Taxonomy of QA Datasets and Tasks

연세대학교 컴퓨터과학과

김승원

Why is QA So Important? (My Ideas)

- **Everything in Deep Learning is (actually) QA**
  - *Classification* => "Upon Cat and Dog, which one do you think this image is?"
  - *Summarization* => "What is a golden summary of this article?"
  - *Relationship Extraction* => "What do you think is the relationship between US and George Washington?"
  - *Translation* => "What is this Korean translation of this English article?"
  - *Generation* => "What are some other likely images of this cat image?"

- **In NLP, QA is the best way to measure a LM's Language Understanding Abilities**
  - **Memorization vs Reasoning**
  - Does the LM understand the question?
  - Can the LM map the best answer given the question?
  - Can the LM logically reason to give the best answer when a complex question is given?
  - Can the LM give an expressive answer when a question that requires generative explanation is given?

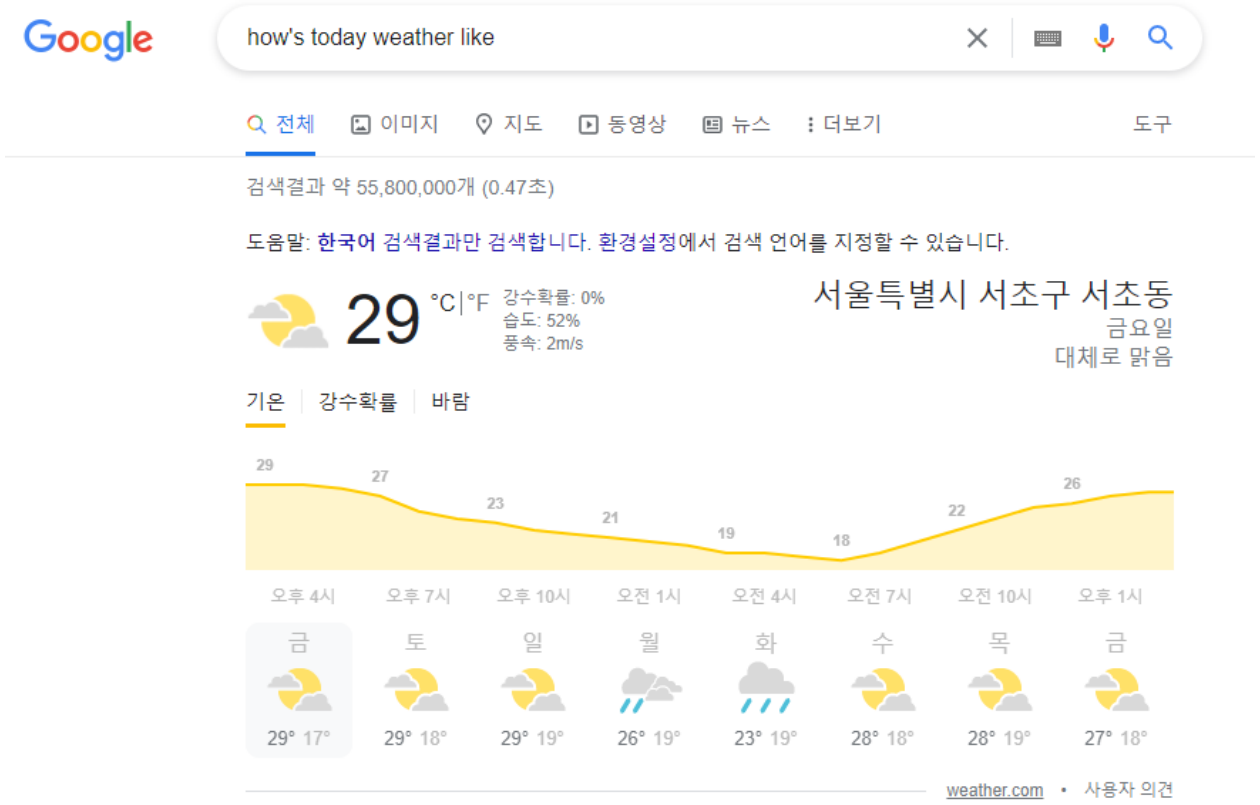
What we will study in Yonsei NLP Study

- **Research Interests**
  - Injection of Commonsense Knowledge & Domain Knowledge into Language Models
  - Logical Reasoning for better Reading Comprehension with Language Models

- **Presentation Plan**
  - **9/8 : Taxonomy of QA Datasets and Tasks**
  - **9/22 : Knowledge Bases and Language Models**
  - **10/27 : Information Retrieval Systems**
  - **11/10 : Commonsense & Domain Knowledge Injection**
  - **11/24 : Machine Reasoning**

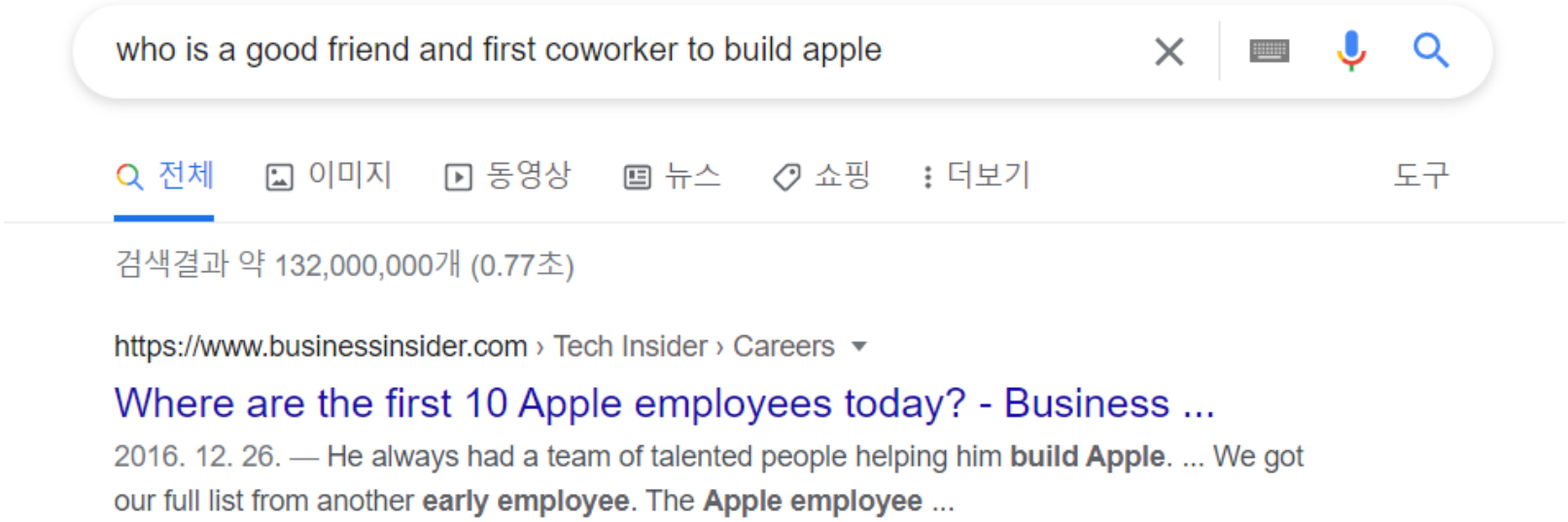
Why not just GOOGLE?

- What Google is Good At
  - Providing answers to the given questions with visualization in a summarized format



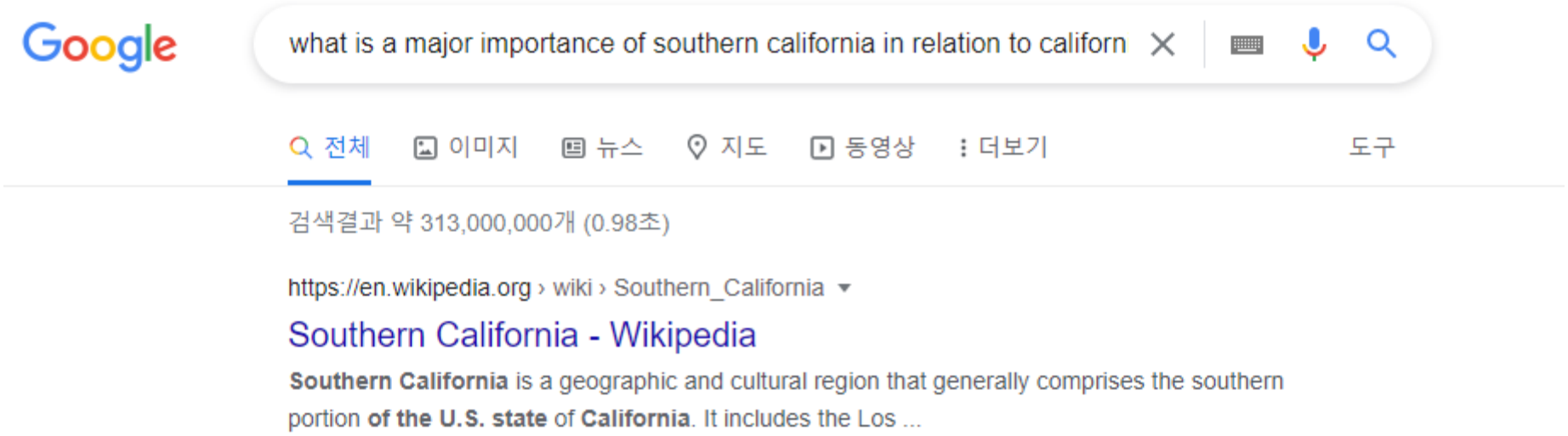
Why not just GOOGLE?

- What Google is **NOT** Good At
  - If the question varies a little bit, it cannot understand the intention of the question
  - This is where Deep Learning based QA Models could tackle into



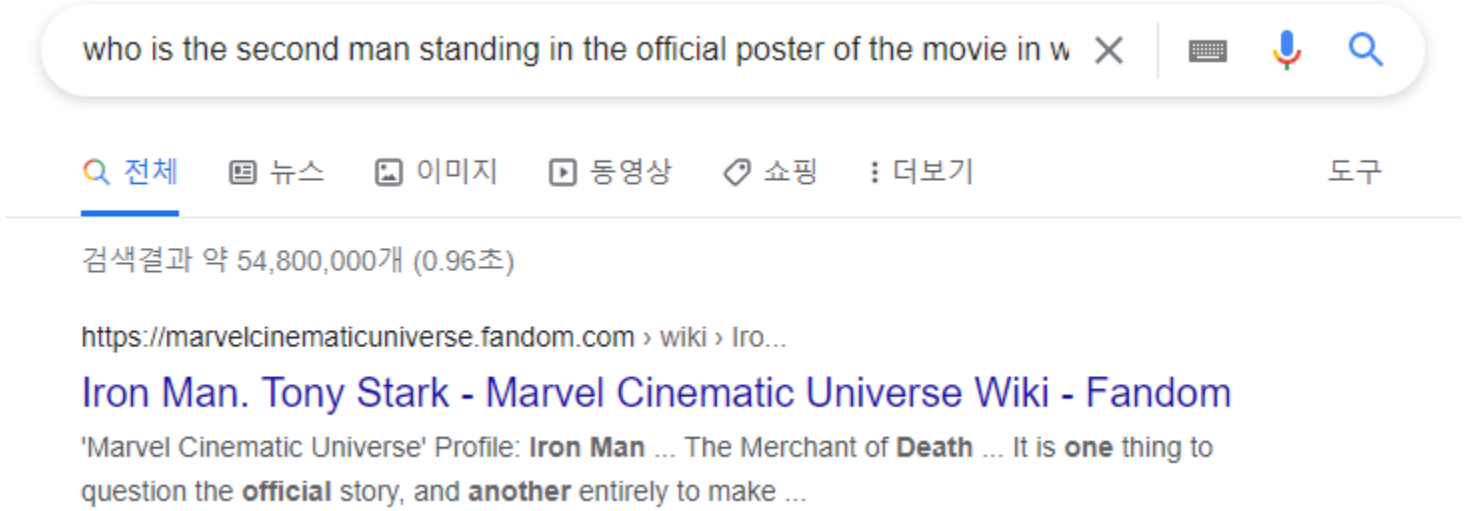
Why not just GOOGLE?

- What Google is **NOT** Good At
  - If the question is too long and specific, it CANNOT even find a relevant context passage.
  - This is where Deep Learning based QA Models could tackle into



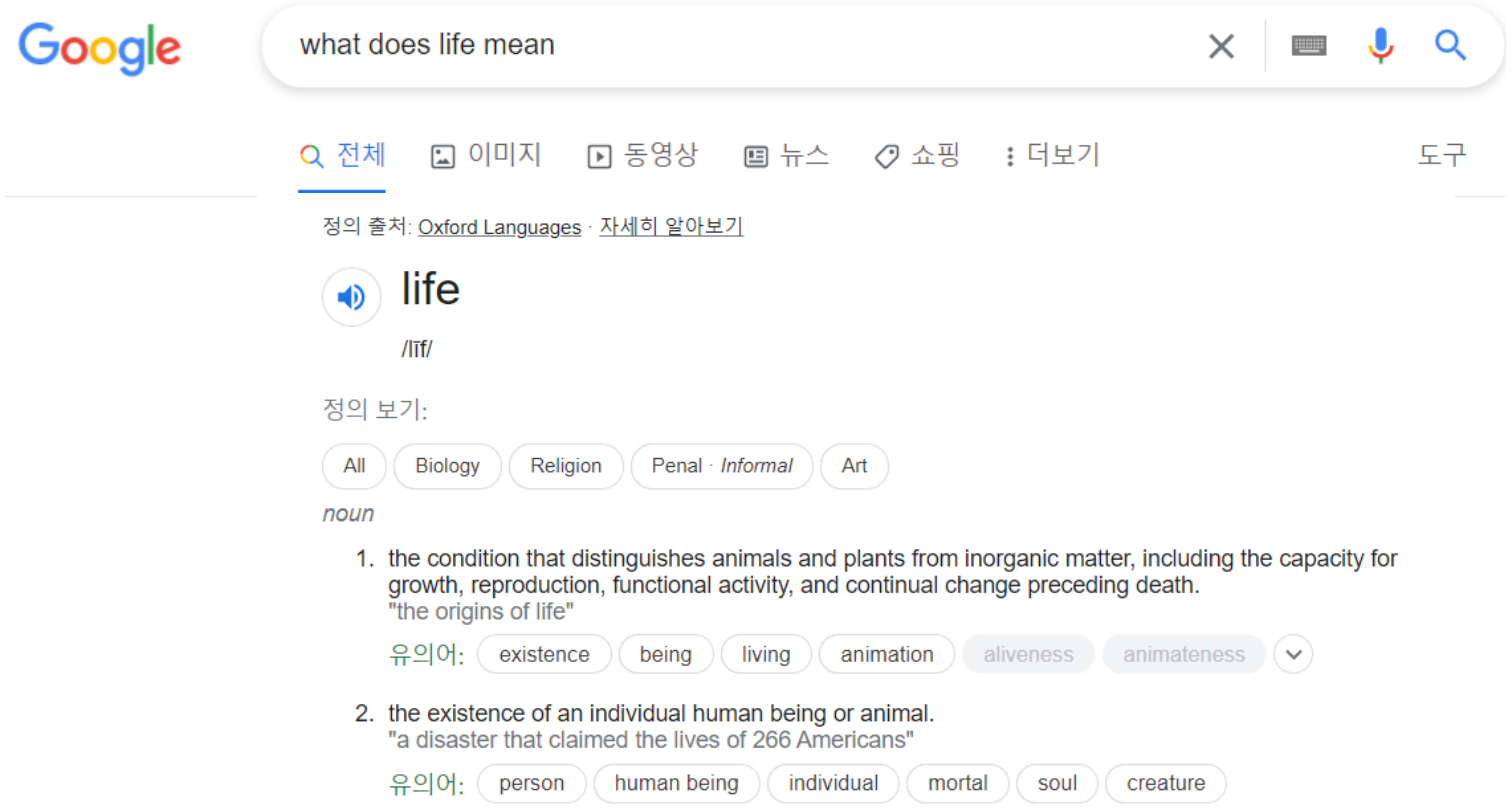
Why not just GOOGLE?

