

CSE 156 | Lecture 16 - Applications: Code Generation & Question Answering

Ndapa Nakashole

November 21, 2024

Reminders

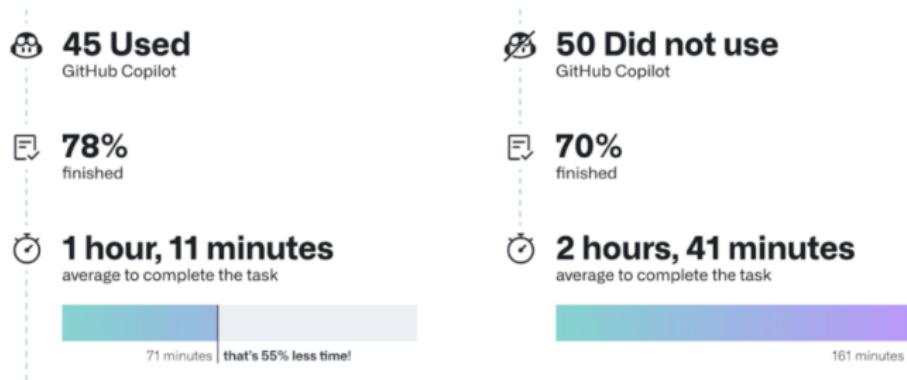
- ▶ PA2 grades out
- ▶ Quiz 3 due today, Quiz 4 will be out today

Today

- ① Application 1: Code Generation
- ② Application 2: Question Answering

Code Generation: Motivation

- ▶ As of January 30, 2024, GitHub CoPilot has over 1.3 million paying subscribers; over 50,000 organizations using enterprise version (source: Microsoft Earnings Report, 2024)



Code LLMs: Expanding Use Cases

- ▶ Originally focused on translating natural language to code
- ▶ But now can be used for many phases of the software development cycle:
 - Code completion
 - Code documentation
 - Bug detection and fixing,
 - Code translation
 - Code refactoring
 - ...

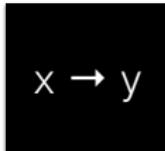
Inputs and Outputs



Natural language



Code context



Inputs and outputs



Screenshots

Code

```
class Person:
    def __init__(self, name, age):
        self.name = name
        self.age = age

    def display(self):
        print(f"Name: {self.name}, Age: {self.age}")

    def years_until_100(self):
        years_left = 100 - self.age
        if years_left <= 0:
            print(f"{self.name} is already 100 years old or more!")
        else:
            print(f"{self.name} has {years_left} years until they turn 100.")
```

Example: Github Copilot

- ▶ A code completion engine within your IDE (VSCode)

```
# load hello.csv
data = np.loadtxt('hello.csv', delimiter=',', skiprows=1)

# add together the two largest numbers in the first column
result = np.sum(np.sort(data[:, 0])[-2:])

# print how much bigger these two numbers are than the average of the first column
result = result - np.mean(data[:, 0])
```

Code LLMs

Codex GitHub Copilot

(powers/power?)

Jul 2021

Evaluating Large Language Models Trained on Code

Mark Chen^{* 1} Jerry Tworek^{* 1} Heewoo Jun^{* 1} Qiming Yuan^{* 1} Henrique Ponde de Oliveira Pinto^{* 1}
Jared Kaplan^{* 2} Harri Edwards¹ Yuri Burda¹ Nicholas Joseph² Greg Brockman¹ Alex Ray¹ Raul Puri¹
Gretchen Krueger¹ Michael Petrov¹ Heidy Khlaaf³ Girish Sastry¹ Pamela Mishkin¹ Brooke Chan¹
Scott Gray¹ Nick Ryder¹ Mikhail Pavlov¹ Alethea Power¹ Lukasz Kaiser¹ Mohammad Bavarian¹
Clemens Winter¹ Philippe Tillet¹ Felipe Petroski Such¹ Dave Cummings¹ Matthias Plappert¹
Fotios Chantzis¹ Elizabeth Barnes¹ Ariel Herbert-Voss¹ William Hebgen Guss¹ Alex Nichol¹ Alex Paino¹
Nikolas Tezak¹ Jie Tang¹ Igor Babuschkin¹ Suchir Balaji¹ Shantanu Jain¹ William Saunders¹
Christopher Hesse¹ Andrew N. Carr¹ Jan Leike¹ Josh Achiam¹ Vedant Misra¹ Evan Morikawa¹
Alec Radford¹ Matthew Knight¹ Miles Brundage¹ Mira Murati¹ Katie Mayer¹ Peter Welinder¹
Bob McGrew¹ Dario Amodei² Sam McCandlish² Ilya Sutskever¹ Wojciech Zaremba¹

Abstract

1. Introduction

Early analysis suggested GPT-3 could generate programs from Python docstrings. This was despite the fact that GPT-3 was not trained for code generation

Codex

- ▶ **Code is massively available** as training data from open source projects (e.g., over 120M public repositories on **Github**)
- ▶ Idea in OpenAI Codex: train large language model on majority code data
- ▶ Codex (v1): Same architecture as GPT-3, but with 12B parameters (vs 175B), and a different tokenization scheme to better handle whitespace

- ▶ Synthesis challenge: given a Python docstring, generate the function implementation

```
def incr_list(l: list):
    """Return list with elements incremented by 1.
    >>> incr_list([1, 2, 3])
    [2, 3, 4]
    >>> incr_list([5, 3, 5, 2, 3, 3, 9, 0, 123])
    [6, 4, 6, 3, 4, 4, 10, 1, 124]
    """
    return [i + 1 for i in l]
```

```
def solution(lst):
    """Given a non-empty list of integers, return the sum of all of the odd elements
    that are in even positions.
```

Examples

```
solution([5, 8, 7, 1]) =>12
solution([3, 3, 3, 3, 3]) =>9
solution([30, 13, 24, 321]) =>0
"""

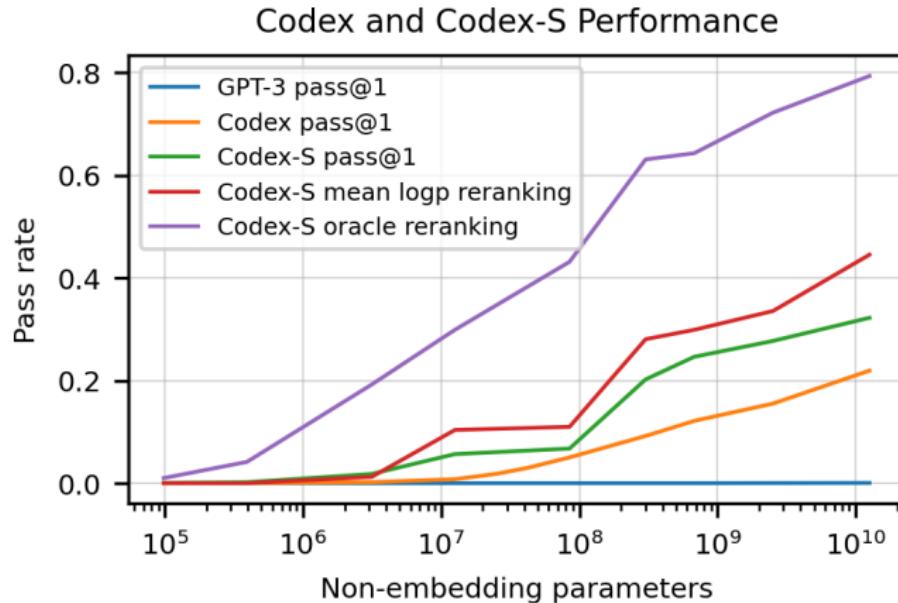
```

```
return sum(lst[i] for i in range(0, len(lst)) if i % 2 == 0 and lst[i] % 2 == 1)
```

Evaluating language models for code generation

- ▶ **Evaluation challenge:** how to ensure problems were not seen during training?
- ▶ Authors introduced **HumanEval**, a manually created dataset of 164 problems
 - Each problem has a set of hidden tests
 - A program is correct if it passes all hidden tests
 - **pass@k**: probability that, out of k samples, at least one is correct

