



Generative Reasoning in Pre-trained Language Models

Junjie Hu

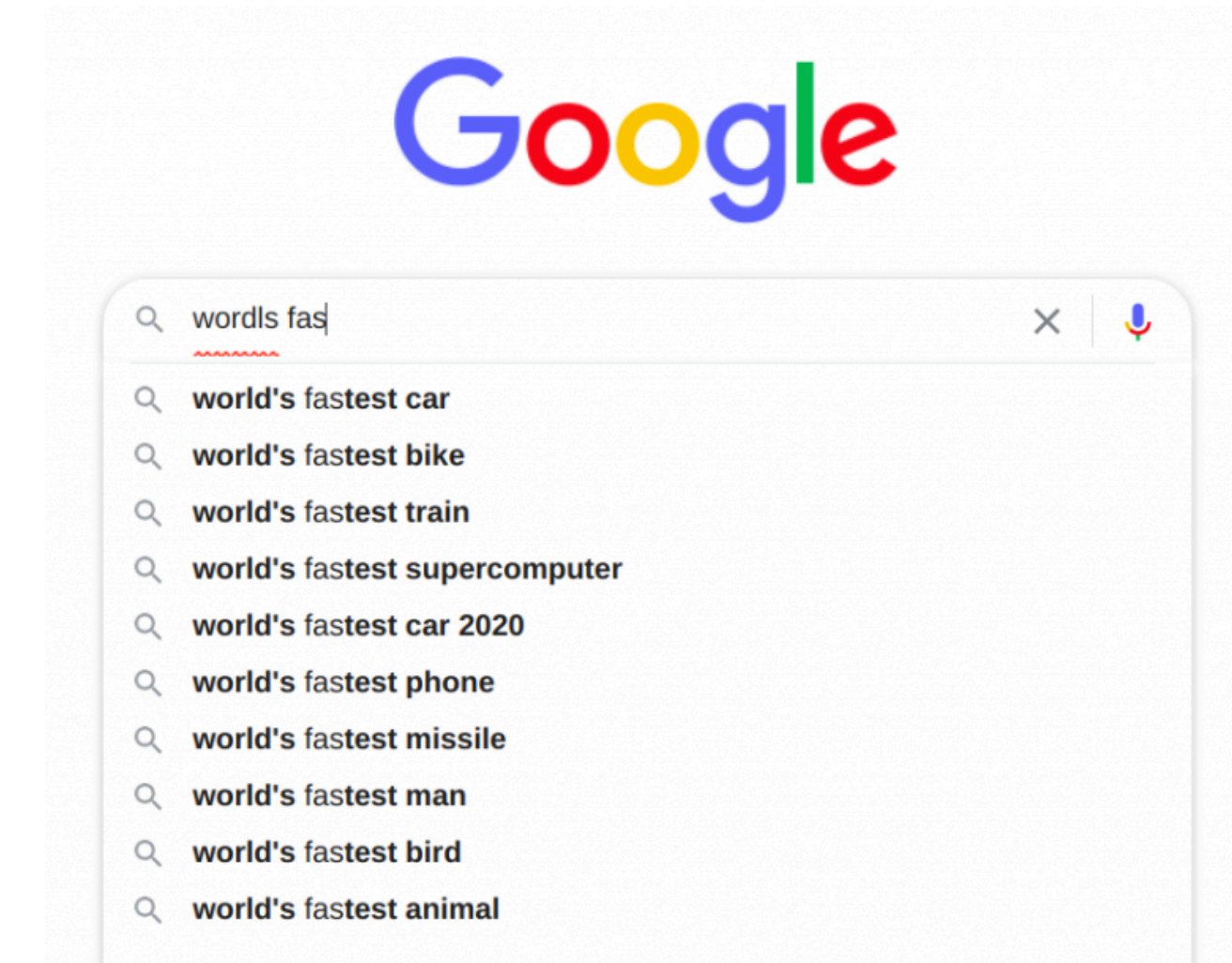
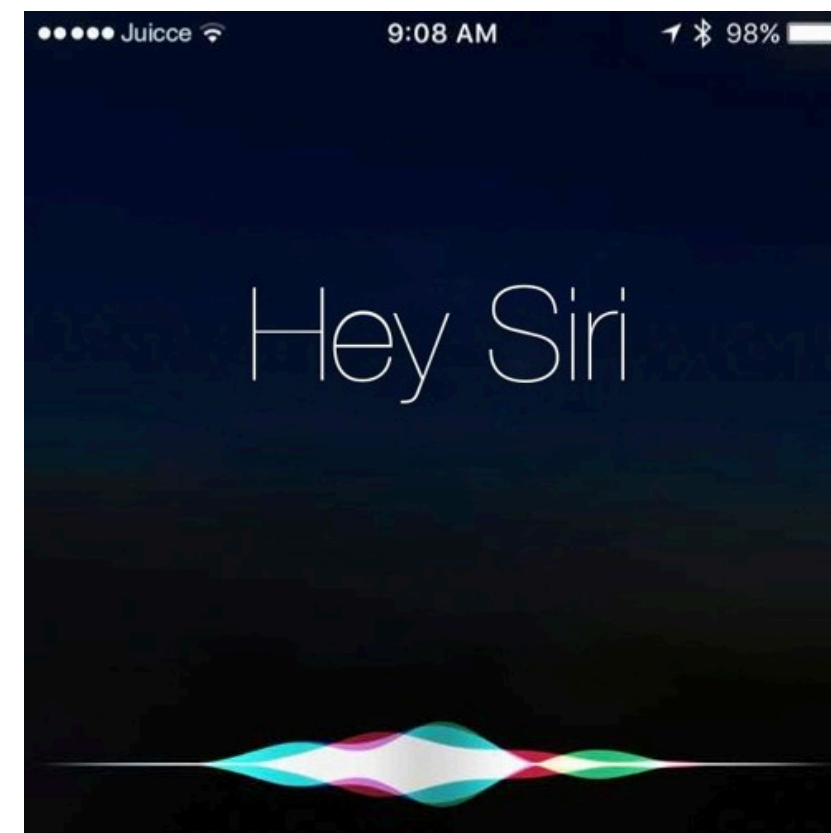
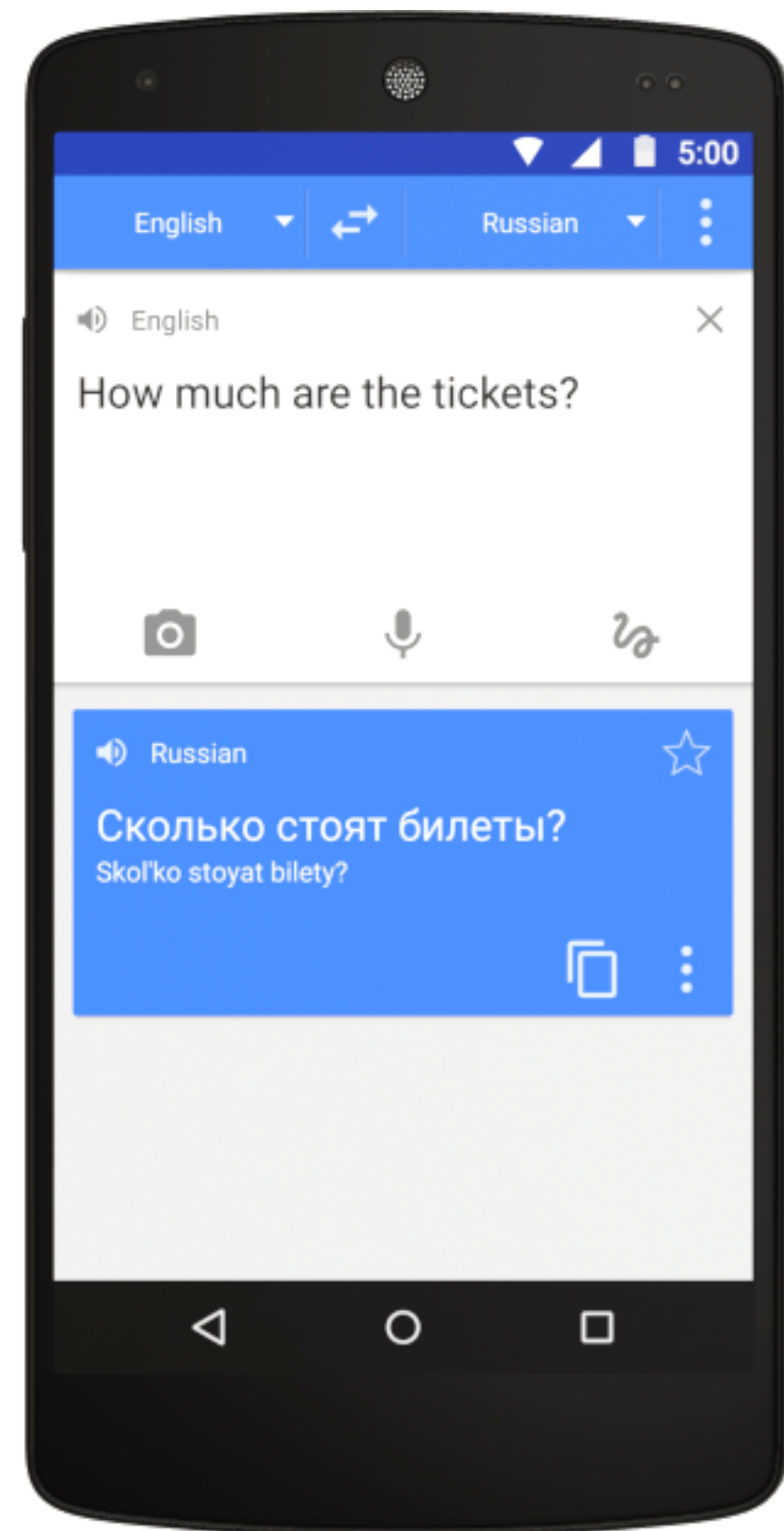
<https://junjihu.github.io/>



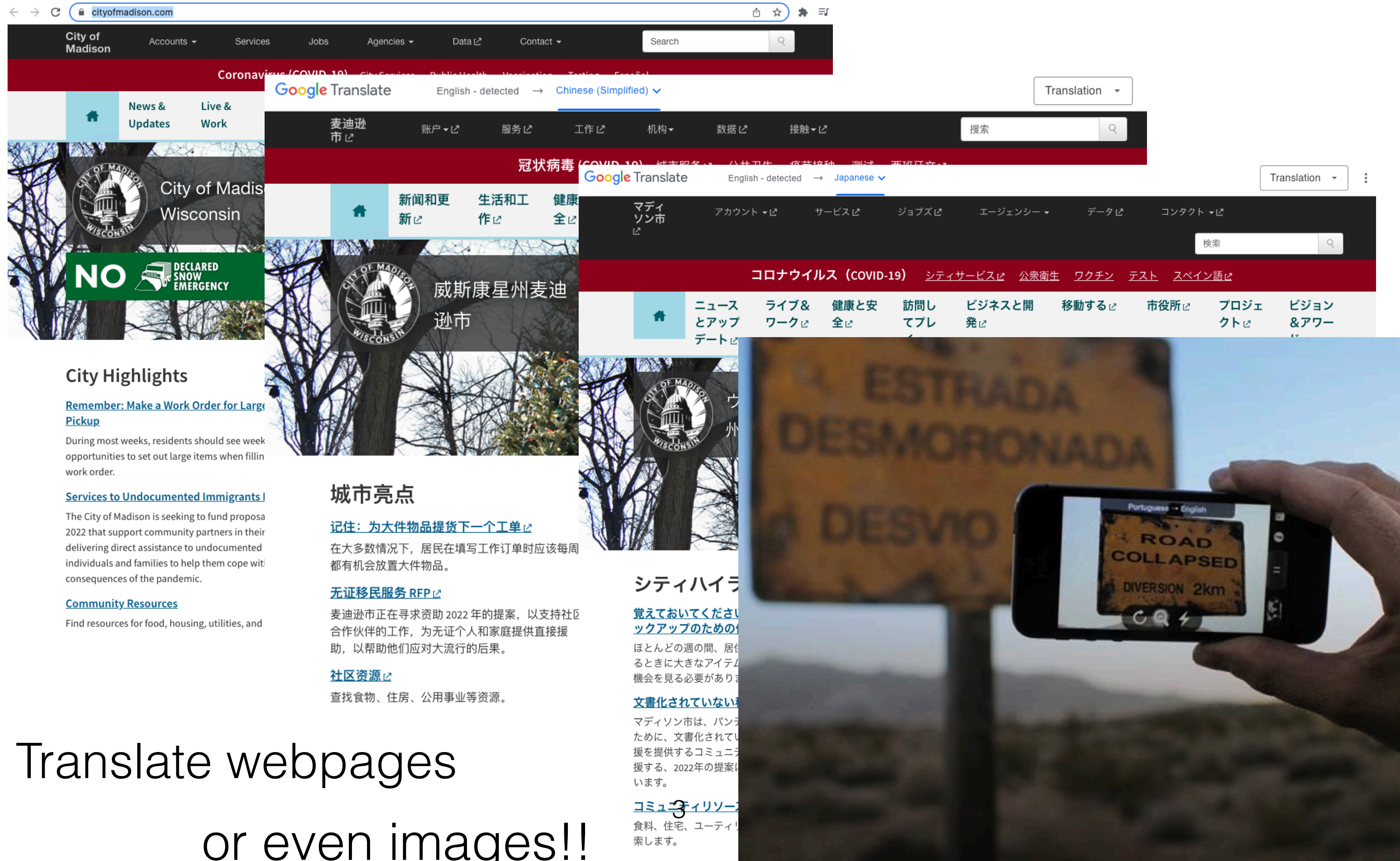
WISCONSIN
UNIVERSITY OF WISCONSIN-MADISON

Natural Language Processing

Process and analyze natural language data to enhance communications



NLP can Translate Text

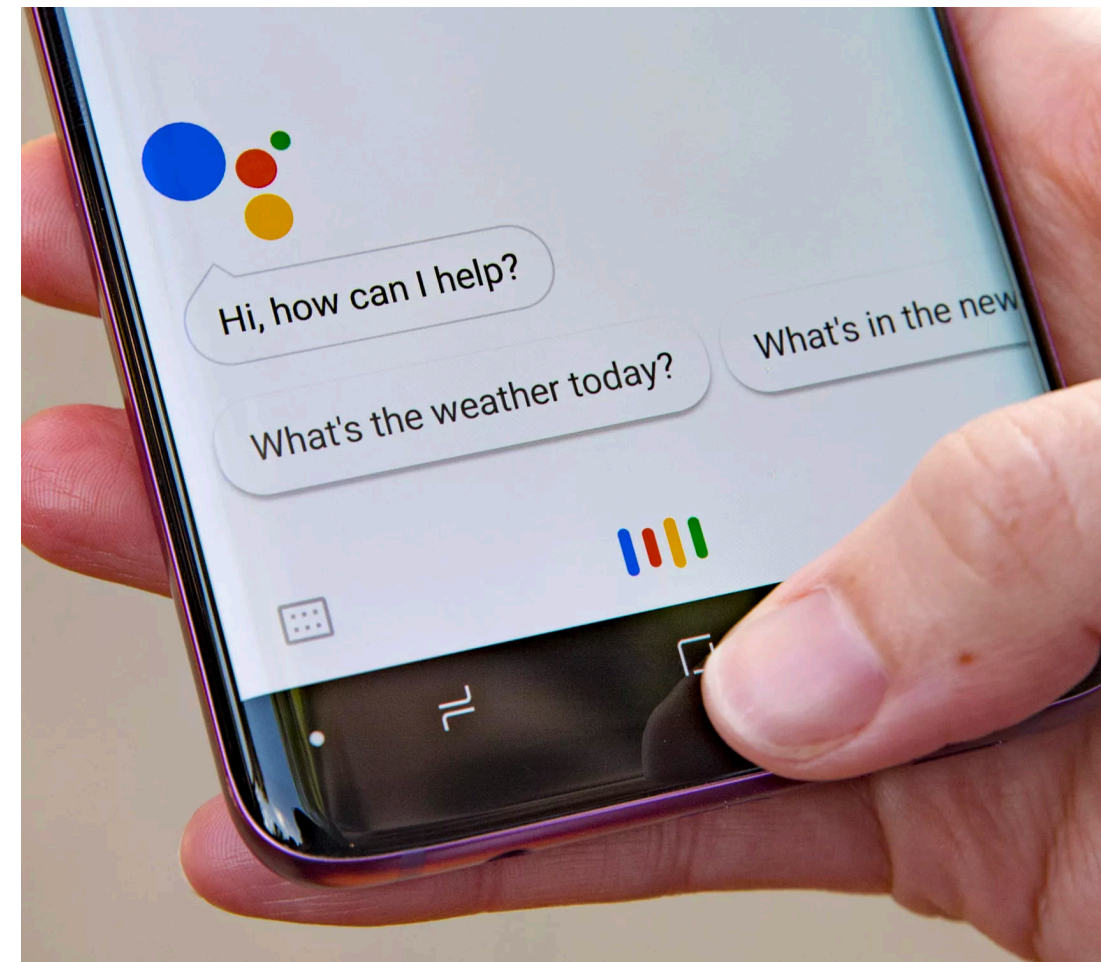


Translate webpages
or even images!!