- What Google is **NOT** Good At
  - If the question contains complex relationships between entities, it cannot understand the intention of the question
  - This is where Deep Learning based QA Models could tackle into



Why not just GOOGLE?

- What Google is **NOT** Good At
  - If the question requires a descriptive explanation, it cannot understand the intention of the question
  - This is where Deep Learning based QA Models could tackle into



Basic Template in QA Datasets

Question

Southern California, often abbreviated SoCal, is a geographic and cultural region that generally comprises California's southernmost 10 counties. The region is traditionally described as "eight counties", based on demographics and economic ties: Imperial, Los Angeles, Orange, Riverside, San Bernardino, San Diego, Santa Barbara, and Ventura. The more extensive 10-county definition, including Kern and San Luis Obispo counties, is also used based on historical political divisions. Southern California is a major economic center for the state of California and the United States.

What is Southern California often abbreviated as?
Ground Truth Answers: SoCal | SoCal | SoCal
Prediction: SoCal

Answer

Despite being traditionall described as "eight counties", how many counties does this region actually have?
Ground Truth Answers: 10 counties | 10 | 10
Prediction: <No Answer>

What is a major importance of Southern California in relation to California and the United States?
Ground Truth Answers: economic center | major economic center | economic center
Prediction: economic center

What are the ties that best described what the "eight counties" are based on?
Ground Truth Answers: demographics and economic ties | economic | demographics and economic
Prediction: demographics and economic ties

Context Passage
(If not given, use IR system)

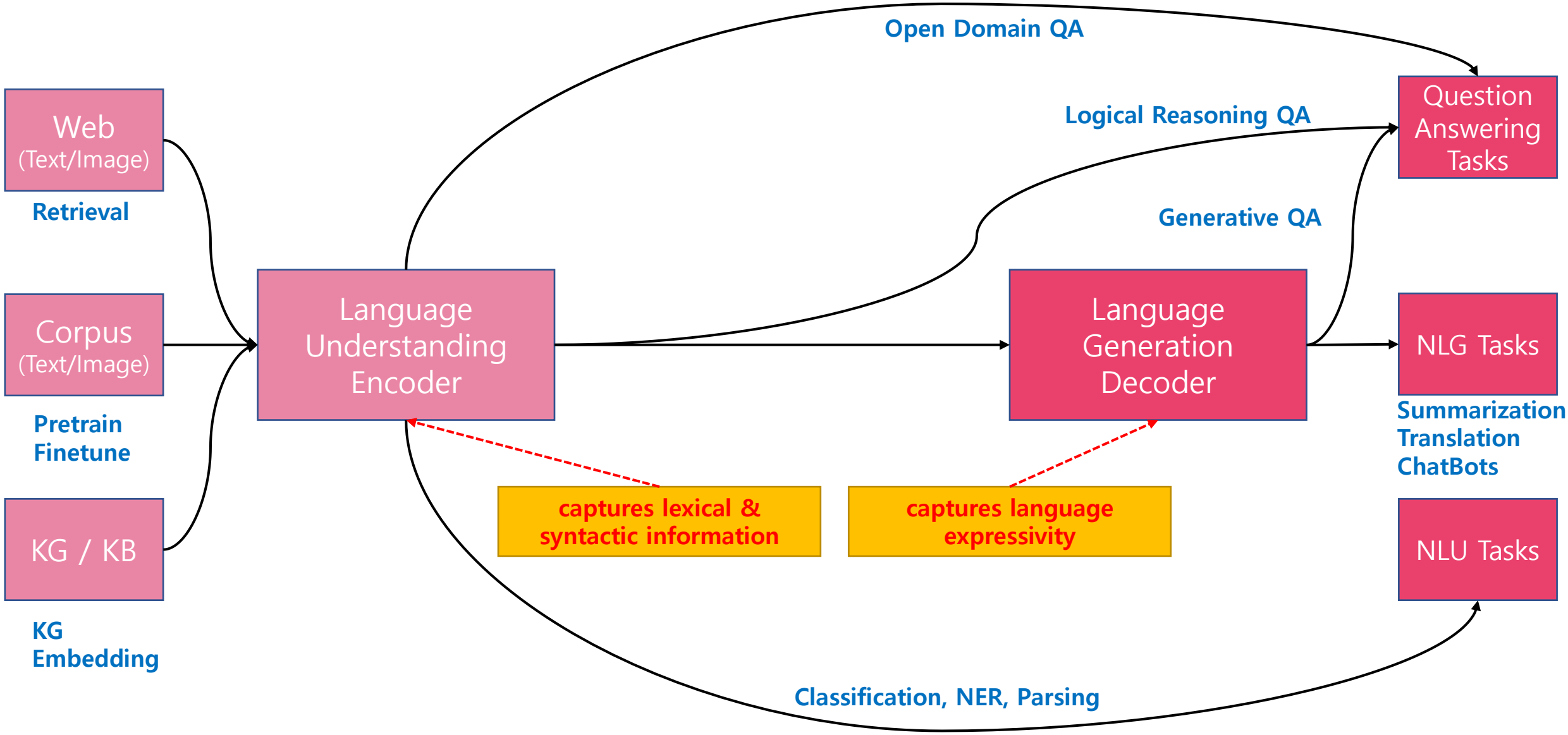
Assume that the answer to the given question is a span within the given context

Variations given to Basic Template

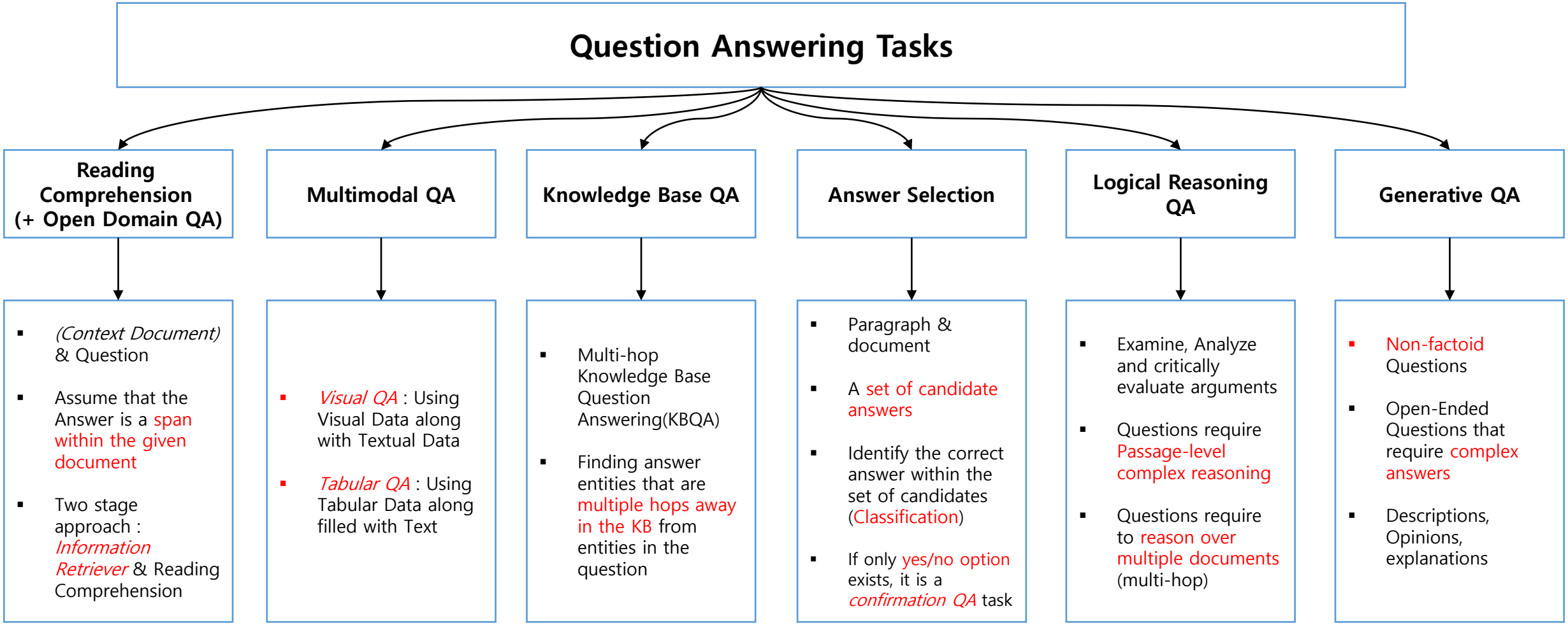
- **What if Context Passage is not given?**
  - Find a Passage on the Internet that will most likely have the answer to the given question => **RETRIEVAL**
  - (Search Engines) ElasticSearch
  - (Retrieval Rank Methods) TF-IDF, BM25, DPR
- **What if the answer is not within the context?**
  - Find other related documents, and multi-hop reason upon multiple documents => **MULTI-HOP REASONING**
  - Understanding the question, perform logical reasoning to answer the question => **LOGICAL REASONING**
- **What about 'yes/no', 'multiple choice' or 'generative' questions?**
  - Think as a classification problem
  - Choose the most likely answer given the candidates of answers => **ANSWER SELECTION / CONFIRMATION QA**
  - Consider whether there might be no answers at all, or there might be multiple of them
  - With generation abilities, give an answer that best maps what the question required => **GENERATIVE ANSWERING**

QA Tasks

Current Approaches

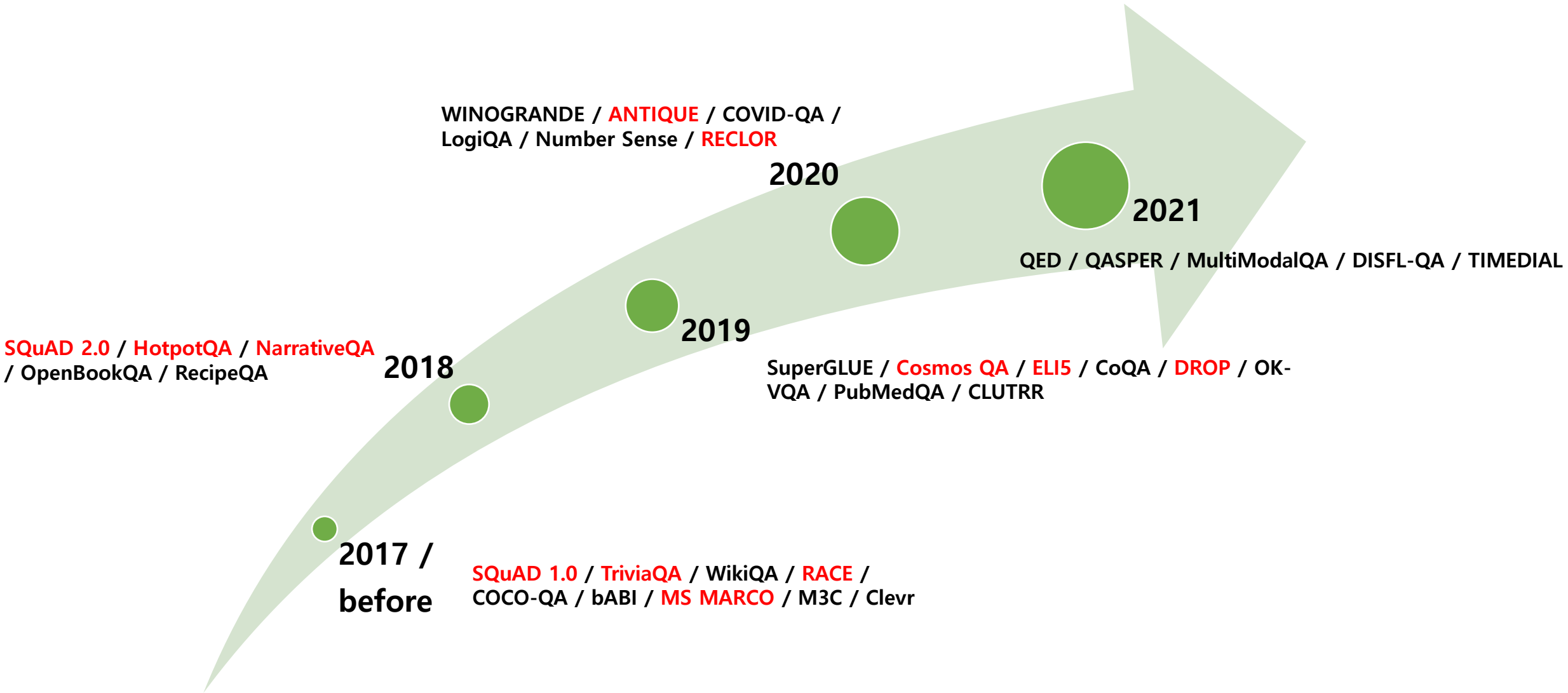


Question Answering Tasks Taxonomy



QA Datasets

33 Question Answering Datasets



Datasets before 2018

Dataset	Query Source	Answer	#Queries	#Context	Size	Tasks	Metric
WikiQA (EMNLP 2015)	User logs	Sentence Selection	3047	29.26K Sentences	12.87MB	QA	MAP / MRR
SQuAD 1.0 (EMNLP 2016)	Crowdsourced	Span of Words	100K	536 Docs	119.27 MB	QA	EM / F1
MS MARCO (NeurIPS 2016)	User logs	Human Generated	100K	1M Passages - 200K+ Docs	575.36 MB	QA - Retrieval	ROUGE-L BLEU-1
Trivia QA (ACL 2017)	Crawled from Websites	Span of Words	174K	662K+ Docs	2662.71 MB	QA	EM / F1
RACE (EMNLP 2017)	From Middle / High School English Exams	Answer Selection	125K	970K Passages	190.90 MB	QA	Accuracy

- Reading Comprehension and Open Domain QA is one of the domains that QA systems struggle to solve.
- **(Query)** Crowdsourced(Engineered by crowd workers, less creativity) vs User Logs(Extracted from real Web documents, people rarely ask interesting questions)
- **(Answer)** Extractive(Limits QA Models to learn expressiveness) vs Human Generated(Hard to evaluate answer; ROUGE)

SQuAD 1.0 (EMNLP 2016)

- The GOLDEN STANDARD to Reading Comprehension
- Provides a public leaderboard to submit models
- The Answer is a span within the given context

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".

