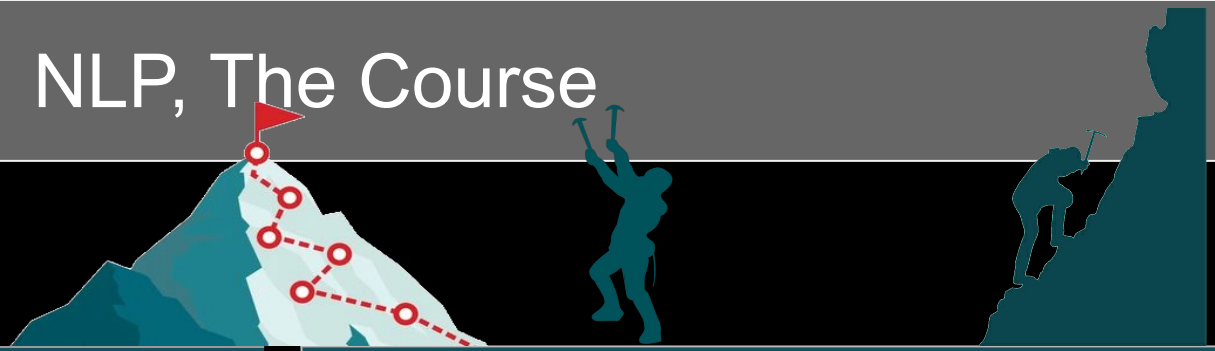
Evaluating and Enhancing Reasoning Capabilities of LMs

Guest Lecture: Vasudha Varadarajan

Slide credits: Yash Kumar Lal, Andrew Schwartz, and Vasudha Varadarajan

NLP, The Course



Overall NLP Concept

I. Syntax

Introduction to NLP; Tokenization; Words Corpora

One-hot, and Multi-hot encoding. Parts-of-Speech; Named Entities;

Parsing; Verbal Predicates;Dependency Parsing

II. Semantics

Dependency Parsing; Word Sense Disambiguation

Vector Semantics (Embeddings), Word2vec

Probabilistic Language Models
Ngram Classifier, Topic Modeling

Overall NLP Concept

III. Language Modeling

Ethical Considerations

Masked Language Modeling (autoencoding)

Generative Language Modeling (autoregressive)

Applying LMs

IV. Applications

Language and Psychology
(advanced sentiment)

Speech and Audio Processing, Dialog (chatbots)

Reasoning Capabilities of LLMs

Reasoning Capabilities of LLMs

Fact-based reasoning (Objective)

Reasoning Capabilities of LLMs

Fact-based reasoning (Objective)

Question (Q) →  → Answer (A)

Cognition and Theory-of-Mind (Subjective)

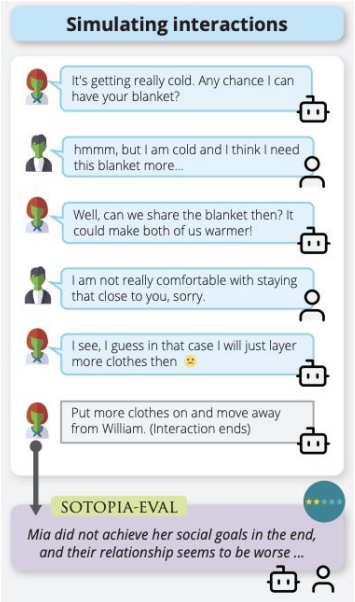
Why did this person *behave* or *think* this way?

Reasoning Capabilities of LLMs

Fact-based reasoning (Objective)

Question (Q) →  → Answer (A)

Social-intelligence (Subjective)



Cognition and Theory-of-Mind (Subjective)

Why did this person *behave* or *think* this way?

1. Question Answering

Question (Q) →  → Answer (A)

Information

Documents (corpus)

Document

Knowledge Base

Other modalities of data (image, video...)

Question

Factoid vs non-factoid

Open vs closed domain

Simple vs multi-step

Answer

Single fact

Explanation

Document

Extracted span

Image or other object

What is Question Answering?

What is Question Answering?

The goal of question answering is to build systems that automatically answer questions posed by humans in a natural language

What is Question Answering?

Question (Q) →  → Answer (A)

The goal of question answering is to build systems that automatically answer questions posed by humans in a natural language

Who is the first person to go to Mariana Trench?

The first person to go to the Mariana Trench was the American oceanographer and adventurer Don Walsh, who descended to its deepest point, the Challenger Deep, in 1960.

What is Question Answering?

Question (Q) →  → Answer (A)

The goal of question answering is to build systems that automatically answer questions posed by humans in a natural language

Who is the first person to go to Mariana Trench?

The first person to go to the Mariana Trench was the American oceanographer and adventurer Don Walsh, who descended to its deepest point, the Challenger Deep, in 1960.

Q: From a user's perspective, are you happy with the answer?

What is Question Answering?

Question (Q) →  → Answer (A)

The goal of question answering is to build systems that automatically answer questions posed by humans in a natural language

GPT-4 visual input example, Extreme Ironing:

User What is unusual about this image?

Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

GPT-4 The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

What is Question Answering?

Question (Q) →  → Answer (A)

The goal of question answering is to build systems that automatically answer questions posed by humans in a natural language

GPT-4 visual input example, Extreme Ironing:

User What is unusual about this image?

Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

GPT-4 The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

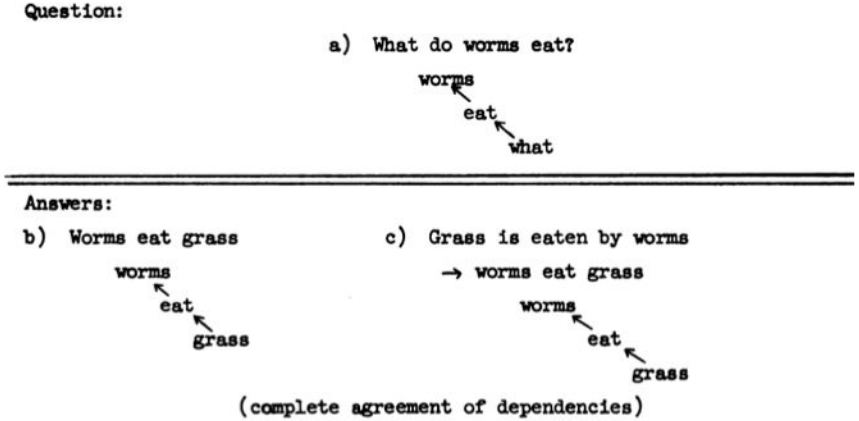
Q: From a user's perspective, are you happy with the answer?

What is Question Answering?

Question (Q) →  → Answer (A)

The goal of question answering is to build systems that automatically answer questions posed by humans in a natural language

The earliest QA systems were built in the 1960s!
(Simmons et al., 1964)



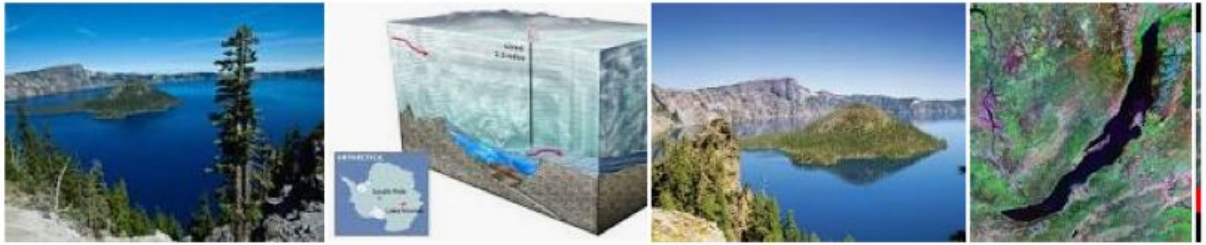
Real-World Applications Everywhere!

Google

Where is the deepest lake in the world?

All | Maps | Images | News | Videos | More | Settings | Tools

About 21,100,000 results (0.71 seconds)



Siberia

Lake **Baikal**, in Siberia, holds the distinction of being both the deepest lake in the world and the largest freshwater lake, holding more than 20% of the unfrozen fresh water on the surface of Earth.

Real-World Applications Everywhere!

Google

How can I protect myself from COVID-19?

All | Images | News | Shopping | Videos | More | Settings | Tools

The best way to prevent illness is to avoid being exposed to this virus. Learn how COVID-19 spreads and practice these actions to help prevent the spread of this illness.

To help prevent the spread of COVID-19:

- Cover your mouth and nose with a mask when around people who don't live with you. Masks work best when everyone wears one.
- Stay at least 6 feet (about 2 arm lengths) from others.
- Avoid crowds. The more people you are in contact with, the more likely you are to be exposed to COVID-19.
- Avoid unventilated indoor spaces. If indoors, bring in fresh air by opening windows and doors.
- Clean your hands often, either with soap and water for 20 seconds or a hand sanitizer that contains at least 60% alcohol.
- Get vaccinated against COVID-19 when it's your turn.
- Avoid close contact with people who are sick.
- Cover your cough or sneeze with a tissue, then throw the tissue in the trash.
- Clean and disinfect frequently touched objects and surfaces daily.

Learn more on cdc.gov

For informational purposes only. Consult your local medical authority for advice.

Areas in Question Answering

Reading Comprehension	● Answer based on a document ● Context is a one (or more) document(s)
Open-Domain QA	● Answer based on encyclopedic knowledge ● Context is the Internet (all knowledge)
Visual QA	● Answer is simple and factual ● Context is one/multiple image(s)

Areas in Question Answering

Reading Comprehension	• Answer based on a document • Context is a one (or more) document(s)
Open-Domain QA	• Answer based on encyclopedic knowledge • Context is the Internet (all knowledge)
Visual QA	• Answer is simple and factual • Context is one/multiple image(s)

Reading Comprehension

Comprehend a passage of text and answer questions about its content

Passage (P) + Question (Q) → Answer (A)

Reading Comprehension (MCTest)

Comprehend a passage of text and answer questions about its content

Passage (*P*) + Question (*Q*) → Answer (*A*)

P

Alyssa got to the beach after a long trip. She's from Charlotte. She traveled from Atlanta. She's now in Miami. She went to Miami to visit some friends. But she wanted some time to herself at the beach, so she went there first. After going swimming and laying out, she went to her friend Ellen's house. Ellen greeted Alyssa and they both had some lemonade to drink. Alyssa called her friends Kristin and Rachel to meet at Ellen's house.

Q Why did Alyssa go to Miami?

A To visit some friends

Richardson et al, 2013.

Reading Comprehension (MCTest)

Comprehend a passage of text and answer questions about its content

Passage (*P*) + Question (*Q*) → Answer (*A*)

P

Alyssa got to the beach after a long trip. She's from Charlotte. She traveled from Atlanta. She's now in Miami. She went to Miami to visit some friends. But she wanted some time to herself at the beach, so she went there first. After going swimming and laying out, she went to her friend Ellen's house. Ellen greeted Alyssa and they both had some lemonade to drink. Alyssa called her friends Kristin and Rachel to meet at Ellen's house.

Q Why did Alyssa go to Miami?

A To visit some friends

- ~3k questions from ~1k articles
- Multiple-choice questions
- Need for paraphrase, coreference resolution and dealing with many distractors

Richardson et al, 2013.

Reading Comprehension (SQuAD)

Comprehend a passage of text and answer questions about its content

Passage (*P*) + Question (*Q*) → Answer (*A*)

P

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".

Q

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

A

Within a cloud

Rajpurkar et al, 2016.

Reading Comprehension (SQuAD)

Comprehend a passage of text and answer questions about its content

Passage (*P*) + Question (*Q*) → Answer (*A*)

P

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".

Q

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

A

Within a cloud

- 100k annotated (passage, question, answer) triples
- Answer is a short segment of text (or span) in passage
- Questions are crowd-sourced, passages are from English Wikipedia, usually 100-150 words long

Rajpurkar et al, 2016.

Evaluating Reading Comprehension

Passage (*P*) + Question (*Q*) → Answer (*A*)

P

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”.

Q

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

A

Within a cloud

Inside clouds

Clouds

M

Collide inside clouds

- Exact match (EM): 0 or 1
- max{0, 0, 0} = 0

Rajpurkar et al, 2016.

Evaluating Reading Comprehension

Passage (*P*) + Question (*Q*) → Answer (*A*)

P

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”.

