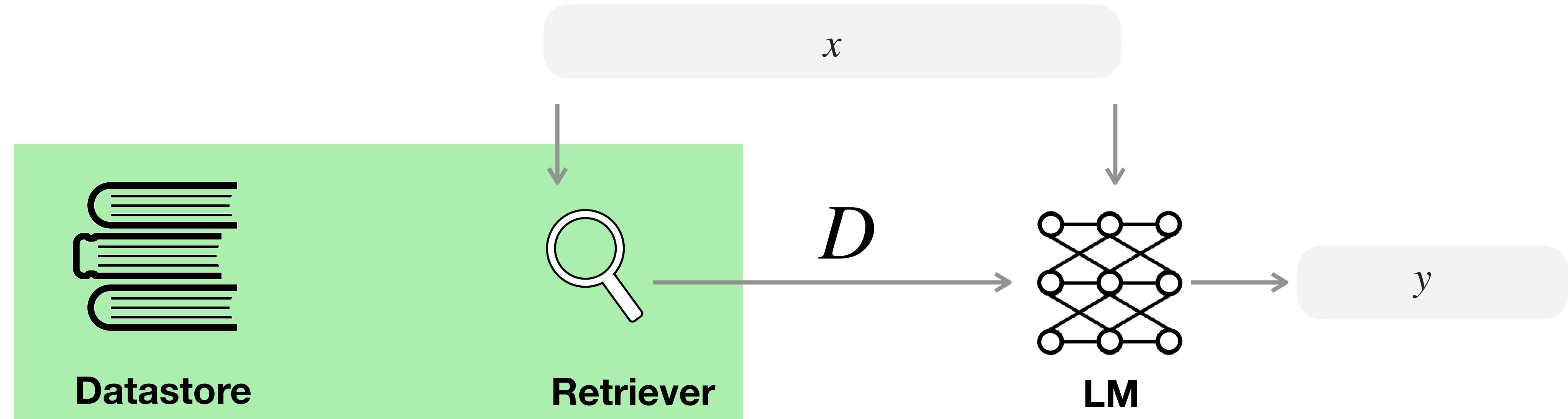
OpenScholar for Scientific Literature Synthesis



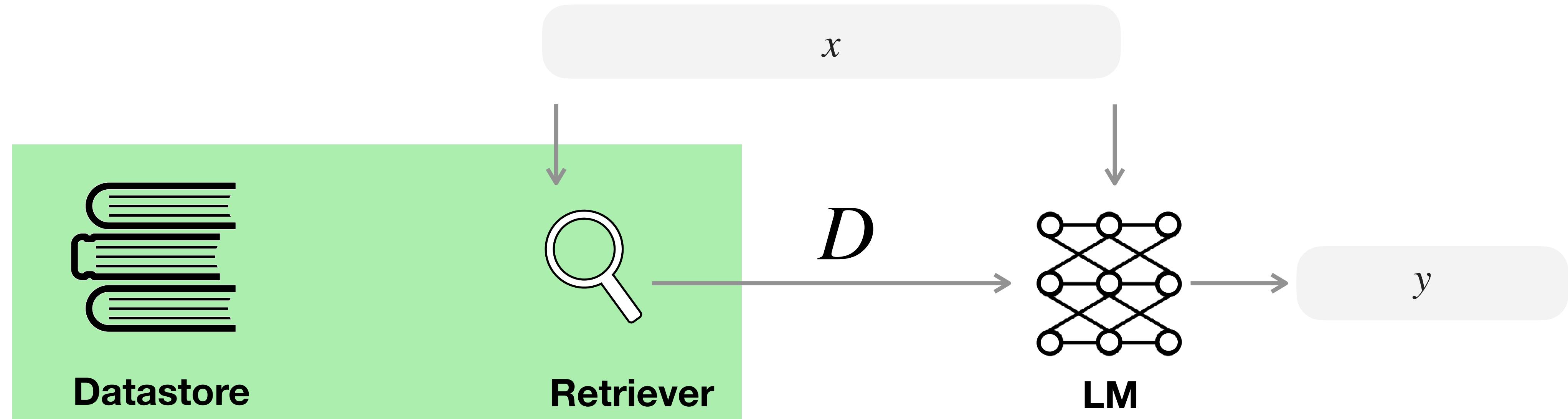
Datastore



OpenScholar: Retrieval

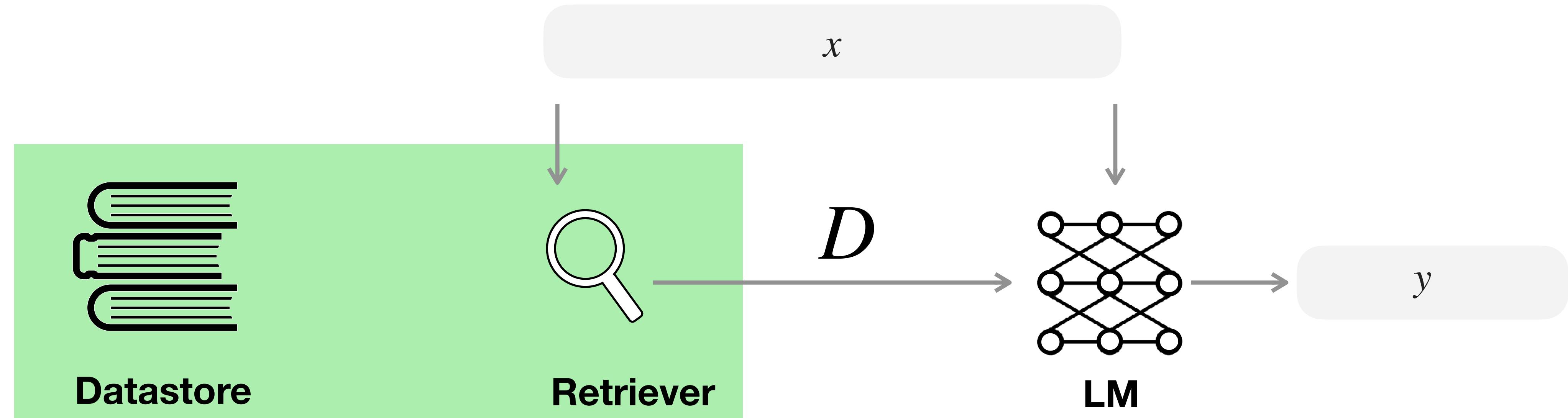


OpenScholar: Retrieval



45 million
open-access papers

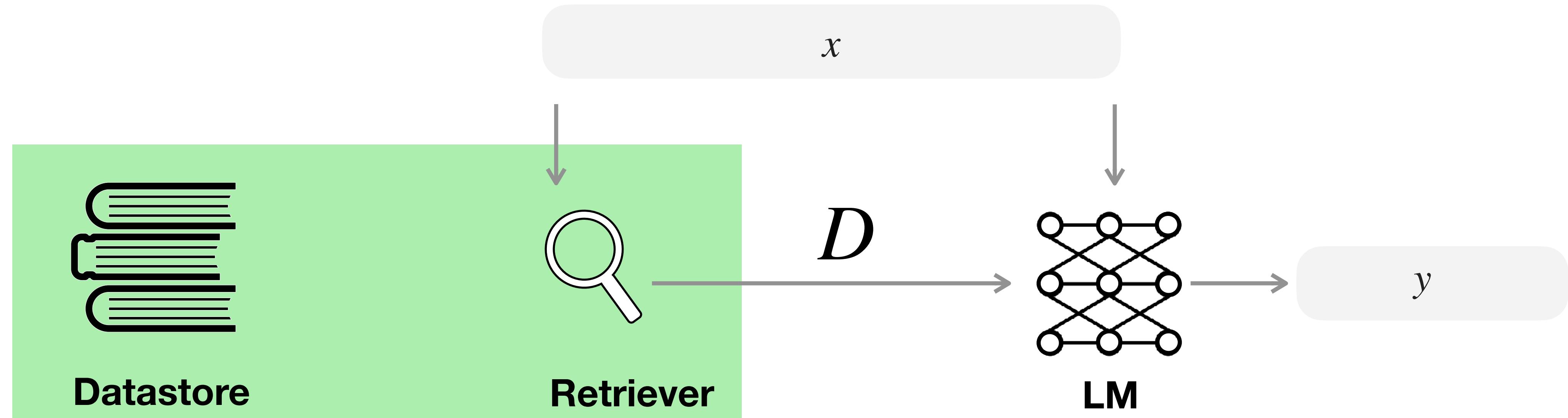
OpenScholar: Retrieval



45 million
open-access papers

Shao, He, [Asai](#) et al. NeurIPS 2024. Scaling Retrieval-Based Language Models with a Trillion-Token Datastore.

OpenScholar: Retrieval

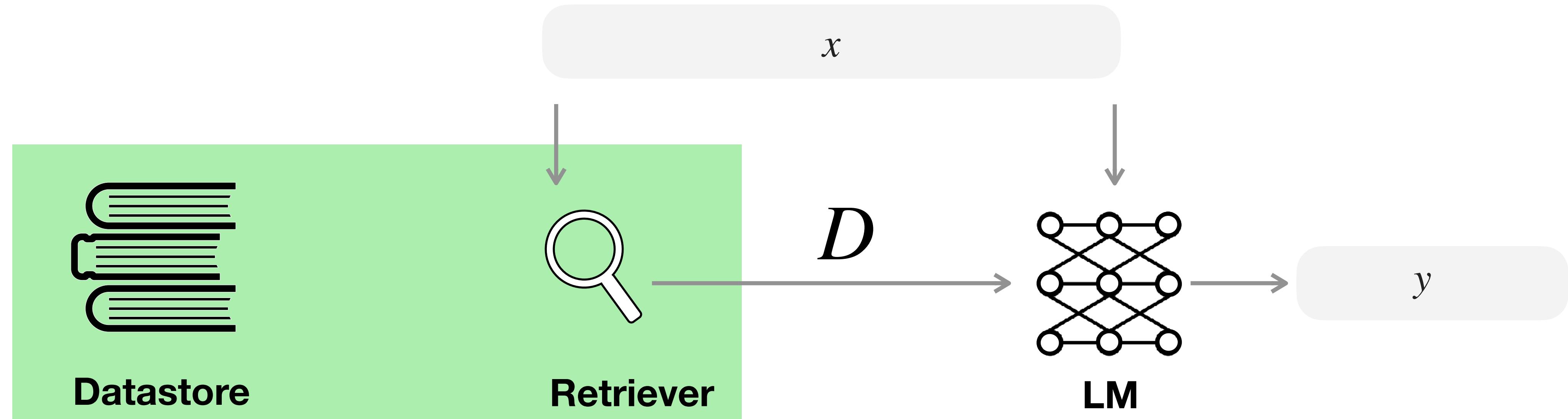


45 million
open-access papers

Updated weekly

Shao, He, [Asai](#) et al. NeurIPS 2024. Scaling Retrieval-Based Language Models with a Trillion-Token Datastore.

OpenScholar: Retrieval



45 million
open-access papers

Updated weekly

Text snippet search

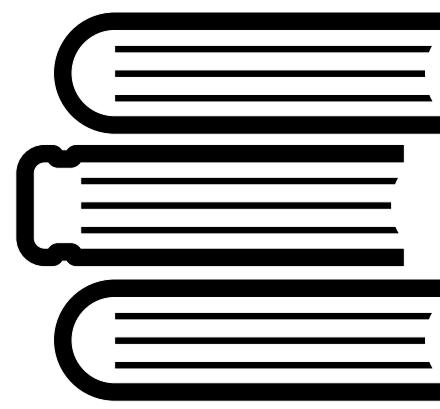
Return the text snippets that most closely match the query. Text snippets are excerpts of approximately 500 words, drawn from a paper's title, abstract, and body text, but excluding figure captions and the bibliography. It will return the highest ranked snippet first, as well as some basic data about the paper it was found in. Examples:

- [https://api.semanticscholar.org/graph/v1/snippet/search?
query=The literature graph is a property graph with directed
edges&limit=1](https://api.semanticscholar.org/graph/v1/snippet/search?query=The%20literature%20graph%20is%20a%20property%20graph%20with%20directed%20edges&limit=1)

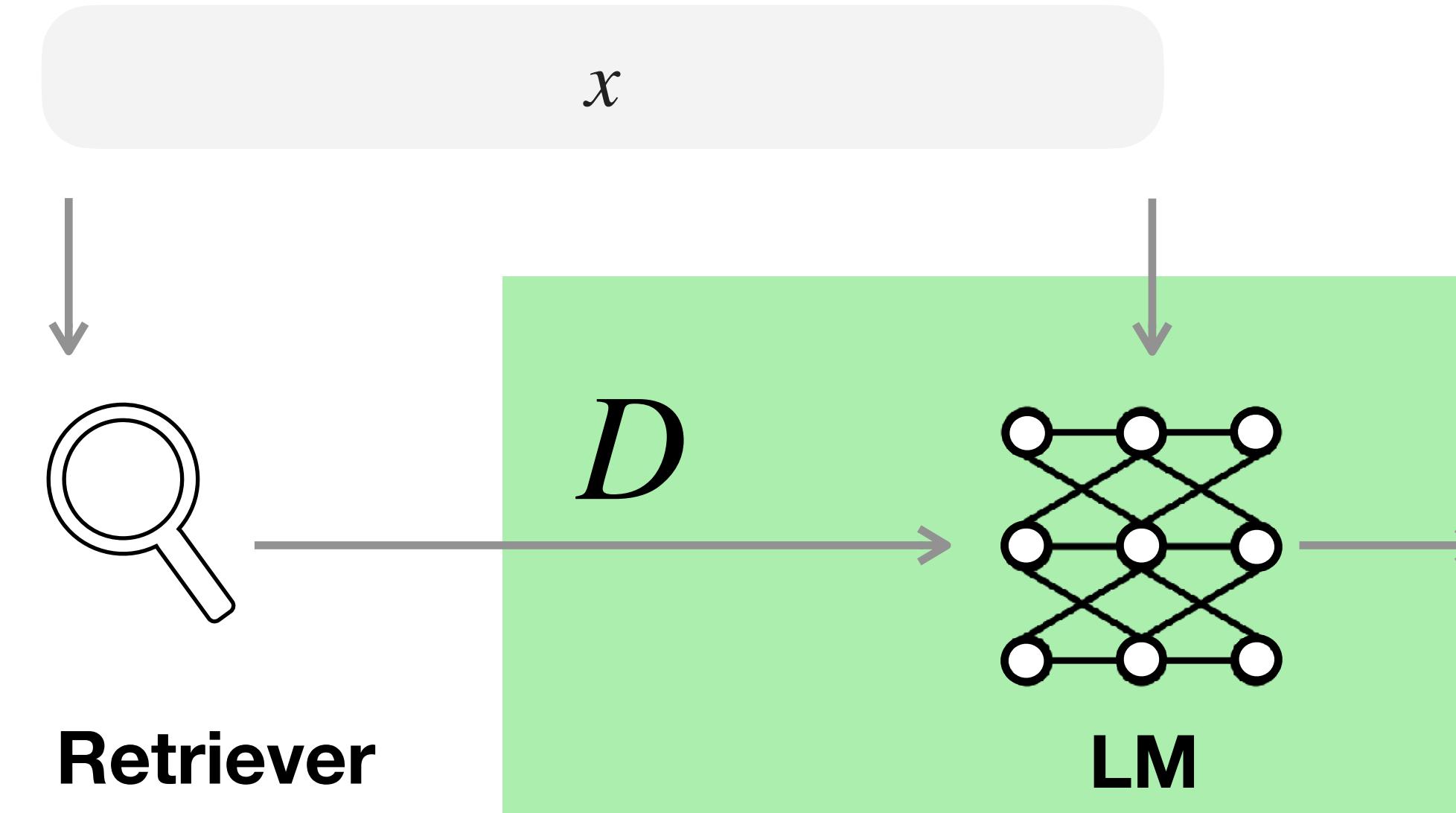
<https://api.semanticscholar.org/api-docs/>

Shao, He, **Asai** et al. NeurIPS 2024. Scaling Retrieval-Based Language Models with a Trillion-Token Datastore.

OpenScholar: Self-Feedback guided Generation

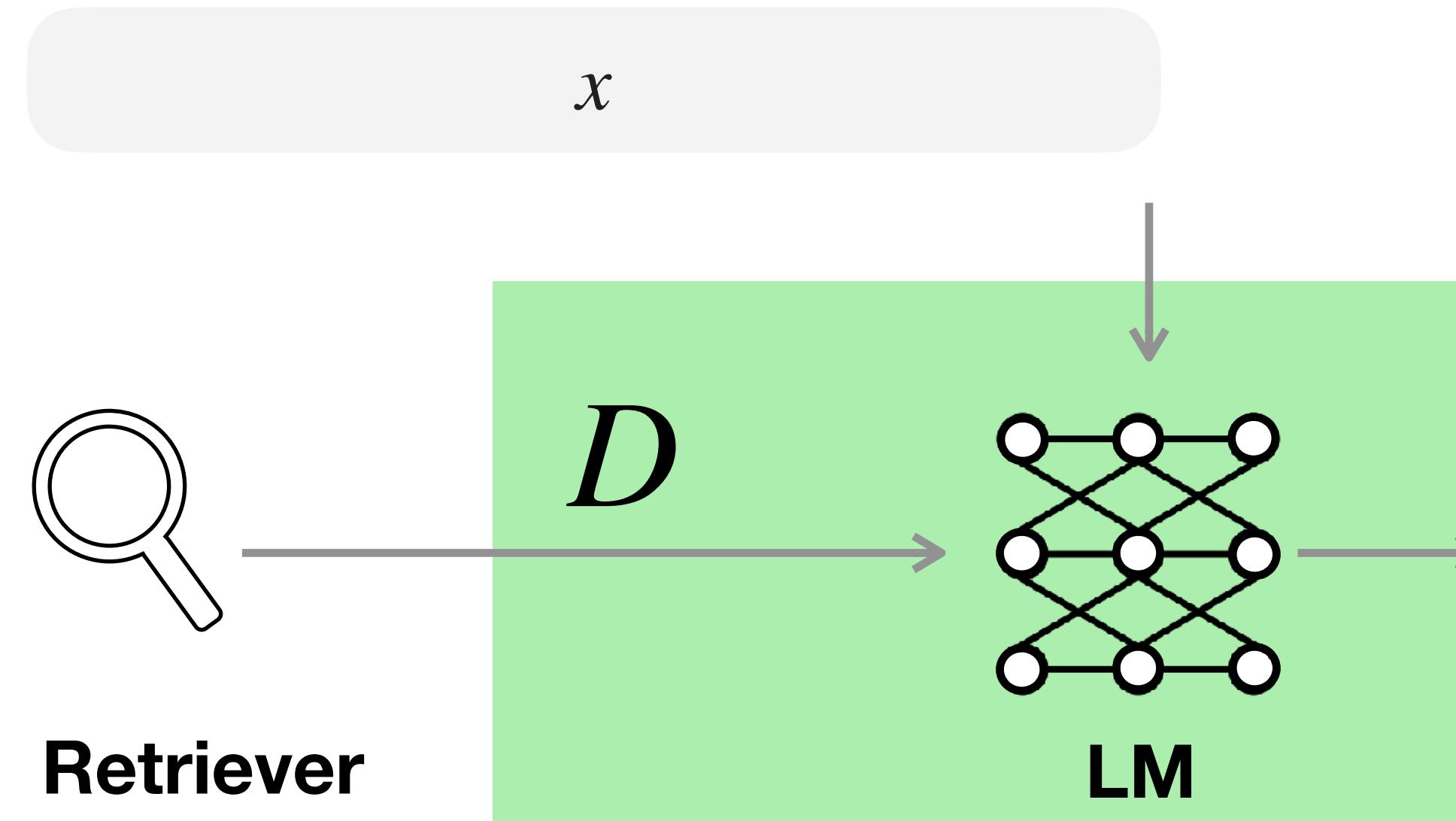


Datastore



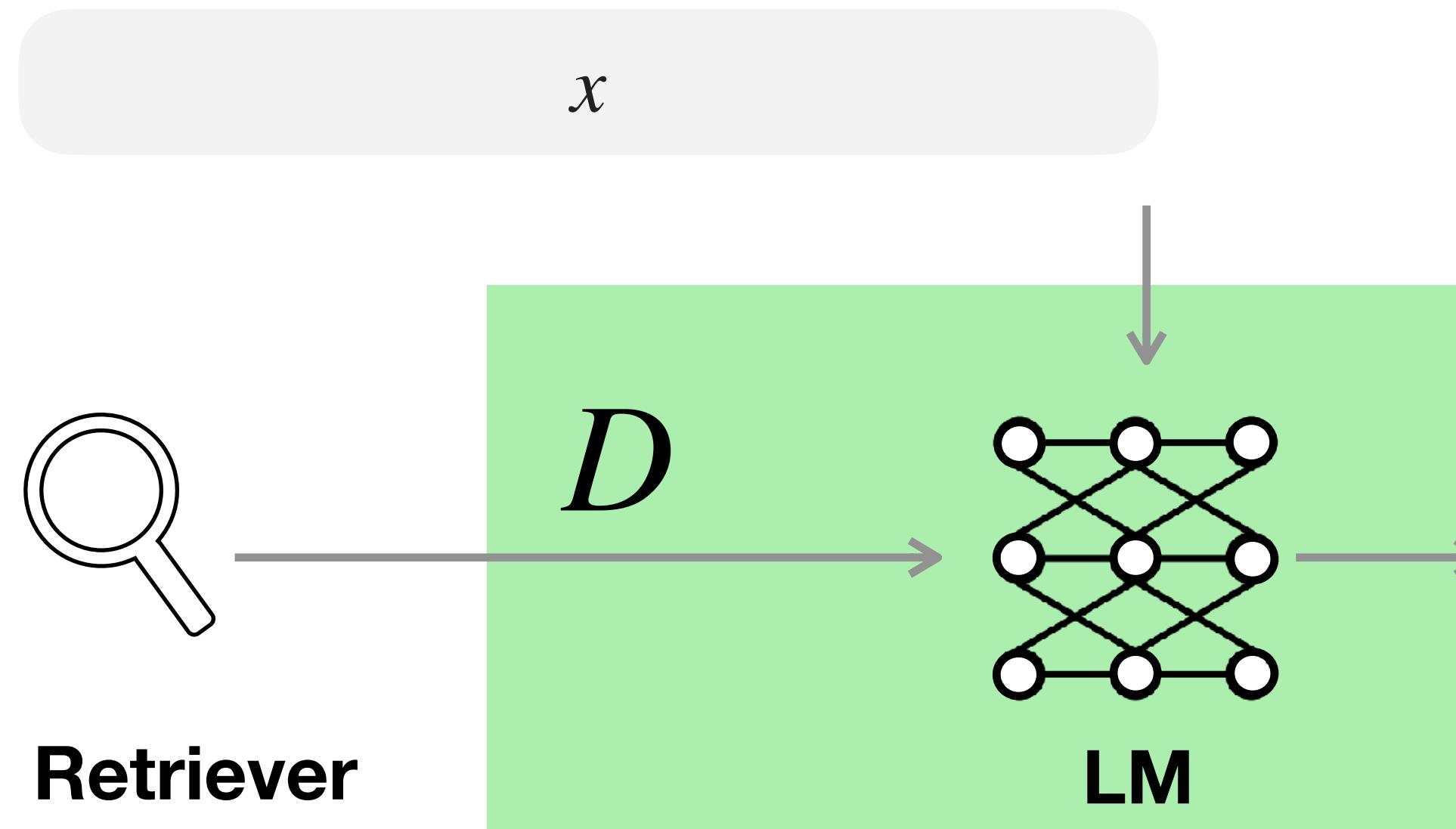
Asai et al. ICLR 2024 (Oral - Top 1%). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

OpenScholar: Self-Feedback guided Generation



Asai et al. ICLR 2024 (Oral - Top 1%). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

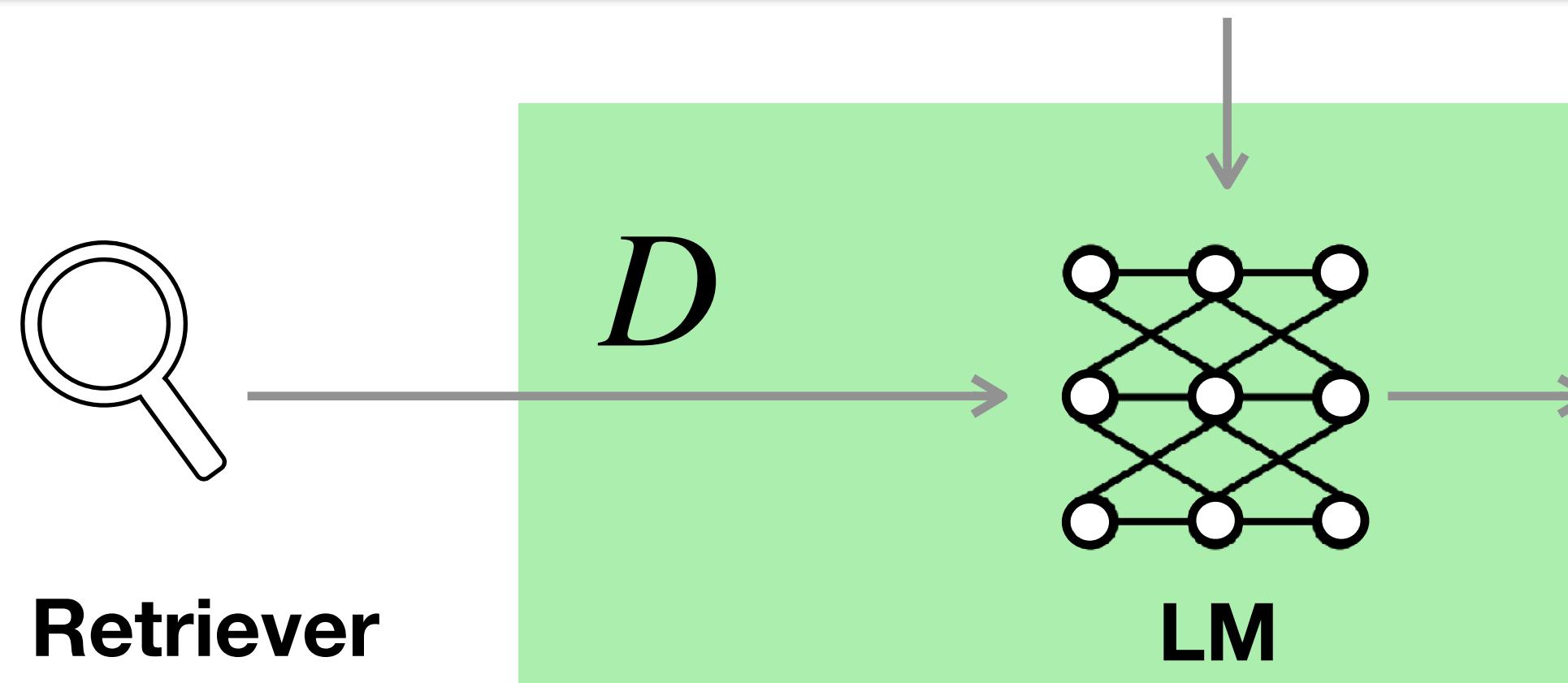
OpenScholar: Self-Feedback guided Generation



Asai et al. ICLR 2024 (Oral - Top 1%). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

OpenScholar: Self-Feedback guided Generation

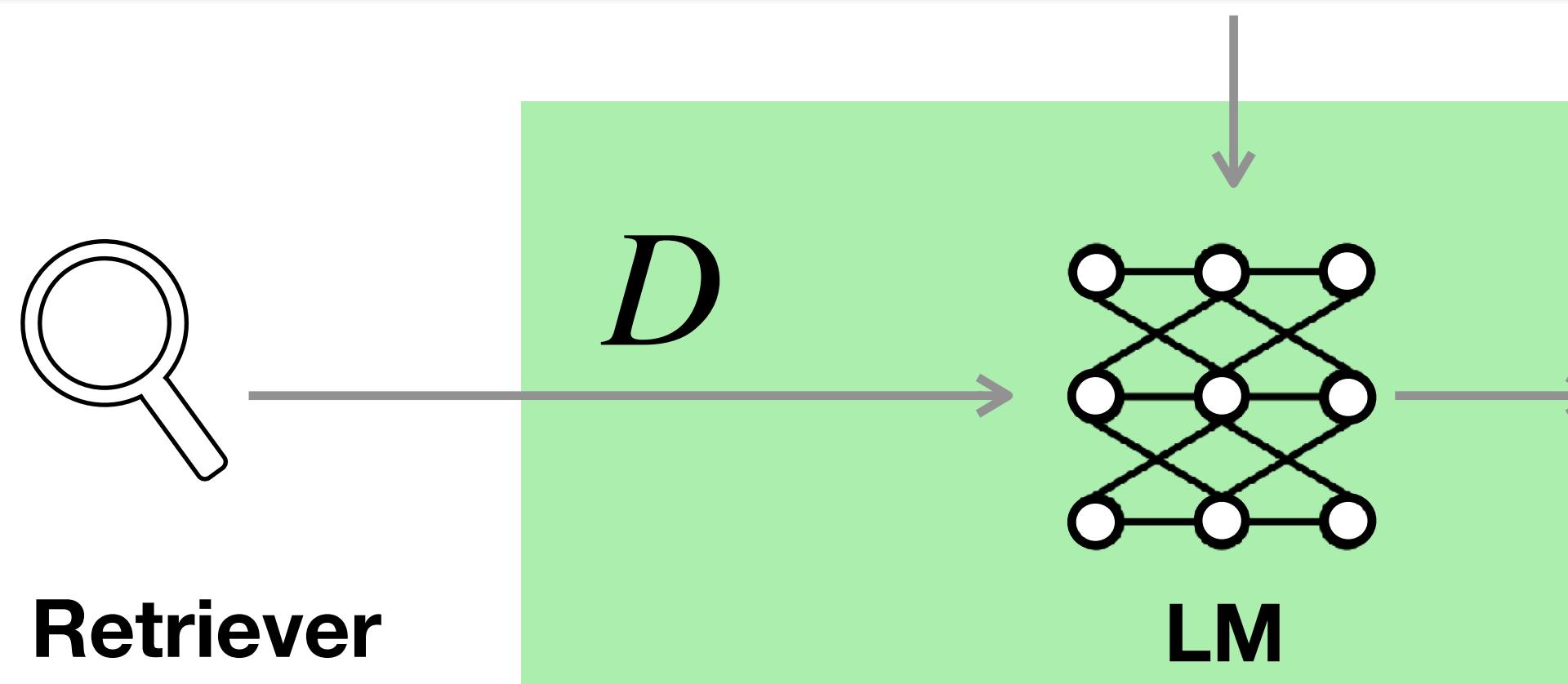
Has anyone evaluated LMs' abilities of understanding scientific literature?



Asai et al. ICLR 2024 (Oral - Top 1%). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

OpenScholar: Self-Feedback guided Generation

Has anyone evaluated LMs' abilities of understanding scientific literature?



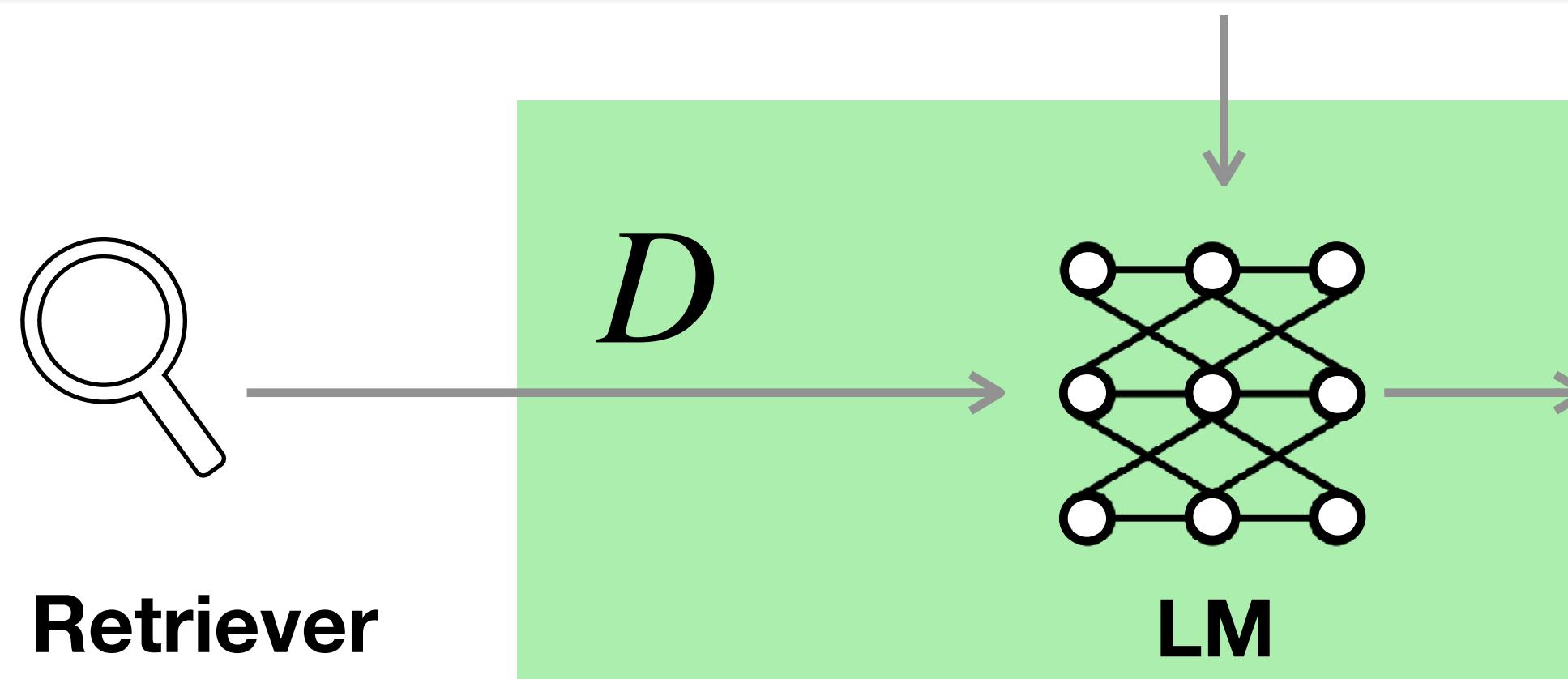
Yes, researchers shows that GPT4o can assist with idea generations [1]

Asai et al. ICLR 2024 (Oral - Top 1%). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

OpenScholar: Self-Feedback guided Generation

Has anyone evaluated LMs' abilities of understanding scientific literature?

Yes, researchers shows that GPT4o can assist with idea generations [1]

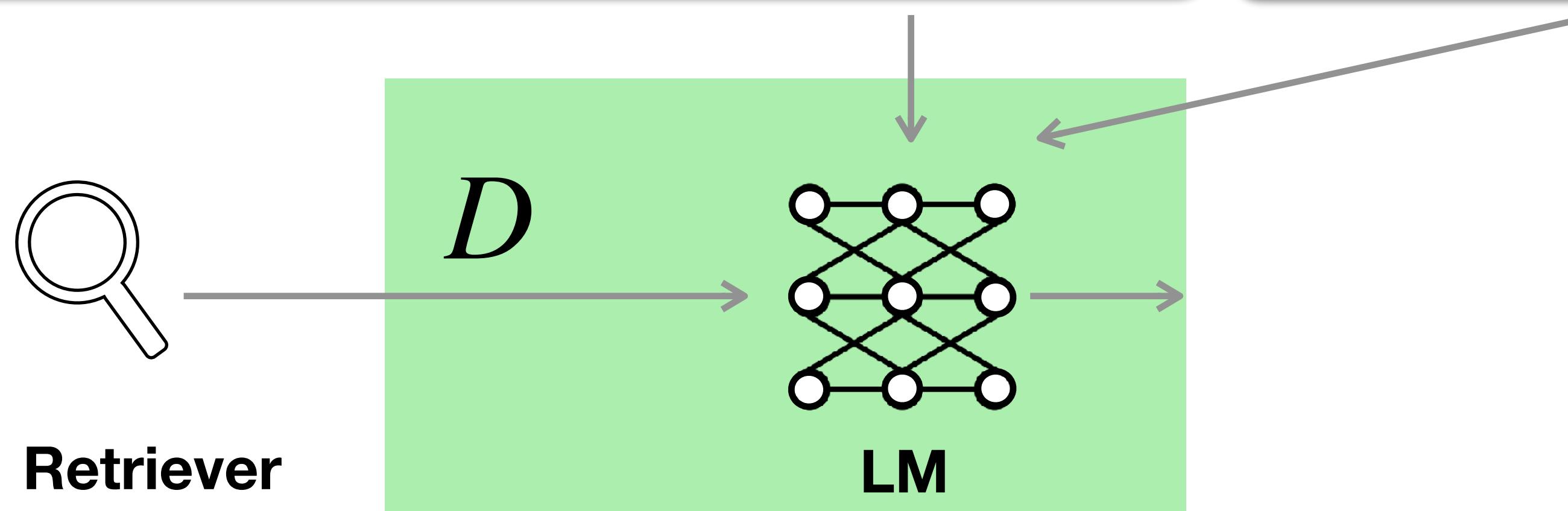


Asai et al. ICLR 2024 (Oral - Top 1%). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

OpenScholar: Self-Feedback guided Generation

Has anyone evaluated LMs' abilities of understanding scientific literature?

Yes, researchers shows that GPT4o can assist with idea generations [1]

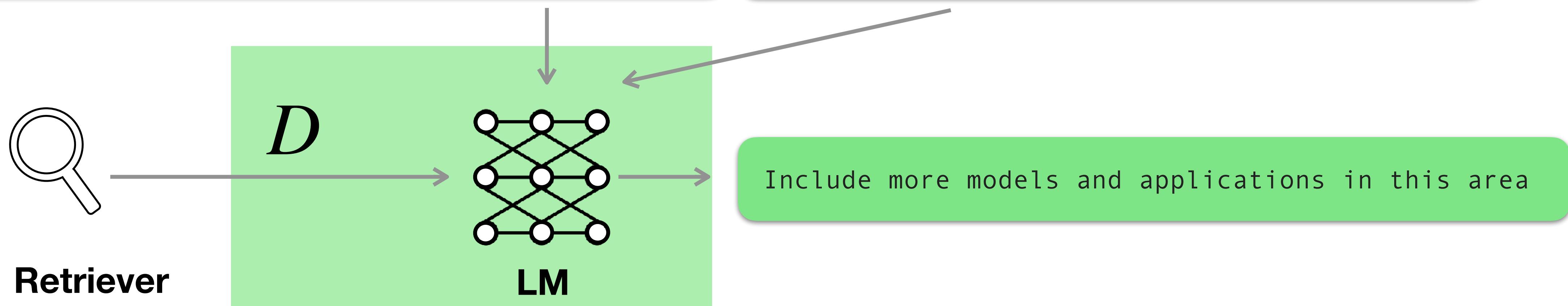


Asai et al. ICLR 2024 (Oral - Top 1%). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

OpenScholar: Self-Feedback guided Generation

Has anyone evaluated LMs' abilities of understanding scientific literature?

Yes, researchers shows that GPT4o can assist with idea generations [1]

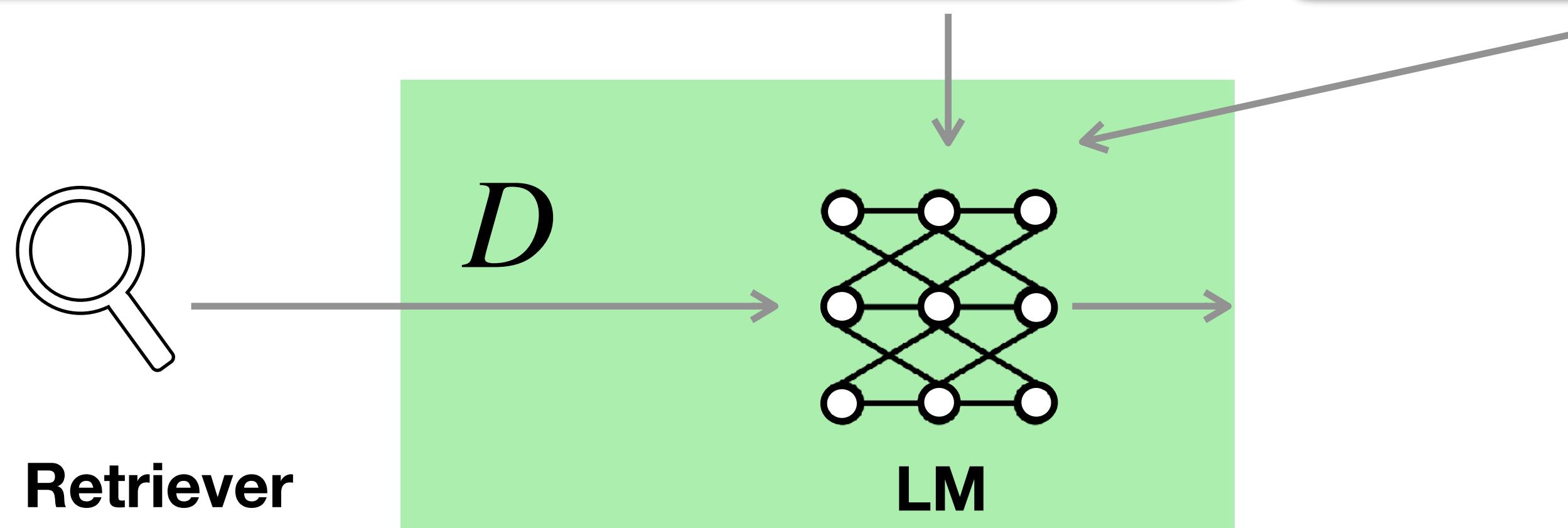


Asai et al. ICLR 2024 (Oral - Top 1%). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

OpenScholar: Self-Feedback guided Generation

Has anyone evaluated LMs' abilities of understanding scientific literature?