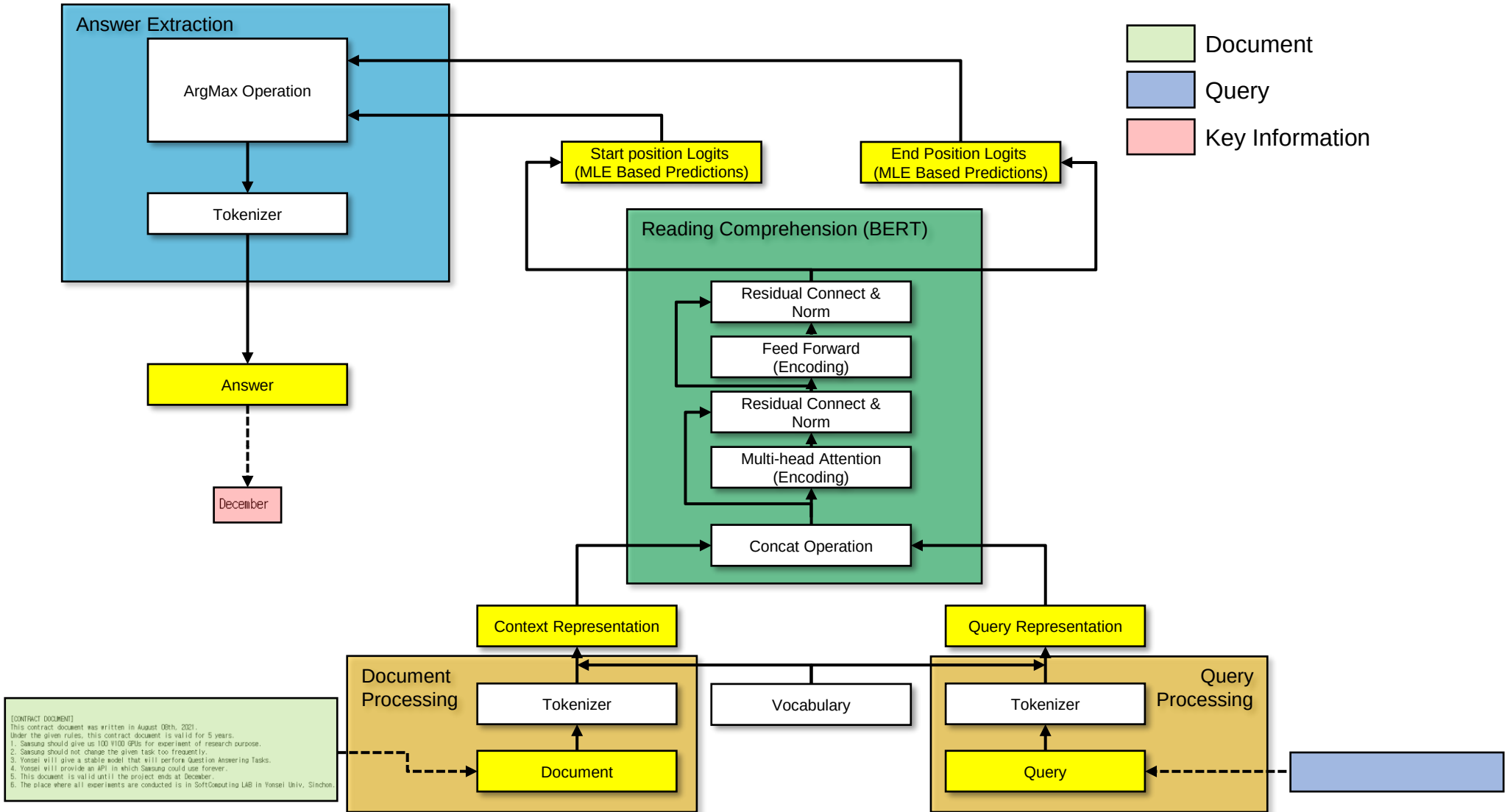
What causes precipitation to fall?
gravity

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

Answer type	Percentage	Example
Date	8.9%	19 October 1512
Other Numeric	10.9%	12
Person	12.9%	Thomas Coke
Location	4.4%	Germany
Other Entity	15.3%	ABC Sports
Common Noun Phrase	31.8%	property damage
Adjective Phrase	3.9%	second-largest
Verb Phrase	5.5%	returned to Earth
Clause	3.7%	to avoid trivialization
Other	2.7%	quietly

Finetuning BERT for SQuAD



MS MARCO Dataset (NeurIPS 2016)

- A harder dataset compared to SQuAD 1.0
- Requires Generative Answers
- The Answer might be in multiple passages, in the question itself, or even not in either of them

The answer is an exact text span in the passage.

Q: how tall is jack griffo

P: Jack Griffo Height : **5'6 (167.64 cm)**. Standing at a height of 5 feet, 6 inches tall Jack Griffo is taller than 11.9% of all men, as reflected by the figure's fill height %.Conversely, at this height Jack Griffo is not as tall as 88.1% of all men.

A: 5'6 (167.64 cm)

All words in the answer are in the passage but from multiple text spans.

Q: who did odysseus see in the underworld

P_1: The souls that Odysseus saw in the Underworld On seeing **Achilles'** soul, said Odysseus: Achilles, the most fortunate man that ever was or will be honored as though you were a god and now you are a mighty prince among the dead.

P_2: Odysseus talked to his mother Anticlea, who died of grief when he did not return home after the Trojan War. Odysseus was also surprised to see **Elphenor**, the youngest member of his crew, in the Underworld.

A: Elphenor and Achilles.

All words in the answer are in the passage and question.

Q: what do **producers need to make food**

P: Plants are producers. Producers are living things that can make their own food using **air, light, soil, and water**. Plants use a process called photosynthesis to make food.

A: Producers need air, light, soil, and water to make food.

Part of words in the answer are not found in the passage or question.

Q: why conversion observed in body

P: Conversion disorder **symptoms** may appear suddenly after a stressful event or trauma, whether physical or psychological. Signs and symptoms that affect movement function may include: 1 Weakness or paralysis. 2 Abnormal movement, such as tremors or difficulty walking. 3 Loss of balance.

A: Due to symptoms in the body

All Words in the answer are not found in the passages or question.

Q: is there an age limit for learning speech

P: Age is not a detriment to language learning, and by all accounts, learning a second (or third etc) language actually keeps the older language learners mind active. People of all ages can benefit from learning languages.

A: No

Query contains	Percentage of queries
what	42.2%
how	15.3%
where	4.4%
when	2.0%
why	1.8%
who	1.7%
which	1.4%

Table 3: Percentage of queries containing question keywords

Answer type	Percentage of queries
Description	52.6%
Numeric	28.4%
Entity	10.5%
Location	5.7%
Person	2.7%

Table 4: Distribution of queries based on answer-type classifier

MS MARCO Dataset (NeurIPS 2016)

- Questions from Search engines will better represent actual human information seeking needs
- Questions from Search engines are more complex to answer compared to artificially generated questions

To solve for these types of questions we need a system with human level reading comprehension and reasoning abilities. E.g., given a query such as *{will I qualify for osap if i'm new in canada}* as shown in figure 2 one of the relevant passages include:

You must be a 1. Canadian citizen, 2. Permanent Resident or 3. Protected person

A RC model needs to parse and understand that being new to a country is usually the opposite of citizen, permanent resident, etc. This is not a simple task to do in a general way. As part of our dataset quality control process, we noticed that even human judges had a hard time reaching this type of conclusions, especially for content belonging to areas they were not familiar with.

Trivia QA Dataset (ACL 2017)

- A harder dataset compared to SQuAD
- Like SQuAD, the Answer is a span within the Context(Excerpt)
- The questions are complex(have compositional semantics), finding correct answer requires complex reasoning(combining facts from multiple sentences or background knowledge) and individual facts can be difficult to recover from text (due to lexical and syntactic variation)

Question: The Dodecanese Campaign of WWII that was an attempt by the Allied forces to capture islands in the Aegean Sea was the inspiration for which acclaimed 1961 commando film?
Answer: The Guns of Navarone
Excerpt: The Dodecanese Campaign of World War II was an attempt by Allied forces to capture the Italian-held Dodecanese islands in the Aegean Sea following the surrender of Italy in September 1943, and use them as bases against the German-controlled Balkans. The failed campaign, and in particular the Battle of Leros, inspired the 1957 novel **The Guns of Navarone** and the successful 1961 movie of the same name.

Question: American Callan Pinckney's eponymously named system became a best-selling (1980s-2000s) book/video franchise in what genre?
Answer: Fitness
Excerpt: Callan Pinckney was an American fitness professional. She achieved unprecedented success with her Callanetics exercises. Her 9 books all became international best-sellers and the video series that followed went on to sell over 6 million copies. Pinckney's first video release "Callanetics: 10 Years Younger In 10 Hours" outsold every other **fitness** video in the US.

Property	Example annotation	Statistics
Avg. entities / question	Which politician won the **Nobel Peace Prize** in 2009?	1.77 per question
Fine grained answer type	What **fragrant essential oil** is obtained from Damask Rose?	73.5% of questions
Coarse grained answer type	**Who** won the Nobel Peace Prize in 2009?	15.5% of questions
Time frame	What was photographed for the first time in **October 1959**	34% of questions
Comparisons	What is the appropriate name of the **largest** type of frog?	9% of questions

Type	Percentage
Numerical	4.17
Free text	2.98
Wikipedia title	92.85
Person	32
Location	23
Organization	5
Misc.	40

Table 4: Distribution of answer types on 200 annotated examples.

Trivia QA Dataset (ACL 2017)

Reasoning	**Lexical variation (synonym)**
	Major correspondences between the question and the answer sentence are synonyms.
Frequency	41% in Wiki documents, 39% in web documents.
Examples	Q What is solid CO2 commonly called?
	S The frozen solid form of CO2, known as **dry ice** ...
	Q Who wrote the novel The Eagle Has landed?
	S The Eagle Has Landed is a book by British writer **Jack Higgins**
Reasoning	**Lexical variation and world knowledge**
	Major correspondences between the question and the document require common sense or external knowledge.
Frequency	17% in Wiki documents, 17% in web documents.
Examples	Q What is the first name of Madame Bovary in Flaubert's 1856 novel?
	S Madame Bovary (1856) is the French writer Gustave Flaubert's debut novel. The story focuses on a doctor's wife, **Emma** Bovary
	Q Who was the female member of the 1980's pop music duo, Eurythmics?
	S Eurythmics were a British music duo consisting of members **Annie Lennox** and David A. Stewart.
Reasoning	**Syntactic Variation**
	After the question is paraphrased into declarative form, its syntactic dependency structure does not match that of the answer sentence
Frequency	69% in Wiki documents, 65% in web documents.
Examples	Q In which country did the Battle of El Alamein take place?
	S The 1942 Battle of El Alamein in **Egypt** was actually two pivotal battles of World War II
	Q Whom was Ronald Reagan referring to when he uttered the famous phrase evil empire in a 1983 speech?
	S The phrase evil empire was first applied to the **Soviet Union** in 1983 by U.S. President Ronald Reagan.
Reasoning	**Multiple sentences**
	Requires reasoning over multiple sentences.
Frequency	40% in Wiki documents, 35% in web documents.
Examples	Q Name the Greek Mythological hero who killed the gorgon Medusa.
	S **Perseus** asks god to aid him. So the goddess Athena and Hermes helps him out to kill Medusa.
	Q Who starred in and directed the 1993 film A Bronx Tale?
	S **Robert De Niro** To Make His Broadway Directorial Debut With A Bronx Tale: The Musical. The actor starred and directed the 1993 film.
Reasoning	**Lists, Table**
	Answer found in tables or lists
Frequency	7% in web documents.
Examples	Q In Moh's Scale of hardness, Talc is at number 1, but what is number 2?
	Q What is the collective name for a group of hawks or falcons?

RACE Dataset (EMNLP 2017)

- A harder dataset compared to SQuAD
- Among the 4 possible answers, after reading the Context(Passage) and the question, the model should choose the most likely answer
- Questions require Detail reasoning, Whole Picture Reasoning, Passage Summarization, Attitude Analysis and World Knowledge

Passage:

In a small village in England about 150 years ago, a mail coach was standing on the street. It didn't come to that village often. People had to pay a lot to get a letter. The person who sent the letter didn't have to pay the postage, while the receiver had to.
"Here's a letter for Miss Alice Brown," said the mailman.
" I'm Alice Brown," a girl of about 18 said in a low voice.
Alice looked at the envelope for a minute, and then handed it back to the mailman.
"I'm sorry I can't take it, I don't have enough money to pay it", she said.
A gentleman standing around were very sorry for her. Then he came up and paid the postage for her.
When the gentleman gave the letter to her, she said with a smile, " Thank you very much, This letter is from Tom. I'm going to marry him. He went to London to look for work. I've waited a long time for this letter, but now I don't need it, there is nothing in it."
"Really? How do you know that?" the gentleman said in surprise.
"He told me that he would put some signs on the envelope. Look, sir, this cross in the corner means that he is well and this circle means he has found work. That's good news."
The gentleman was Sir Rowland Hill. He didn't forgot Alice and her letter.
"The postage to be paid by the receiver has to be changed," he said to himself and had a good plan.
"The postage has to be much lower, what about a penny? And the person who sends the letter pays the postage. He has to buy a stamp and put it on the envelope." he said . The government accepted his plan. Then the first stamp was put out in 1840. It was called the "Penny Black". It had a picture of the Queen on it.

Questions:

1): The first postage stamp was made _.
A. in England B. in America C. by Alice D. in 1910

2): The girl handed the letter back to the mailman because _.
A. she didn't know whose letter it was
B. she had no money to pay the postage
C. she received the letter but she didn't want to open it
D. she had already known what was written in the letter

3): We can know from Alice's words that _.
A. Tom had told her what the signs meant before leaving
B. Alice was clever and could guess the meaning of the signs
C. Alice had put the signs on the envelope herself
D. Tom had put the signs as Alice had told him to

4): The idea of using stamps was thought of by _.
A. the government
B. Sir Rowland Hill
C. Alice Brown
D. Tom

5): From the passage we know the high postage made _.
A. people never send each other letters
B. lovers almost lose every touch with each other
C. people try their best to avoid paying it
D. receivers refuse to pay the coming letters

Answer: ADABC

RACE Dataset (EMNLP 2017)

- Detail Reasoning
  - To answer the question, the agent should be clear about the details of the passage
  - Answers are not found by simply matching questions with the passage
- Whole-picture Reasoning
  - Agents needs to understand the whole picture of the story to obtain the correct answer
  - Agent is required to comprehend the entire story
- Passage Summarization
  - Question requires the agent to select the best summarization of the passage among four candidate summarizations
  - The main idea of this passage is [MASK]
- Attitude Analysis
  - Asks about the opinions / attitudes of the author or a character
  - LEFT EXAMPLE
- World Knowledge
  - Certain external knowledge is required
  - RIGHT EXAMPLE

- *Evidence*: "...Many people optimistically thought industry awards for better equipment would stimulate the production of quieter appliances. It was even suggested that noise from building sites could be alleviated ..."
- *Question*: What was the author's attitude towards the industry awards for quieter?
- *Options*: A.suspicious B.positive C.enthusiastic D.indifferent

- *Evidence*: "The park is open from 8 am to 5 pm."
- *Question*: The park is open for __ hours a day.
- *Options*: A.eight B.nine C.ten D.eleven

Other Datasets before 2018

Dataset	Query Source	Answer	#Queries	#Context / #Images	Size	Tasks	Metric
COCO-QA (NeurIPS 2015)	Object, Number, Color, Location	One Word	117K+	123K+ Images	2MB (Text Only) 37GB (Image)	VQA	Accuracy
TQA (CVPR 2017)	Middle School Science Curricula	Answer Selection	26K	1076 Lessons	(Not provided in HuggingFace)	VQA	Accuracy
Clevr (CVPR 2017)	Exist, Count, Compare	One Word	1M	100K Images	(Not provided in HuggingFace)	QA	EM / F1
bABI (ICLR 2016)	Crowdsourced	One Word	100K	536 Docs	(Not mentioned in HuggingFace)	QA	EM / F1

- Visual Question Answering is a popular area of research in Computer Vision
- **(Query)** Crowdsourced(Engineered by crowd workers, less creativity) vs User Logs(Extracted from real Web documents, people rarely ask interesting questions)
- **(Answer)** Extractive(Limits QA Models to learn expressiveness) vs Human Generated(Hard to evaluate answer; ROUGE)

Other Datasets before 2018

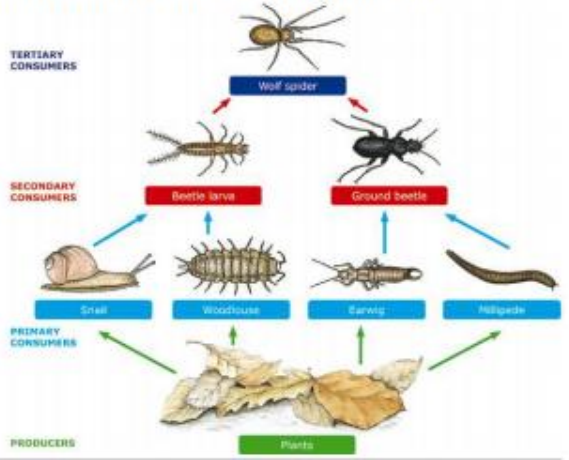
DAQUAR 1553
What is there in front of the sofa?
Ground truth: table
IMG+BOW: table (0.74)
2-VIS+BLSTM: table (0.88)
LSTM: chair (0.47)

COCOQA 5078
How many leftover donuts is the red bicycle holding?
Ground truth: three
IMG+BOW: two (0.51)
2-VIS+BLSTM: three (0.27)
BOW: one (0.29)