Q

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

A

Within a cloud

Inside clouds

Clouds

M

Collide inside clouds

- F1: Partial credit
- max{0.33, 0.67, 0.33} = 0.67

Rajpurkar et al, 2016.

Models for Reading Comprehension

Passage (P) + Question (Q) Answer (A)

Input: $P = (p_1, p_2, \ldots, p_N)$, $Q = (q_1, q_2, \ldots, q_M)$ N~100, M~15

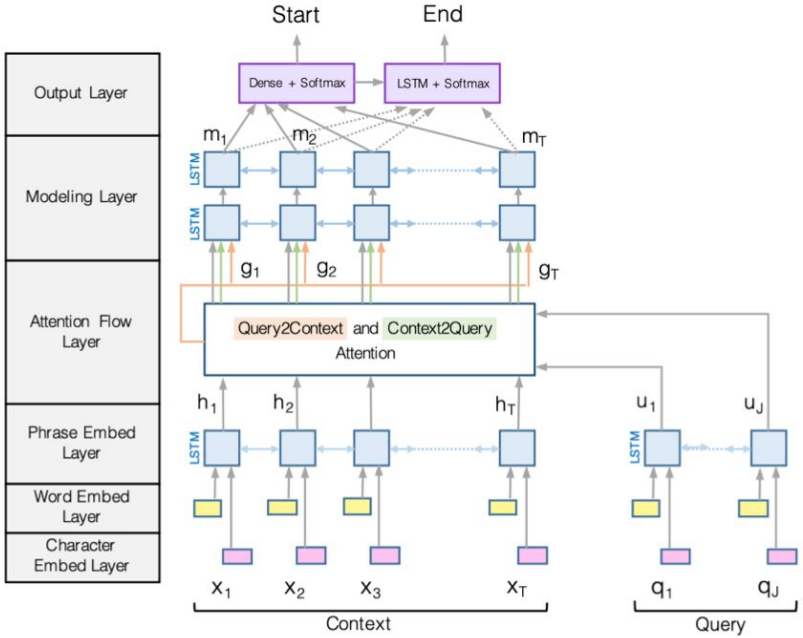
Output: $0 < \text{start} < \text{end} < N+1$ answer is a span in the passage

LSTM-Based Models for Reading Comprehension

Passage (P) + Question (Q) Answer (A)

Input: $P = (p_1, p_2, \ldots, p_N)$, $Q = (q_1, q_2, \ldots, q_M)$ N~100, M~15

Output: $0 < \text{start} < \text{end} < N+1$ answer is a span in the passage



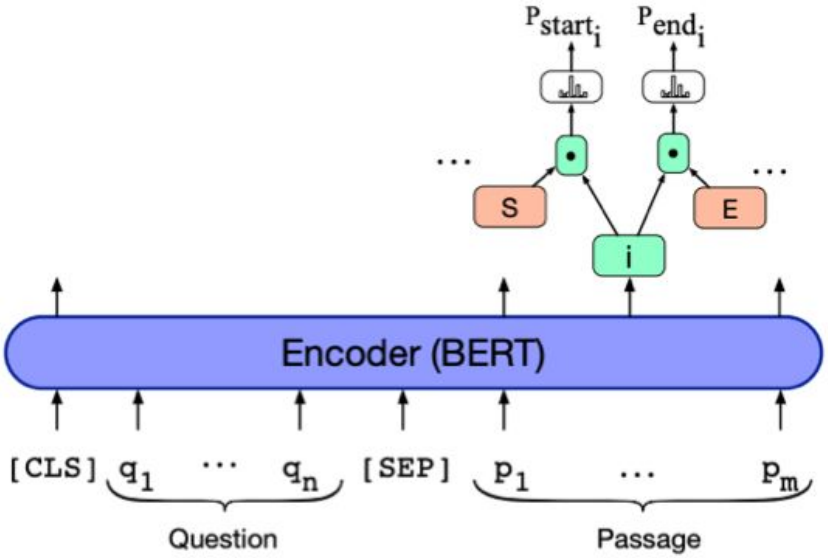
Seo et al, 2017.

BERT-Based Models for Reading Comprehension

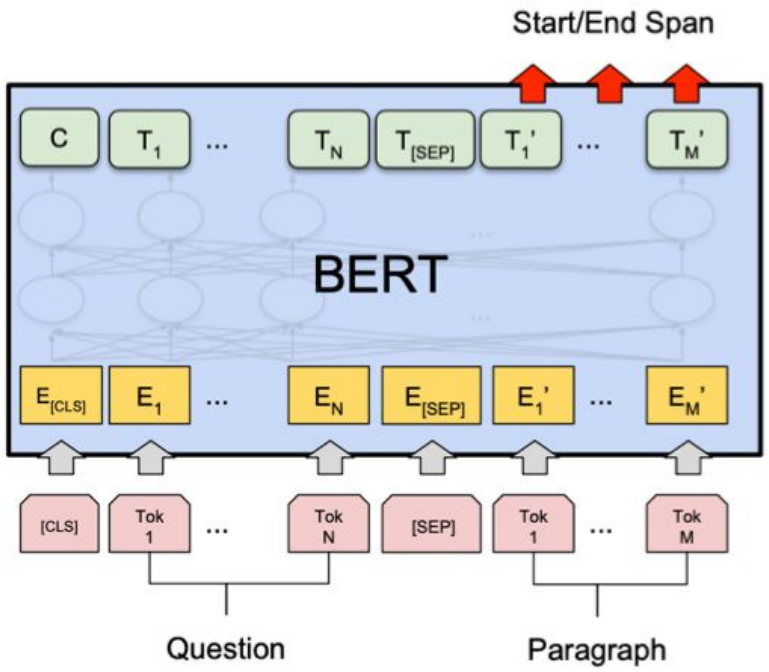
Passage (P) + Question (Q) Answer (A)

Input: $P = (p_1, p_2, \ldots, p_N)$, $Q = (q_1, q_2, \ldots, q_M)$ N~100, M~15

Output: $0 < \text{start} < \text{end} < N+1$ answer is a span in the passage

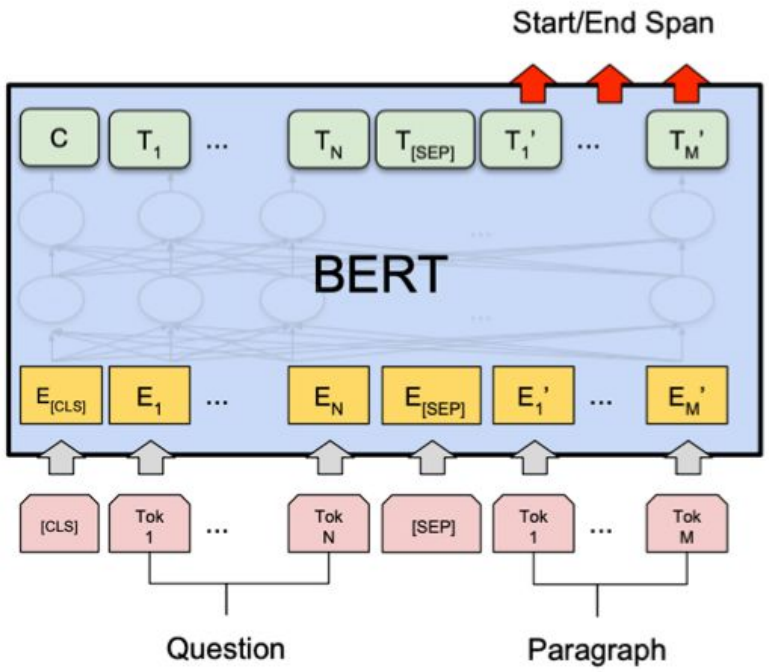


BERT-Based Models for Reading Comprehension

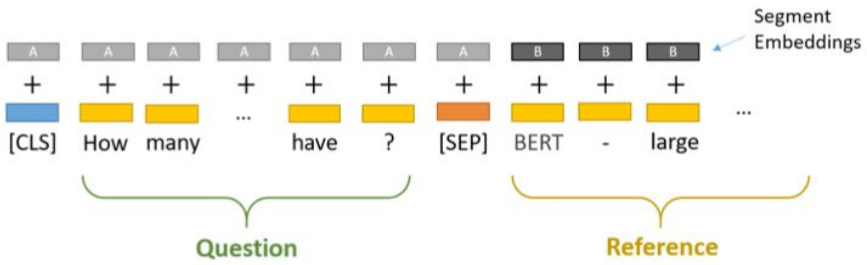


Question = Segment A
Passage = Segment B
Answer = predicting two endpoints in segment B

BERT-Based Models for Reading Comprehension



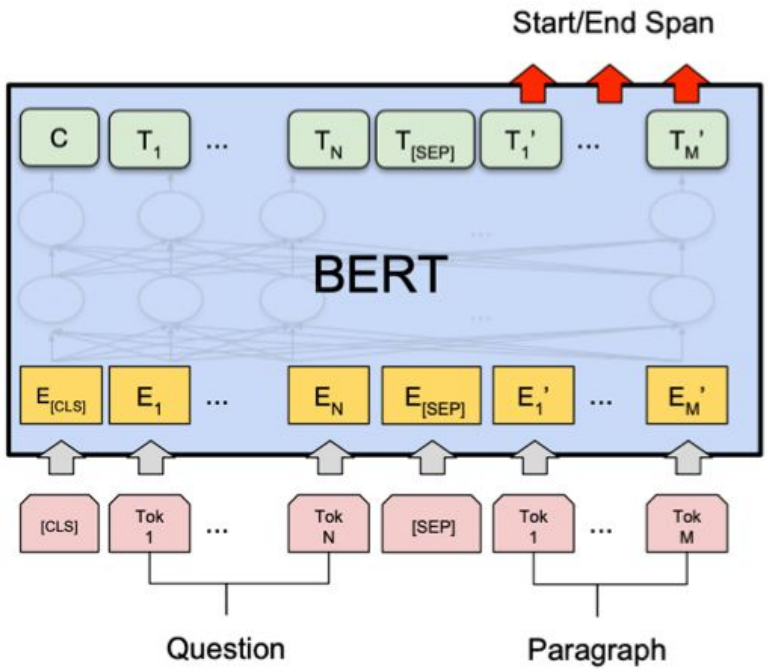
Question = Segment A
Passage = Segment B
Answer = predicting two endpoints in segment B



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-Based Models for Reading Comprehension



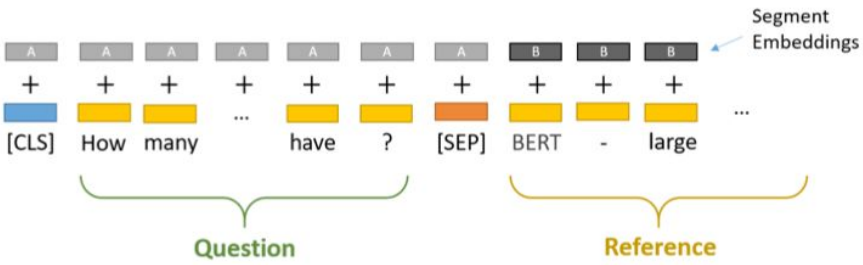
$$\mathcal{L} = -\log p_{\text{start}}(s^*) - \log p_{\text{end}}(e^*)$$

$$p_{\text{start}}(i) = \text{softmax}_i(\mathbf{w}_{\text{start}}^{\intercal}\mathbf{h}_i)$$

$$p_{\text{end}}(i) = \text{softmax}_i(\mathbf{w}_{\text{end}}^{\intercal}\mathbf{h}_i)$$

where $\mathbf{h}_i$ is the hidden vector of c_i, returned by BERT

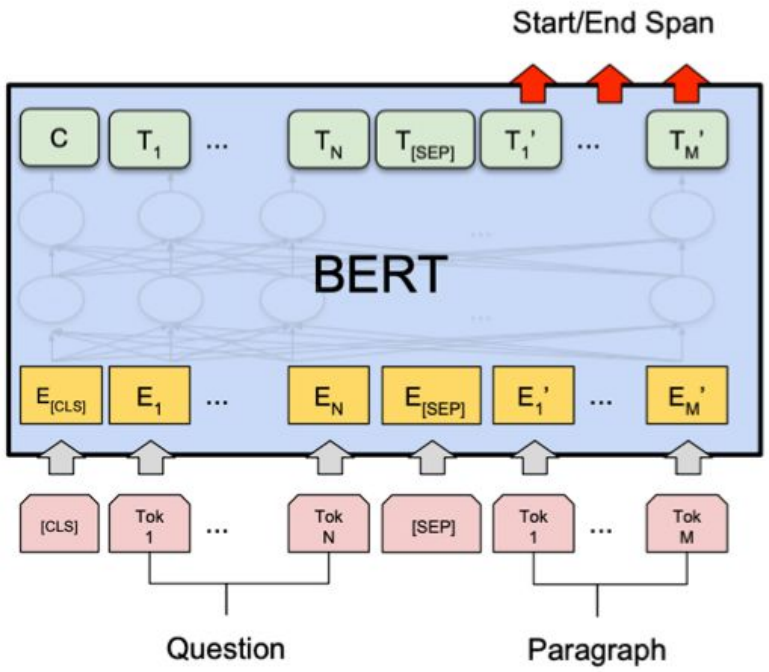
Question = Segment A
Passage = Segment B
Answer = predicting two endpoints in segment B



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.

