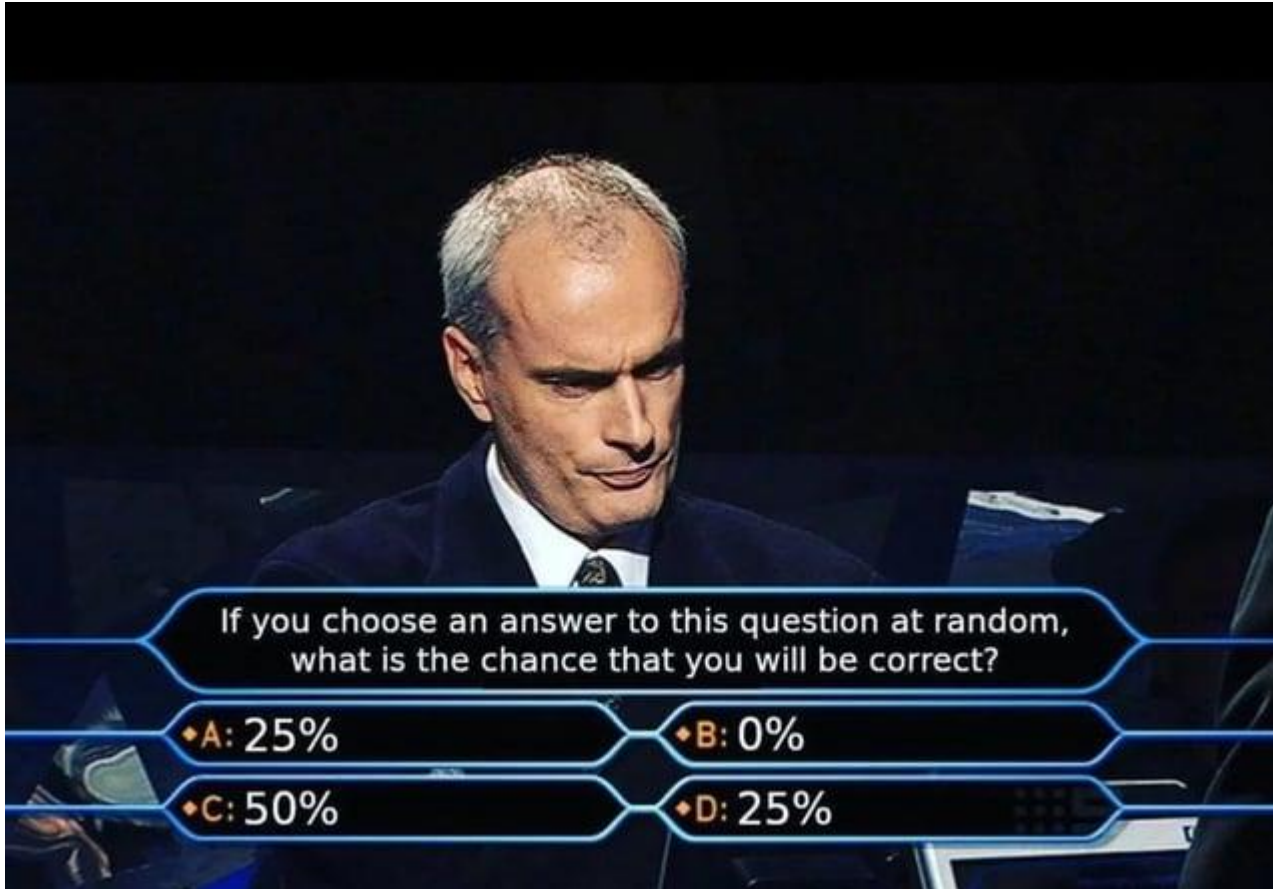


# Question Answering

CSC485



# Annoucement

- Last Lecture:
- Tuesday:
  - 10 – 11
  - 12 – 13

# Modern QA from text

The common person's view? [From a novel]

“I like the Internet. Really, I do. Any time I need a piece of shareware or I want to find out the weather in Bogota ... I'm the first guy to get the modem humming. But **as a source of information, it sucks**. You got **a billion pieces of data**, struggling to be heard and seen and downloaded, and **anything I want to know seems to get trampled underfoot in the crowd.**”

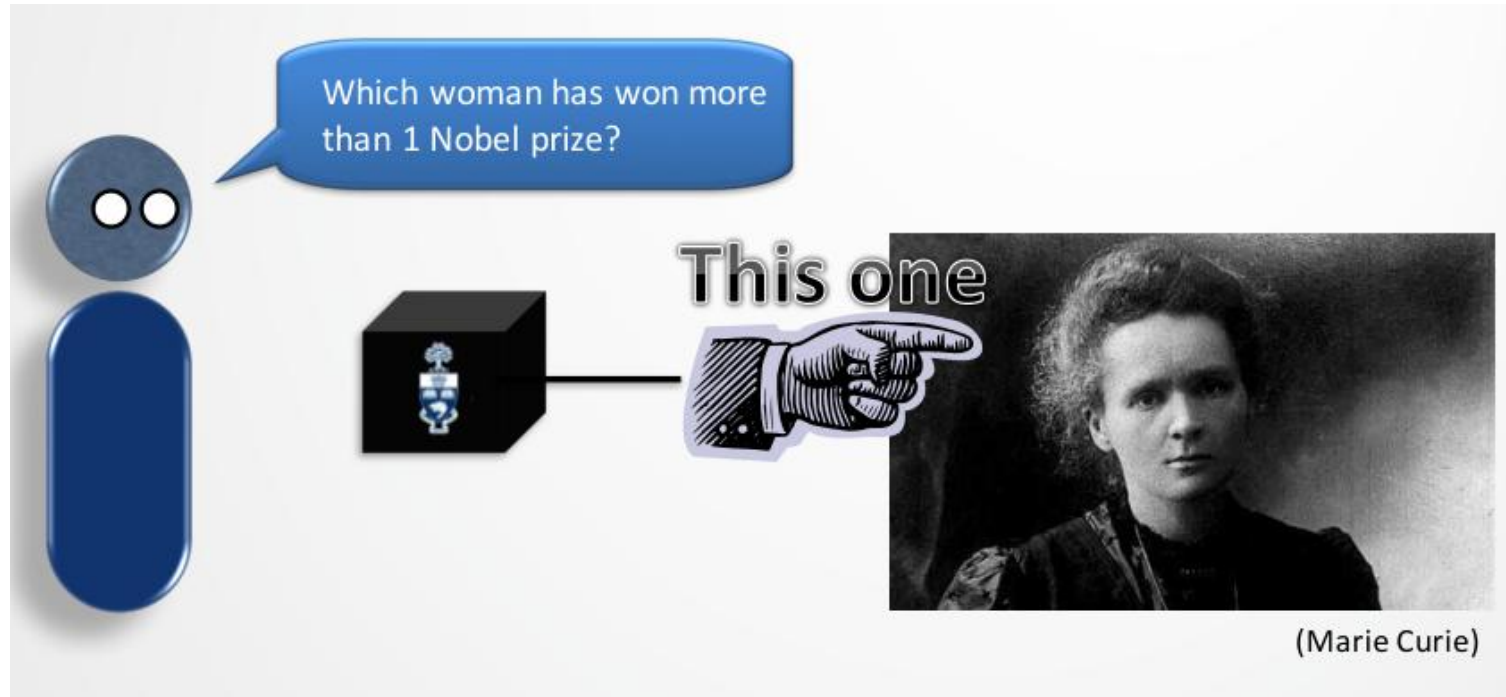
M. Marshall. The Straw Men. HarperCollins Publishers, 2002.

- An idea originating from the IR community.
- With massive collections of full-text documents, simply finding **relevant documents** is of limited use: we want **answers** from textbases.
- QA: give the user a (short) answer to their question, perhaps supported by evidence.

# Outline

- Intro to QA
- QA & IR before deep learning
- QA & IR with deep learning
- RAG: QA with LLM
  - More LLM stuff: post-training & prompt-engineering.

# Question Answering (QA)



- Question Answering (QA) usually involves a specific answer to a question.

# Information Retrieval (IR) and QA

A screenshot of a Google search results page for the query "which woman has won more than 1 nobel prize?". The search bar at the top shows the query. Below the search bar, there are tabs for "All", "Images", "News", "Videos", "Shopping", "Web", "Books", and "More". The "All" tab is selected. The search results are displayed in a list format. The first result is from Wikipedia, titled "List of female Nobel laureates", with a snippet stating "Curie is also the first person and the only woman to have won multiple Nobel Prizes; in 1911, she won the Nobel Prize in Chemistry." The second result is also from Wikipedia, titled "Nobel Prize", with a snippet stating "Multiple laureates Five people have received two Nobel Prizes. Marie Curie received the Physics Prize in 1903 for her work on radioactivity and the Chemistry ...". The third result is from Rincón educativo, titled "The magnificent four who repeated Nobel", with a snippet stating "By Elena Sanz - The first person in history to achieve the feat of receiving a double Nobel was the Polish Marie Skłodowska Curie, laureate first in Physics and, ...". The fourth result is from Phys.org, titled "The five scientists who won two Nobel prizes", with a snippet stating "Oct 5, 2022 — Marie Curie (1903, 1911) The mother of modern physics was the first woman ever to win not one, but two, Nobel prizes for her seminal ...". The fifth result is from The Conversation, titled "The five scholars who won two Nobel prizes", with a snippet stating "Jul 9, 2024 — Marie Curie is the most famous of these five scholars and for good reason. The world today, as well as science in general, is different because ...". The sixth result is from Statista, titled "Chart: The Nobel Prize Gender Gap", with a snippet stating "https://www.statista.com > ... > Global status of women > ...".

A screenshot of a Google search results page for the query "which woman has won more than 1 nobel prize?". The search bar at the top shows the query. Below the search bar, there are tabs for "All", "Images", "News", "Videos", "Shopping", "Web", "Books", and "More". The "All" tab is selected. The search results are displayed in a list format. The first result is from Wikipedia, titled "List of female Nobel laureates", with a snippet stating "Curie is also the first person and the only woman to have won multiple Nobel Prizes; in 1911, she won the Nobel Prize in Chemistry." The second result is also from Wikipedia, titled "Nobel Prize", with a snippet stating "Multiple laureates Five people have received two Nobel Prizes. Marie Curie received the Physics Prize in 1903 for her work on radioactivity and the Chemistry ...". The third result is from Rincón educativo, titled "The magnificent four who repeated Nobel", with a snippet stating "By Elena Sanz - The first person in history to achieve the feat of receiving a double Nobel was the Polish Marie Skłodowska Curie, laureate first in Physics and, ...". The fourth result is from Phys.org, titled "The five scientists who won two Nobel prizes", with a snippet stating "Oct 5, 2022 — Marie Curie (1903, 1911) The mother of modern physics was the first woman ever to win not one, but two, Nobel prizes for her seminal ...". The fifth result is from The Conversation, titled "The five scholars who won two Nobel prizes", with a snippet stating "Jul 9, 2024 — Marie Curie is the most famous of these five scholars and for good reason. The world today, as well as science in general, is different because ...". The sixth result is from Statista, titled "Chart: The Nobel Prize Gender Gap", with a snippet stating "https://www.statista.com > ... > Global status of women > ...".

One strategy is to turn QA into information retrieval (IR) and let the human complete the task.

# Question Answering (QA)



A screenshot of the WolframAlpha website. The header features the WolframAlpha logo with the tagline "computational knowledge engine". A search bar contains the query: "How much potassium is in 450,000 cubic kilometers of bananas?". Below the search bar, the "Input interpretation:" section shows the query broken down into components: "banana", "amount", "450 000 km<sup>3</sup> (cubic kilometers)", and "potassium". The "Result:" section displays the answer: "1.5 × 10<sup>12</sup> t (metric tons)".

WolframAlpha™ computational knowledge engine

How much potassium is in 450,000 cubic kilometers of bananas?

Input interpretation:

banana	amount	450 000 km <sup>3</sup> (cubic kilometers)	potassium
--------	--------	--	-----------

Result:

1.5 × 10<sup>12</sup> t (metric tons)


# Knowledge-based QA



1. Build a structured semantic representation of the query.
  - Extract times, dates, locations, entities using regular expressions.
  - Fit to well-known templates.
2. Query databases with these semantics.
  - Ontologies (Wikipedia infoboxes).
  - Restaurant review databases.
  - Calendars.
  - ...





# IR-based QA




which woman has won more than 1 nobel prize?

×







All

Images

News

Videos

Shopping

Forums


Web


⋮ More

Tools

## Marie Curie

Only one woman, **Marie Curie**, has been honoured twice, with the Nobel Prize in Physics 1903 and the Nobel Prize in Chemistry 1911. This means that 65 women in total have been awarded the Nobel Prize between 1901 and 2024.



 Nobel Prize

<https://www.nobelprize.org/prizes/lists/nobel-prize...>

[Nobel Prize awarded women - NobelPrize.org](#)

?

About featured snippets • 

Feedback

Results for **Paris, France** · [Choose area](#) ⋮

## 75001 Paris, France

Louvre Museum, Address



Wikipedia




<https://en.wikipedia.org/wiki/Louvre> ⋮

### Louvre

The Louvre museum is located **inside the Louvre Palace, in the center of Paris**, adjacent to the Tuileries Gardens. The two nearest Métro stations are Louvre ...

[Louvre Palace](#) · [Louvre Pyramid](#) · [Louvre Abu Dhabi](#) · [Art museum](#)

### People also ask ⋮

Where is the Louvre located exactly? How far apart are the Louvre and Eiffel Towers? Is the Louvre where the Mona Lisa is? How much does it cost to get into the Louvre? [Feedback](#)

Le Louvre

<https://www.louvre.fr/visit/map-entrances-directions> ⋮

### Map, entrances & directions - - All roads lead to the Louvre

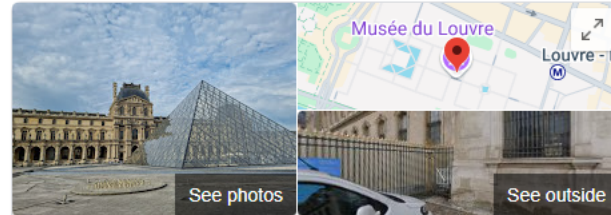
An underground car park is located at 1 **Avenue du Général Lemonnier**, from which you can access the museum via the Galerie du Carrousel entrance.



Britannica

[https://www.britannica.com/Visual\\_Arts/Painting](https://www.britannica.com/Visual_Arts/Painting) ⋮

### Louvre | History, Collections, & Facts



## Louvre Museum

[Website](#)[Directions](#)[Save](#)

4.7 ★★★★★ 325,915 Google reviews

Museum in Paris, France

[SEE TICKETS](#)

### Sponsored

Withlocals ⋮

[Louvre at Night: Explore with a Local](#)

\$125 · 5.0 ★ (1.4K)



The Louvre, or the Louvre Museum, is a national art museum in Paris, France, and one of the most famous museums in the world.  
[Wikipedia](#)

**Departments:** [Librairie-Boutique du Musée du Louvre](#)**Address:** 75001 Paris, France**Founded:** August 10, 1793**Hours:** Closed · Opens 9 a.m. · [More hours](#)**Director:** [Laurence des Cars](#)**Visitors:** 8.9 million (2023): Ranked 1st nationally; Ranked 1st globally**Phone:** +33 1 40 20 53 17**Subsidiary:** [Louvre Conservation Center](#)**Curator:** [Marie-Laure de Rochebrune](#)[Suggest an edit](#)

## Louvre Museum / Artworks



Mona Lisa  
Leonardo da Vinci



Venus de Milo  
Alexandros of An...