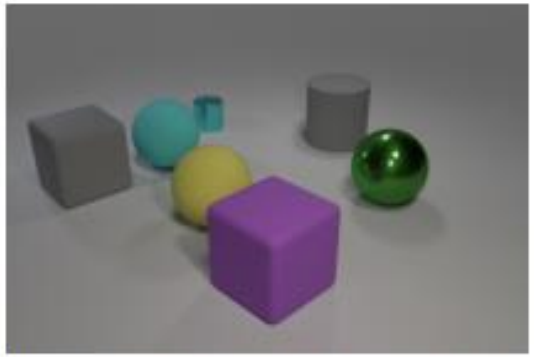
COCO-QA

(f) Hypothetical Question

Q: If the population of beetle larva decreases, what happens with the snail population?
a. Decreases
b. Increases
c. Decreases slightly
d. Stays the same



TQA



Q: How many things are either green things or matte cubes behind the green ball?
A: 2
Q-type: count
Size: 11

Q: There is a cyan object that is to the right of the cyan rubber sphere; is its size the same as the gray rubber cylinder?
A: no
Q-type: equal_size
Size: 16

Q: There is a sphere to the right of the large yellow ball; what material is it?
A: metal
Q-type: query_material
Size: 9

Q: Are there the same number of tiny cylinders that are behind the cyan metal object and purple blocks right of the gray cube?
A: no
Q-type: equal_integer
Size: 17

CLEVR

Other Datasets before 2018

Task 1: Single Supporting Fact
Mary went to the bathroom.
John moved to the hallway.
Mary travelled to the office.
Where is Mary? A:office

Task 2: Two Supporting Facts
John is in the playground.
John picked up the football.
Bob went to the kitchen.
Where is the football? A:playground

Task 3: Three Supporting Facts
John picked up the apple.
John went to the office.
John went to the kitchen.
John dropped the apple.
Where was the apple before the kitchen? A:office

Task 4: Two Argument Relations
The office is north of the bedroom.
The bedroom is north of the bathroom.
The kitchen is west of the garden.
What is north of the bedroom? A: office
What is the bedroom north of? A: bathroom

Task 5: Three Argument Relations
Mary gave the cake to Fred.
Fred gave the cake to Bill.
Jeff was given the milk by Bill.
Who gave the cake to Fred? A: Mary
Who did Fred give the cake to? A: Bill

Task 6: Yes/No Questions
John moved to the playground.
Daniel went to the bathroom.
John went back to the hallway.
Is John in the playground? A:no
Is Daniel in the bathroom? A:yes

Task 7: Counting
Daniel picked up the football.
Daniel dropped the football.
Daniel got the milk.
Daniel took the apple.
How many objects is Daniel holding? A: two

Task 8: Lists/Sets
Daniel picks up the football.
Daniel drops the newspaper.
Daniel picks up the milk.
John took the apple.
What is Daniel holding? milk, football

Task 9: Simple Negation
Sandra travelled to the office.
Fred is no longer in the office.
Is Fred in the office? A:no
Is Sandra in the office? A:yes

Task 10: Indefinite Knowledge
John is either in the classroom or the playground.
Sandra is in the garden.
Is John in the classroom? A:maybe
Is John in the office? A:no

Task 11: Basic Coreference
Daniel was in the kitchen.
Then he went to the studio.
Sandra was in the office.
Where is Daniel? A:studio

Task 12: Conjunction
Mary and Jeff went to the kitchen.
Then Jeff went to the park.
Where is Mary? A: kitchen
Where is Jeff? A: park

Task 13: Compound Coreference
Daniel and Sandra journeyed to the office.
Then they went to the garden.
Sandra and John travelled to the kitchen.
After that they moved to the hallway.
Where is Daniel? A: garden

Task 14: Time Reasoning
In the afternoon Julie went to the park.
Yesterday Julie was at school.
Julie went to the cinema this evening.
Where did Julie go after the park? A:cinema
Where was Julie before the park? A:school

Task 15: Basic Deduction
Sheep are afraid of wolves.
Cats are afraid of dogs.
Mice are afraid of cats.
Gertrude is a sheep.
What is Gertrude afraid of? A:wolves

Task 16: Basic Induction
Lily is a swan.
Lily is white.
Bernhard is green.
Greg is a swan.
What color is Greg? A:white

Task 17: Positional Reasoning
The triangle is to the right of the blue square.
The red square is on top of the blue square.
The red sphere is to the right of the blue square.
Is the red sphere to the right of the blue square? A:yes
Is the red square to the left of the triangle? A:yes

Task 18: Size Reasoning
The football fits in the suitcase.
The suitcase fits in the cupboard.
The box is smaller than the football.
Will the box fit in the suitcase? A:yes
Will the cupboard fit in the box? A:no

Task 19: Path Finding
The kitchen is north of the hallway.
The bathroom is west of the bedroom.
The den is east of the hallway.
The office is south of the bedroom.
How do you go from den to kitchen? A: west, north
How do you go from office to bathroom? A: north, west

Task 20: Agent's Motivations
John is hungry.
John goes to the kitchen.
John grabbed the apple there.
Daniel is hungry.
Where does Daniel go? A:kitchen
Why did John go to the kitchen? A:hungry

bABI

Fixed Template in QA Datasets(1)

LONG / COMPLEX
Question
(Requires Reasoning)

Southern California, often abbreviated SoCal, is a geographic and cultural region that generally comprises California's southernmost 10 counties. The region is traditionally described as "eight counties", based on demographics and economic ties: Imperial, Los Angeles, Orange, Riverside, San Bernardino, San Diego, Santa Barbara, and Ventura. The more extensive 10-county definition, including Kern and San Luis Obispo counties, is also used based on historical political divisions. Southern California is a major economic center for the state of California and the United States.

What is Southern California often abbreviated as?
Ground Truth Answers: SoCal | SoCal | SoCal
Prediction: SoCal

Despite being traditionall described as "eight counties", how many counties does this region actually have?
Ground Truth Answers: 10 counties | 10 | 10
Prediction: <No Answer>

What is a major importance of Southern California in relation to California and the United States?
Ground Truth Answers: economic center | major economic center | economic center
Prediction: economic center

What are the ties that best described what the "eight counties" are based on?
Ground Truth Answers: demographics and economic ties | economic | demographics and economic
Prediction: demographics and economic ties

Answer
(Generative, Selection, Span)

Multiple Context Passage
(If not given, use IR system)

(Requires Multi-hop Reasoning)

Answer might be divided in multiple documents,
in the question, or even not exist anywhere

Datasets in 2018

Dataset	Query Source	Answer	#Queries	#Context	Size	Tasks	Metric
SQuAD 2.0 (ACL Short 2018)	Crowdsourced	Span of Words	100K(1.0) + 50K (Added)	853 Docs	166.91 MB	QA	EM / F1
HotpotQA (EMNLP 2018)	Crowdsourced based on Wikipedia	Span of Words	113K	5M Paragraphs	2400.69 MB	QA - Retrieval	EM / F1
NarrativeQA (ACL 2018)	Crowdsourced	Human Generated	46K+	1572 Stories (Books/Scripts)	(Not mentioned in HuggingFace)	QA - Retrieval	BLEU-1 BLEU-4 METEOR
OpenBookQA (EMNLP 2018)	Elementary-Level Science	Answer Selection	5957	1326 science facts	2.75 MB	QA	Accuracy
Recipe QA (EMNLP 2018)	Cooking recipes from Instructables	Answer Selection	36K	20K Recipes & Images	(Not provided in HuggingFace)	QA	Accuracy

SQuAD 2.0 (ACL 2018 Short)

- Compared to 1.0, should determine when there is no answer by the paragraph (Negative Examples)
- Provides a public leaderboard to submit models
- The Answer is a span within the given context

Article: Endangered Species Act
Paragraph: " . . . *Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised.*"

Question 1: "*Which laws faced significant opposition?*"
Plausible Answer: *later laws*

Question 2: "*What was the name of the 1937 treaty?*"
Plausible Answer: *Bald Eagle Protection Act*

	SQuAD 1.1	SQuAD 2.0
Train		
Total examples	87,599	130,319
Negative examples	0	43,498
Total articles	442	442
Articles with negatives	0	285
Development		
Total examples	10,570	11,873
Negative examples	0	5,945
Total articles	48	35
Articles with negatives	0	35
Test		
Total examples	9,533	8,862
Negative examples	0	4,332
Total articles	46	28
Articles with negatives	0	28

SQuAD 2.0 (ACL 2018 Short)

Reasoning	Description	Example	Percentage
Negation	Negation word inserted or removed.	Sentence: "*Several hospital pharmacies have decided to outsource high risk preparations . . .*" Question: "*What types of pharmacy functions have* ***never*** *been outsourced?*"	9%
Antonym	Antonym used.	S: "*the extinction of the dinosaurs. . .allowed the tropical rainforest to spread out across the continent.*" Q: "*The extinction of what led to the* ***decline*** *of rainforests?*"	20%
Entity Swap	Entity, number, or date replaced with other entity, number, or date.	S: "*These values are much greater than the 9–88 cm as projected . . . in its Third Assessment Report.*" Q: "*What was the projection of sea level increases in the* ***fourth assessment report****?*"	21%
Mutual Exclusion	Word or phrase is mutually exclusive with something for which an answer is present.	S: "*BSkyB. . . waiv[ed] the charge for subscribers whose package included two or more premium channels.*" Q: "*What service did BSkyB* ***give away for free unconditionally****?*"	15%
Impossible Condition	Asks for condition that is not satisfied by anything in the paragraph.	S: "*Union forces left Jacksonville and confronted a Confederate Army at the Battle of Olustee. . . Union forces then retreated to Jacksonville and held the city for the remainder of the war.*" Q: "*After what battle did Union forces leave Jacksonville* ***for good****?*"	4%
Other Neutral	Other cases where the paragraph does not imply any answer.	S: "*Schuenemann et al. concluded in 2011 that the Black Death. . . was caused by a variant of Y. pestis. . .*" Q: "*Who* ***discovered*** *Y. pestis?*"	24%
Answerable	Question is answerable (i.e. dataset noise).		7%

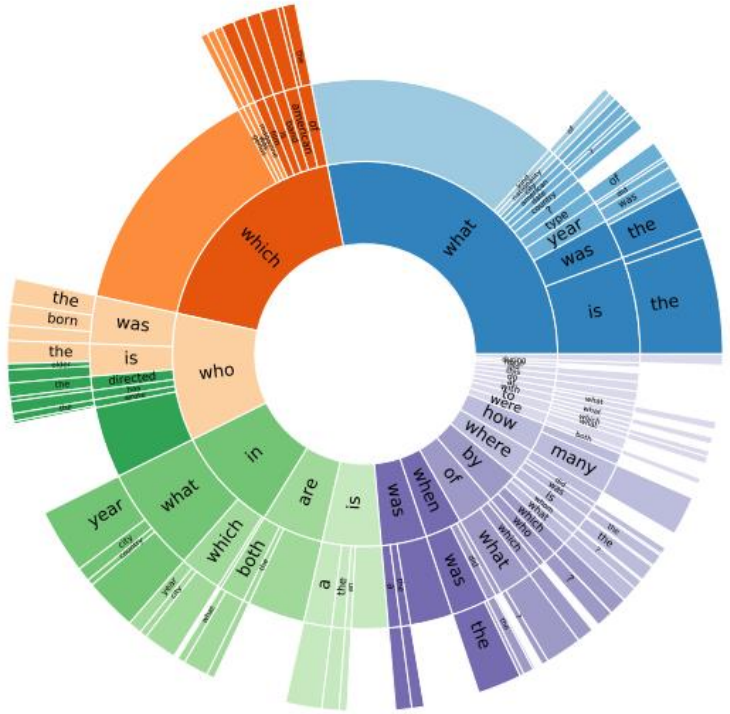
Hotpot QA (EMNLP 2018)

- Each question in the dataset comes with two gold paragraphs
- Questions require finding and reasoning over multiple supporting documents to answer
- Sentence-level supporting facts required for reasoning are provided

Paragraph A, Return to Olympus:
[1] *Return to Olympus is the only album by the alternative rock band Malfunkshun.* [2] *It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990.* [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

Paragraph B, Mother Love Bone:
[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene.* [7] *Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success.* [8] The album was finally released a few months later.

Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?
A: Malfunkshun
Supporting facts: 1, 2, 4, 6, 7



Answer Type	%	Example(s)
Person	30	King Edward II, Rihanna
Group / Org	13	Cartoonito, Apalachee
Location	10	Fort Richardson, California
Date	9	10th or even 13th century
Number	8	79.92 million, 17
Artwork	8	Die schweigsame Frau
Yes/No	6	-
Adjective	4	conservative
Event	1	Prix Benois de la Danse
Other proper noun	6	Cold War, Laban Movement Analysis
Common noun	5	comedy, both men and women

Hotpot QA (EMNLP 2018)

Reasoning Type	%	Example(s)
Inferring the *bridge entity* to complete the 2nd-hop question (Type I)	42	**Paragraph A:** The 2015 Diamond Head Classic was a college basketball tournament ... *Buddy Hield was named the tournament's MVP.* **Paragraph B:** *Chavano Rainier "Buddy" Hield* is a Bahamian professional basketball player for the **Sacramento Kings** of the NBA... **Q:** Which team does the player named 2015 Diamond Head Classic's MVP play for?
Comparing two entities (Comparison)	27	**Paragraph A:** LostAlone were a British rock band ... consisted of *Steven Battelle, Alan Williamson, and Mark Gibson*... **Paragraph B:** Guster is an American alternative rock band ... Founding members *Adam Gardner, Ryan Miller, and Brian Rosenworcel* began... **Q:** Did LostAlone and Guster have the same number of members? **(yes)**
Locating the **answer entity** by checking multiple properties (Type II)	15	**Paragraph A:** Several *current and former members of the Pittsburgh Pirates* ... John Milner, **Dave Parker**, and Rod Scurry... **Paragraph B:** **David Gene Parker**, *nicknamed "The Cobra"*, is an American former player in Major League Baseball... **Q:** Which former member of the Pittsburgh Pirates was nicknamed "The Cobra"?
Inferring about the property of an entity in question through a *bridge entity* (Type III)	6	**Paragraph A:** *Marine Tactical Air Command Squadron 28* is a United States Marine Corps aviation command and control unit based at *Marine Corps Air Station Cherry Point*... **Paragraph B:** *Marine Corps Air Station Cherry Point* ... is a United States Marine Corps airfield located in **Havelock, North Carolina**, USA ... **Q:** What city is the Marine Air Control Group 28 located in?
Other types of reasoning that require more than two supporting facts (Other)	2	**Paragraph A:** ... the towns of Yodobashi, **Okubo, Totsuka, and Ochiai town** *were merged into Yodobashi ward. ... Yodobashi Camera is a store with its name taken from the town and ward.* **Paragraph B:** *Yodobashi Camera* Co., Ltd. is *a major Japanese retail chain specializing in electronics, PCs, cameras and photographic equipment.* **Q:** Aside from Yodobashi, what other towns were merged into the ward which gave the major Japanese retail chain specializing in electronics, PCs, cameras, and photographic equipment it's name?

Narrative QA (ACL 2018)

- Testing understanding of text requires creation of questions that examine high-level abstractions instead of just facts occurring in one sentence at a time
- A summary / full script is given as context to answer the question
- Using the summary is similar to RC, while using IR+Reader is similar to Open Domain QA

Title: Jacob's Ladder
Question: What is the fatal injury that Jacob sustains which ultimately leads to his death ?
Answer: A bayonete stabbing to his gut.
Summary snippet: A terrified Jacob flees into the jungle, only to be bayoneted in the gut by an unseen assailant.
[. . .]
In a wartime triage tent in 1971, military doctors fruitlessly treating Jacob reluctantly declare him dead
Story snippet: As he spins around one of the attackers jams all eight inches of his bayonet blade into Jacob's stomach. Jacob screams. It is a loud and piercing wail.
[. . .]
Int. Vietnam Field Hospital - Day
A doctor leans his head in front of the lamp and removes his mask. His expression is somber. He shakes his head. His words are simple and final.
DOCTOR
He's gone.
Cut to Jacob Singer . . .
The doctor steps away. A nurse rudely pulls a green sheet up over his head. The doctor turns to one of the aides and throws up his hands in defeat.