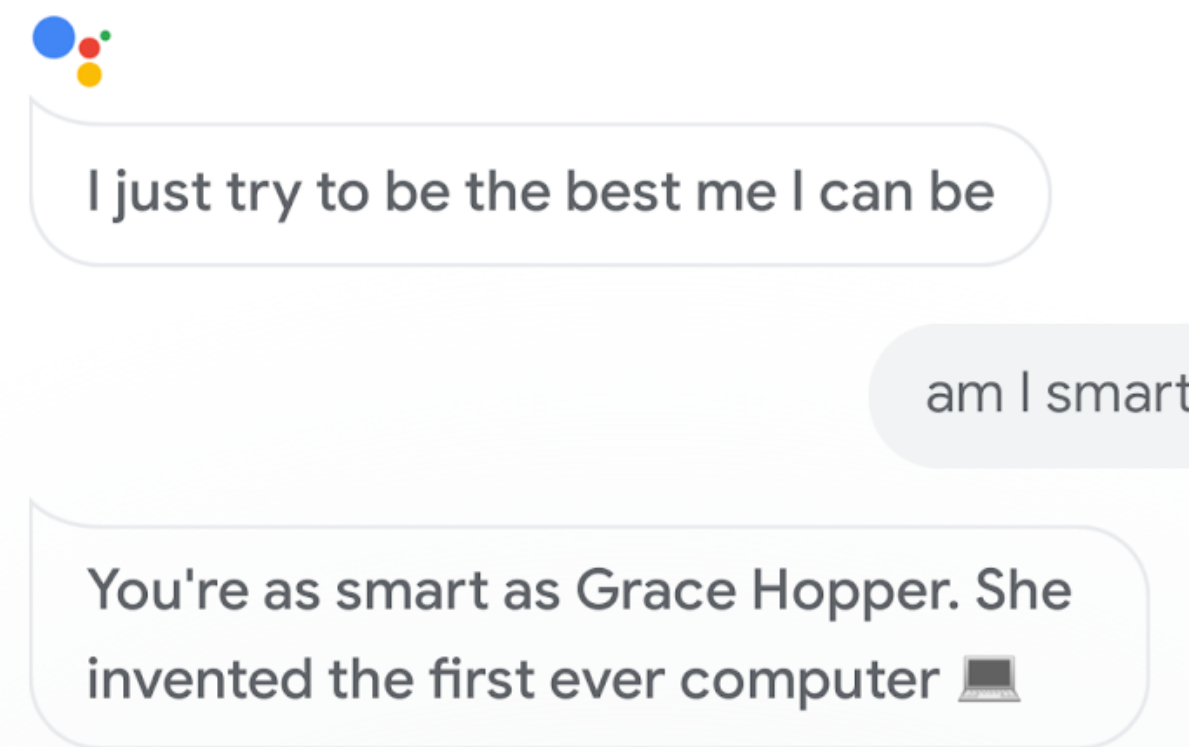
NLP can Chat with You

Conversational agents:

- Speech recognition
- Language analysis
- Dialogue processing
- Information retrieval
- Text to speech

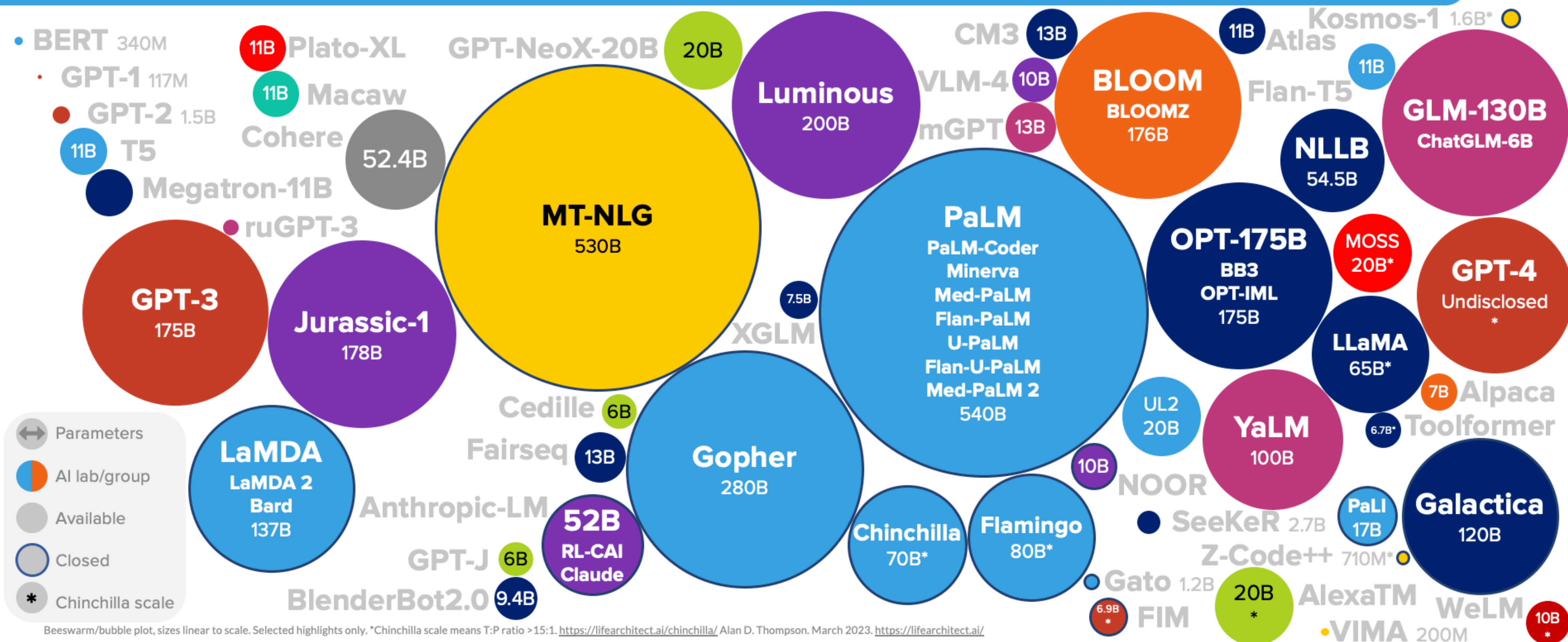



works with the
Google Assistant



Recent Advances driven by Pre-trained Language Models (PLM)

LANGUAGE MODEL SIZES TO MAR/2023



Beeswarm/bubble plot, sizes linear to scale. Selected highlights only. *Chinchilla scale means T:P ratio >15:1. <https://lifearchitect.ai/chinchilla/> Alan D. Thompson, March 2023. <https://lifearchitect.ai/>

PLMs can Answer our Questions

A screenshot of a Google search interface. The search bar contains the text "how many lakes does Madison have". Below the search bar, there are navigation tabs for "All", "News", "Images", "Maps", "Shopping", and "More", with "All" selected. The search results show "About 64,200,000 results (0.84 seconds)". The main heading is "five lakes". Below this, there is a snippet from a website: "Lake Kegonsa. From fishing to watersports, runs, bike rides, or simply nature watching, each of the **five** lakes around Madison provides a different experience for tourists and locals. Nov 29, 2018". The URL is "https://www.smarttoyota.com › blogs › madison-wi-lakes". The main link is "Madison Lakes | Explore the Incredible Lakes of Madison, WI". At the bottom, there are links for "About featured snippets" and "Feedback".

how many lakes does Madison have

All News Images Maps Shopping More Tools

About 64,200,000 results (0.84 seconds)

five lakes

Lake Kegonsa

From fishing to watersports, runs, bike rides, or simply nature watching, each of the **five** lakes around Madison provides a different experience for tourists and locals. Nov 29, 2018

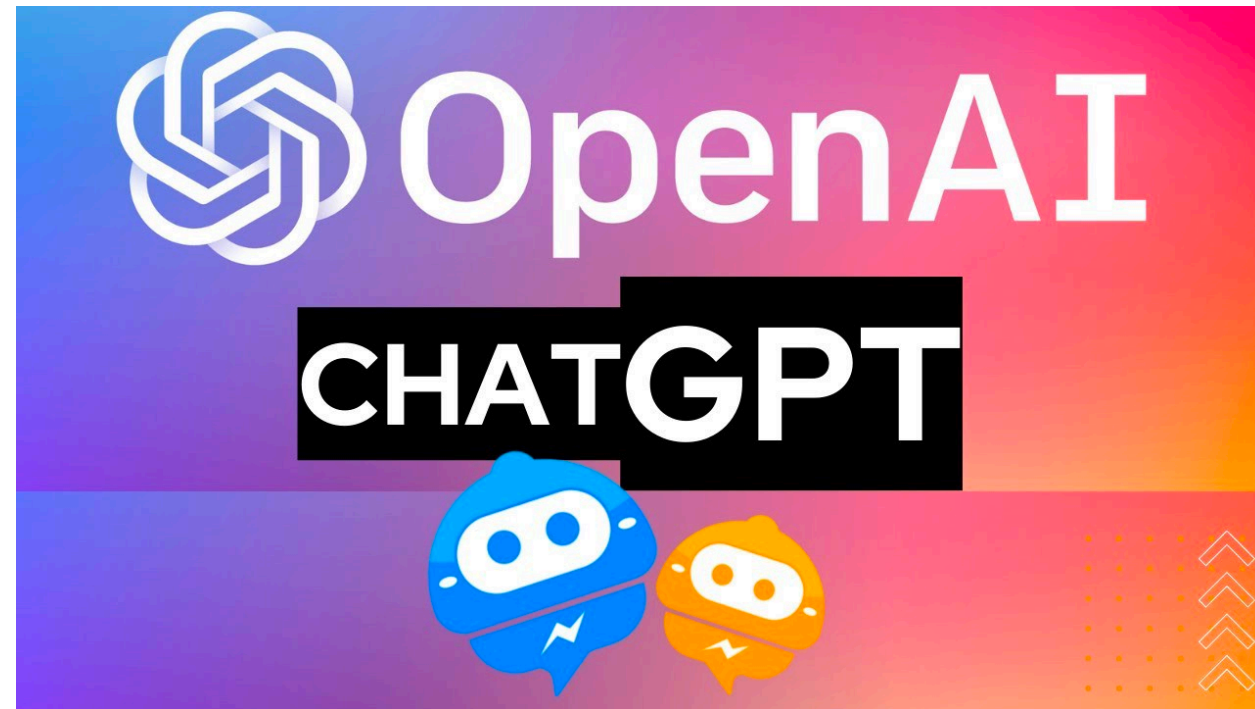
<https://www.smarttoyota.com › blogs › madison-wi-lakes>

[Madison Lakes | Explore the Incredible Lakes of Madison, WI](#)

About featured snippets Feedback

Retrieved Jan. 25, 2022

LLMs can be prompted with language instructions



- ChatGPT may save our time on text editing.

JU

Could you help me to write an announcement for the talk from Professor Junjie Hu on large language models and language reasoning?



Of course, I'd be happy to help you write an announcement for Professor Junjie Hu's talk on large language models and language reasoning. Here's a sample announcement you could use:

[Your Institution/Event Name] is proud to present a talk on "Large Language Models and Language Reasoning" by Professor Junjie Hu, a leading researcher in the field of natural language processing.

With the advent of large language models like GPT-3, there has been a surge of interest in the potential of these models to perform complex language reasoning tasks. In this talk, Professor Hu will discuss the latest developments in large language models and their applications in natural language understanding, question answering, and more.

Through his research, Professor Hu has made significant contributions to the field of natural language processing, particularly in the areas of machine reading comprehension and knowledge representation. He has published numerous papers in top-tier conferences such as ACL, EMNLP, and NeurIPS, and his work has been recognized with several prestigious awards.

This talk is suitable for anyone with an interest in natural language processing or machine learning, whether you're a student, researcher, or industry professional. We hope you'll join us for this exciting opportunity to learn from one of the leading experts in the field.

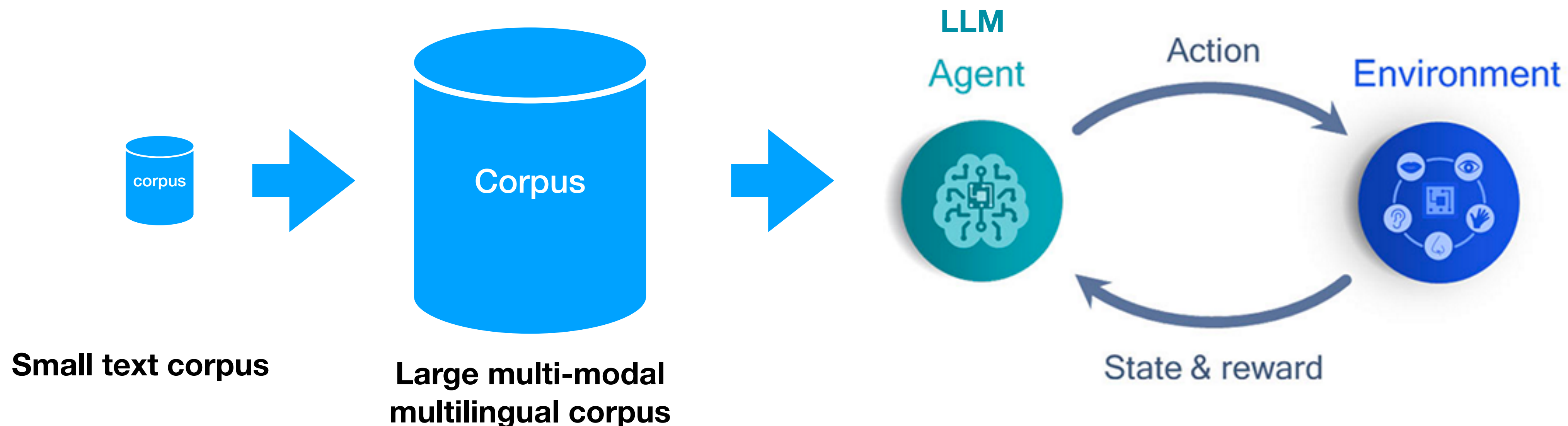
Date: [Insert date and time]

Location: [Insert location or online platform]

RSVP: [Insert RSVP information if applicable]

What makes pertained LLM so powerful?

- **Pre-training** on massive raw **texts** (and even **images**) in **100+ languages**
 - BERT, GPT-4, PaLM, T5, LLaMA, ...
- **Fine-tuning** on language **instructions** with **supervised learning** or **RL with human feedback**
 - ChatGPT, Bard, FlanT5, Alpaca, ...



What Can we Do w/ LMs?

- Score sentences, e.g., $P(X = \text{"Jane went to the store"})$:

Jane went to the store . \rightarrow high

store to Jane went the . \rightarrow low

- Learning meaningful language embeddings

$$g : \mathcal{X} \rightarrow \mathcal{Z}$$

- **Generate sentences:**

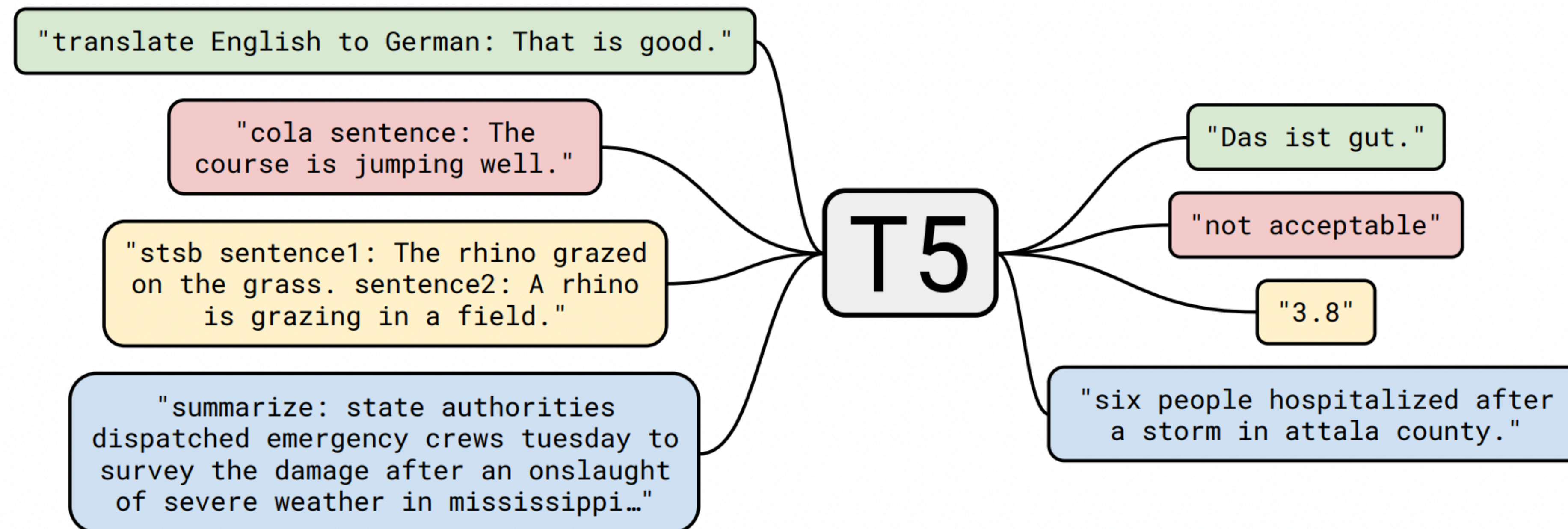
while didn't choose end-of-sentence symbol, i.e., [EOS]:

calculate probability $P(\text{Next Word} \mid \text{Context})$

sample a new word from the probability distribution

LLM Prompting: One generative model for ALL

- Convert all NLP tasks into a autoregressive generative task.
- Append an instruction (e.g., “translate English to German”) before the real input sentence.