Winged Victory  
of Samothrace



Liberty Leading  
the People  
Eugène Delacroix



Psyche Revived  
by Cupid's Kiss  
Antonio Canova



The Raft of the  
Medusa  
Théodore Géricault



The Coronation  
of Napoleon  
Jacques-Louis D...



The Wedding at  
Cana  
Paolo Veronese



The Seated  
Scribe



The Virgin of the  
Rocks  
Leonardo da Vinci



La Belle  
Ferronnière  
Leonardo da Vinci



Oath of the  
Horatii  
Jacques-Louis D...

Feedback



Paris City Vision

<https://www.pariscityvision.com> > ... > Louvre museum

## Louvre artwork : top masterpieces and paintings

How can we not mention the **Mona Lisa**? The portrait assumed to be of the wife of Francesco del Giocondo is considered to be the most famous painting in the world ...

## People also ask

What is the most famous artwork in Louvre?



What are the three masterpieces of the Louvre?



What are the big 3 at the Louvre?



Where is the real Mona Lisa painting?



Feedback



Le Louvre

<https://www.louvre.fr> > explore > visitor-trails > the-lou...

## The Louvre's Masterpieces - What exactly is a ...

The palace is home to some of the **world's most** iconic pieces – **paintings**, sculptures, architectural elements and **art** objects by **famous** or anonymous artists.



See photos

See outside

## Louvre Museum

Website

Directions

Save

4.7 ★★★★★ 325,915 Google reviews

Museum in Paris, France

SEE TICKETS

### Sponsored

Withlocals

Withlocals Your Way! - Paris City Tour

\$87 · 5.0 ★ (2.2K)



The Louvre, or the Louvre Museum, is a national art museum in Paris, France, and one of the most famous museums in the world.  
[Wikipedia](#)

Departments: Librairie-Boutique du Musée du Louvre



Wikipedia

[https://en.wikipedia.org/wiki/Salon\\_des\\_Refusés](https://en.wikipedia.org/wiki/Salon_des_Refusés)

## Salon des Refusés

Today, by extension, salon des refusés refers to any exhibition of works rejected from a juried art show.



### People also ask

Where is the Salon Carre in the Louvre?



What happened with the works entered in the Salon of the Refused?



Which painting was included in the first Salon des Rejectés Salon of the Rejected?



Does the Salon in Paris still exist?

[Feedback](#)

Artland Magazine

<https://magazine.artland.com/articles-and-features>

## Contemporary Art History: The Salon Des Refusés

Discover the 1863 **Salon des Refusés**: first of a string of landmark contemporary art shows that have radically changed the course of Art History.



Le Louvre

<https://www.louvre.fr/explore/visitor-trails/the-louvre>

## The Louvre's Masterpieces - What exactly is a ...

The palace is home to some of the world's most iconic **pieces** – **paintings**, **sculptures**, architectural elements and art objects by famous or anonymous artists.



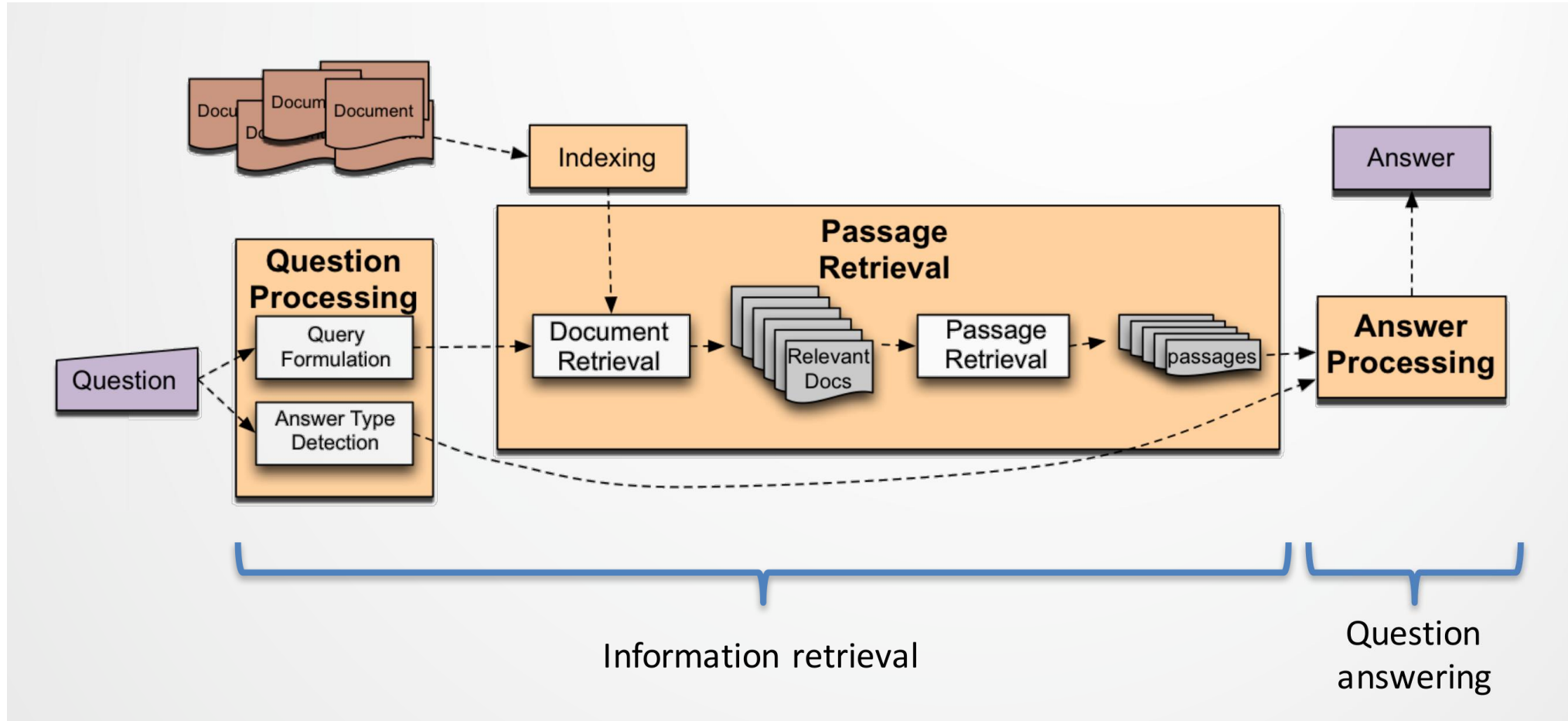
The Tour Guy

<https://thetourguy.com/france/paris/louvre>

## The Louvre Museum's 17 Most Important Works of Art, Paris

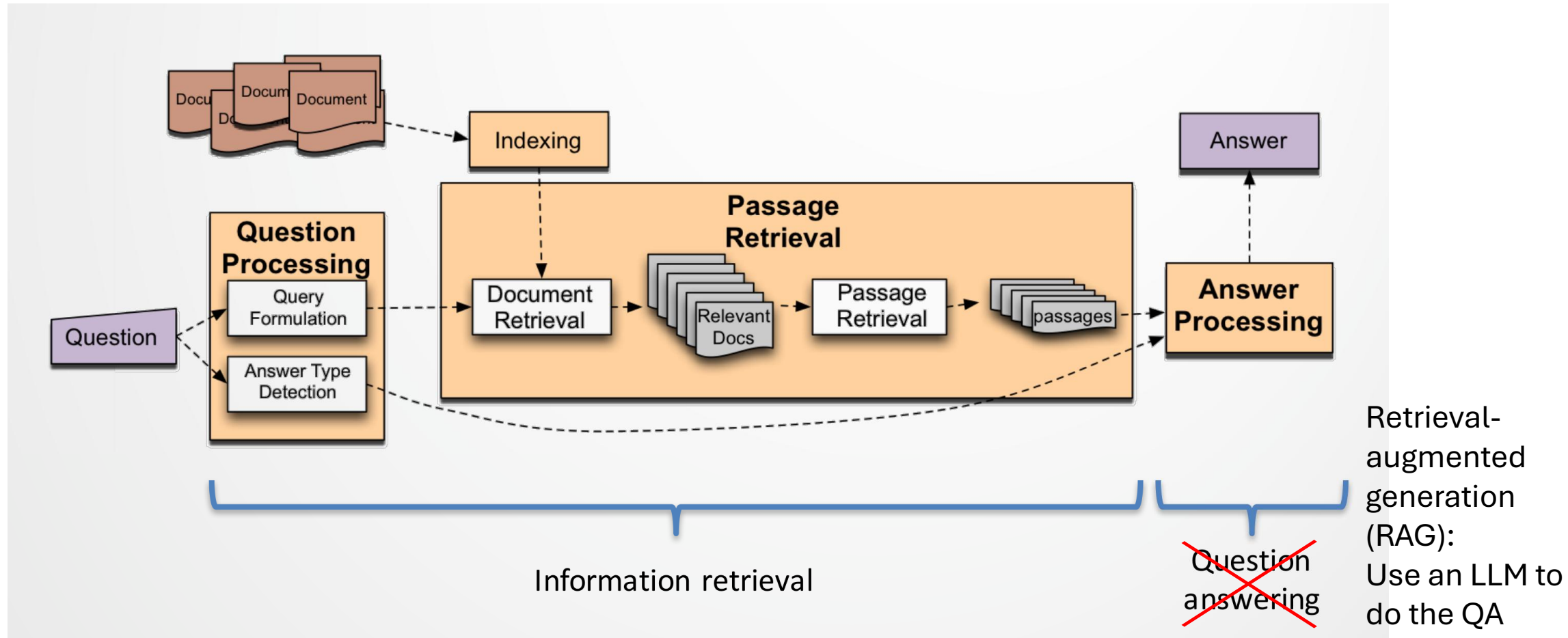
Oct 28, 2024 — The **Louvre** is massive. To make things easy, we've listed 17 famous **paintings** to see in the **Louvre** and explained why they're so important.

# IR-based QA





# IR-based QA with LLM (RAG)



# Sample TREC questions

1. Who is the author of the book, "The Iron Lady: A Biography of Margaret Thatcher"?
2. What was the monetary value of the Nobel Peace Prize in 1989?
3. What does the Peugeot company manufacture?
4. How much did Mercury spend on advertising in 1993?
5. What is the name of the managing director of Apricot Computer?
6. Why did David Koresh ask the FBI for a word processor?
7. What debts did Quintex group leave?
8. What is the name of the rare neurological disease with symptoms such as: involuntary movements (tics), swearing, and incoherent vocalizations (grunts, shouts, etc.)?



# Query types

- Different kinds of questions can be asked.
  - Factoid questions, e.g.,
    - *How often were the peace talks in Ireland delayed or disrupted as a result of acts of violence?*
  - Narrative (open-ended) questions, e.g.
    - *Can you tell me about contemporary interest in the Greek philosophy of stoicism?*
  - Complex/hybrid questions, e.g.,
    - *Who was involved in the Schengen agreement to eliminate border controls in Western Europe and what did they hope to accomplish?*



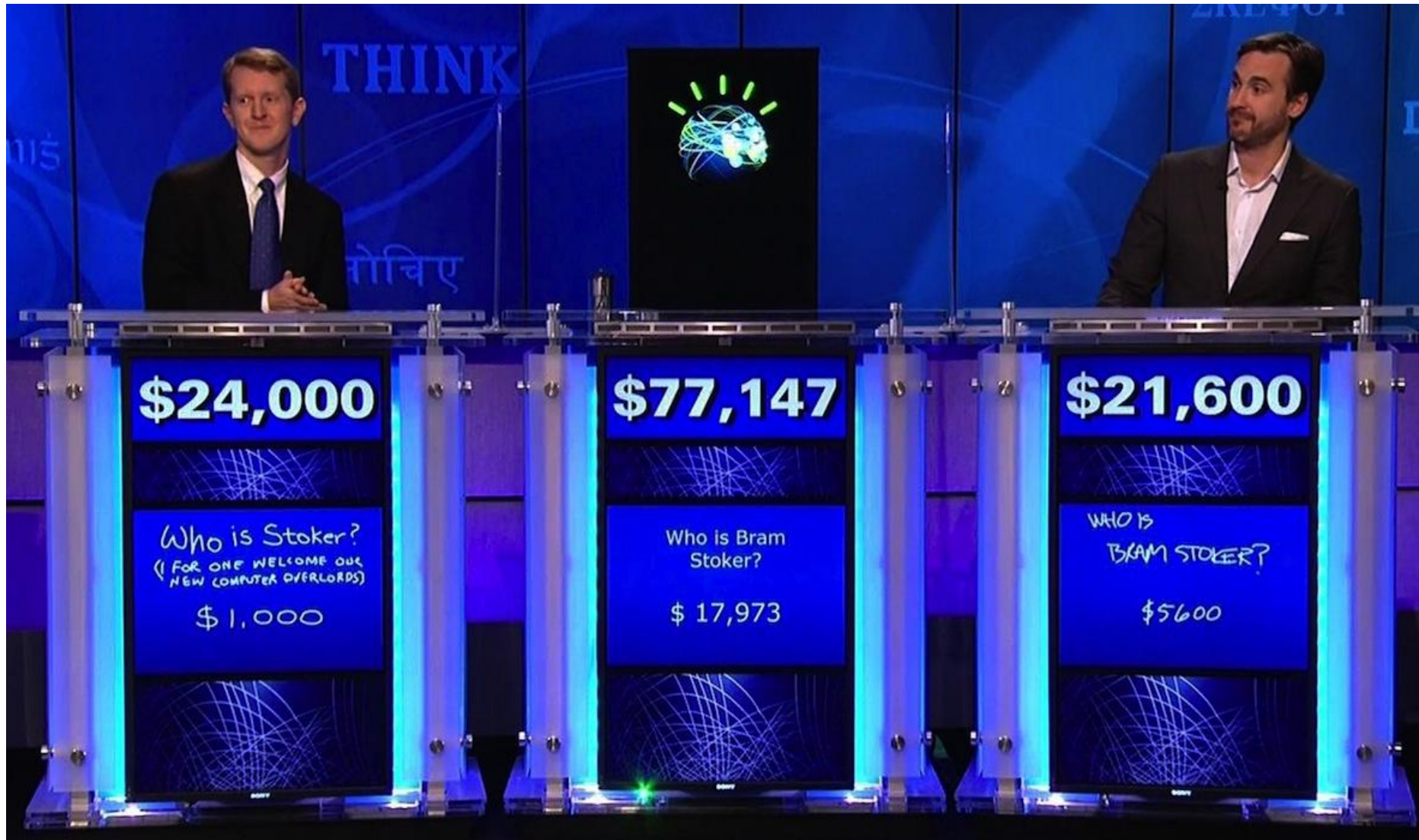
# People **want** to ask questions...