Evaluating LMs for code generation : HumanEval



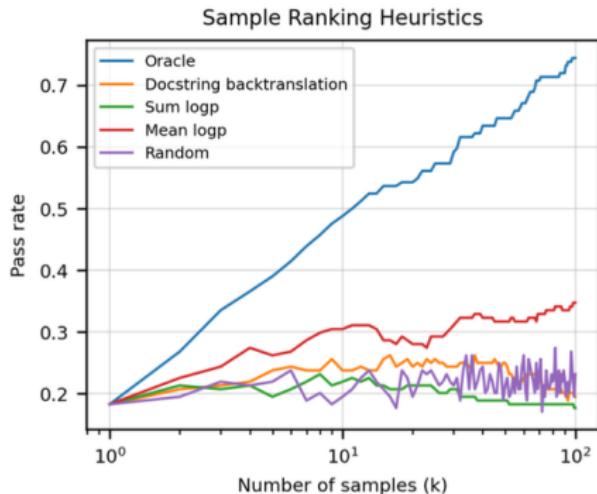
- ▶ GPT-3 fails on all problems in HumanEval
- ▶ Fine-tuning on problems with this format of function synthesis (Codex-S) improves performance
- ▶ Sampling 100 programs, reranking and choosing best improves performance

Ranking

For end-users, you don't want to present 100 choices

Sample a large number of programs, but re-rank and only show top k for small k

- ▶ **Oracle reranker:** run on all hidden tests and return the program that passes all, if any
- ▶ **Model's log-probabilities:** use models log probabilities to re-rank programs
- ▶ **Back-translation:** use P (ground truth docstring|generated sample)



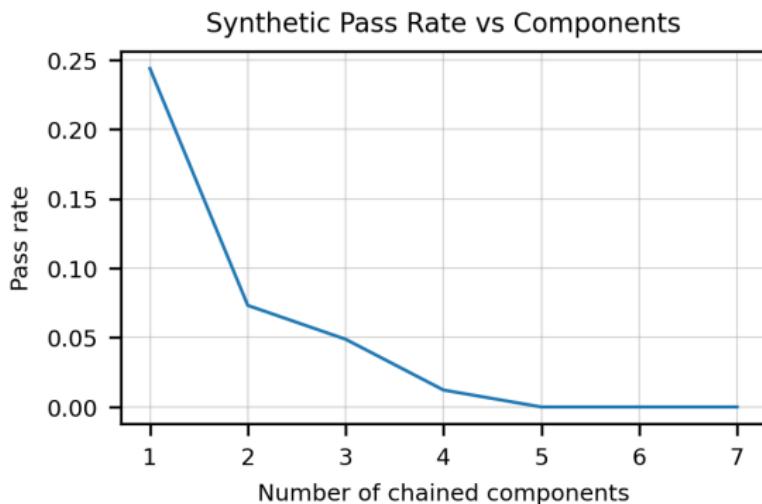
From code to natural language

Code documentation

- ▶ The first experiments sampled from $P(\text{code} \mid \text{docstring})$.
What if we sample $P(\text{docstring} \mid \text{code})$?
- ▶ Less frequent in training data since docstring comes before
- ▶ Can synthetically create a dataset by inverting this order and fine-tune Codex to obtain Codex-D
- ▶ Pass@k estimated by human evaluation
- ▶ Lower performance (35%) than in the other direction!

A note on compositionality

- ▶ If a human can trivially solve problem X (e.g., reverse a string), and also problem Y (e.g., compute string length), the problem “do X then Y ” is still trivial
- ▶ Result: performance on synthetic tasks made by chaining simple components decays exponentially as components increase,



RESEARCH

COMPUTER SCIENCE

Competition-level code generation with AlphaCode

Yujia Li^{*†}, David Choi^{*†}, Junyoung Chung[†], Nate Kushman[†], Julian Schrittwieser[†], Rémi Leblond[†], Tom Eccles[†], James Keeling[†], Felix Gimeno[†], Agustin Dal Lago[†], Thomas Hubert[†], Peter Choy[†], Cyprien de Masson d'Autume[†], Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, Oriol Vinyals^{*}

- ▶ In 2022, DeepMind published AlphaCode, a system combining & expanding these ideas to solve competitive programming problems

AlphaCode: example problem

Backspace

You are given two strings s and t , both consisting of lowercase English letters. You are going to type the string s character by character, from the first character to the last one.

When typing a character, instead of pressing the button corresponding to it, you can press the “Backspace” button. It deletes the last character you have typed among those that aren’t deleted yet (or does nothing if there are no characters in the current string). For example, if s is “`abcdbd`” and you press Backspace instead of typing the first and the fourth characters, you will get the string “`bd`” (the first press of Backspace deletes no character, and the second press deletes the character ‘c’). Another example, if s is “`abcaa`” and you press Backspace instead of the last two letters, then the resulting text is “`a`”.

Your task is to determine whether you can obtain the string t , if you type the string s and press “Backspace” instead of typing several (maybe zero) characters of s .

Input

The first line contains a single integer q ($1 \leq q \leq 10^5$) the number of test cases. The first line of each test case contains the string s ($1 \leq |s| \leq 10^5$). Each character of s is a lowercase English letter.

The second line of each test case contains the string t ($1 \leq |t| \leq 10^5$). Each character of t is a lowercase English letter.

It is guaranteed that the total number of characters in the strings over all test cases does not exceed $2 \cdot 10^5$.

Output

For each test case, print “YES” if you can obtain the string t by typing the string s and replacing some characters with presses of “Backspace” button, or “NO” if you cannot.

You may print each letter in any case (YES, yes, Yes will all be recognized as positive answer, NO, no and nO will all be recognized as negative answer).

AlphaCode: example problem input-output pairs

Example Input

4
ababa
ba
ababa
bb
aaa
aaaa
aababa
ababa

Example Output

YES
NO
NO
YES

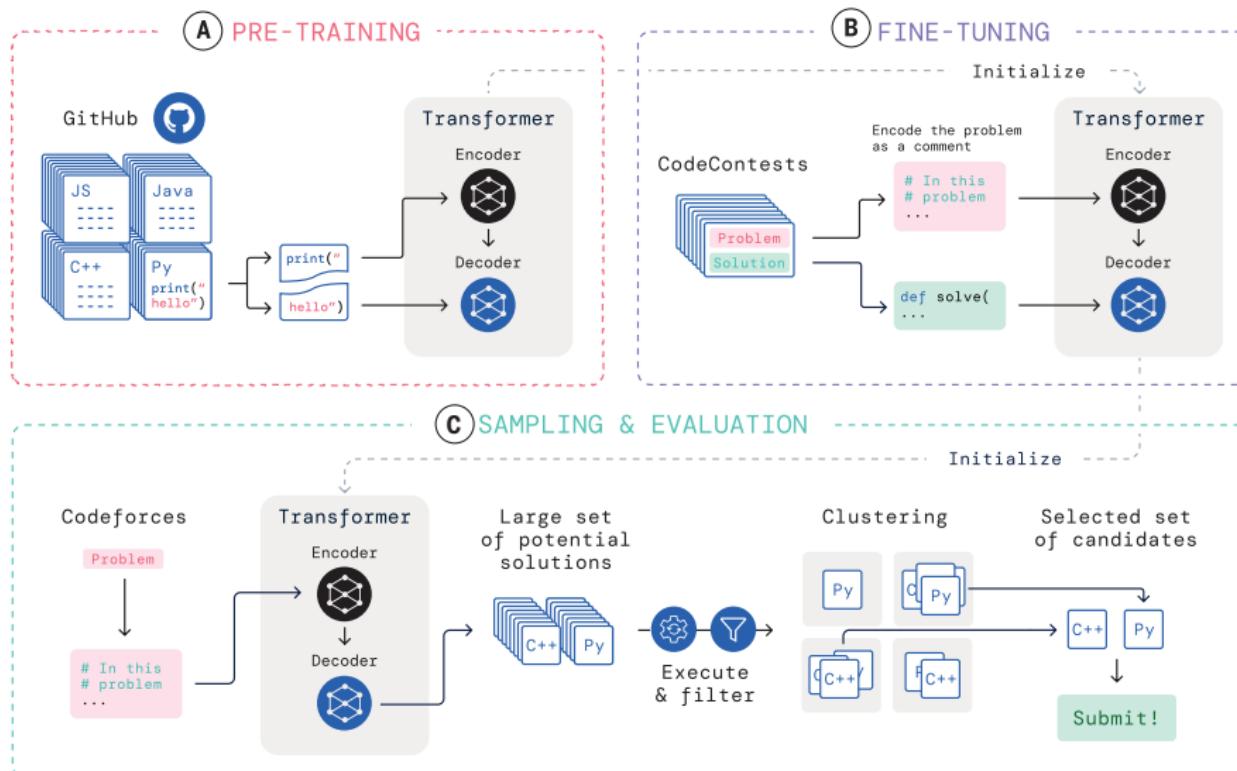
AlphaCode: example problem solution

```
1 t=int(input())
2 for i in range(t):
3     s=input()
4     t=input()
5     a=[]
6     b=[]
7     for j in s:
8         a.append(j)
9     for j in t:
10        b.append(j)
11    a.reverse()
12    b.reverse()
13    c=[]
14    while len(b)!=0 and len(a)!=0:
15        if a[0]==b[0]:
16            c.append(b.pop(0))
17            a.pop(0)
18        elif a[0]!=b[0] and len(a)!=1:
19            a.pop(0)
20            a.pop(0)
21        elif a[0]!=b[0] and len(a)==1:
22            a.pop(0)
23    if len(b)==0:
24        print("YES")
25    else:
26        print("NO")
```