Is Reading Comprehension Solved?



	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

dev set, except for human performance

Is Reading Comprehension Solved?

Questions that require long answers

Question: How do Jellyfish function without brains or nervous systems? [...] (60 words)

Answer: Jellyfish may not have a brain, but they have a rough nervous system and innate behaviours. However, they are very simple creatures. They're invertebrate: creatures without a backbone. Most jellyfish have really short life spans. Sometimes just a couple of hours. [...] As their name implies, they are largely composed of basically jelly inside a thin membrane. They're over 95% water. (327 words)

Documents: [...] Jellyfish do not have brains, and most barely have nervous systems. They have primitive nerve cells that help them orient themselves in the water and sense light and touch. [...] While they dont possess brains, the animals still have neurons that send all sorts of signals throughout their body. [...] They may accomplish this through the assistance of their nerve rings. Jellyfish don't have brains, and that's just where things begin. They don't have many of the body parts that are typical in other animals. [...] (1070 words)

ELI5 (Fan et al., 2019)

Is Reading Comprehension Solved?

Questions that require long answers

Question: How do Jellyfish function without brains or nervous systems? [...] (60 words)

Answer: Jellyfish may not have a brain, but they have a rough nervous system and innate behaviours. However, they are very simple creatures. They're invertebrate: creatures without a backbone. Most jellyfish have really short life spans. Sometimes just a couple of hours. [...] As their name implies, they are largely composed of basically jelly inside a thin membrane. They're over 95% water. (327 words)

Documents: [...] Jellyfish do not have brains, and most barely have nervous systems. They have primitive nerve cells that help them orient themselves in the water and sense light and touch. [...] While they dont possess brains, the animals still have neurons that send all sorts of signals throughout their body. [...] They may accomplish this through the assistance of their nerve rings. Jellyfish don't have brains, and that's just where things begin. They don't have many of the body parts that are typical in other animals. [...] (1070 words)

ELI5 (Fan et al., 2019)

Questions that require discrete reasoning

Q: Where did Charles travel to first, Castile or Barcelona?

In 1517, the seventeen-year-old King sailed to Castile, where he was formally recognised as King of Castile. There, his Flemish court provoked much scandal, … In May 1518, Charles traveled to Barcelona in Aragon, where he would remain for nearly two years.

DROP (Dua et al., 2019)

Is Reading Comprehension Solved?

As of April 2025: Somewhat!

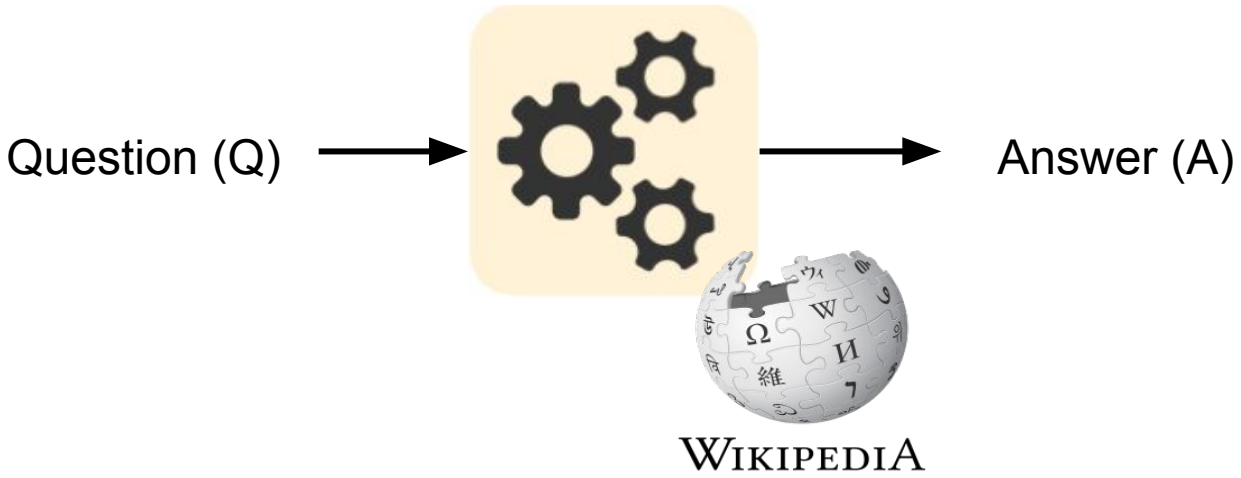
(2023;older work)

Setting	CoQA	DROP	QuAC	SQuADv2	RACE-h	RACE-m
Fine-tuned SOTA	**90.7**[a]	**89.1**[b]	**74.4**[c]	**93.0**[d]	**90.0**[e]	**93.1**[e]
GPT-3 Zero-Shot	81.5	23.6	41.5	59.5	45.5	58.4
GPT-3 One-Shot	84.0	34.3	43.3	65.4	45.9	57.4
GPT-3 Few-Shot	85.0	36.5	44.3	69.8	46.8	58.1

Areas in Question Answering

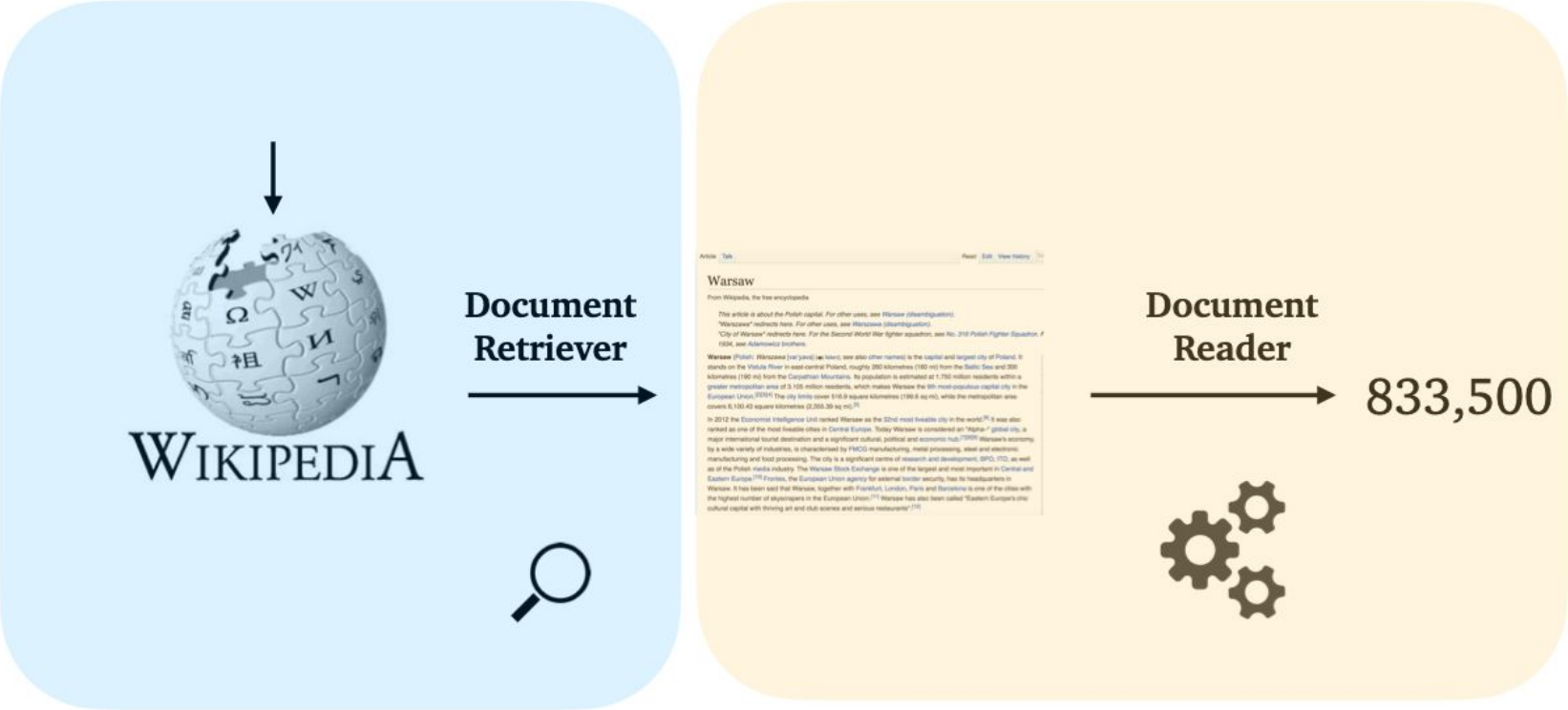
Reading Comprehension	● Answer based on a document ● Context is a one (or more) document(s)
Open-Domain QA	● Answer based on encyclopedic knowledge ● Context is the Internet (all knowledge)
Visual QA	● Answer is simple and factual ● Context is one/multiple image(s)

Open-Domain Question Answering



- No given passage, just a large collection of documents (e.g., Wikipedia)
- No idea where answer is located
- Have to answer any open-domain questions
- Very challenging, but more practical

Open-Domain Question Answering



Chen et al, 2017.

Retriever-Reader Framework

Input: $D = (D_1, D_2, ..., D_N)$, Q

Output: an answer string A

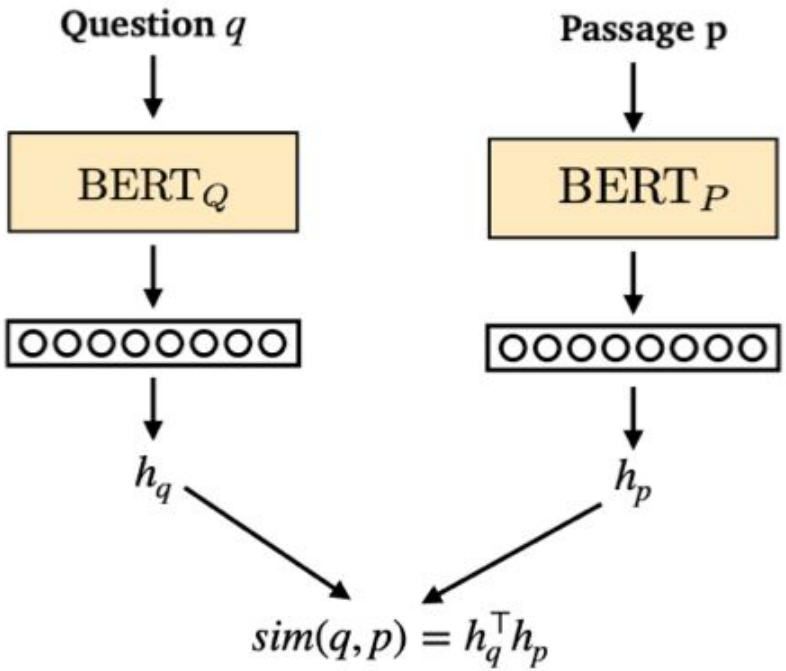
Retriever: $f(D, Q) \rightarrow (P_1, P_2, ..., P_K)$