Yes, researchers shows that GPT4o can assist with idea generations [1]



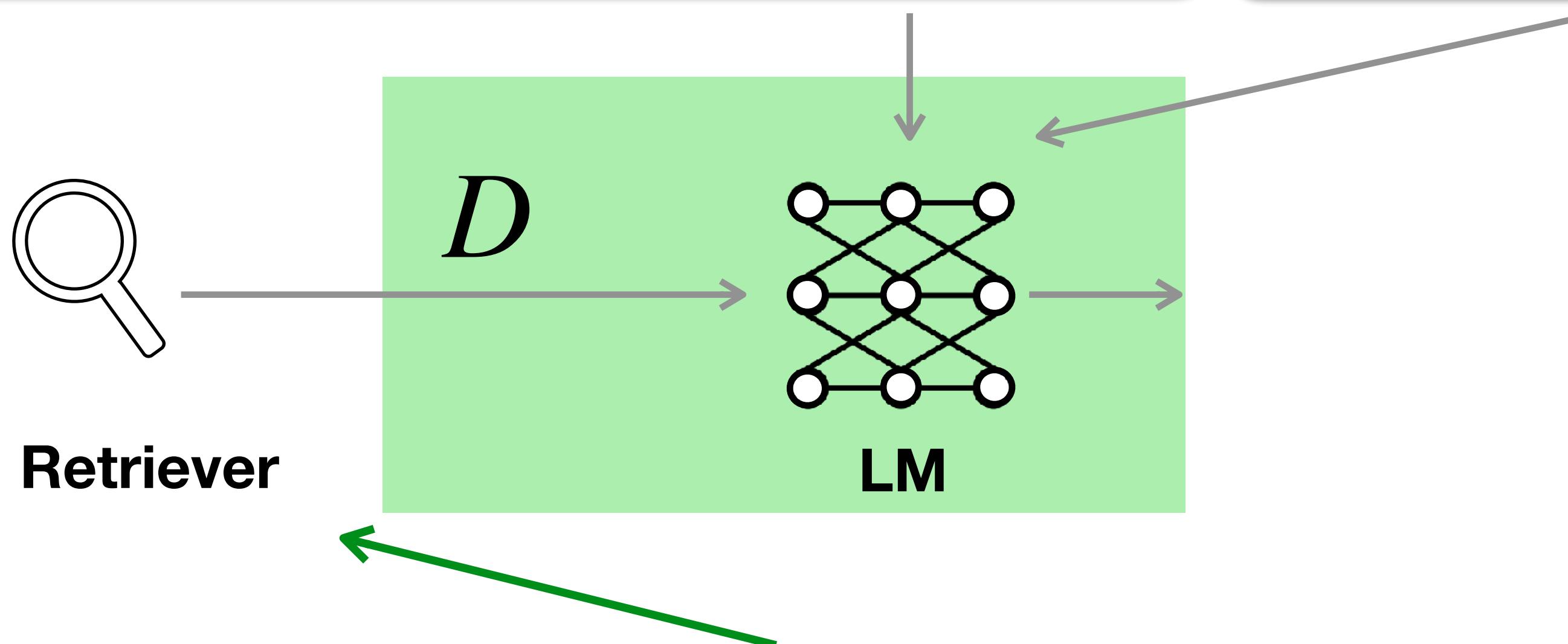
Include more models and applications in this area

Asai et al. ICLR 2024 (Oral - Top 1%). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

OpenScholar: Self-Feedback guided Generation

Has anyone evaluated LMs' abilities of understanding scientific literature?

Yes, researchers shows that GPT4o can assist with idea generations [1]



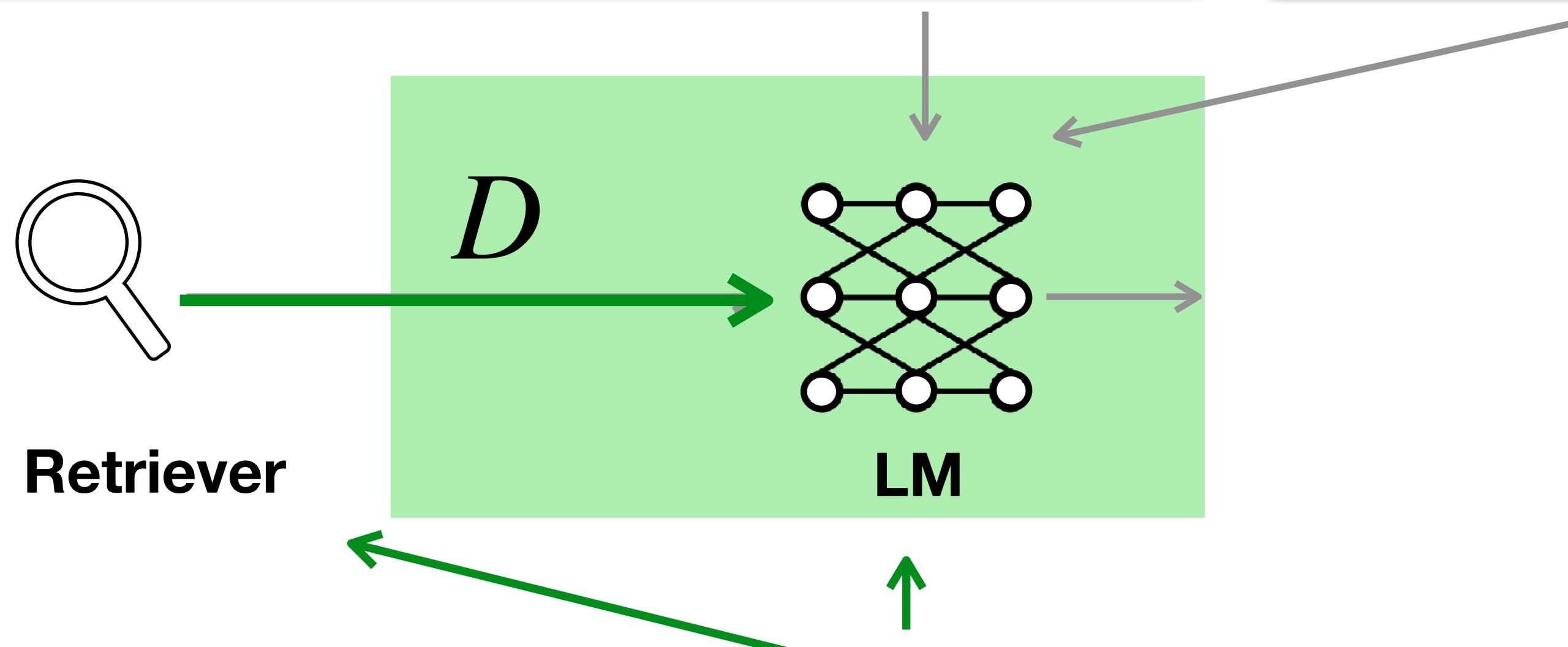
Include more models and applications in this area

Asai et al. ICLR 2024 (Oral - Top 1%). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

OpenScholar: Self-Feedback guided Generation

Has anyone evaluated LMs' abilities of understanding scientific literature?

Yes, researchers shows that GPT4o can assist with idea generations [1]

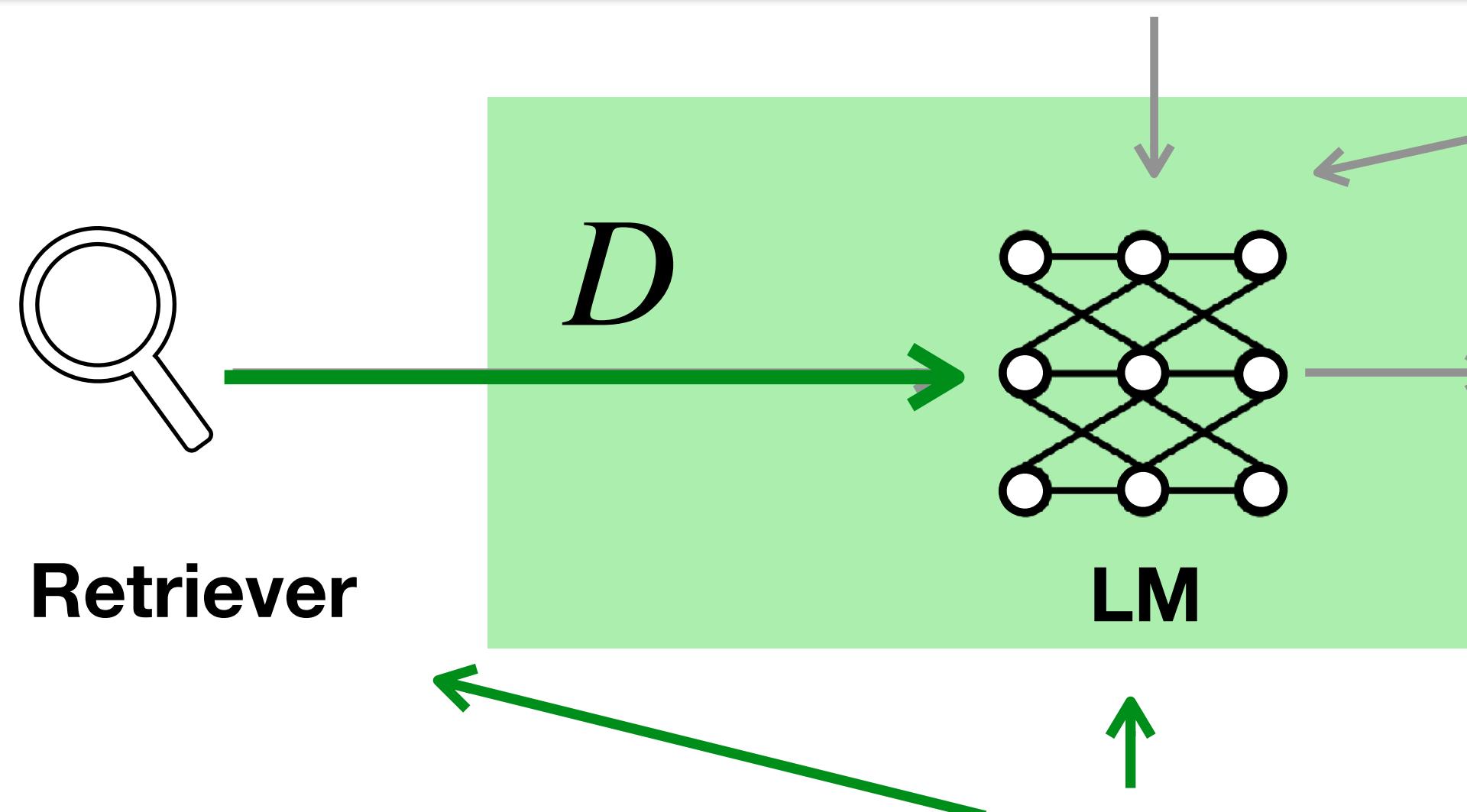


Asai et al. ICLR 2024 (Oral - Top 1%). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

OpenScholar: Self-Feedback guided Generation

Has anyone evaluated LMs' abilities of understanding scientific literature?

Yes, researchers shows that GPT4o can assist with idea generations [1]

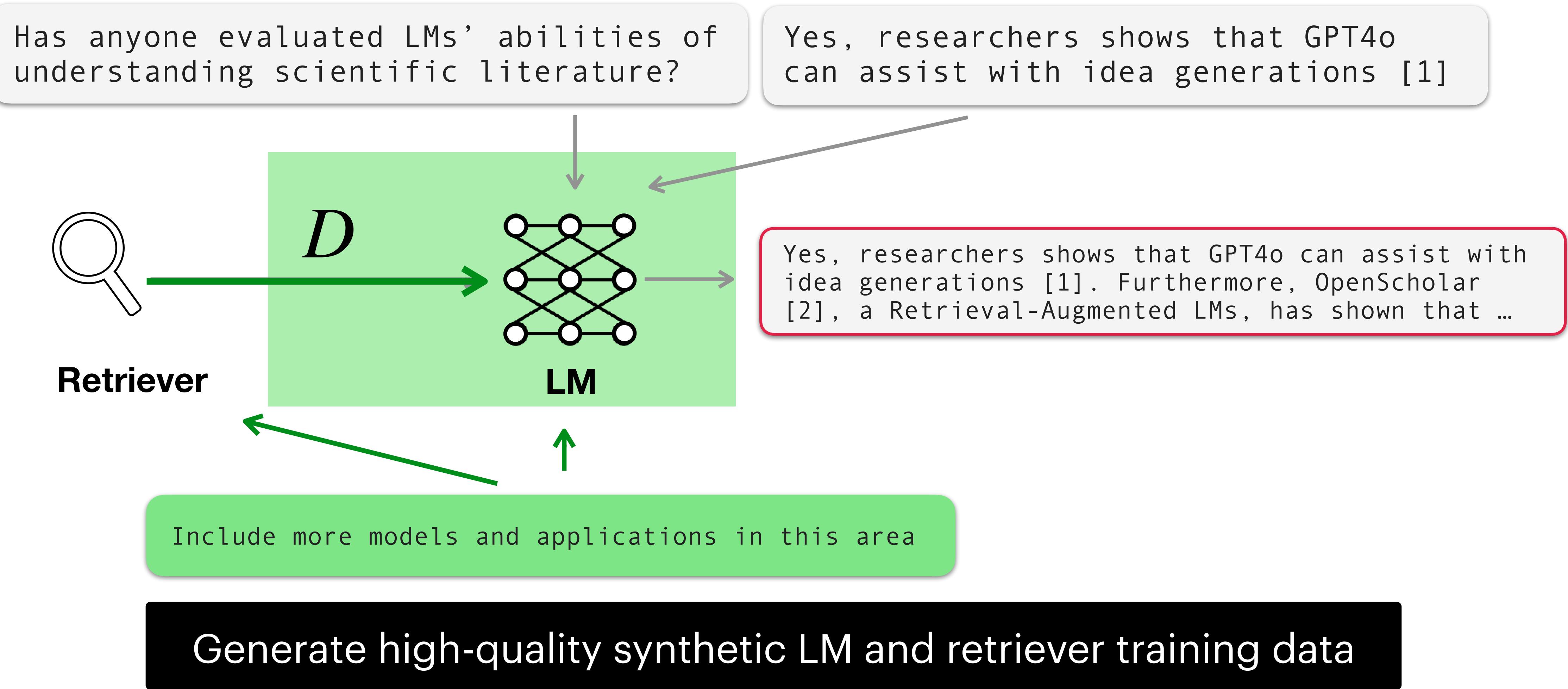


Yes, researchers shows that GPT4o can assist with idea generations [1]. Furthermore, OpenScholar [2], a Retrieval-Augmented LMs, has shown that ...

Include more models and applications in this area

Asai et al. ICLR 2024 (Oral - Top 1%). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

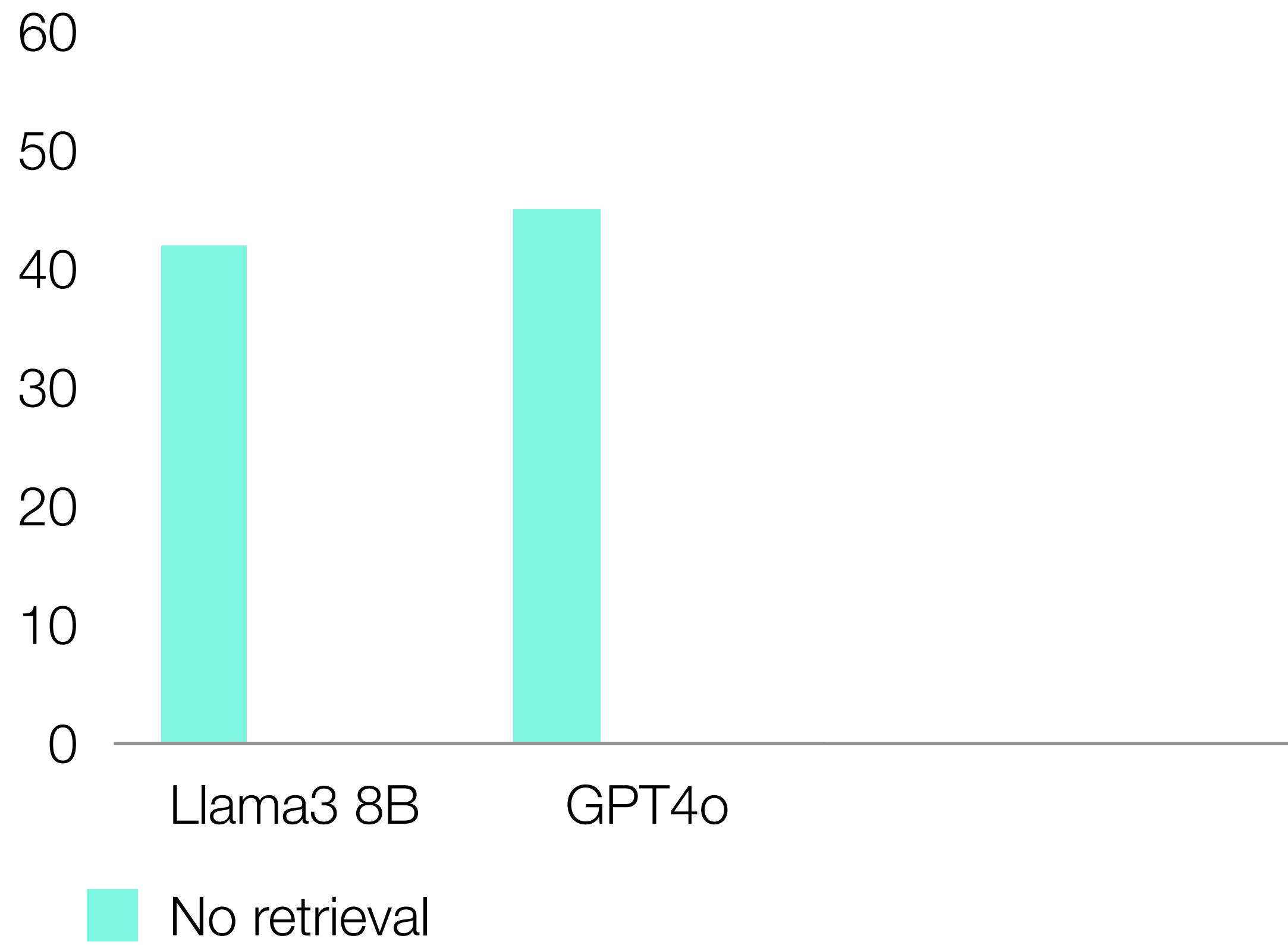
OpenScholar: Self-Feedback guided Generation



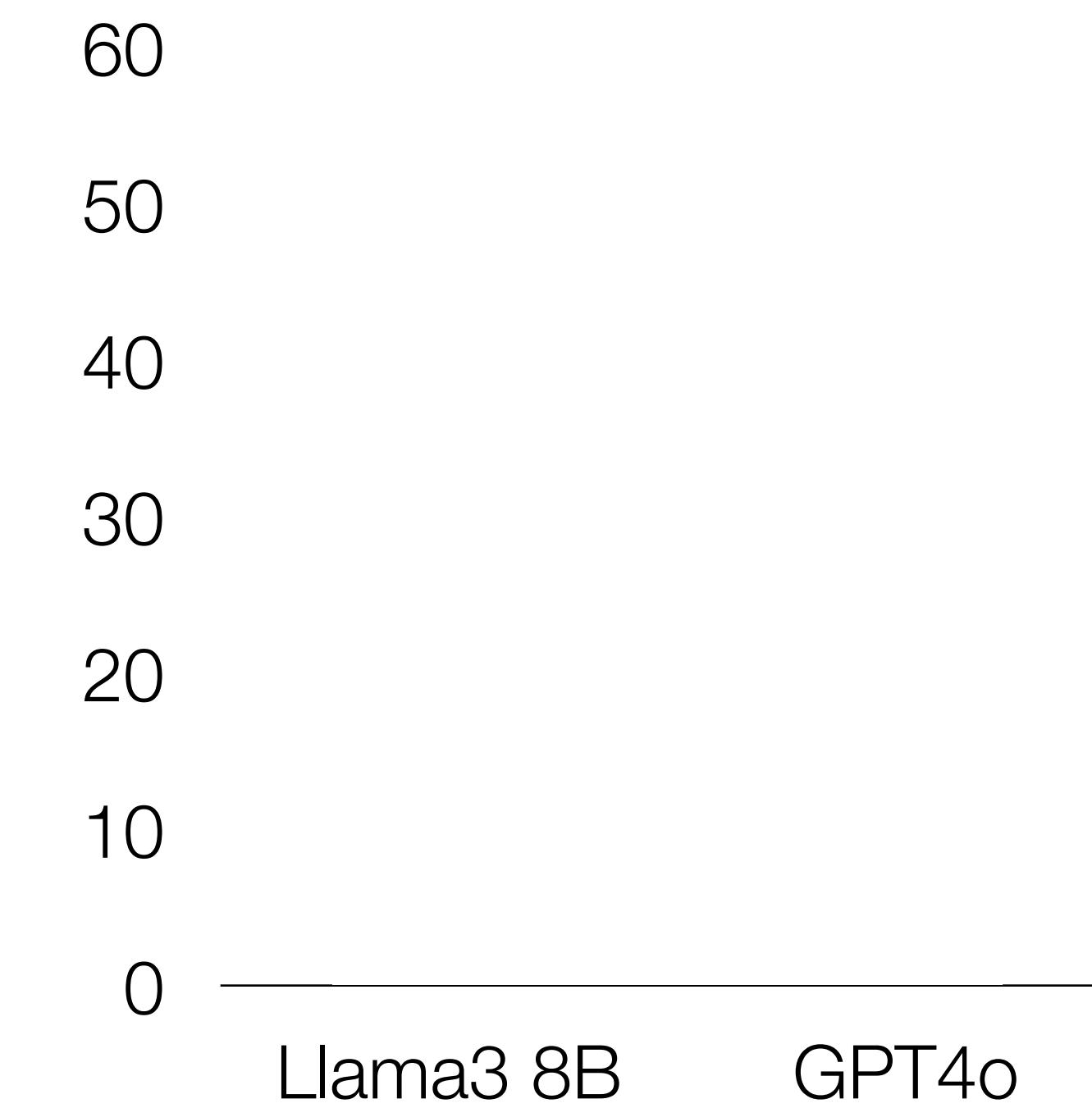
Asai et al. ICLR 2024 (Oral - Top 1%). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection.

Results: ScholarQABench (CS)

Rubric Accuracy (%)

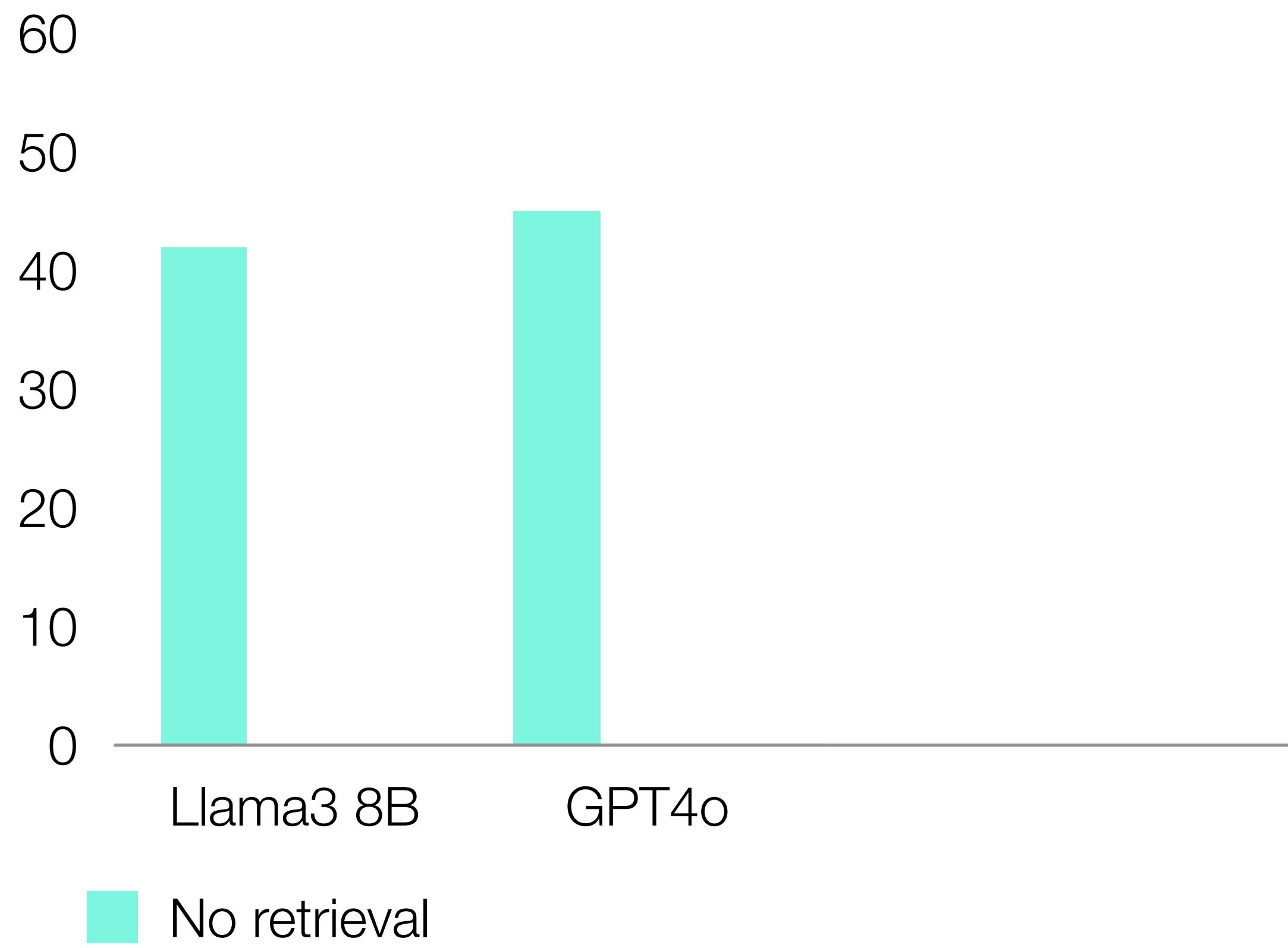


Citation Accuracy (%)

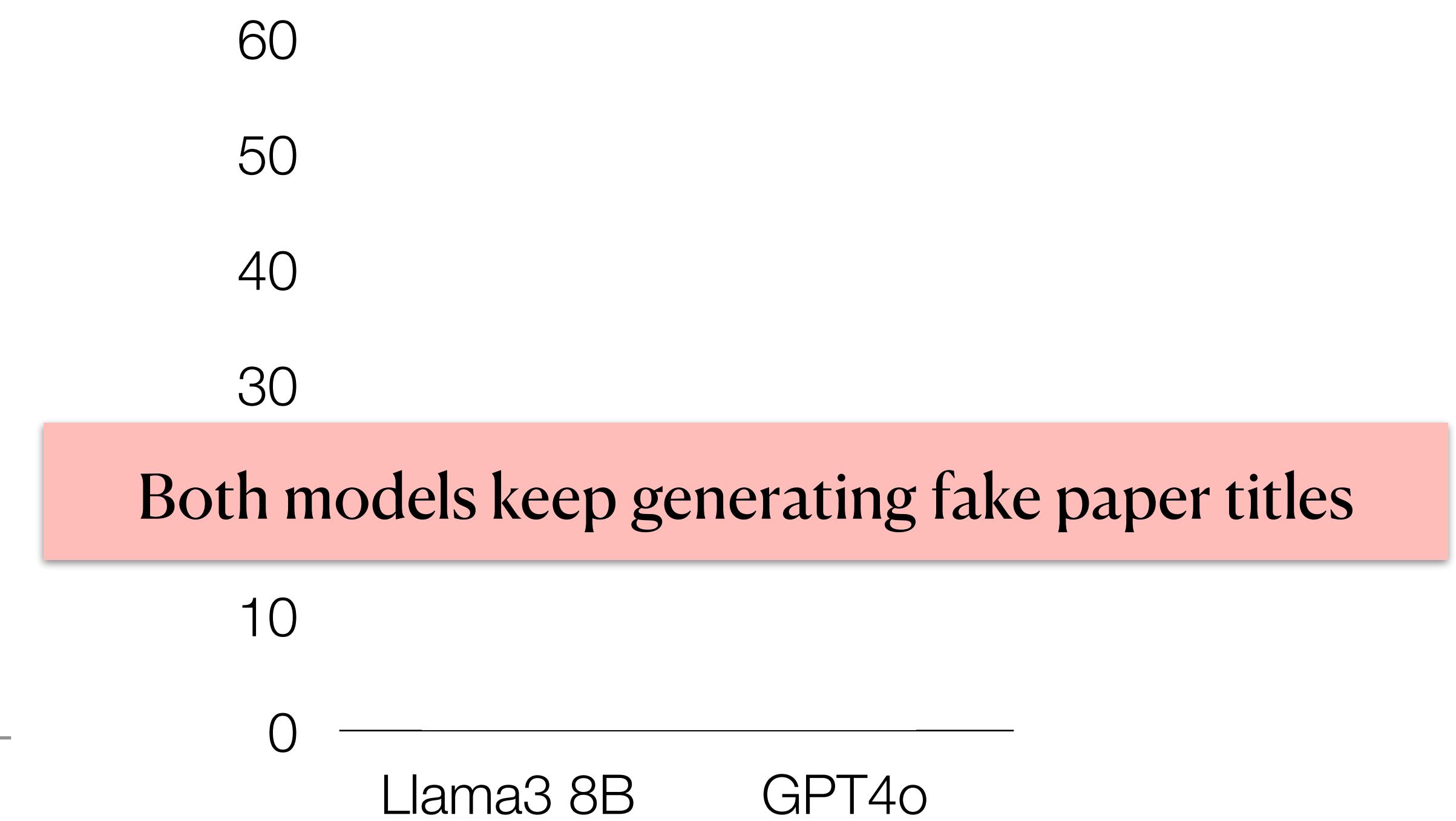


Results: ScholarQABench (CS)

Rubric Accuracy (%)

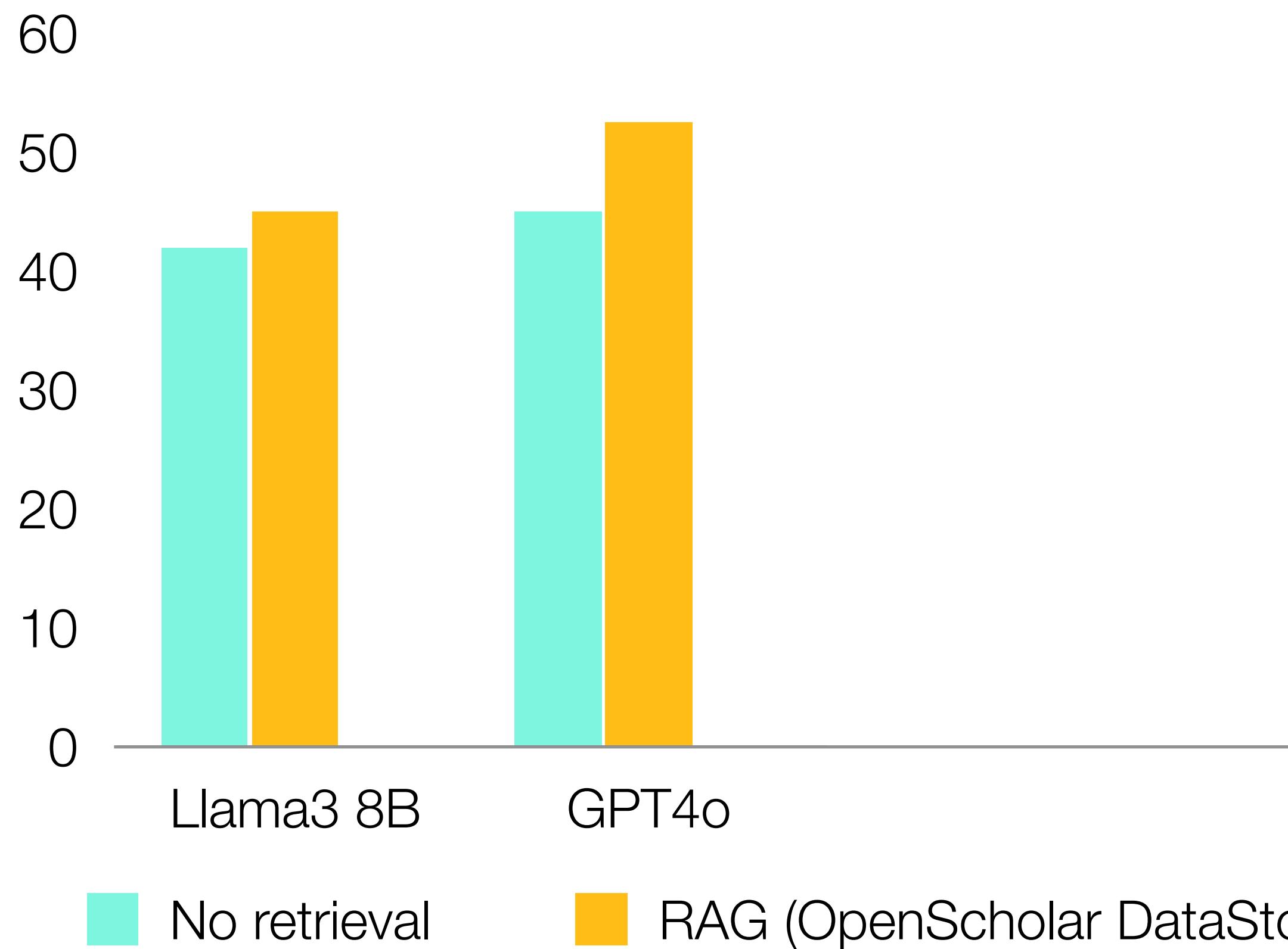


Citation Accuracy (%)

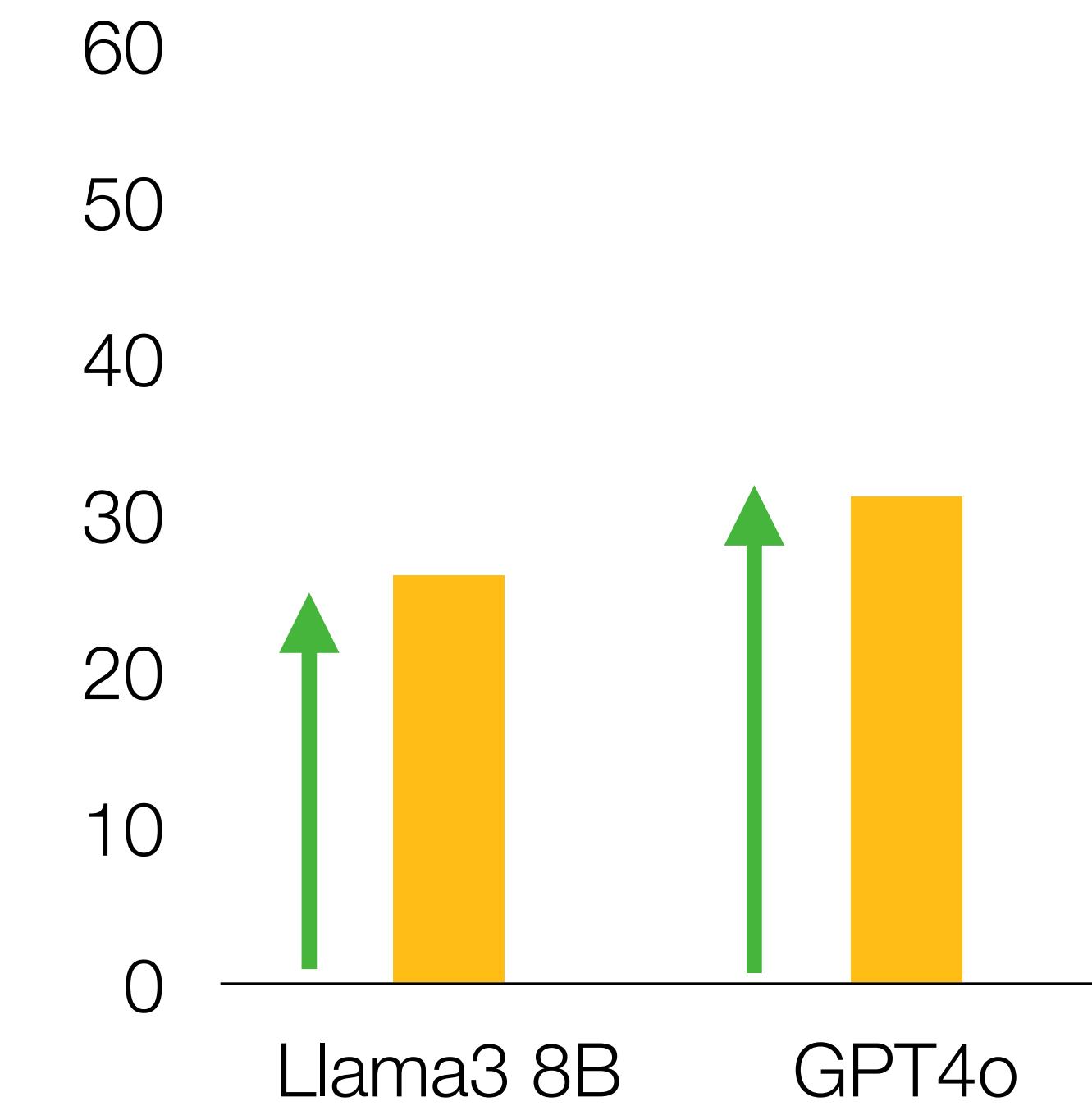


Results: ScholarQABench (CS)

Rubric Accuracy (%)

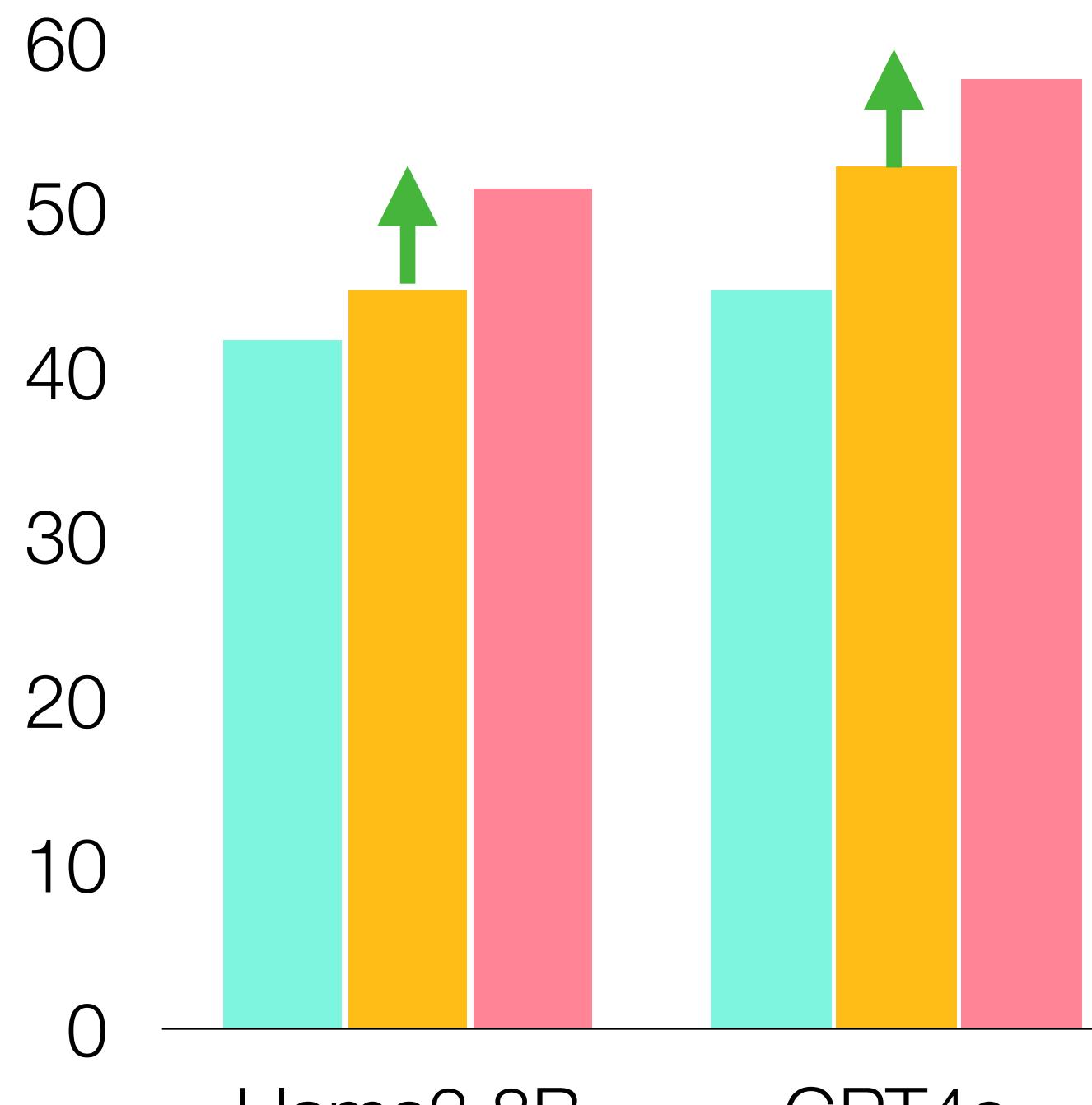


Citation Accuracy (%)