Title: Armageddon 2419 A.D.
Question: In what year did Rogers awaken from his deep slumber?
Answer: 2419
Summary snippet: . . . Rogers remained in sleep for 492 years. He awakes in 2419 and,. . .
Story snippet: I should state therefore, that I, Anthony Rogers, am, so far as I know, the only man alive whose normal span of eighty-one years of life has been spread over a period of 573 years. To be precise, I lived the first twenty-nine years of my life between 1898 and 1927; the other fifty-two since 2419. The gap between these two, a period of nearly five hundred years, I spent in a state of suspended animation, free from the ravages of katabolic processes, and without any apparent effect on my physical or mental faculties. When I began my long sleep, man had just begun his real conquest of the air. . .

First token	Frequency
What	38.04%
Who	23.37%
Why	9.78%
How	8.85%
Where	7.53%
Which	2.21%
How many/much	1.80%
When	1.67%
In	1.19%
OTHER	5.57%

Category	Frequency
Person	30.54%
Description	24.50%
Location	9.73%
Why/reason	9.40%
How/method	8.05%
Event	4.36%
Entity	4.03%
Object	3.36%
Numeric	3.02%
Duration	1.68%
Relation	1.34%

Other Datasets in 2018

Text Cloze Style Question	Context Modalities: Images and Descriptions of Steps

Recipe: Last-Minute Lasagna

1. Heat oven to 375 degrees F. Spoon a thin layer of sauce over the bottom of a 9-by-13-inch baking dish.
2. Cover with a single layer of ravioli.
3. Top with half the spinach half the mozzarella and a third of the remaining sauce.
4. Repeat with another layer of ravioli and the remaining spinach mozzarella and half the remaining sauce.
5. Top with another layer of ravioli and the remaining sauce not all the ravioli may be needed. Sprinkle with the Parmesan.
6. Cover with foil and bake for 30 minutes. Uncover and bake until bubbly, 5 to 10 minutes.
7. Let cool 5 minutes before spooning onto individual plates.


Step 1 Step 2 Step 3 Step 4


Step 5 Step 6 Step 7

Question	Choose the best text for the missing blank to correctly complete the recipe Cover. _____. Bake. Cool, serve.
Answer	**A. Top, sprinkle** B. Finishing touches C. Layer it up D. Ravioli bonus round

RECIPE QA

Question:
Which of these would let the most heat travel through?
A) a new pair of jeans.
B) a steel spoon in a cafeteria.
C) a cotton candy at a store.
D) a calvin klein cotton hat.

Science Fact:
Metal is a thermal conductor.

Common Knowledge:
Steel is made of metal.
Heat travels through a thermal conductor.

Open Book QA

Fixed Template in QA Datasets(2)

LONG / COMPLEX / UNANSWERABLE Question (Requires Reasoning & Memory)

Southern California, often abbreviated SoCal, is a geographic and cultural region that generally comprises California's southernmost 10 counties. The region is traditionally described as "eight counties", based on demographics and economic ties: Imperial, Los Angeles, Orange, Riverside, San Bernardino, San Diego, Santa Barbara, and Ventura. The more extensive 10-county definition, including Kern and San Luis Obispo counties, is also used based on historical political divisions. Southern California is a major economic center for the state of California and the United States.

What is Southern California often abbreviated as?
Ground Truth Answers: SoCal | SoCal | SoCal
Prediction: SoCal

Despite being traditionall described as "eight counties", how many counties does this region actually have?
Ground Truth Answers: 10 counties | 10 | 10
Prediction: <No Answer>

What is a major importance of Southern California in relation to California and the United States?
Ground Truth Answers: economic center | major economic center | economic center
Prediction: economic center

What are the ties that best described what the "eight counties" are based on?
Ground Truth Answers: demographics and economic ties | economic | demographics and economic
Prediction: demographics and economic ties

Answer (Generative, Selection, Span)

Multiple Context Passage (If not given, use IR system)

(Requires Multi-hop Reasoning)

Answer might be divided in multiple documents, in the question, or even not exist anywhere

Datasets in 2019 (1)

Dataset	Query Source	Answer	#Queries	#Context	Size	Tasks	Metric
COSMOS QA (EMNLP 2019)	Crowdsourced	Answer Selection	35600	21866	46.64 MB	QA	Accuracy
DROP (NAACL 2019)	Crowdsourced (Adversarially-Created)	One Word = Arithmetic	100K	7000 Passages	113.69 MB	Symbolic based QA	EM / F1
ELI5 (ACL 2019)	From Reddit Forums	Human Generated	100K	270K Threads	(Not mentioned in HuggingFace)	Abstractive QA	ROUGE-1 ROUGE-2 ROUGE-K

COSMOS QA (EMNLP-IJCNLP 2019)

- Most existing reading comprehension datasets have questions that focus on factual and literal understanding of context
- Asking questions that require reasoning is essential for everyday narratives
- E.g. Why, What may happen, What will happen types of questions

P1: It's a very humbling experience when you need someone to dress you every morning, tie your shoes, and put your hair up. Every menial task takes an unprecedented amount of effort. It made me appreciate Dan even more. But anyway I shan't dwell on this (I'm not dying after all) and not let it detract from my lovely 5 days with my friends visiting from Jersey.

Q: *What's a possible reason the writer needed someone to dress him every morning?*

A: The writer doesn't like putting effort into these tasks.

✓ **B: The writer has a physical disability.**

C: The writer is bad at doing his own hair.

D: None of the above choices.

P2: A woman had topped herself by jumping off the roof of the hospital she had just recently been admitted to. She was there because the first or perhaps latest suicide attempt was unsuccessful. She put her clothes on, folded the hospital gown and made the bed. She walked through the unit unimpeded and took the elevator to the top floor.

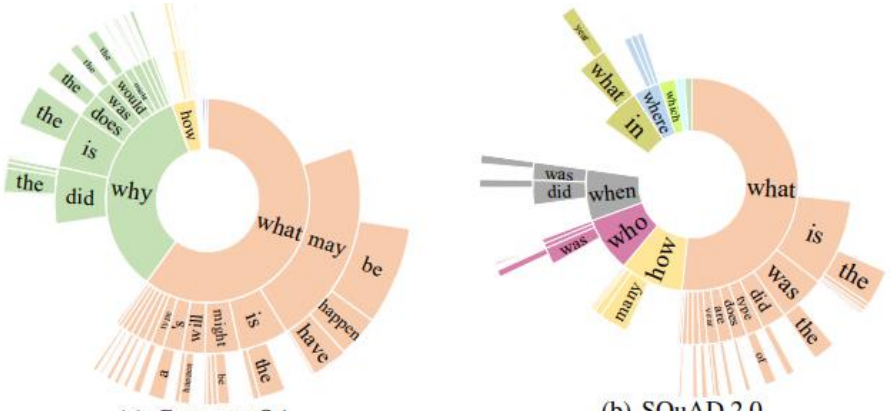
Q: *What would have happened to the woman if the staff at the hospital were doing their job properly?*

✓ **A: The woman would have been stopped before she left to take the elevator to the top floor and she would have lived.**

B: She would have been ushered to the elevator with some company.

C: She would have managed to get to the elevator quicker with some assistance.

D: None of the above choices.



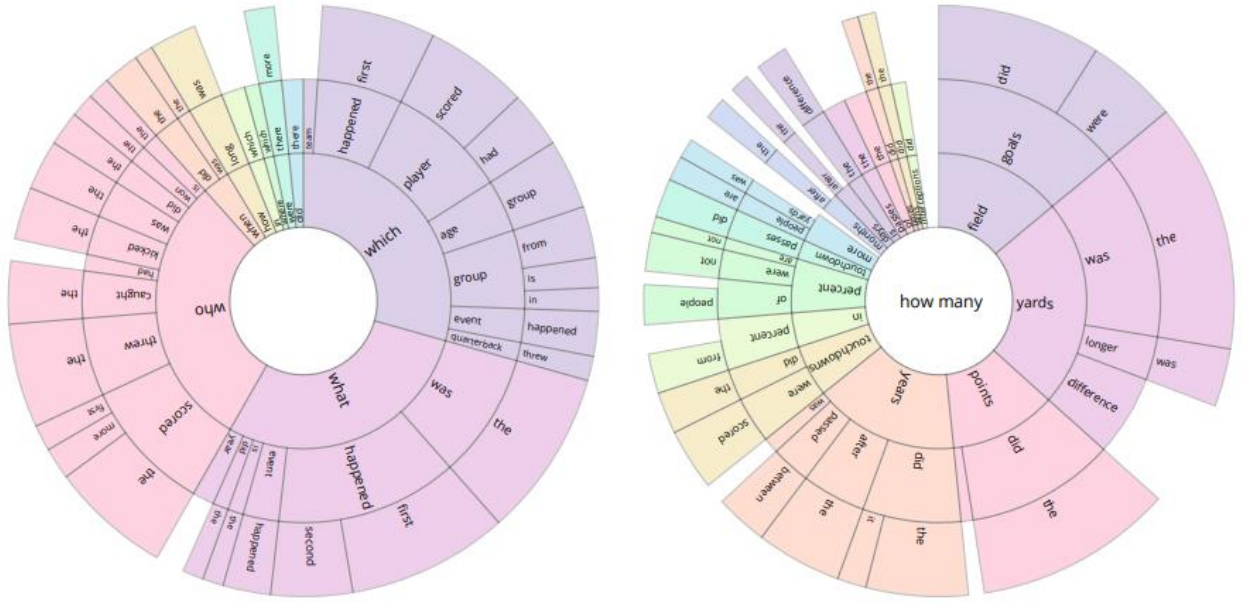
(a) COSMOS QA (b) SQuAD 2.0

Type	Percentage (%)
MRC w/o commonsense	6.2
MRC w/ commonsense	93.8
Pre-/Post- Condition	27.2
Motivation	16.0
Reaction	13.2
Temporal Events	12.4
Situational Fact	23.8
Counterfactual	4.4
Other	12.6

DROP (NAACL-HLT 2019)

- DROP is designed to encourage research on methods that combine NN with discrete, symbolic reasoning

Reasoning	Passage (some parts shortened)	Question	Answer	BiDAF
Subtraction (28.8%)	That year, his **Untitled (1981)**, a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was **sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate**.	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000	$16.3 million
Comparison (18.2%)	In **1517, the seventeen-year-old King sailed to Castile**. There, his Flemish court .... **In May 1518, Charles traveled to Barcelona in Aragon**.	Where did Charles travel to first, Castile or Barcelona?	Castile	Aragon
Selection (19.4%)	In 1970, to commemorate the 100th anniversary of the founding of Baldwin City, **Baker University professor and playwright Don Mueller and Phyllis E. Braun, Business Manager, produced a musical play entitled The Ballad Of Black Jack** to tell the story of the events that led up to the battle.	Who was the University professor that helped produce The Ballad Of Black Jack, Ivan Boyd or Don Mueller?	Don Mueller	Baker
Addition (11.7%)	Before the UNPROFOR fully deployed, the HV clashed with an armed force of the RSK in the village of Nos Kalik, located in a pink zone near Šibenik, and captured the village at 4:45 p.m. on **2 March 1992**. The JNA formed a battlegroup to counterattack the **next day**.	What date did the JNA form a battlegroup to counterattack after the village of Nos Kalik was captured?	3 March 1992	2 March 1992
Count (16.5%) and Sort (11.7%)	Denver would retake the lead with kicker **Matt Prater nailing a 43-yard field goal**, yet Carolina answered as kicker **John Kasay ties the game with a 39-yard field goal**. ... Carolina closed out the half with **Kasay nailing a 44-yard field goal**. ... In the fourth quarter, Carolina sealed the win with **Kasay's 42-yard field goal**.	Which kicker kicked the most field goals?	John Kasay	Matt Prater
Coreference Resolution (3.7%)	**James Douglas** was the second son of Sir George Douglas of Pittendreich, and Elizabeth Douglas, daughter David Douglas of Pittendreich. Before **1543 he married Elizabeth**, daughter of James Douglas, 3rd Earl of Morton. **In 1553 James Douglas succeeded to the title and estates of his father-in-law**.	How many years after he married Elizabeth did James Douglas succeed to the title and estates of his father-in-law?	10	1553
Other Arithmetic (3.2%)	Although the movement initially gathered some **60,000 adherents**, the subsequent establishment of the Bulgarian Exarchate **reduced their number by some 75%**.	How many adherents were left after the establishment of the Bulgarian Exarchate?	15000	60,000
Set of spans (6.0%)	According to some sources 363 civilians were killed in **Kavadarci**, 230 in **Negotino** and 40 in **Vatasha**.	What were the 3 villages that people were killed in?	Kavadarci, Negotino, Vatasha	Negotino and 40 in Vatasha
Other (6.8%)	This **Annual Financial Report** is our principal financial statement of accountability. The **AFR gives a comprehensive view** of the Department's financial activities ...	What does AFR stand for?	Annual Financial Report	one of the Big Four audit firms



(a) For span type answers

(b) For number type answers

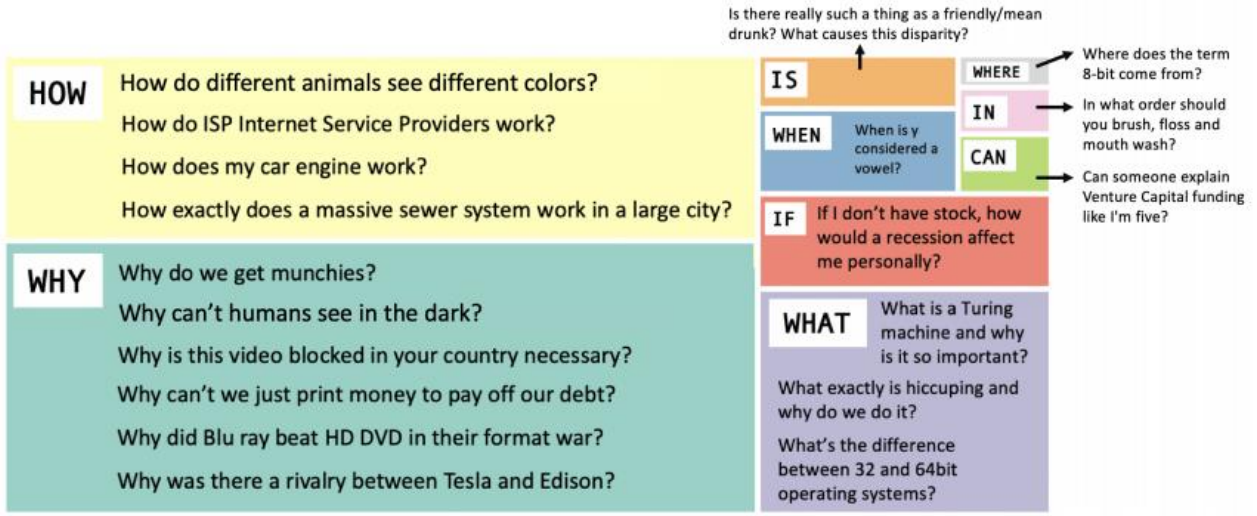
ELI5 (ACL 2019)

- ELI5 comprises diverse questions requiring multi-sentence answers
- Abstractive model training is required
- Can be treated as a form of query-based multi-document summarization (text generation)

Question: Why do TV shows hide logos ?

Document: Why is that? Why do logos get blurred on television, or in film, or even in music videos? The answer, it turns out, is complicated, but is mostly about money in various forms. A whole lot of dysfunction here. Quick Pick: HBO TV Shows Pick the missing word in the title of these HBO shows. September is season premiere month so these logos should be fresh in your mind. TV Shows Venn Diagram II Can you click on the most accurate section of the Venn Diagram for each of the following TV Shows? [...]

ELI5 Answer: nothing is free. In most cases, it is a prop for the show, but because apple did NOT pay them for the product placement, the show isn't going to give it away. In other cases, apple may not want their brand used in association with that media.