Reader: $g(Q, \{P_1, P_2, ..., P_K\}) \rightarrow A$

D: large collection of documents

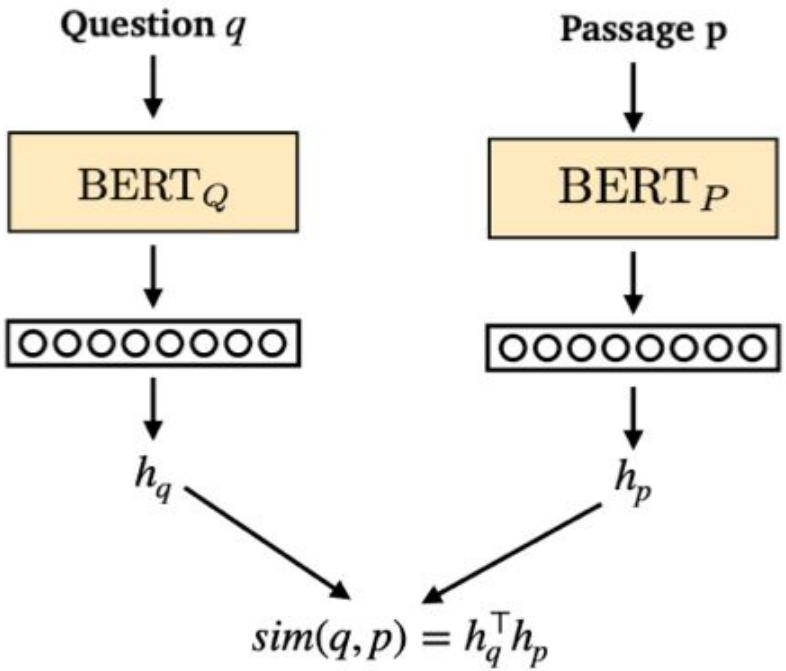
K is pre-defined

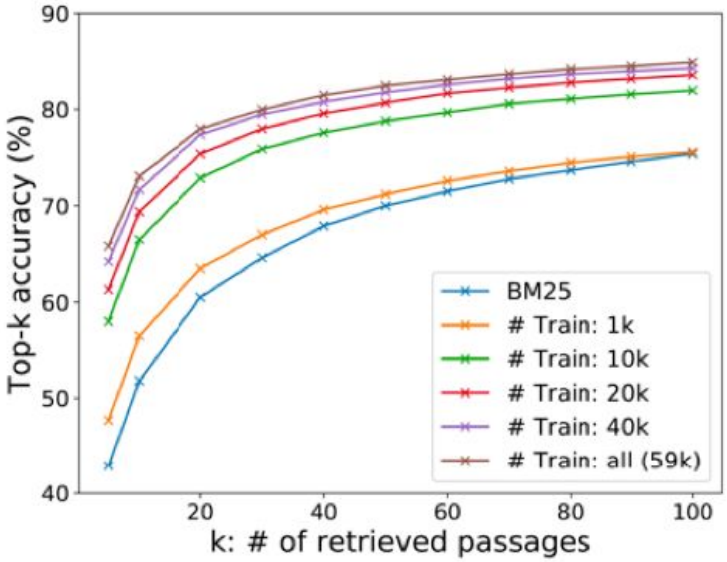
Dense Passage Retrieval



Karpukhin et al, 2020.

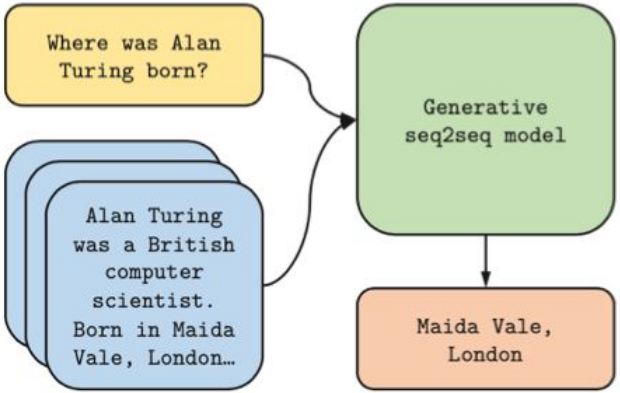
Dense Passage Retrieval





Karpukhin et al, 2020.

Dense Retrieval + Generative Models

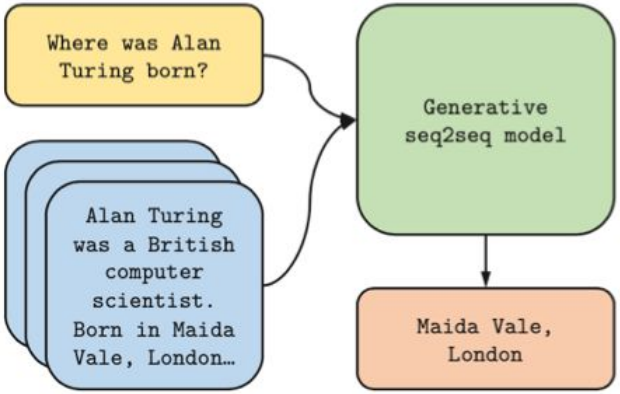


Fusion-in-decoder (FID)

DPR + T5

Izacard and Grave, 2020.

Dense Retrieval + Generative Models



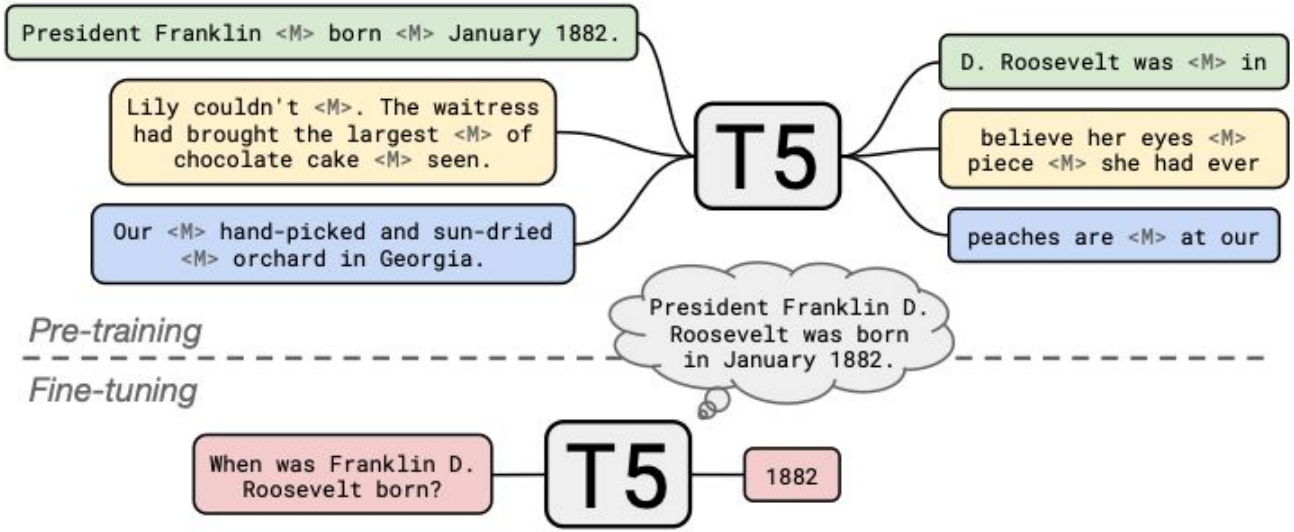
Model	NaturalQuestions	TriviaQA	
ORQA (Lee et al., 2019)	31.3	45.1	-
REALM (Guu et al., 2020)	38.2	-	-
DPR (Karpukhin et al., 2020)	41.5	57.9	-
SpanSeqGen (Min et al., 2020)	42.5	-	-
RAG (Lewis et al., 2020)	44.5	56.1	68.0
T5 (Roberts et al., 2020)	36.6	-	60.5
GPT-3 few shot (Brown et al., 2020)	29.9	-	71.2
Fusion-in-Decoder (base)	48.2	65.0	77.1
Fusion-in-Decoder (large)	**51.4**	**67.6**	**80.1**

Fusion-in-decoder (FID)

DPR + T5

Izacard and Grave, 2020.

Generative Models for Open-Domain QA



Roberts et al, 2020.

Multi-Step Questions Answering

Questions to answer which we need multiple steps of reasoning.

What are Multi-Step Questions

Questions to answer which we need multiple steps of reasoning.

Where did OpenAI's CEO go for undergrad?

- Who is OpenAI's CEO ⇒ Sam Altman
- Where did Sam Altman go for undergrad? ⇒ Stanford University

Stanford University

Why should we care about Multi-Step Questions?

Why should we care about Multi-Step Questions?

Many of our day-to-day information needs require multi-step reasoning.

Why should we care about Multi-Step Questions?

Many of our day-to-day information needs require multi-step reasoning.

Which vegetarian restaurants near me are open if I've a peanut allergy?

- Find a list of open restaurants near me.
- Select the ones which have vegetarian options in the menu.
- Select the ones which have peanut-free options in the menu.

Why should we care about Multi-Step Questions?

Many of our day-to-day information needs require multi-step reasoning.

Can I finish GOT Season 7 if I've 10 hours this weekend?

- Get a list of episodes and duration of GOT from season 7.
- Sum the time duration of GOT for all the episodes.
- Check if the total duration is less than 10 hours.

To satisfy such information needs, we need models that perform multi-step reasoning.

Chain-of-thought (CoT) Prompting

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. ✓

Wei et al. (2022)

Zero-shot Chain-of-thought Prompting

(a) Few-shot

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: 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:

(Output) The answer is 8. ✗

(b) Few-shot-CoT

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: 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:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are 16 / 2 = 8 golf balls. Half of the golf balls are blue. So there are 8 / 2 = 4 blue golf balls. The answer is 4. ✓

(c) Zero-shot

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: The answer (arabic numerals) is

(Output) 8 ✗

(d) Zero-shot-CoT (Ours)

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.***

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

Kojima et al. (2022)

What are the Challenges of Multi-Step QA?

Challenges of Multi-Step QA

Challenges of Multi-Step QA

Reading Comprehension QA

Where did OpenAI's CEO go for undergrad?