AlphaCode: Architecture

- ▶ Unlike Codex, decoder only, AlphaCode used an encoder-decoder Transformer
- ▶ To speed things up, they used **multi-query attention** instead of full multi-head attention blocks
 - Several query heads but single set of key-value pairs

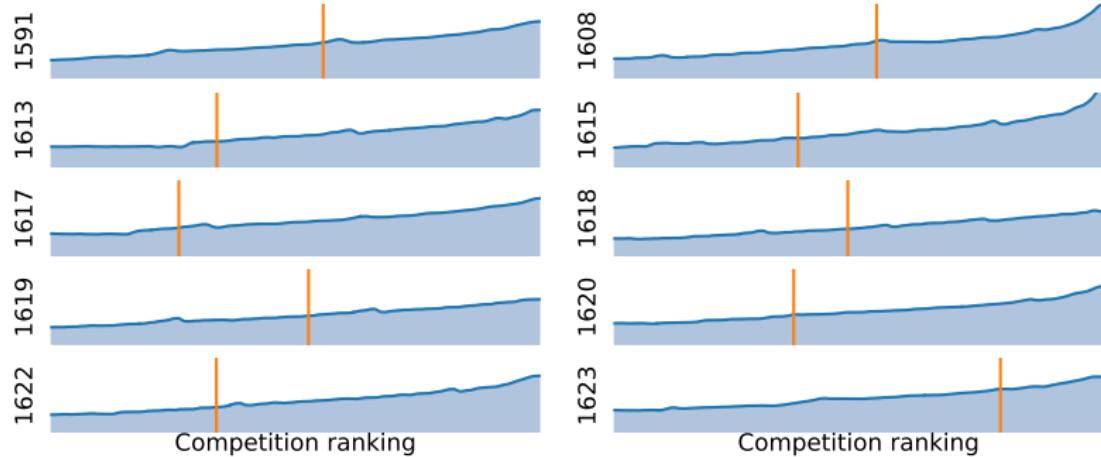
AlphaCode: Pipeline



AlphaCode: Pipeline

- ▶ **Pretraining:** Similar data to Codex, open-source code from GitHub
- ▶ **Fine-tuning:** programming competition data, human solutions to 13k competitive programming problems
- ▶ **Sampling:** Up to 100k samples per problem (!)
- ▶ **Filtering:** discard samples that fail public tests
- ▶ **Clustering:** trained a separate model to generate inputs; clustered generated solution by behavior on those inputs; submitted one example from **10 largest clusters**

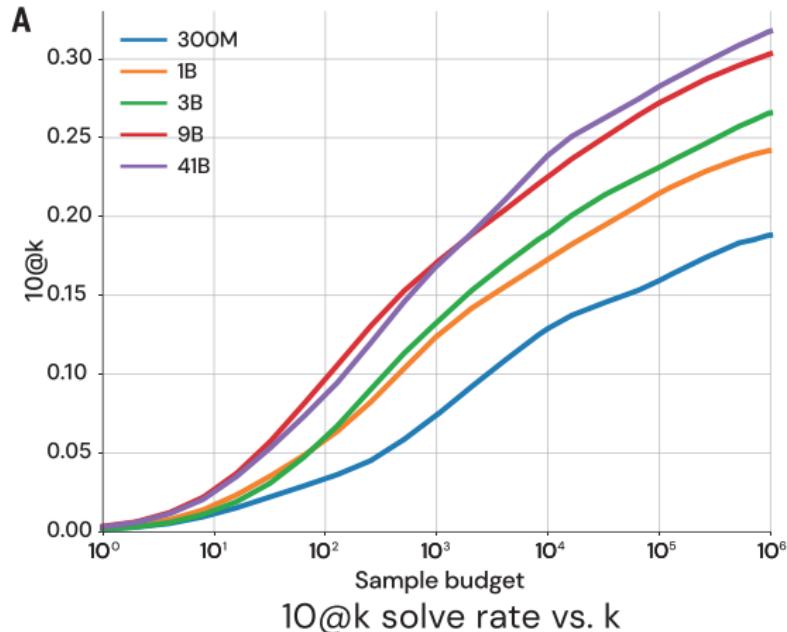
AlphaCode: Ranking among contest participants



(a) AlphaCode's ranking in 10 contests

"Approximately corresponds to a novice programmer with a few months to a year of training"

AlphaCode: Sampling more is good



(10x samples approx. +6% solve rate)

AlphaCode: takeaways

- ▶ Sampling more is the largest contributor. Most of their methods don't improve things at 1k samples

Fine-tuning setting	10@1K	10@10K	Solve rate 10@100K
No Enhancements	6.7% (6.5-6.8)	10.4% (9.6-11.0)	15.2% (14.3-15.9)
+ MLM	6.6% (6.2-7.0)	12.5% (12.1-12.7)	17.0% (16.5-17.2)
+ Tempering	7.7% (7.2-8.5)	13.3% (12.5-13.8)	18.7% (18.0-19.2)
+ Tags and Ratings	6.8% (6.4-7.0)	13.7% (12.8-14.9)	19.3% (18.1-20.0)
+ Value	10.6% (9.8-11.1)	16.6% (16.4-16.9)	20.2% (19.6-20.7)
+ GOLD	12.4% (12.0-13.0)	17.3% (16.9-17.6)	21.5% (20.5-22.2)
+ Clustering	12.2% (10.8-13.4)	18.0% (17.3-18.8)	24.1% (23.2-25.0)

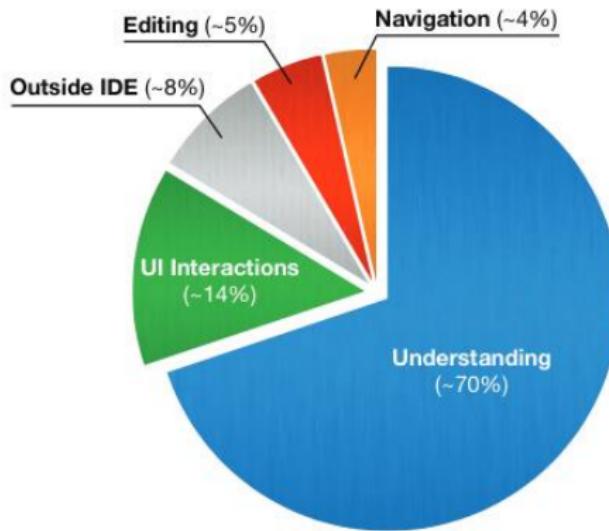
Other Code LLMs

- ▶ **StarCoder 2** (Lozhkov et al. 2024)
- ▶ **DeepSeek Coder** (Guo et al. 2024)
- ▶ **CodeLLaMA** (Roziere et al. 2023)
- ▶ **Codex** (Chen et al. 2022)

Which one to use?

- ▶ Somewhat similar performance. See comparisons in StarCoder paper (Lozhkov et al. 2024)

How much of a programmer's job can current language models automate?