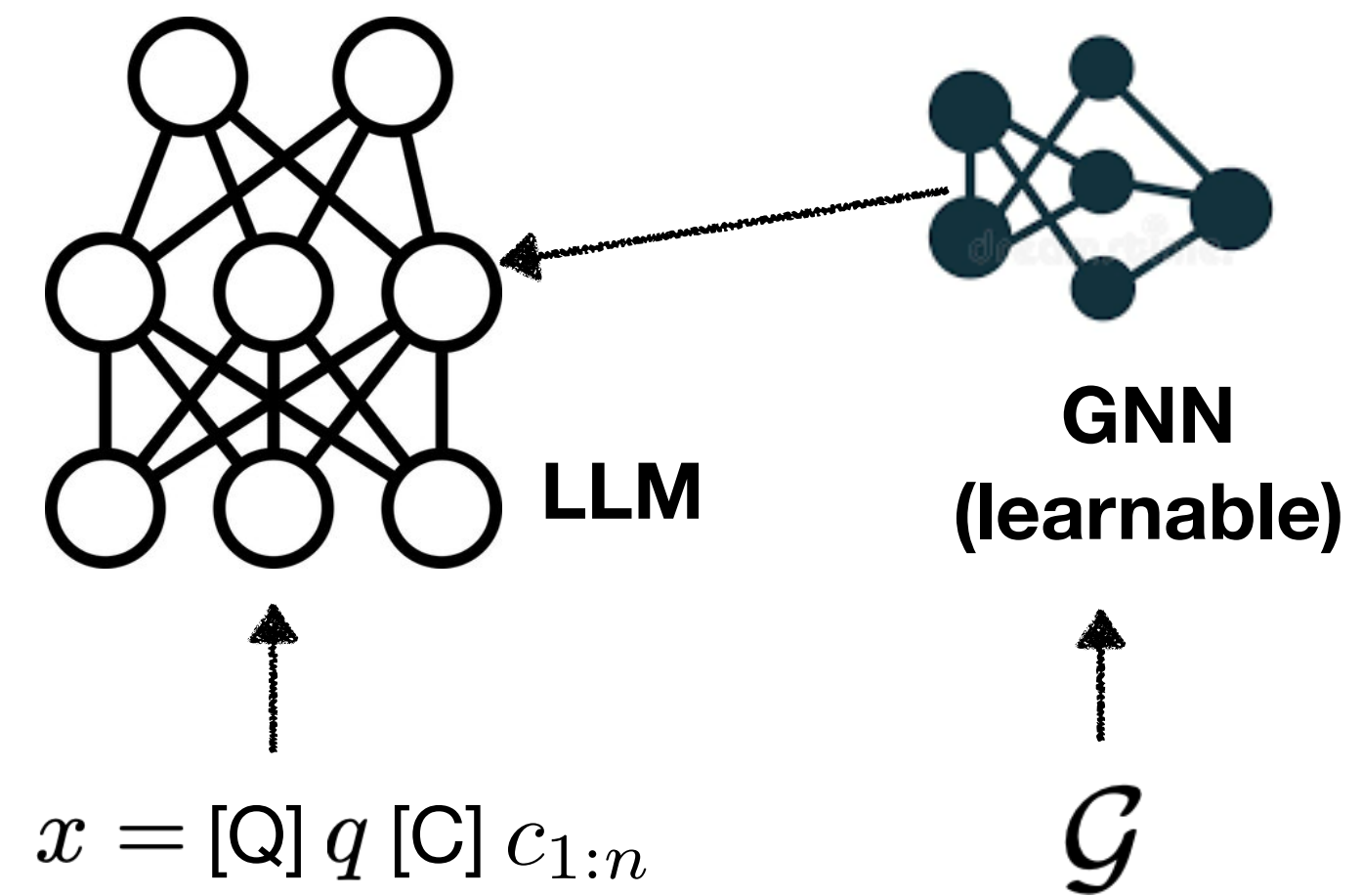
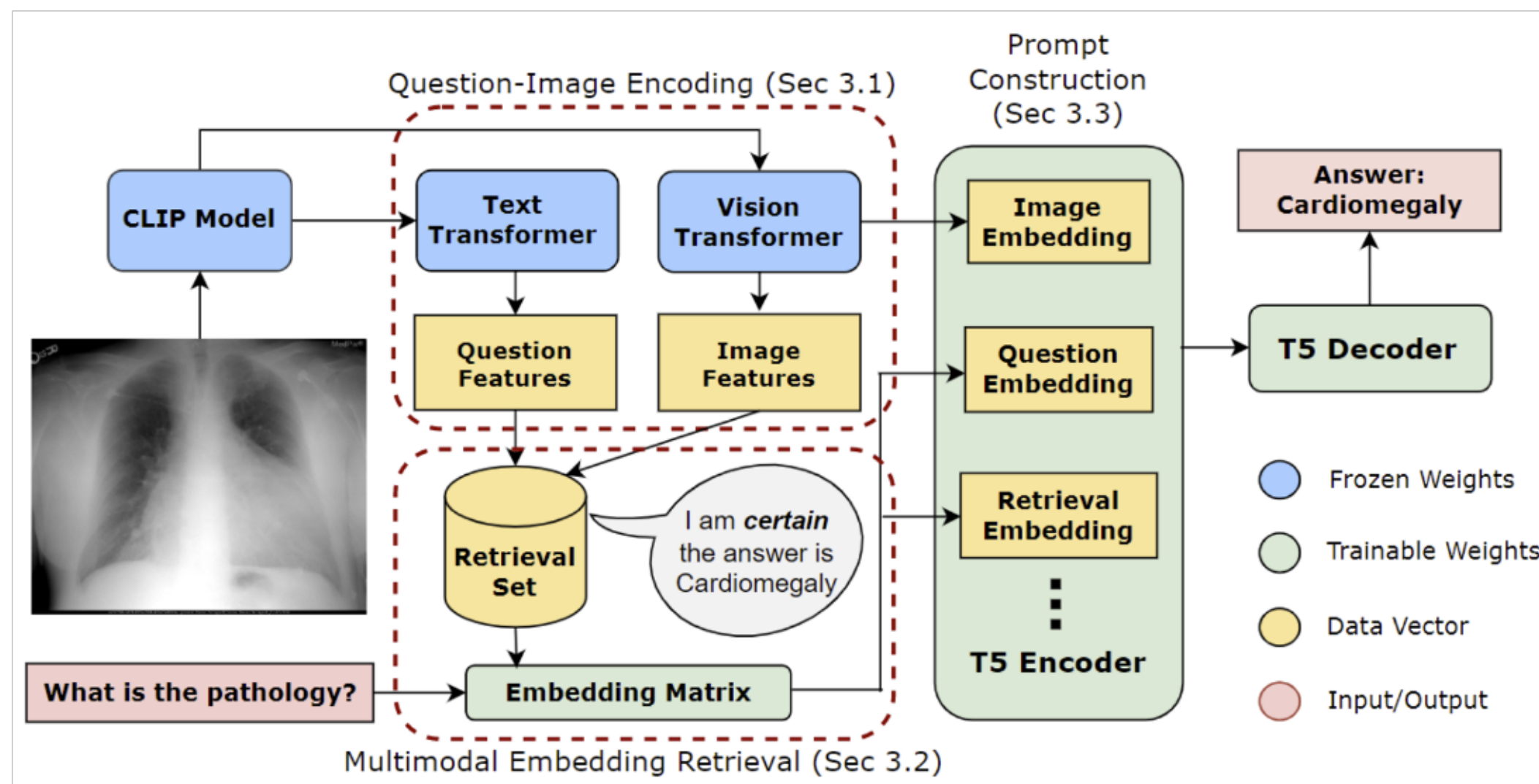
But, is PLM alone sufficient for generative reasoning?

- **Internal v.s. External Knowledge:** PLM's internal knowledge may be outdated, and external knowledge is more up-to-date.
- **Structured v.s. Unstructured Data:** Informations can be better represented in a structured graph than in unstructured raw text
- **Single-Modality v.s. Multi-Modality:** Other modality provides extra information in addition to PLM's embedded knowledge

Augmented PLMs for Generative Reasoning

- Multimodal PLMs via prompt retrieval

- Structure-augmented PLMs via graph reasoning



Structured CoT Tuning

[OH, ACL 2023 Findings]

[RSH, ACL 2023]

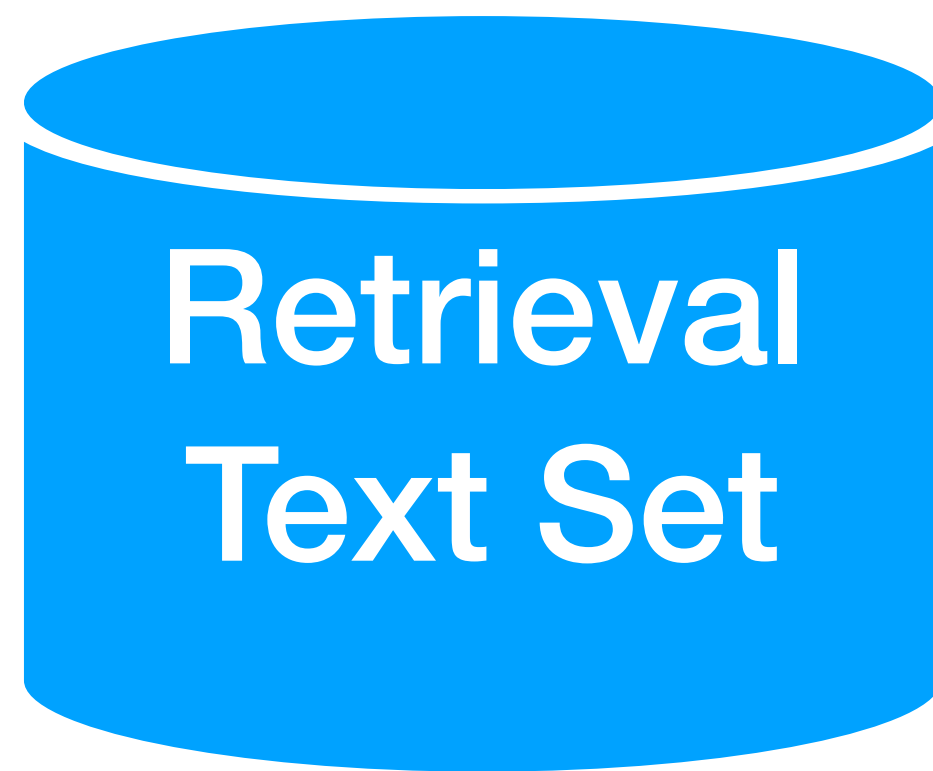
Multimodal PLMs for Generative Visual Question Answering

Retrieval-augmented LMs

- KNN LMs: Use k-nearest neighbor search to find similar context for next word prediction

- Example:

Obama's birthplace is _____



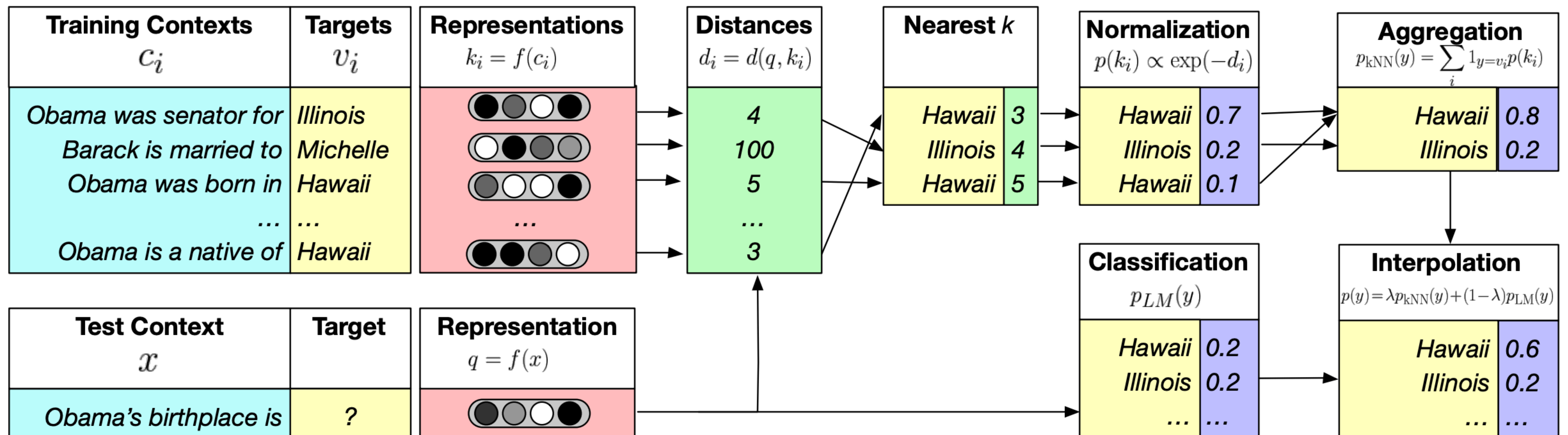
| Training Contexts c_i | Targets v_i | Representations $k_i = f(c_i)$ |
|----------------------------|------------------|-----------------------------------|
| Obama was senator for | Illinois | |
| Barack is married to | Michelle | |
| Obama was born in | Hawaii | |
| ... | ... | ... |
| Obama is a native of | Hawaii | |

| Test Context x | Target | Representation $q = f(x)$ |
|-----------------------|--------|------------------------------|
| Obama's birthplace is | ? | |