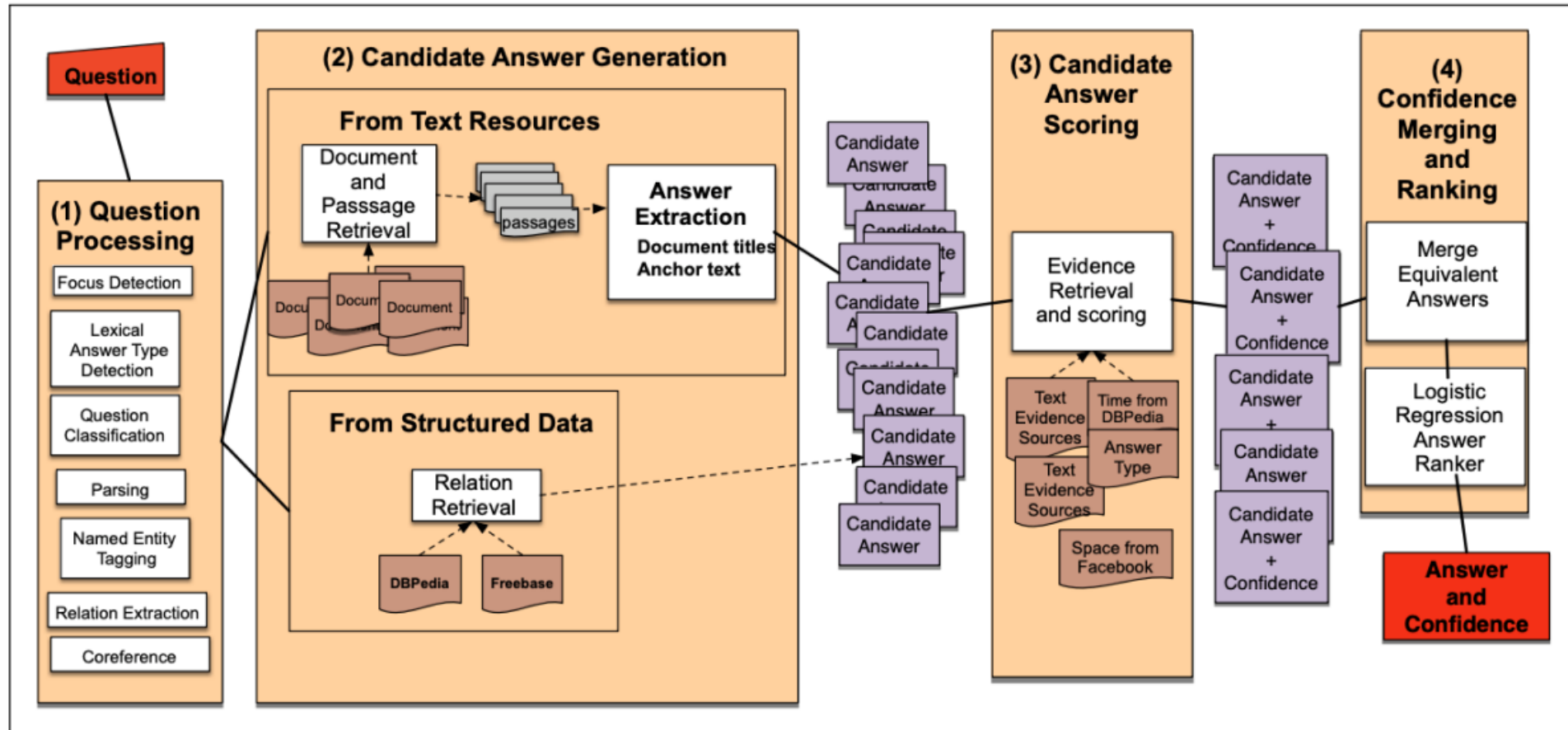
- **Examples from AltaVista query log (late 1990s)**
  - who invented surf music?
  - how to make stink bombs
  - where are the snowdens of yesteryear?
  - which english translation of the bible is used in official catholic liturgies?
  - how to do clayart
  - how to copy psx
  - how tall is the sears tower?
- **Examples from Excite query log (12/1999)**
  - how can i find someone in texas
  - where can i find information on puritan religion?
  - what are the 7 wonders of the world
  - how can i eliminate stress
  - What vacuum cleaner does Consumers Guide recommend

**Around 10% of early query logs are QUESTIONS.**

# 2011: IBM Watson beat Jeopardy! champions



# IBM Watson: Search



# QA at TREC

- Question answering competition at TREC started with answering a set of 500 fact-based questions:
  - E.g., “*When was Mozart born?*”
- For the first three years systems were allowed to return 5 ranked answer snippets (50/250 bytes) to each question.
  - Mean Reciprocal Rank (MRR) scoring:
    - 1, 0.5, 0.33, 0.25, 0.2, 0 for 1, 2, 3, 4, 5, 6+ rankings
  - Mainly Named Entity answers (person, place, date, ...)
- From 2002 the systems were only allowed to return a single **exact** answer and the notion of confidence was introduced.

# The TREC Document Collection

- Each task features a collection from a domain,
  - E.g., news articles:
    - AP newswire, 1998-2000
    - New York Times newswire, 1998-2000
    - Xinhua News Agency newswire, 1996-2000
- Usually about 1,000,000 documents in the collection. Roughly 3GB of text.
- This was once a lot of text to process entirely using advanced NLP techniques, so the systems usually consisted of an initial information retrieval phase followed by more advanced processing.
- Allowed to supplement this text with use of the web, and other knowledge bases?
- See also SQUAD (1.1 and 2.0/open).

# Top Performing Systems

- Best TREC vanilla QA systems score ~60-80% !!!
- Approaches and successes have varied a fair deal
- AskMSR (2001): first wildly successful purely statistical system, stressing how much could be achieved by very simple methods with enough text (and now various copycats)




# AskMSR: Shallow Approach

- *In what year did Abraham Lincoln die?*
- Ignore hard documents and find easy ones.

**Abraham Lincoln, 1809-1865**


**\*LINCOLN, ABRAHAM** was born near Hodgenville, Kentucky, on February 12, 1809. In 1816, the Lincoln family moved to Pigeon Creek in Perry (now Spencer) County. Two years later, Abraham Lincoln's mother died and his father married a woman who became his "angel" mother. Lincoln attended a formal school for only a few months but acquired knowledge through the reading of books. He moved to Illinois, in 1830 where he obtained a job as a store clerk and the local postmaster. He served without distinction in the Black Hawk War. He lost his attempt at the state legislature, but two years later he tried again, was successful, and Lincoln was admitted to the bar and became noteworthy as a witty, honest, competent circuit lawyer. He served a one-year term in the U.S. House in 1846, at which time he opposed the war with Mexico. By 1858, Lincoln had gained national attention for his series of debates with Stephen A. Douglas. He lost the election but became a significant figure in his party. On the day of his inauguration on March 4, seven southern states had seceded, for a total of 11. Lincoln immediately took action to suppress the rebellion. He called for 75,000 volunteers (approximately 43,000 were accepted). The Emancipation Proclamation which expanded the purpose of the war to the abolition of slavery was issued. The dedication of a national cemetery in Gettysburg, Lincoln's explanation of the war, and his final address to the nation are among his most famous speeches. He died of disease on April 15, 1865, at Ford's Theatre in Washington, D.C.



**ABRAHAM LINCOLN**

**Sixteenth President of the United States**

**Born in 1809 - Died in 1865**




**Sixteenth President**  
1861-1865  
Married to Mary Todd Lincoln

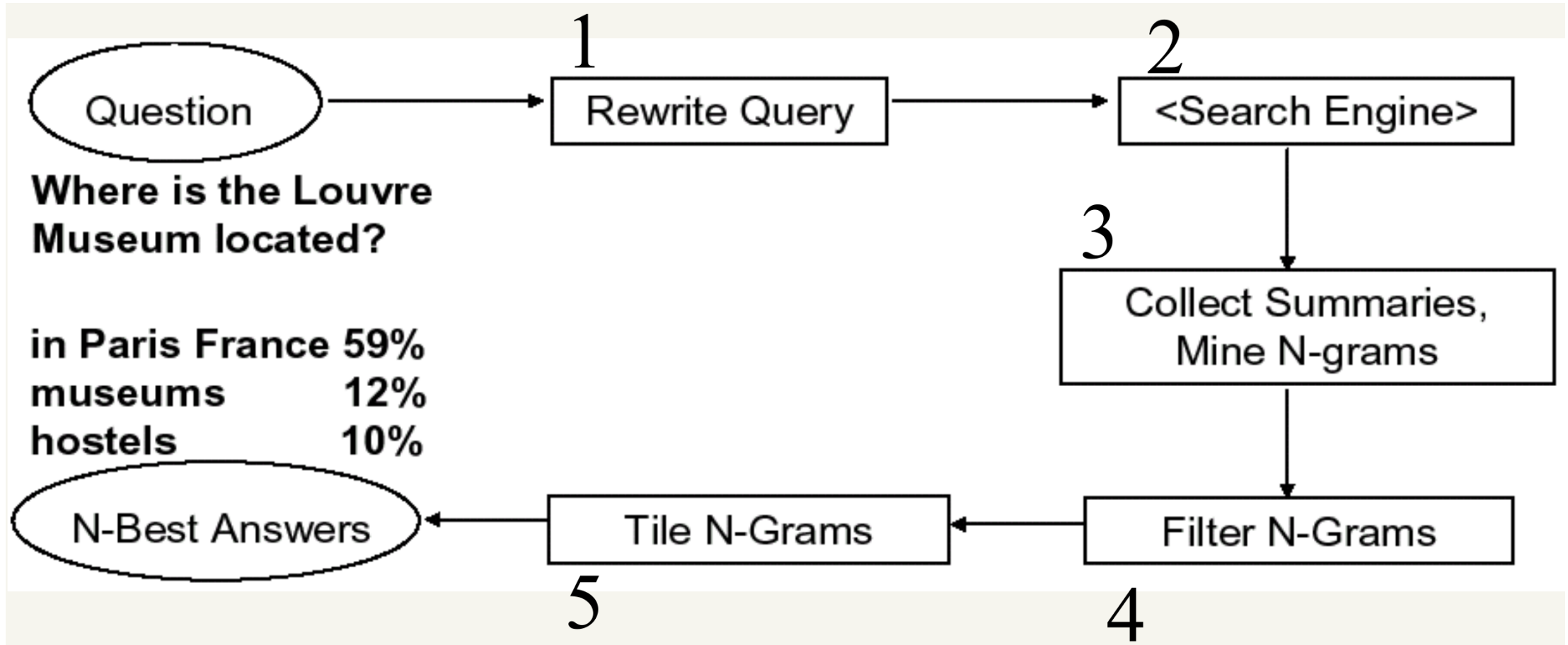
**Abraham Lincoln**

**16th President of the United States (March 4, 1861 to April 15, 1865)**  
Born: February 12, 1809, in Hardin County, Kentucky  
Died: April 15, 1865, at Petersen's Boarding House in Washington, D.C.

"I was born February 12, 1809, in Hardin County, Kentucky. My parents were both born in Virginia, of undistinguished families, perhaps I should say. My mother, who died in my tenth year, was of a family of the name of Hanks."



# AskMSR: Details





# Step 1: Query rewriting: Answer similar to Question

- Classify question into seven categories.

- **Who** is/was/are/were...?
- **When** is/did/will/are/were ...?
- **Where** is/are/were ...?

## 1. Category-specific transformation rules.

- “Where is the Louvre Museum located”
  - “is the Louvre Museum located”
  - “the is Louvre Museum located”
  - “the Louvre is Museum located”
  - “the Louvre Museum is located”
  - “the Louvre Museum located is”

**Nonsense,**  
but who  
cares? It’s  
only a few  
more queries  
to Google.

## 2. Expected answer “Datatype” (eg, Date, Person, Location)

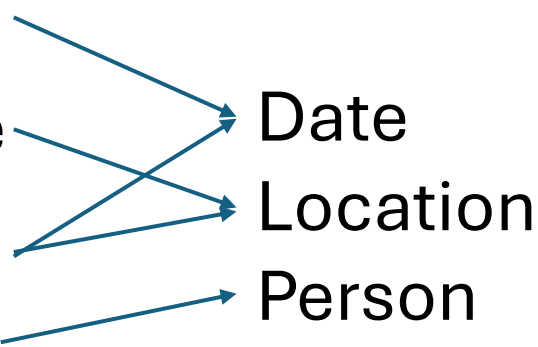
- When was the French Revolution? → DATE

## 3. Hand-crafted classification/rewrite/datatype rules

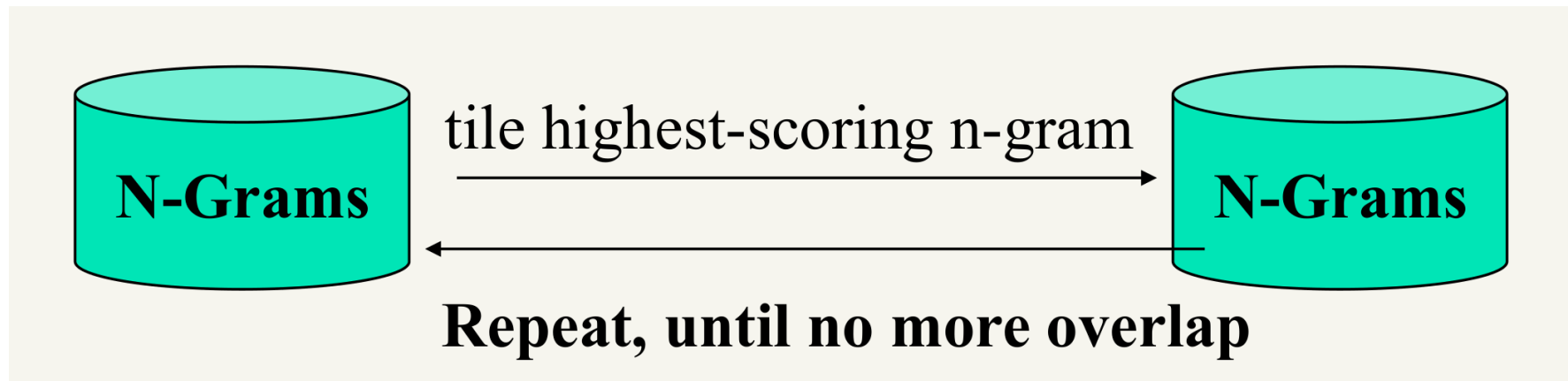
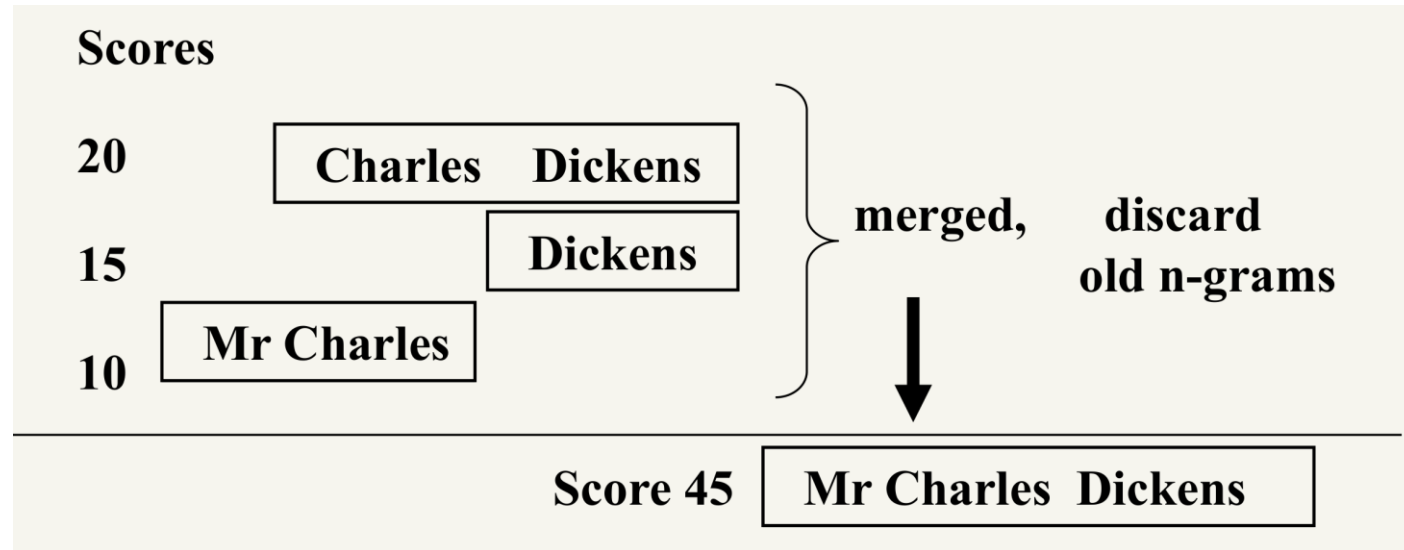
# Step 3: Mining N-Grams