Results: ScholarQABench (CS)

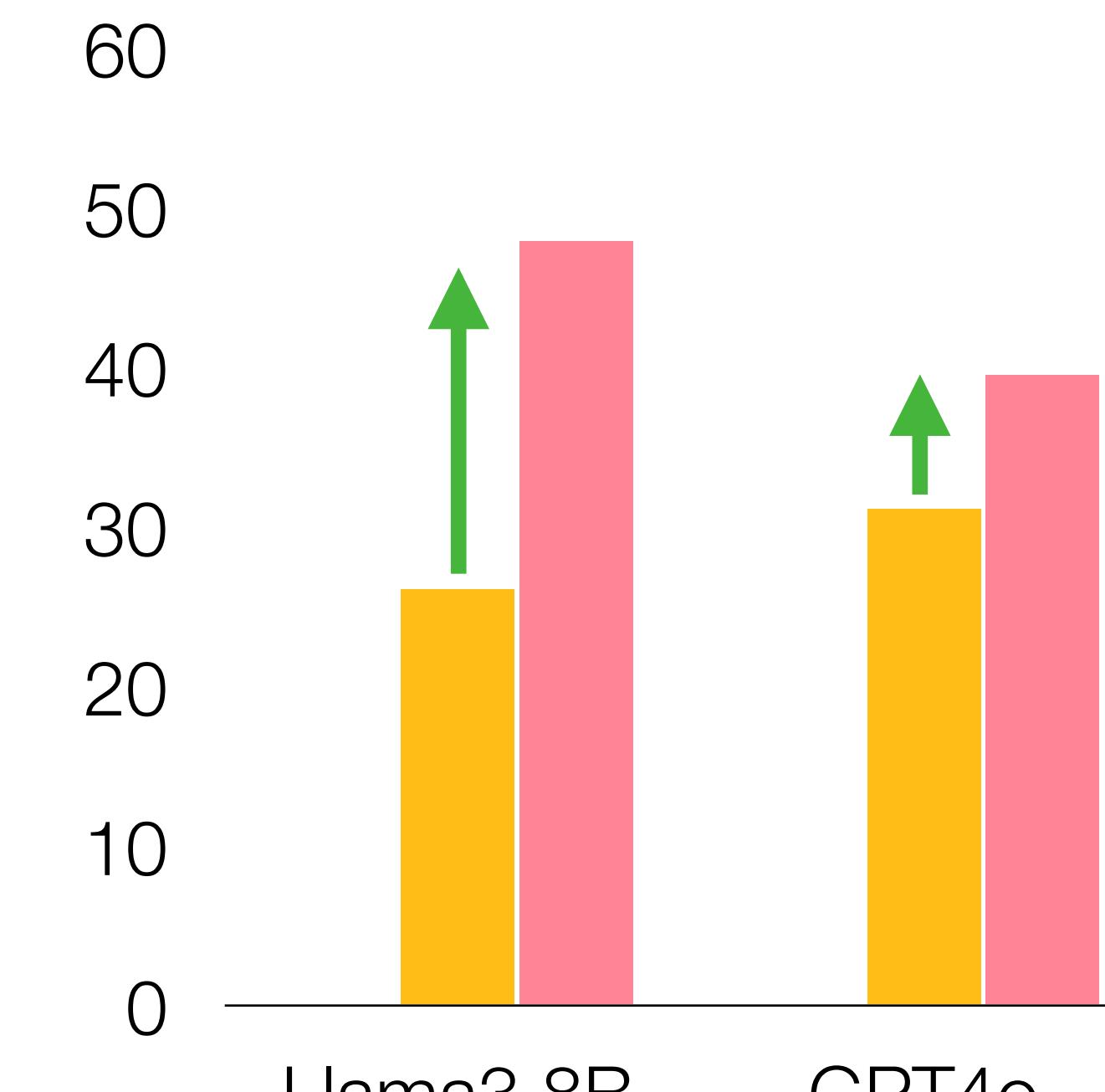
Rubric Accuracy (%)



No retrieval

RAG (OpenScholar DataStore)

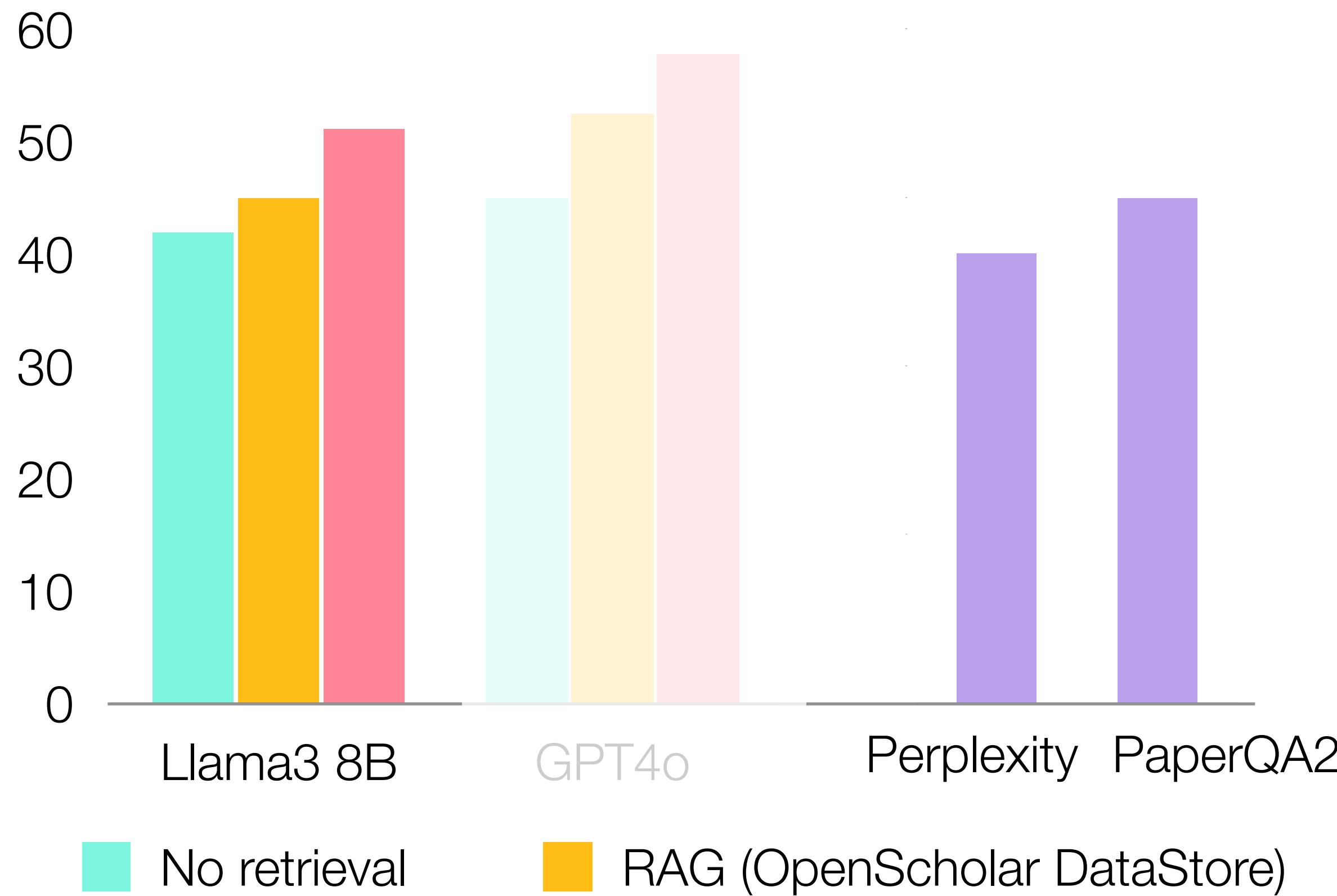
Citation Accuracy (%)



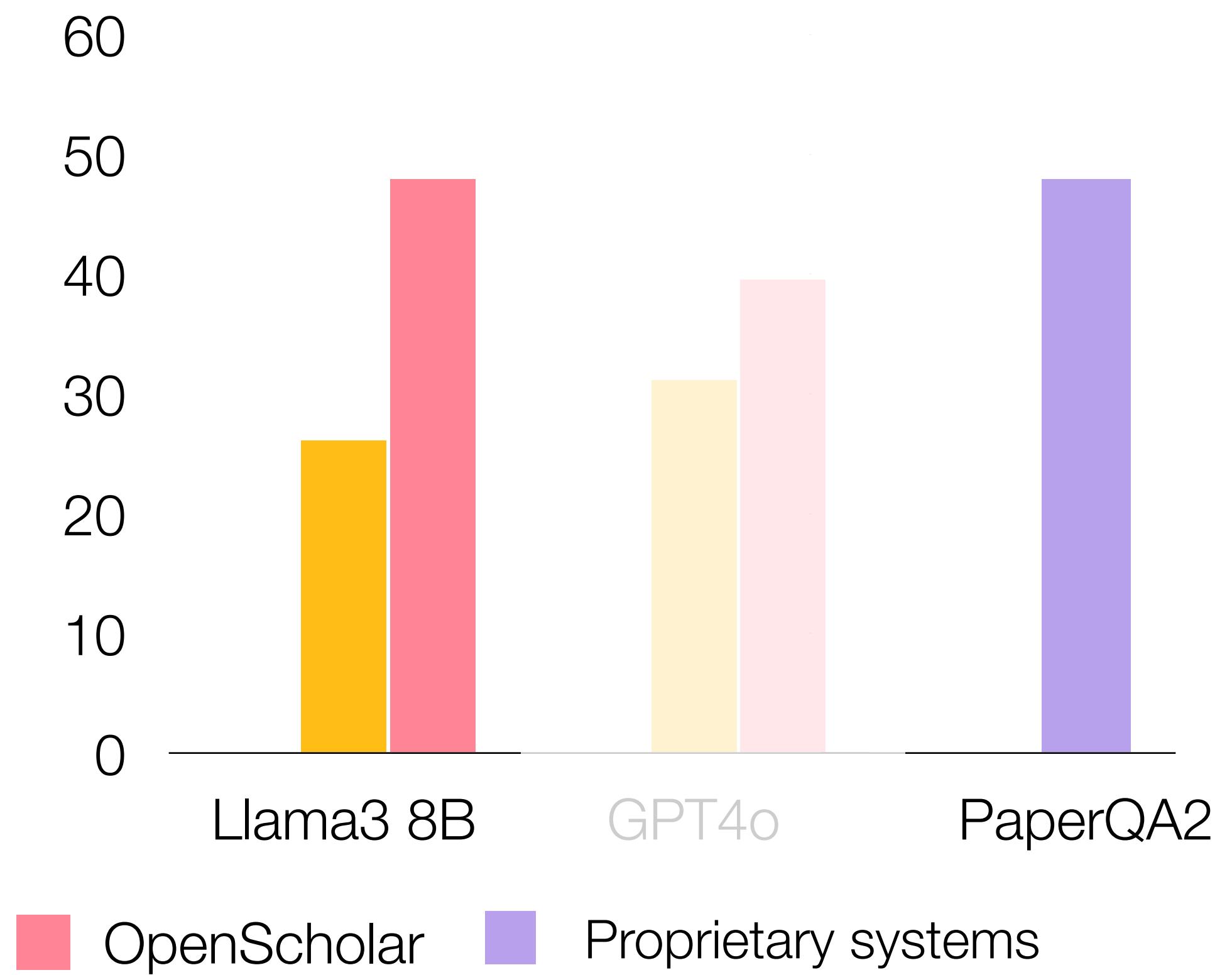
OpenScholar

Results: ScholarQABench (CS)

Rubric Accuracy (%)

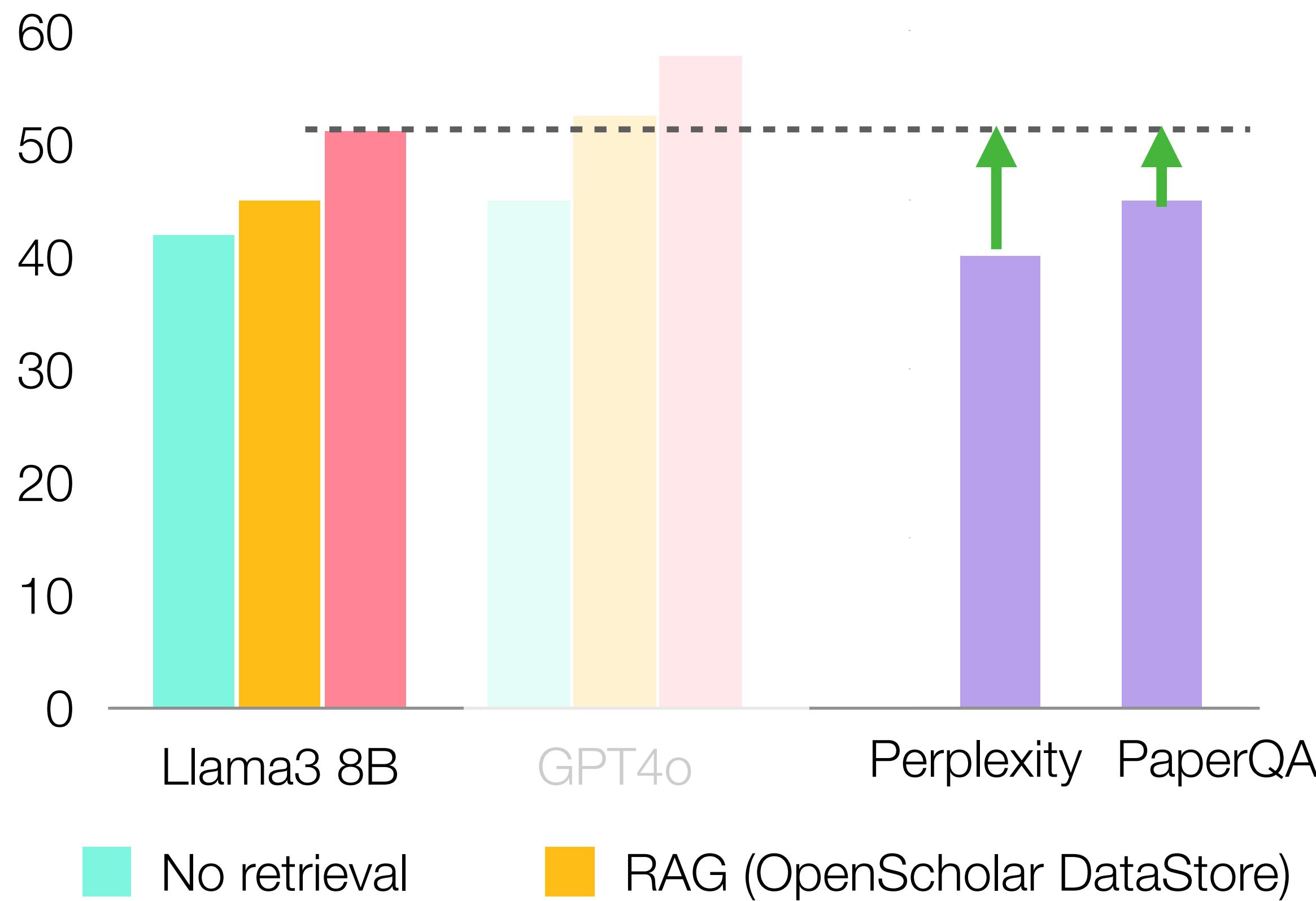


Citation Accuracy (%)

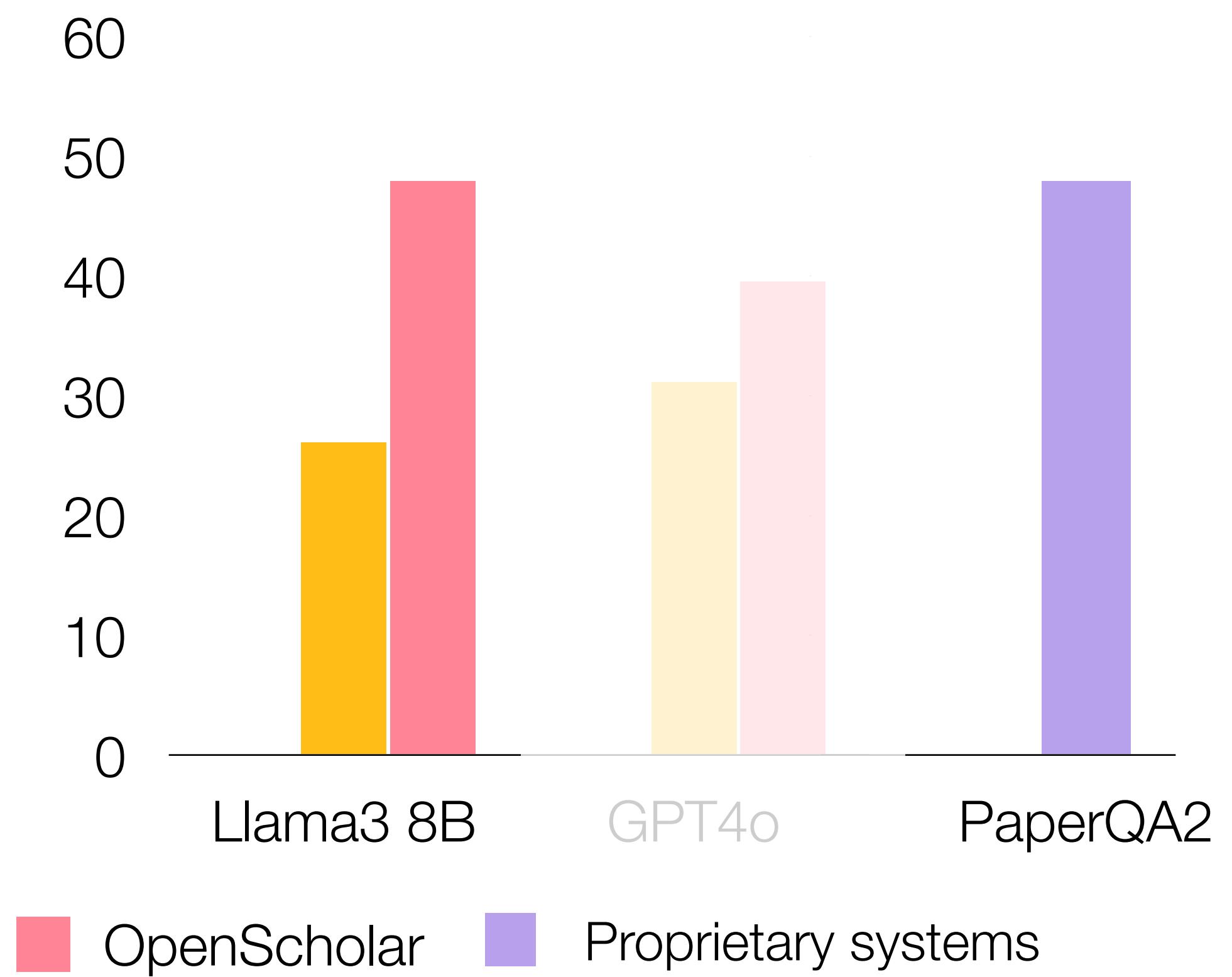


Results: ScholarQABench (CS)

Rubric Accuracy (%)



Citation Accuracy (%)



Expert Evaluations

■ Model win ■ Tie ■ Human (ScholarQAMulti, Gold Answer) Win

GPT4o vs Human

OS-8B vs Human

OS-GPT4o vs Human

0 25 50 75 100

%

Preference distribution

Expert Evaluations

■ Model win ■ Tie ■ Human (ScholarQAMulti, Gold Answer) Win

GPT4o vs Human

OS-8B vs Human

OS-GPT4o vs Human

0 25 50 75 100

%

Preference distribution

Expert Evaluations

■ Model win ■ Tie ■ Human (ScholarQAMulti, Gold Answer) Win

GPT4o vs Human

OS-8B vs Human

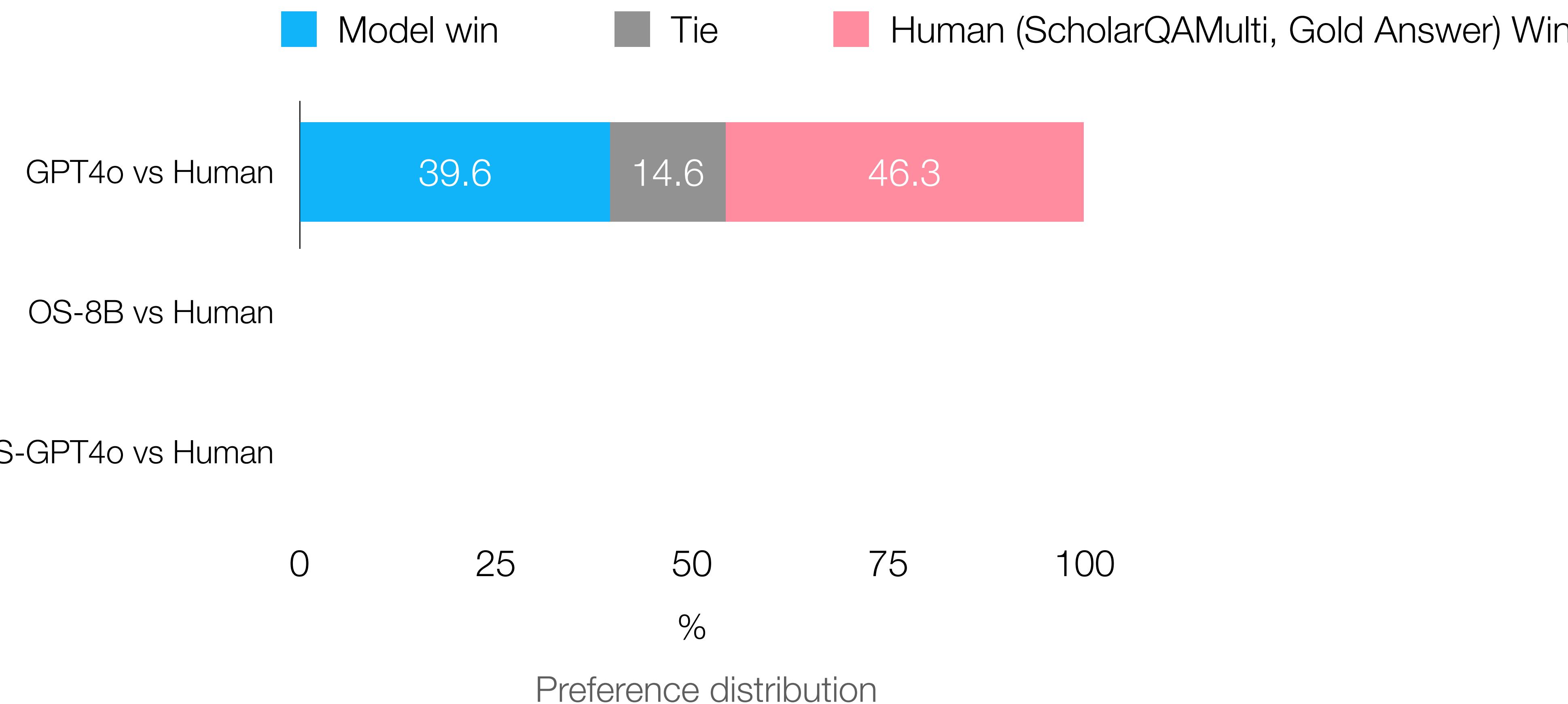
OS-GPT4o vs Human

0 25 50 75 100

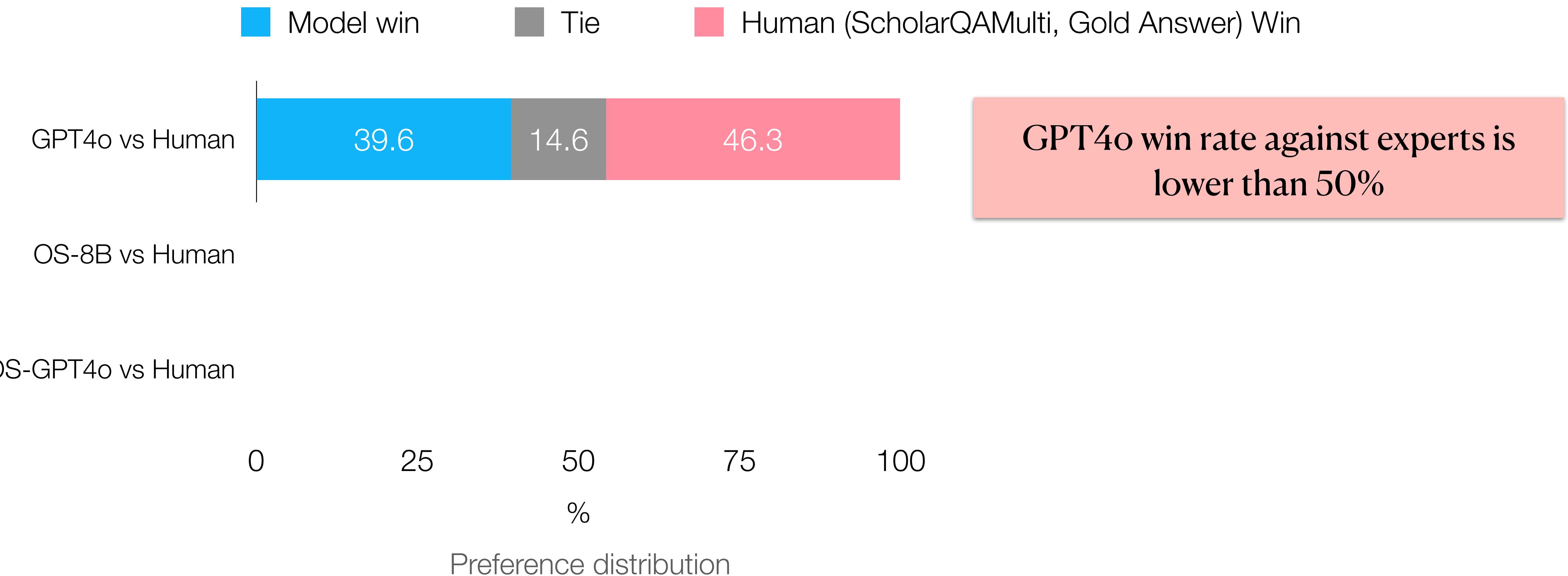
%

Preference distribution

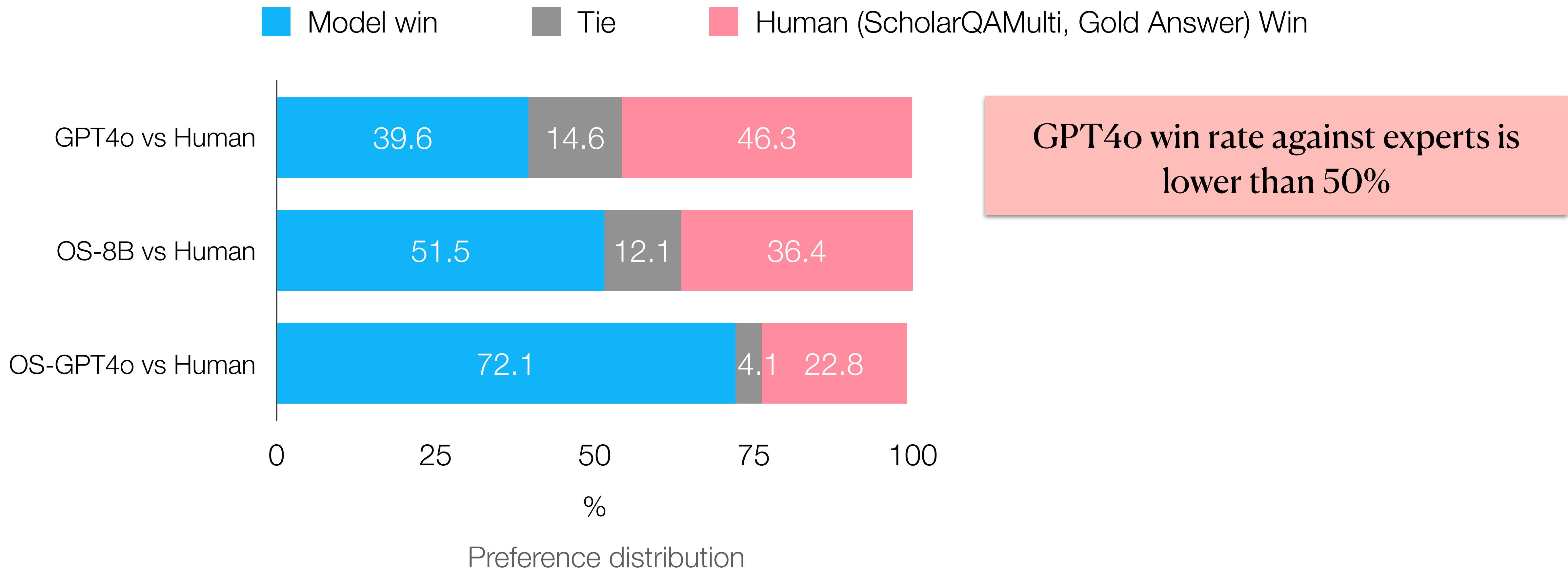
Expert Evaluations



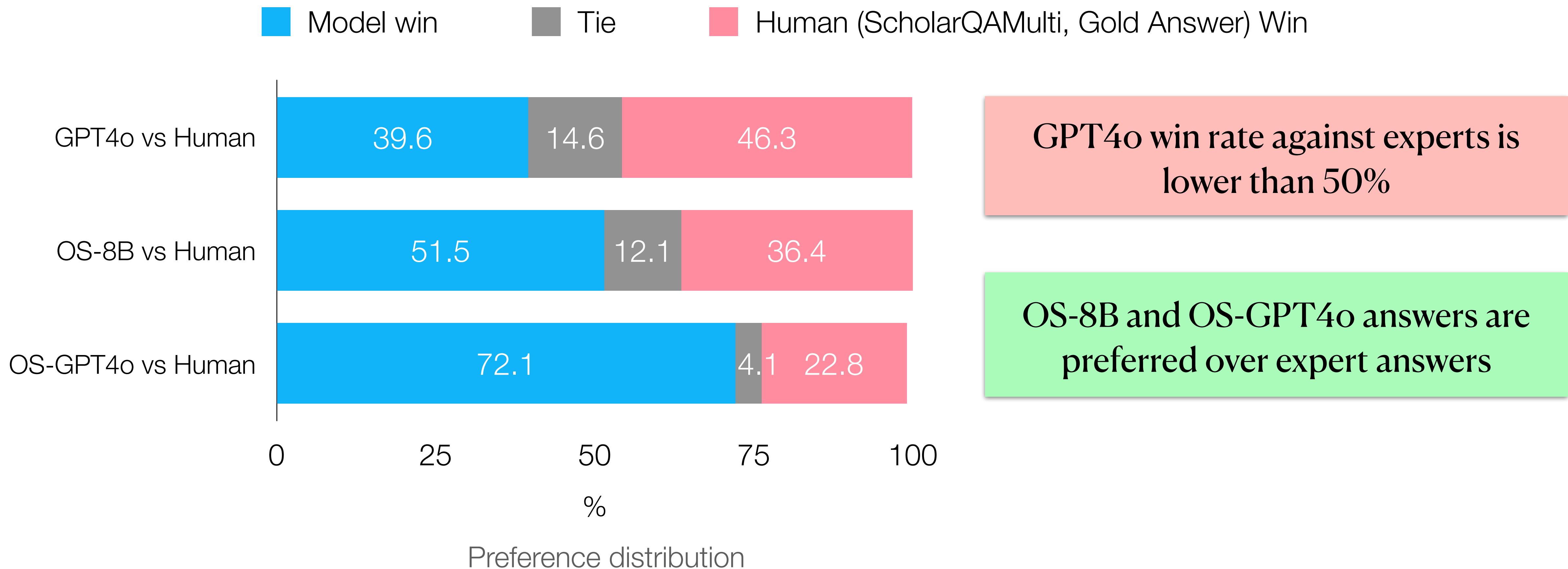
Expert Evaluations



Expert Evaluations



Expert Evaluations



Impact

Impact

First open research synthesis tool

VentureBeat Starred 602 ▾

OpenScholar: The open-source A.I. that's outperforming GPT-4o in scientific research

OpenScholar, an innovative AI system by Allen Institute for AI and University of Washington, revolutionizes scientific research by...



Impact

First open research synthesis tool

 VentureBeat

[OpenScholar: The open-source A.I. that's outperforming GPT-4o in scientific research](#)

OpenScholar, an innovative AI system by Allen Institute for AI and University of Washington, revolutionizes scientific research by...



★ Starred 602 ▾

Reduce errors for medical communicator RAG
(Wong et al., 2025)

Detecting Plagiarism in AI generated ideas
(Gupta et al., 2025)

Impact

First open research synthesis tool

VentureBeat Starred 602 ▾

OpenScholar: The open-source A.I. that's outperforming GPT-4o in scientific research

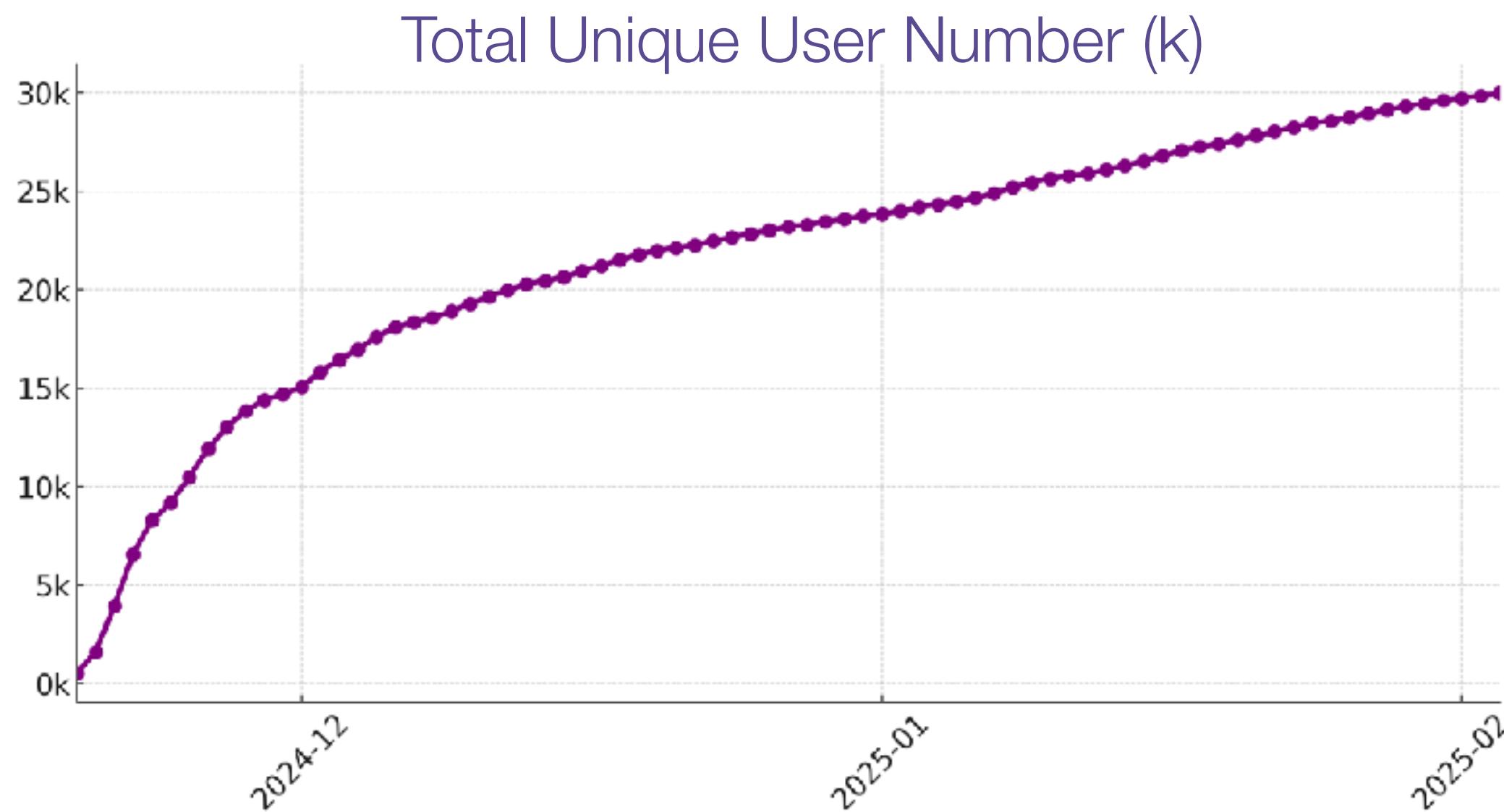
OpenScholar, an innovative AI system by Allen Institute for AI and University of Washington, revolutionizes scientific research by...



Reduce errors for medical communicator RAG
(Wong et al., 2025)

Detecting Plagiarism in AI generated ideas
(Gupta et al., 2025)

Over 30k users across disciplines



Impact

First open research synthesis tool

VentureBeat Starred 602 ▾

OpenScholar: The open-source A.I. that's outperforming GPT-4o in scientific research

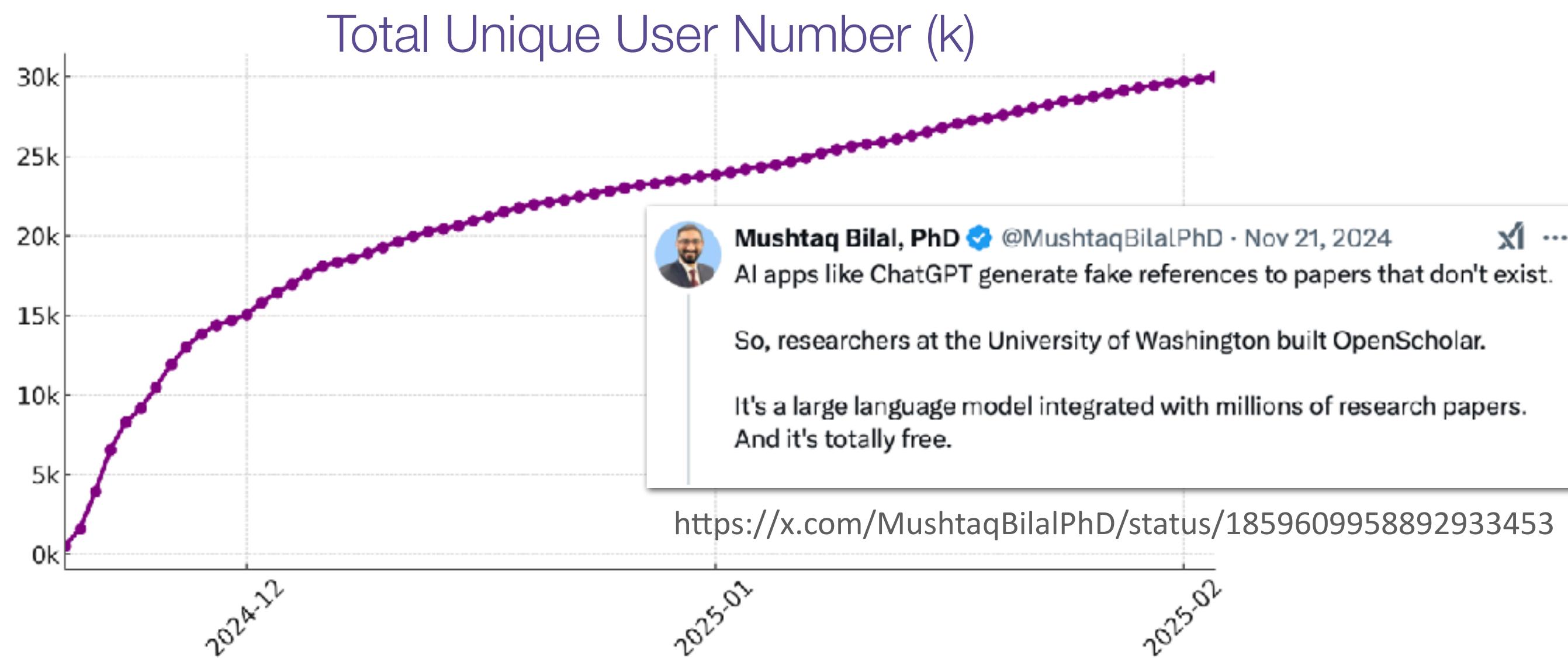
OpenScholar, an innovative AI system by Allen Institute for AI and University of Washington, revolutionizes scientific research by...