KNN-LM Example

- **Rapid Adaptation:** it's easy to replace the retrieval set for other domains of texts without further fine-tuning LMs
- **However,** it's hard to balance between KNN outputs and LM outputs by λ

$$p(y|x) = \lambda p_{\text{kNN}}(y|x) + (1 - \lambda) p_{\text{LM}}(y|x)$$

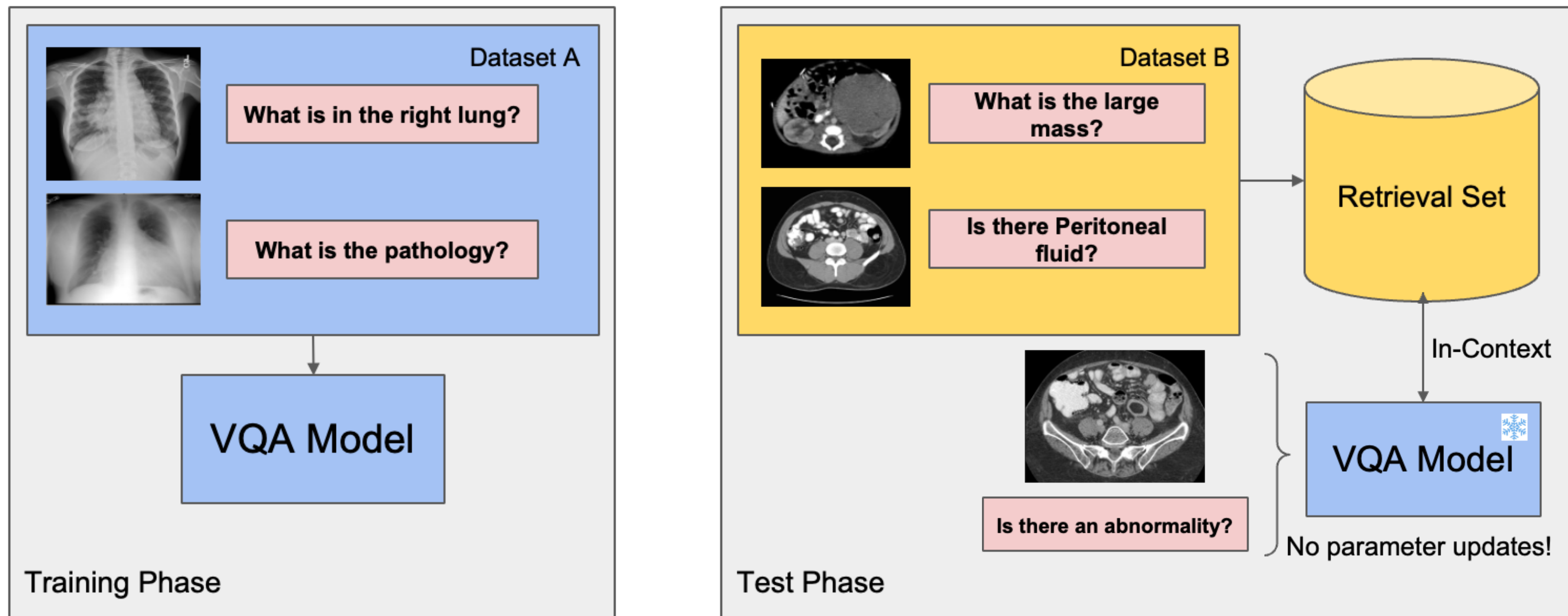


Research Questions

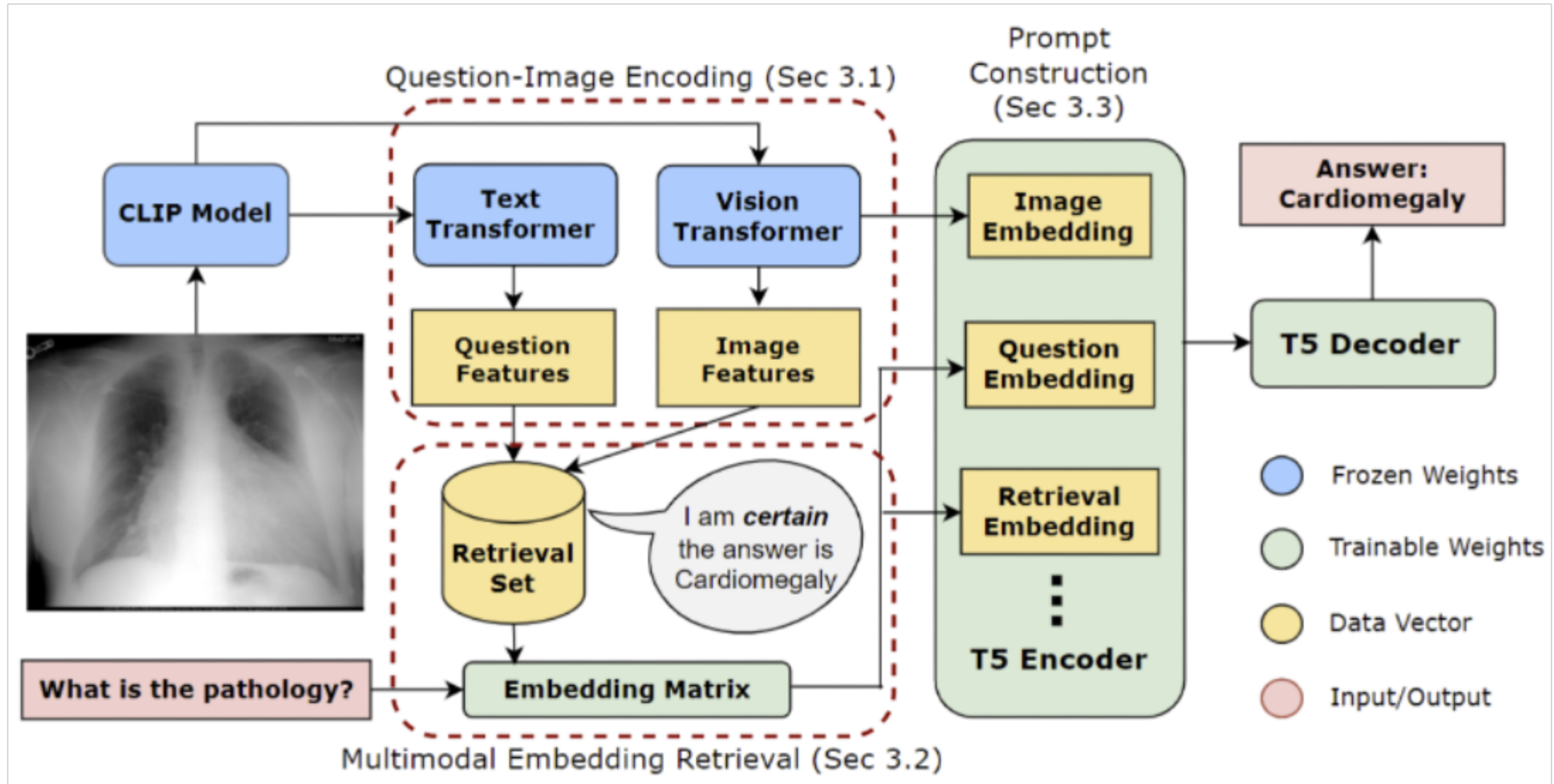
- Can we extend this to multimodality, e.g., visual QA tasks?
- Can we leverage retrievals for rapid adaptation to low-resourced domains, e.g., medicine?
- Can we let LMs automatically learn to balance the KNN and LM outputs?

Multimodal Prompt Retrieval (MPR)

- **Goal:** Adaptation to a new, similar VQA domain at test time
- **Data setting:** Train a VQA model on a source dataset, and adapt to a target dataset w/o fine-tuning at test time

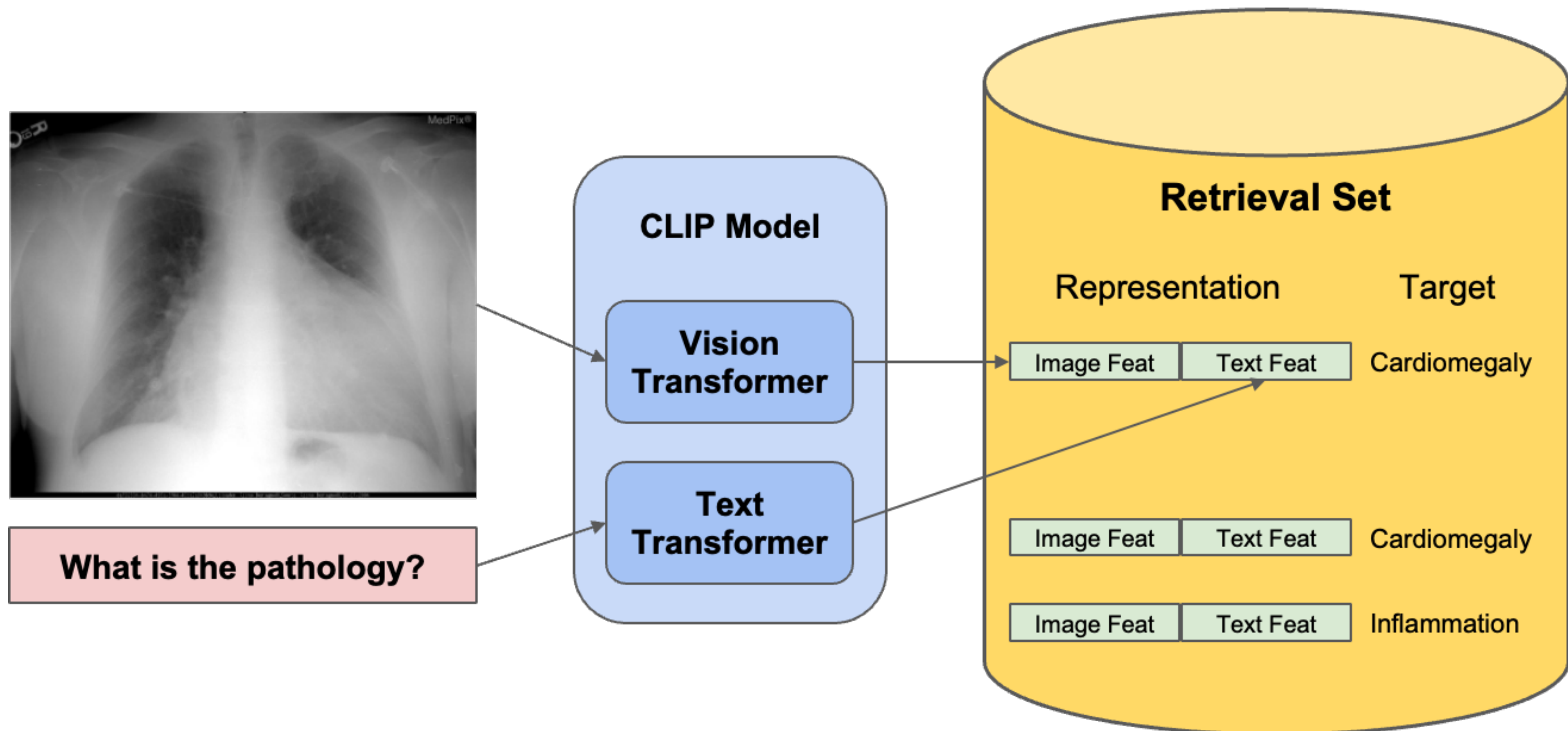


MPR Overview



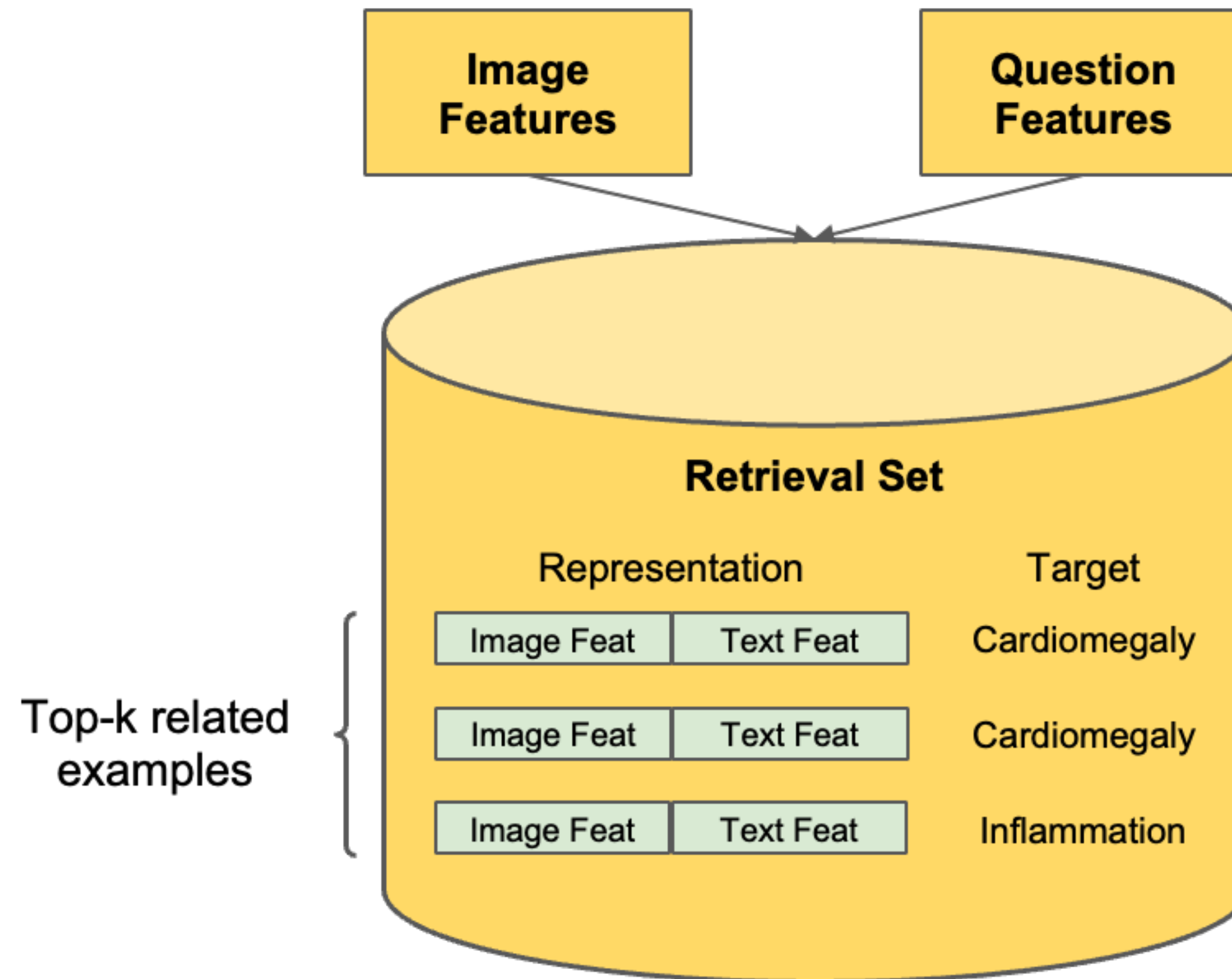
Construct a multimodal retrieval set

- Each image-question pair is encoded with a CLIP model and added to the retrieval set along with its corresponding answer



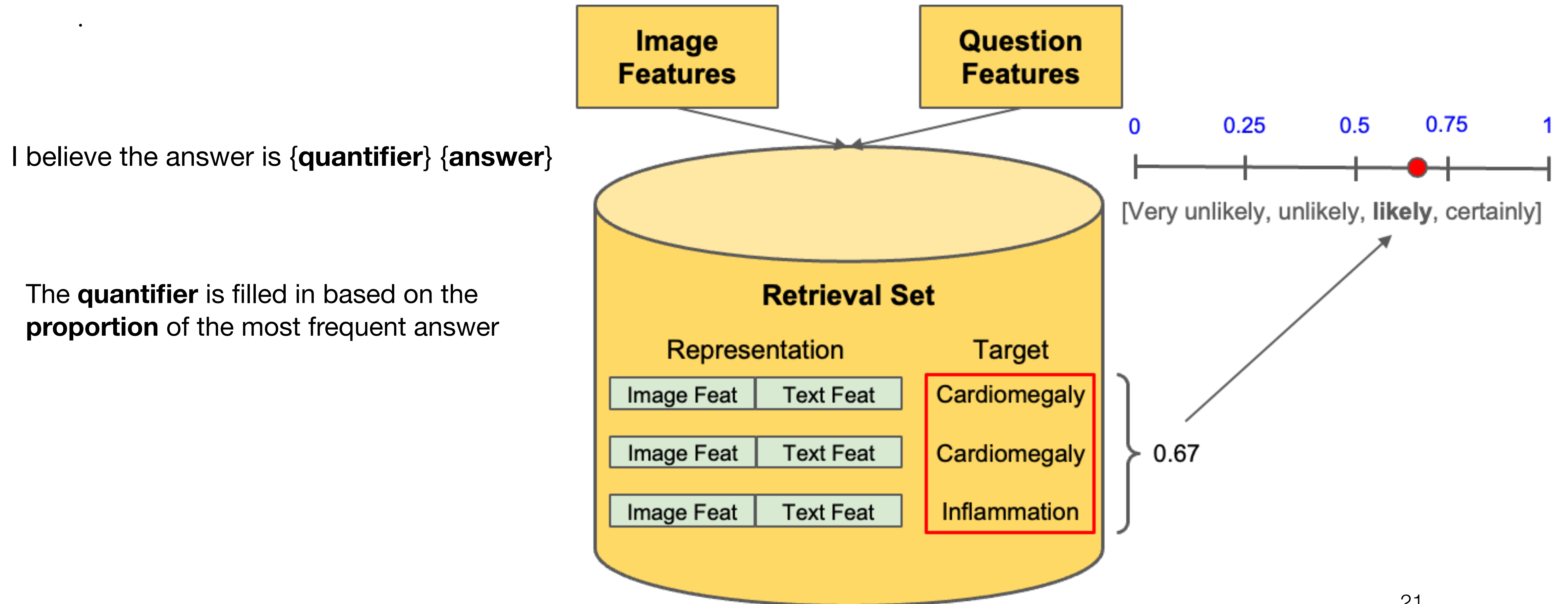
Multimodal Embedding Retrieval

- Given a question on an image, embed them and find the Top-K related examples in the retrieval set