[Minelli et al, 2015]

- ▶ Relatively very little!
- ▶ Much time is spent on deciding & discussing what to build rather than actually building it

Pretraining LMs on Code language models: takeaways

- ▶ Applying language models to code has led to exciting developments in tools for programmers.
- ▶ **Sampling**, testing and filtering can get quite far (similar ideas in other generation settings e.g. DALLE)
- ▶ It's now standard to include some code in LMs trained primarily on natural language text. Some attribute the models' apparent reasoning capabilities to this (Madaan et al., 2022)
- ▶ Still, many of these experiments assume a setting that fundamentally differs from realworld programming
 - Well-defined, self-contained, short problems; extensive existing correctness & performance tests; only need standard libraries;
 - ...

Limitations and Discussion

- ▶ Lots of public code to pre-train, but does not cover everything:
 - New or in-house libraries
 - New programming languages
 - New language features
- ▶ Public code repositories have lots of code with bugs
- ▶ Generated code often has functional or security bugs. Still need to understand it!
- ▶ Perry & Srivastava et al, 2022 ran a user study where participants solved programming tasks with and without Codex
 - "Overall, we find that participants who had access to an AI assistant based on OpenAI's codex-davinci-002 model wrote significantly less secure code than those without access"

Application 2: Question Answering

CSE 156, Fall 2024

Lecture 16b: Question Answering

Ndapa Nakashole, UCSD
November 21 2024





Question Answering

- “Question answering” as a term is so broad as to be meaningless
- Facts, math, MT, common sense, ...
 - What states border Mississippi?
 - When was the current US president born?
 - How is Advil different from Tylenol?
 - What is $4+5$?
 - What is the translation of [sentence] into French?
 - Is it okay to use a blender in 2 AM in an apartment?



Why do we Study Answering?

- As a testbed to evaluate how machines **understand** text

THE PROCESS OF QUESTION ANSWERING

May 1977

Research Report #88

Wendy Lehnert

When a person understands a story, he can demonstrate his understanding by answering questions about the story. Since questions can be devised to query any aspect of text comprehension, the ability to answer questions is the strongest possible demonstration of understanding. Question answering is therefore a task criterion for evaluating reading skills.

If a computer is said to understand a story, we must demand of the computer the same demonstrations of understanding that we require of people. Until such demands are met, we have no way of evaluating text understanding programs. Any computer programmer can write a program which inputs text. If the programmer assures us that his program 'understands' text, it is a bit like being reassured by a used car salesman about a suspiciously low speedometer reading. Only when we can ask a program to answer questions about what it reads will we be able to begin to assess that program's comprehension.

“Since questions can be devised to query **any aspect** of text comprehension, the ability to answer questions is the **strongest possible demonstration of understanding**.”



Many Practical Applications

Google

Where is the deepest lake in the world?



All

Maps

Images

News

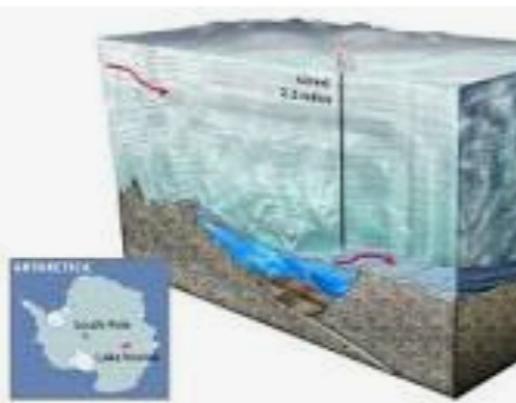
Videos

More

Settings

Tools

About 21,100,000 results (0.71 seconds)



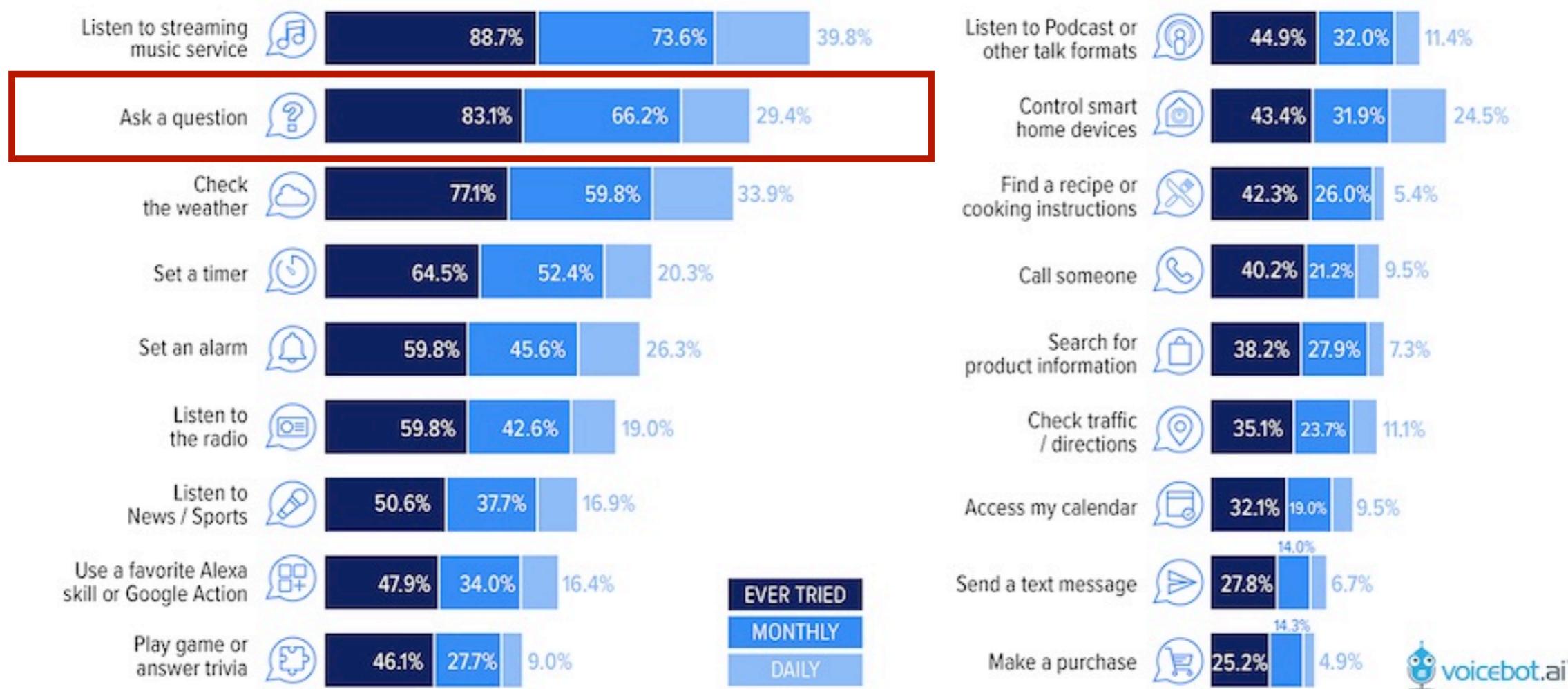
Siberia

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.