Reduce errors for medical communicator RAG
(Wong et al., 2025)

Detecting Plagiarism in AI generated ideas
(Gupta et al., 2025)

Over 30k users across disciplines



Impact

First open research synthesis tool

VentureBeat Starred 602 ▾

OpenScholar: The open-source A.I. that's outperforming GPT-4o in scientific research

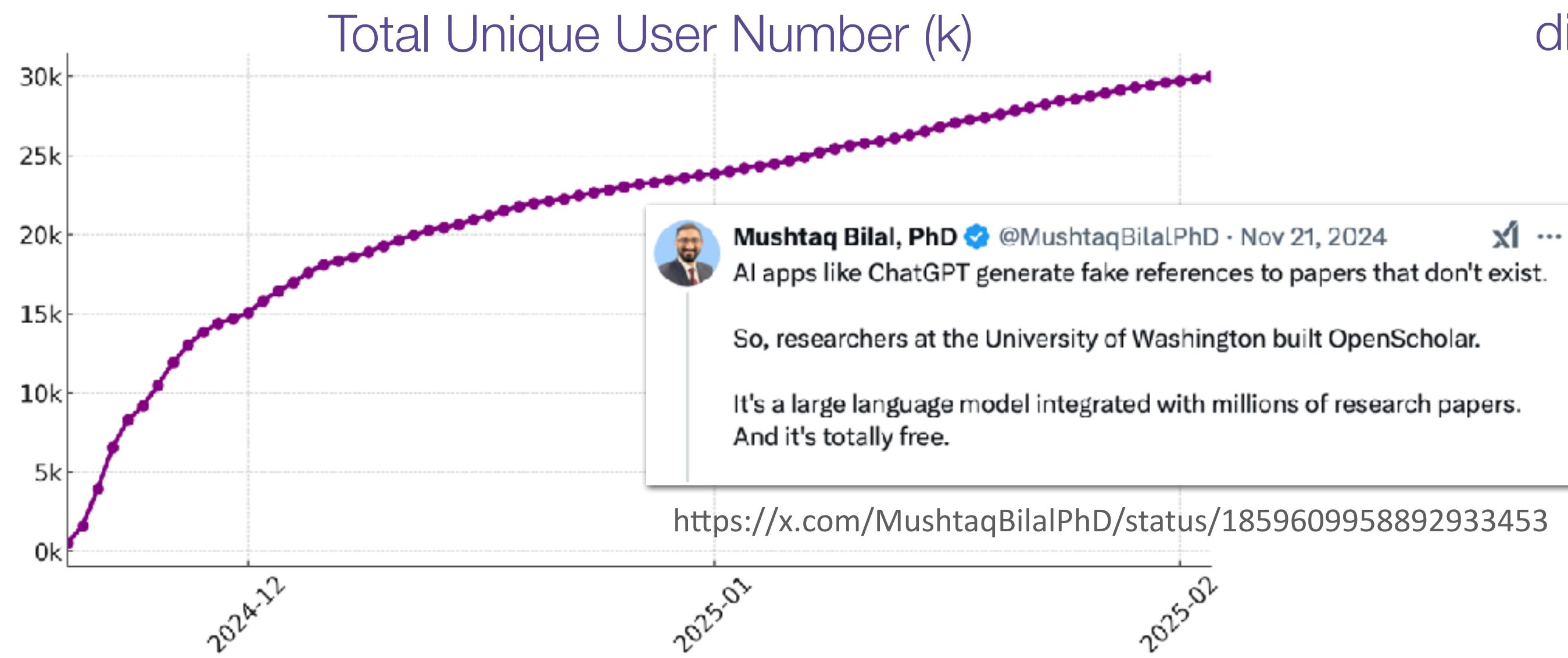
OpenScholar, an innovative AI system by Allen Institute for AI and University of Washington, revolutionizes scientific research by...



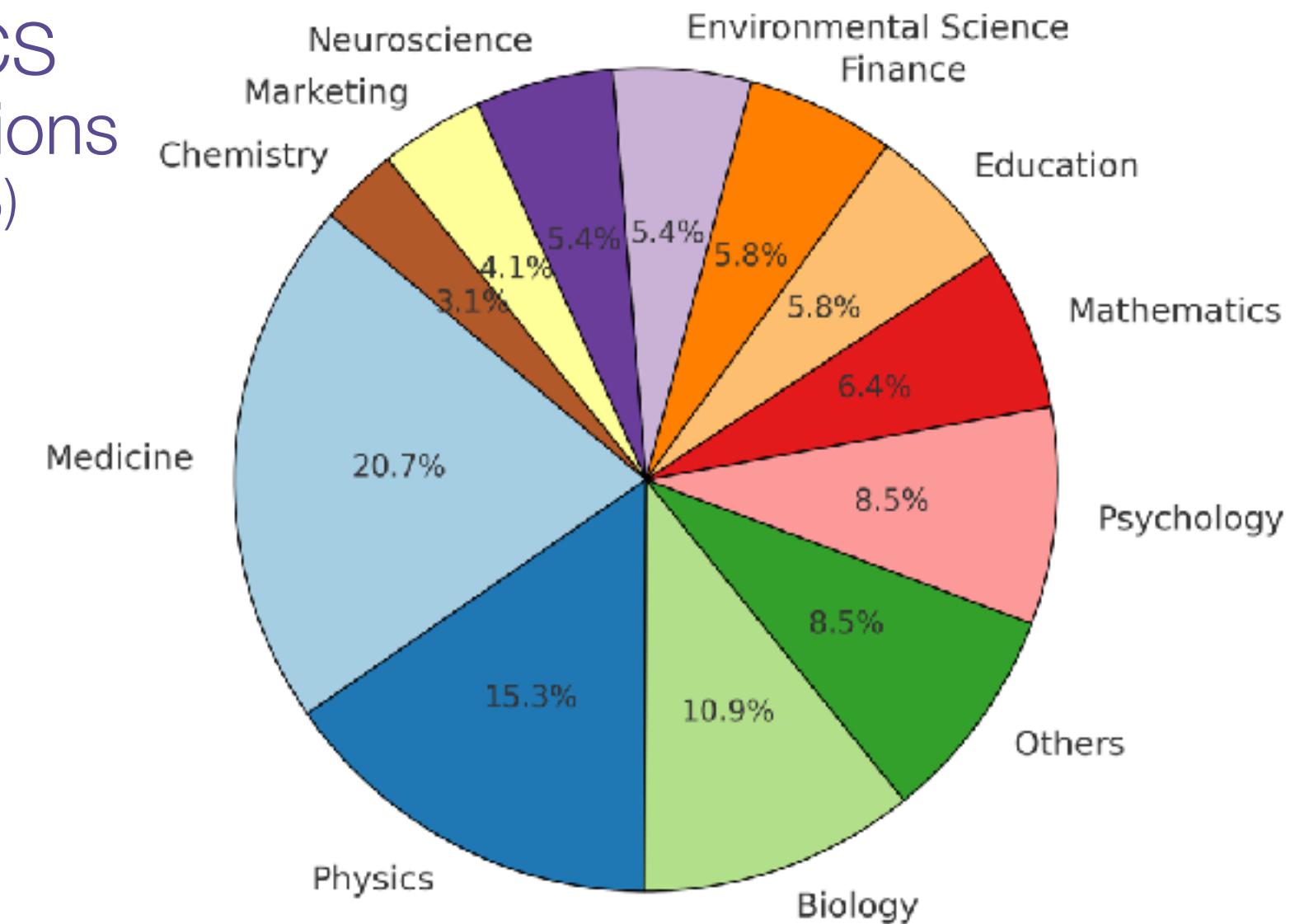
Reduce errors for medical communicator RAG
(Wong et al., 2025)

Detecting Plagiarism in AI generated ideas
(Gupta et al., 2025)

Over 30k users across disciplines



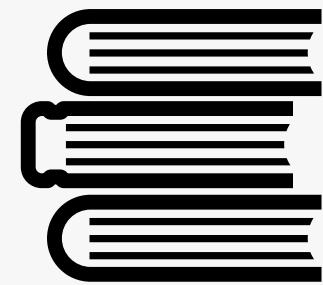
Non-CS distributions (~55%)



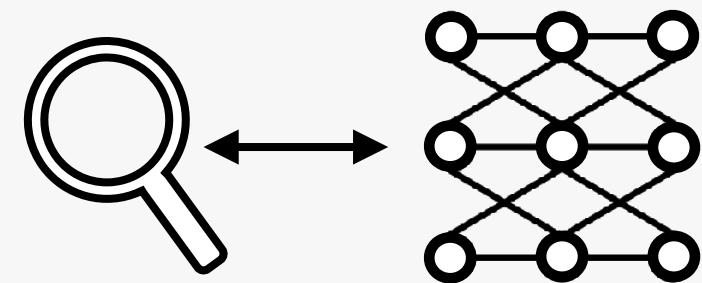
Summary



Develop new benchmarks with domain experts



Carefully curate **new datastore** for new applications



Training LM and retrievers using unlabeled datastore

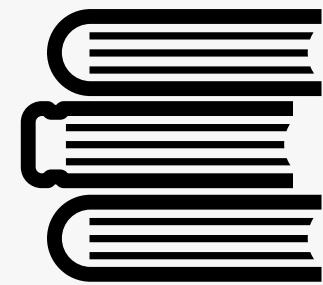


Open-source models and engage with **community**

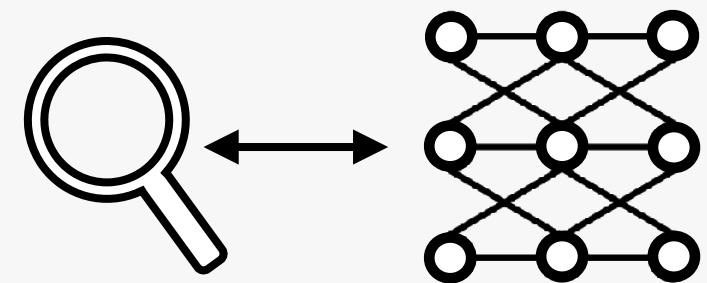
Summary



Develop new benchmarks with domain experts



Carefully curate **new datastore** for new applications



Training LM and retrievers using unlabeled datastore

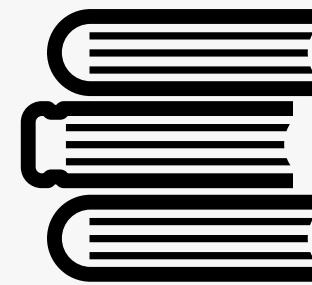


Open-source models and engage with **community**

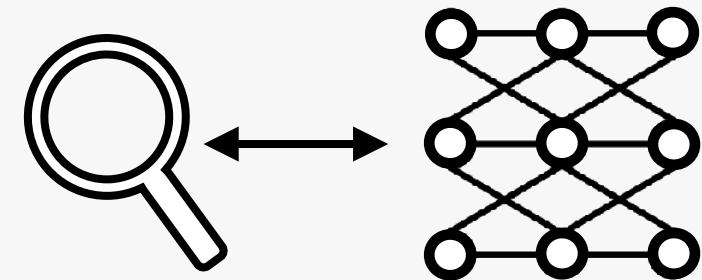
Summary



Develop new benchmarks with domain experts



Carefully curate **new datastore** for new applications



Training LM and retrievers using unlabeled datastore



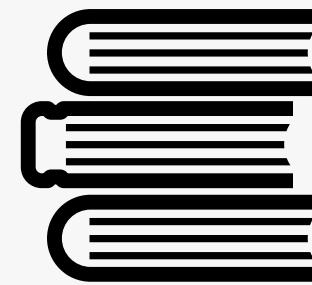
Open-source models and engage with **community**

Wang*, **Asai*** (contributed equally) et al. Findings of NAACL 2025. CodeRAG-Bench: Can Retrieval Augment Code Generation?

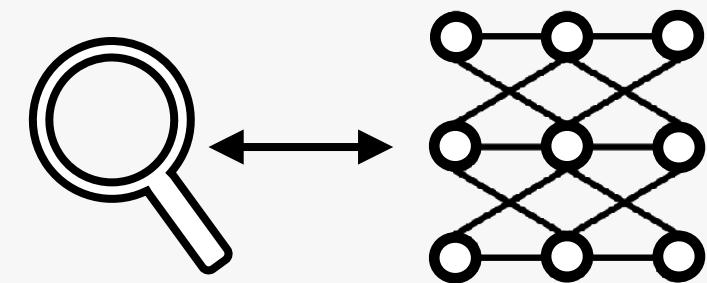
Summary



Develop new benchmarks with domain experts



Carefully curate **new datastore** for new applications



Training LM and retrievers using unlabeled datastore



Open-source models and engage with **community**

Wang*, **Asai*** (contributed equally) et al. Findings of NAACL 2025. CodeRAG-Bench: Can Retrieval Augment Code Generation?

Asai et al. NeurIPS 2021. One Question Answering Model for Many Languages with Cross-lingual Dense Passage Retrieval & **Asai** et al. NAACL 2021 (Oral)

Beyond Scaling: Frontiers of Retrieval-Augmented LMs

Beyond Scaling: Frontiers of Retrieval-Augmented LMs



From Monolithic



Why?

- Systematically analyze **limitations of scaling monolithic LMs**
- Demonstrate their **reliability, adaptability** and **efficiency**

Beyond Scaling: Frontiers of Retrieval-Augmented LMs



↓
Why?

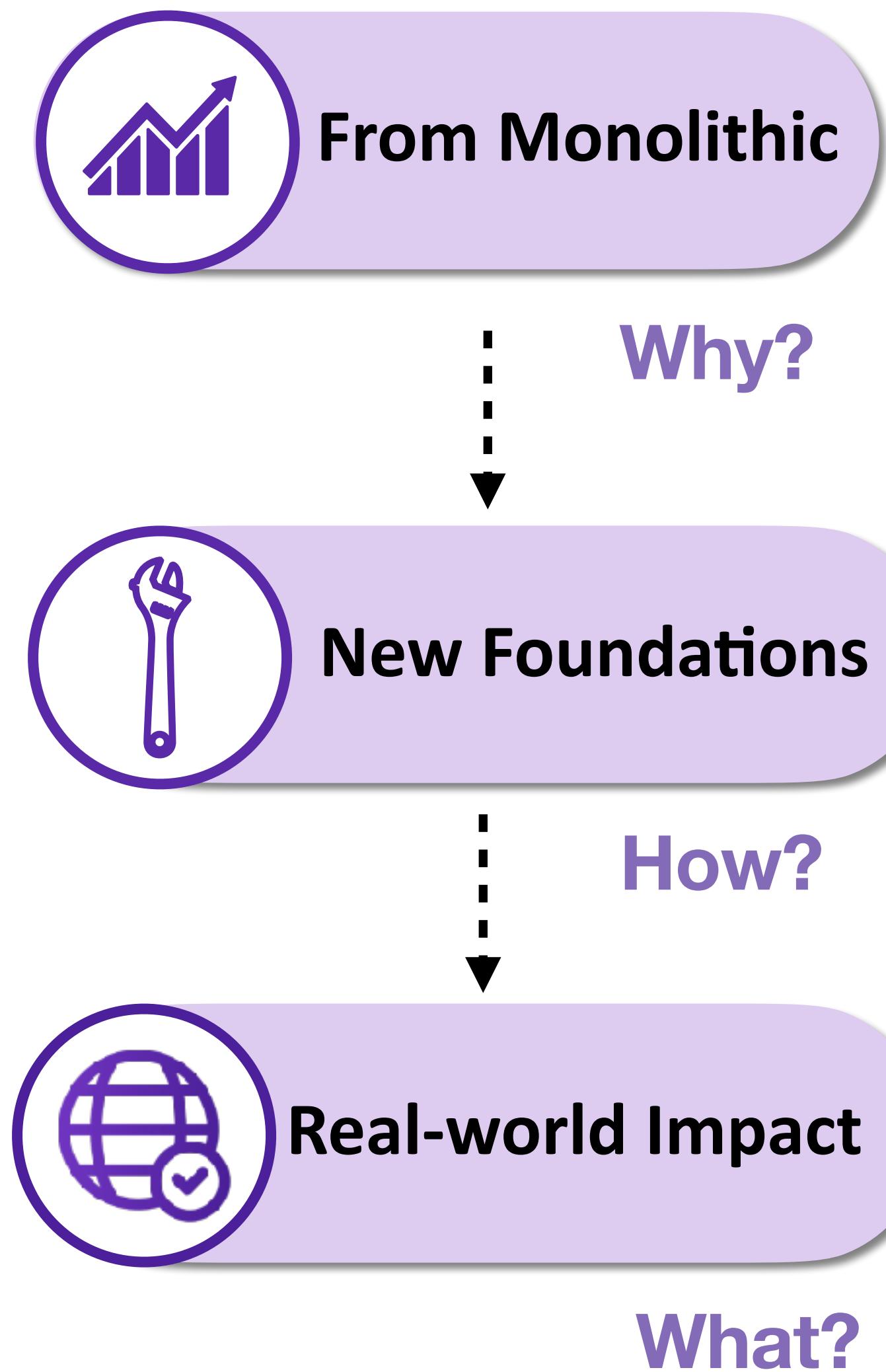


↓
How?

- Systematically analyze **limitations of scaling monolithic LMs**
- Demonstrate their **reliability, adaptability** and **efficiency**

- **Design** and **train** core components for Retrieval-Augmented LMs
- Integrated into major libraries, leading to state-of-the-art systems

Beyond Scaling: Frontiers of Retrieval-Augmented LMs

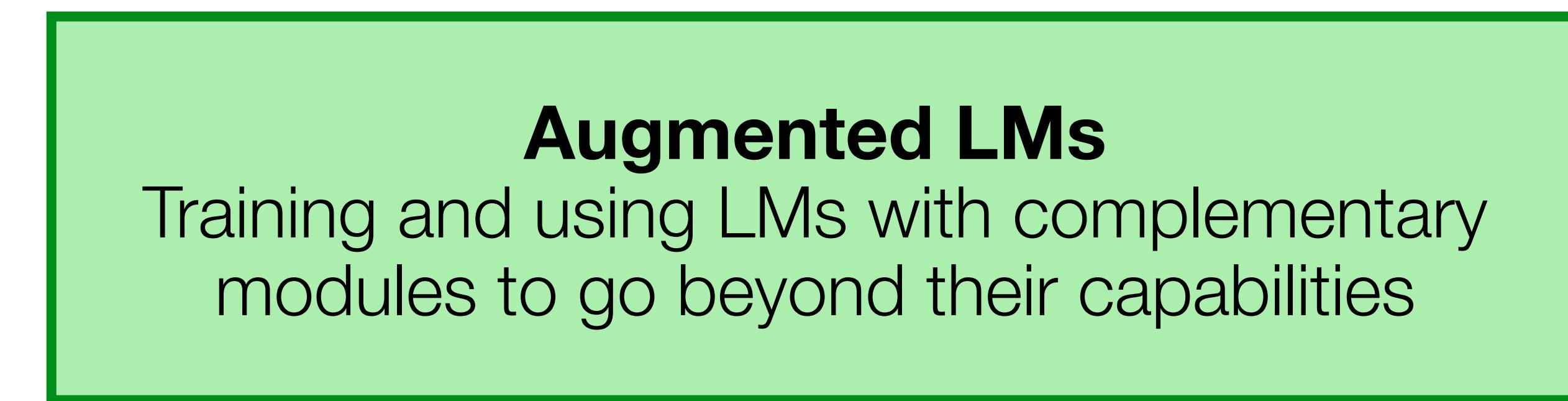
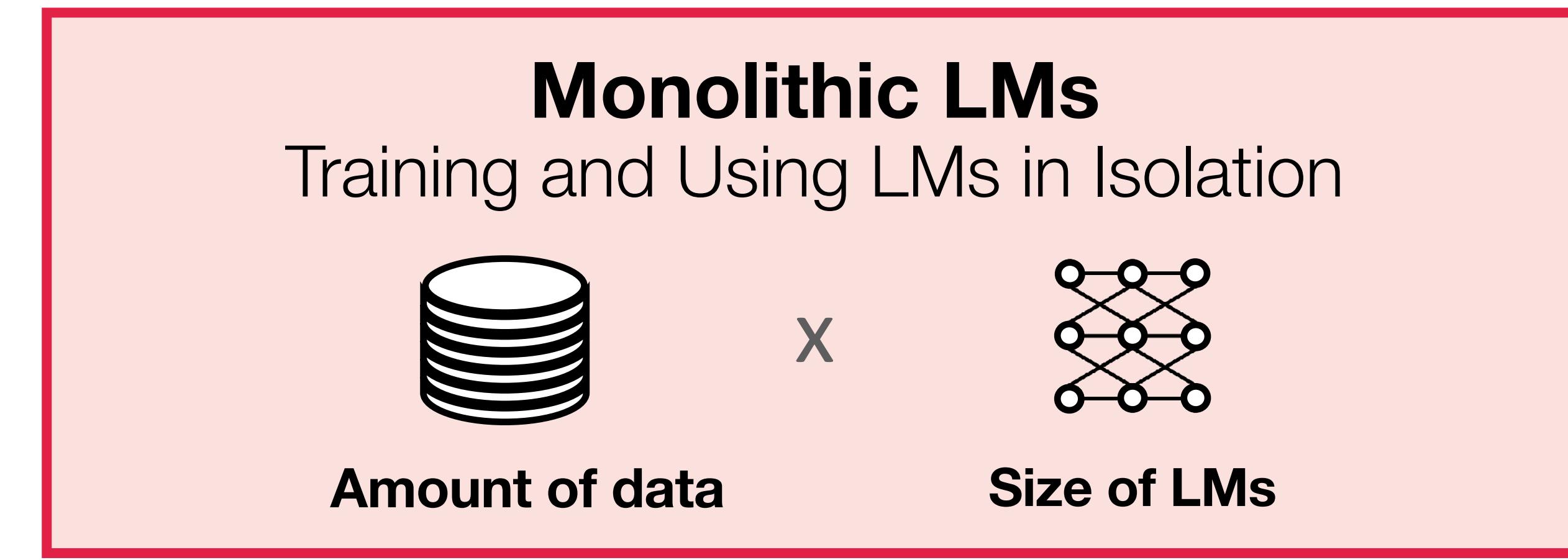
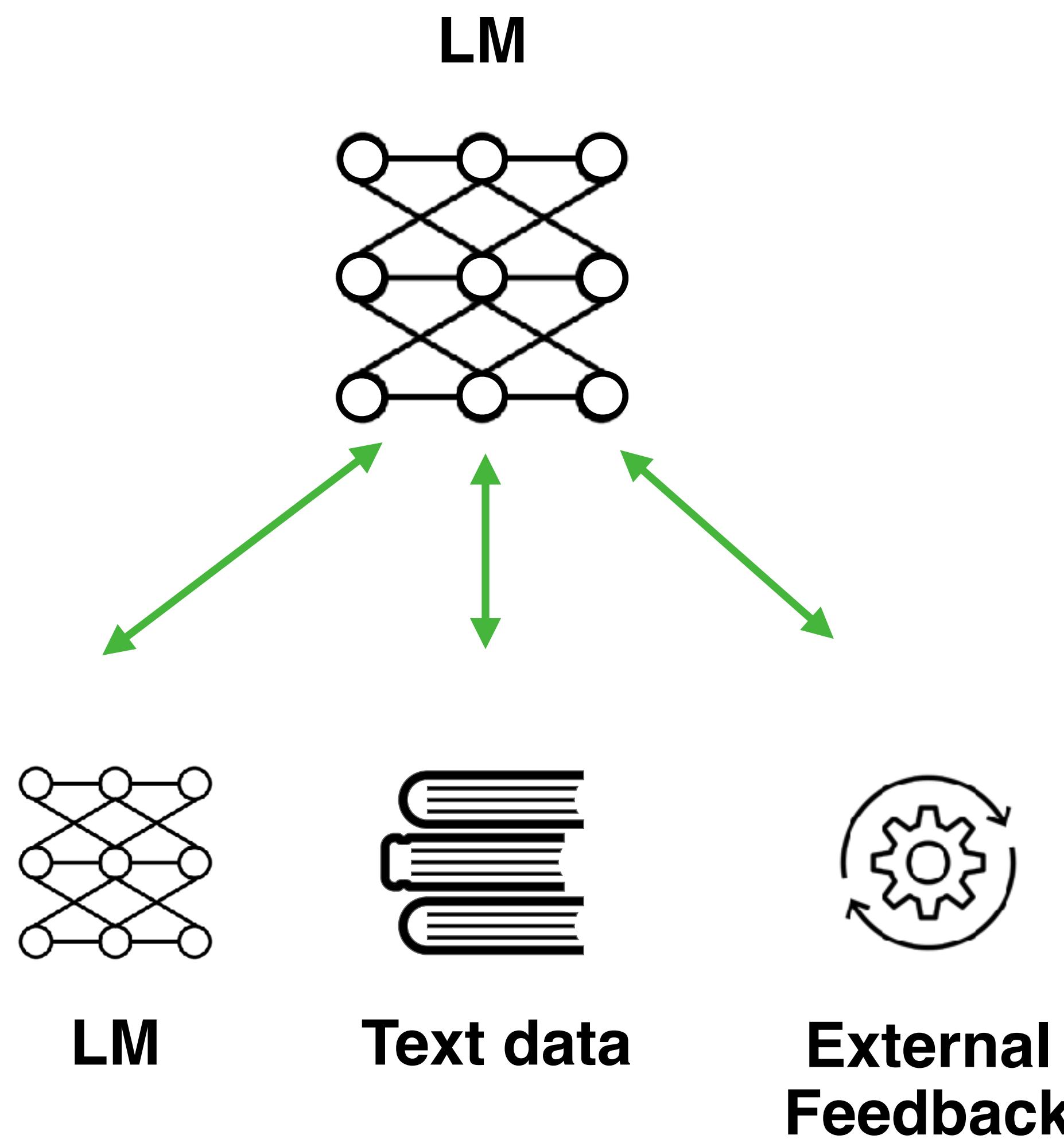


- Systematically analyze **limitations of scaling monolithic LMs**
- Demonstrate their **reliability, adaptability** and **efficiency**

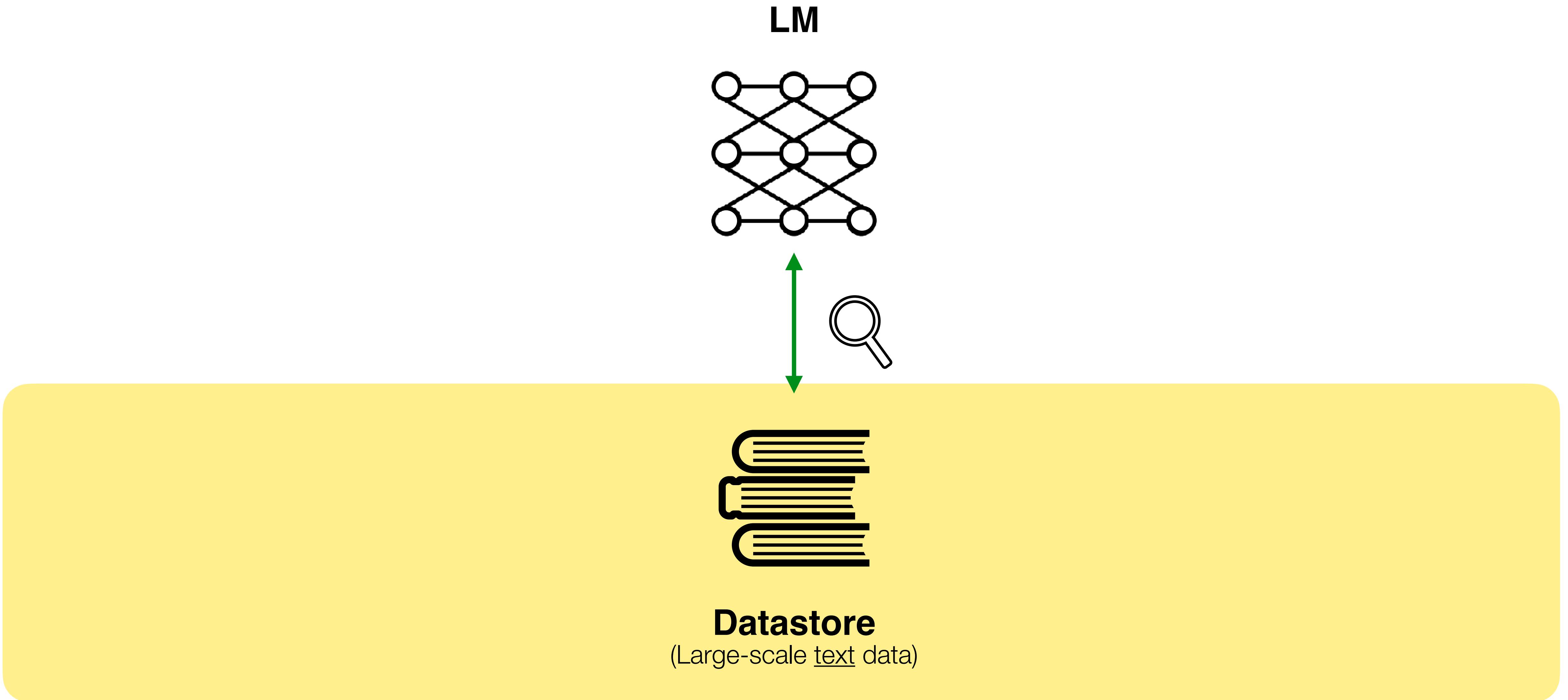
- **Design** and **train** core components for Retrieval-Augmented LMs
- Integrated into major libraries, leading to state-of-the-art systems

- Extend the foundations to drive impact in **scientific literature synthesis**
- Outperformed **GPT-4o** and **experts**; now used by **30K researchers**

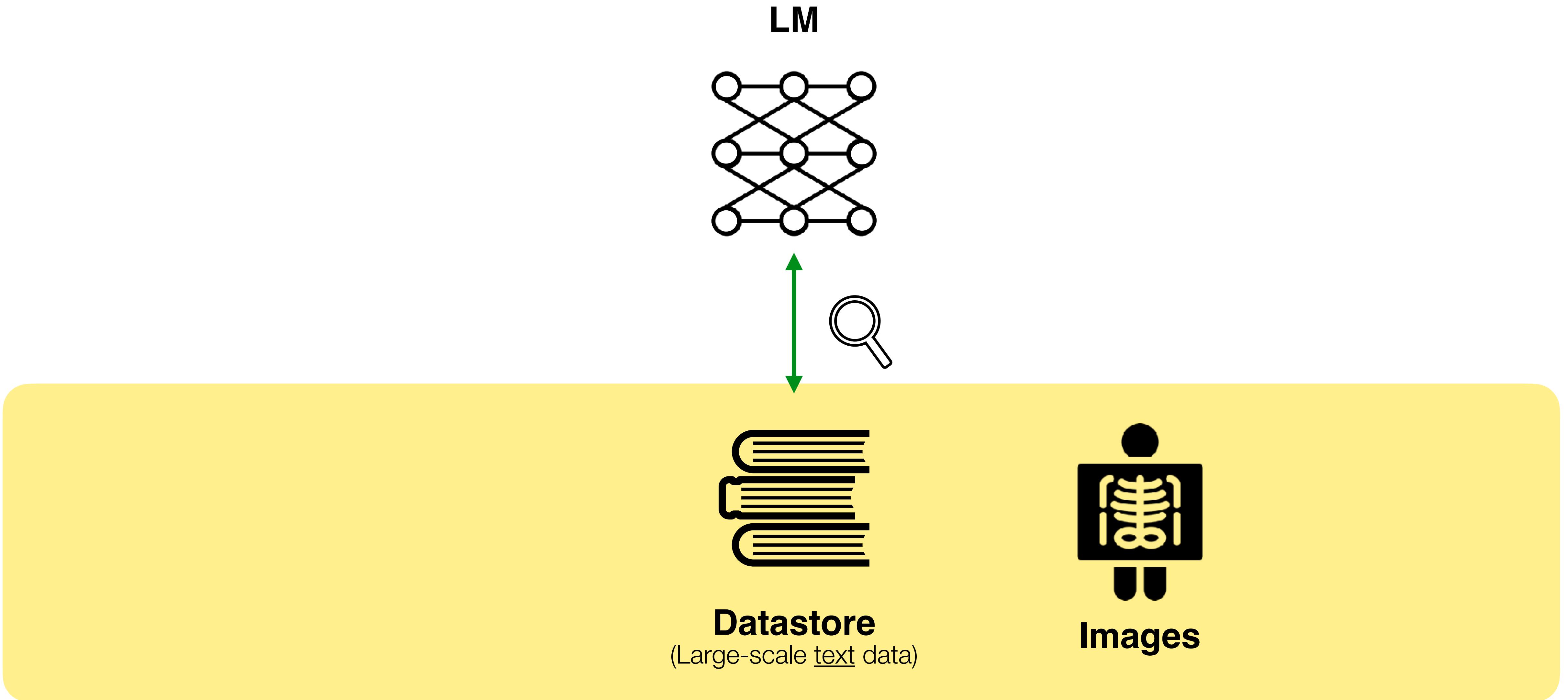
Vision: From Scaling **Monolithic LMs** to **Augmented LMs**



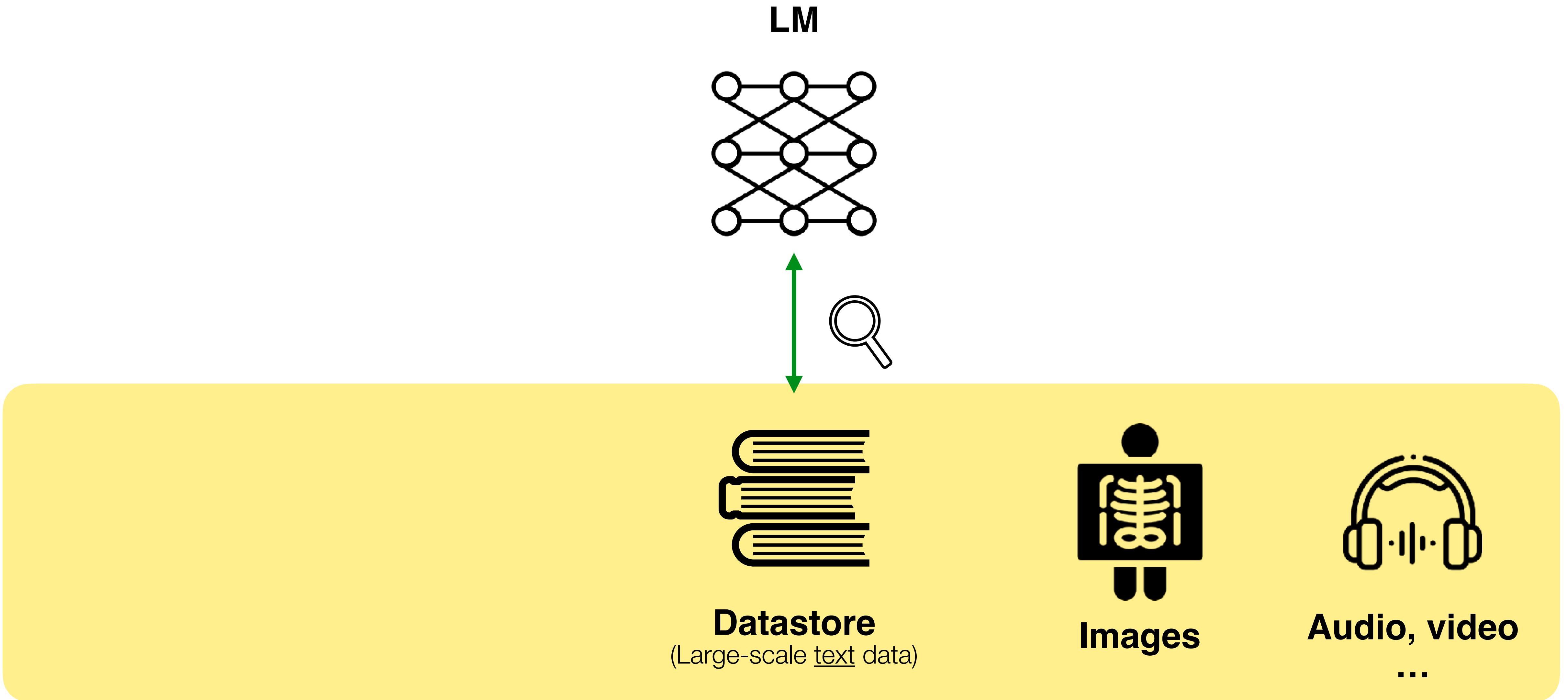
Future Work: Searching Over Heterogeneous Knowledge



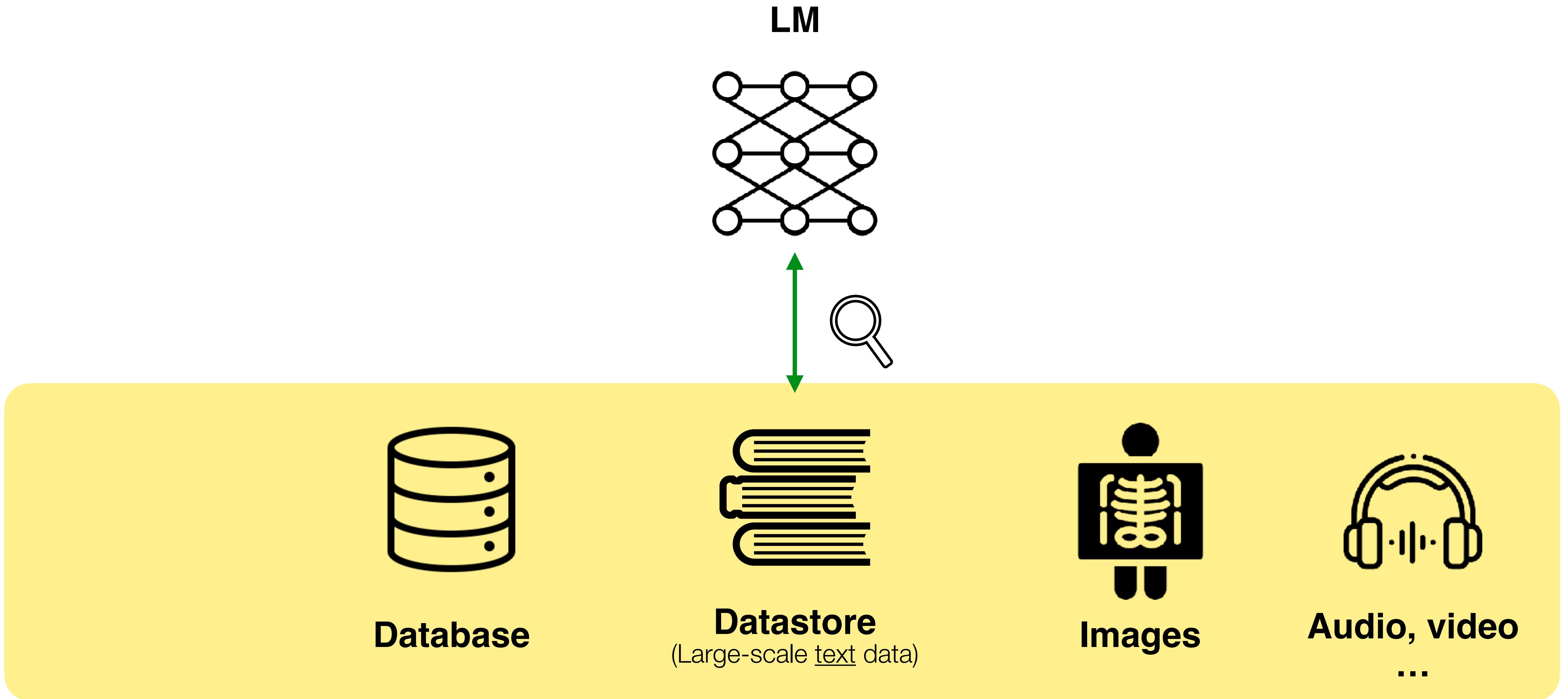
Future Work: Searching Over Heterogeneous Knowledge



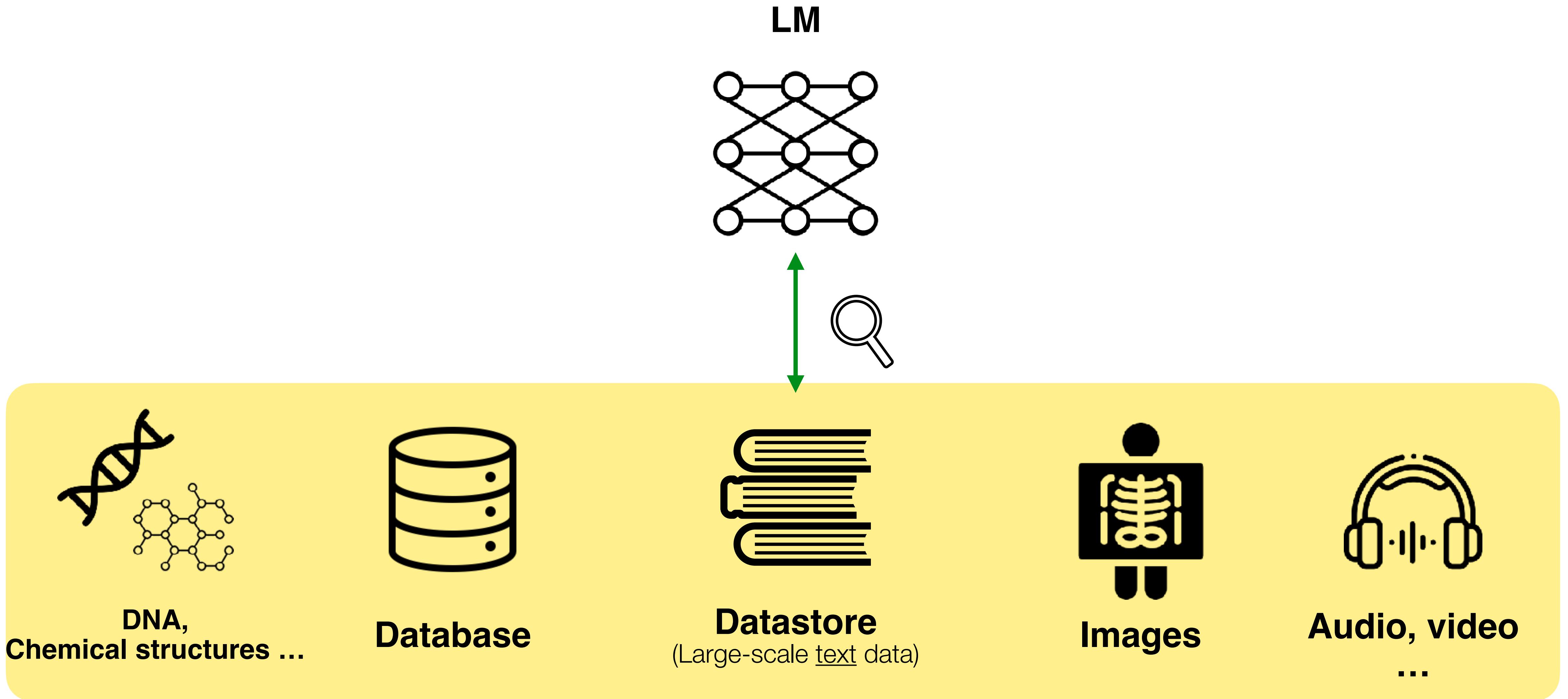
Future Work: Searching Over Heterogeneous Knowledge



Future Work: Searching Over Heterogeneous Knowledge



Future Work: Searching Over Heterogeneous Knowledge



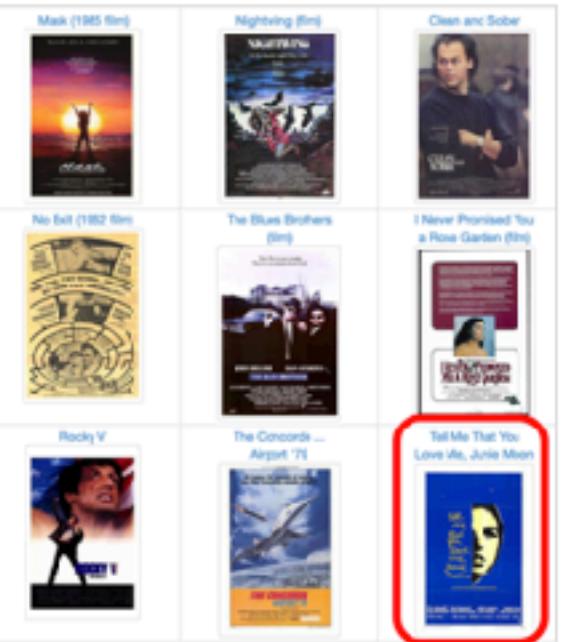
Future Work: Searching Over Heterogeneous Knowledge

Multimodal Context

Steal This Movie!	La licale
The film follows Hoffman's (D'Onofrio) relationship with his second wife Anita (Garofalo) and their "awakening" and subsequent conversion to an activist life. The title of the film is a play on Hoffman's 1970 counter-culture guidebook titled "Steal This Book".	La licale (internationally released as "The Teasers", "Under-graduate Girls", "Sophomore Swingers" and "Teasers") is a 1975 commedia sexy all'italiana directed by Michele Massimo Tarantini.
—	—
Sage Stallone	Pierino contro tutti
Stallone made his acting debut alongside his father in Rocky V (1990), the fifth installment of the Rocky franchise, playing Robert Balbo Jr., the onscreen son of his father's title character. He did not, however,	Pierino contro tutti (also known as "Desirable Teacher") is a 1981 comedy film directed by Marino Girolami. The main character of the film is Pierino, an
—	—
After that, he acted in lesser profile films.	I as a short lived subgenre of joke-films in which the plot basically consists of a series of jokes placed side by side.