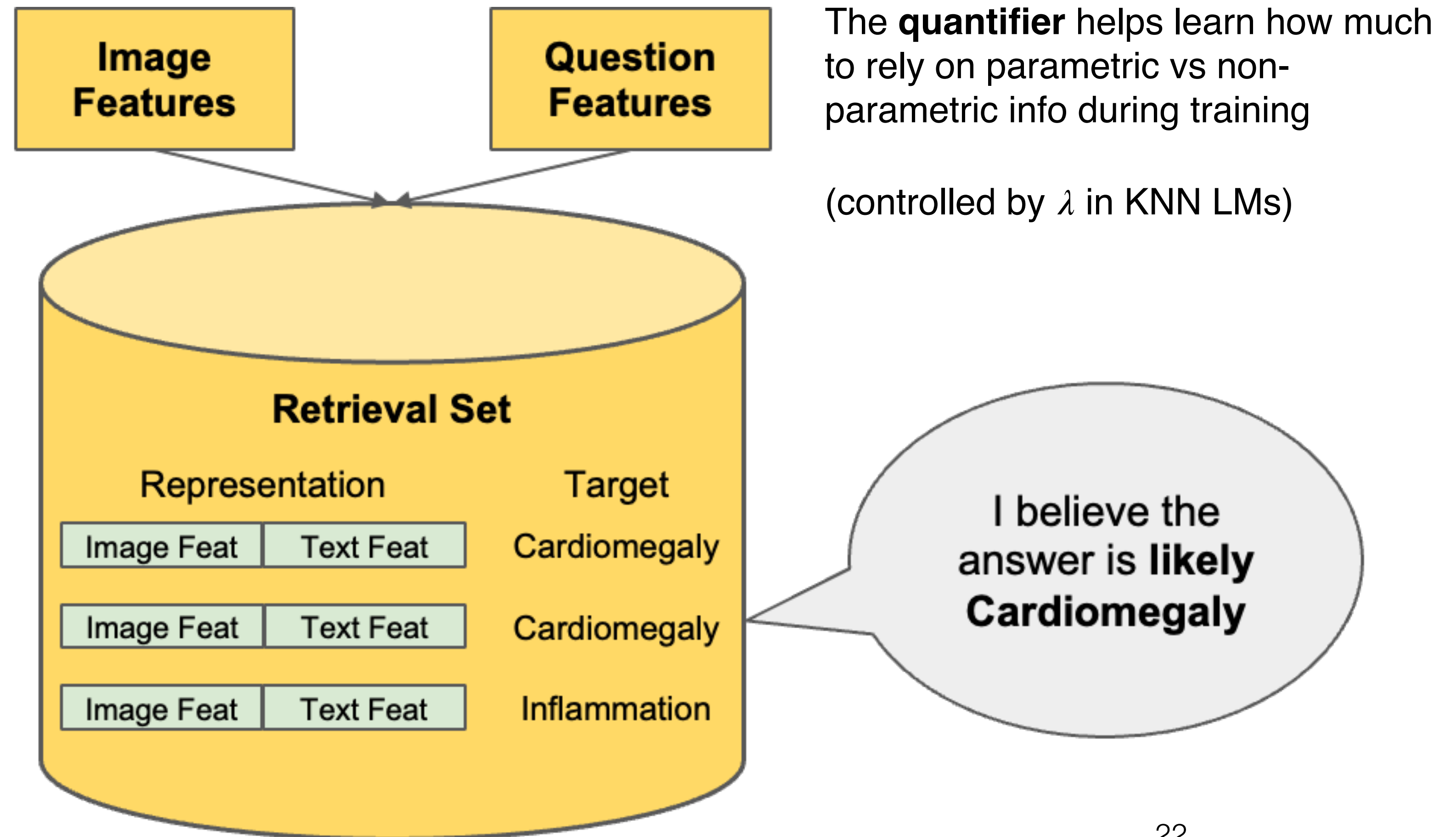
Retrieval Prompt Construction

- We use a **prompt template** to construct the prompt based on the retrieved examples:



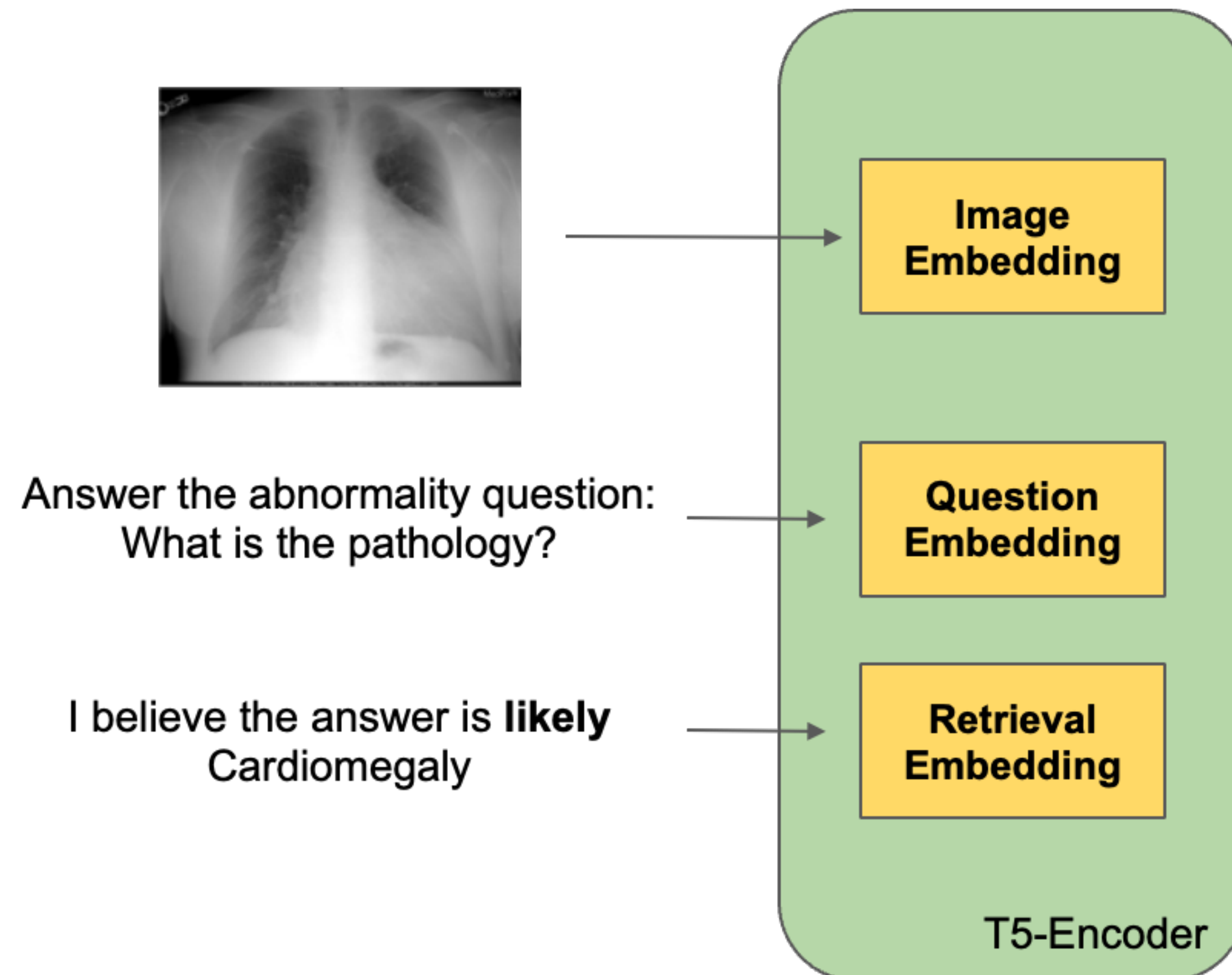
Retrieval Prompt Construction

- We use a **prompt template** to construct the prompt based on the retrieved examples:



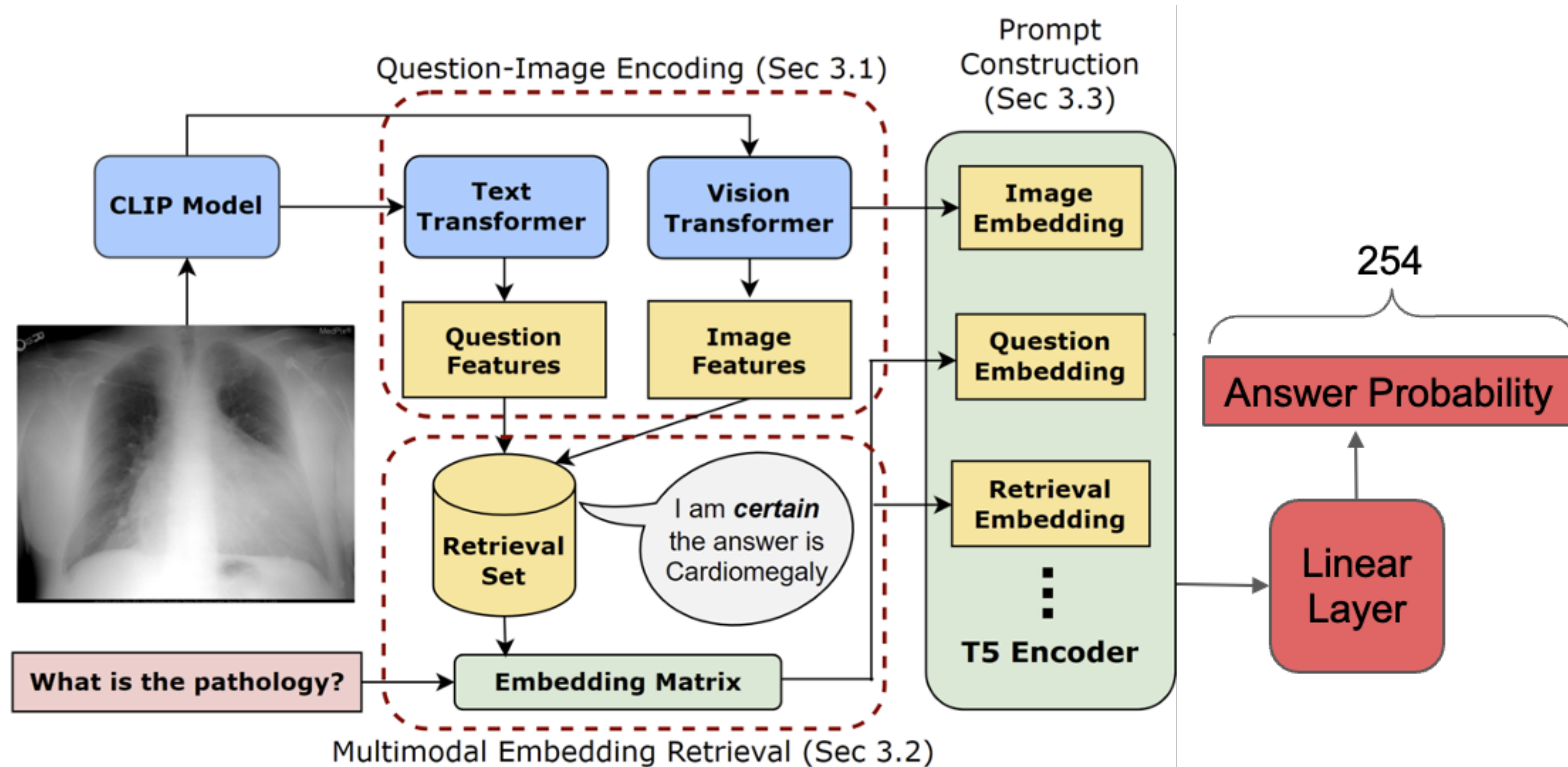
Prompt Construction & Encoding

- Feed the image, question and retrieval output as a prompt to the model for encoding
- Use a T5 decoder to generate the answer



Baseline

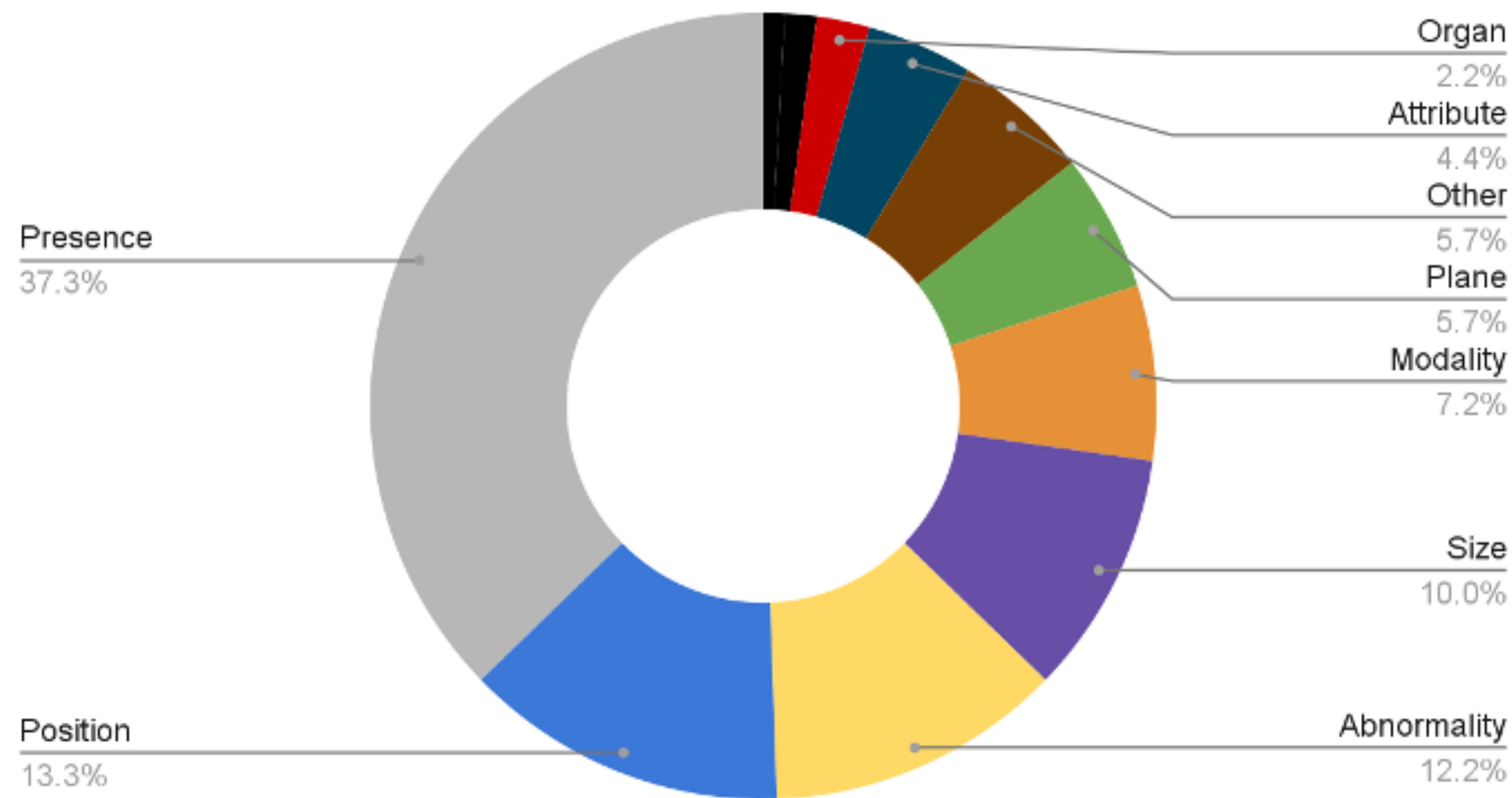
- Replace the T5 decoder with a linear layer to predict a fixed set of answers