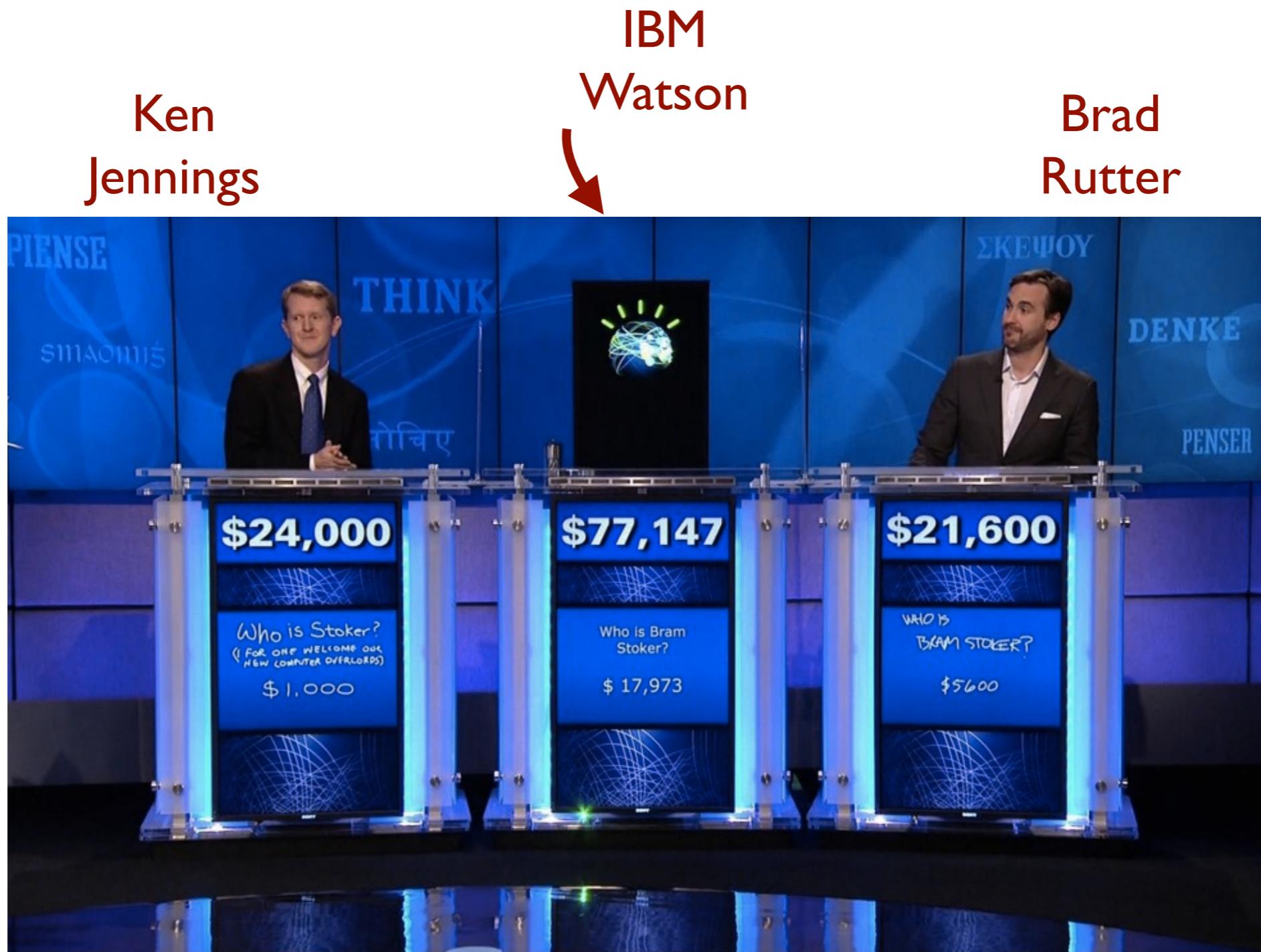
Many Practical Applications

Smart Speaker Use Case Frequency January 2020





IBM Watson defeated two of Jeopardy's human champions in 2011

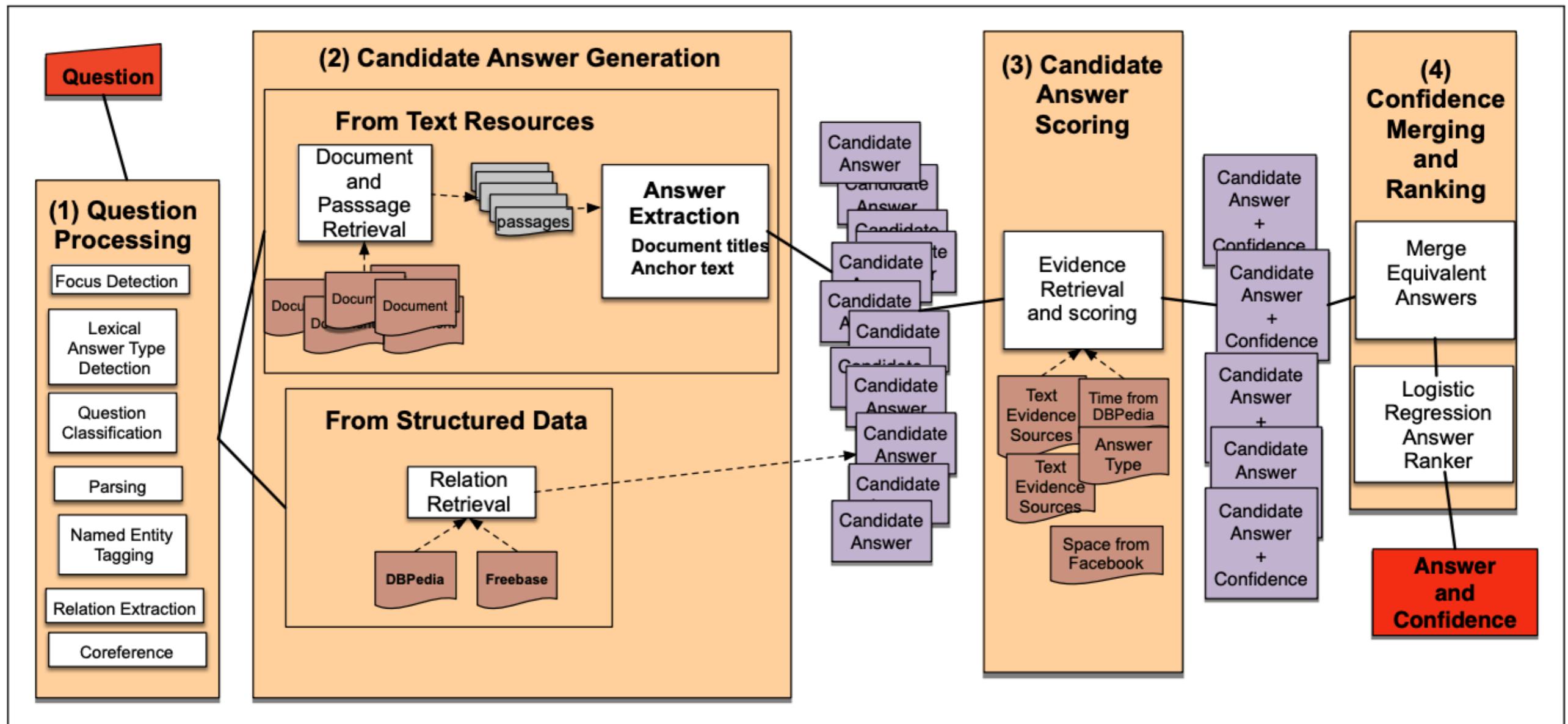


D. Ferrucci et al.: Building Watson. AI Magazine, Fall 2010.

IBM Journal of R&D 56(3/4), 2012: This is Watson



IBM Watson QA



(1) Question processing, (2) Candidate answer generation, (3) Candidate answer scoring, and (4) Confidence merging and ranking.



Conversational Question Answering

The Virginia governor's race, billed as the marquee battle of an otherwise anticlimactic 2013 election cycle, is shaping up to be a foregone conclusion. Democrat Terry McAuliffe, the longtime political fixer and moneymen, hasn't trailed in a poll since May. Barring a political miracle, Republican Ken Cuccinelli will be delivering a concession speech on Tuesday evening in Richmond. In recent ...

Q: What are the candidates **running** for?

A: Governor

Q: **Where?**

A: Virginia

Q: Who is the democratic candidate?

A: Terry McAuliffe

Q: Who is **his** opponent?

A: Ken Cuccinelli

Q: What party does **he** belong to?

A: Republican

Q: Which of **them** is winning?



Table-based Question Answering

Year	City	Country	Nations
1896	Athens	Greece	14
1900	Paris	France	24
1904	St. Louis	USA	12
...
2004	Athens	Greece	201
2008	Beijing	China	204
2012	London	UK	204

x = Greece held its last Summer Olympics in which year?

y = 2004



Visual Question Answering



What color are her eyes?

What is the mustache made of?

(Antol et al, 2015): Visual Question Answering



How many slices of pizza are there?
Is this a vegetarian pizza?



how many slices of pizza are there? is this a vegetarian pizza?