Ben Piazza - Filmography

Year	Title	Role
1957	A Dangerous Age	David
1959	The Hanging Tree	Rune
1962	No Exit	Camarero
1970	Tell Me That You Love Me, Junie Moon	Jesse
1972	The Outside Man	Desk Clerk
---	---	---
1985	Mask	Mr. Simms
1988	Clean and Sober	Kramer
1990	Rocky V	Doctor
1991	Guilty by Suspicion	Darryl Zanuck



Q: Which B. Piazza title came earlier: the movie S. Stallone's son starred in or the movie with half of a lady's face on the poster?
A: Tell Me That You Love Me, Junie Moon

Talmor, ..., **Asai** et al., ICLR 2021. MultiModalQA: Complex Question Answering Over Text, Table and Images

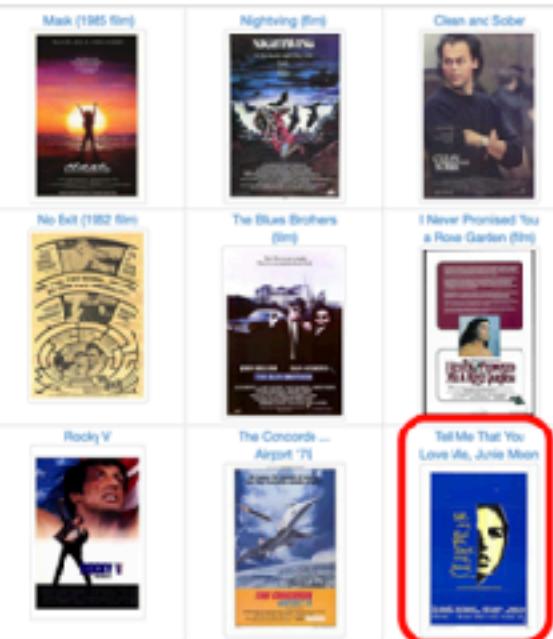
Future Work: Searching Over Heterogeneous Knowledge

Multimodal Context

Steal This Movie!	La licale
The film follows Hoffman's (D'Onofrio) relationship with his second wife Anita (Garofalo) and their "awakening" and subsequent conversion to an activist life. The title of the film is a play on Hoffman's 1970 counter-culture guidebook titled "Steal This Book".	La licale (internationally released as "The Teasers", "Under-graduate Girls", "Sophomore Swingers" and "Teasers") is a 1975 commedia sexy all'italiana directed by Michele Massimo Tarantini.
—	Guida. It was followed by "La licale nella classe dei ripetenti".
Sage Stallone	Pierino contro tutti
Stallone made his acting debut alongside his father in Rocky V (1990), the fifth installment of the Rocky franchise, playing Robert Balbo Jr., the onscreen son of his father's title character. He did not, however, —	Pierino contro tutti (also known as "Desirable Teacher") is a 1981 comedy film directed by Marino Girolami. The main character of the film is Pierino, an —
After that, he acted in lesser profile films.	I as a short lived subgenre of joke-films in which the plot basically consists of a series of jokes placed side by side.

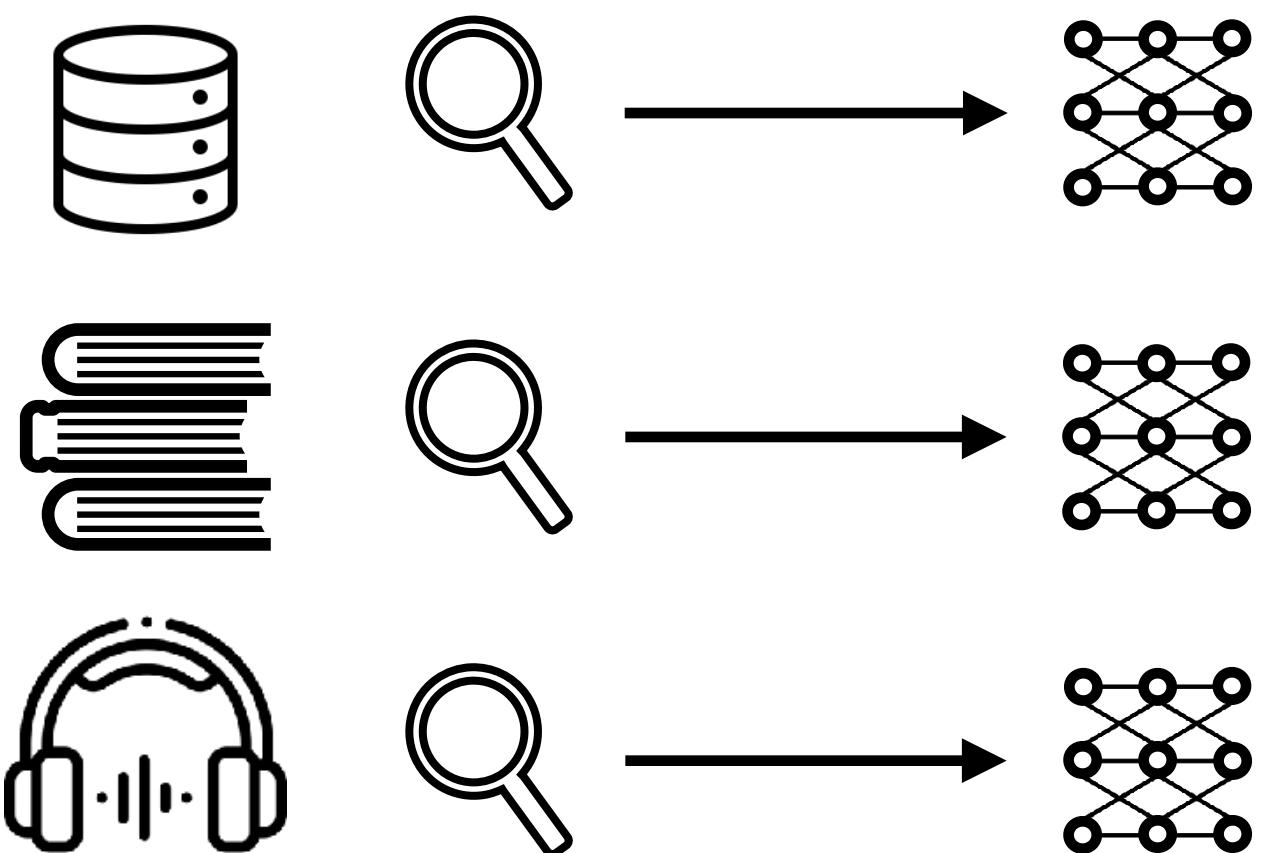
Ben Piazza - Filmography

Year	Title	Role
1957	A Dangerous Age	David
1959	The Hanging Tree	Rune
1962	No Exit	Camarero
1970	Tell Me That You Love Me, Junie Moon	Jesse
1972	The Outside Man	Desk Clerk
---	---	---
1985	Mask	Mr. Simms
1988	Clean and Sober	Kramer
1990	Rocky V	Doctor
1991	Guilty by Suspicion	Darryl Zanuck



Q: Which B. Piazza title came earlier: the movie S. Stallone's son starred in or the movie with half of a lady's face on the poster?
A: Tell Me That You Love Me, Junie Moon

Talmor, ..., Asai et al., ICLR 2021. MultiModalQA: Complex Question Answering Over Text, Table and Images



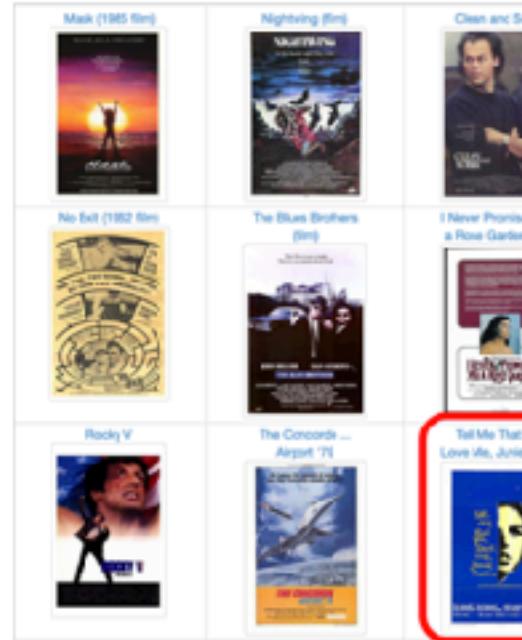
Future Work: Searching Over Heterogeneous Knowledge

Multimodal Context

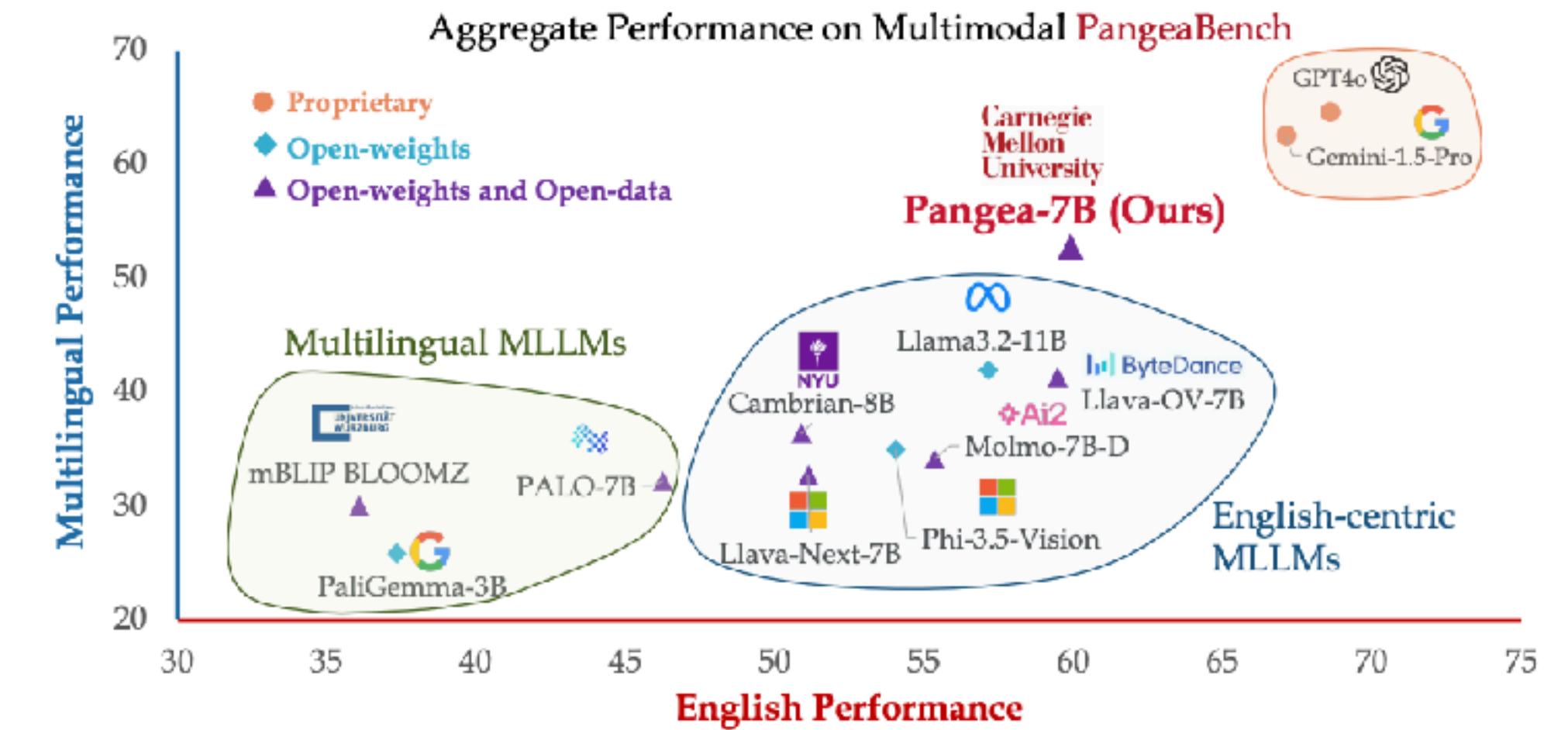
Steal This Movie!	La liciale
The film follows Hoffman's (D'Onofrio) relationship with his second wife Anita (Garofalo) and their "awakening" and subsequent conversion to an activist life. The title of the film is a play on Hoffman's 1970 counter-culture guidebook titled "Steal This Book".	La liciale (internationally released as The Teasers, "Under-graduate Girls", "Sophomore Swingers" and "Teasers") is a 1975 commedia sexy all'italiana directed by Michele Massimo Tarantini.
—	—
Sage Stallone	Pierino contro tutti
Stallone made his acting debut alongside his father in Rocky V (1990), the fifth installment of the Rocky franchise, playing Robert Balbo Jr., the onscreen son of his father's title character. He did not, however, —	Pierino contro tutti (also known as "Desirable Teacher") is a 1981 comedy film directed by Marino Girolami. The main character of the film is Pierino, an —
After that, he acted in lesser profile films.	I as a short lived subgenre of joke-films in which the plot basically consists of a series of jokes placed side by side.

Ben Piazza - Filmography

Year	Title	Role
1957	A Dangerous Age	David
1959	The Hanging Tree	Rune
1962	No Exit	Camarero
1970	Tell Me That You Love Me, Junie Moon	Jesse
1972	The Outside Man	Desk Clerk
...
1985	Mask	Mr. Simms
1988	Clean and Sober	Kramer
1990	Rocky V	Doctor
1991	Guilty by Suspicion	Darryl Zanuck

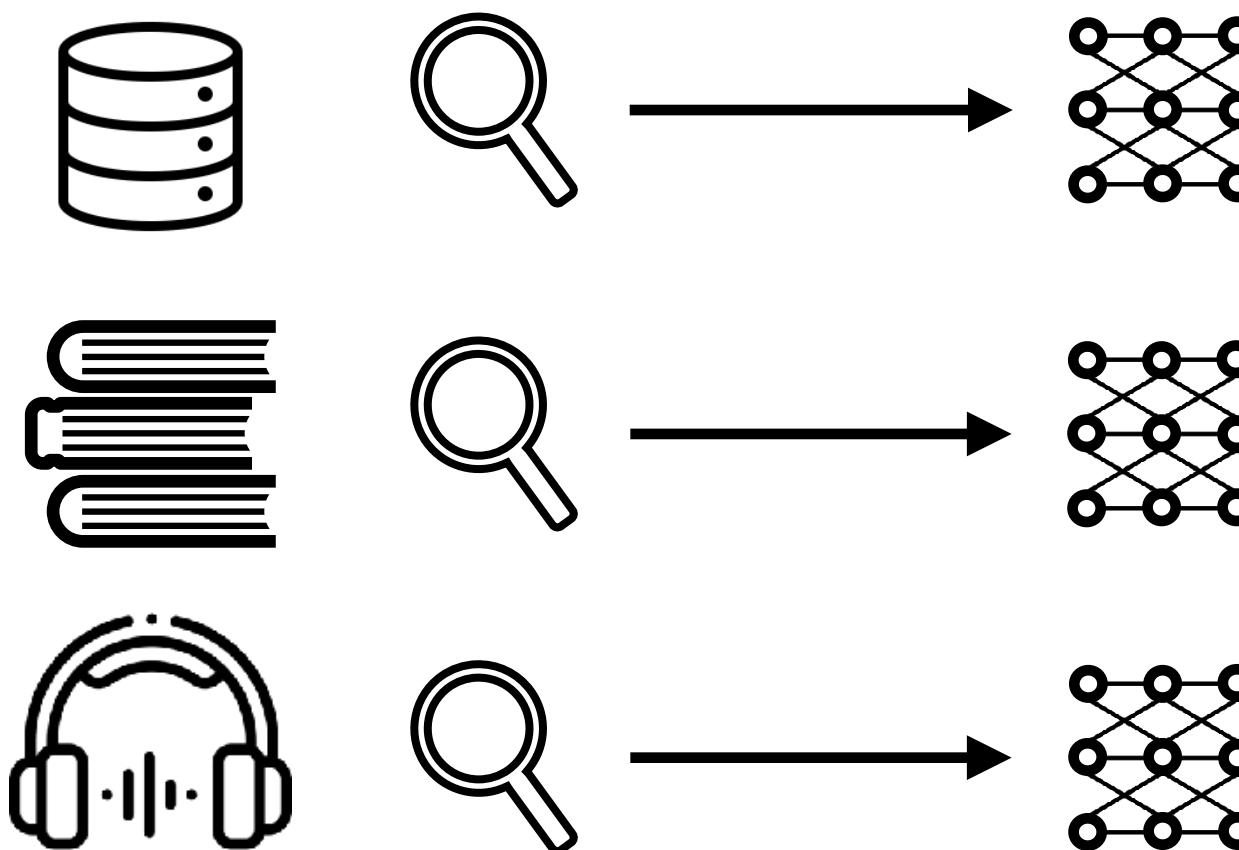


Q: Which B. Piazza title came earlier: the movie S. Stallone's son starred in or the movie with half of a lady's face on the poster?
A: Tell Me That You Love Me, Junie Moon



Talmor, ..., **Asai** et al., ICLR 2021. MultiModalQA: Complex Question Answering Over Text, Table and Images

Yue, Song, **Asai** et al., ICLR 2025. Pangea: A Fully Open Multilingual Multimodal LLM for 39 Languages.



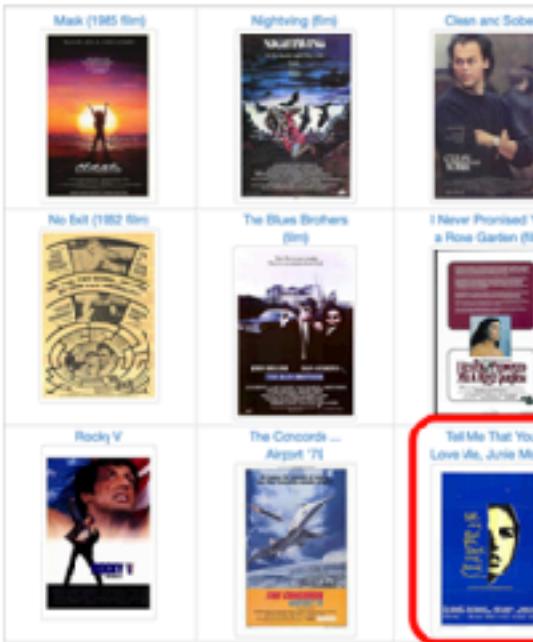
Future Work: Searching Over Heterogeneous Knowledge

Multimodal Context

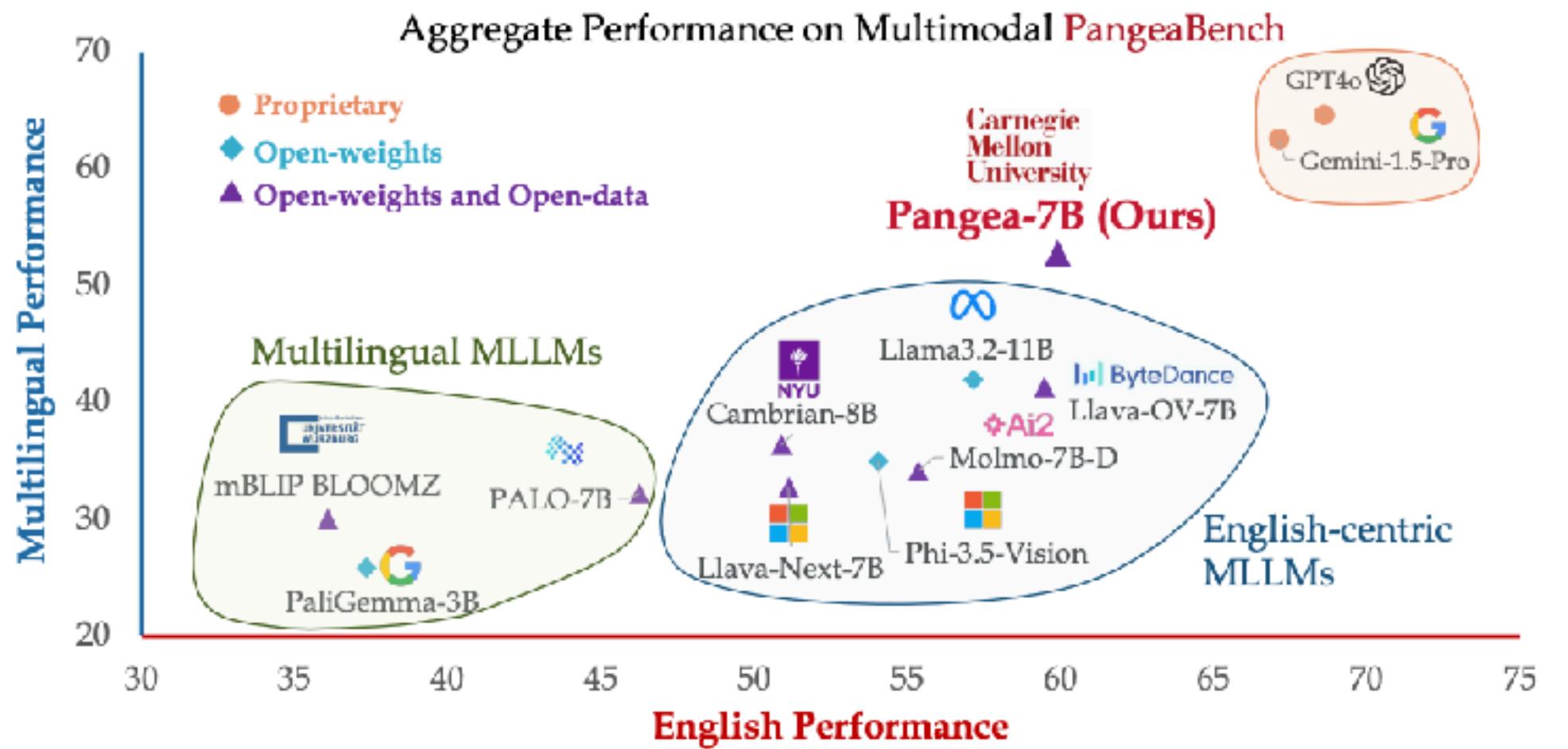
Steal This Movie!	
The film follows Hoffman's (D'Onofrio) relationship with his second wife Anita (Garofalo) and their "awakening" and subsequent conversion to an activist life. The title of the film is a play on Hoffman's 1970 counter-culture guidebook titled "Steal This Book".	
La liciale	
La liciale (internationally released as The Teasers, "Under-graduate Girls", "Sophomore Swingers" and "Teasers") is a 1975 commedia sexy all'italiana directed by Michele Massimo Tarantini.	
—	
Guida. It was followed by "La liciale nella classe dei ripetenti".	
Sage Stallone	
Stallone made his acting debut alongside his father in Rocky V (1990), the fifth installment of the Rocky franchise, playing Robert Balbo Jr., the onscreen son of his father's title character. He did not, however,	
—	
After that, he acted in lesser profile films.	
Pierino contro tutti	
Pierino contro tutti (also known as "Desirable Teacher") is a 1981 comedy film directed by Marino Girolami. The main character of the film is Pierino, an	
—	
I as a short lived subgenre of joke-films in which the plot basically consists of a series of jokes placed side by side.	

Ben Piazza - Filmography

Year	Title	Role
1957	A Dangerous Age	David
1959	The Hanging Tree	Rune
1962	No Exit	Camarero
1970	Tell Me That You Love Me, Junie Moon	Jesse
1972	The Outside Man	Desk Clerk
...
1985	Mask	Mr. Simms
1988	Clean and Sober	Kramer
1990	Rocky V	Doctor
1991	Guilty by Suspicion	Darryl Zanuck



Q: Which B. Piazza title came earlier: the movie S. Stallone's son starred in or the movie with half of a lady's face on the poster?
A: Tell Me That You Love Me, Junie Moon



Talmor, ..., **Asai** et al., ICLR 2021. MultiModalQA: Complex Question Answering Over Text, Table and Images

Yue, Song, **Asai** et al., ICLR 2025. Pangea: A Fully Open Multilingual Multimodal LLM for 39 Languages.

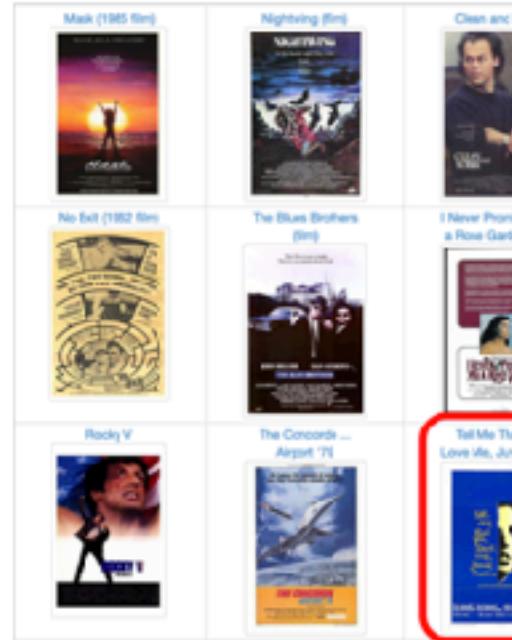
Future Work: Searching Over Heterogeneous Knowledge

Multimodal Context

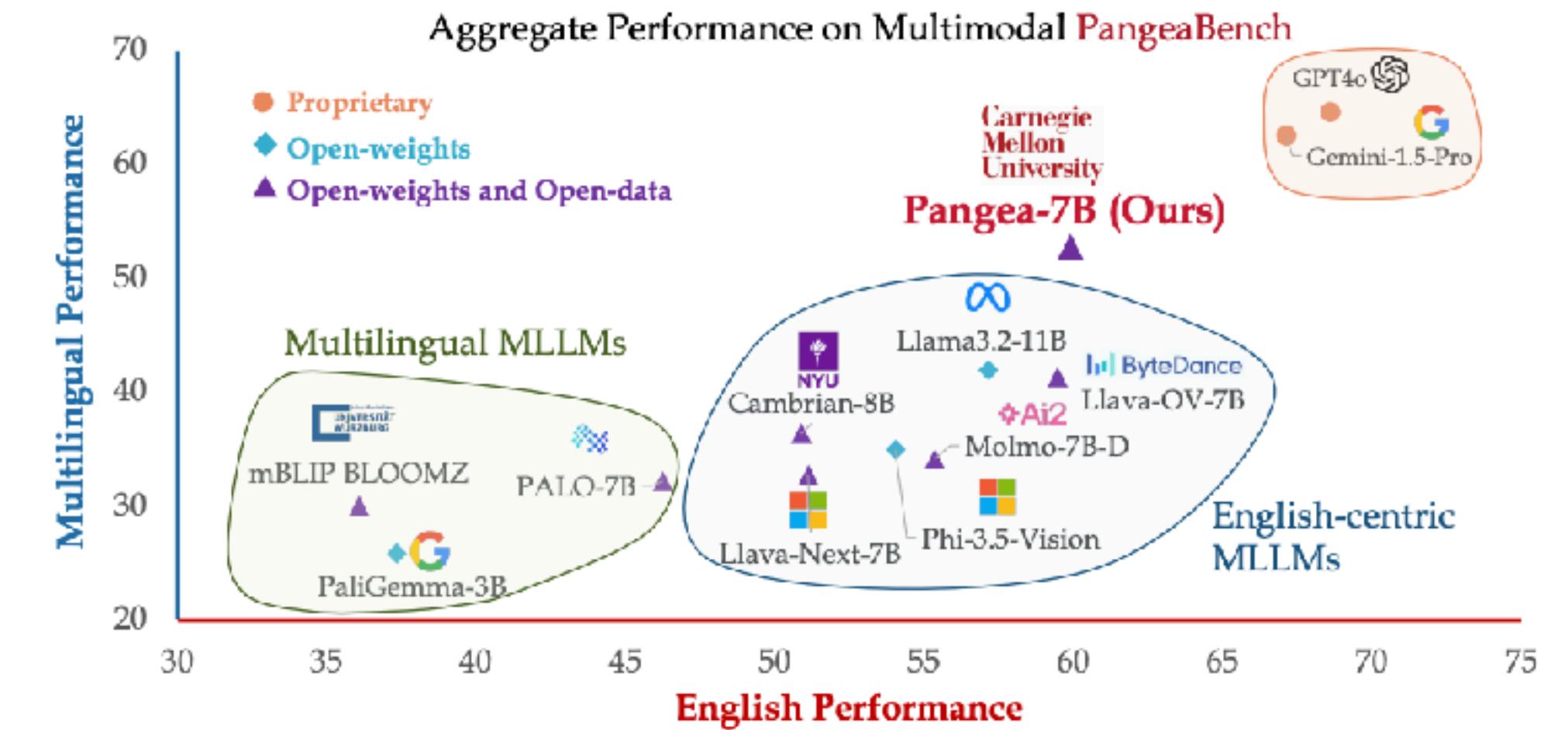
Steal This Movie!	La liciale
The film follows Hoffman's (D'Onofrio) relationship with his second wife Anita (Garofalo) and their "awakening" and subsequent conversion to an activist life. The title of the film is a play on Hoffman's 1970 counter-culture guidebook titled "Steal This Book".	La liciale (internationally released as The Teasers, "Under-graduate Girls", "Sophomore Swingers" and "Teasers") is a 1975 commedia sexy all'italiana directed by Michele Massimo Tarantini.
—	Guida. It was followed by "La liciale nella classe dei ripetenti".
Sage Stallone	Pierino contro tutti
Stallone made his acting debut alongside his father in Rocky V (1990), the fifth installment of the Rocky franchise, playing Robert Balbo Jr., the onscreen son of his father's title character. He did not, however, —	Pierino contro tutti (also known as "Desirable Teacher") is a 1981 comedy film directed by Marino Girolami. The main character of the film is Pierino, an —
After that, he acted in lesser profile films.	I as a short lived subgenre of joke-films in which the plot basically consists of a series of jokes placed side by side.

Ben Piazza - Filmography

Year	Title	Role
1957	A Dangerous Age	David
1959	The Hanging Tree	Rune
1962	No Exit	Camarero
1970	Tell Me That You Love Me, Junie Moon	Jesse
1972	The Outside Man	Desk Clerk
...
1985	Mask	Mr. Simms
1988	Clean and Sober	Kramer
1990	Rocky V	Doctor
1991	Guilty by Suspicion	Darryl Zanuck

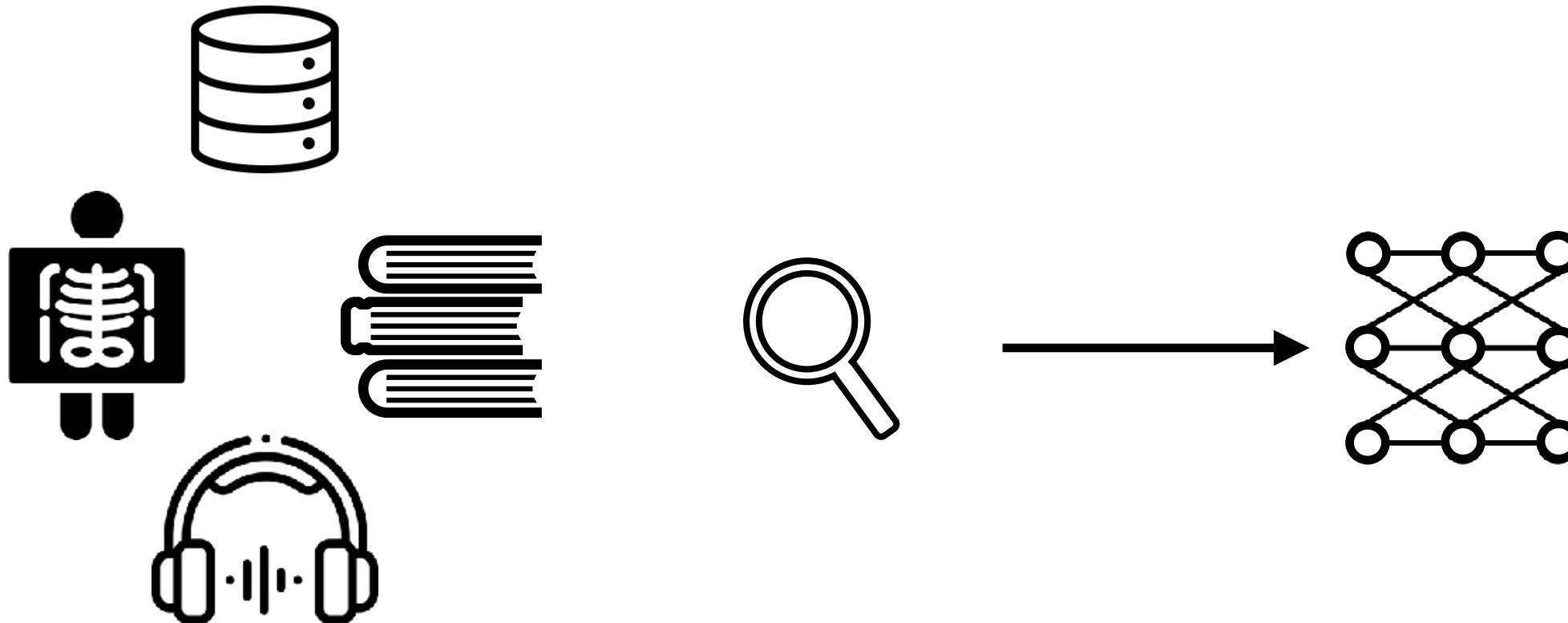


Q: Which B. Piazza title came earlier: the movie S. Stallone's son starred in or the movie with half of a lady's face on the poster?
A: Tell Me That You Love Me, Junie Moon



Talmor, ..., **Asai** et al., ICLR 2021. MultiModalQA: Complex Question Answering Over Text, Table and Images

Yue, Song, **Asai** et al., ICLR 2025. Pangea: A Fully Open Multilingual Multimodal LLM for 39 Languages.