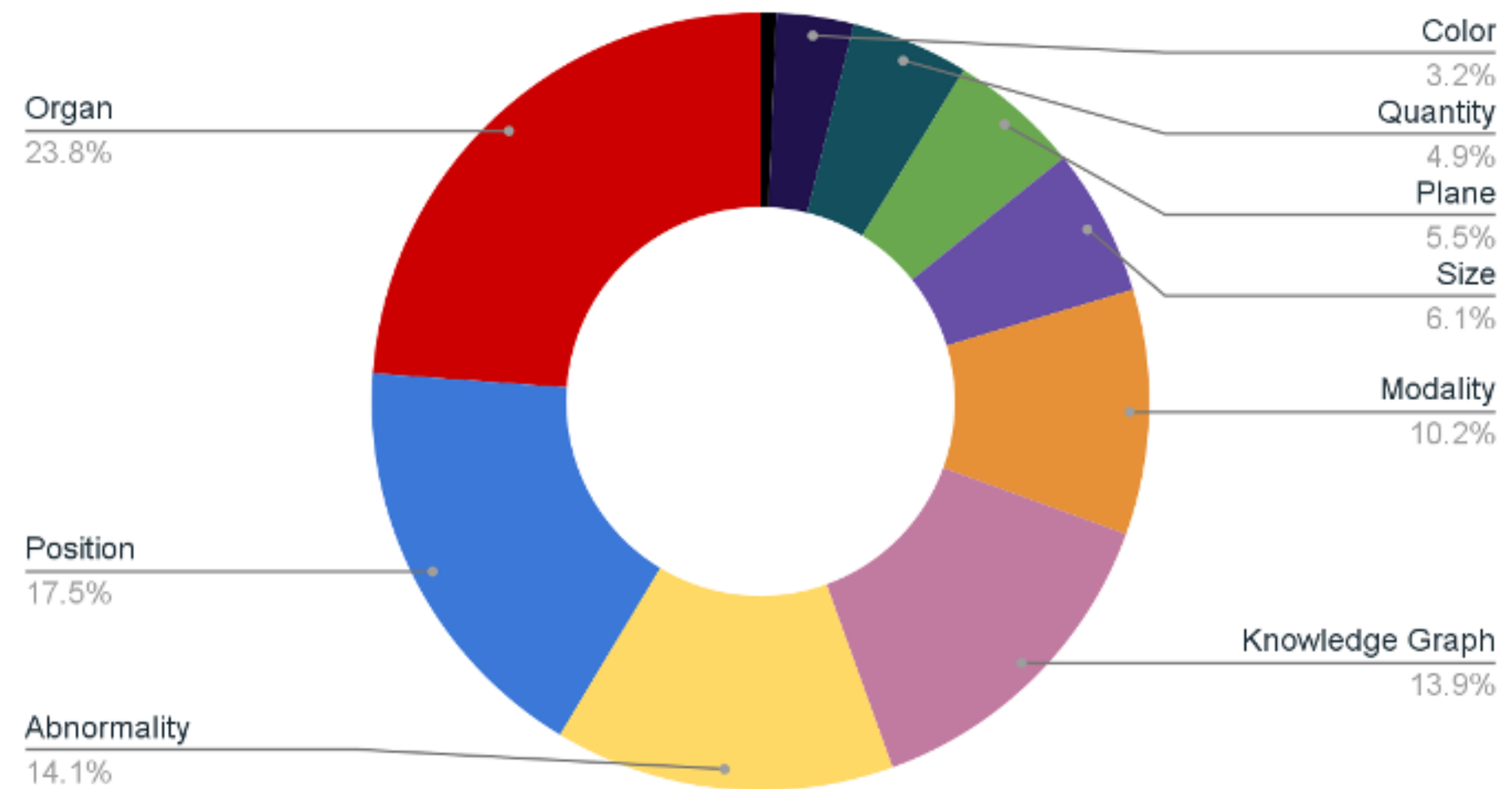
Datasets

- Two medical VQA datasets used as the source-target domain pairs

VQA-RAD



SLAKE



Domain Adaptation Setting

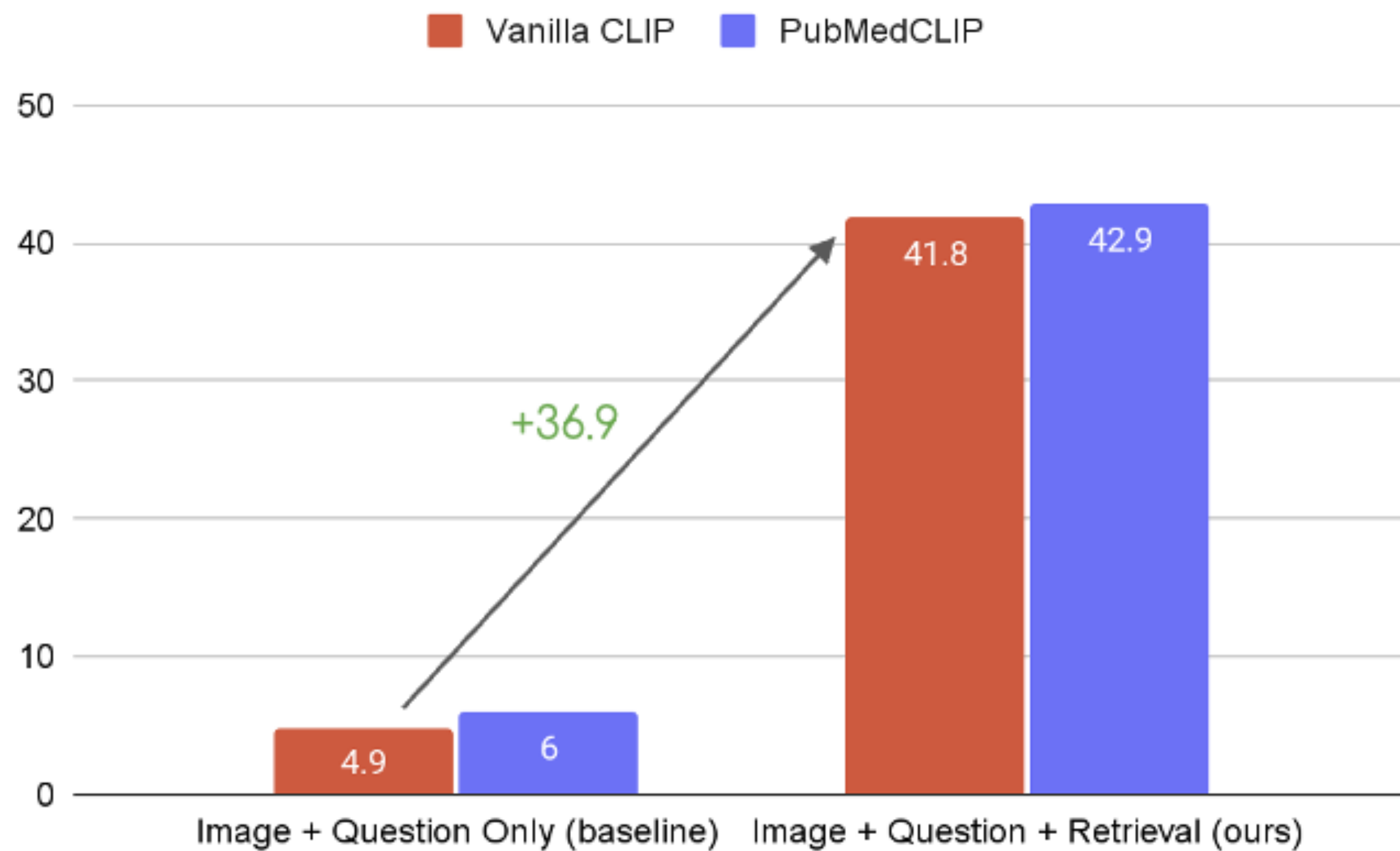
1. **In-context prediction:** Model is frozen when testing on target data
2. **Further fine-tuning:** Model is further fine-tuned w/ extra target data

For all evaluation settings, the retrieval set comes from the corresponding training dataset unless otherwise specified.

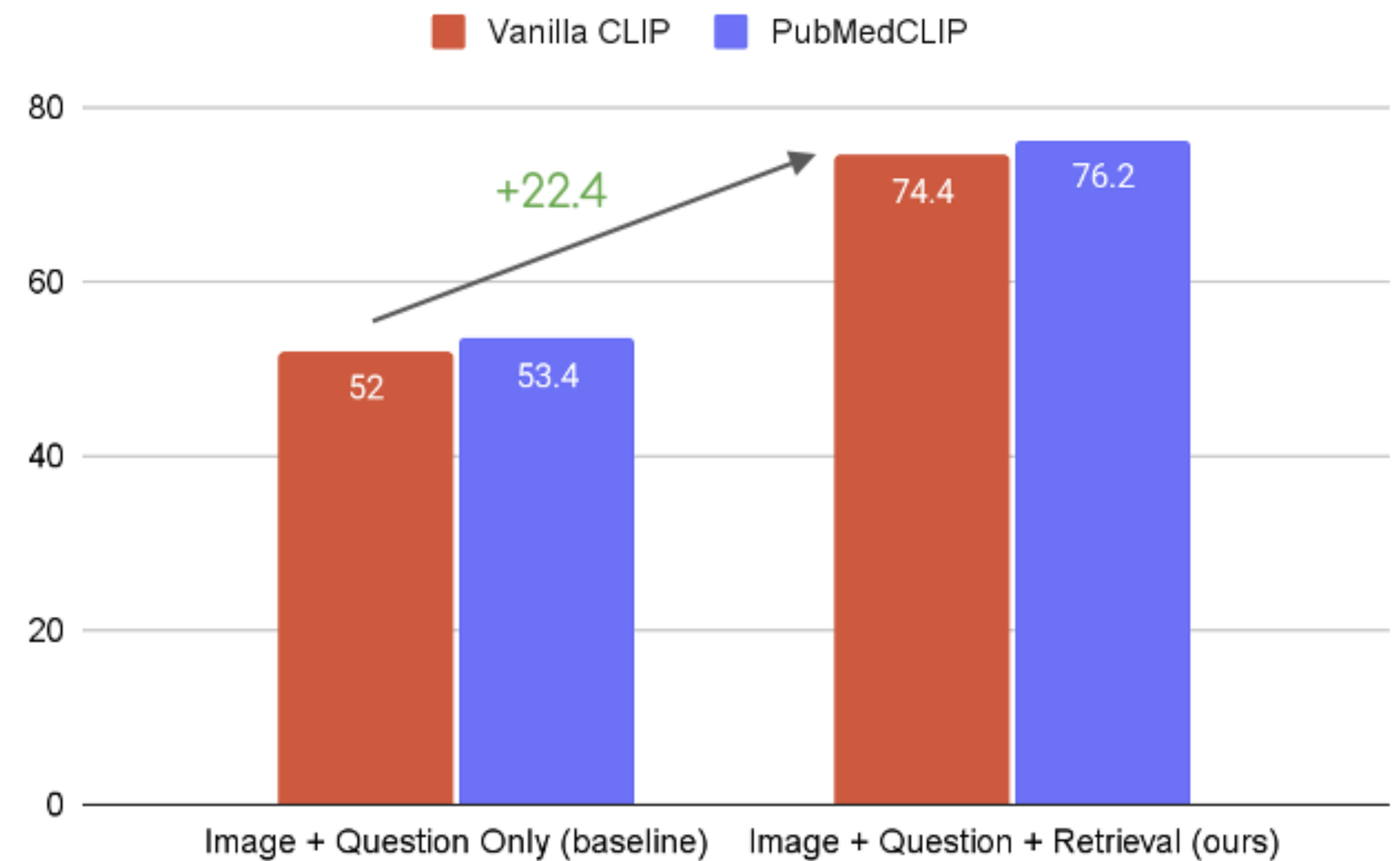
Domain Adaptation Performance

- Using **retrieval greatly** improves performance
- Using a domain-adapted vision encoder helps

SLAKE → VQA-RAD Performance (Open Questions)

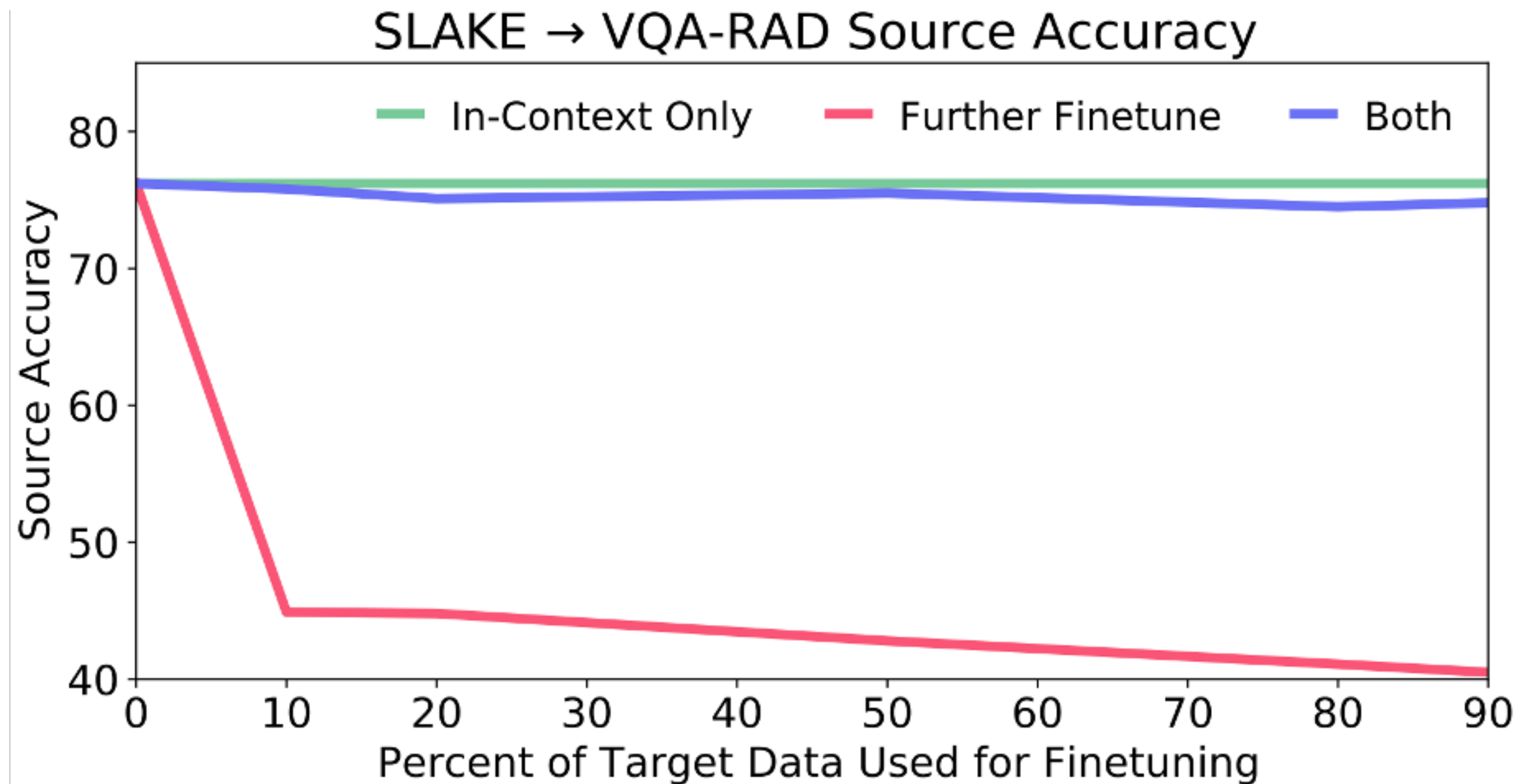


SLAKE → VQA-RAD Performance (Closed Questions)



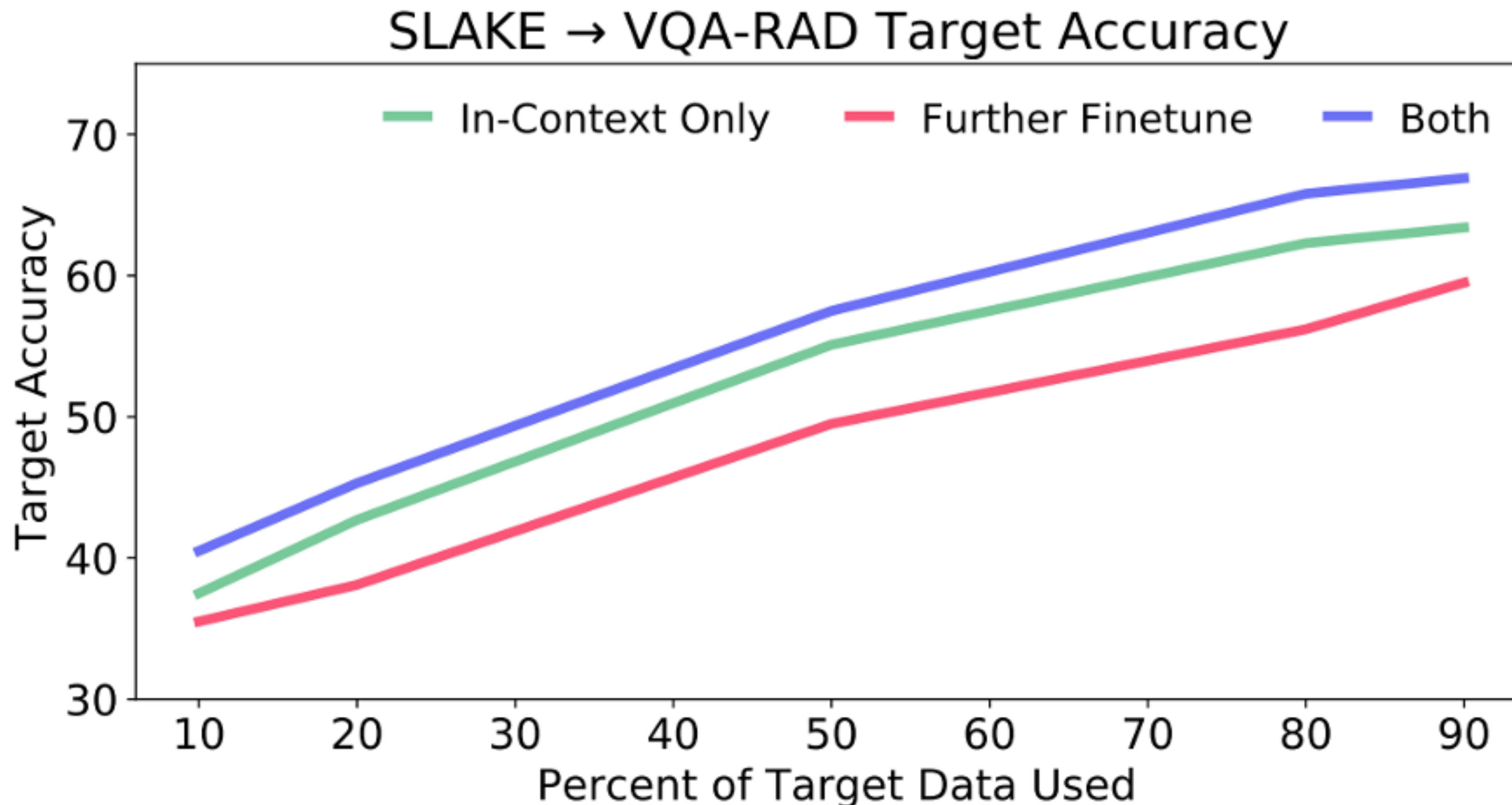
Further Fine-tuning v.s. In-Context Prediction

- **Takeaway:** Fine-tuning **encourages forgetting**. In-context prediction maintains stable performance



Further Fine-tuning v.s. In-Context Prediction

- **Takeaway:** Benefit from in-context prediction is **orthogonal** to benefit from fine-tuning



Structured-Augmented PLMs for Multi-hop QA

Single-hop QA

- Answers can be extracted directly from one single text sequence
 - Example: How many episodes in season 2 breaking bad?