- Send query to search engine; use result snippets
- Enumerate all N-grams in all retrieved snippets
  - Use hash table and other fancy footwork to make this efficient
- Weight of an n-gram: occurrence count, each weighted by “reliability” (weight) of rewrite that fetched the document.
- Example: “Who created the character of Scrooge?”
  - Dickens - 117
  - Christmas Carol - 78
  - Charles Dickens - 75
  - Disney - 72
  - Carl Banks - 54
  - A Christmas - 41
  - Christmas Carol - 45
  - Uncle - 31

# Step 4: Filtering N-Grams

- Each question type is associated with one or more “data-type filters” = regular expression
  - When
  - Where
  - What
  - Who
- 
- ```
graph LR; When --> Date; Where --> Location; What --> Date; What --> Location; Who --> Person;
```
- Date
  - Location
  - Person
- Boost score of n-grams that do match regexp
  - Lower score of n-grams that don't match regexp

# Step 5: Tiling the Answers



# Results

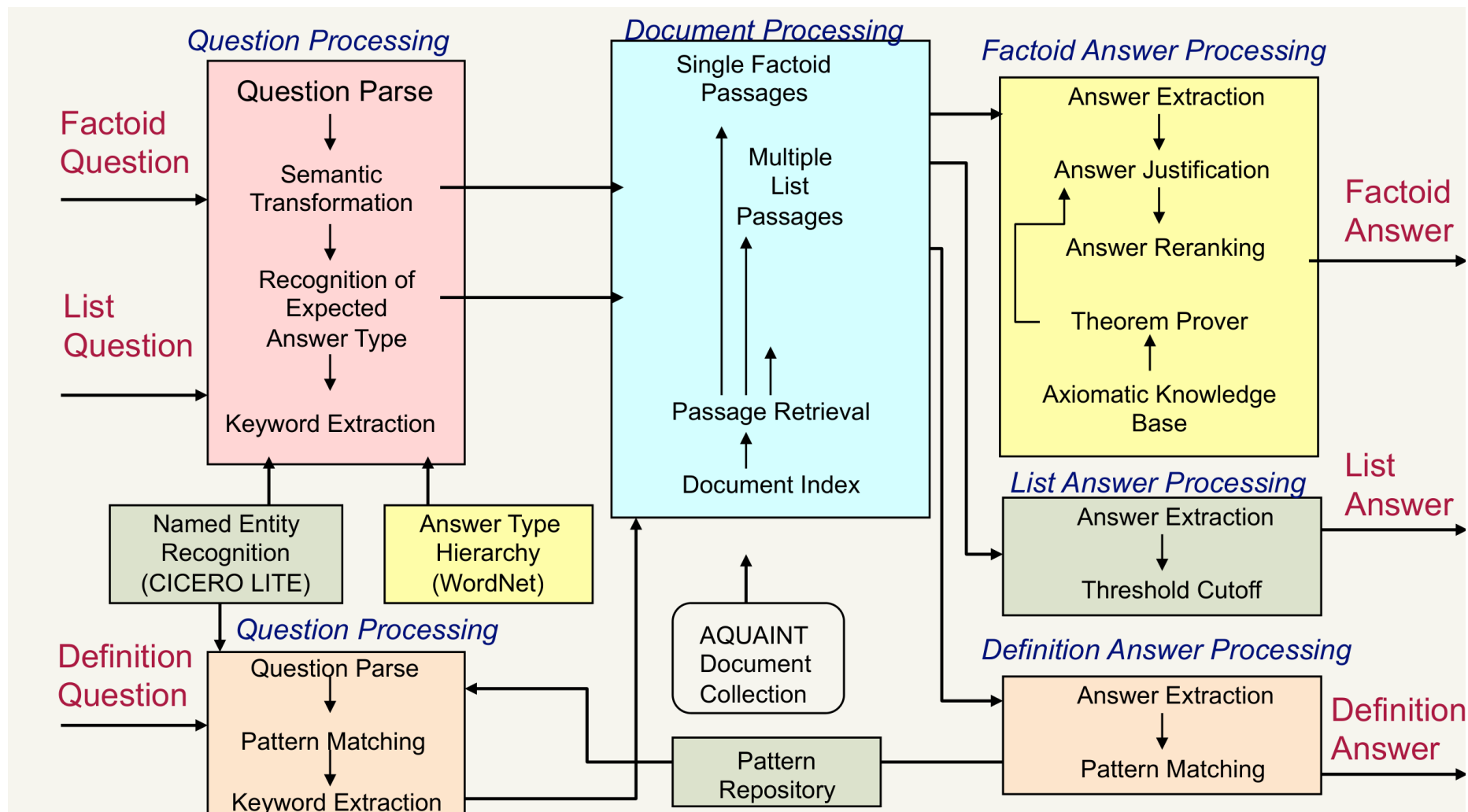
- Standard TREC contest test-bed:
  - ~1M documents; 900 questions.
- Doesn't do so well (but in top 9 of ~30 participants)
  - MRR = 0.262
    - Right answer ranked about #4–5 on average
  - Why? Because it relies on the enormity of the Web
- Using the Web as a whole, not just TREC's 1M documents
  - MRR = 0.42
  - On average, right answer is ranked about #2–3

# Limitations

- In many scenarios we only have a small set of documents
  - e.g., monitoring an individuals email...
- Works best/only for trivia-style fact-based questions
- Limited/brittle repertoire of
  - question categories
  - answer data types/filters
  - query rewriting rules

# Full NLP QA: LCC (Harabagiu/Moldovan)

below is the Architecture of LCC's QA system circa 2003



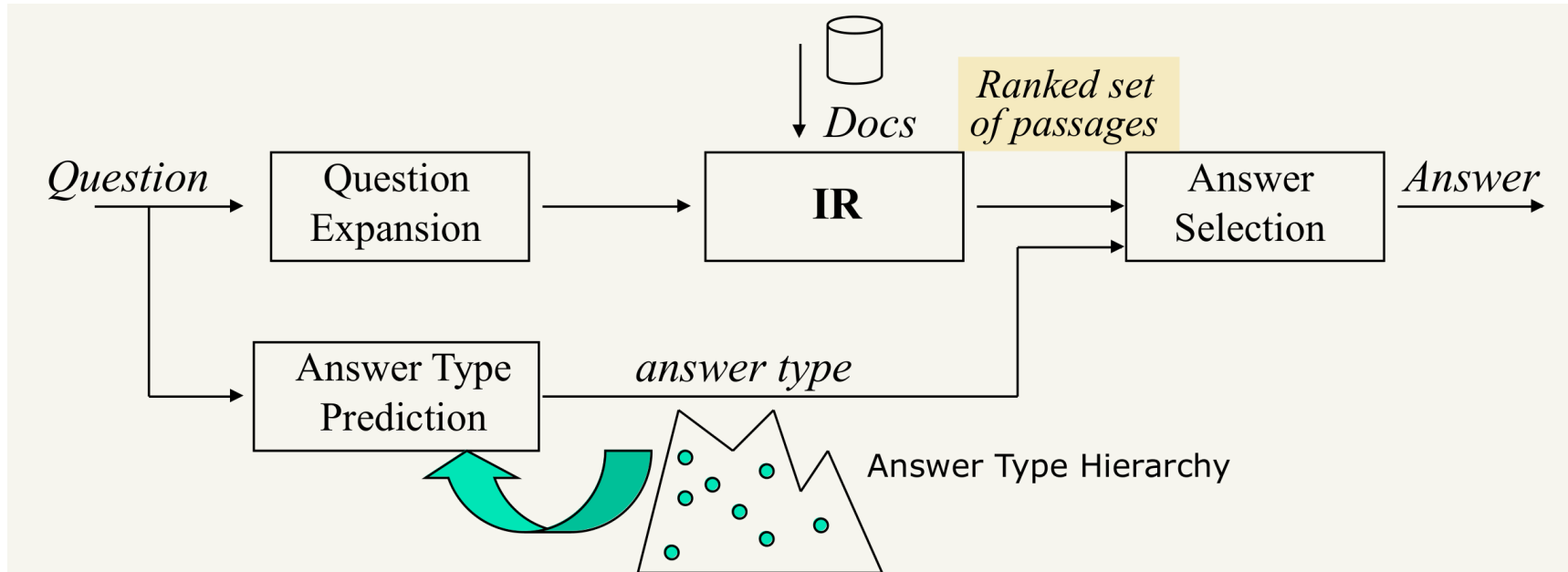
# Value from sophisticated NLP

Pasca and Harabagiu (2001)

- Good IR is needed: SMART paragraph retrieval
- Large taxonomy of question types and expected answer types is crucial
- Statistical parser used to parse questions and relevant text for answers, and to build KB
- Further value comes from deeper NLP and inferencing



# Answer types in LCC QA systems



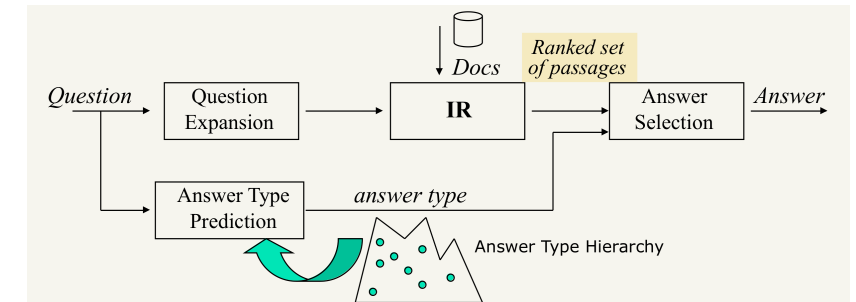
- Answer type
  - Labels questions with answer type based on a taxonomy
  - Person, location, weight, temperature, year, vehicle
  - Classifies questions (e.g. by using a maximum entropy model)

# Answer Types

- Of course, determining the answer type isn't that easy...
- **Who** questions can have organizations as answers
  - *Who sells the most hybrid cars?*
- **Which** questions can have people as answers
  - *Which president went to war with Mexico?*

# Lexical Term Extraction:

## Input to Information Retrieval



- Questions approximated by sets of unrelated words (lexical terms)
- Similar to bag-of-word IR models: but choose nominal non-stop words and verbs

| Question (from TREC QA track)                                       | Lexical terms                              |
|---------------------------------------------------------------------|--------------------------------------------|
| Q002: What was the monetary value of the Nobel Peace Prize in 1989? | monetary, value, Nobel, Peace, Prize, 1989 |
| Q003: What does the Peugeot company manufacture?                    | Peugeot, company, manufacture              |
| Q004: How much did Mercury spend on advertising in 1993?            | Mercury, spend, advertising, 1993          |

# Keyword Selection Algorithm

1. Select all non-stopwords in quotations
2. Select all NNP words in recognized named entities
3. Select all complex nominals with their adjectival modifiers
4. Select all other complex nominals
5. Select all nouns with adjectival modifiers
6. Select all other nouns
7. Select all verbs
8. Select the answer type word

# Passage Extraction Loop

- Passage Extraction Component
  - Extracts passages that contain all selected keywords
  - Passage size dynamic
  - Start position dynamic
- Passage quality and keyword adjustment
  - In the first iteration use the first 6 keyword selection heuristics
  - If the number of passages is lower than a threshold
    - ⇒ query is too strict
    - ⇒ drop a keyword
  - If the number of passages is higher than a threshold
    - ⇒ query is too relaxed
    - ⇒ add a keyword

# Passage Scoring

- Passage ordering is performed using a sort that involves three scores:
  - The number of words from the question that are recognized in the same sequence in the window
  - The number of words that separate the most distant keywords in the window
  - The number of unmatched keywords in the window

# Rank candidate answers in retrieved passages

Q066: Name the first private citizen to fly in space.