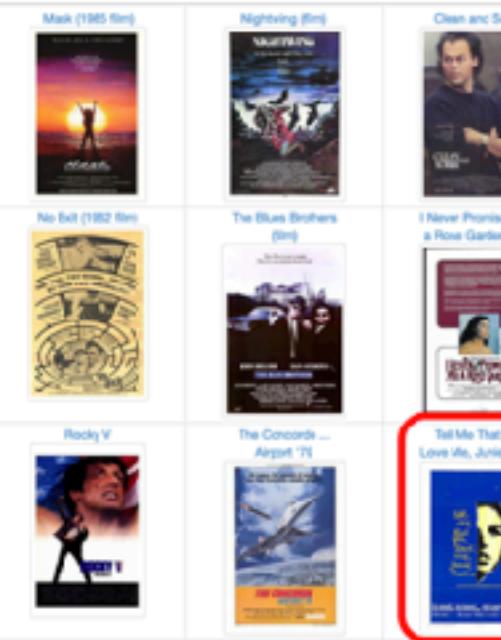
Future Work: Searching Over Heterogeneous Knowledge

Multimodal Context

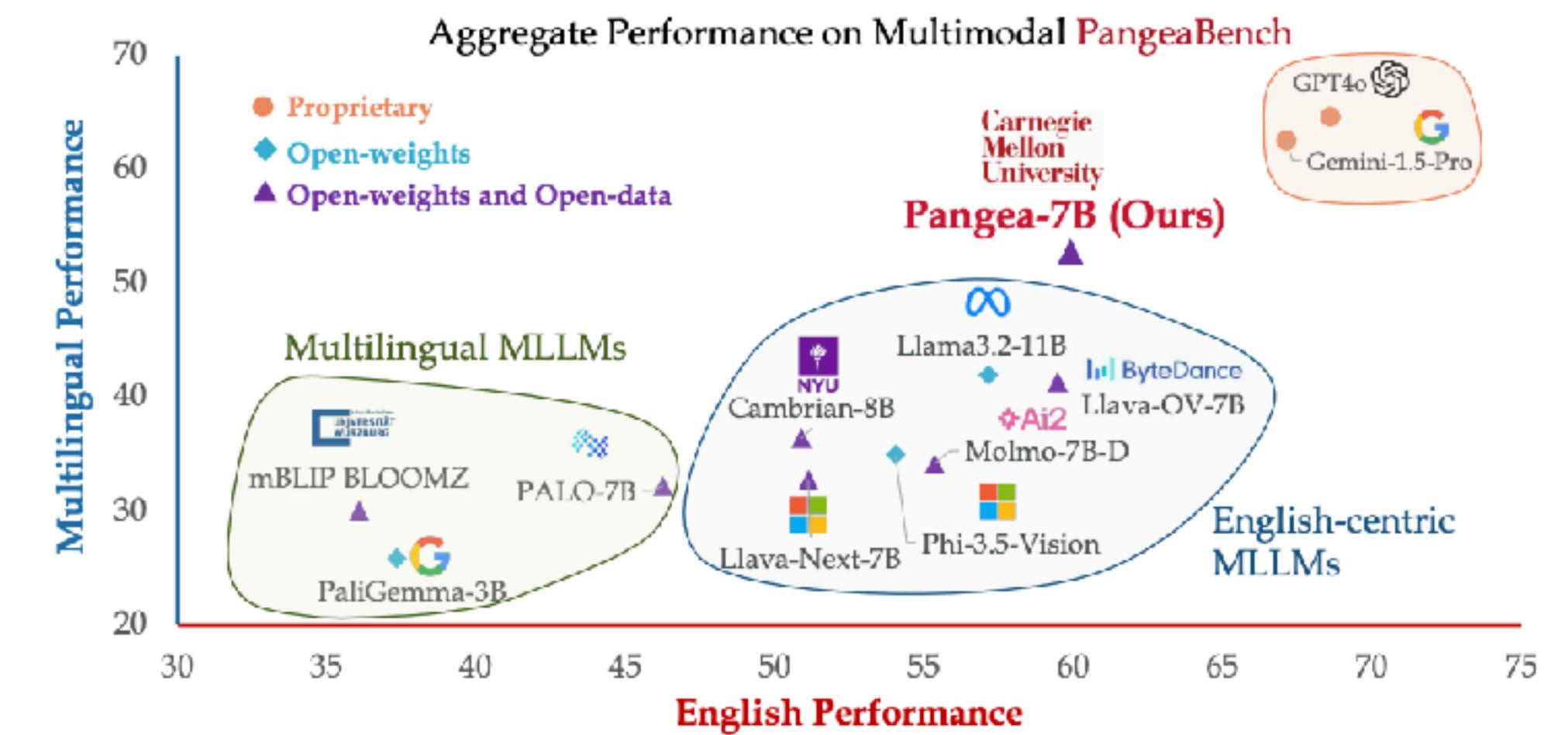
Steal This Movie!	La liciale
The film follows Hoffman's (D'Onofrio) relationship with his second wife Anita (Garofalo) and their "awakening" and subsequent conversion to an activist life. The title of the film is a play on Hoffman's 1970 counter-culture guidebook titled "Steal This Book".	La liciale (internationally released as The Teasers, "Under-graduate Girls", "Sophomore Swingers" and "Teasers") is a 1975 commedia sexy all'italiana directed by Michele Massimo Tarantini.
—	—
Sage Stallone	Tell Me That You Love Me, Junie Moon
Stallone made his acting debut alongside his father in Rocky V (1990), the fifth installment of the Rocky franchise, playing Robert Balbo Jr., the onscreen son of his father's title character. He did not, however, —	Pierino contro tutti (also known as "Desirable Teacher") is a 1981 comedy film directed by Marino Girolami. The main character of the film is Pierino, an —
After that, he acted in lesser profile films.	I as a short lived subgenre of joke-films in which the plot basically consists of a series of jokes placed side by side.

Ben Piazza - Filmography

Year	Title	Role
1957	A Dangerous Age	David
1959	The Hanging Tree	Rune
1962	No Exit	Camarero
1970	Tell Me That You Love Me, Junie Moon	Jesse
1972	The Outside Man	Desk Clerk
...
1985	Mask	Mr. Simms
1988	Clean and Sober	Kramer
1990	Rocky V	Doctor
1991	Guilty by Suspicion	Darryl Zanuck

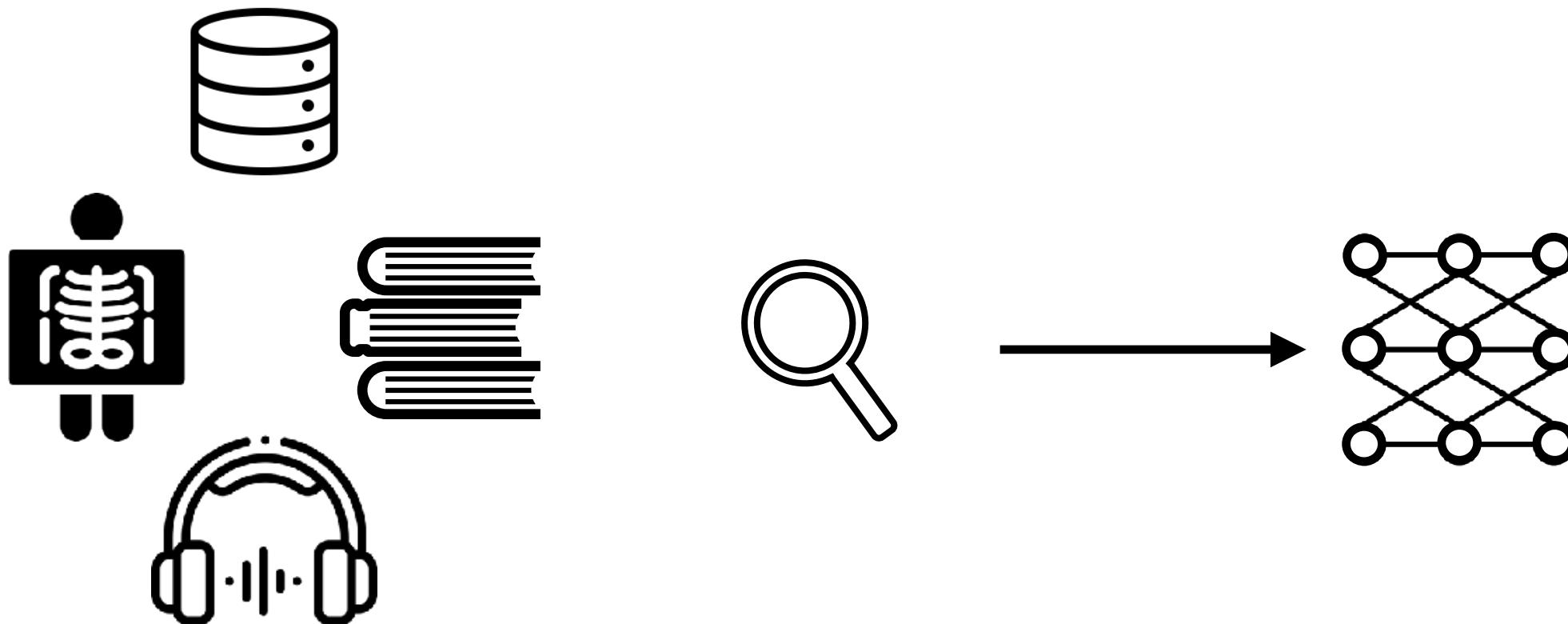


Q: Which B. Piazza title came earlier: the movie S. Stallone's son starred in or the movie with half of a lady's face on the poster?
A: Tell Me That You Love Me, Junie Moon



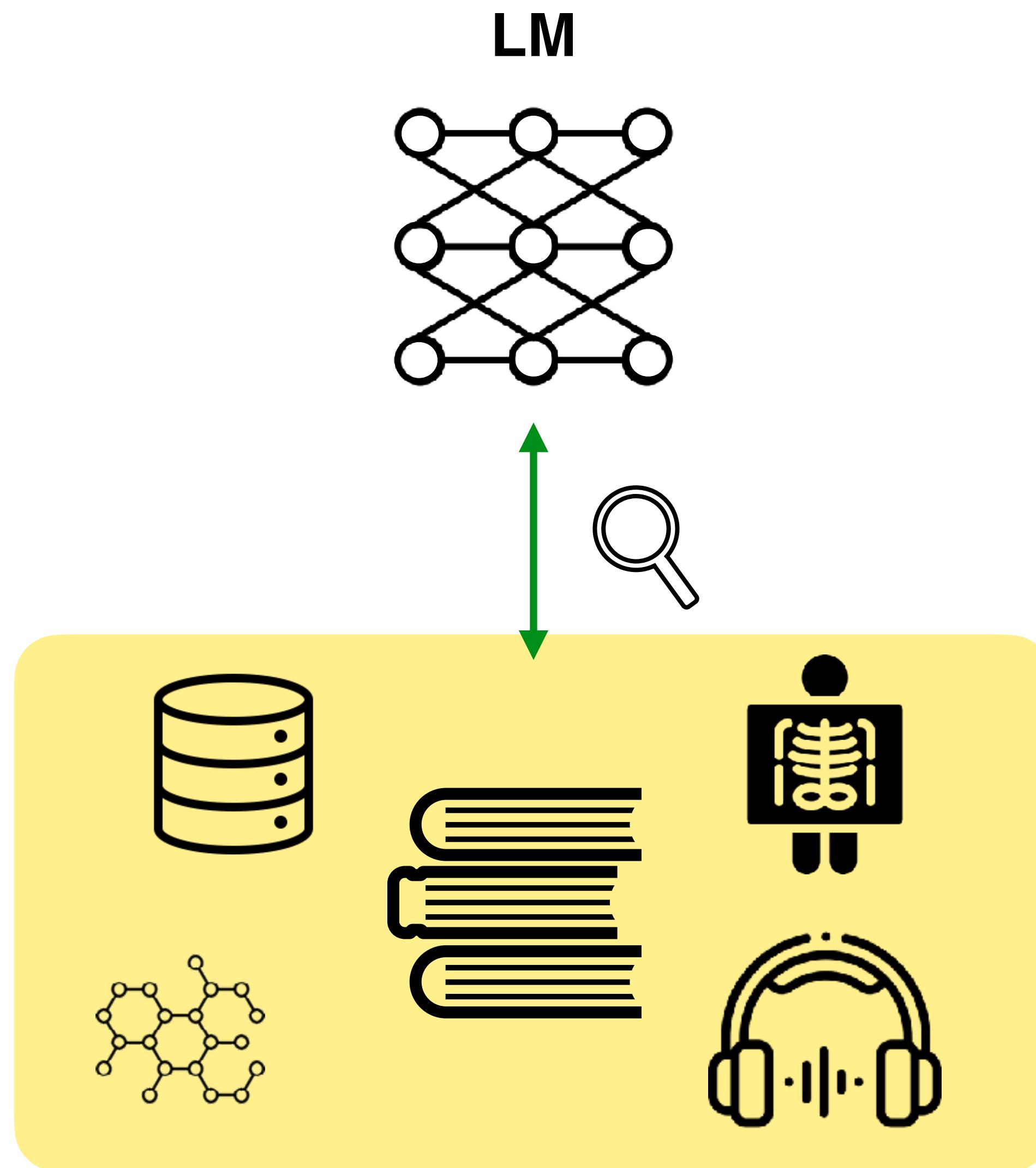
Talmor, ..., **Asai** et al., ICLR 2021. MultiModalQA: Complex Question Answering Over Text, Table and Images

Yue, Song, **Asai** et al., ICLR 2025. Pangea: A Fully Open Multilingual Multimodal LLM for 39 Languages.

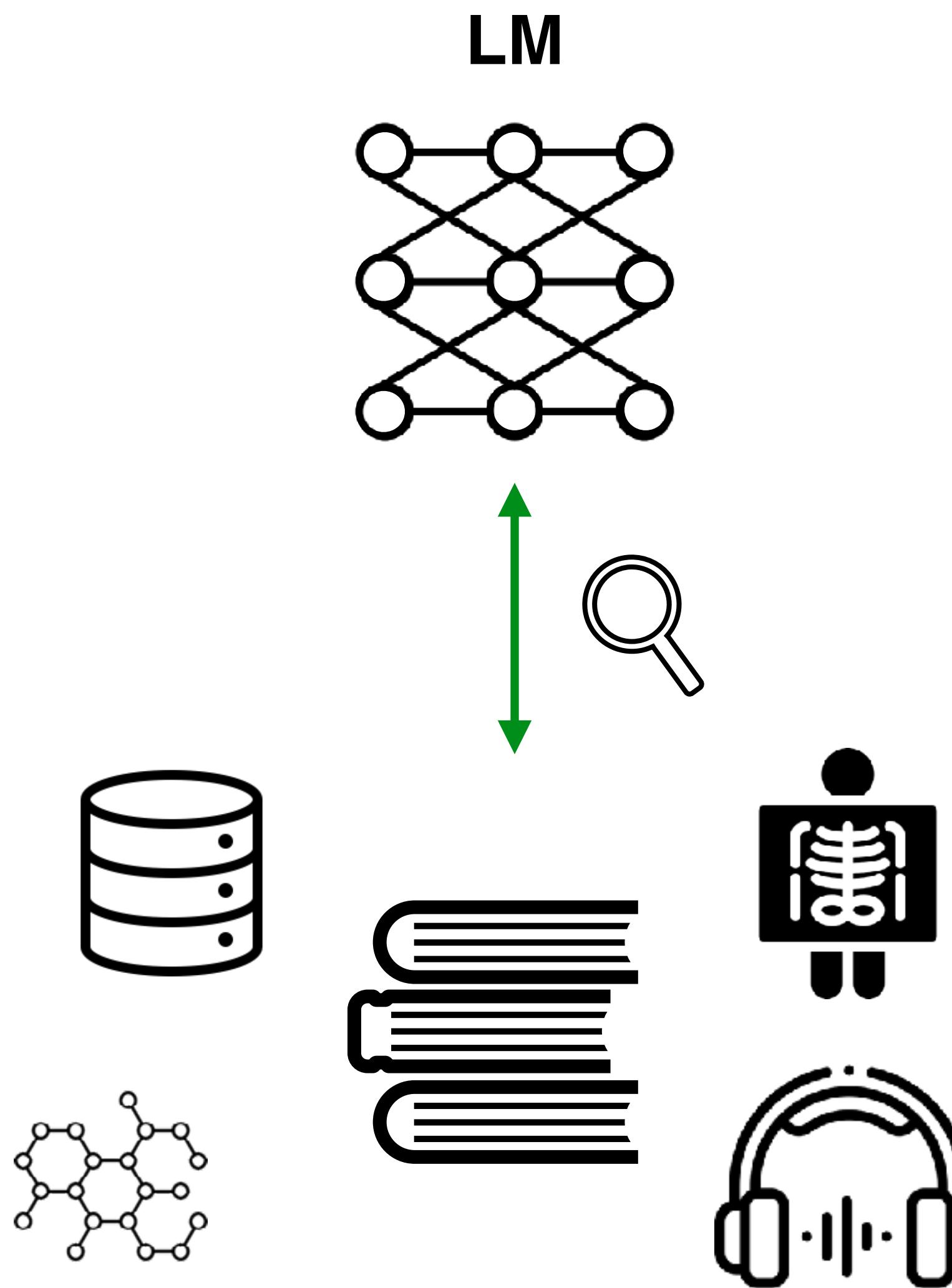


Unified multimodal retrievers and LMs
for heterogeneous knowledge

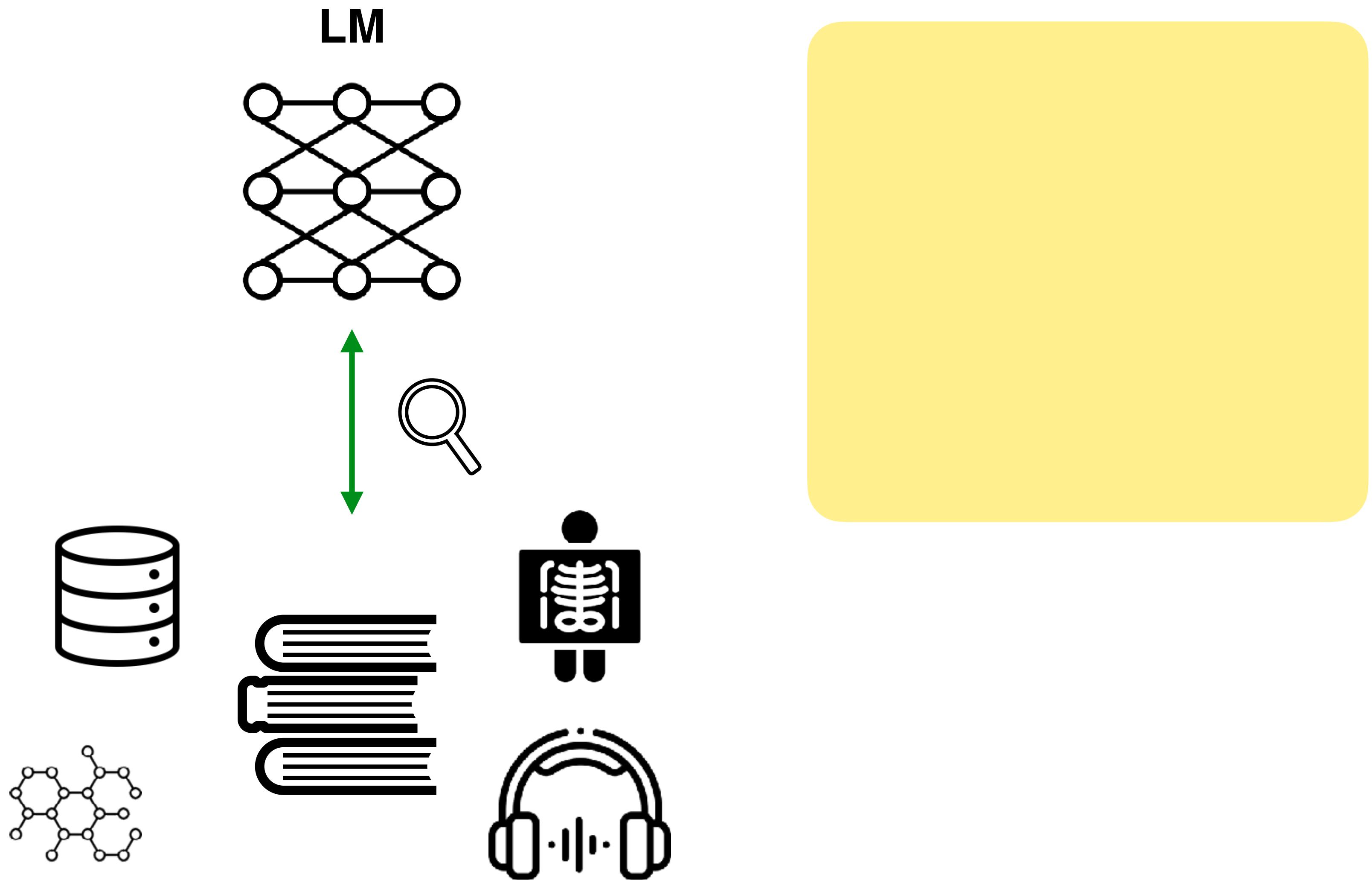
Future Work: Towards Module Orchestration



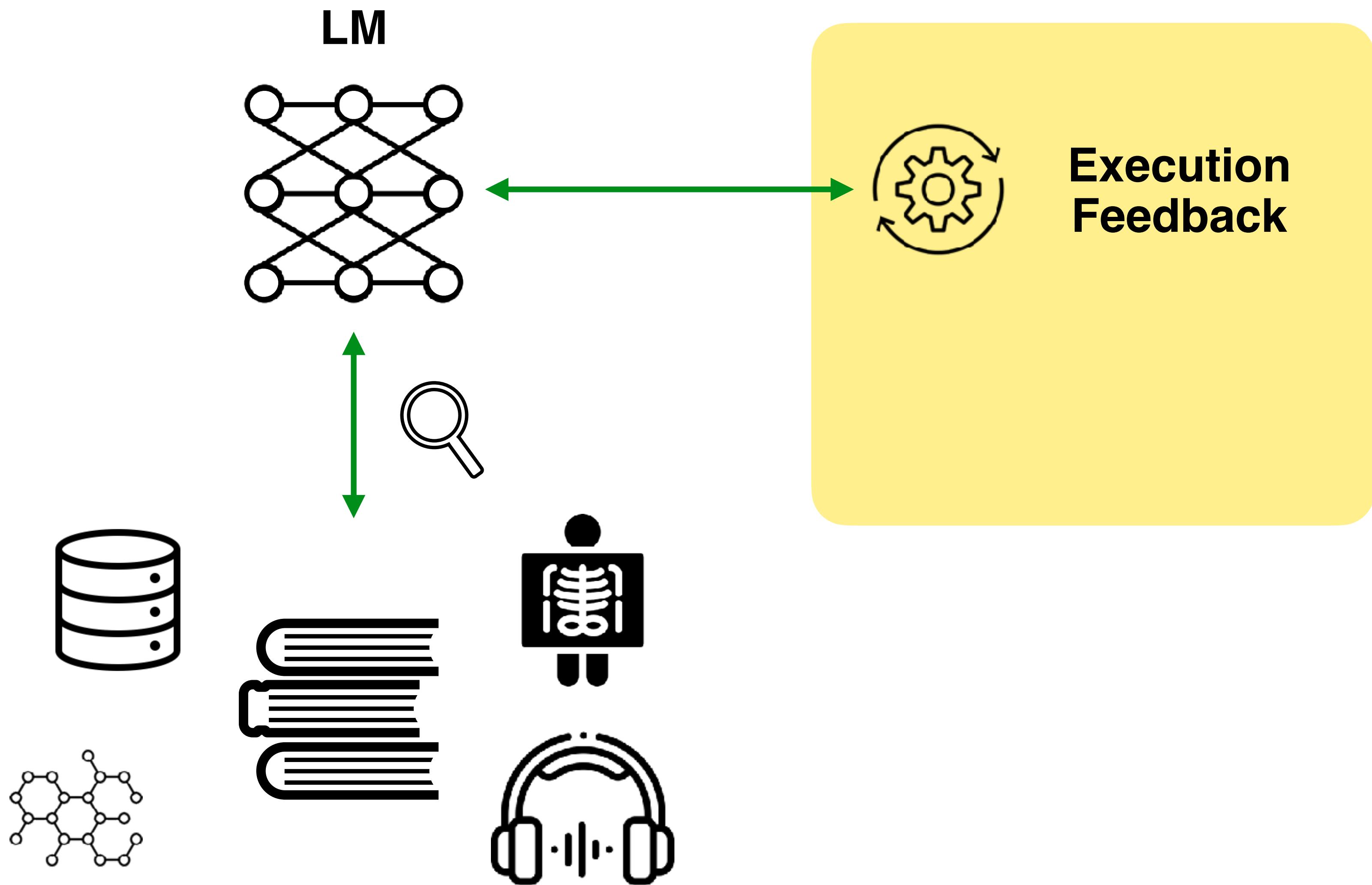
Future Work: Towards Module Orchestration



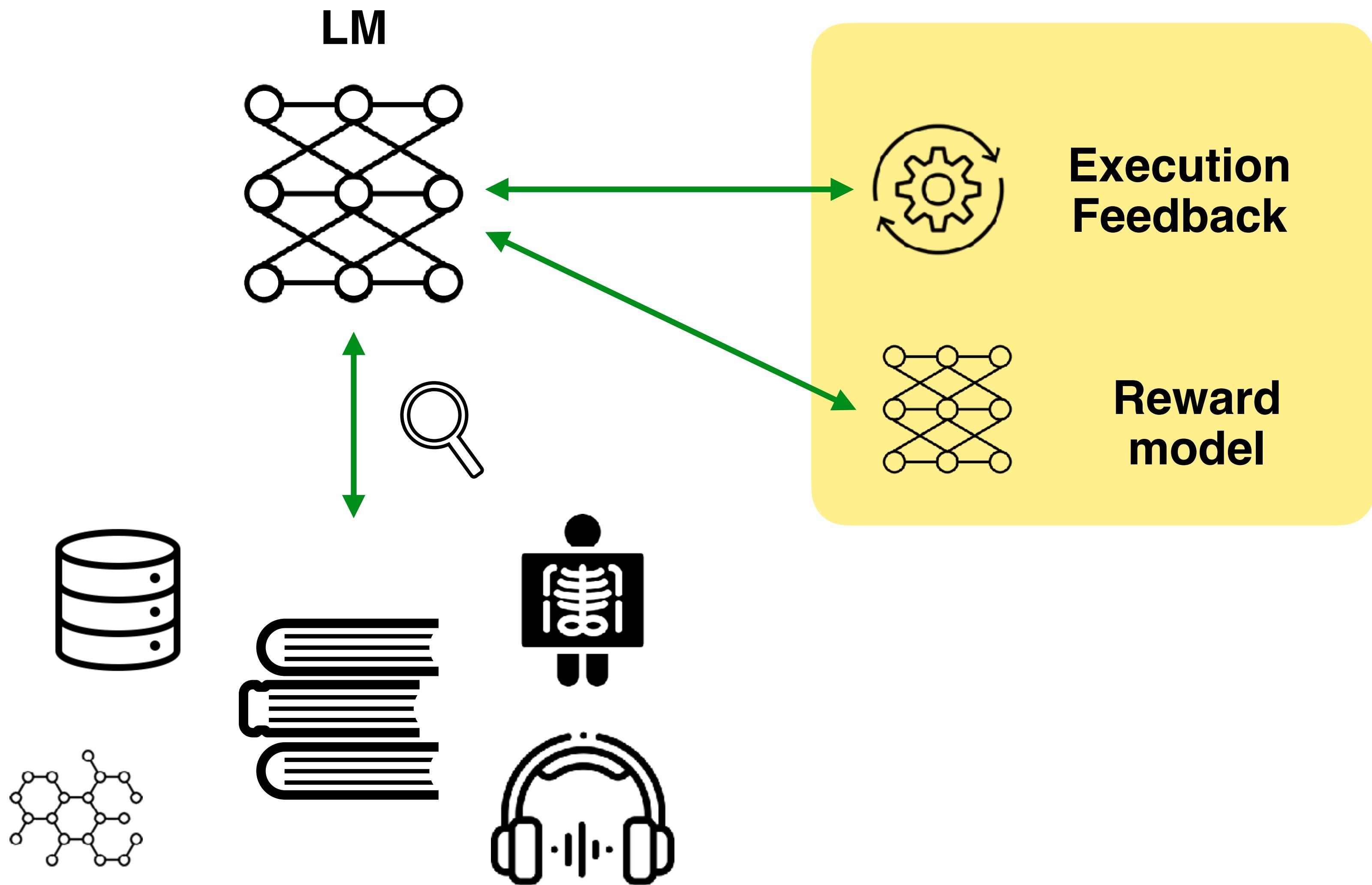
Future Work: Towards Module Orchestration



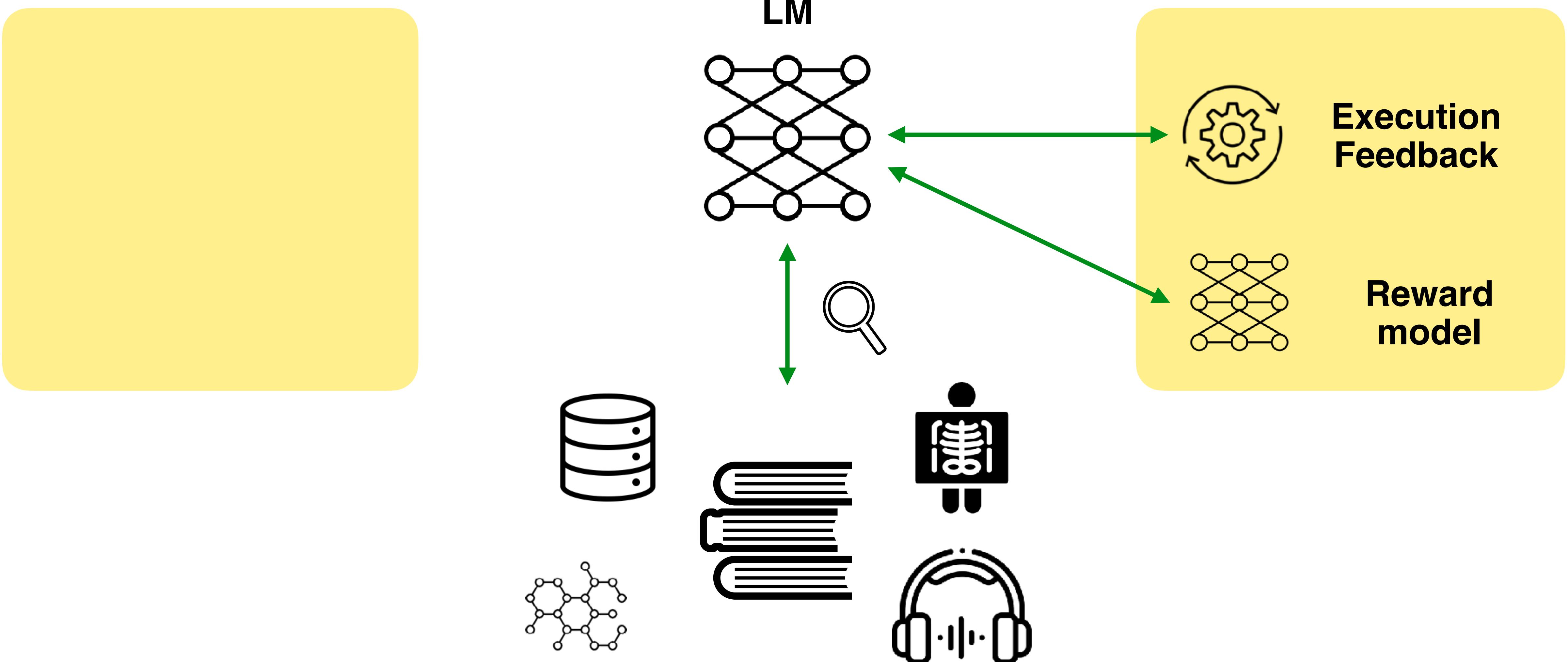
Future Work: Towards Module Orchestration



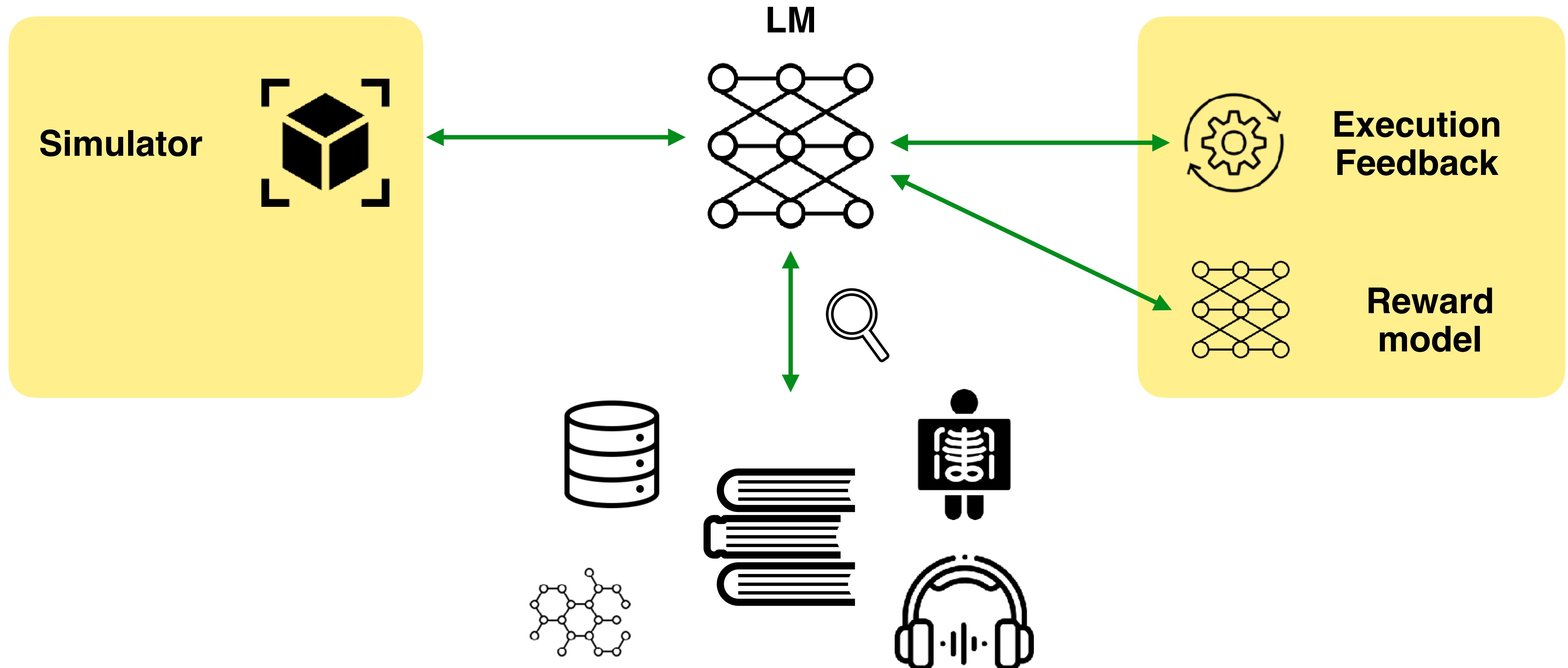
Future Work: Towards Module Orchestration



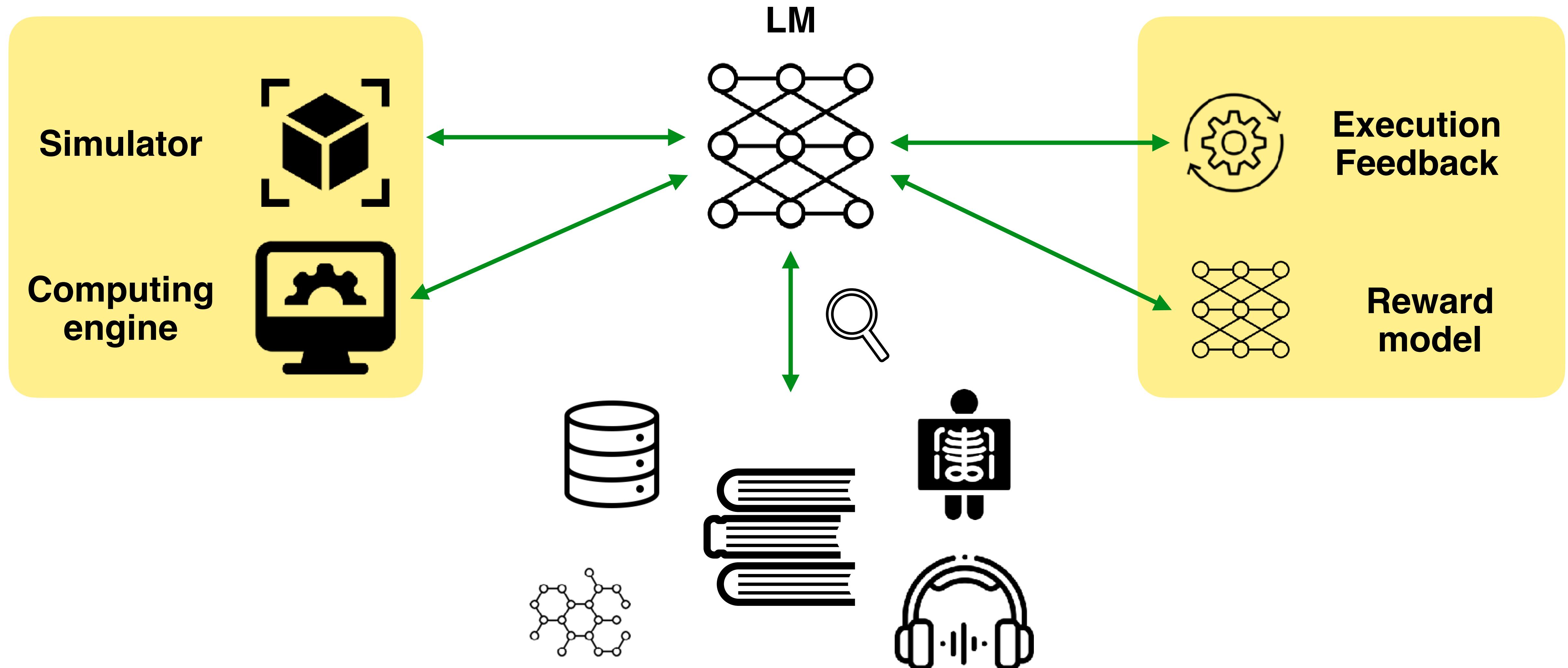
Future Work: Towards Module Orchestration



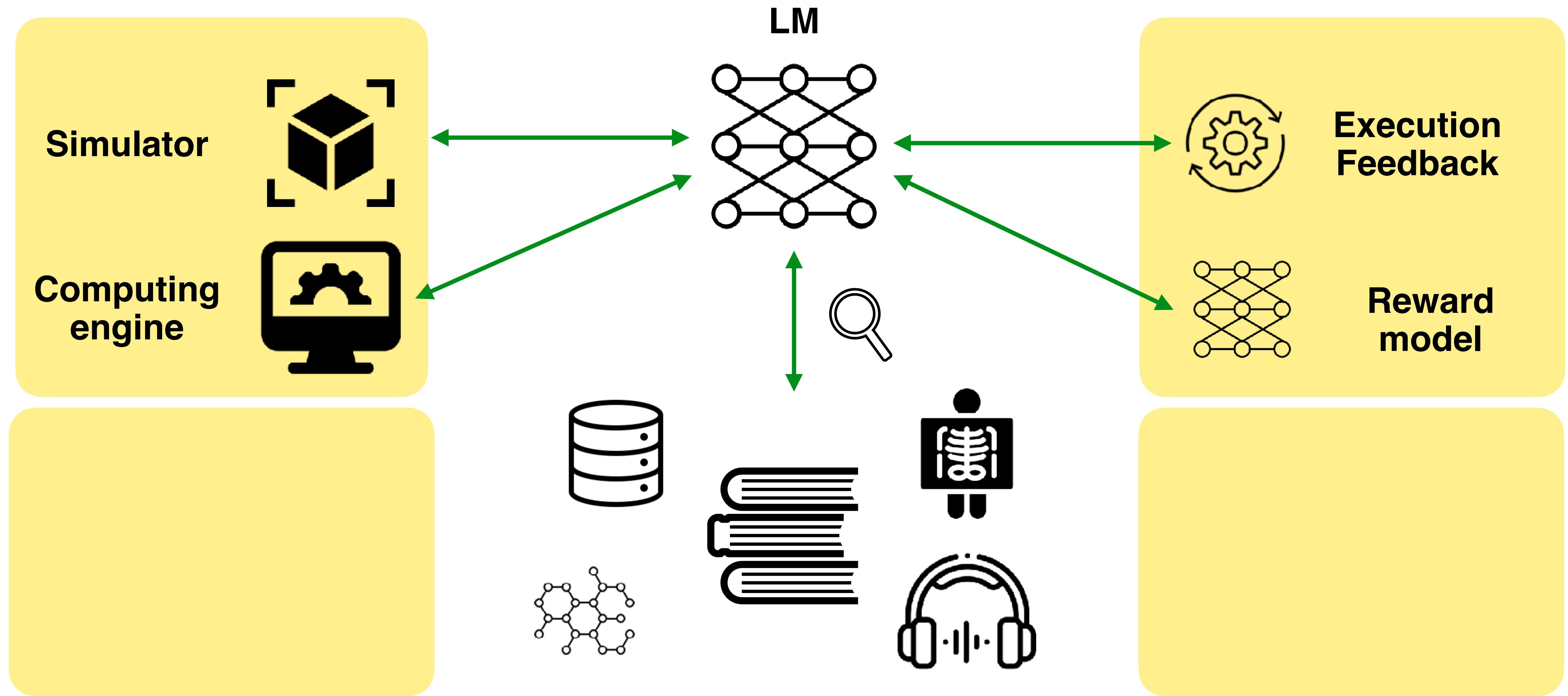
Future Work: Towards Module Orchestration