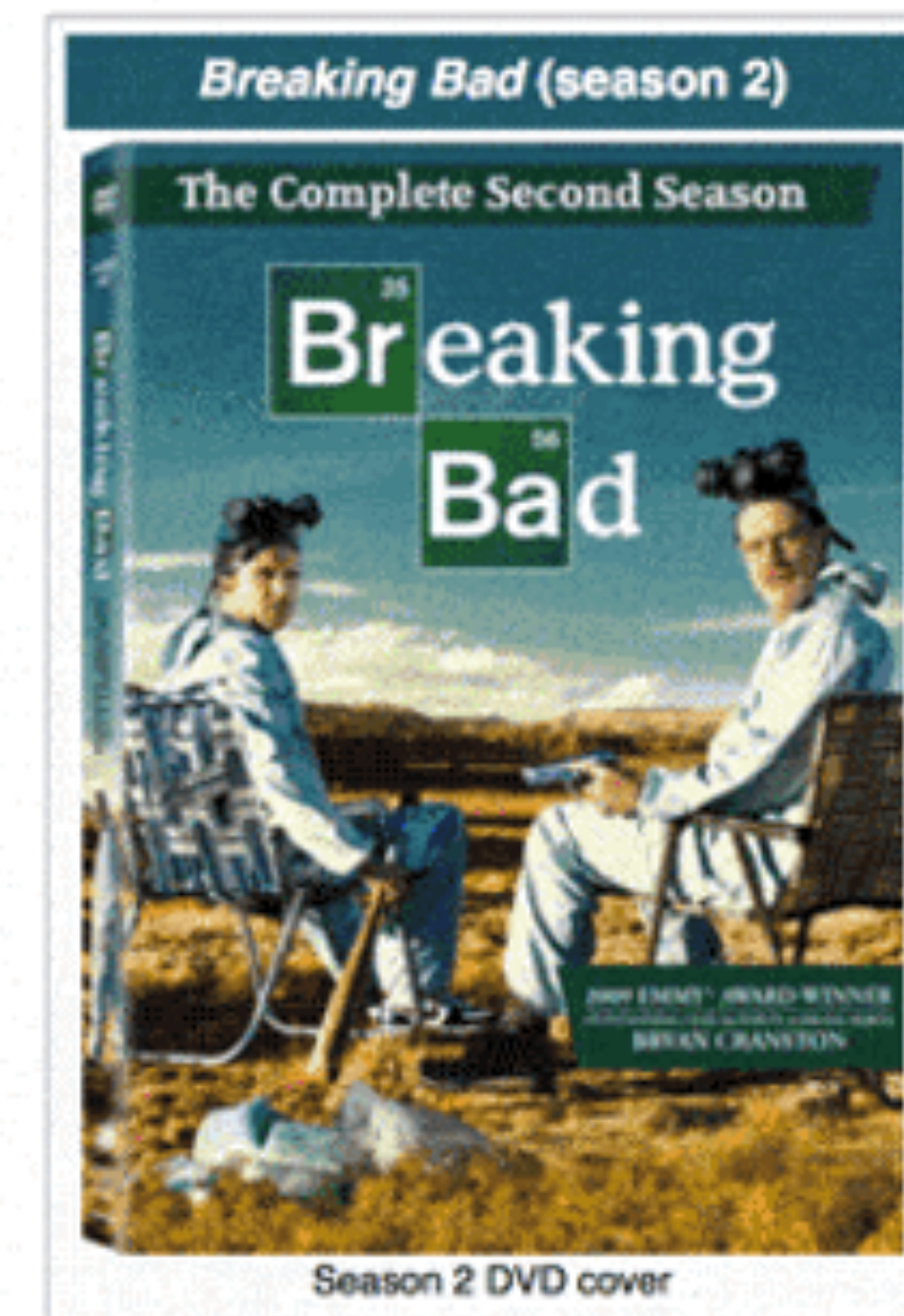
Breaking Bad (season 2)

From Wikipedia, the free encyclopedia

The second season of the American [television drama](#) series *Breaking Bad* premiered on March 8, 2009 and concluded on May 31, 2009. **It consisted of 13 episodes** each running approximately 47 minutes in length. AMC broadcast the second season on Sundays at 10:00 pm in the United States. The complete second season was released on Region 1 DVD and Region A Blu-ray on March 16, 2010.^[1]

Contents [\[hide\]](#)

- 1 [Cast](#)
 - 1.1 [Main](#)
 - 1.2 [Recurring](#)
- 2 [Episodes](#)
- 3 [Production](#)
- 4 [Music](#)
- 5 [Home video releases](#)
- 6 [Reception](#)
 - 6.1 [Reviews](#)
 - 6.2 [Awards and nominations](#)
- 7 [References](#)
- 8 [External links](#)



Multi-hop QA

- Answers can be derived after reading multiple text pieces

Q: What was the nickname of **Judy Lewis's** father ?

P1: Judy Lewis (born Judith Young; November 6, 1935 – November 25, 2011) was an American actress, writer, producer, and therapist. She was the secret biological daughter of actor **Clark Gable** and actress Loretta Young.

P1: William Clark Gable (February 1, 1901 – November 16, 1960) was an American film actor, often referred to as "**The King of Hollywood**". He had roles in more than 60 motion pictures in a wide variety of genres during a career that lasted 37 years...

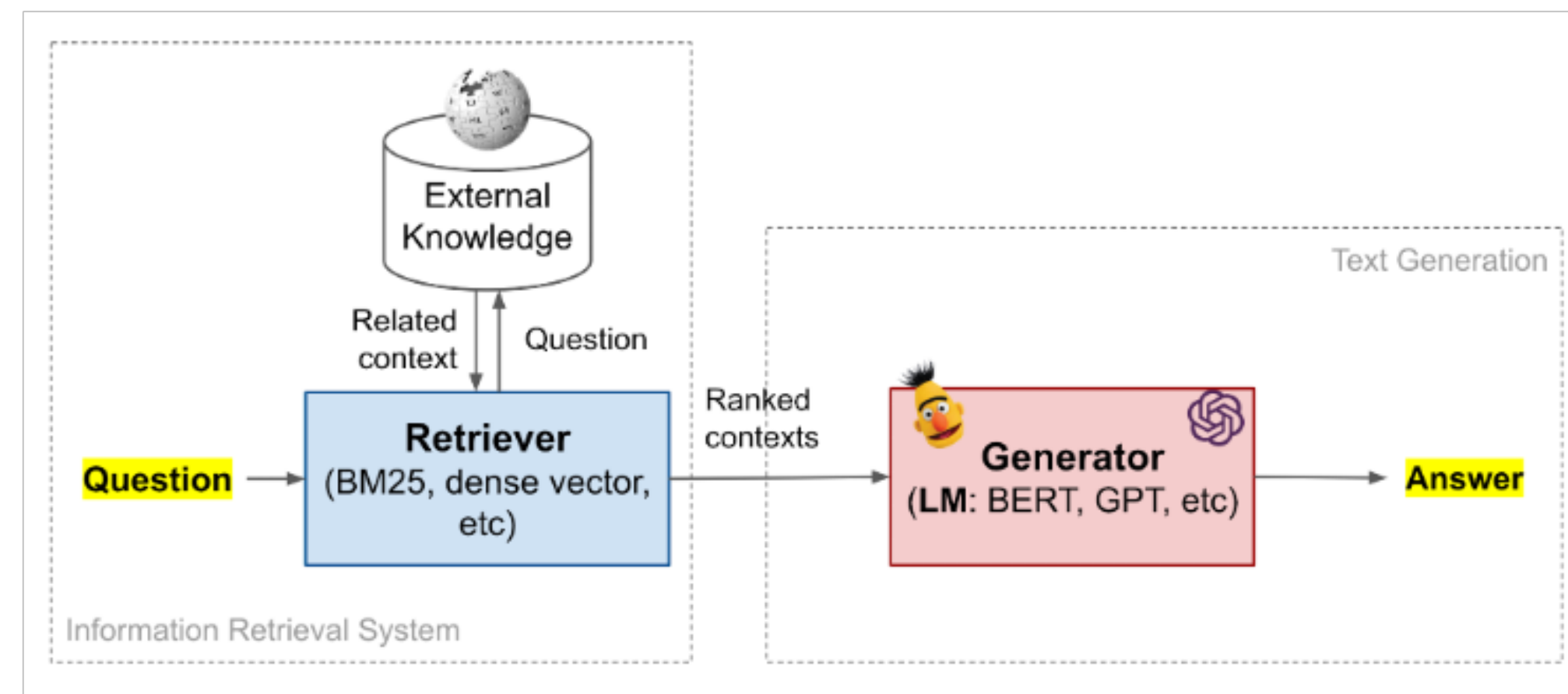
HotpotQA (2 hops)

| Graph | Question | Decomposition |
|-------|--|---|
| | Who succeeded the first President of Namibia? Hifikepunye Pohamba | <ol style="list-style-type: none"> Who was the first President of Namibia? Sam Nujoma Who succeeded Sam Nujoma? Hifikepunye Pohamba |
| | What currency is used where Billy Giles died? pound sterling | <ol style="list-style-type: none"> At what location did Billy Giles die? Belfast What part of the UK is Belfast located in? Northern Ireland What is the unit of currency in Northern Ireland? pound sterling |
| | When was the first establishment that McDonaldization is named after, open in the country Horndean is located? 1974 | <ol style="list-style-type: none"> What is McDonaldization named after? McDonald's Which state is Horndean located in? England When did the first McDonald's open in England? 1974 |
| | When did Napoleon occupy the city where the mother of the woman who brought Louis XVI style to the court died? 1805 | <ol style="list-style-type: none"> Who brought Louis XVI style to the court? Marie Antoinette Who's mother of Marie Antoinette? Maria Theresa In what city did Maria Theresa die? Vienna When did Napoleon occupy Vienna? 1805 |
| | How many Germans live in the colonial holding in Aruba's continent that was governed by Prazeres's country? 5 million | <ol style="list-style-type: none"> What continent is Aruba in? South America What country is Prazeres? Portugal Colonial holding in South America governed by Portugal? Brazil How many Germans live in Brazil? 5 million |
| | When did the people who captured Malakoff come to the region where Philipsburg is located? 1625 | <ol style="list-style-type: none"> What is Philipsburg capital of? Saint Martin Saint Martin is located on what terrain feature? Caribbean Who captured Malakoff? French When did the French come to the Caribbean? 1625 |

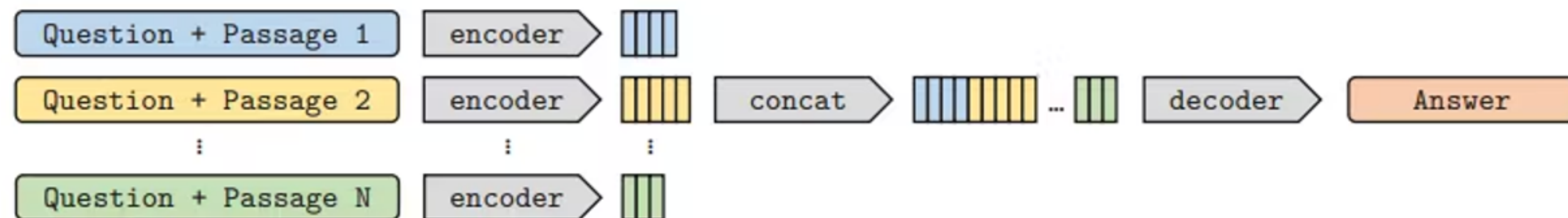
MuSiQue (2,3,4 hops)

Fusion in Decoder (FiD)

- Retrieve text passages given a question
- Encode each passage (with the question) separately, concatenate them for decoding



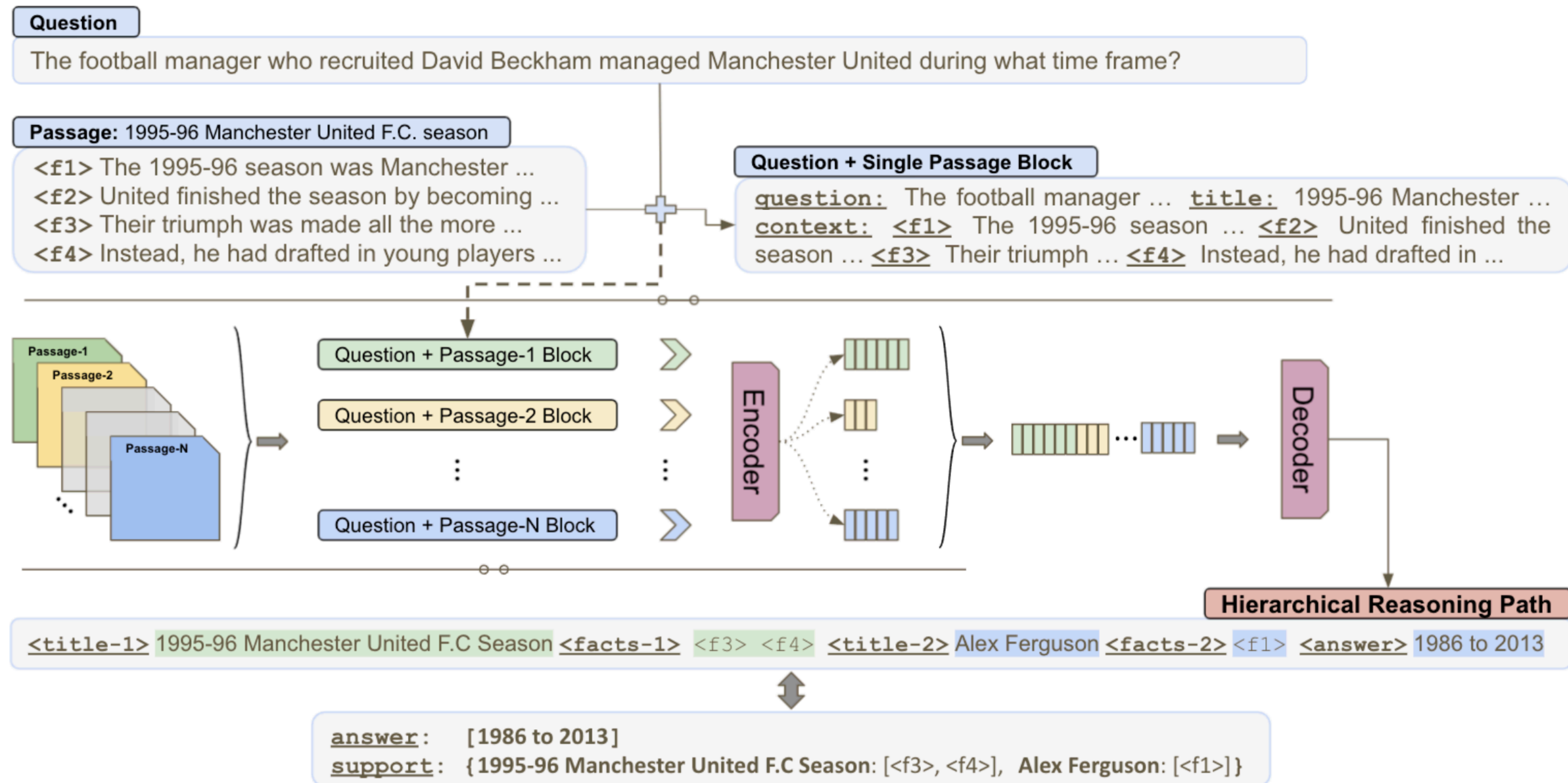
Retriever + Generator Architecture



Detailed Implementations

PathFiD

- Predict not only the final answer but also the reasoning path



Disconnected Reasoning Path

- The predicted reasoning path may not be connected!

Q: Who is the executive producer of the film that has a score composed by Jerry Goldsmith?

C1: The iconic, avant-garde score to the film "Alien" was composed by Jerry Goldsmith

C2: Alien is a 1979 science-fiction horror film directed by Ridley Scott ...

C2: Dan O'Bannon, ..., wrote the screenplay from a story he co-authored with Ronald Shusett.

C3: Shusett was executive producer.