- Answer type: **Person**
- Text passage:

“Among them was **Christa McAuliffe**, the first private citizen to fly in space. **Karen Allen**, best known for her starring role in “Raiders of the Lost Ark”, plays **McAuliffe**. **Brian Kerwin** is featured as shuttle pilot **Mike Smith**...”
- Best candidate answer: **Christa McAuliffe**

# Extracting Answers for Factoid Questions: NER!

- In TREC 2003 the LCC QA system extracted 289 correct answers for factoid questions
- The Name Entity Recognizer was responsible for 234 of them
  - Current QA is largely based on the high accuracy recognition of a large variety of Named Entity types

|                 |    |               |    |              |   |
|-----------------|----|---------------|----|--------------|---|
| QUANTITY        | 55 | ORGANIZATION  | 15 | PRICE        | 3 |
| NUMBER          | 45 | AUTHORED WORK | 11 | SCIENCE NAME | 2 |
| DATE            | 35 | PRODUCT       | 11 | ACRONYM      | 1 |
| PERSON          | 31 | CONTINENT     | 5  | ADDRESS      | 1 |
| COUNTRY         | 21 | PROVINCE      | 5  | ALPHABET     | 1 |
| OTHER LOCATIONS | 19 | QUOTE         | 5  | URI          | 1 |
| CITY            | 19 | UNIVERSITY    | 3  |              |   |



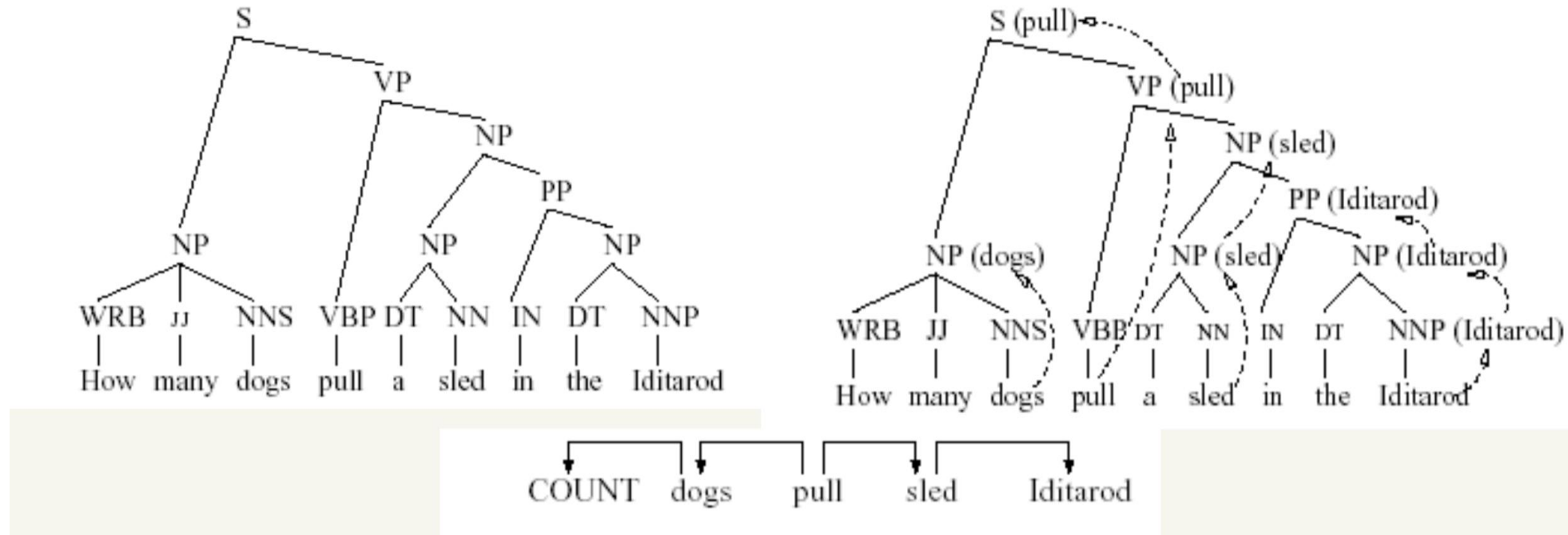
# Semantics and Reasoning for QA:

## Predicate-argument structure

- Q336: *When was Microsoft established?*
- This question is difficult because Microsoft tends to establish lots of things...
  - **Microsoft** plans to **establish** manufacturing partnerships in Brazil and Mexico in **May**.
- Need to be able to detect sentences in which ‘Microsoft’ is **object** of ‘establish’ or close synonym.
- Matching sentence:
  - *Microsoft Corp was founded in the US in 1975, incorporated in 1981, and established in the UK in 1982.*
- Requires analysis of sentence syntax/semantics

# Semantics and Reasoning for QA:

## Syntax to Logical Forms



- Syntactic analysis plus semantic => logical form
- Mapping of question and potential answer LFs to find the best match

# Abductive inference

- System attempts inference to justify an answer (often following lexical chains)
- Their inference is a kind of funny middle ground between logic and pattern matching
- But very effective: 30% improvement
  - Q: When was the internal combustion engine invented?
  - A: The first internal-combustion engine was built in 1867.
  - invent → create\_mentally → create → build

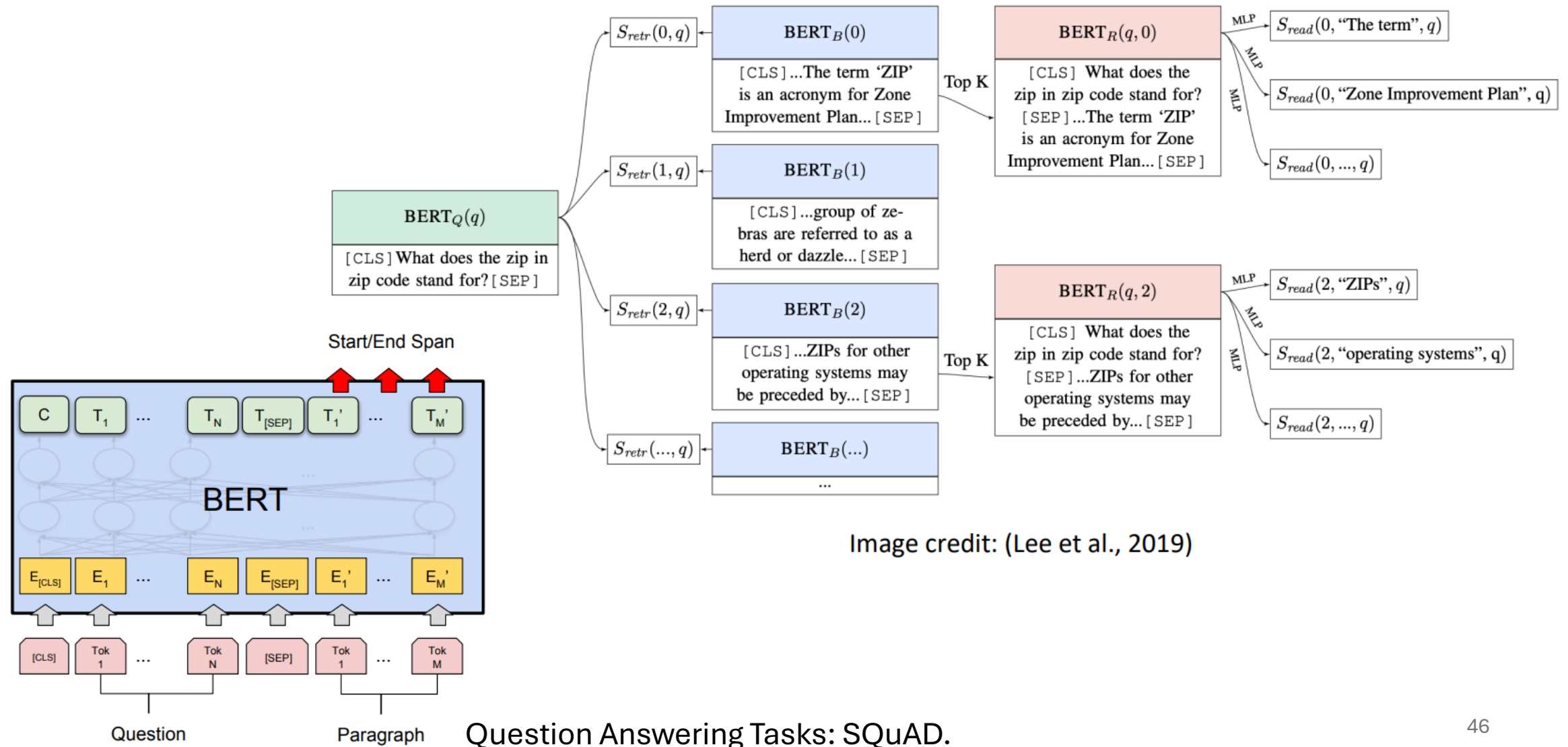
# Question Answering and Inference

- How hot does the inside of an active **volcano** get?
  - `get(TEMPERATURE, inside(volcano(active)))`
- A: “**lava** fragments belched out of the **mountain** were as hot as 300 degrees Fahrenheit”
- `fragments(X, lava, temperature(degrees(300)), belched(X, mountain))`
  - volcano IS\_A mountain
  - lava IS\_PART\_OF volcano
  - lava inside volcano
  - fragments of lava HAVE\_PROPERTIES\_OF lava
- The needed semantic information is in WordNet definitions, and was successfully translated into a form that was used for rough “proofs.”

# Not all problems are solved by these

- Where do lobsters like to live?
  - on a Canadian airline
- Where are zebras most likely found?
  - near dumps
  - in the dictionary
- Why can't ostriches fly?
  - Because of American economic sanctions
- What's the population of Mexico?
  - Three
- What can trigger an allergic reaction?
  - ..something that can **trigger** an allergic reaction

# Question answering in deep learning era



# SQuAD: Stanford question answering dataset

- 100k annotated (passage, question, answer) triples
  - Large-scale supervised datasets are also a key ingredient for training effective neural models for reading comprehension!
- Passages are selected from English Wikipedia, usually 100~150 words.
- Questions are crowd-sourced.
- Each answer is a short segment of text (or span) in the passage.
  - This is a limitation— not all the questions can be answered in this way!
- SQuAD was for years the most popular reading comprehension dataset; it is “almost solved” today (though the underlying task is not,) and the state-of-the-art exceeds the estimated human performance.
- SQuAD 2.0: some questions can’t be answered.

---

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called “showers”.

What causes precipitation to fall?

**gravity**

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?

**graupel**

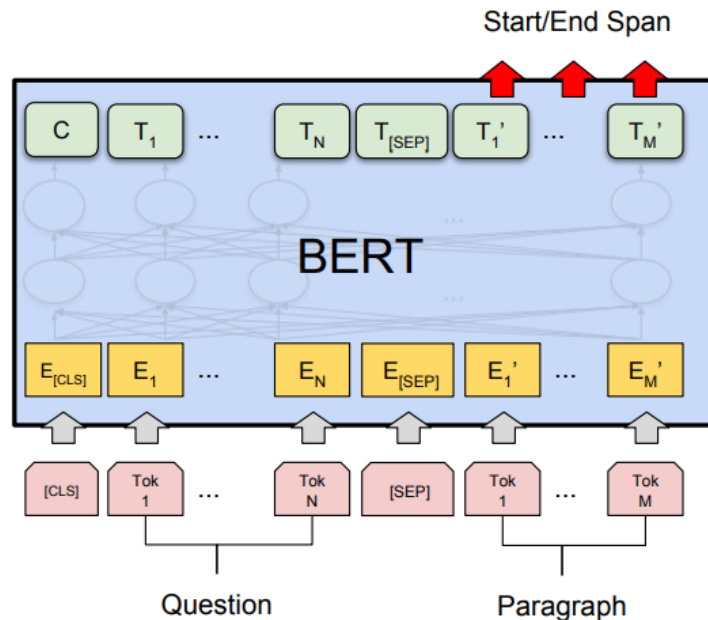
Where do water droplets collide with ice crystals to form precipitation?

**within a cloud**

---

<https://rajpurkar.github.io/SQuAD-explorer/>

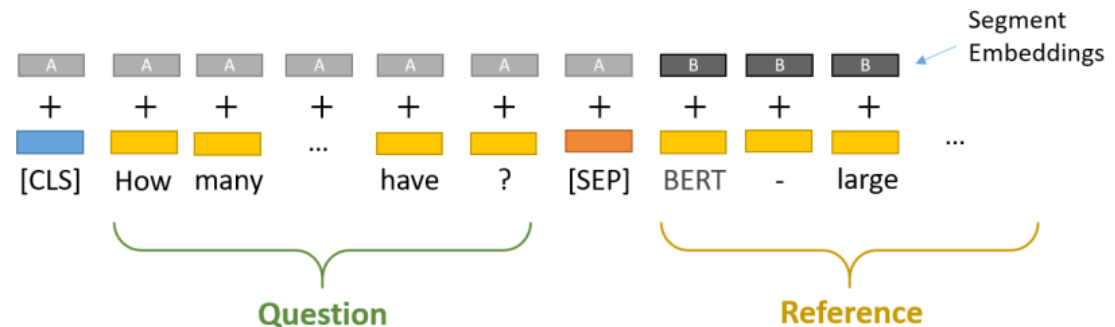
# BERT for Reading Comprehension



$$p_{\text{start}}(i) = \text{softmax}(\mathbf{W}_s \mathbf{h}_i)$$

$$p_{\text{end}}(i) = \text{softmax}(\mathbf{W}_e \mathbf{h}_i)$$

- This simplified version of QA aka **Reading Comprehension**.
  - (Passage, Question)  $\Rightarrow$  Answer

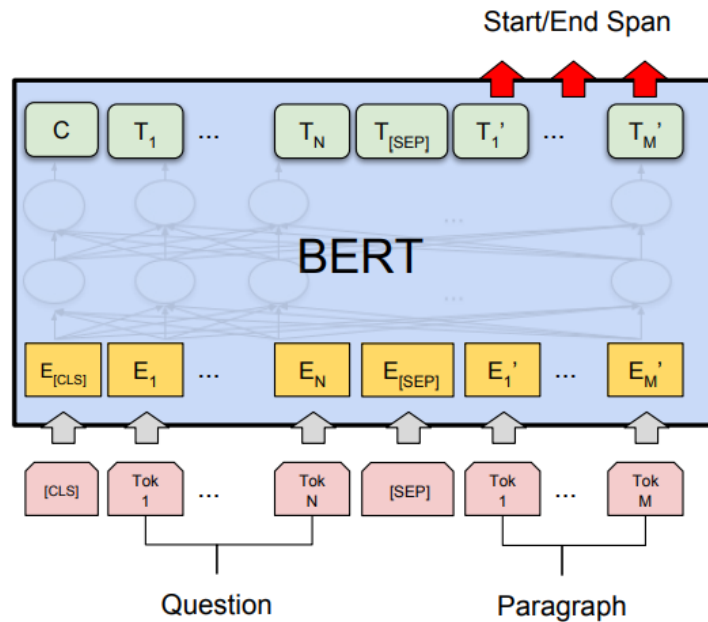


**Question:** How many parameters does BERT-large have?

**Reference Text:** BERT-large is really big... it has 24 layers and an embedding size of 1,024, for a total of 340M parameters! Altogether it is 1.34GB, so expect it to take a couple minutes to download to your Colab instance.