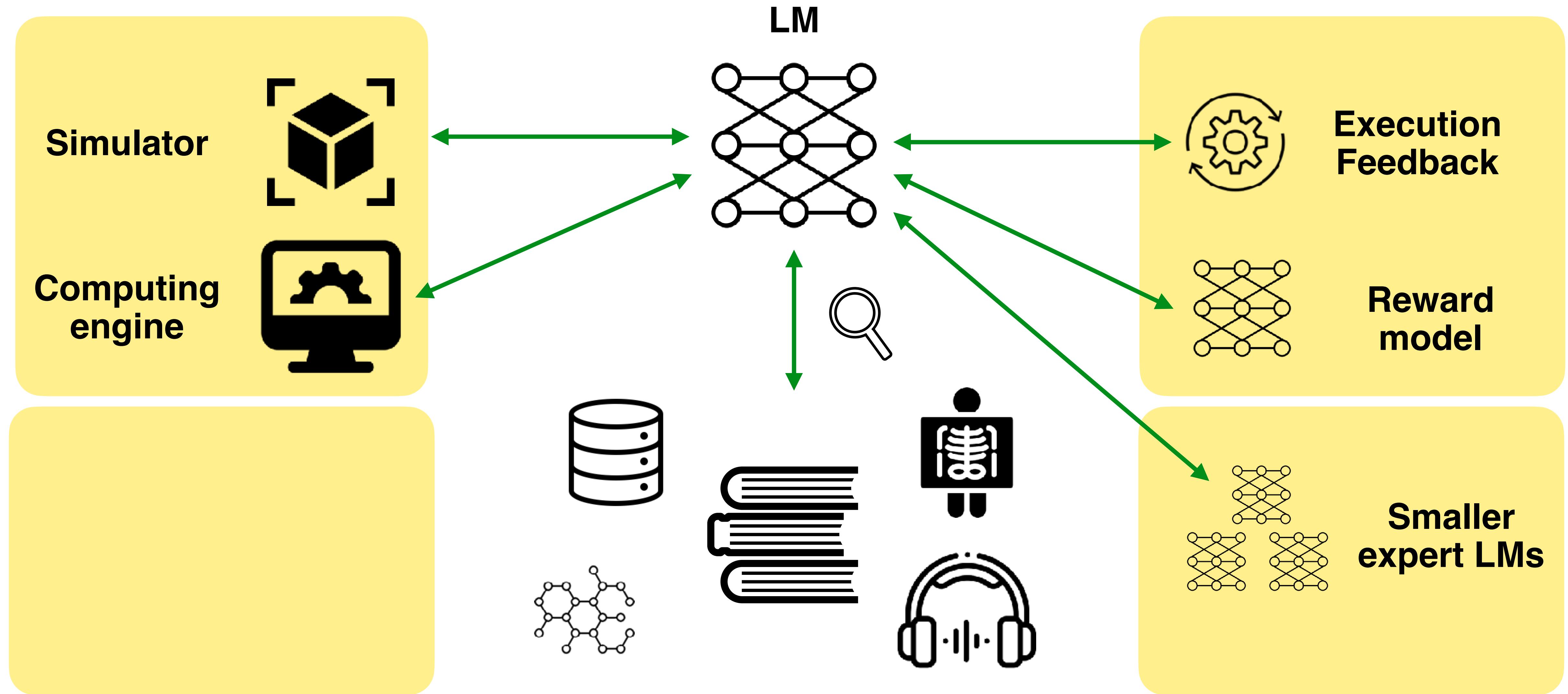
Future Work: Towards Module Orchestration



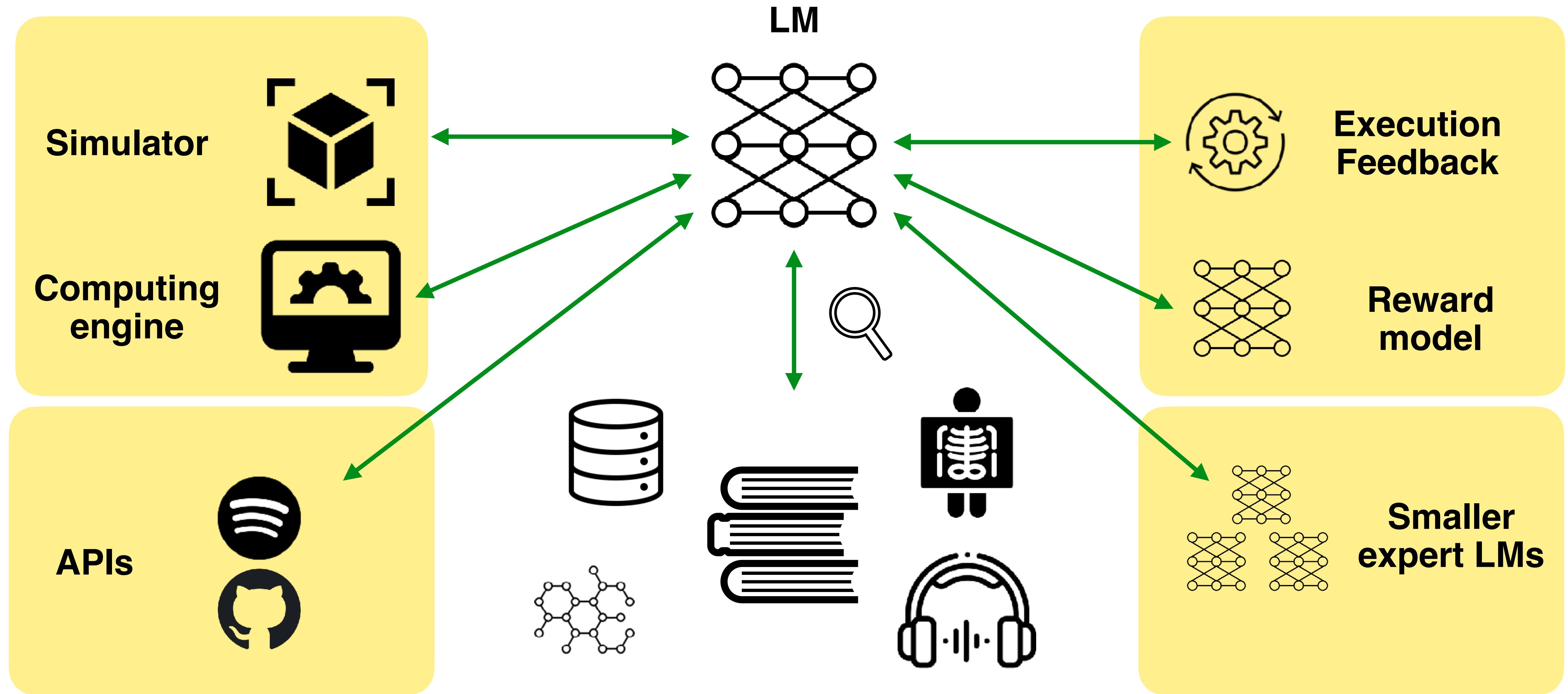
Future Work: Towards Module Orchestration



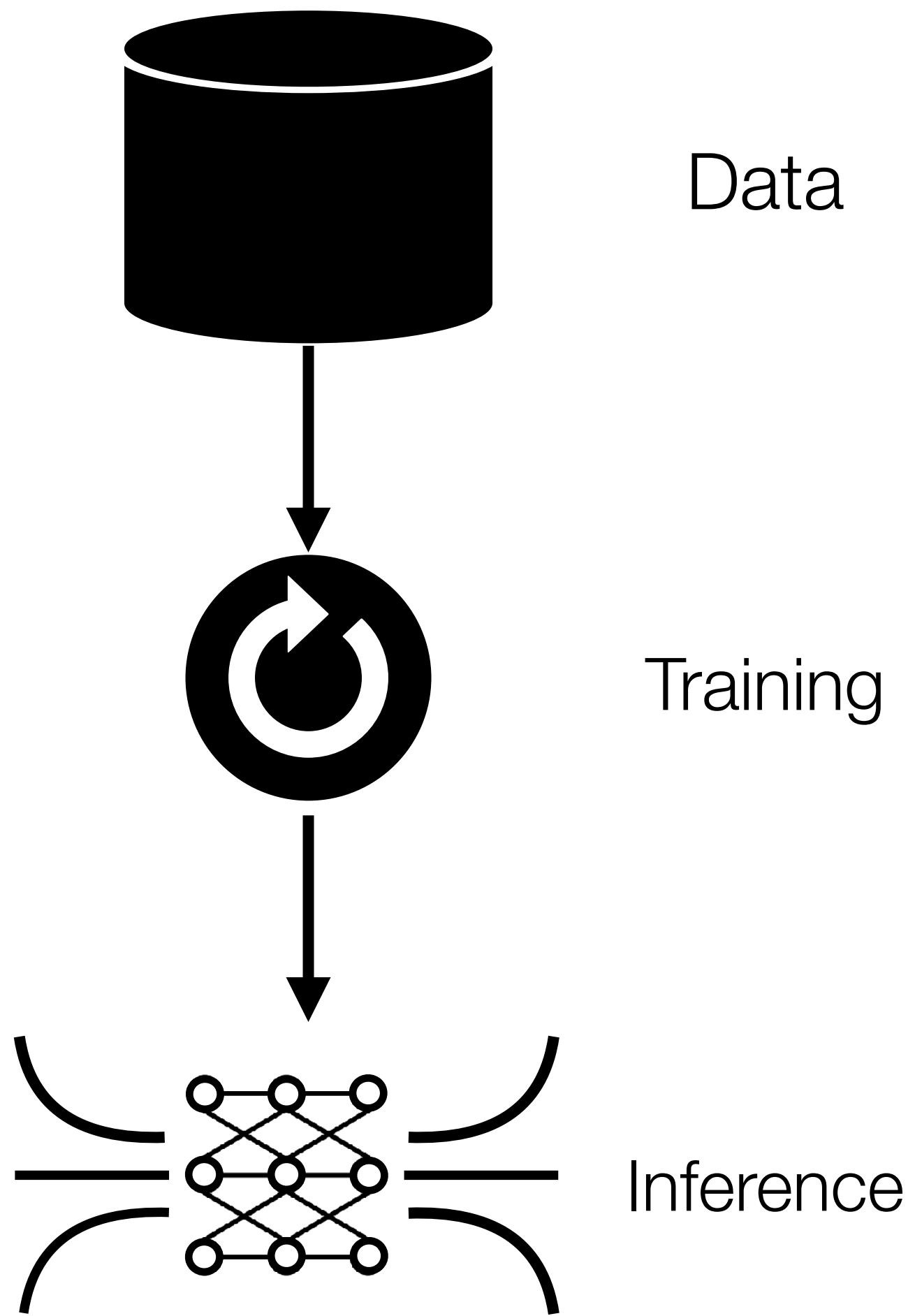
Future Work: Towards Module Orchestration



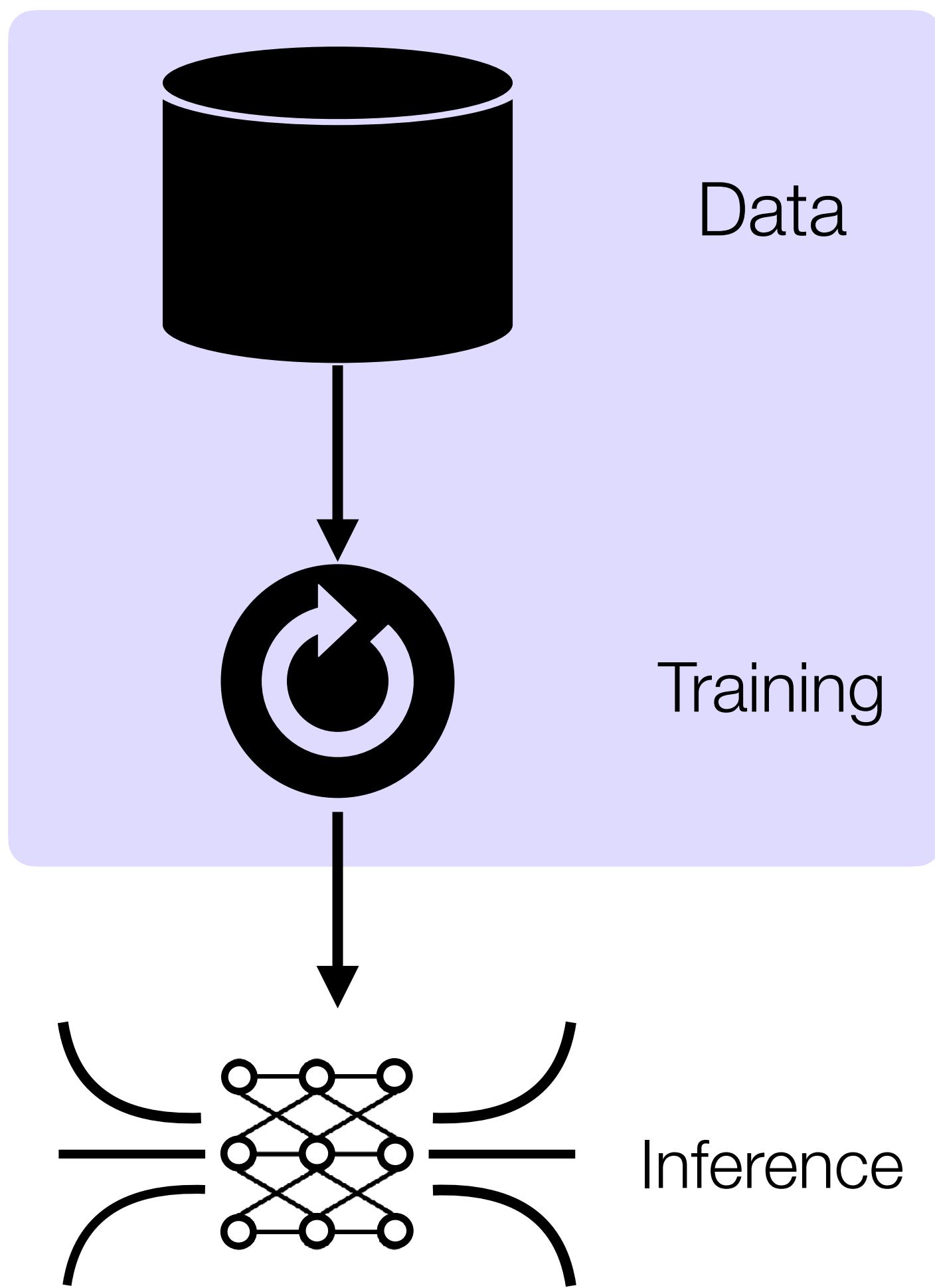
Future Work: Towards Module Orchestration



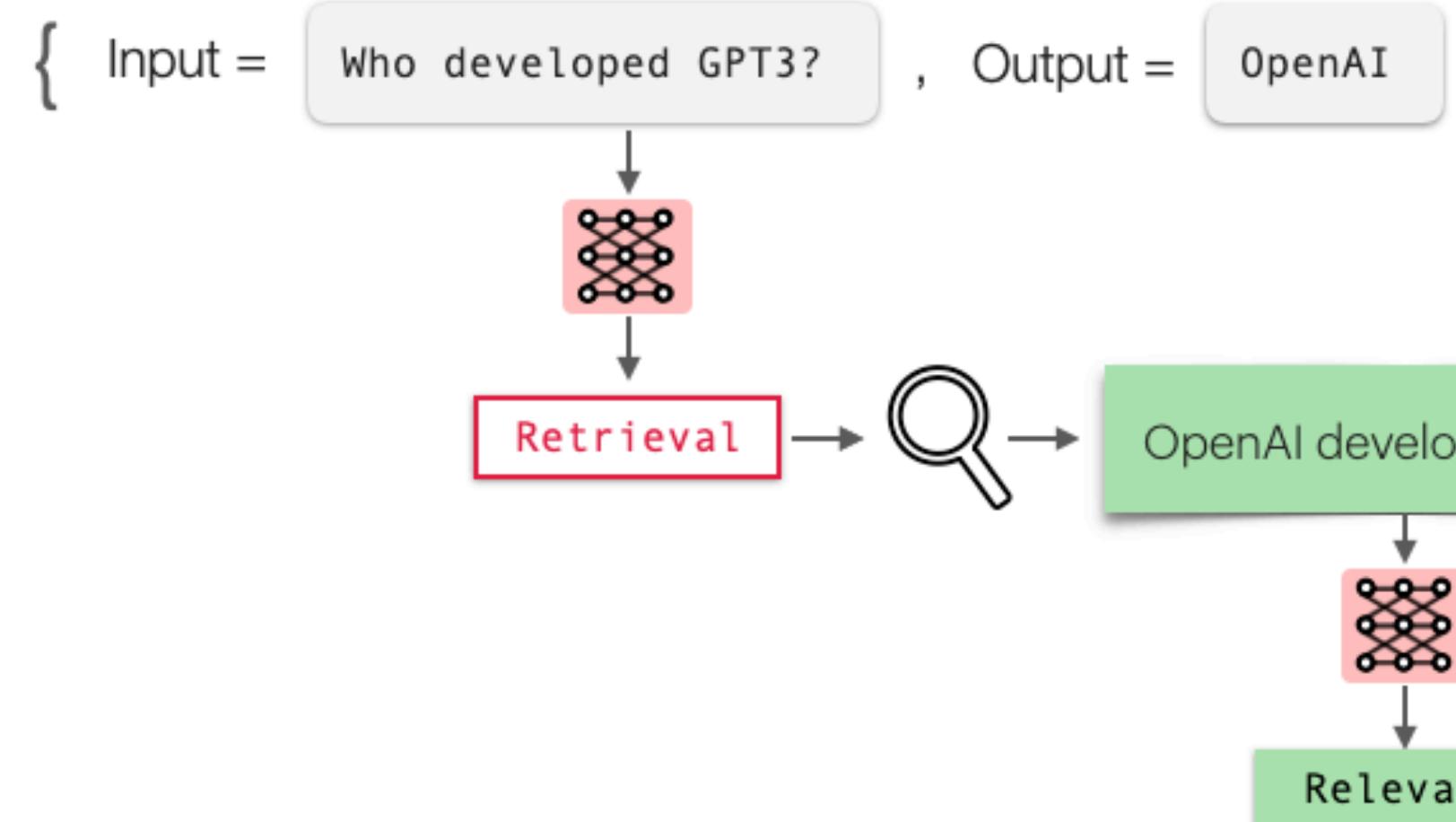
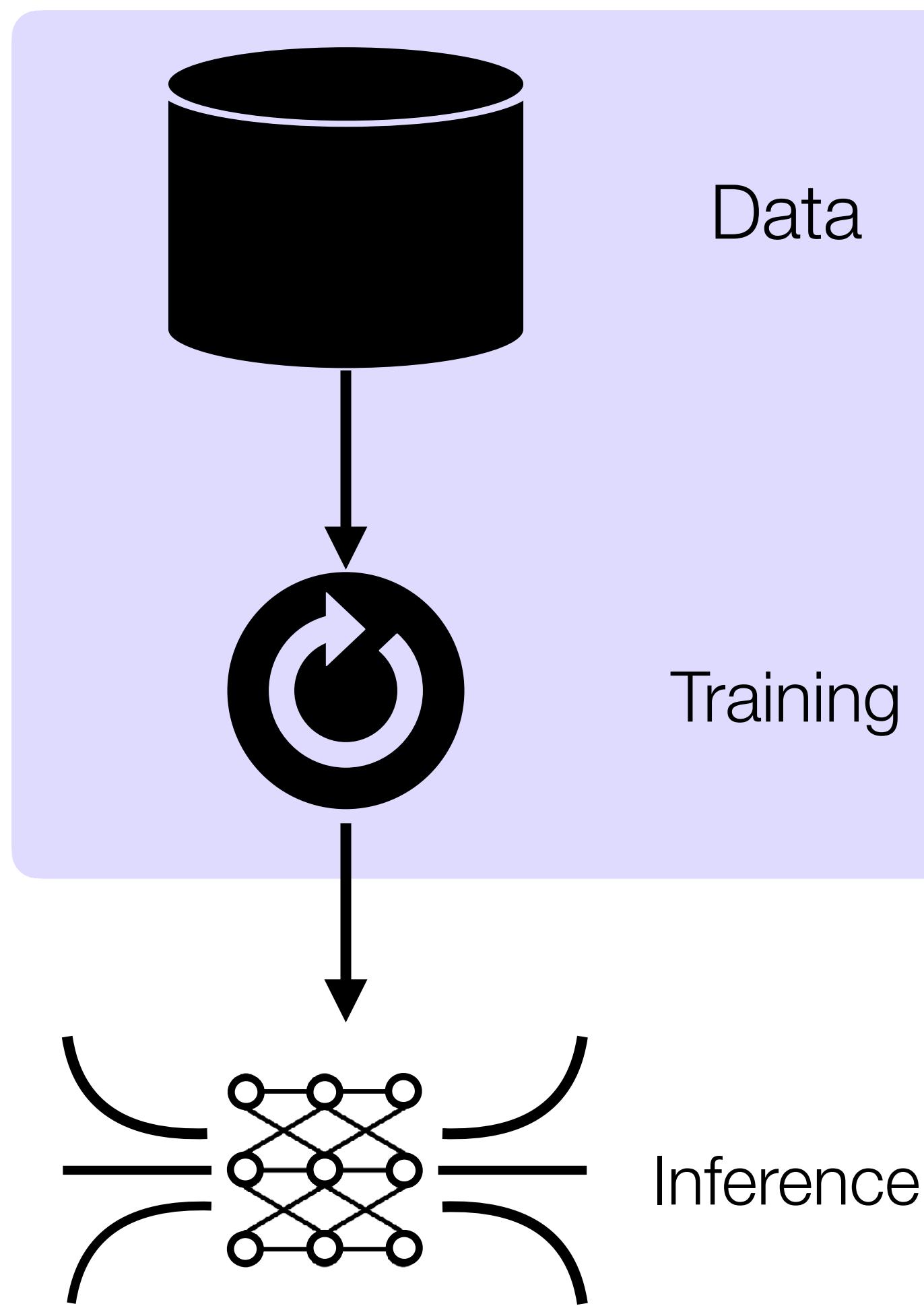
Future Work: Towards Module Orchestration



Future Work: Towards Module Orchestration

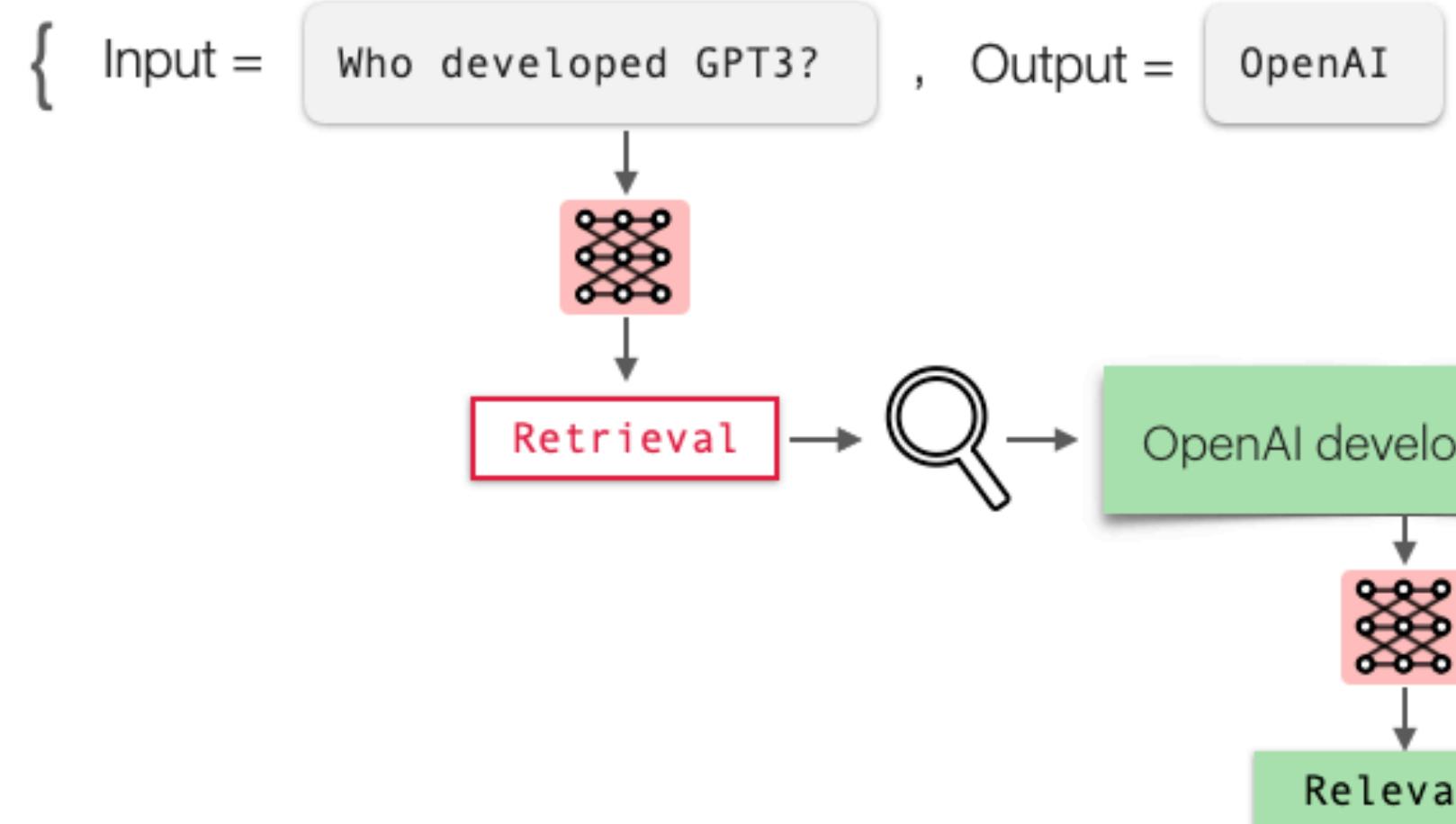
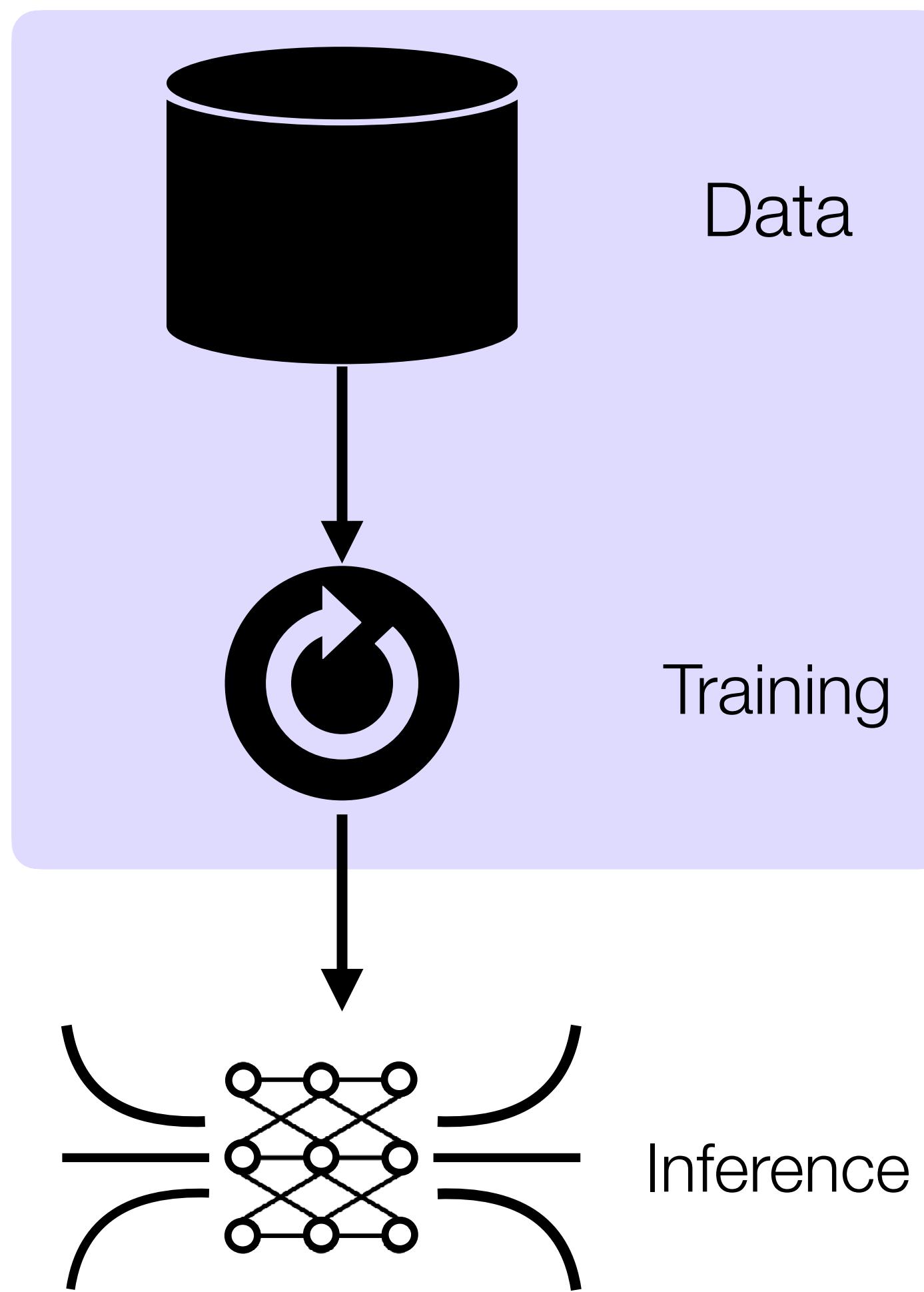


Future Work: Towards Module Orchestration



Asai et al., ICLR 2024 (Oral - Top 1%). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection

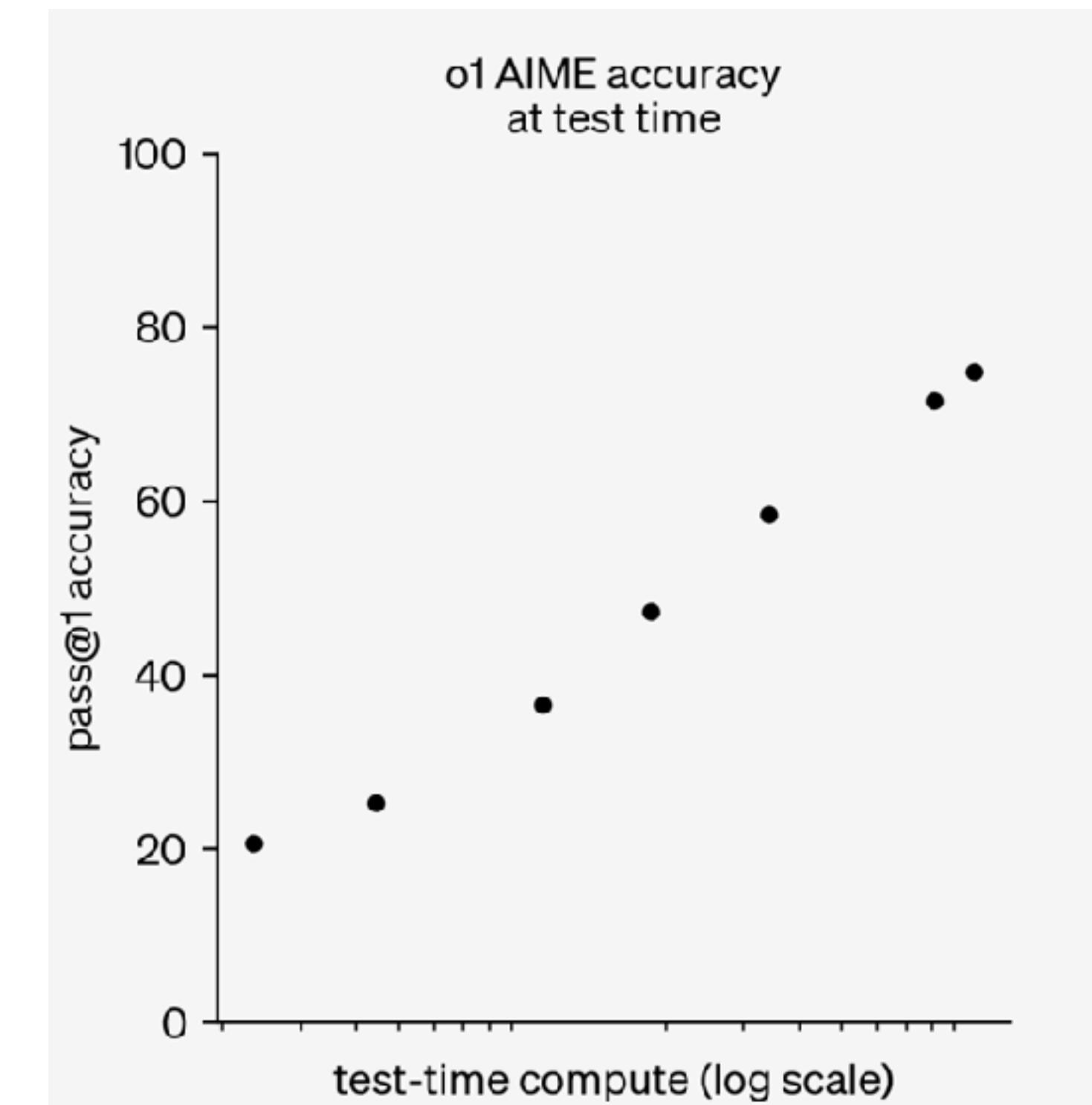
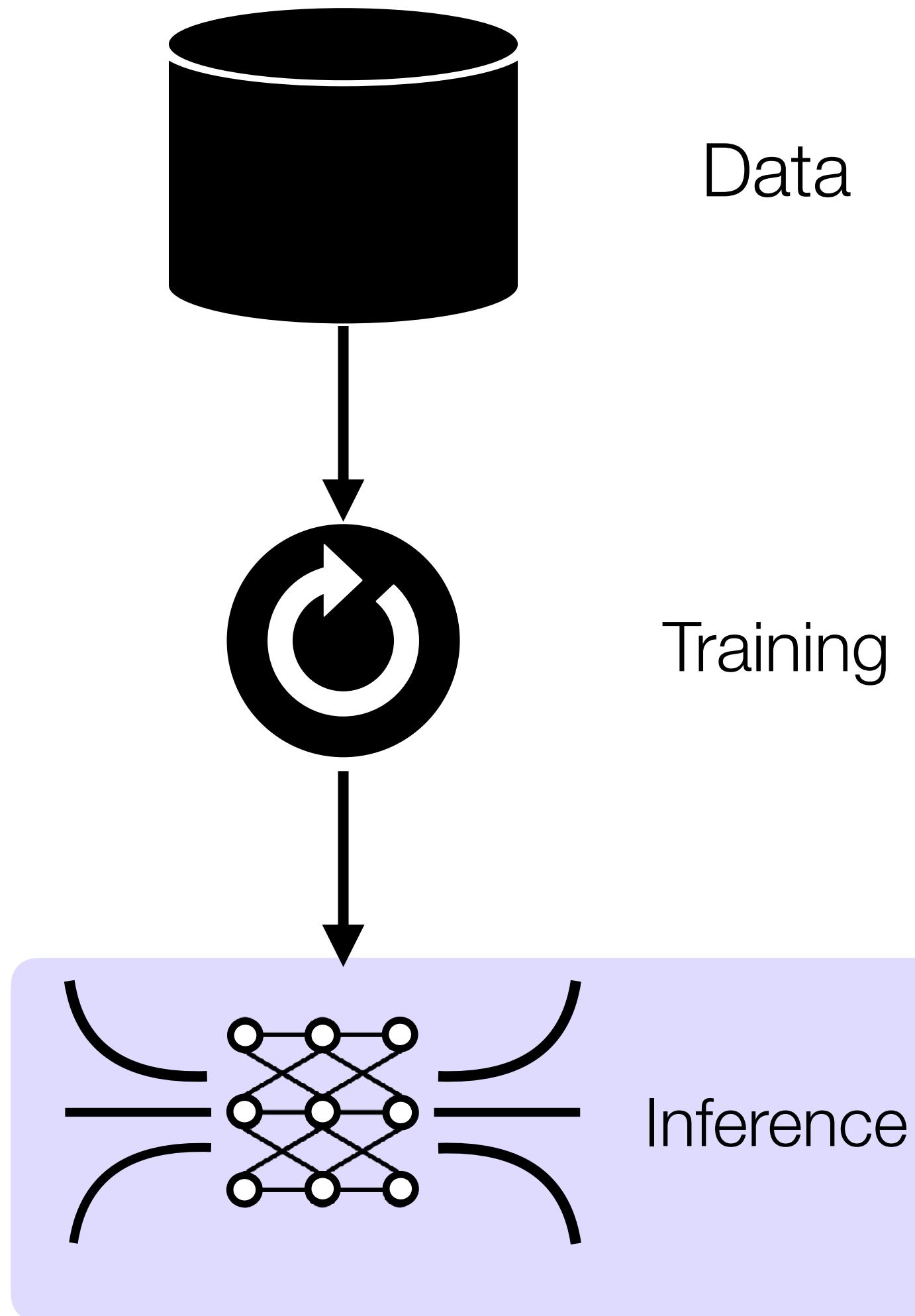
Future Work: Towards Module Orchestration



Asai et al., ICLR 2024 (Oral - Top 1%). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection

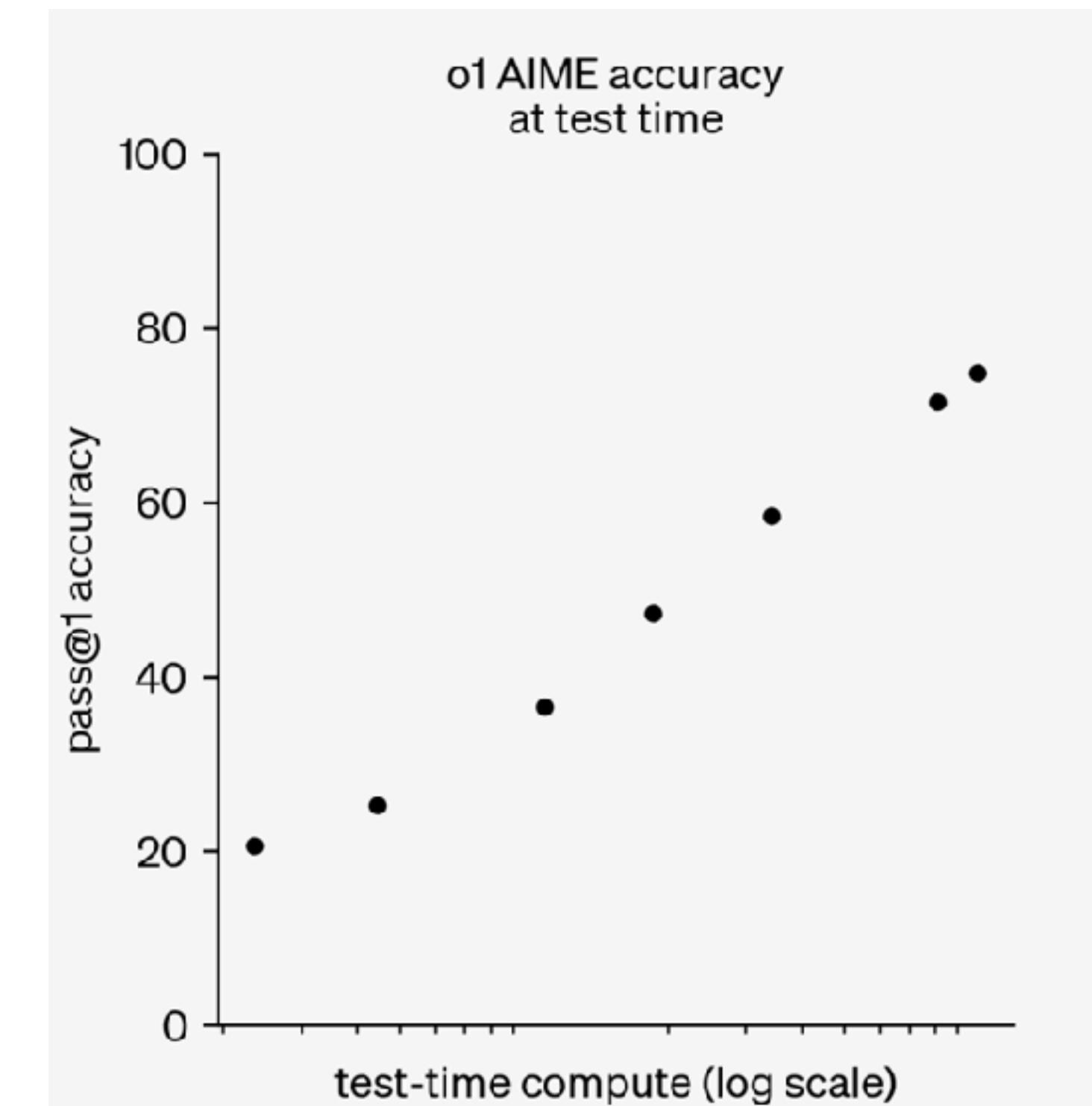
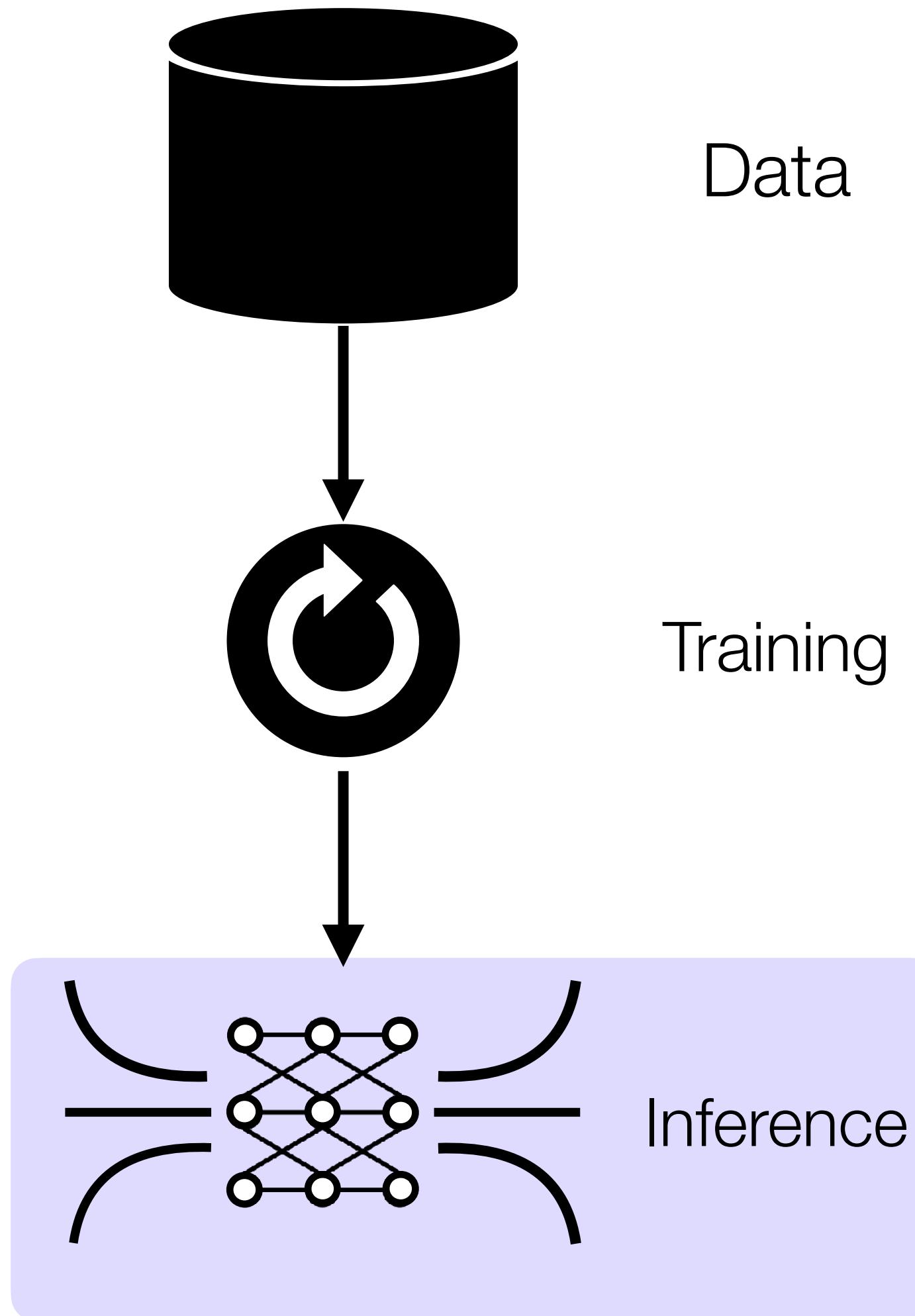
How can we effectively train Augmented LMs?