Datasets in 2019 (2)

Dataset	Query Source	Answer	#Queries	#Context	Size	Tasks	Metric
COQA (ACL 2019)	Crowdsourced	Free-form text	127K	8K Conversations	73.75MB	Conversational QA	F1
PubMedQA (EMNLP 2019)	PubMed (Medical Domain)	Yes/No (Confirmative)	1K(expert) 61.2K(unlabel) 211.3K (generated)	1K(expert) 61.2K(unlabel) 211.3K (generated)	(Not mentioned in HuggingFace)	QA	Accuracy / F1
SuperGLUE (NeurIPS 2019) [**MultiRC / ReCORD**]	Crowdsourced	True/False (Confirmative) = Cloze Style	10K+ = 120K+	1K+ paragraphs = 80K+ Passages	293.67 MB	QA = Commonsense Reasoning	EM / F1
OK-VQA (ACL 2019)	Crowdsourced	One Word	14K	14K Images	(Not provided in HuggingFace)	VQA	Accuracy
CLUTRR (EMNLP 2019)	Graph Generation Based	Cloze Style	Data Generation	Data Generation	-	Logic Programming	Accuracy

Other Datasets in 2019

Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well. Jessica had . . .

Q_1: Who had a birthday?
A_1: Jessica
R_1: Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80.

Q_2: How old would she be?
A_2: 80
R_2: she was turning 80

Q_3: Did she plan to have any visitors?
A_3: Yes
R_3: Her granddaughter Annie was coming over

Q_4: How many?
A_4: Three
R_4: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

Q_5: Who?
A_5: Annie, Melanie and Josh
R_5: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.

COQA

Original Statement Title	Converted Question	Label	%
Spontaneous electrocardiogram alterations ***predict*** ventricular fibrillation in Brugada syndrome.	***Do*** spontaneous electrocardiogram alterations ***predict*** ventricular fibrillation in Brugada syndrome?	***yes***	92.8
Liver grafts from selected older donors ***do not have*** significantly more ischaemia reperfusion injury.	***Do*** liver grafts from selected older donors ***have*** significantly more ischaemia reperfusion injury?	***no***	7.2

PubMed QA

MultiRC

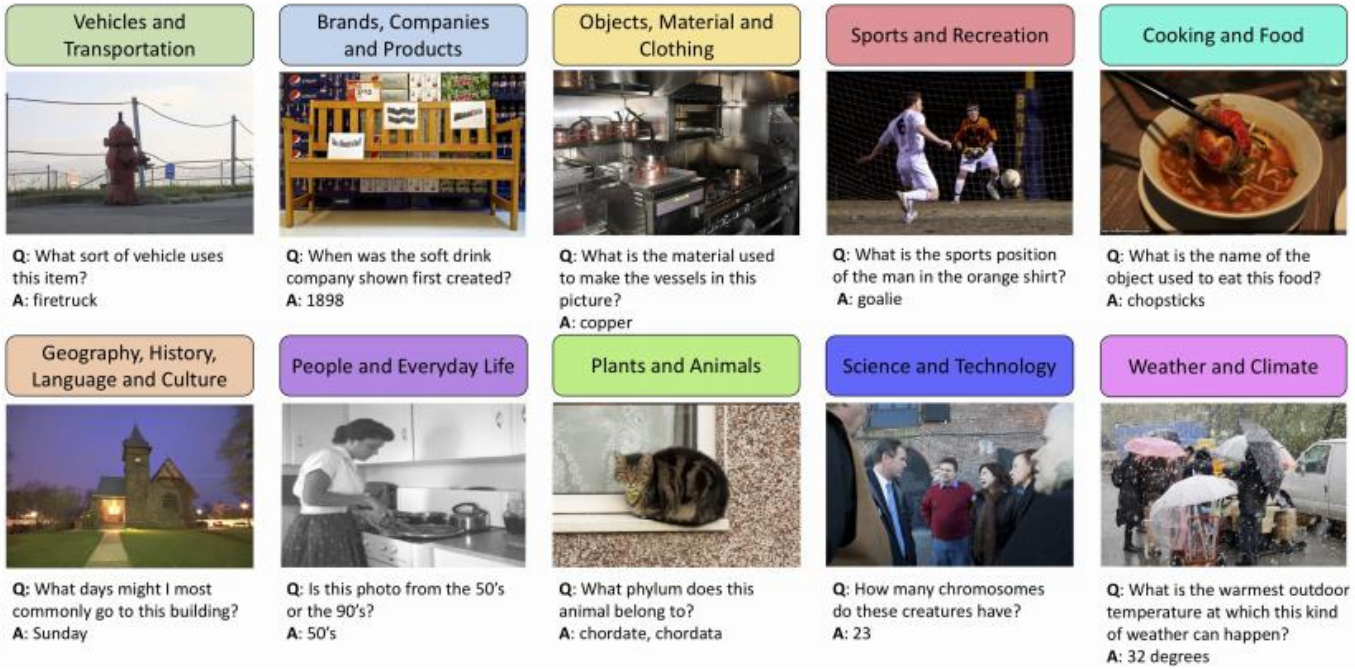
Paragraph: *Susan wanted to have a birthday party. She called all of her friends. She has five friends. Her mom said that Susan can invite them all to the party. Her first friend could not go to the party because she was sick. Her second friend was going out of town. Her third friend was not so sure if her parents would let her. The fourth friend said maybe. The fifth friend could go to the party for sure. Susan was a little sad. On the day of the party, all five friends showed up. Each friend had a present for Susan. Susan was happy and sent each friend a thank you card the next week*
Question: *Did Susan's sick friend recover?* **Candidate answers:** *Yes, she recovered* (T), *No* (F), *Yes* (T), *No, she didn't recover* (F), *Yes, she was at Susan's party* (T)

ReCoRD

Paragraph: *(CNN) Puerto Rico on Sunday overwhelmingly voted for statehood. But Congress, the only body that can approve new states, will ultimately decide whether the status of the US commonwealth changes. Ninety-seven percent of the votes in the nonbinding referendum favored statehood, an increase over the results of a 2012 referendum, official results from the State Electorcal Commission show. It was the fifth such vote on statehood. "Today, we the people of Puerto Rico are sending a strong and clear message to the US Congress ... and to the world ... claiming our equal rights as American citizens, Puerto Rico Gov. Ricardo Rossello said in a news release. @highlight Puerto Rico voted Sunday in favor of US statehood*
Query For one, they can truthfully say, "Don't blame me, I didn't vote for them, " when discussing the <placeholder> presidency **Correct Entities:** US

SuperGLUE (MultiRC / ReCoRD)

Other Datasets in 2019



OK-VQA

Puzzle	Question	Gender	Answer
Roger was playing baseball with his sons *Sam* and *Leon*. *Sam* had to take a break though because he needed to call his sister *Robin*.	Leon is the _____ of Robin	Robin:female, Sam:male, Roger:male, Leon:male	brother
Elvira and her daughter *Nancy* went shopping together last Monday and they bought new shoes for *Elvira*'s kids. *Pedro* and his sister *Allison* went to the fair. *Pedro*'s mother, *Nancy*, was out with friends for the day.	Elvira is the _____ of Allison	Allison:female, Pedro:male, Nancy:female, Elvira:female	grandmother
Roger met up with his sister *Nancy* and her daughter *Cynthia* at the mall to go shopping together. *Cynthia*'s brother *Pedro* was going to be the star in the new show.	Pedro is the _____ of Roger	Roger:male, Nancy:female, Cynthia:female, Pedro:male	nephew

CLUTRR

Fixed Template in QA Datasets(3)

LONG / COMPLEX / UNANSWERABLE Question (Requires Reasoning & Memory)

Southern California, often abbreviated SoCal, is a geographic and cultural region that generally comprises California's southernmost 10 counties. The region is traditionally described as "eight counties", based on demographics and economic ties: Imperial, Los Angeles, Orange, Riverside, San Bernardino, San Diego, Santa Barbara, and Ventura. The more extensive 10-county definition, including Kern and San Luis Obispo counties, is also used based on historical political divisions. Southern California is a major economic center for the state of California and the United States.

What is Southern California often abbreviated as?
Ground Truth Answers: SoCal | SoCal | SoCal
Prediction: SoCal

Despite being traditionall described as "eight counties", how many counties does this region actually have?
Ground Truth Answers: 10 counties | 10 | 10
Prediction: <No Answer>

Answer (Generative, Selection, Span)

What is a major importance of Southern California in relation to California and the United States?
Ground Truth Answers: economic center | major economic center | economic center
Prediction: economic center

What are the ties that best described what the "eight counties" are based on?
Ground Truth Answers: demographics and economic ties | economic | demographics and economic
Prediction: demographics and economic ties

Multiple Context Passage (If not given, use IR system)

(Requires Multi-hop Reasoning)

Answer might be divided in multiple documents, in the question, or even not exist anywhere

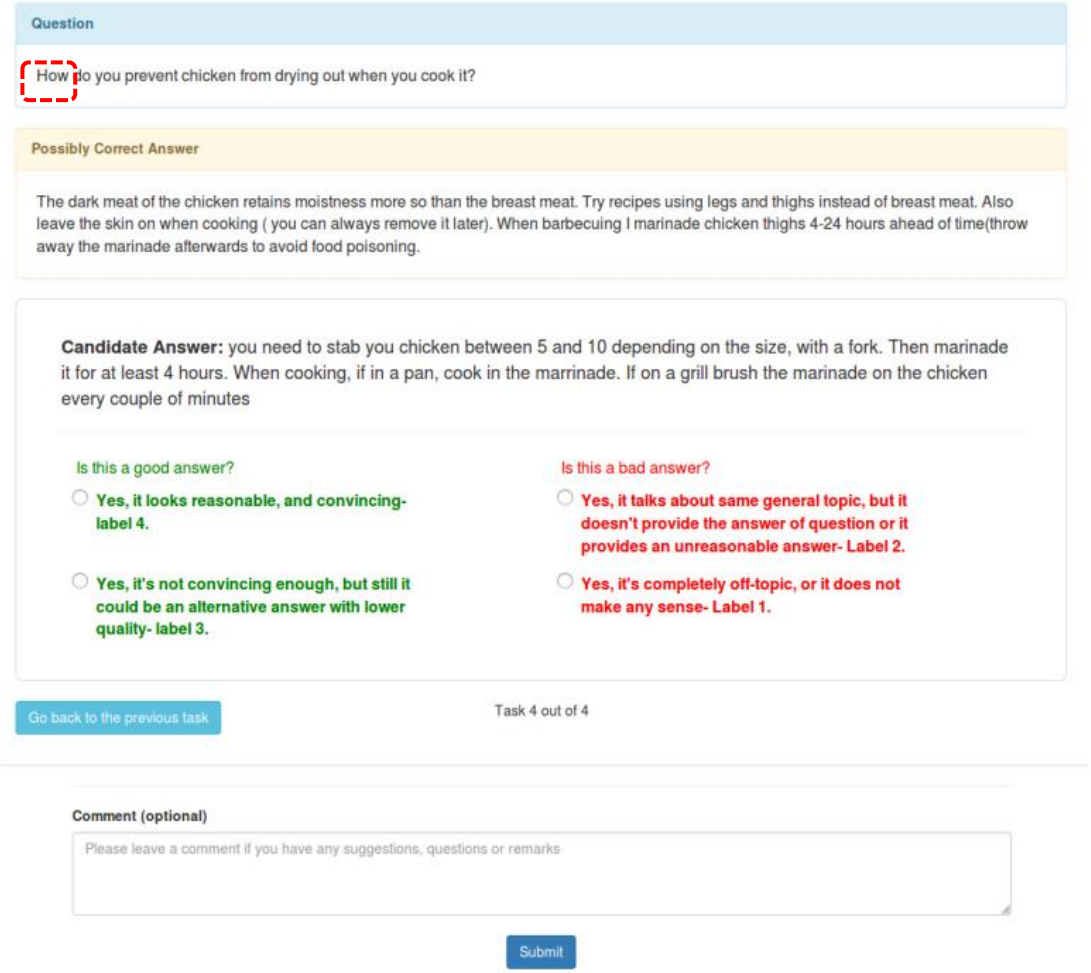
Answer requires Logical Reasoning (e.g. Arithmetic)

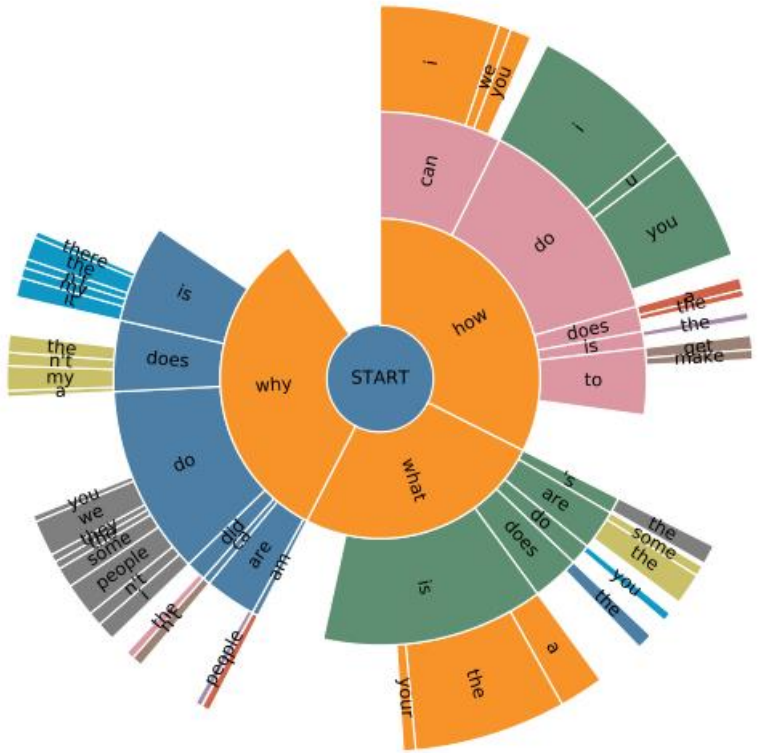
Datasets in 2020

Dataset	Query Source	Answer	#Queries	#Context	Size	Tasks	Metric
ANTIQUE (ECIR 2020)	Community QA (Yahoo)	Sentence Selection	2626	34,011 annotations	(Not supported by HuggingFace)	Answer Quality Ranking	MAP / MRR
RECLOR (ICLR 2020)	Law School Admission Council (GMAT, LSAT)	Answer Selection	6138	6138	(Not mentioned in HuggingFace)	QA	Accuracy
WINOGRANDE (AAAI 2020)	CrowdSourced	Cloze-Style & Answer Selection between two	77K	-	4.87 MB	QA	Accuracy
LogiQA (IJCAI 2020)	Civil Servants Examination of China	Answer Selection	8678	8687	(Not supported by HuggingFace)	Logic Programming	Accuracy
Machine Numer Sense (AAAI 2020)	Graph Generation based	Cloze-style	Data Generation	Data Generation	(Not supported by HuggingFace)	Logic Programming	Accuracy
COVID-QA (ACL 2020 Workshop)	Kaggle	Answer Selection	124	124	190.90 MB	QA	Accuracy

ANTIQUE (ECIR 2020)

- Using community QA Systems sample questions & answers, and using Crowdsourcing workers rank the candidate answers





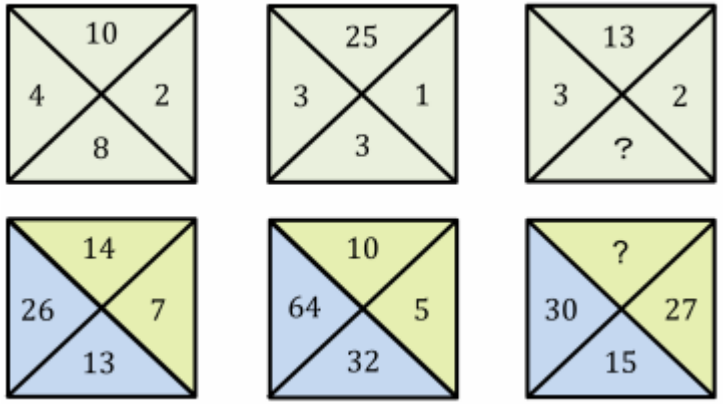
RECLOR (ICLR 2020)

- In Natural Language Understanding, Logical Reasoning is an important ability to *examine, analyze, and critically evaluate* arguments as they occur in ordinary language *[Law School Admission Council (2019a)]*
- In RECLOR Dataset, to answer the question, readers need to identify the logical connections between the lines to pinpoint the conflict, and then understand each of the options and select an option that solve the conflict.