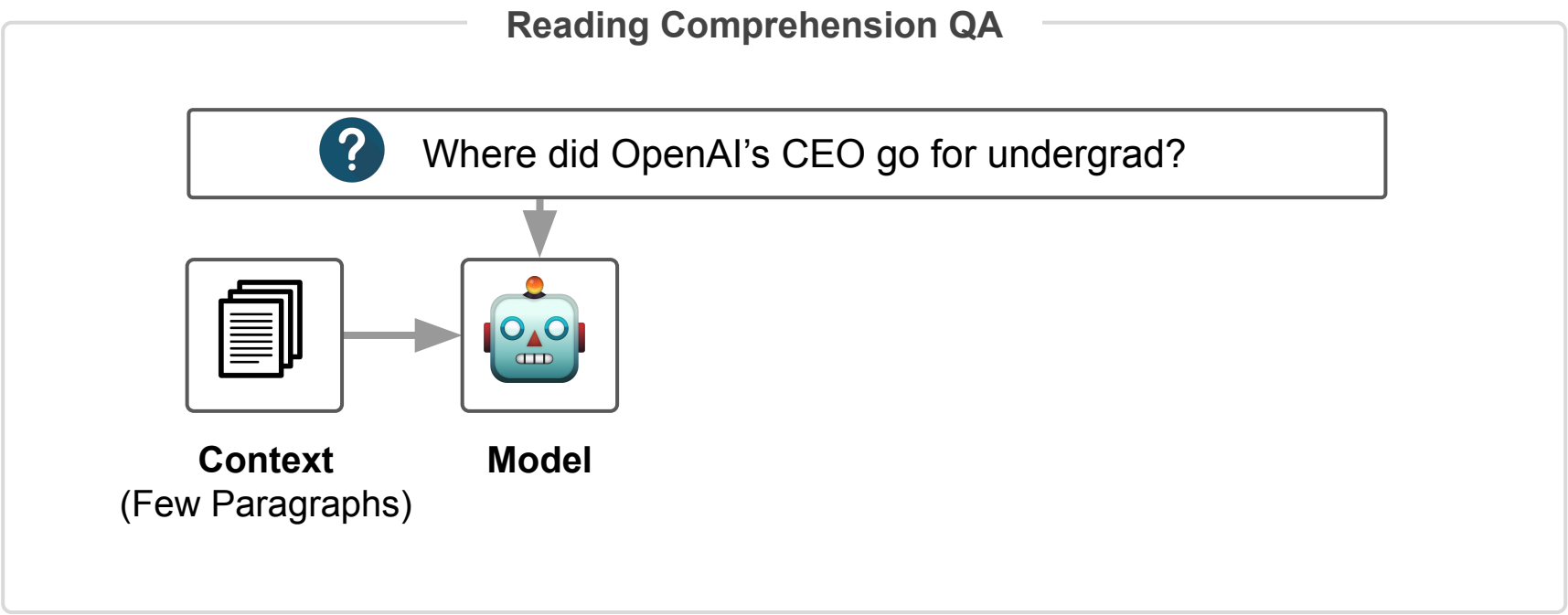
Challenges of Multi-Step QA

Reading Comprehension QA

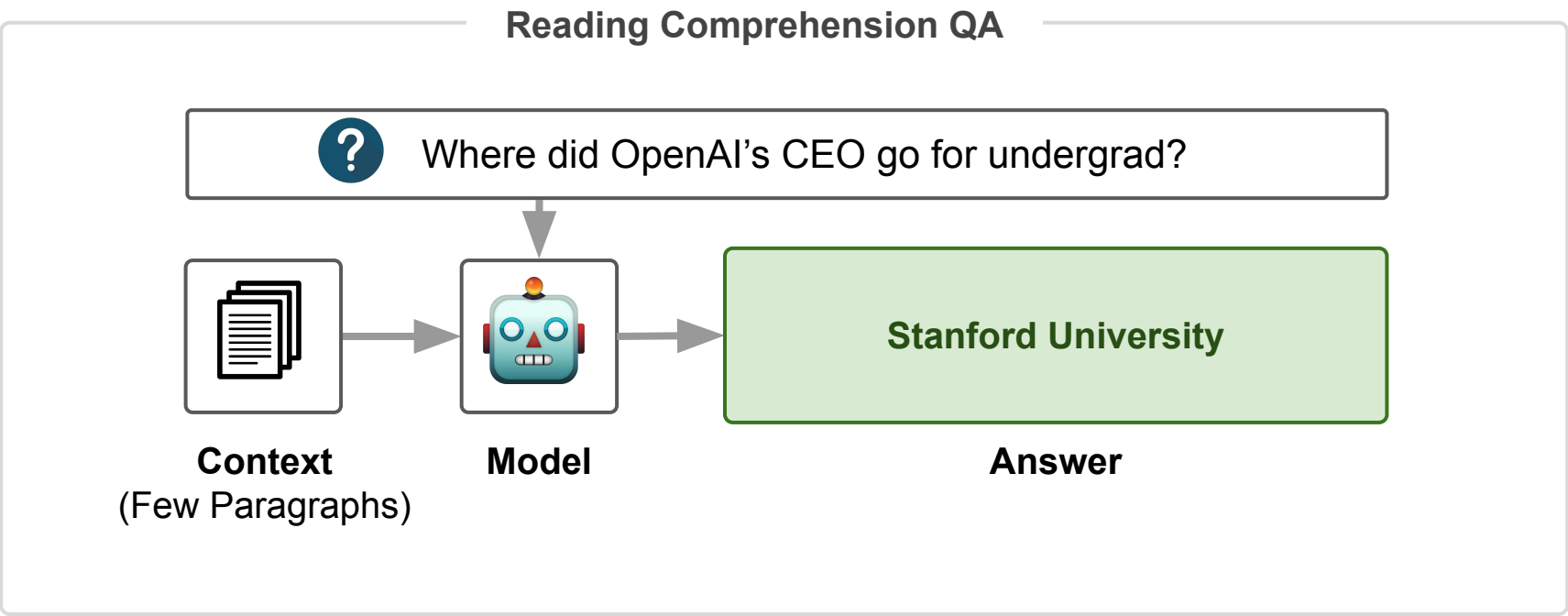
Where did OpenAI's CEO go for undergrad?

Context
(Few Paragraphs)