ChatGPT

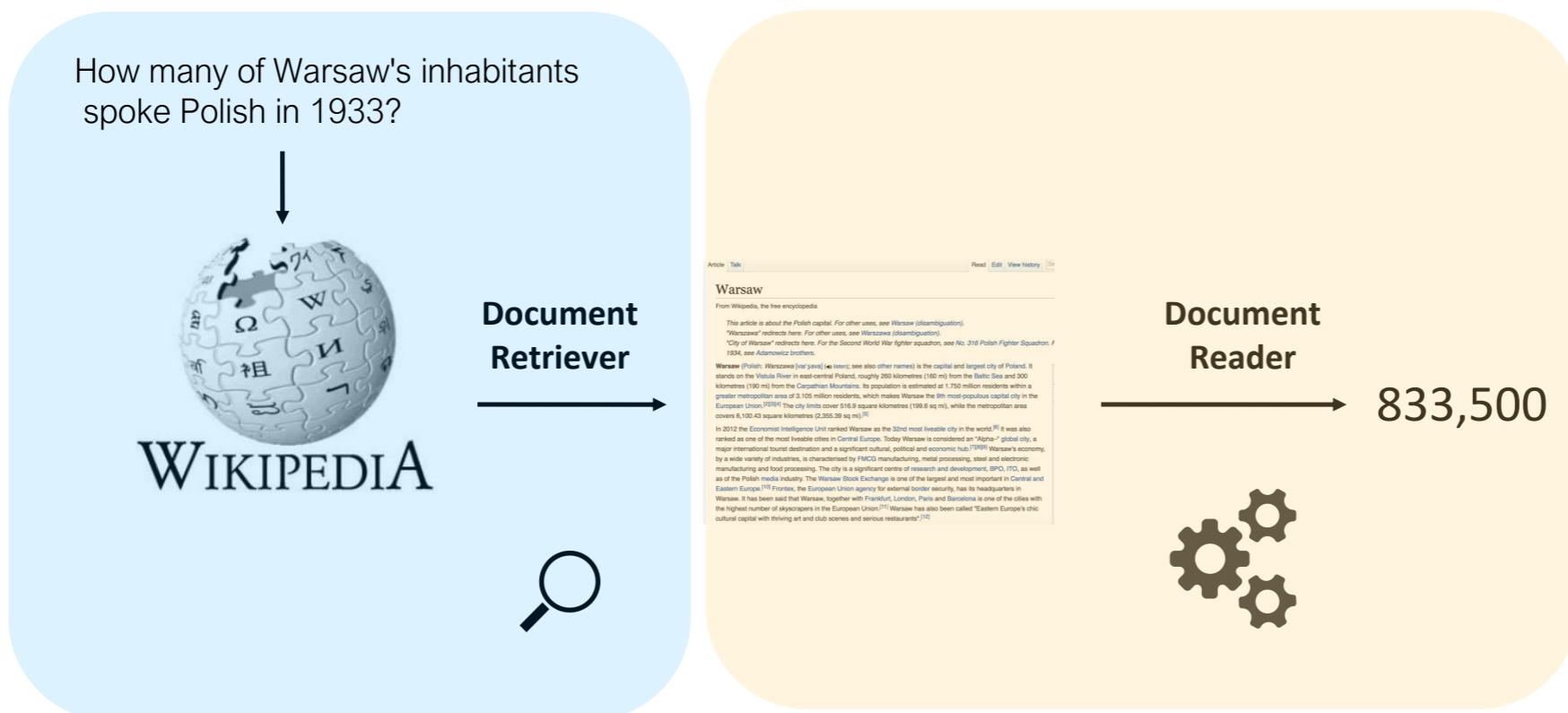
There are 8 slices of pizza in the image. As for whether it's a vegetarian pizza, it appears to have toppings that could be vegetables, such as pieces that look like diced tomatoes and possibly green peppers. However, without a clearer image or a list of ingredients, it's impossible to be certain if there are no meat products on the pizza. Vegetarian pizzas typically do not include meat, but they can include a variety of vegetable toppings like onions, peppers, mushrooms, olives, and more. If there is any meat or meat-based toppings visible, then it would not be a vegetarian pizza.

GPT-4
As of Feb 15th, 2024



Retriever-reader framework for QA

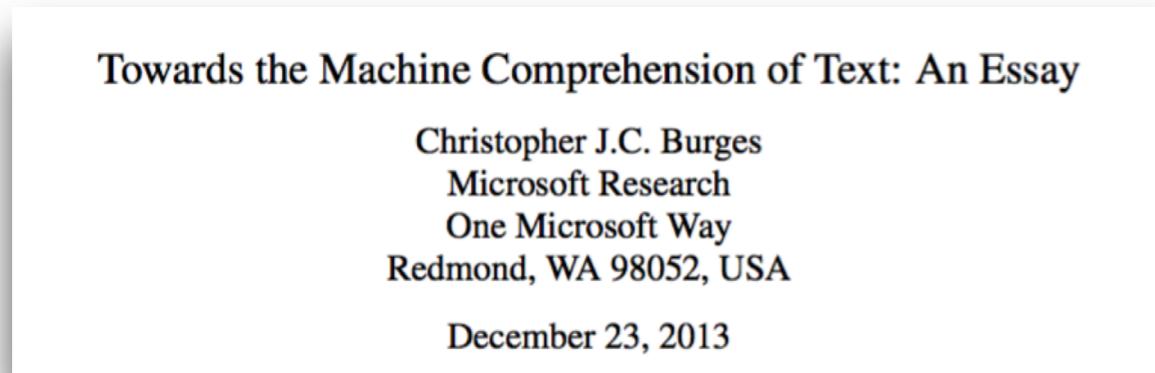
1. Finding documents that (might) contain an answer
 - Which can be handled by traditional information retrieval/web search
2. Finding an answer in a paragraph or a document
 - This problem is often termed **Reading Comprehension**





Machine Comprehension (Burges 2013)

- “A machine **comprehends** a passage of **text** if, for any **question** regarding that text that can be **answered** correctly by a majority of native speakers, that machine can provide a string which those speakers would agree both answers that question, and does not contain information irrelevant to that question.”





Reader: Reading Comprehension in the Deep Learning Era

- Floodgates opened in 2015/16 with the production of large datasets which permit supervised neural systems to be built
 - Hermann et al. (NIPS 2015) DeepMind CNN/DM dataset
 - Rajpurkar et al. (EMNLP 2016) SQuAD
 - MS MARCO, TriviaQA, RACE, NewsQA, NarrativeQA, ...

CNN/Daily Mail Datasets

CNN Entertainment » 'Star Wars' universe gets its first gay character Hide Caption



Story highlights

Official "Star Wars" universe gets its first gay character, a lesbian governor

The character appears in the upcoming novel "Lords of the Sith"

Characters in "Star Wars" movies have gradually become more diverse

(CNN) — If you feel a ripple in the Force today, it may be the news that the official Star Wars universe is getting its first gay character.

According to the sci-fi website Big Shiny Robot, the upcoming novel "Lords of the Sith" will feature a capable but flawed Imperial official named Moff Mors who "also happens to be a lesbian."

The character is the first gay figure in the official Star Wars universe -- the movies, television shows, comics and books approved by Star Wars franchise owner Disney -- according to Shelly Shapiro, editor of "Star Wars" books at Random House imprint Del Rey Books.

CNN/Daily Mail Datasets

CNN Entertainment » 'Star Wars' universe gets its first gay character Hide caption > <



Story highlights

Official "Star Wars" universe gets its first gay character, a lesbian governor

The character appears in the upcoming novel "Lords of the Sith"

Characters in [REDACTED] movies have gradually become more diverse

(CNN) — If you feel a ripple in the Force today, it may be the news that the official Star Wars universe is getting its first gay character.

According to the sci-fi website Big Shiny Robot, the upcoming novel "Lords of the Sith" will feature a capable but flawed Imperial official named Moff Mors who "also happens to be a lesbian."

The character is the first gay figure in the official Star Wars universe -- the movies, television shows, comics and books approved by Star Wars franchise owner Disney -- according to Shelly Shapiro, editor of "Star Wars" books at Random House imprint Del Rey Books.