Future Work: Towards Module Orchestration



Open AI. Learning to reason with LLMs.
<https://openai.com/index/learning-to-reason-with-langs>

Future Work: Towards Module Orchestration

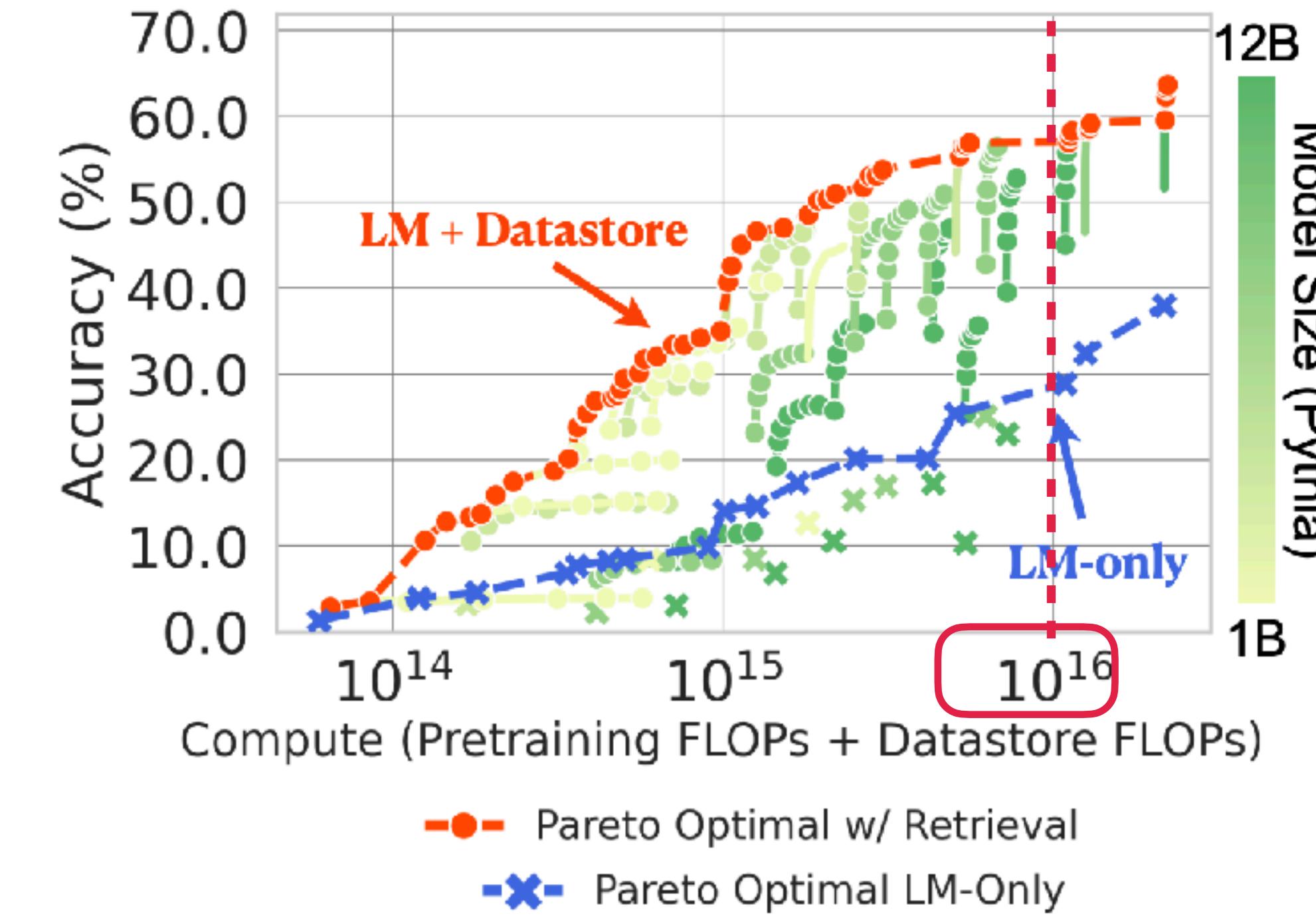
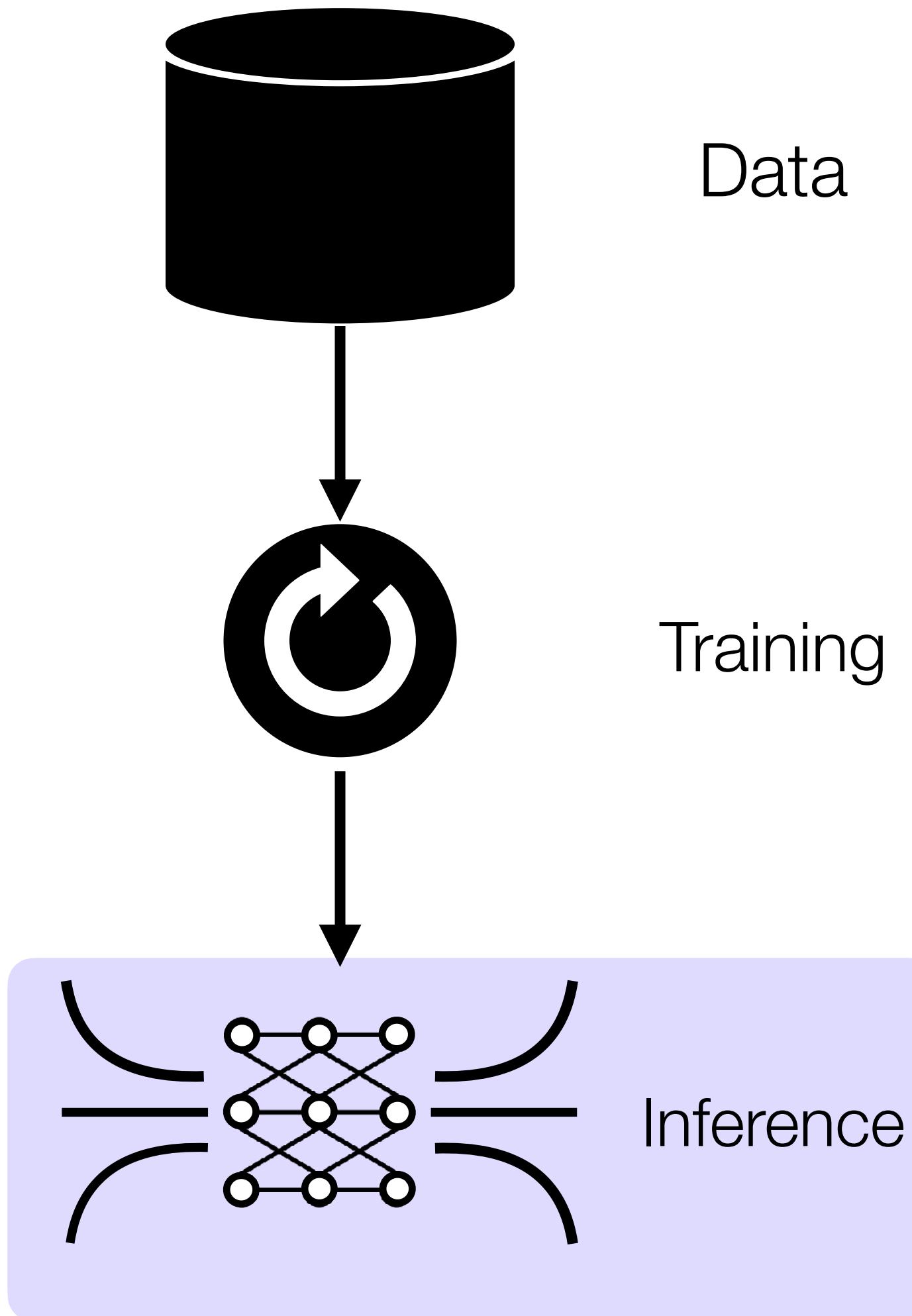


Open AI. Learning to reason with LLMs.
<https://openai.com/index/learning-to-reason-with-langs>



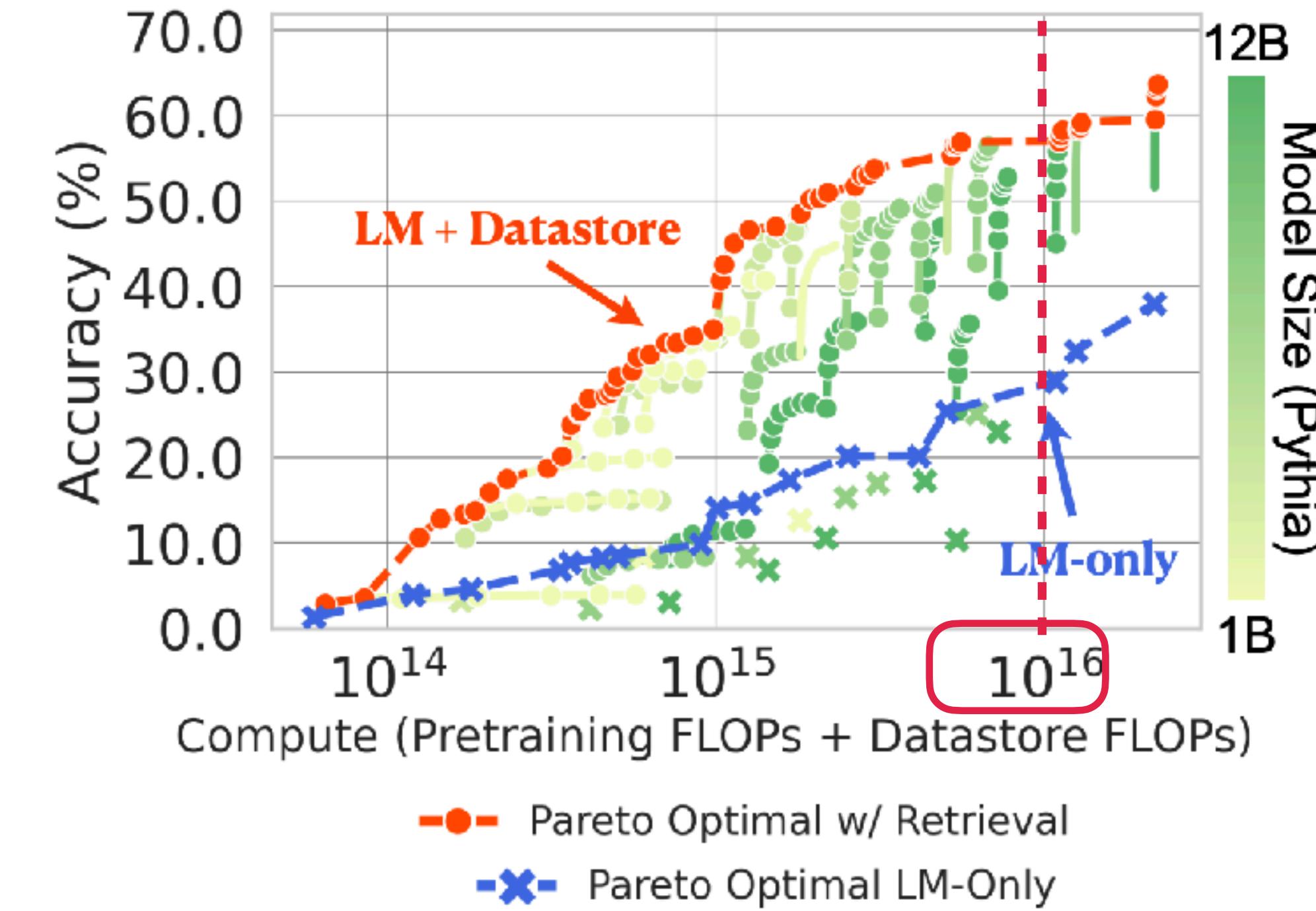
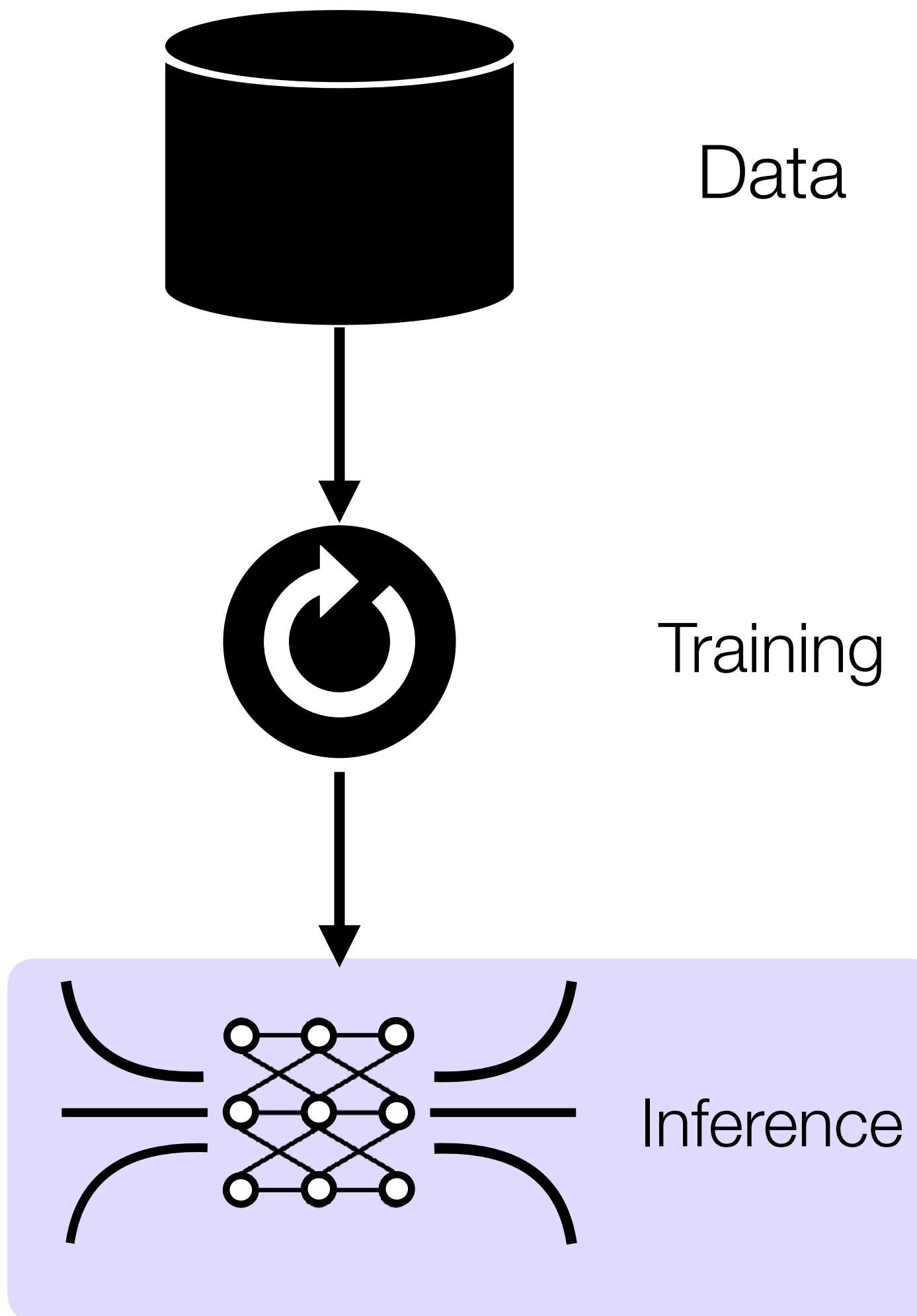
Reasoned about exam solution for 2m 56s

Future Work: Towards Module Orchestration



Shao, He, Asai et al. NeurIPS 2024. Scaling Retrieval-Based Language Models with a Trillion-Token Datastore.

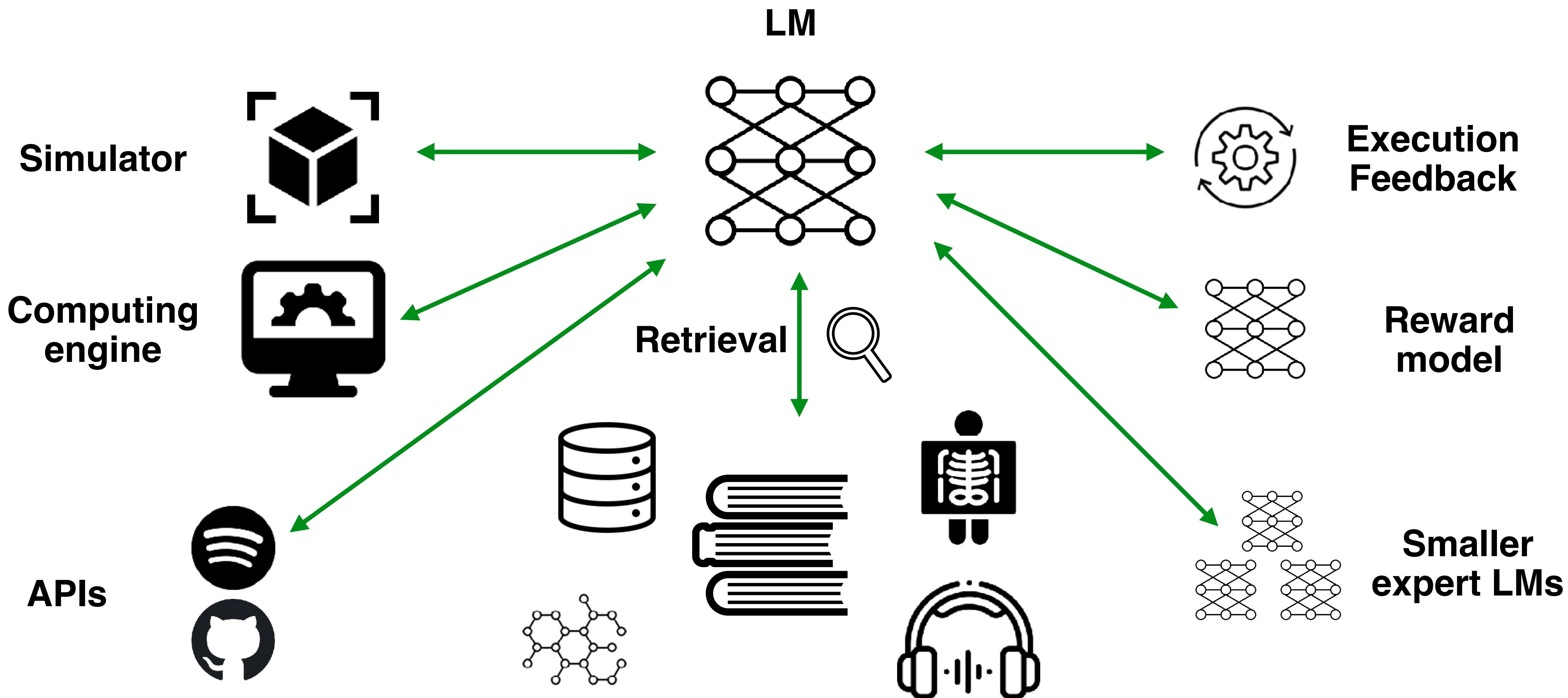
Future Work: Towards Module Orchestration



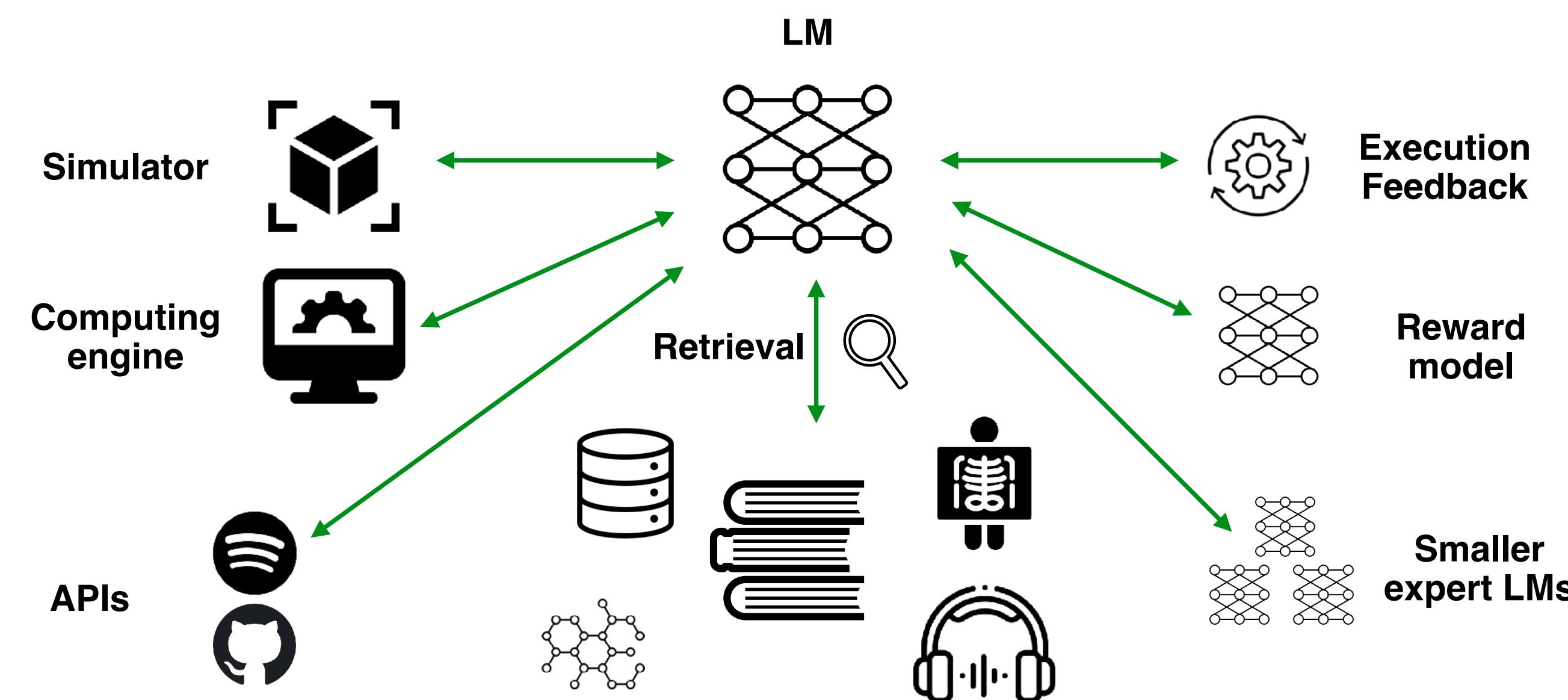
Shao, He, Asai et al. NeurIPS 2024. Scaling Retrieval-Based Language Models with a Trillion-Token Datastore.

How can we optimize test-time performance and efficiency of Augmented LMs?

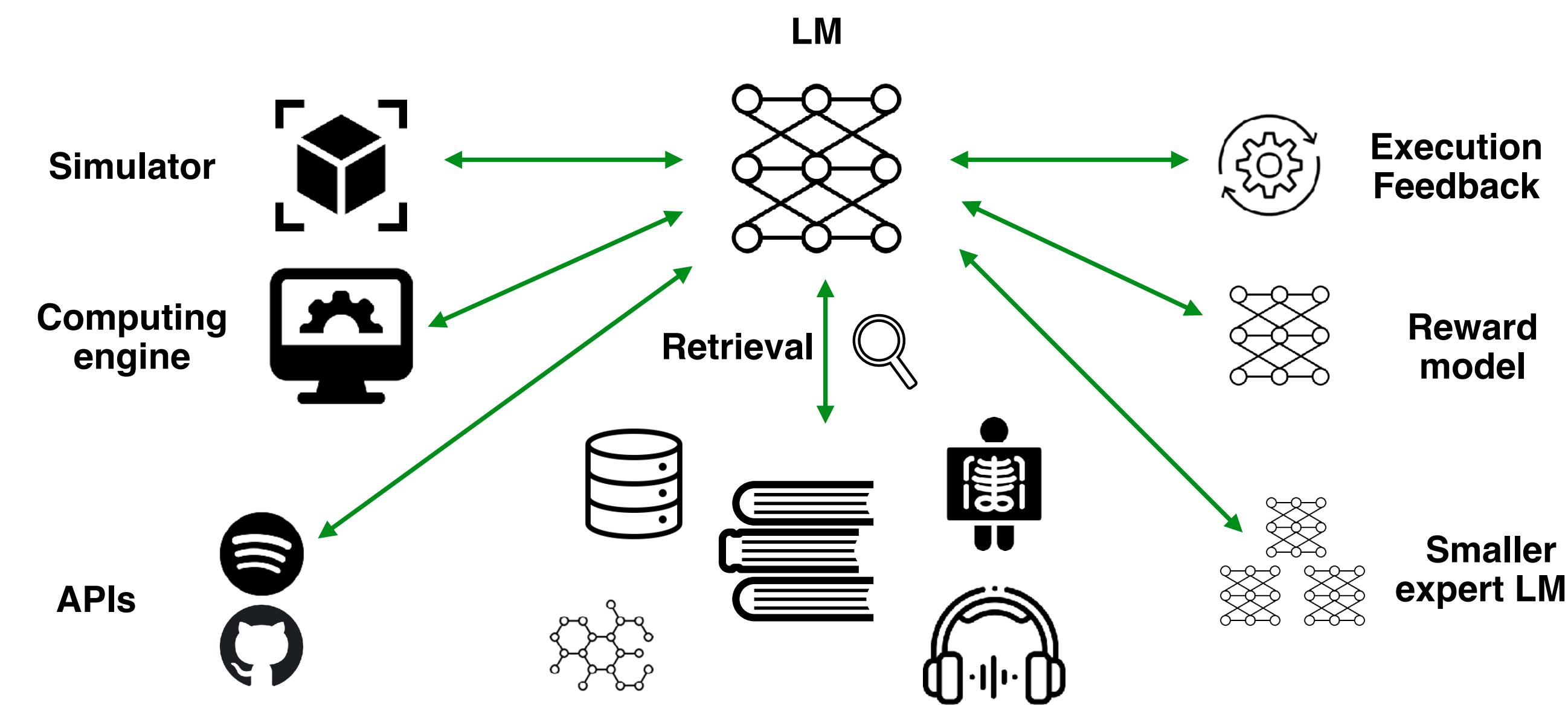
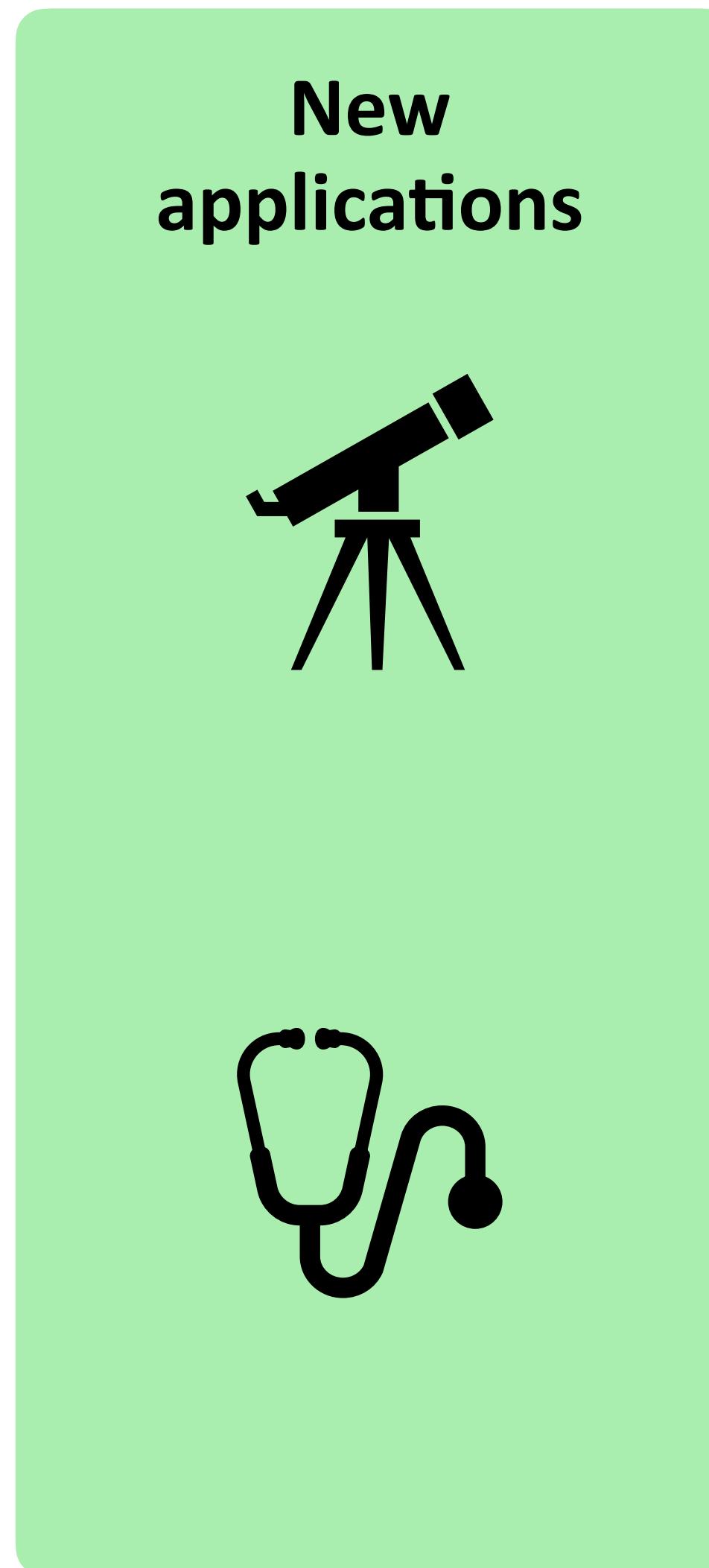
Future work: Responsible Evaluations in New Frontiers



Future work: Responsible Evaluations in New Frontiers

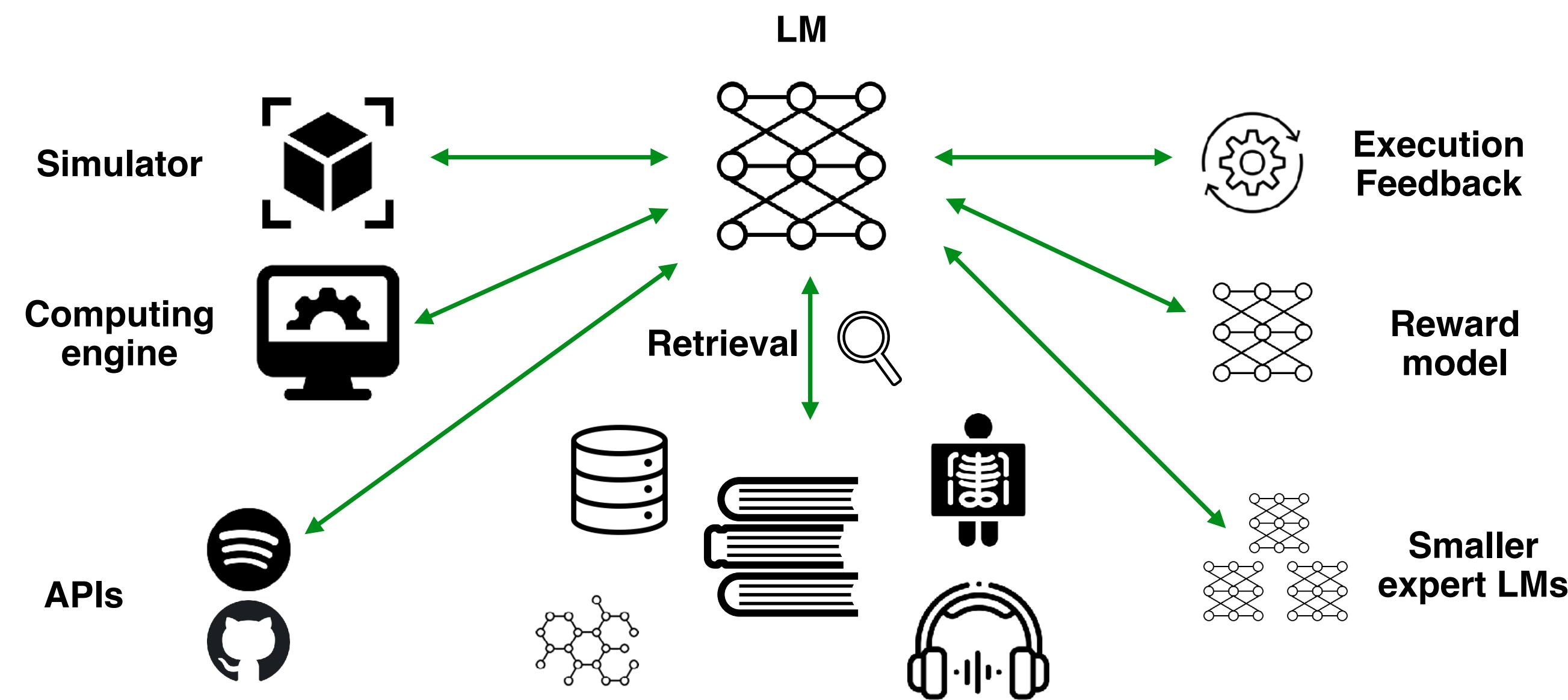


Future work: Responsible Evaluations in New Frontiers

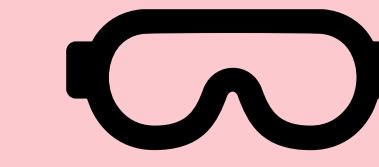


Future work: Responsible Evaluations in New Frontiers

New applications

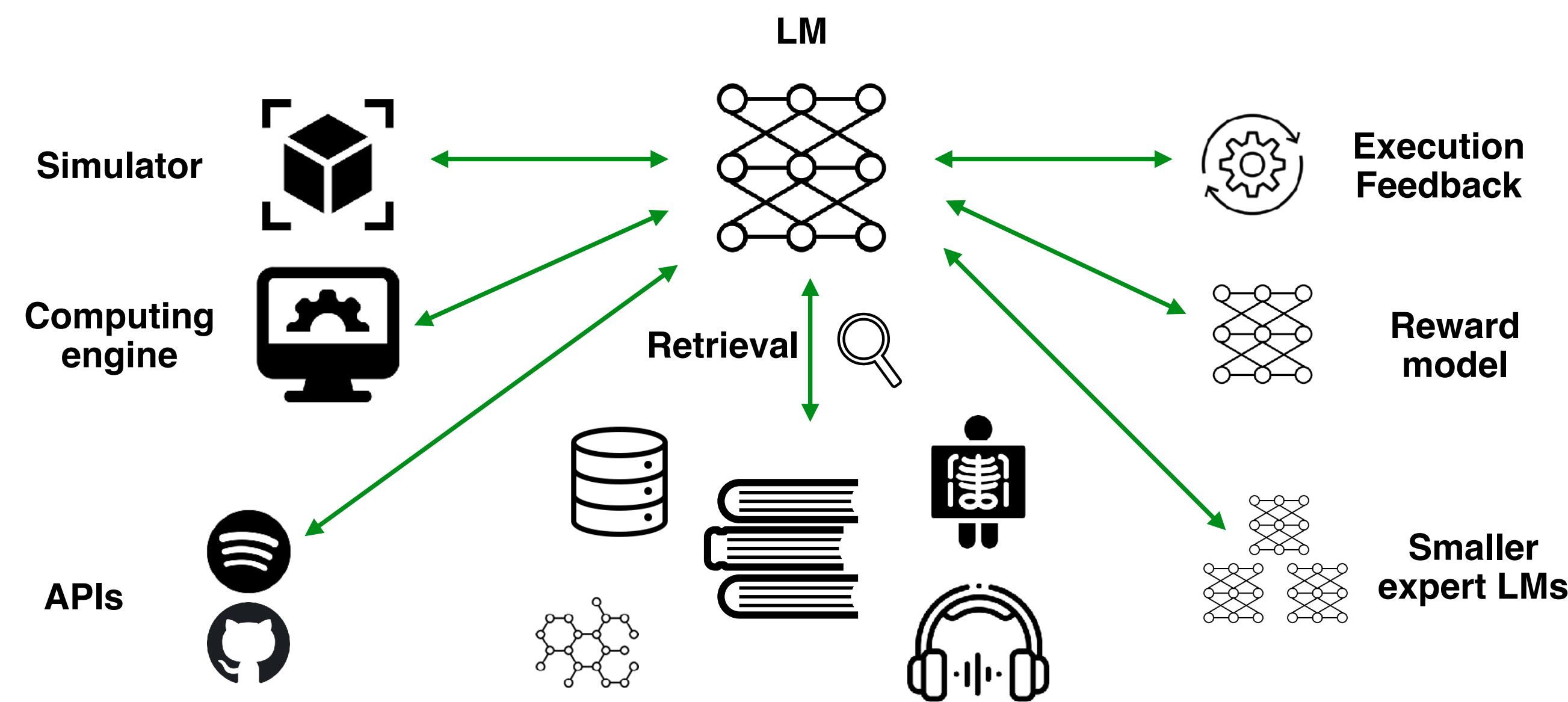


New risks



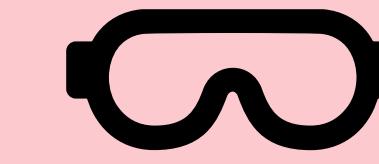
Future work: Responsible Evaluations in New Frontiers

New applications



Expert-in-the-loop evaluations for new frontiers

New risks





Cornell University.



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL



UNIVERSITY OF
ILLINOIS
URBANA-CHAMPAIGN



WISCONSIN
UNIVERSITY OF WISCONSIN-MADISON

東北大學
TOHOKU
UNIVERSITY



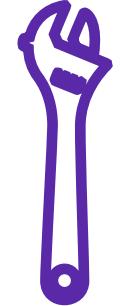
Beyond Scaling: Frontiers of Retrieval-Augmented LMs

Akari Asai (University of Washington)



From Monolithic

- Systematically analyze **limitations of scaling monolithic LMs**
- Demonstrate their **reliability, adaptability** and **efficiency**



New Foundations

- **Design** and **train** core components for Retrieval-Augmented LMs
- Integrated into major libraries, leading to state-of-the-art systems



Real-world Impact

- Extend the foundations to drive impact in **scientific literature synthesis**
- Outperformed **GPT-4o** and **experts**, now used by **30K researchers**

Website

akariasai.github.io/

Email

akari@cs.washington.edu

Feedback / Questions (Google form)

<https://tinyurl.com/akaritalkfb>