Challenges of Multi-Step QA



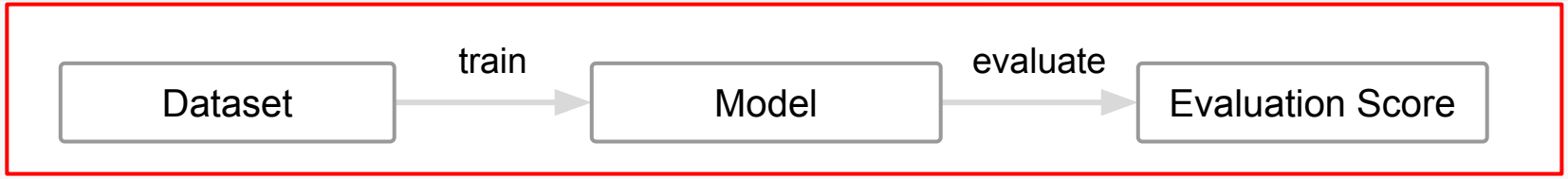
Challenges of Multi-Step QA

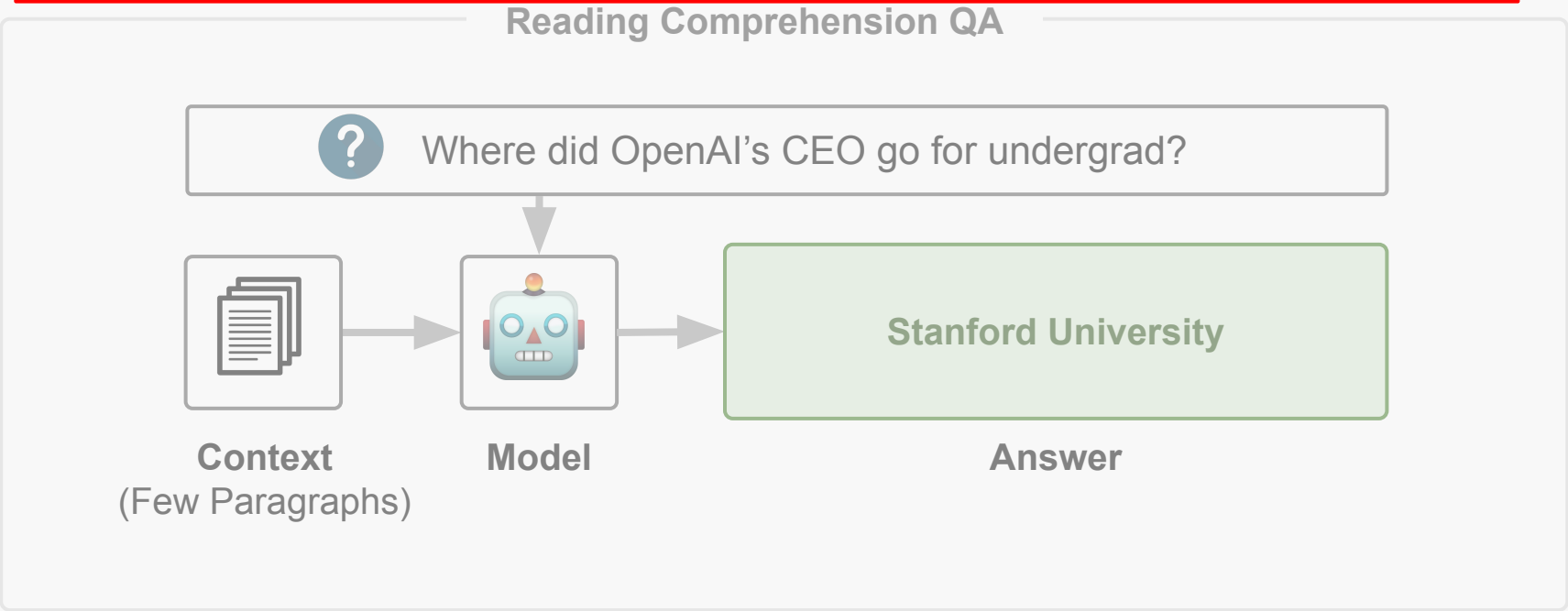


Challenges of Multi-Step QA

Reading Comprehension QA

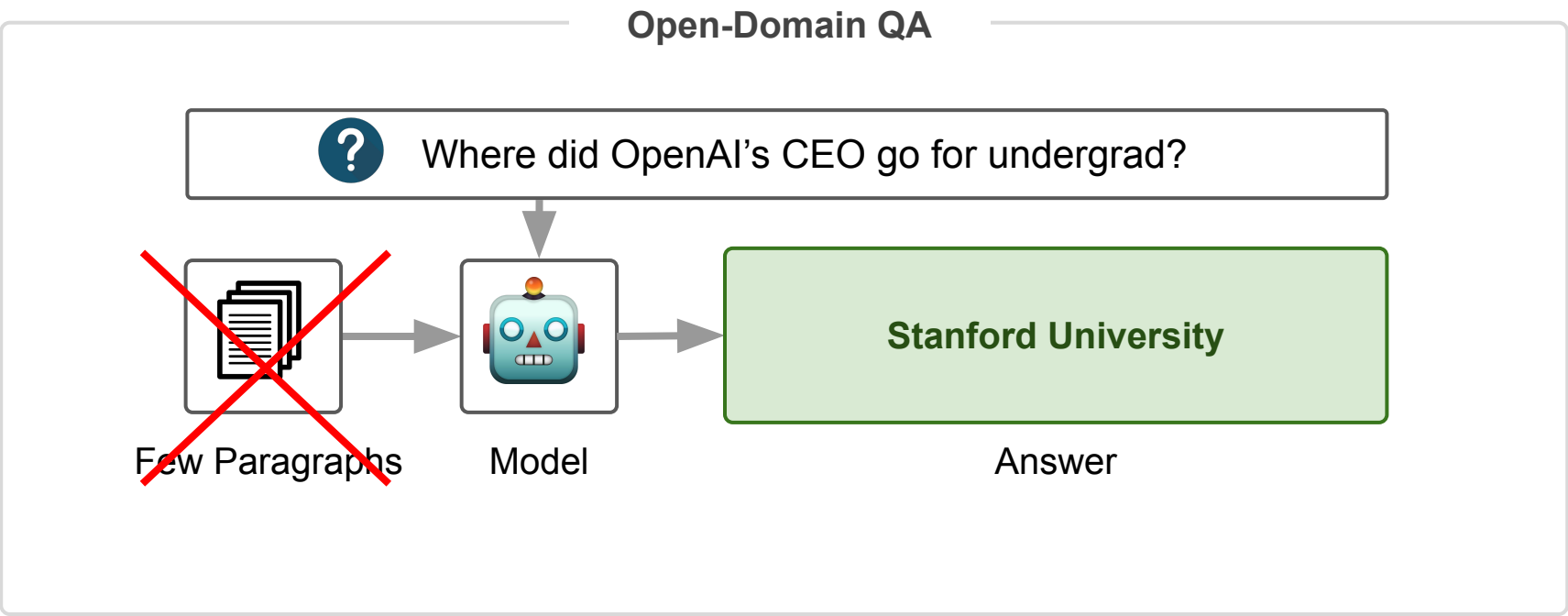
Open-Domain QA



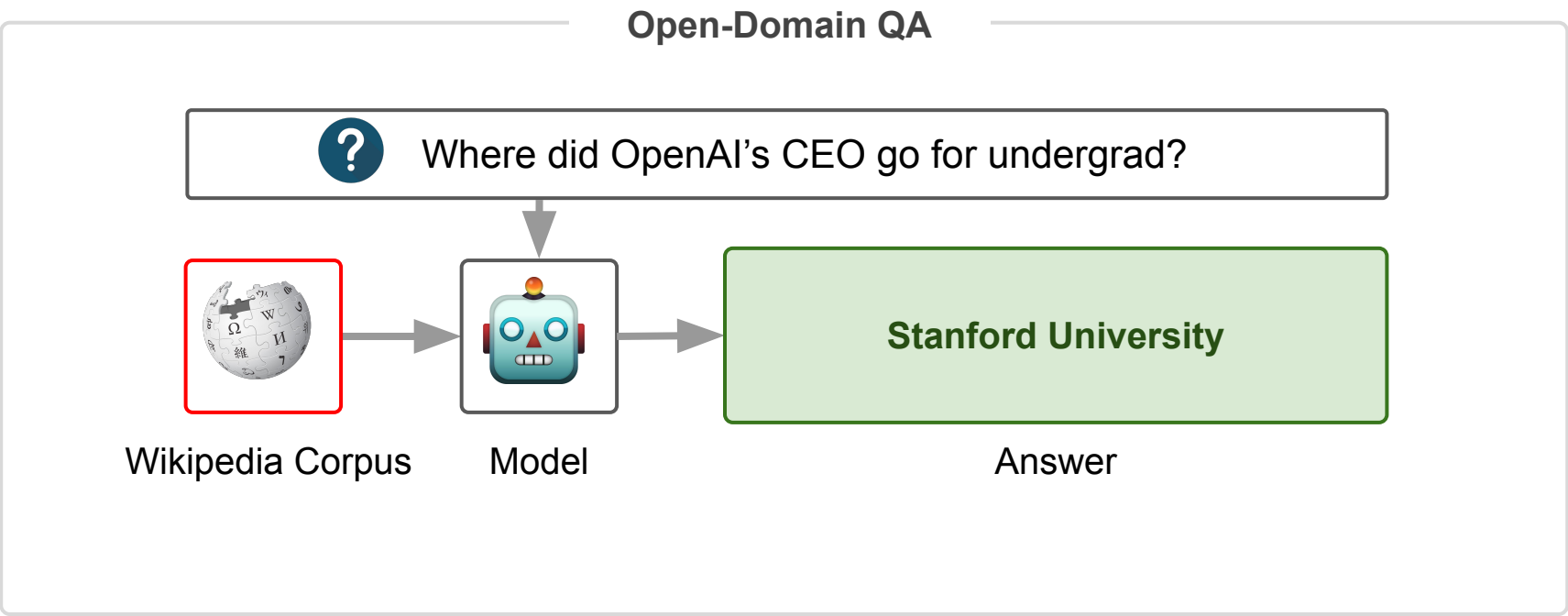


Challenges of Multi-Step QA

Challenges of Multi-Step QA



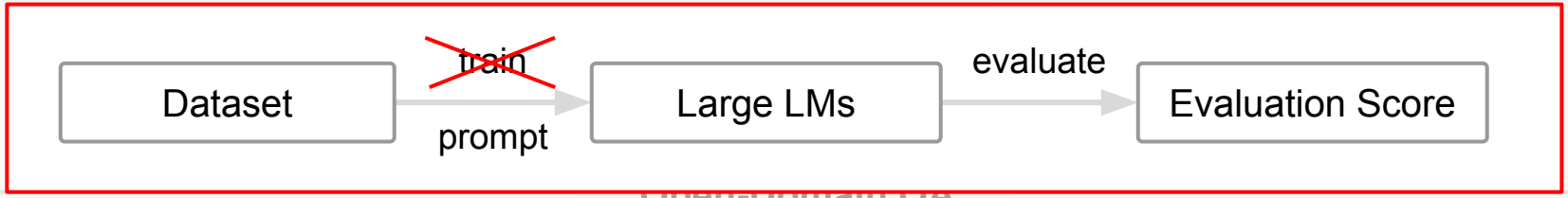
Challenges of Multi-Step QA



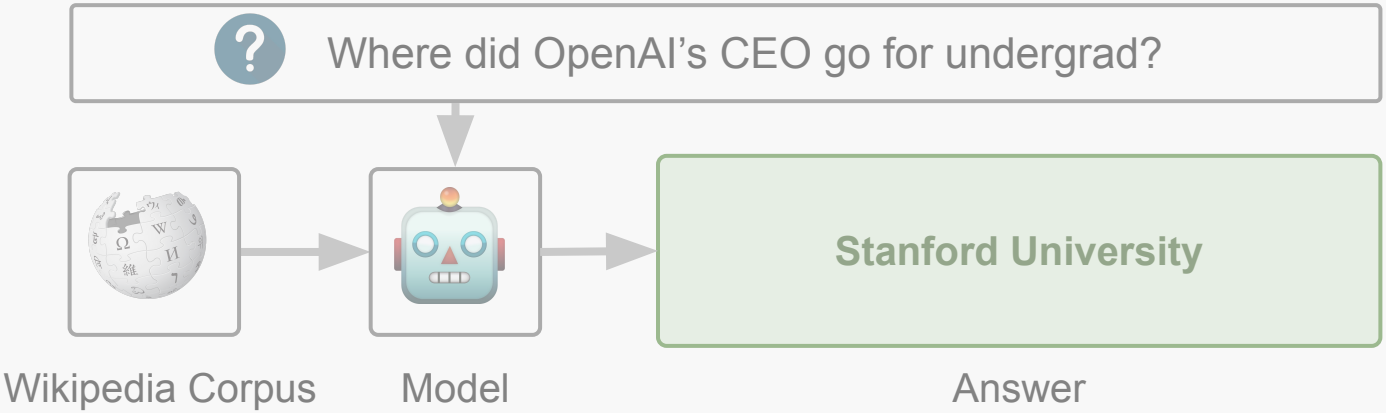
Challenges of Multi-Step QA

Reading Comprehension QA

Open-Domain QA



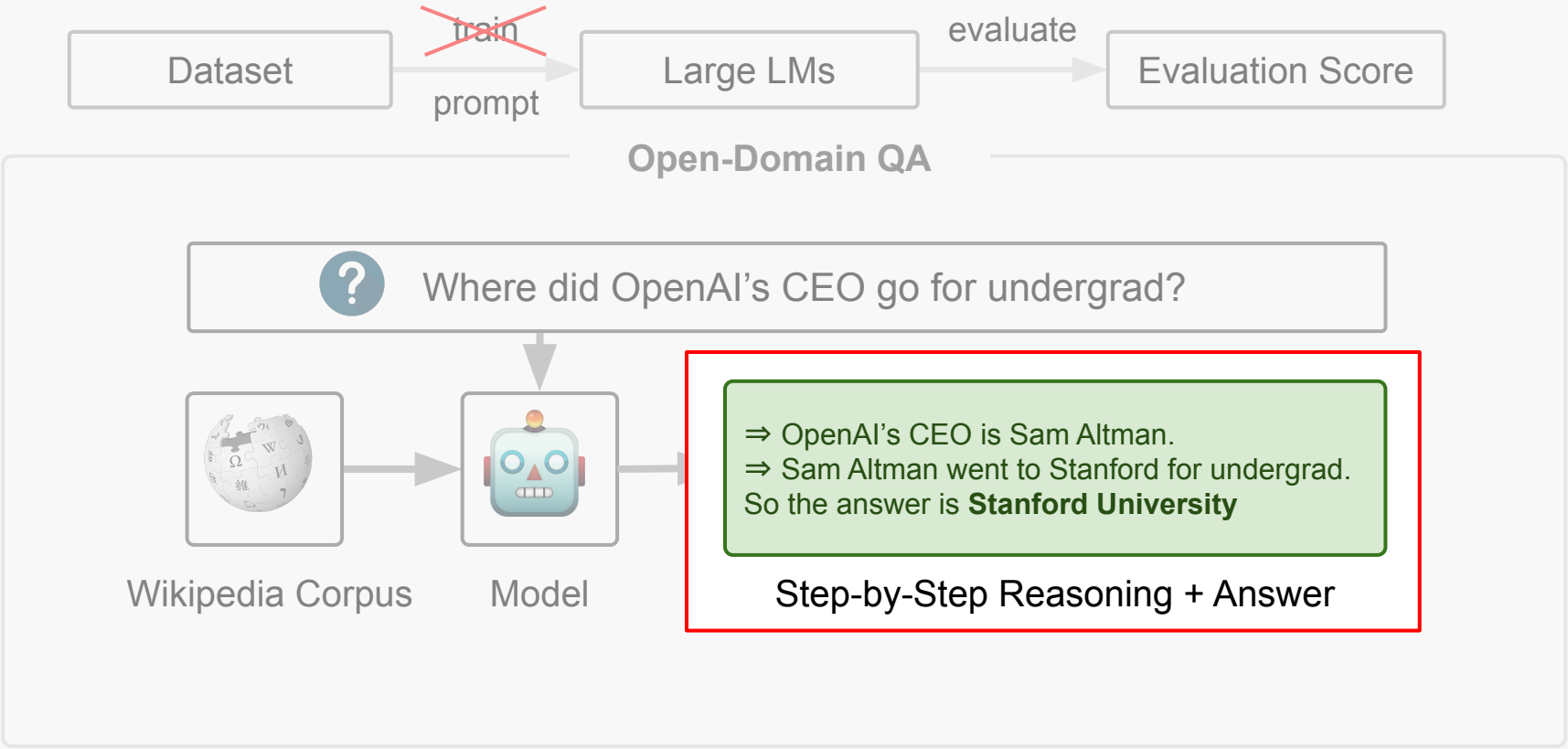
Open-Domain QA



Challenges of Multi-Step QA

Reading Comprehension QA

Open-Domain QA



State of Few-Shot Multi-Step Open-Domain QA

Model	$HpQA^{Br}$	HpQA	2WikiMQA	MQ^{2H}
InterAug	– | –	30.3 | –	– | –	– | –
ReAct	– | –	35.1 | –	– | –	– | –
SelfAsk	– | –	– | –	40.1 | –	15.2 | –
DecomP	– | 50.0	– | –	– | 59.3	– | –
IRCoT QA	**45.8 | 58.5**	**49.3 | 60.7**	**57.7 | 68.0**	**34.2 | 43.8**

EM | F1

⇒ InterAug: (Internet-augmented LMs through few-shot prompting for ODQA) Lazaridou et. al.
⇒ ReAct: (ReAct: Synergizing Reasoning and Acting in Language Models) Yao et. al
⇒ SelfAsk: (Measuring and Narrowing the Compositionality Gap in Language Models) Press et. al.
⇒ DecomP: (Decomposed Prompting: A Modular Approach for Solving Complex Tasks) Khot et. al.

Reasoning Capabilities of LLMs

Fact-based reasoning (Objective)

Question (Q) →  → Answer (A)

Social-intelligence (Subjective)

Cognition and Theory-of-Mind (Subjective)

Why did this person *behave* or *think* this way?

Theory-of-Mind

Sean puts the book in the box and leaves to get something to eat in the kitchen. While he is away, Anna moves the book from the box to the basket. Sean comes back into the room and wants to read more of his book.

Q: Sean thinks the book is in the ___.

Theory-of-Mind

Sean puts the book in the box and leaves to get something to eat in the kitchen. While he is away, Anna moves the book from the box to the basket. Sean comes back into the room and wants to read more of his book.

Q: Sean thinks the book is in the ___. *box* ✓ *other* ✗

What is Theory-of-Mind?

- Theory of mind is the ability to understand the *thoughts*, *beliefs*, *desires*, and *emotions* of other people.
- In this case, it is the ability of models to understand human beliefs, cognition, emotions and behaviors.

Benchmarks for Theory-of-Mind

False Belief (Wimmer & Perner, 1983) Free-text

Sean puts the book in the box and leaves to get something to eat in the kitchen. While he is away, Anna moves the book from the box to the basket. Sean comes back into the room and wants to read more of his book.

Q: Sean thinks the book is in the ___. *box* ✓ *other* ✗

Recursive Mindreading (O'Grady et al., 2015) 2AFC

[Story containing recursively embedded mental states]

Q: Which continuation is consistent with the story?

A) John thinks Sheila hasn't realised that he likes her. ✓

B) John thinks Sheila has realised that he likes her. ✗

Short Stories (Dodell-Feder et al., 2013) Manual Scoring

[*The End of Something* by Ernest Hemingway]

Q: Why does Nick say to Marjorie, "You know everything"?

He's being sarcastic to provoke a fight ✓

He thinks Marjorie is a know-it-all ✗

Strange Stories (Happé, 1994) Manual Scoring

Peter thinks Aunt Jane's hat is very ugly indeed. But when Aunt Jane asks Peter, "How do you like my new hat?", Peter says, "Oh, its very nice".

Q: Why does Peter say that?

He's lying to spare her feelings ✓ *Because he's nice* ✗

Indirect Request (Trott & Bergen, 2020) 2AFC

You and Jonathan both notice a blinking light, which indicates that the car's heating system is broken... Jonathan shivers in his seat. He turns to you and says, "Man, it's really cold in here."