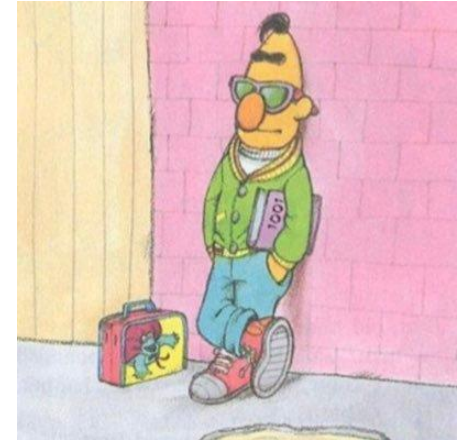
# BERT for Reading Comprehension



$$p_{\text{start}}(i) = \text{softmax}(\mathbf{W}_s \mathbf{h}_i)$$

$$p_{\text{end}}(i) = \text{softmax}(\mathbf{W}_e \mathbf{h}_i)$$

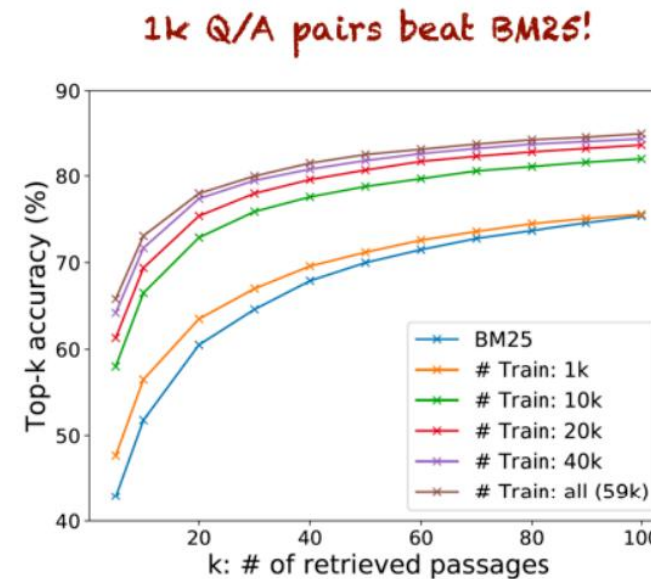
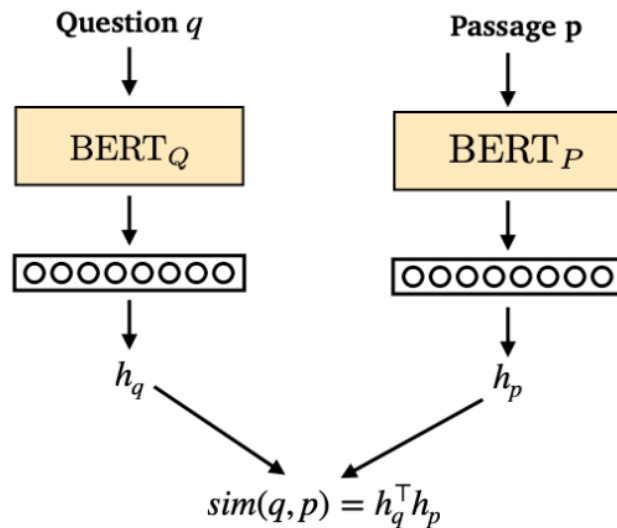
- This simplified version of QA aka **Reading Comprehension**.
  - (Passage, Question)  $\Rightarrow$  Answer



|                   | F1    | EM    |
|-------------------|-------|-------|
| Human performance | 91.2* | 82.3* |
| BiDAF             | 77.3  | 67.7  |
| BERT-base         | 88.5  | 80.8  |
| BERT-large        | 90.9  | 84.1  |
| XLNet             | 94.5  | 89.0  |
| RoBERTa           | 94.6  | 88.9  |
| ALBERT            | 94.8  | 89.3  |

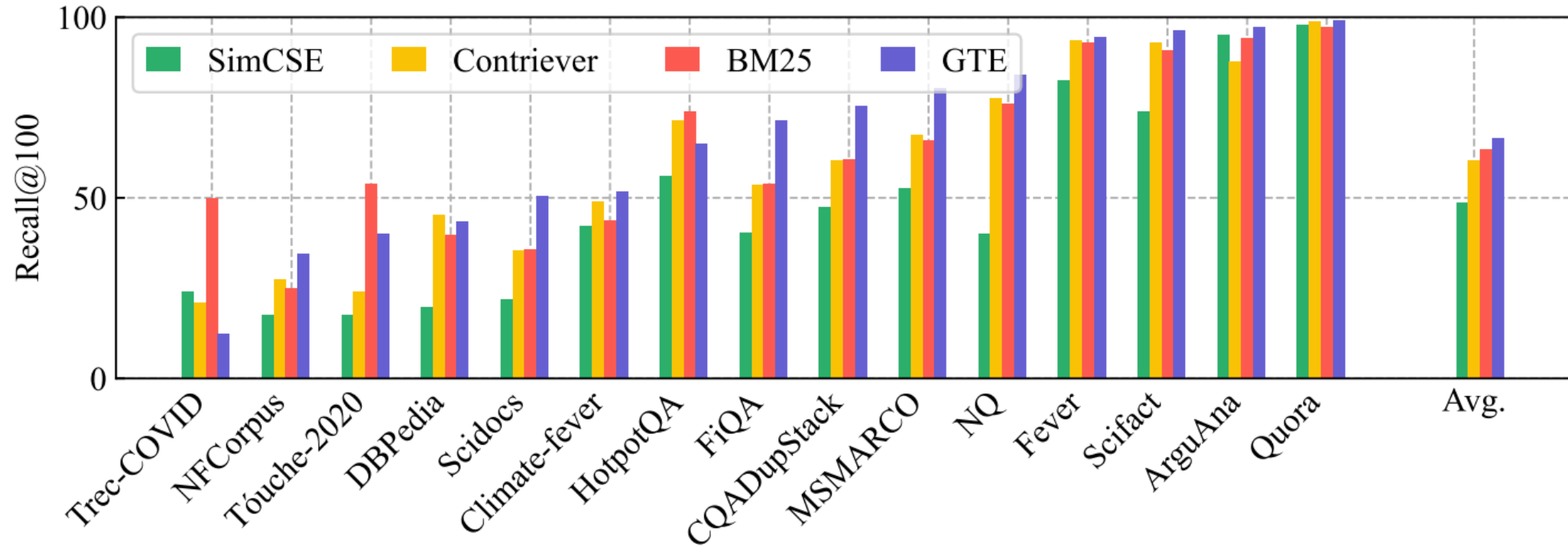
# BERT for IR

- Dense passage retrieval (DPR)
  - We can also just train the retriever using question-answer pairs!



- Trainable retriever (using BERT) largely outperforms traditional IR retrieval models.

# Improvements



- GTE: general-purpose text embedding
  - Multi-stage contrastive learning.
  - Recall SBERT, SimCSE.

# Retrieval-Augmented Generation (RAG)

- Sounds fancy, but actually very simple.
- RAG:
  - Step 1: retrieve N documents using some IR algorithm.
  - Step 2: write the augmented query.

```
Context information is below.
```

```
-----
```

```
{context_str}
```

```
-----
```

```
Given the context information and not  
prior knowledge, answer the query.
```

```
Query: {query_str}
```

```
Answer:
```

- Step 3: Profit.

Prompt templates:

[https://github.com/  
run-llama/llama\\_index](https://github.com/run-llama/llama_index)

# Distraction in RAG

- Distraction:
  - When a piece of irrelevant context is provided, the model generates an incorrect response.



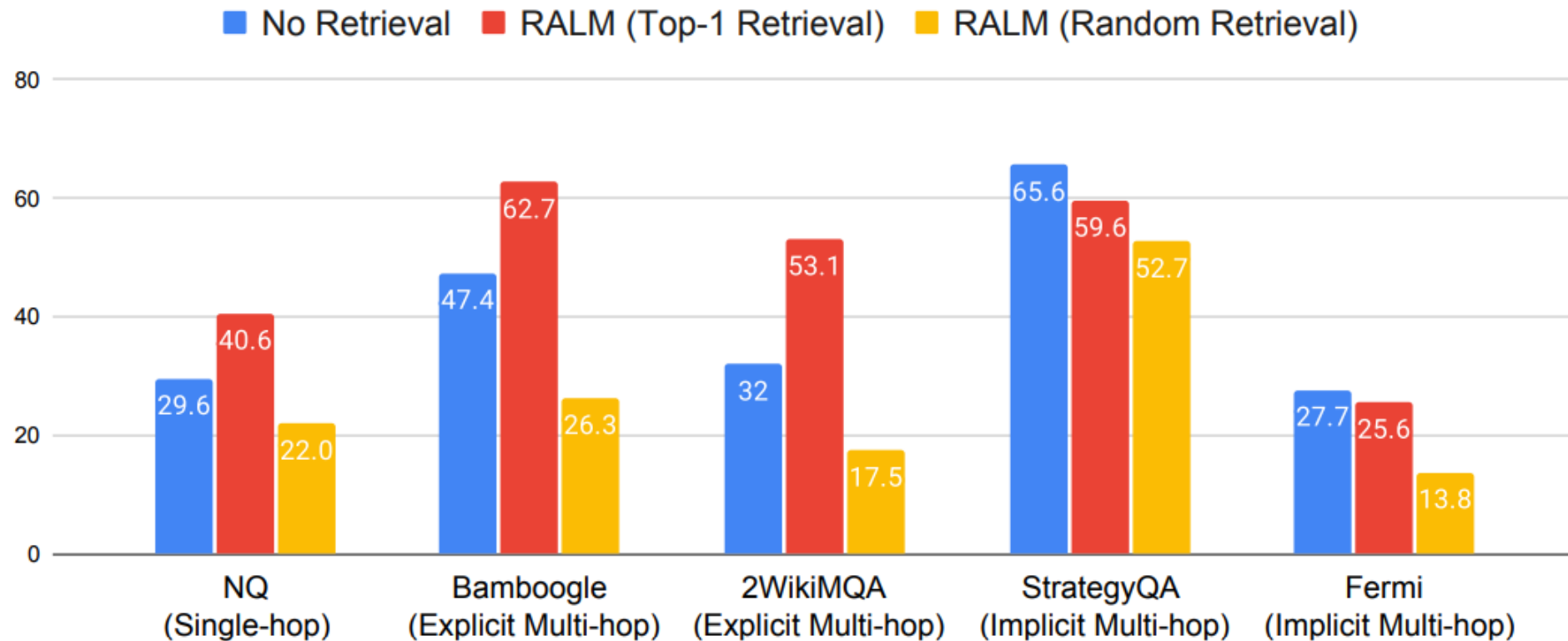
Q: Who is the actor playing Jason on general hospital?

| Large Language Model (no retrieval)         | Retrieval Augmented Language Model                                                                                                                                                                                       |
|---------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>The answer is: Steve Burton</p> <p>✓</p> | <p>E: Jason Gerhardt (born April 21, 1974) is an American actor. He is known for playing the role of Cooper Barrett in General Hospital and Zack Kilmer in Mistresses.</p> <p>The answer is: Jason Gerhardt</p> <p>✗</p> |



- **Relevant** refers to whether the correct answer is in the prompt or not.

# Distraction in RAG



# Solution #1: Use NLI to filter irrelevant context

- Review: NLI models

- Premise:

- *If you help the needy, God will reward you.*

- Hypotheses:

- *Giving money to a poor man has good consequences.*

Entailment

- *Giving money to a poor man has no consequences.*

Contradiction

- *Giving money to a poor man will make you a better person.*

Neutral

- NLI against distraction:

- Remove context sentence if it contradicts the question.

# Solution #2: Finetuning

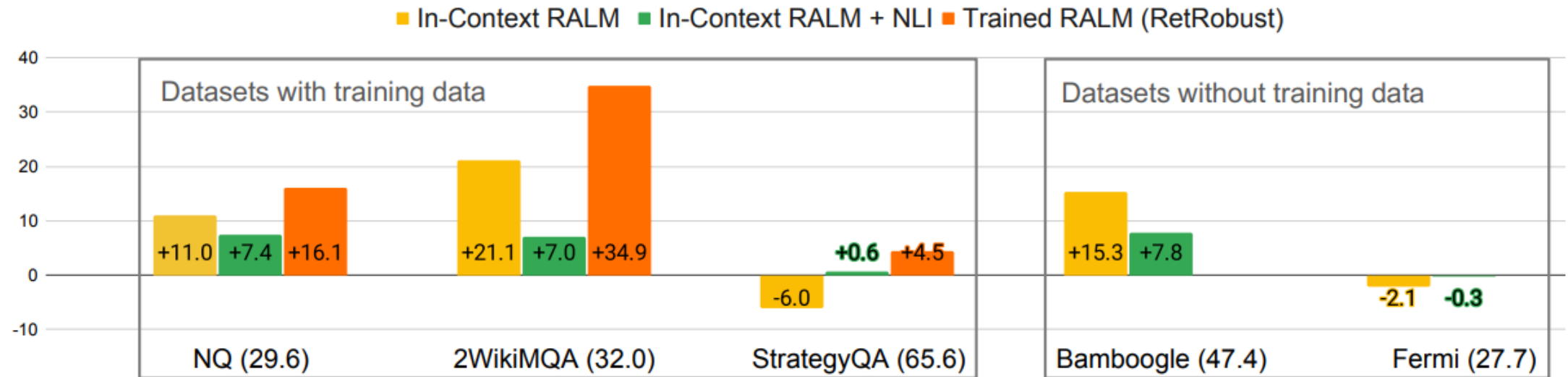
- Fine-tune the LM with:
  - Both relevant and irrelevant contexts

Q: Who is the actor playing Jason on general hospital?

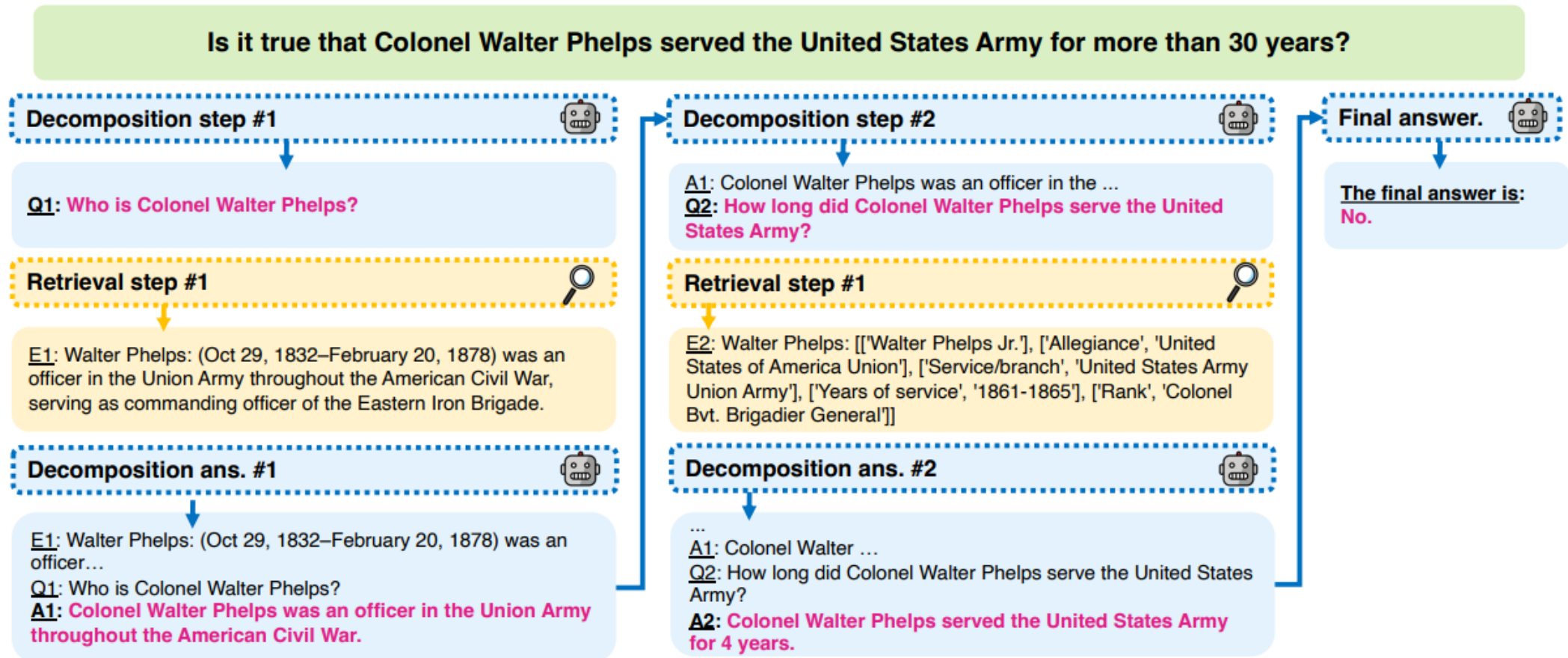
| Large Language Model (no retrieval)         | Retrieval Augmented Language Model                                                                                                                                                                                                               |
|---------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>The answer is: Steve Burton</p> <p>✓</p> | <p>E: Jason Gerhardt (born April 21, 1974) is an American actor. He is known for playing the role of Cooper Barrett in General Hospital and Zack Kilmer in Mistresses.</p> <p>The answer is: <del>Jason Gerhardt</del> Steve Burton</p> <p>✗</p> |