Context:
In jurisdictions where use of headlights is optional when visibility is good, drivers who use headlights at all times are less likely to be involved in a collision than are drivers who use headlights only when visibility is poor. Yet Highway Safety Department records show that making use of headlights mandatory at all times does nothing to reduce the overall number of collisions.
Question: Which one of the following, if true, most helps to resolve the apparent discrepancy in the information above?
Options:
A. In jurisdictions where use of headlights is optional when visibility is good, one driver in four uses headlights for daytime driving in good weather.
B. Only very careful drivers use headlights when their use is not legally required.
C. The jurisdictions where use of headlights is mandatory at all times are those where daytime visibility is frequently poor.
D. A law making use of headlights mandatory at all times is not especially difficult to enforce.
Answer: B

Type	Description
Necessary Assumptions (11.4%)	identify the claim that must be true or is required in order for the argument to work.
Sufficient Assumptions (3.0%)	identify a sufficient assumption, that is, an assumption that, if added to the argument, would make it logically valid.
Strengthen (9.4%)	identify information that would strengthen an argument
Weaken (11.3%)	identify information that would weaken an argument
Evaluation (1.3%)	identify information that would be useful to know to evaluate an argument
Implication (4.6%)	identify something that follows logically from a set of premises
Conclusion/Main Point (3.6%)	identify the conclusion/main point of a line of reasoning
Most Strongly Supported (5.6%)	find the choice that is most strongly supported by a stimulus
Explain or Resolve (8.4%)	identify information that would explain or resolve a situation
Principle (6.5%)	identify the principle, or find a situation that conforms to a principle, or match the principles
Dispute (3.0%)	identify or infer an issue in dispute
Technique (3.6%)	identify the technique used in the reasoning of an argument
Role (3.2%)	describe the individual role that a statement is playing in a larger argument
Identify a Flaw (11.7%)	identify a flaw in an arguments reasoning
Match Flaws (3.1%)	find a choice containing an argument that exhibits the same flaws as the passages argument
Match the Structure (3.0%)	match the structure of an argument in a choice to the structure of the argument in the passage
Others (7.3%)	other types of questions which are not included by the above

Other Datasets in 2020



Machine Number Sense

category: Asymptomatic shedding
subcategory: Proportion of patients who were asymptomatic
query: proportion of patients who were asymptomatic
question: What proportion of patients are asymptomatic?

Answers

id: `56zhxd6e`
title: Epidemiological parameters of coronavirus disease 2019: a pooled analysis of publicly reported individual data of 1155 cases from seven countries
answer: 49 (14.89%) were asymptomatic

id: `rjm1dqk7`
title: Epidemiological characteristics of 2019 novel coronavirus family clustering in Zhejiang Province
answer: 54 asymptomatic infected cases

Covid QA

		Twin sentences	Options (**answer**)
✓ (1)	a	The trophy doesn't fit into the brown suitcase because **it**'s too *large*.	**trophy** / suitcase
	b	The trophy doesn't fit into the brown suitcase because **it**'s too *small*.	trophy / **suitcase**
✓ (2)	a	Ann asked Mary what time the library closes, *because* **she** had forgotten.	**Ann** / Mary
	b	Ann asked Mary what time the library closes, *but* **she** had forgotten.	Ann / **Mary**
✗ (3)	a	The tree fell down and crashed through the roof of my house. Now, I have to get **it** *removed*.	**tree** / roof
	b	The tree fell down and crashed through the roof of my house. Now, I have to get **it** *repaired*.	tree / **roof**
✗ (4)	a	The lions ate the zebras because **they** are *predators*.	**lions** / zebras
	b	The lions ate the zebras because **they** are *meaty*.	lions / **zebras**

WINOGRANDE

P1: David, Jack and Mark are colleagues in a company. David supervises Jack, and Jack supervises Mark. David gets more salary than Jack.

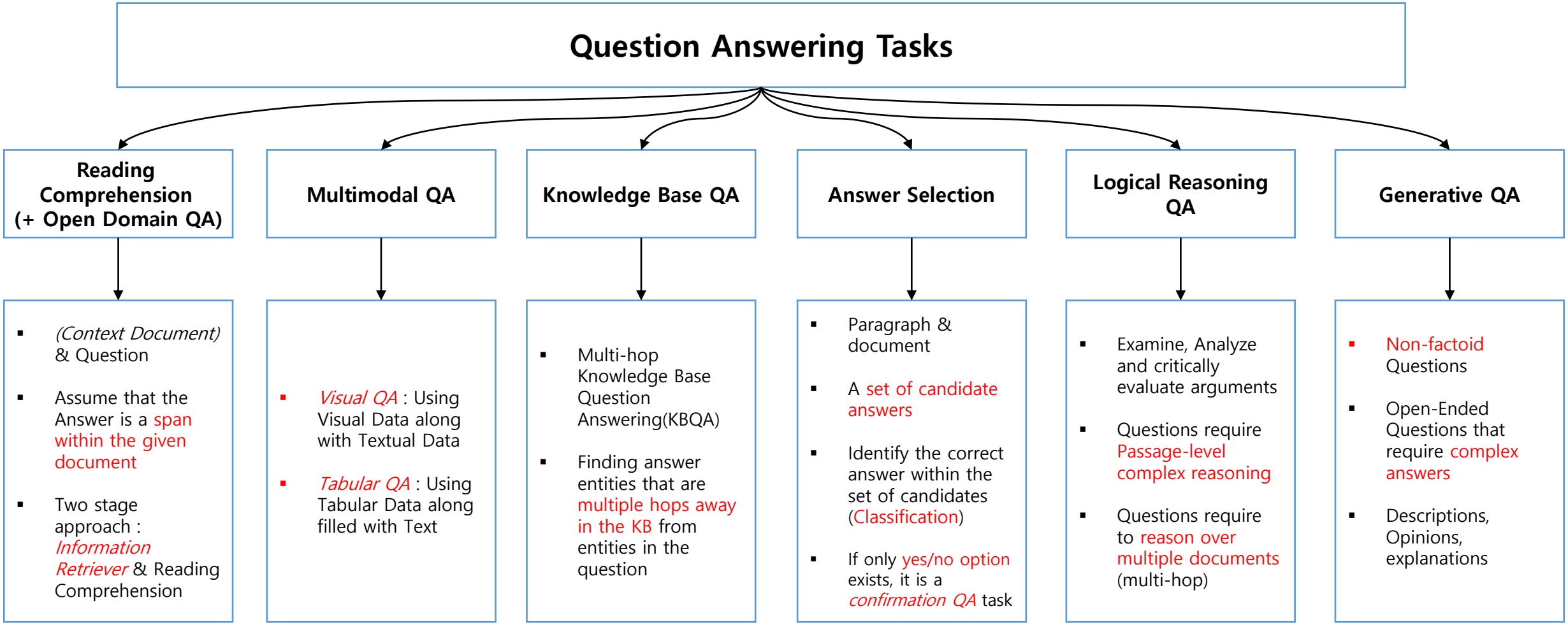
Q: *What can be inferred from the above statements?*
A. Jack gets more salary than Mark.
B. David gets the same salary as Mark.
C. One employee supervises another who gets more salary than himself.
✓ **D. One employee supervises another who gets less salary than himself.**

P2: Our factory has multiple dormitory areas and workshops. None of the employees who live in dormitory area A are textile workers. We conclude that some employees working in workshop B do not live in dormitory area A.

Q: *What may be the missing premise of the above argument?*
A. Some textile workers do not work in workshop B.
B. Some employees working in workshop B are not textile workers.
✓ **C. Some textile workers work in workshop B.**
D. Some employees living in dormitory area A work in the workshop B.

LogiQA

Question Answering Tasks Taxonomy

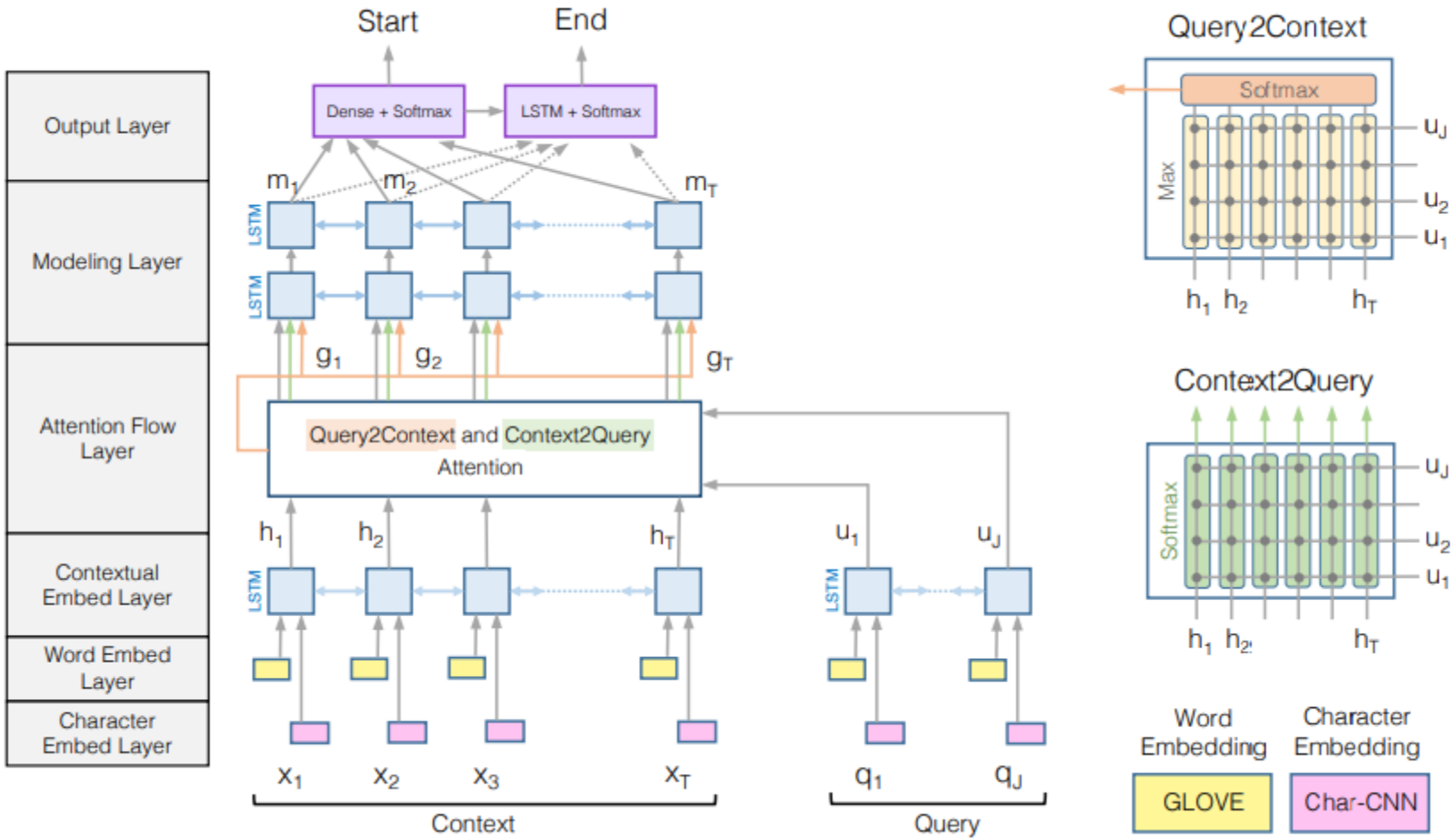


QA Baselines

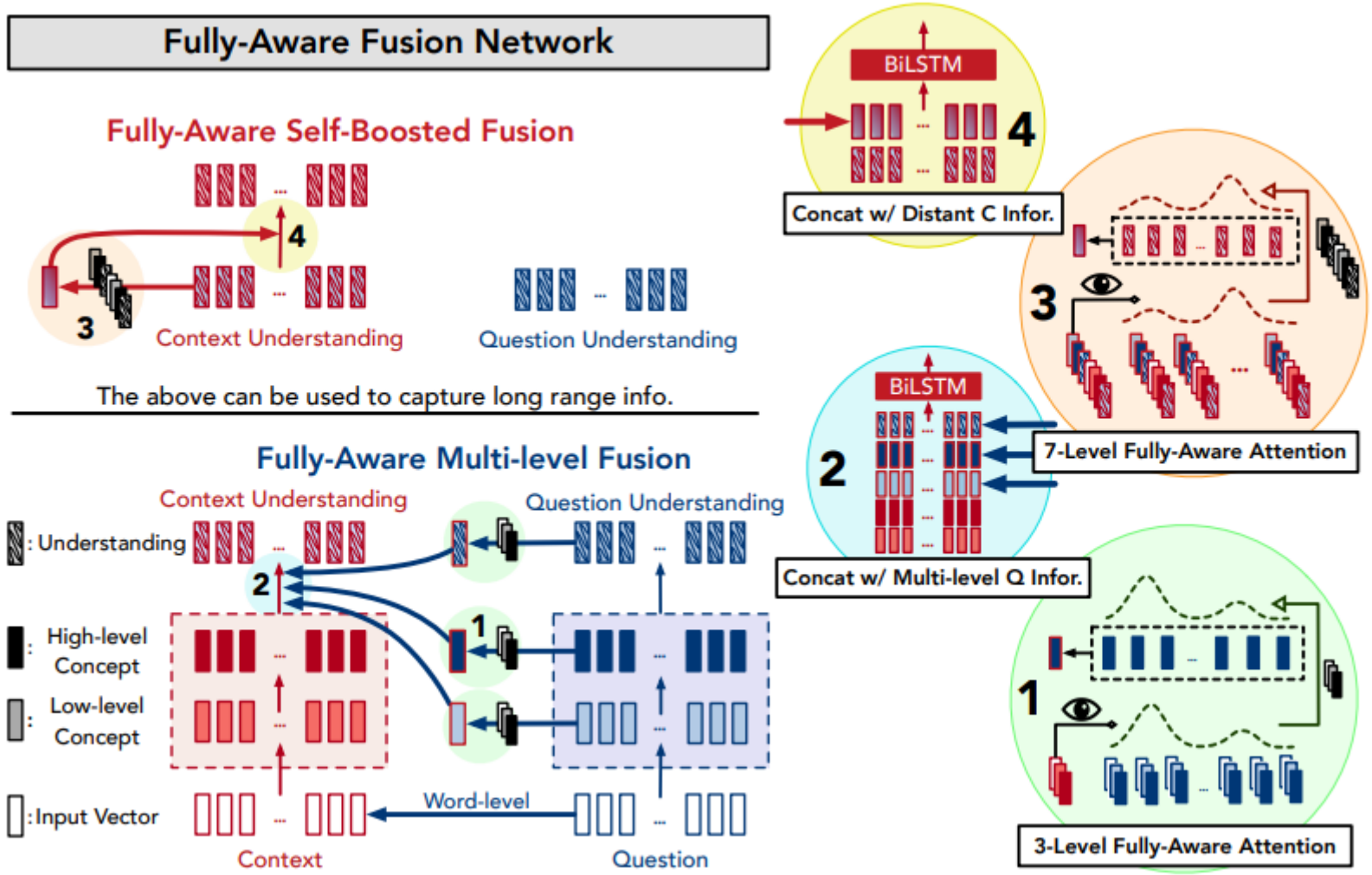
Baselines

- Bi-DAF
  - *Seo, Minjoon, et al. "Bidirectional attention flow for machine comprehension." (ICLR 2017)*
- FusionNet
  - *Hsin-Yuan Hunag et al. FusionNet : Fusing via Fully-aware Attention with Application to Machine Comprehension (ICLR 2018)*
- BERT
  - *Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." (HLT-NAACL 2019)*
- LUKE
  - *Yamada, Ikuya, et al. "LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention.“ (EMNLP 2020)*
- SGNet
  - *Zhang, Zhuosheng, et al. "SG-Net: Syntax-guided machine reading comprehension." (AAAI 2020)*
- Retro-Reader
  - *Zhang, Zhuosheng, Junjie Yang, and Hai Zhao. "Retrospective reader for machine reading comprehension." (AAAI 2021)*