Q: Do you think he is making a request? *No* ✓ *Yes* ✗

Scalar Implicature (Goodman & Stuhlmüller, 2013) Bet

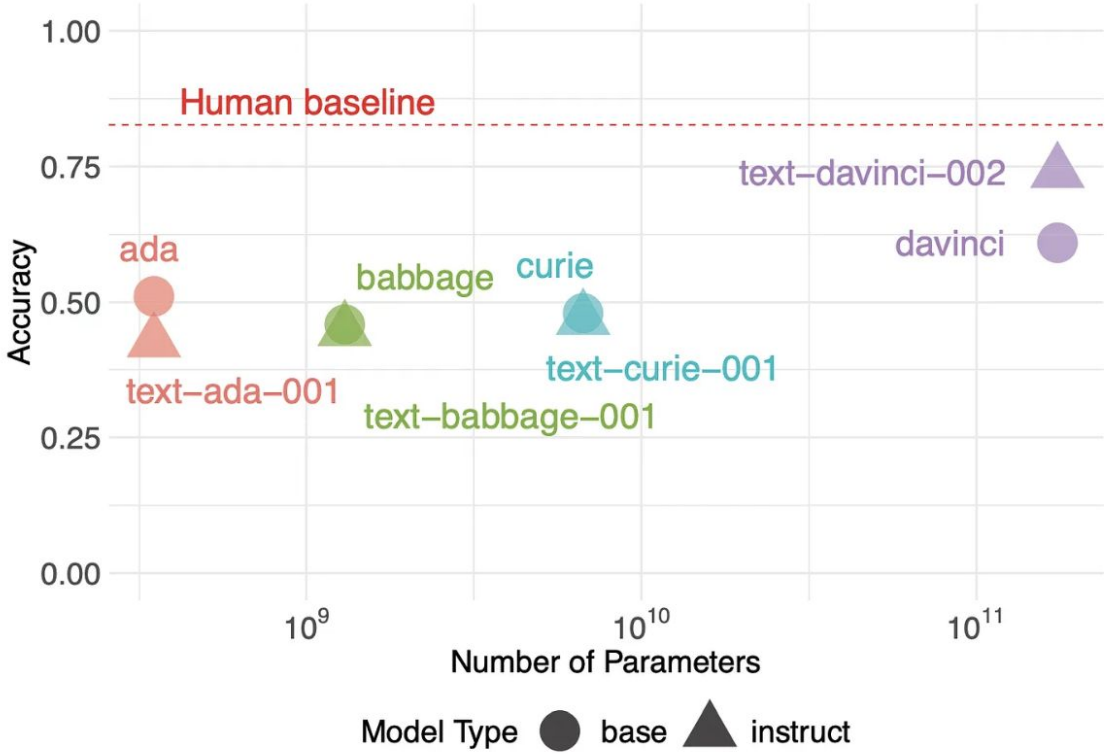
David ordered 3 pizzas which almost always have cheese in the crust. David tells you: "I have looked at 3 of the 3 pizzas. Some of the pizzas have cheese in the crust."

Q: How many pizzas do you think have cheese in the crust

Bet $100 across 4 options (0,1,2,3) *p(3)* ↓ ✓ *other* ✗

Jones et al. (2023)

Performance of recent LLMs



Trott et al. (2023).

Cognitive Styles – or Thinking Patterns

Surface-form of Language or “What They Say”

- Simple, lexical models can pick these signals up.

I am depressed.

I am feeling great.

Cognitive Styles or “How They Think”

Dissonance

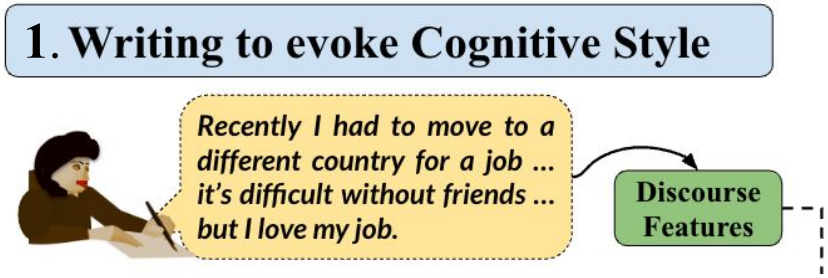
I know smoking could kill you, but *I need it for my job*!

Catastrophizing

My friends went to a party, they forgot to invite me. *They probably hate me*. :(

- Modern LLMs are able to pick this up to some extent.
- These expressions reveal cognitive processes.
- ***Deep Semantic Modeling*** can capture these complex relationships more explicitly.

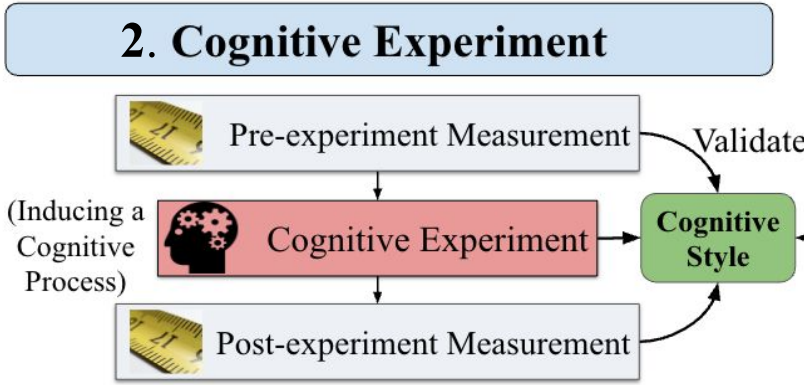
Cognitive Styles: An Experimental Validation



In this work:
Decision-making
Cognitive Style

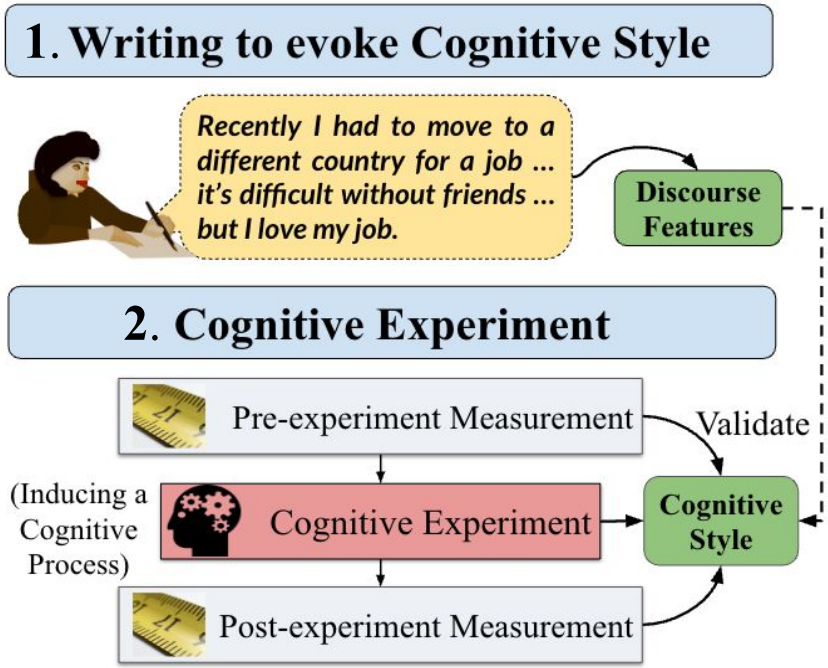
Cognitive Styles: An Experimental Validation

In this work:
Decision-making
Cognitive Style



Capturing Human Cognitive Styles with Language: Towards an Experimental Evaluation Paradigm (Varadarajan et al., NAACL 2025)

Cognitive Styles: An Experimental Validation



In this work:
Decision-making
Cognitive Style

Capturing Human Cognitive Styles with Language: Towards an Experimental Evaluation Paradigm (Varadarajan et al., NAACL 2025)

1. Writing Task: Describe a recent decision

Staying in a relationship with my partner, despite him being controlling. … becomes super passive aggresive and tells me how thats selfish … and he loves my family,... im not ready to let him go...

My mom was in a serious car accident and I was her healthcare proxy so I needed to make decisions regarding her health while in the ICU. The most difficult decision … to put her in a medically induced coma … The risks involved were definitely severe … my mom would most likely be okay ...

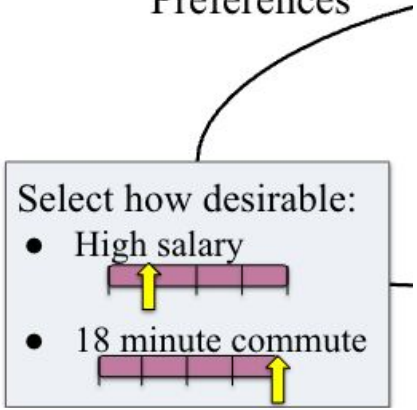
I had to spend less time with my best friends…The amount of time I spent on the mobile and desktop app Discord needed to decrease …

Capturing Human Cognitive Styles with Language: Towards an Experimental Evaluation Paradigm (Varadarajan et al., NAACL 2025)