SQuAD: Stanford Question Answering Dataset

SQuAD: 100,000+ Questions for Machine Comprehension of Text

Pranav Rajpurkar and **Jian Zhang** and **Konstantin Lopyrev** and **Percy Liang**

{pranavsr, zjian, klopyrev, pliang}@cs.stanford.edu

Computer Science Department
Stanford University

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

Dataset	Question source	Formulation	Size
SQuAD	crowdsourced	RC, spans in passage	100K
MCTest (Richardson et al., 2013)	crowdsourced	RC, multiple choice	2640
Algebra (Kushman et al., 2014)	standardized tests	computation	514
Science (Clark and Etzioni, 2016)	standardized tests	reasoning, multiple choice	855
WikiQA (Yang et al., 2015)	query logs	IR, sentence selection	3047
TREC-QA (Voorhees and Tice, 2000)	query logs + editor	IR, free form	1479
CNN/Daily Mail (Hermann et al., 2015)	summary cloze	+ RC, fill in single entity	1.4M
CBT (Hill et al., 2015)	cloze	RC, fill in single word	688K

Reasoning	Description	Example	Percentage
Lexical variation (synonymy)	Major correspondences between the question and the answer sentence are synonyms.	Q: What is the Rankine cycle sometimes called ? Sen.: The Rankine cycle is sometimes referred to as a practical Carnot cycle .	33.3%
Lexical variation (world knowledge)	Major correspondences between a question and the answer sentence require world knowledge to resolve.	Q: Which governing bodies have veto power? Sen.: The European Parliament and the Council of the European Union have powers of amendment and veto during the legislative process.	9.1%
Syntactic variation	After a question is paraphrased into declarative form, its syntactic dependency structure does not match that of the answer sentence after light pruning.	Q: What Shakespeare scholar is currently on the faculty ? Sen.: Current faculty include the anthropologist Marshall Sahlins, ..., Shakespeare scholar David Bevington.	64.1%
Multiple sentence reasoning	There is a coreference or answering requires higher-level fusion of multiple sentences.	Q: What collection does the V&A Theatre & Performance galleries hold? Sen.: The V&A Theatre & Performance galleries opened in March 2009. ... They hold the UK's biggest national collection of material about live performance.	13.6%
Ambiguous	We don't agree with the crowd-workers' answer or the question does not have a unique answer.	Q: What is the main goal of criminal punishment? Sen.: Achieving crime control via incapacitation and deterrence is a major goal of criminal punishment.	6.1%



SQuAD Limitations

SQuAD has a number of key limitations:

- Only span-based answers (no yes/no, counting, implicit why)
- Questions were constructed looking at the passages
 - Not genuine information needs
 - Generally greater lexical and syntactic matching between questions and answer span than you get IRL
- Barely any multi-fact/sentence inference beyond coreference

Nevertheless, it is a well-targeted, well-structured, clean dataset

- It has been the most used and competed on QA dataset
- It has also been a useful starting point for building systems in industry (though in-domain data always really helps!)



Other question answering datasets

TriviaQA: Questions and answers by trivia enthusiasts. Independently collected web paragraphs that contain the answer and seem to discuss question, but no human verification that paragraph supports answer to question

Distant supervision

Natural Questions: Question drawn from frequently asked Google search questions. Answers from Wikipedia paragraphs. Answer can be substring, yes, no, or NOT_PRESENT. Verified by human annotation.

HotpotQA. Constructed questions to be answered from the whole of Wikipedia which involve getting information from two pages to answer a multistep query:
Q: Which novel by the author of “Armada” will be adapted as a feature film by Steven Spielberg? A: *Ready Player One*



Neural Models for Reading Comprehension

- How can we build a model to solve SQuAD?



Neural Models for Reading Comprehension

Problem formulation

- Input: $C = (c_1, c_2, \dots, c_N), Q = (q_1, q_2, \dots, q_M), c_i, q_i \in V$ $N \sim 100, M \sim 15$
- Output: $1 \leq \text{start} \leq \text{end} \leq N$ answer is a span

A family of LSTM-based models with attention (2016-2018)

Attentive Reader (Hermann et al., 2015), Stanford Attentive Reader (Chen et al., 2016), Match-LSTM (Wang et al., 2017), BiDFA (Seo et al., 2017), Dynamic coattention network (Xiong et al., 2017), DrQA (Chen et al., 2017), R-Net (Wang et al., 2017), ReasoNet (Shen et al., 2017)..

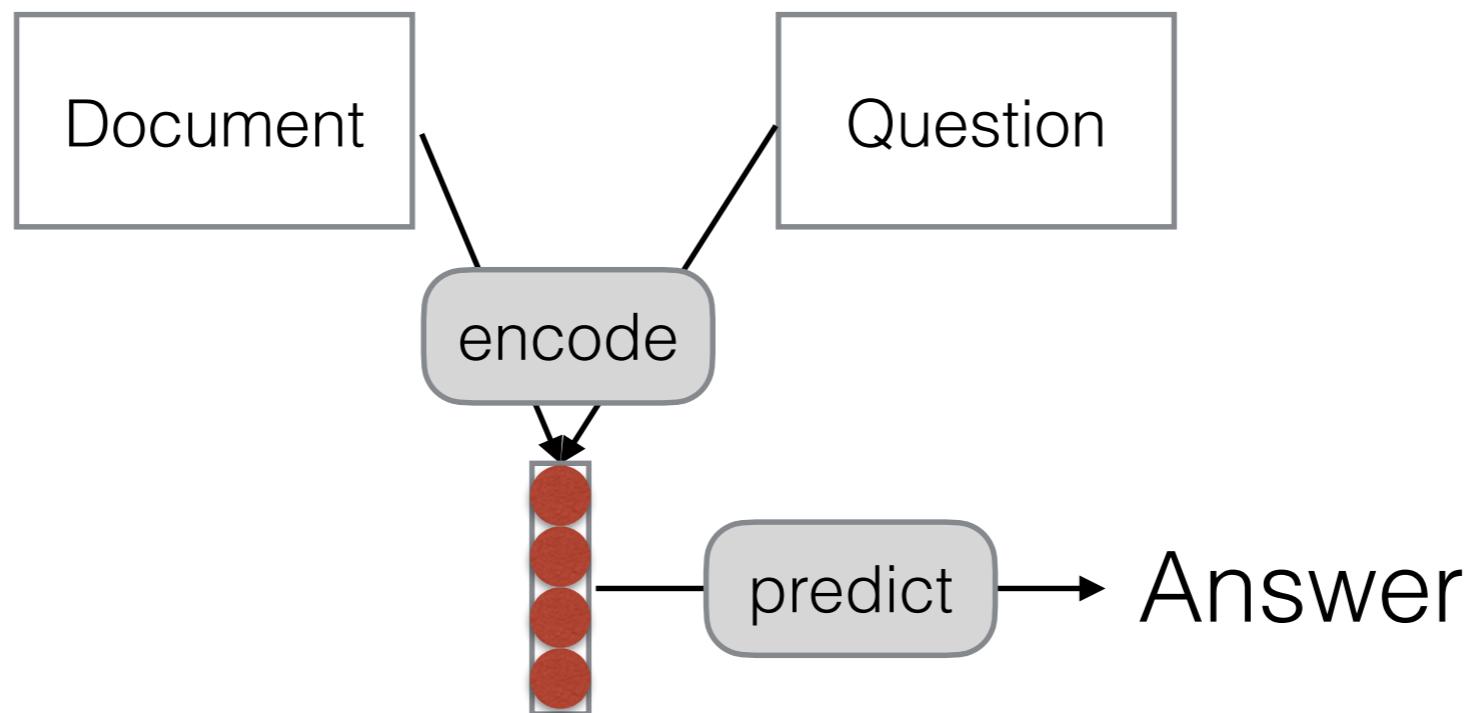
Fine-tuning BERT-like models for reading comprehension (2019+)

What approach is missing here?



A Basic Neural QA Model

- Encode the **document** and **question**; then generate an answer (sentence or single word)



- Problem: encoding whole documents with high accuracy and coverage is hard!