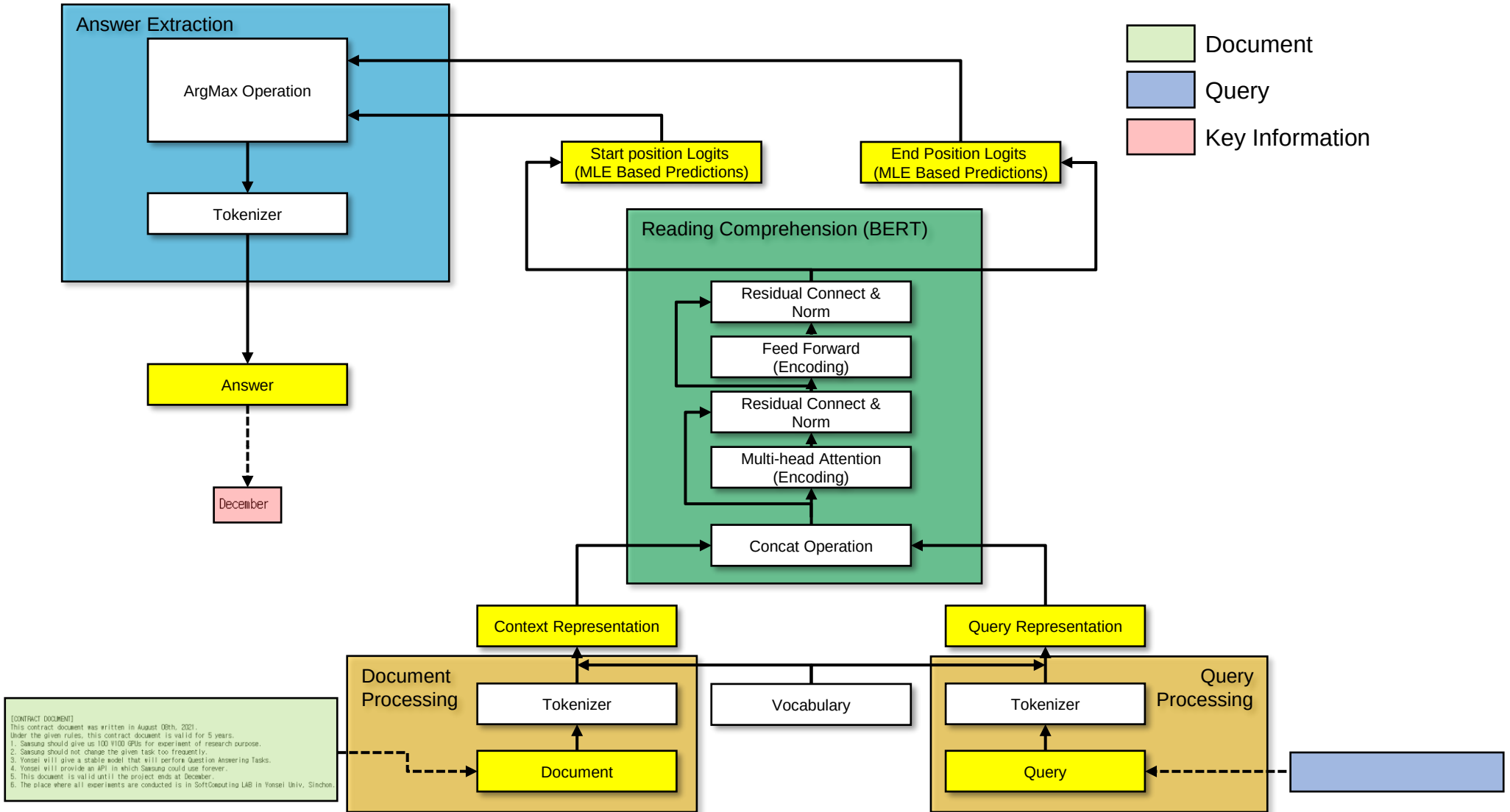
BiDAF (ICLR 2017)



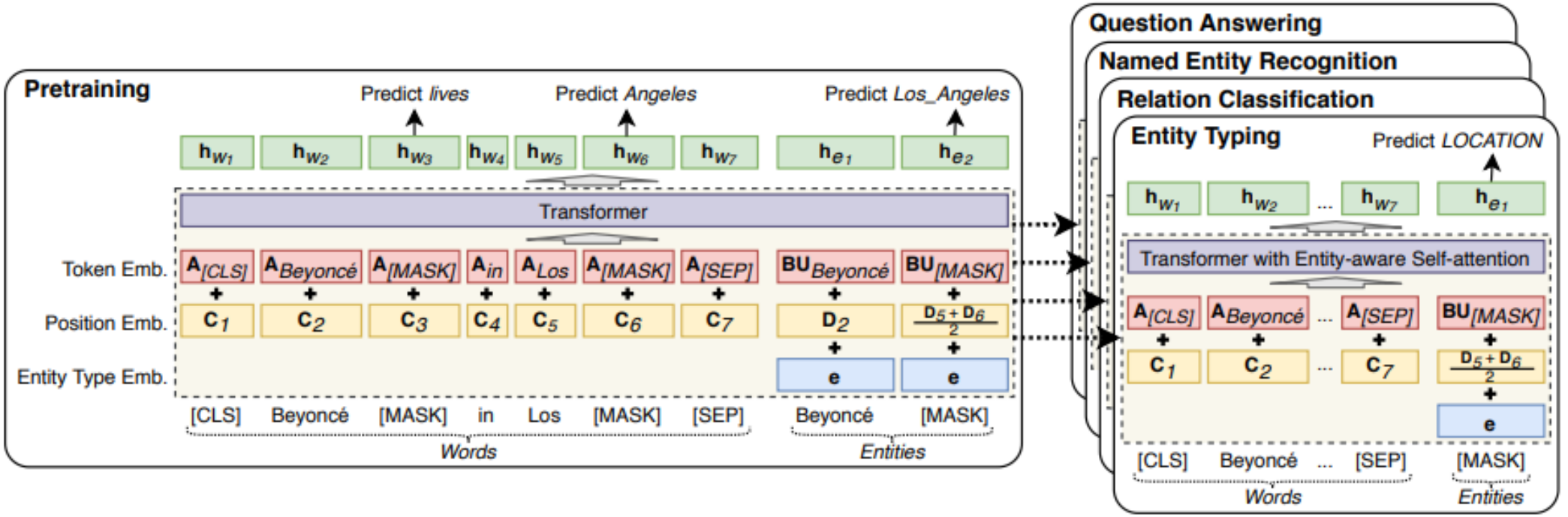
FusionNet (ICLR 2018)



BERT (NAACL-HLT 2019)



LUKE (EMNLP 2020)



SGNet (AAAI 2020)

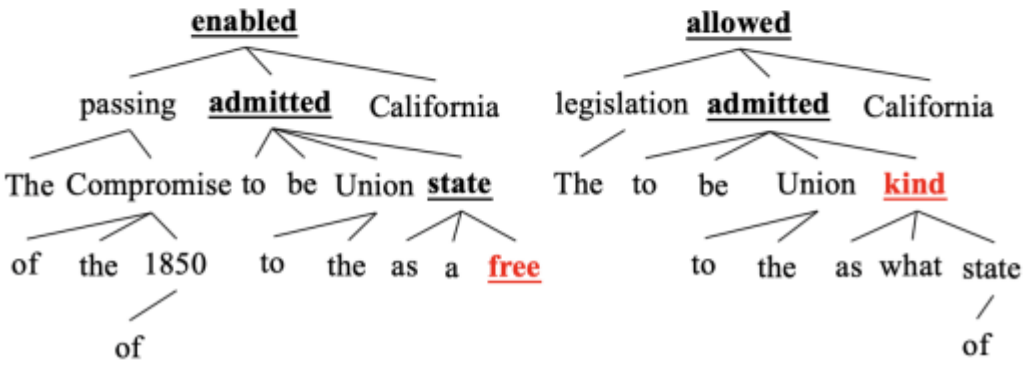
Passage:

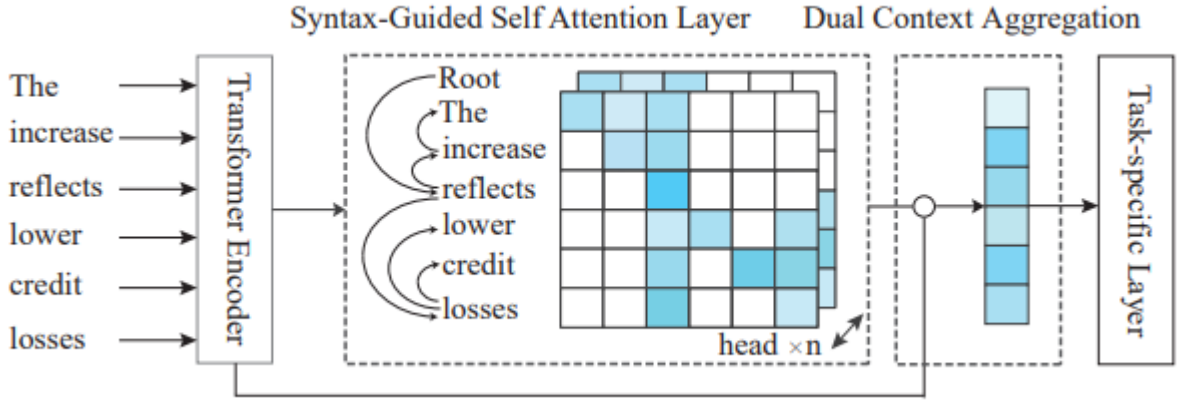
The passing of the Compromise of 1850 **enabled** California to be **admitted** to the Union as a **free** **state**, preventing southern California from becoming its own separate slave state…

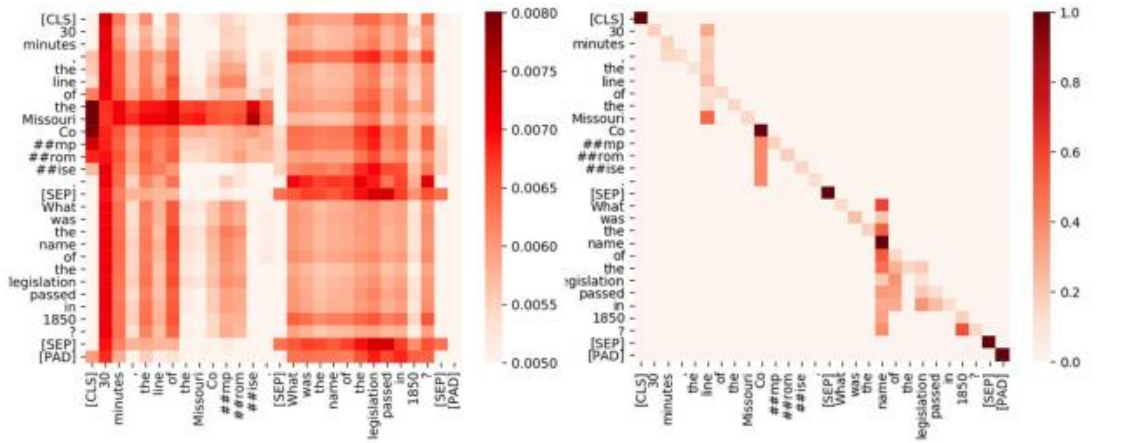
Question: The legislation **allowed** California to be **admitted** to the Union as what **kind** of state?

Answer: **free**

(a)

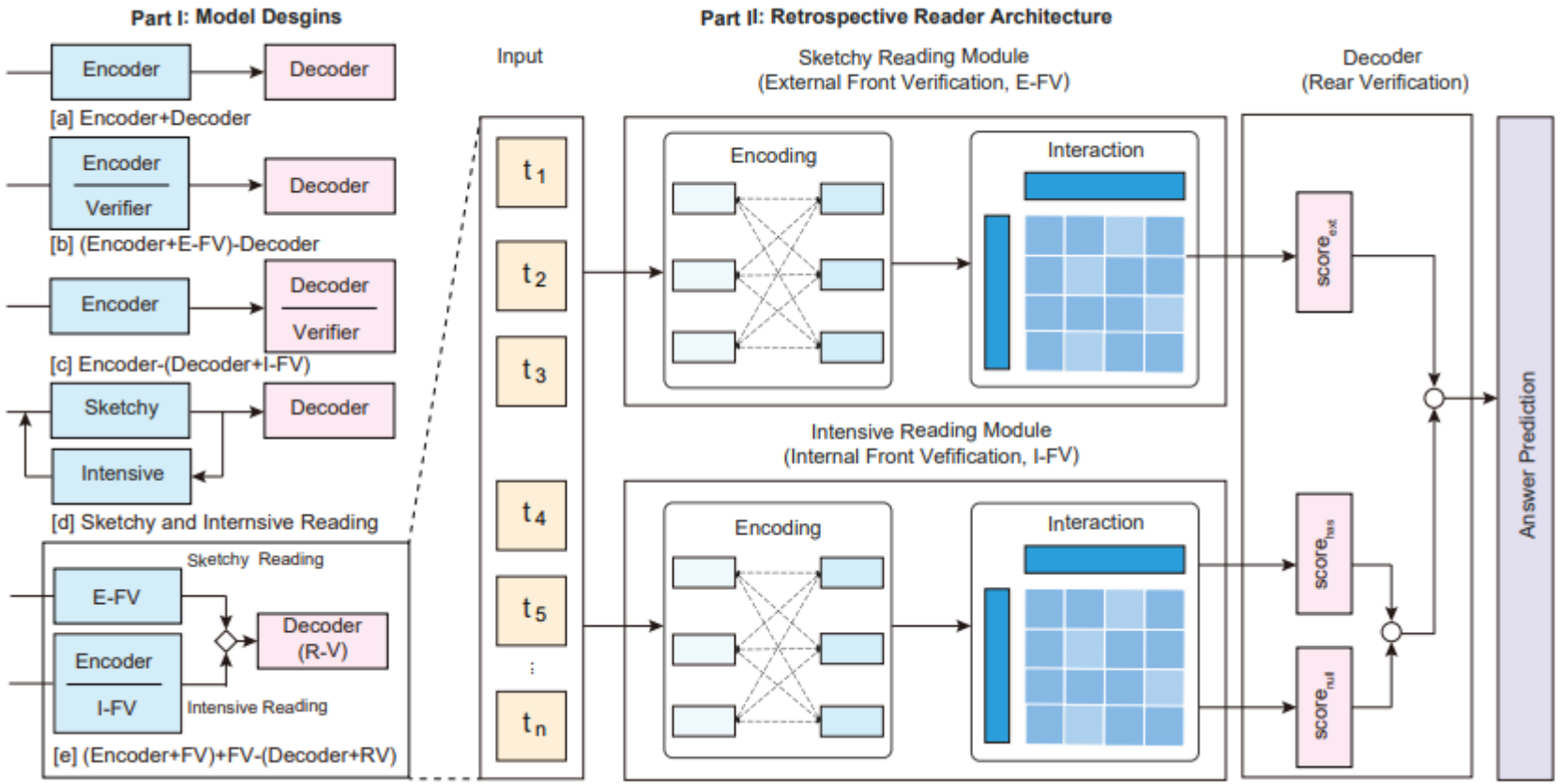






Passage (extract):*...30 minutes, the line of the Missouri Compromise...* ***Question***:*What was the name of the legislation passed in 1850?* ***Answer***:*the Missouri Compromise*

Retrospective Reader (AAAI 2021)



Passage:
Southern California consists of a heavily developed urban environment, home to some of the largest urban areas in the state, along with vast areas that have been left undeveloped. It is the third most populated megalopolis in the United States, after the Great Lakes Megalopolis and the Northeastern megalopolis. Much of southern California is famous for its large, spread-out, suburban communities and use of automobiles and highways. The dominant areas are Los Angeles, Orange County, San Diego, and Riverside-San Bernardino, each of which are the centers of their respective metropolitan areas...

Question:
What are the second and third most populated megalopolis after Southern California?

Answer:
Gold: ⟨no answer⟩
ALBERT (+TAV): Great Lakes Megalopolis and the Northeastern megalopolis.
Retro-Reader over ALBERT: ⟨no answer⟩
$score_{has} = 0.03, score_{na} = 1.73, \delta = -0.98$