# Distraction in RAG: Mitigation Result



# Solution #3: Interleaving decomposition

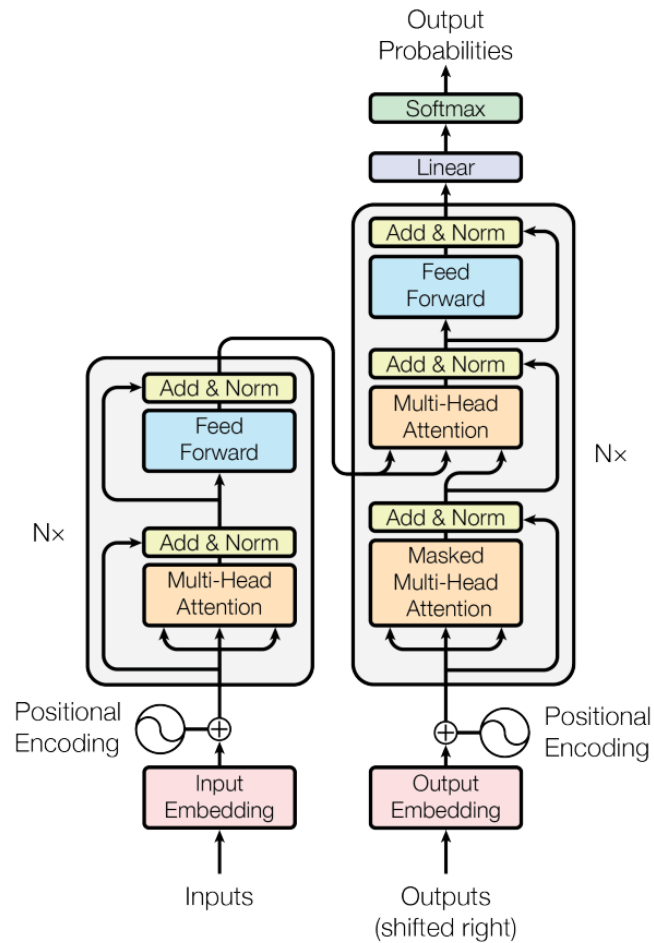


# Prompt Engineering

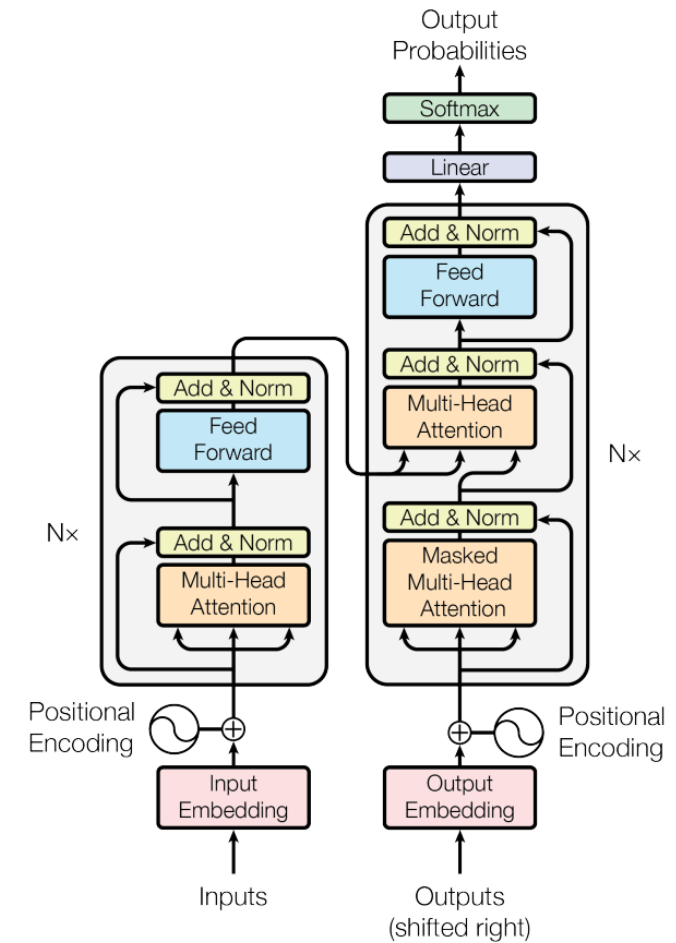
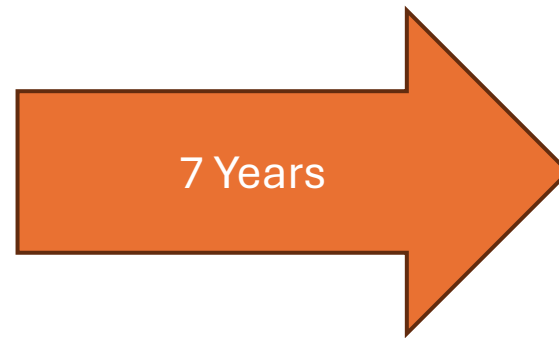
We will return to interleaving decomposition soon



# LLM Architecture Development Progress



Vaswani et al. (2017)  
Attention is All You Need



LLaMA 3.1 (2024)  
GPT-4 (2024)

...

<https://github.com/meta-llama/llama3/blob/main/llama/model.py>

```
class TransformerBlock(nn.Module):
    def __init__(self, layer_id: int, args: ModelArgs):
        super().__init__()
        self.n_heads = args.n_heads
        self.dim = args.dim
        self.head_dim = args.dim // args.n_heads
        self.attention = Attention(args)
        self.feed_forward = FeedForward(
            dim=args.dim,
            hidden_dim=4 * args.dim,
            multiple_of=args.multiple_of,
            ffn_dim_multiplier=args.ffn_dim_multiplier,
        )
        self.layer_id = layer_id
        self.attention_norm = RMSNorm(args.dim, eps=args.norm_eps)
        self.ffn_norm = RMSNorm(args.dim, eps=args.norm_eps)

    def forward(
        self,
        x: torch.Tensor,
        start_pos: int,
        freqs_cis: torch.Tensor,
        mask: Optional[torch.Tensor],
    ):
        h = x + self.attention(self.attention_norm(x), start_pos, freqs_cis, mask)
        out = h + self.feed_forward(self.ffn_norm(h))
        return out
```

Large Language Model

Introducing Meta Llama 3: The most capable  
openly available LLM to date

April 18, 2024





Architectural Changes



Post-training

# Limits of prompting for harder tasks?

- Some tasks seem too hard for even large LMs to learn through prompting alone.
- Especially tasks involving richer, multi-step reasoning.

$$19583 + 29534 = 49117$$

$$98394 + 49384 = 147778$$

$$29382 + 12347 = 41729$$

$$93847 + 39299 = ?$$

**Improvement: change the prompt!**

# Chain-of-thought

## Standard Prompting

### Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

### Model Output

A: The answer is 27. ❌

## Chain-of-Thought Prompting

### Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

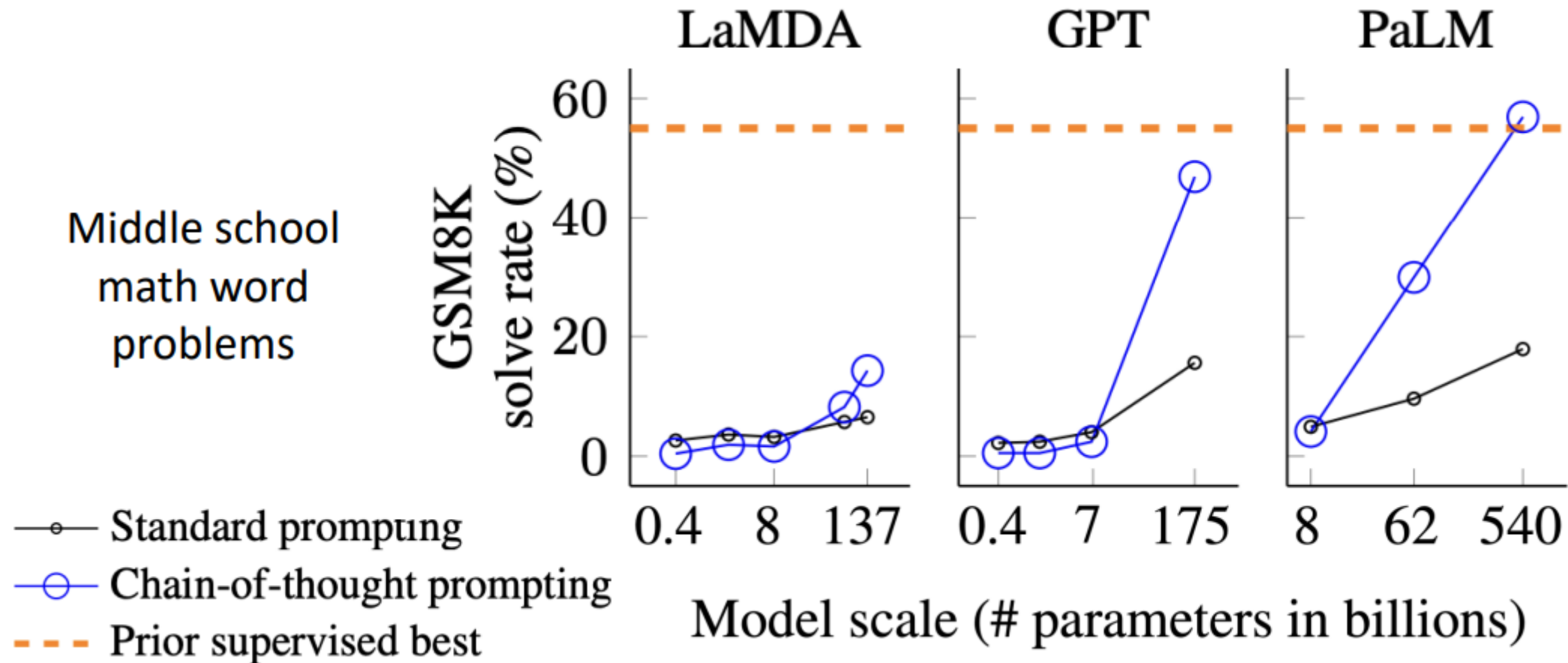
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

### Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had  $23 - 20 = 3$ . They bought 6 more apples, so they have  $3 + 6 = 9$ . The answer is 9. ✅



# Chain-of-thought



[Wei et al., 2022; also see Nye et al., 2021]

# Chain-of-thought

## Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

## Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had  $23 - 20 = 3$ . They bought 6 more apples, so they have  $3 + 6 = 9$ . The answer is 9. ✓

**Do we even need examples of reasoning?**

**Can we just ask the model to reason through things?**

# Chain-of-thought

## Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

## Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had  $23 - 20 = 3$ . They bought 6 more apples, so they have  $3 + 6 = 9$ . The answer is 9. ✓

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step**. There are 16 balls in total. Half of the balls are golf balls. That means there are 8 golf balls. Half of the golf balls are blue. That means there are 4 blue golf balls. ✓

# Zero-shot CoT prompting

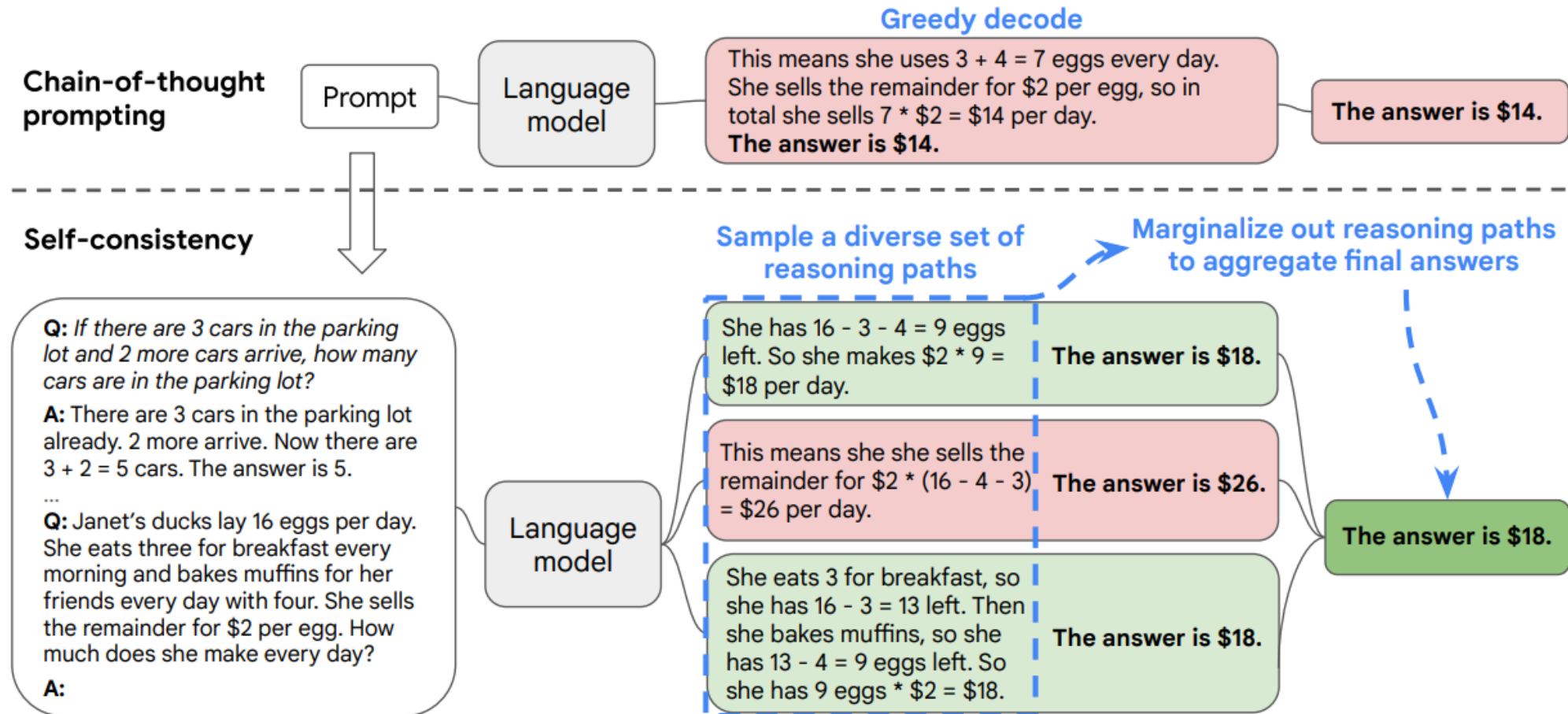
|                                                | MultiArith  | GSM8K       |
|------------------------------------------------|-------------|-------------|
| <b>Zero-Shot</b>                               | <b>17.7</b> | <b>10.4</b> |
| Few-Shot (2 samples)                           | 33.7        | 15.6        |
| Few-Shot (8 samples)                           | 33.8        | 15.6        |
| <b>Zero-Shot-CoT</b>                           | <b>78.7</b> | <b>40.7</b> |
| Few-Shot-CoT (2 samples)                       | 84.8        | 41.3        |
| Few-Shot-CoT (4 samples : First) (*1)          | 89.2        | -           |
| Few-Shot-CoT (4 samples : Second) (*1)         | 90.5        | -           |
| Few-Shot-CoT (8 samples)                       | 93.0        | 48.7        |
| <b>Zero-Plus-Few-Shot-CoT (8 samples) (*2)</b> | <b>92.8</b> | <b>51.5</b> |

Greatly outperforms zero-shot!