Unstructured Texts $c_{1:n}$

PathFiD: L.A. Confidential -> ... -> Lionheart (1987 film) -> ... -> Steven Spiel

Incorrect final answer

No direct relation link between the two text spans in the context passages

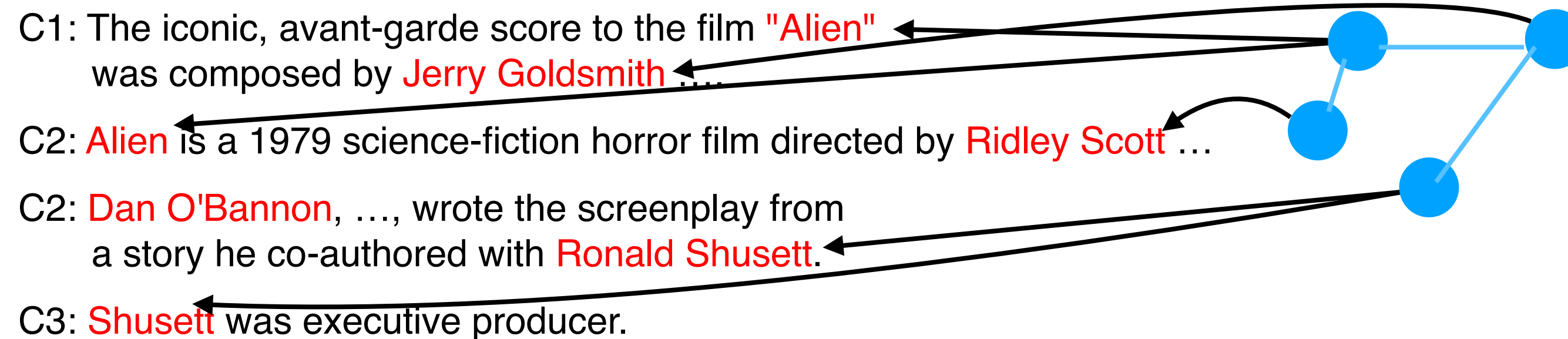
SeqGraph: Integrating structures to prompt LLMs

- Given a text query q , we have access to a search engine to retrieve a set of related text passages $c_{1:n} = \{c_1, \dots, c_n\}$ and an entity graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ constructed from the retrieved unstructured texts. We aim to use a LLM to estimate:

Reasoning probability $p(r|q, c_{1:n}, \mathcal{G})$

Answer probability $p(a|q, c_{1:n}, \mathcal{G}, r)$

Q: Who is the executive producer of the film that has a score composed by Jerry Goldsmith?



Unstructured Texts $c_{1:n}$

Structured Graph \mathcal{G}

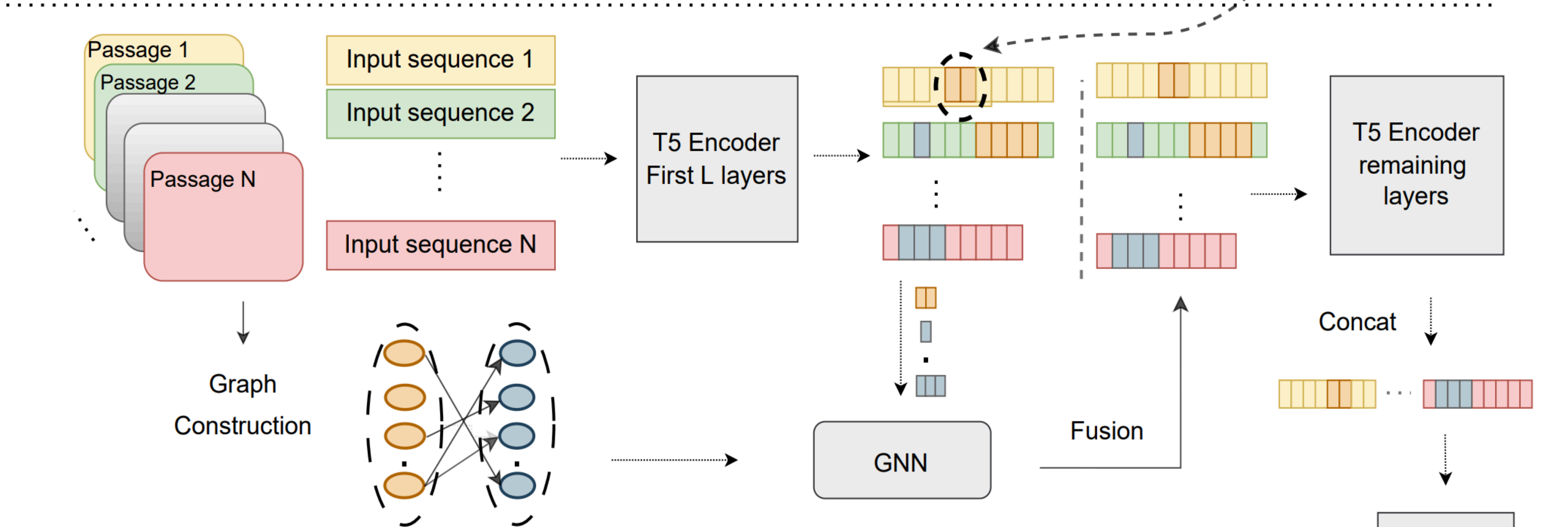
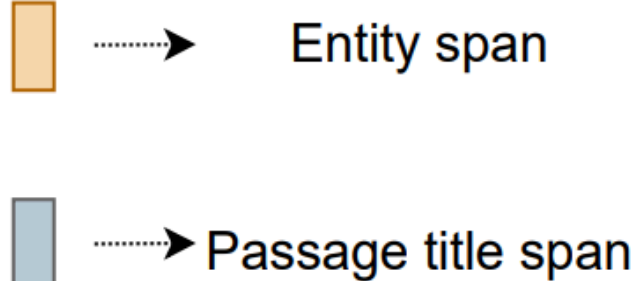
SeqGraph Workflow

Question:

"An American Werewolf in Paris was a partial sequel to the comedy film starring whom?"

Passage 1: An American Werewolf in Paris
 [f1] It follows the ... An American Werewolf in London
 [f2] The film is a... the United States and France.

Input Sequence 1
Question: An American ... whom? **Title:** An American Werewolf ... **Context:** [f1]: It follows ... [f2] The film is a ... the United States and France.



Reasoning path

[title-1] An American Werewolf in Paris [facts-1] [f1] [title-2] An American Werewolf in London [facts-2] [f2] [answer] David Naughton, Jenny Agutter and Griffin Dunne"

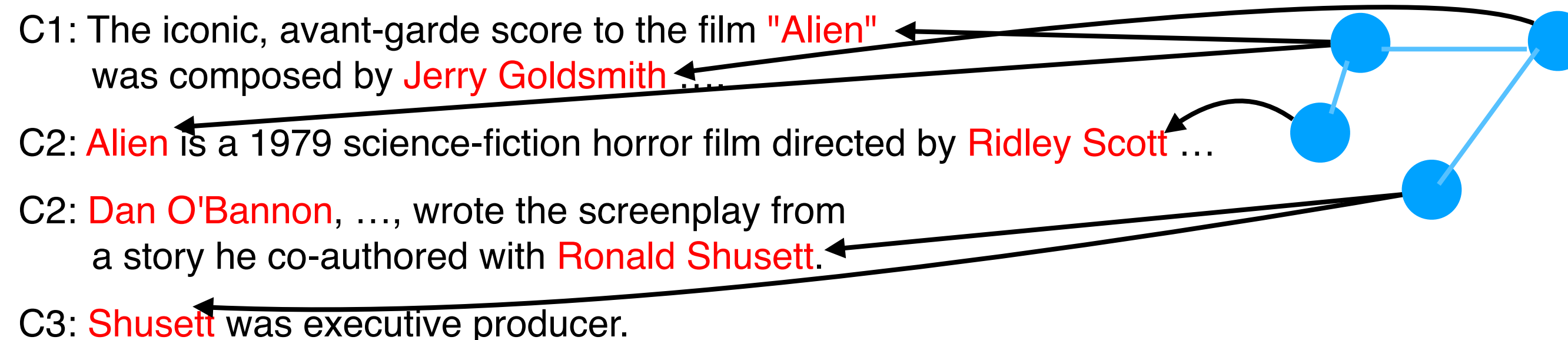
SeqGraph: Integrating structures to prompt LLMs

- Given a text query q , we have access to a search engine to retrieve a set of related text passages $c_{1:n} = \{c_1, \dots, c_n\}$ and an entity graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ constructed from the retrieved unstructured texts. We aim to use a LLM to estimate:

Reasoning probability $p(r|q, c_{1:n}, \mathcal{G})$

Answer probability $p(a|q, c_{1:n}, \mathcal{G}, r)$

Q: Who is the executive producer of the film that has a score composed by Jerry Goldsmith?



Unstructured Texts $c_{1:n}$

Structured Graph \mathcal{G}

SeqGraph: Integrating structures to prompt LLMs

- Different from existing CoT prompting strategies, we apply structures to prompt LLM

$$\text{Manual CoT: } a, r = \max_{a,r} p(a, r | q^{\text{demo}}, c_{1:n}^{\text{demo}}, r^{\text{demo}}, a^{\text{demo}}, q, c_{1:n})$$

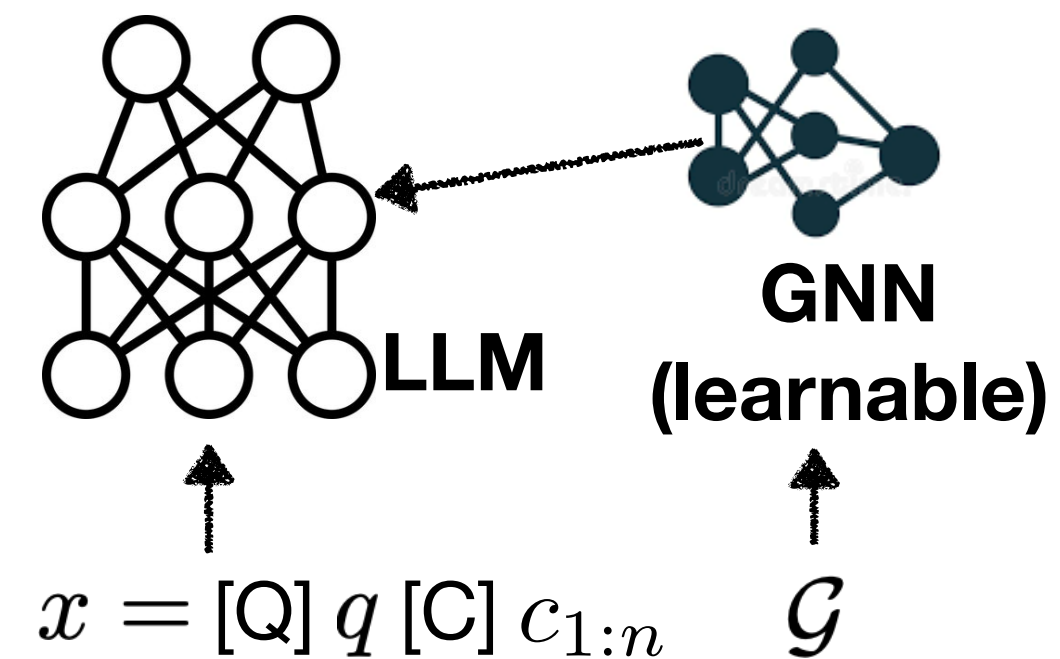
$$\text{Zero-shot CoT: } r = \max_r p(r | q, c_{1:n}), \quad a = \max_a p(a | q, c_{1:n}, r)$$

$$\text{Structure CoT (ours): } r = \max_r p(r | q, c_{1:n}, \mathcal{G}), \quad a = \max_a p(a | q, c_{1:n}, \mathcal{G}, r)$$

- Use a graph neural network to encode **neural symbolic representations** which are easily integrated with **distributed language representations** in large language models

- Then estimate: $p(r | q, c_{1:n}, \mathcal{G}; \theta_{\text{LLM}}, \theta_{\text{GNN}})$

$$p(a | q, c_{1:n}, \mathcal{G}, r; \theta_{\text{LLM}}, \theta_{\text{GNN}})$$



Structured CoT Tuning

- Maximum likelihood estimation:

$$\mathcal{L}_{\text{nl}}(\mathcal{D}, \theta_{\text{LLM}}, \theta_{\text{GNN}}) = -\mathbb{E}_{\mathcal{D}} \log p(q, c_{1:n}, \mathcal{G}, r, a; \theta_{\text{LLM}}, \theta_{\text{GNN}})$$

Fusion of Text and Graph Representations

- Concatenate a question with each retrieved context passage and its title

$$S_i := [\text{Question}] q [\text{Title}] t_i [\text{Content}] c_i$$

- Construct a graph of **entity-title links** from retrieved passages, and then embed text and graph

$$\mathbf{Z}_i^L = \text{TextEncoder}(S_i, L)$$

$$\mathbf{Z}^G = \text{GraphEncoder}(\mathbf{N}, \mathcal{G})$$

- Representation fusion and further encoding

$$\mathbf{S}_i = \text{TextEncoder}(\mathbf{Z}_i^G \oplus \mathbf{Z}_i^L, M - L)$$

- Feed all context passages to a T5 decoder for predicting a reasoning path and a final answer

Multi-step Reasoning with and without Structures

- Evaluate on two multi-step reasoning QA datasets
- Obtain state-of-the-art performance on the Musique dataset

| Model | HOTPOT-QA | | | | MUSIQUE | | | |
|-----------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| | Answer | | Support | | Answer | | Support | |
| | EM | F1 | EM | F1 | EM | F1 | EM | F1 |
| FID-Base | 61.84 | 75.20 | - | - | 29.38 | 39.97 | - | - |
| PATH-FID-Base | 62.03 | 75.69 | 60.45 | 86.00 | 34.71 | 44.93 | 57.30 | 80.18 |
| SEQGRAPH-Base | 64.19 | 77.60 | 62.44 | 87.72 | 37.36 | 47.11 | 58.05 | 80.39 |
| FID-Large | 65.59 | 79.39 | - | - | 36.04 | 46.66 | - | - |
| PATH-FID-Large* | 65.80 | 78.90 | 59.30 | 85.70 | - | - | - | - |
| PATH-FID-Large | 65.33 | 79.00 | 61.52 | 86.88 | 42.28 | 53.86 | 62.14 | 82.45 |
| SEQGRAPH-Large | 66.51 | 81.62 | 63.24 | 88.28 | 46.01 | 56.88 | 65.12 | 83.65 |

- **Take-away: structured contexts** (encoded by a graph neural network) **improved the multi-step reasoning process**

Ongoing Research Directions

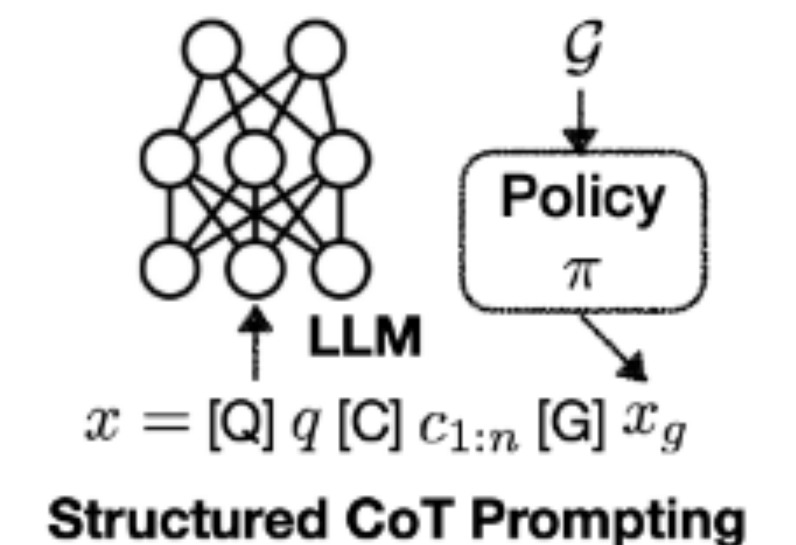
LLMs are powerful, but new challenges arise

- Representation learning under the parameter-efficient learning framework where only a small additional prompt parameters θ_{Prompt} are updated.

$$p(X; \theta_{\text{LLM}}, \theta_{\text{Prompt}})$$

- Prompt engineering requires huge human efforts. Can we automate this process?

$$\max_{\pi} \mathbb{E}_{\pi} R(a', a), \text{ where } a' = \max_a p(a|q, c_{1:n}, x_g, r), \quad r = \max_r p(r|q, c_{1:n}, x_g), \quad x_g \sim \pi(\mathcal{G}, q, c_{1:n}).$$



- Theoretical understanding of in-context learning: How can LLMs learn from in-context examples? What's the sample complexity for LLMs to learn from few-shot demonstrations?

Thanks! QA

Key take-away:

- LLMs as a generative model can be augmented with external multimodal data and structures
- More information can be found in my website: <https://junjiehu.github.io/>