2. Cognitive Experiment: Experimental Job Offer Scenario

Preference change

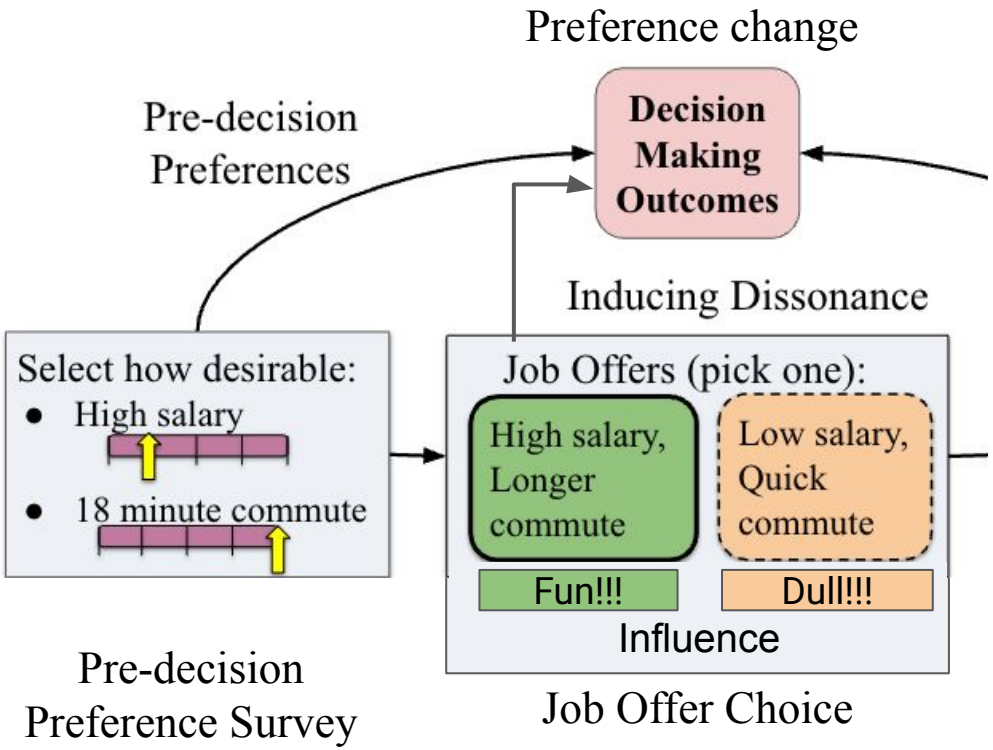
Pre-decision
Preferences



Pre-decision
Preference Survey

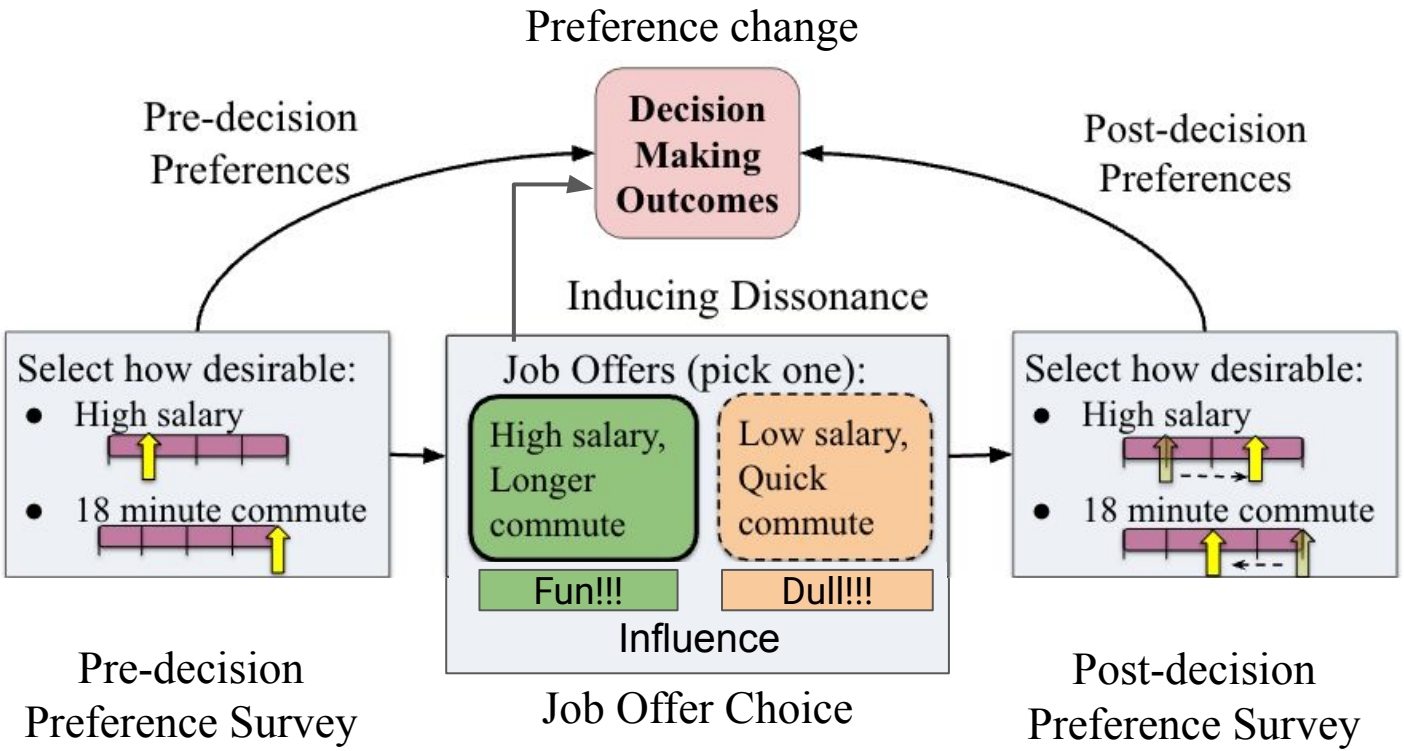
Simon, Dan, Daniel C. Krawczyk, and Keith J. Holyoak. "Construction of preferences by constraint satisfaction." *Psychological Science* 15.5 (2004): 331-336.

2. Cognitive Experiment: Experimental Job Offer Scenario



Simon, Dan, Daniel C. Krawczyk, and Keith J. Holyoak. "Construction of preferences by constraint satisfaction." *Psychological Science* 15.5 (2004): 331-336.

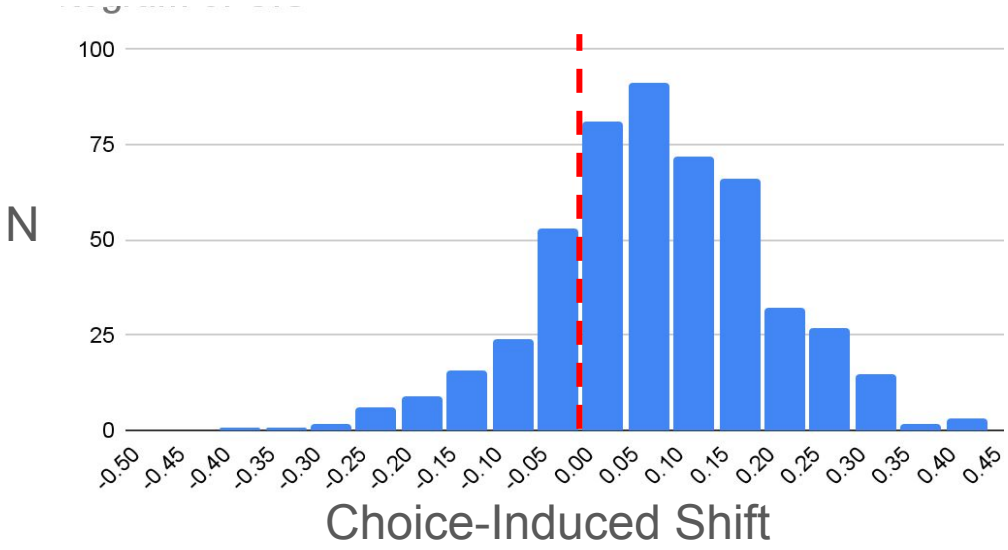
2. Cognitive Experiment: Experimental Job Offer Scenario



Simon, Dan, Daniel C. Krawczyk, and Keith J. Holyoak. "Construction of preferences by constraint satisfaction." *Psychological Science* 15.5 (2004): 331-336.

Choice-Induced Shifts (CIS) aka Preference Change

- People exhibit positive shifts
  - generally adjust preference *towards justifying their choice*



Capturing Human Cognitive Styles with Language: Towards an Experimental Evaluation Paradigm (Varadarajan et al., NAACL 2025)

Results

Baselines	AUC	Discourse feats	AUC	k
Random	0.50	Causal	**0.81**	1
Llama3.1 (0-sh)	0.56	Counterfactual	0.80	1
Gemma (0-sh)	0.56	Consonance	**0.81**	1
Llama3.1 (4-sh)	0.64	Dissonance	0.80	1
Gemma (4-sh)	**0.79**	DiscRE (full)	0.76	845
RoBERTa-L23	0.69	DiscRE (16-D)	0.79	16

➔ Discourse can predict actual changes in cognitive states.

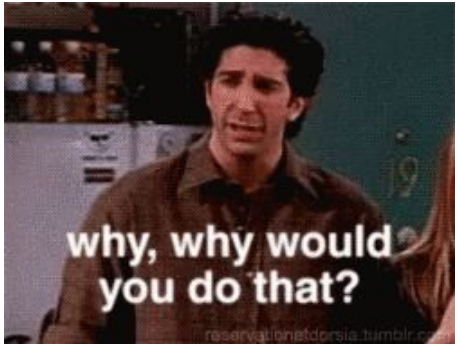
Capturing Human Cognitive Styles with Language: Towards an Experimental Evaluation Paradigm (Varadarajan et al., NAACL 2025)

Supplement

Answering Why-Questions

Matt and Sarah were pregnant.
They wanted to announce it in a fun way.
They wrote it on a cake.
They invited their friends over.
When their friends saw the cake, they were excited.

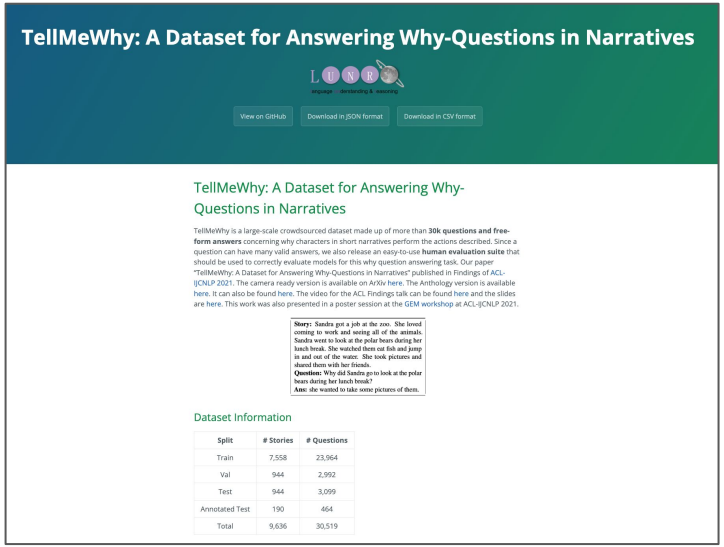
Why were Matt and Sarah pregnant?

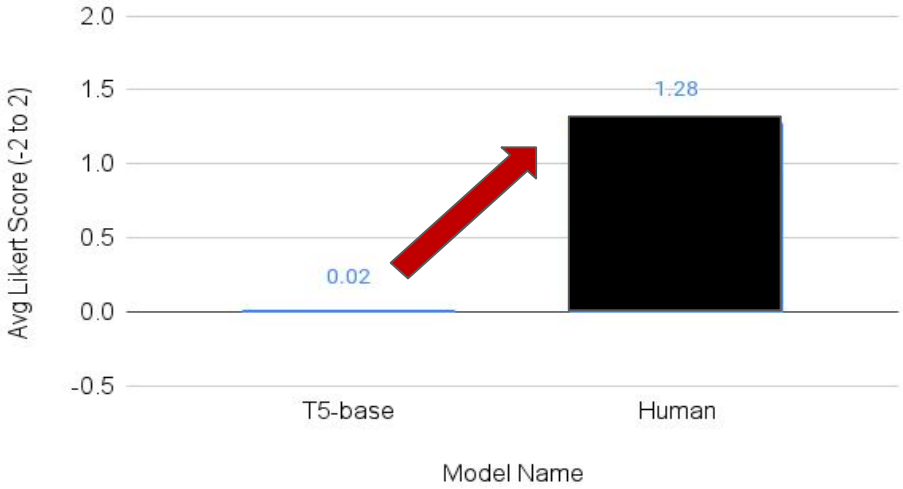


Knowing why is important for reasoning about events

Lal et al, 2021.

Answering Why-Questions





Lal et al, 2021.

Using Commonsense to Answer Why-Questions

Matt and Sarah were pregnant.
They wanted to announce it in a fun way.
They wrote it on a cake.
They invited their friends over.
When their friends saw the cake, they were excited.

Q: Why were Matt and Sarah pregnant?

Commonsense Knowledge:

- ❑ become pregnant to have babies
- ❑ can become pregnant from sexual intercourse

T5

They wanted to have a baby

Lal et al, 2022.

Using Commonsense to Answer Why-Questions

Matt and Sarah were pregnant.
They wanted to announce it in a fun way.
They wrote it on a cake.
They invited their friends over.
When their friends saw the cake, they were excited.

- Larger Model
- External Commonsense Resource

Commonsense Knowledge:
- ❑ become pregnant to have babies
- ❑ can become pregnant from sexual intercourse

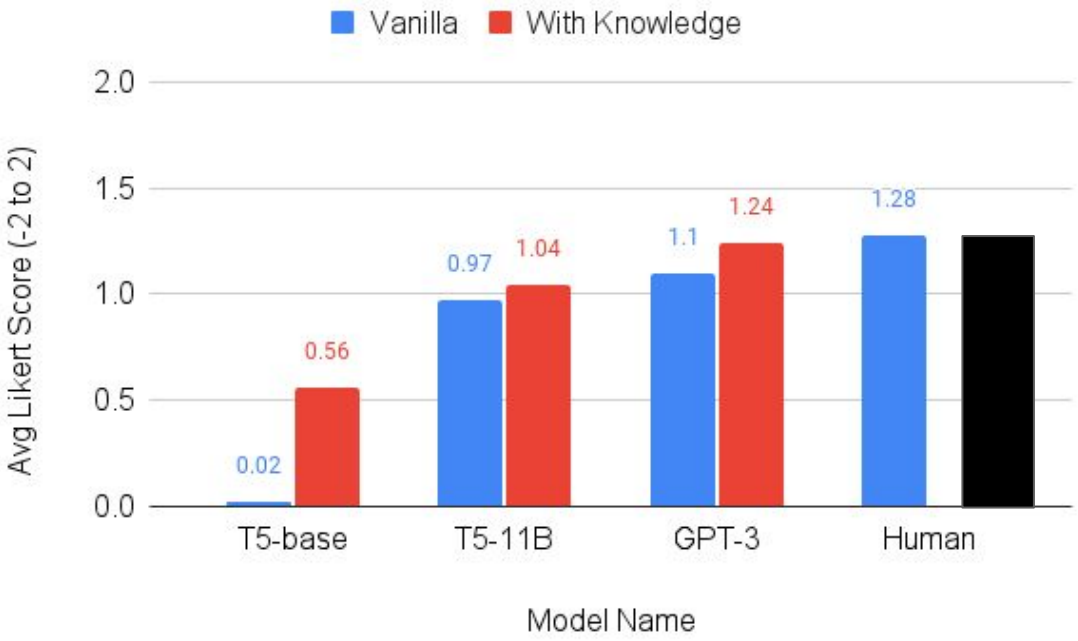
Q: Why were Matt and Sarah pregnant?

T5

They wanted to have a baby

Lal et al, 2022.

How Good are Models?



Lal et al, 2022.