Manual CoT still better

# CoT with “Self-consistency”

- Replace greedy decoding with an ensemble of samples...
- **Main idea**: correct reasoning processes have greater agreement than incorrect processes.



# CoT with “Self-consistency”

|                                | GSM8K          | MultiArith     | AQuA           | SVAMP          | CSQA           | ARC-c          |
|--------------------------------|----------------|----------------|----------------|----------------|----------------|----------------|
| Greedy decode                  | 56.5           | 94.7           | 35.8           | 79.0           | 79.0           | 85.2           |
| Weighted avg (unnormalized)    | 56.3 $\pm$ 0.0 | 90.5 $\pm$ 0.0 | 35.8 $\pm$ 0.0 | 73.0 $\pm$ 0.0 | 74.8 $\pm$ 0.0 | 82.3 $\pm$ 0.0 |
| Weighted avg (normalized)      | 22.1 $\pm$ 0.0 | 59.7 $\pm$ 0.0 | 15.7 $\pm$ 0.0 | 40.5 $\pm$ 0.0 | 52.1 $\pm$ 0.0 | 51.7 $\pm$ 0.0 |
| Weighted sum (unnormalized)    | 59.9 $\pm$ 0.0 | 92.2 $\pm$ 0.0 | 38.2 $\pm$ 0.0 | 76.2 $\pm$ 0.0 | 76.2 $\pm$ 0.0 | 83.5 $\pm$ 0.0 |
| Weighted sum (normalized)      | 74.1 $\pm$ 0.0 | 99.3 $\pm$ 0.0 | 48.0 $\pm$ 0.0 | 86.8 $\pm$ 0.0 | 80.7 $\pm$ 0.0 | 88.7 $\pm$ 0.0 |
| Unweighted sum (majority vote) | 74.4 $\pm$ 0.1 | 99.3 $\pm$ 0.0 | 48.3 $\pm$ 0.5 | 86.6 $\pm$ 0.1 | 80.7 $\pm$ 0.1 | 88.7 $\pm$ 0.1 |

Table 1: Accuracy comparison of different answer aggregation strategies on PaLM-540B.

Out-performs regular CoT on a variety of benchmarks

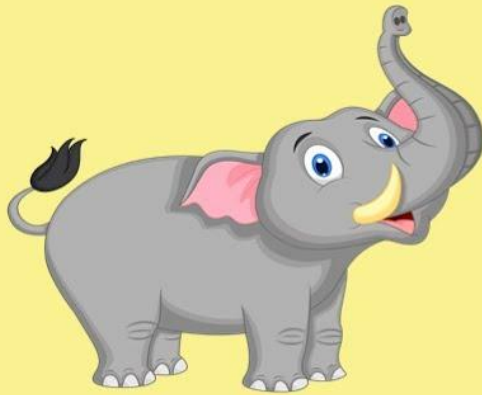
|                                     | GSM8K                            | MultiArith                       | SVAMP                            | ARC-e                            | ARC-c                            |
|-------------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|
| CoT (Wei et al., 2022)              | 17.1                             | 51.8                             | 38.9                             | 75.3                             | 55.1                             |
| Ensemble (3 sets of prompts)        | 18.6 $\pm$ 0.5                   | 57.1 $\pm$ 0.7                   | 42.1 $\pm$ 0.6                   | 76.6 $\pm$ 0.1                   | 57.0 $\pm$ 0.2                   |
| Ensemble (40 prompt permutations)   | 19.2 $\pm$ 0.1                   | 60.9 $\pm$ 0.2                   | 42.7 $\pm$ 0.1                   | 76.9 $\pm$ 0.1                   | 57.0 $\pm$ 0.1                   |
| Self-Consistency (40 sampled paths) | <b>27.7 <math>\pm</math> 0.2</b> | <b>75.7 <math>\pm</math> 0.3</b> | <b>53.3 <math>\pm</math> 0.2</b> | <b>79.3 <math>\pm</math> 0.3</b> | <b>59.8 <math>\pm</math> 0.2</b> |

Self-consistency is doing more than simple ensembling



# Least-to-Most prompting

**How do you put an elephant  
into a refrigerator into 3  
steps?**



# Least-to-Most prompting

- **Idea**: decompose a problem into smaller ones.

## Stage 1: Decompose Question into Subquestions

**Q:** It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The water slide closes in 15 minutes. How many times can she slide before it closes?

Language  
Model

**A:** To solve “How many times can she slide before it closes?”, we need to first solve: “How long does each trip take?”



# Least-to-Most prompting

- **Idea**: decompose a problem into smaller ones.

## Stage 1: Decompose Question into Subquestions

**Q:** It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The water slide closes in 15 minutes. How many times can she slide before it closes?

Language Model

**A:** To solve "How many times can she slide before it closes?", we need to first solve: "How long does each trip take?"

## Stage 2: Sequentially Solve Subquestions

Subquestion 1

It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The slide closes in 15 minutes.

**Q:** How long does each trip take?

Language Model

**A:** It takes Amy 4 minutes to climb and 1 minute to slide down.  $4 + 1 = 5$ . So each trip takes 5 minutes.

Append model answer to Subquestion 1

It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The slide closes in 15 minutes.

**Q:** How long does each trip take?

**A:** It takes Amy 4 minutes to climb and 1 minute to slide down.  $4 + 1 = 5$ . So each trip takes 5 minutes.

Language Model

**A:** The water slide closes in 15 minutes. Each trip takes 5 minutes. So Amy can slide  $15 \div 5 = 3$  times before it closes.

Subquestion 2

**Q:** How many times can she slide before it closes?

# Least-to-Most prompting for Math Reasoning

Q: Elsa has 5 apples. Anna has 2 more apples than Elsa. How many apples do they have together?  
A: Anna has 2 more apples than Elsa, so Anna has  $2 + 5 = 7$  apples. Elsa and Anna have  $5 + 7 = 12$  apples together. The answer is 12.

Standard CoT prompting

Q: Elsa has 5 apples. Anna has 2 more apples than Elsa. How many apples do they have together?  
A: Let's break down this problem: 1. How many apples does Anna have? 2. How many apples do Elsa and Anna have together?  
1. Anna has 2 more apples than Elsa. So Anna has  $2 + 5 = 7$  apples.  
2. Elsa and Anna have  $5 + 7 = 12$  apples together.

Least-to-most prompting

Q: {question}  
A: Let's break down this problem:  
—  
The answer is:

# Least-to-Most prompting for Math Reasoning

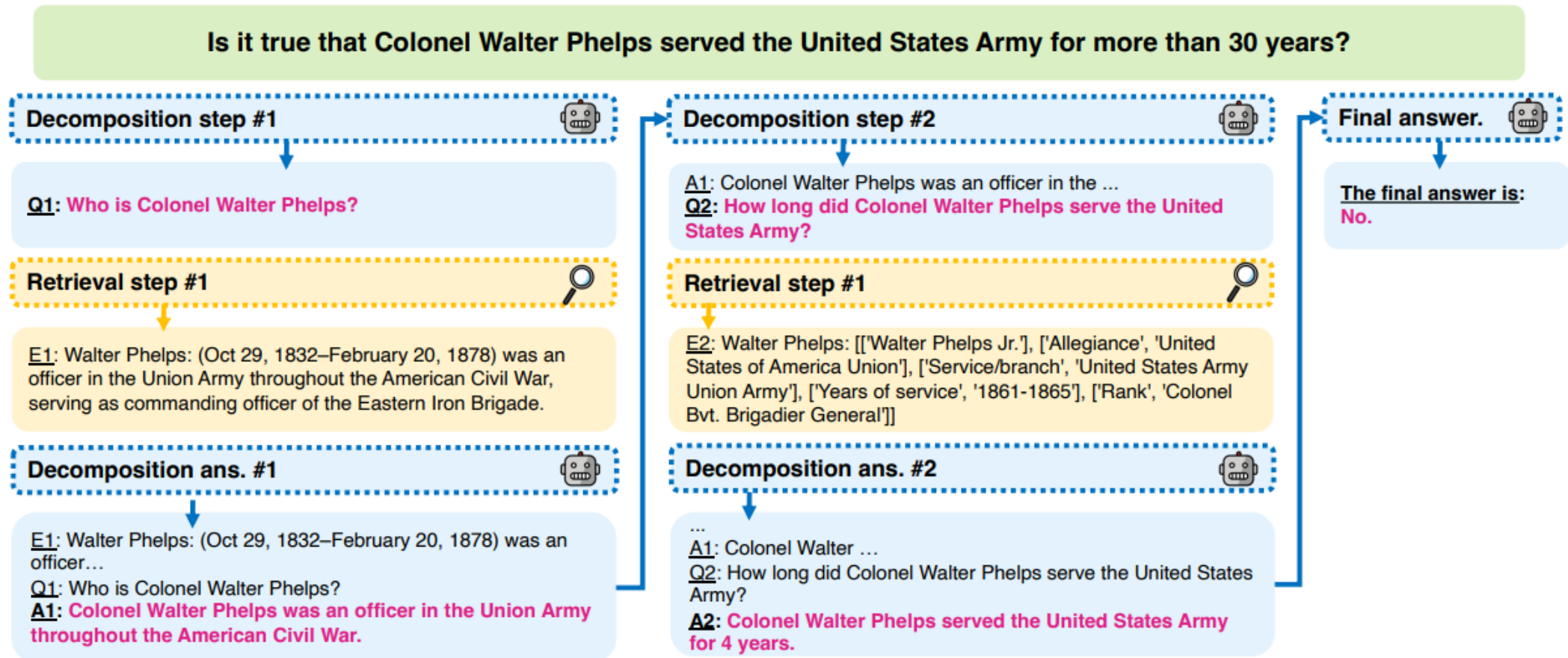
| Accuracy by Steps (GSM8K) | All          | 2 Steps      | 3 Steps      | 4 steps      | $\geq 5$ steps |
|---------------------------|--------------|--------------|--------------|--------------|----------------|
| Least-to-Most             | <b>62.39</b> | 74.53        | <b>68.91</b> | <b>59.73</b> | <b>45.23</b>   |
| Chain-of-Thought          | 60.87        | <b>76.68</b> | 67.29        | 59.39        | 39.07          |

Generalizes to more #steps than in-context example!

| Prompting method            | Accuracy                 |
|-----------------------------|--------------------------|
| Zero-Shot                   | 16.38                    |
| Standard prompting          | 17.06 <sup>3</sup>       |
| Chain-of-Thought (original) | 61.18                    |
| Chain-of-Thought (1-shot)   | 60.88                    |
| Least-to-Most (1-shot)      | 62.39                    |
| Chain-of-Thought (best)     | <b>68.61<sup>3</sup></b> |
| Least-to-Most (best)        | 68.01                    |

But with enough prompt engineering, CoT  $\approx$  Least-to-Most

# Solution #3: Interleaving decomposition



# Solution #3: Interleaving decomposition SelfAsk

## Direct Prompting

GPT-3

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?

Answer: Harry Vaughan Watkins.

Question: Who was president of the U.S. when superconductivity was discovered?

Answer: Franklin D. Roosevelt



## Chain of Thought

GPT-3

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?

Answer: Theodor Haecker was 65 years old when he died. Harry Vaughan Watkins was 69 years old when he died.

So the final answer (the name of the person) is: Harry Vaughan Watkins.

Question: Who was president of the U.S. when superconductivity was discovered?

Answer: Superconductivity was discovered in 1911 by Heike Kamerlingh Onnes. Woodrow Wilson was president of the United States from 1913 to 1921. So the final answer (the name of the president) is: Woodrow Wilson.



## Self-Ask

GPT-3

Question: Who lived longer, Theodor Haecker or Harry Vaughan Watkins?

Are follow up questions needed here: Yes.

Follow up: How old was Theodor Haecker when he died?

Intermediate answer: Theodor Haecker was 65 years old when he died.

Follow up: How old was Harry Vaughan Watkins when he died?

Intermediate answer: Harry Vaughan Watkins was 69 years old when he died.

So the final answer is: Harry Vaughan Watkins

Question: Who was president of the U.S. when superconductivity was discovered?

Are follow up questions needed here: Yes.

Follow up: When was superconductivity discovered?

Intermediate answer: Superconductivity was discovered in 1911.

Follow up: Who was president of the U.S. in 1911?

Intermediate answer: William Howard Taft.

So the final answer is: William Howard Taft.



# Solution #3: Interleaving decomposition

## SelfAsk

|                    | Bamb.       | 2Wiki.      | Musique     |
|--------------------|-------------|-------------|-------------|
| Direct prompting   | 17.6        | 25.4        | 5.6         |
| Chain of Thought   | 46.4        | 29.8        | 12.6        |
| Search             | 0.0         | 2.2         | 1.5         |
| Search + postproc. | -           | 26.3        | 6.5         |
| Self-ask           | 57.6        | 30.0        | 13.8        |
| Self-ask + Search  | <b>60.0</b> | <b>40.1</b> | <b>15.2</b> |

|               | 2Wiki.      |            | Musique     |            |
|---------------|-------------|------------|-------------|------------|
|               | Acc. ↑      | # Toks ↓   | Acc. ↑      | # Toks ↓   |
| Least-to-Most | 29.0        | 844        | <b>16.8</b> | 1020       |
| Self-ask      | <b>35.5</b> | <b>569</b> | <b>16.3</b> | <b>663</b> |



# ChatGPT-o1

- We know very little about how exactly it is built.
- OpenAI released very little about its implementation details.
- But we have an idea:
  - Chain-of-thought
  - Use reinforcement learning (similar to RLHF) to improve the CoT process.

A large, white, sans-serif 'o1' is centered on the right side of the slide. The background is a vibrant yellow with a subtle blue and green circular pattern, resembling a stylized sun or a galaxy.

# Last Quiz

- Which of the following is not a prompt engineering technique?
  - A. Adding retrieved context or examples to a prompt
  - B. Using specific instructions in the prompt
  - C. Changing the programming language used to implement the model
  - D. Asking for outputs with the “thinking process”



May You  
Live In  
Interesting  
Times

BIENNALE ARTE  
2019

11.05—24.11  
VENEZIA  
GIARDINI/ARSENALE