BI-DIRECTIONAL ATTENTION FLOW FOR MACHINE COMPREHENSION

Minjoon Seo^{1*} **Aniruddha Kembhavi²** **Ali Farhadi^{1,2}** **Hananneh Hajishirzi¹**

University of Washington¹, Allen Institute for Artificial Intelligence²

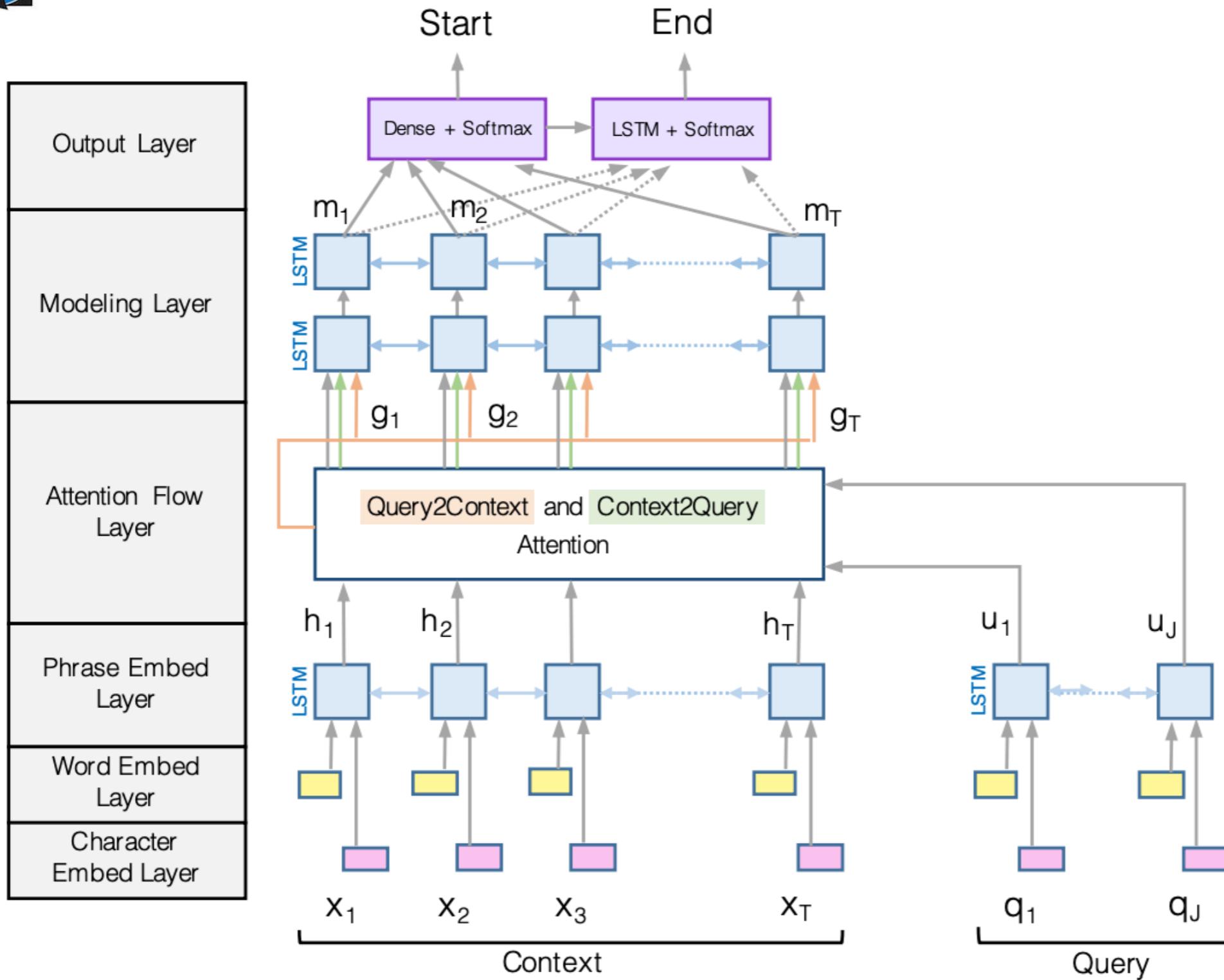
{minjoon, ali, hannaneh}@cs.washington.edu, {anik}@allenai.org

ABSTRACT

Machine comprehension (MC), answering a query about a given context paragraph, requires modeling complex interactions between the context and the query. Recently, attention mechanisms have been successfully extended to MC. Typically these methods use attention to focus on a small portion of the context and summarize it with a fixed-size vector, couple attentions temporally, and/or often form a uni-directional attention. In this paper we introduce the Bi-Directional Attention Flow (BiDAF) network, a multi-stage hierarchical process that represents the context at different levels of granularity and uses bi-directional attention flow mechanism to obtain a query-aware context representation without early summarization. Our experimental evaluations show that our model achieves the state-of-the-art results in Stanford Question Answering Dataset (SQuAD) and CNN/DailyMail cloze test.

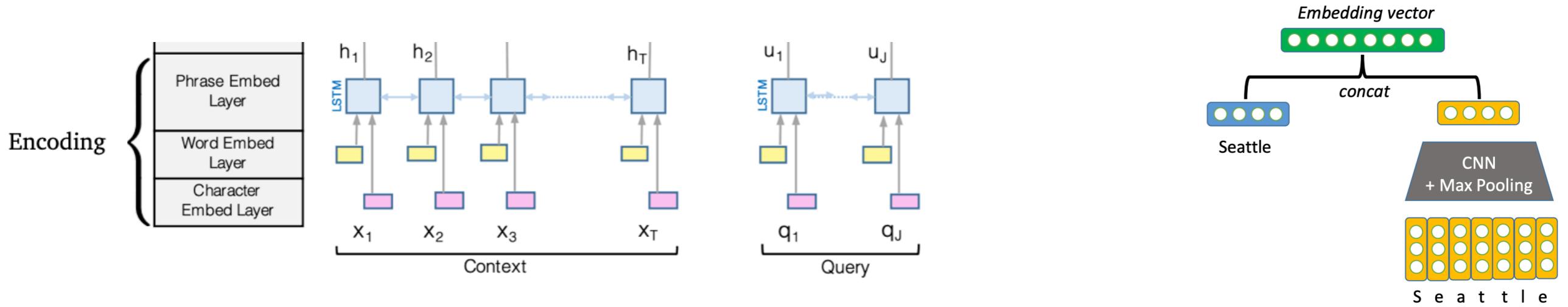


BiDAF: the Bidirectional Attention Flow model





BiDAF Embedding Layers



- Use a concatenation of word embedding (GloVe) and character embedding (CNNs over character embeddings) for each word in context and query.

$$e(c_i) = f([\text{GloVe}(c_i); \text{charEmb}(c_i)])$$

$$e(q_i) = f([\text{GloVe}(q_i); \text{charEmb}(q_i)])$$

- Then, use two **bidirectional** LSTMs separately to produce contextual embeddings for both context and query.

$$\overrightarrow{\mathbf{c}}_i = \text{LSTM}(\overrightarrow{\mathbf{c}}_{i-1}, e(c_i)) \in \mathbb{R}^H$$

$$\overleftarrow{\mathbf{c}}_i = \text{LSTM}(\overleftarrow{\mathbf{c}}_{i+1}, e(c_i)) \in \mathbb{R}^H$$

$$\mathbf{c}_i = [\overrightarrow{\mathbf{c}}_i; \overleftarrow{\mathbf{c}}_i] \in \mathbb{R}^{2H}$$

$$\overrightarrow{\mathbf{q}}_i = \text{LSTM}(\overrightarrow{\mathbf{q}}_{i-1}, e(q_i)) \in \mathbb{R}^H$$

$$\overleftarrow{\mathbf{q}}_i = \text{LSTM}(\overleftarrow{\mathbf{q}}_{i+1}, e(q_i)) \in \mathbb{R}^H$$

$$\mathbf{q}_i = [\overrightarrow{\mathbf{q}}_i; \overleftarrow{\mathbf{q}}_i] \in \mathbb{R}^{2H}$$



Key idea

- There are variants of and improvements to the BiDAF architecture over the years, but the central idea is **the Attention Flow layer**
- **Attention Flow Idea:** attention should flow both ways – from the context to the question and from the question to the context



BiDAF: Attention Flow Layer (Part 0)

- Make similarity matrix

$$S_{ij} = \mathbf{w}_{\text{sim}}^T [\mathbf{c}_i; \mathbf{q}_j; \mathbf{c}_i \circ \mathbf{q}_j] \in \mathbb{R}$$



BiDAF: Attention Flow Layer (Part I)

- Context-to-Question (C2Q) attention:
(which query words are most relevant to each context word)

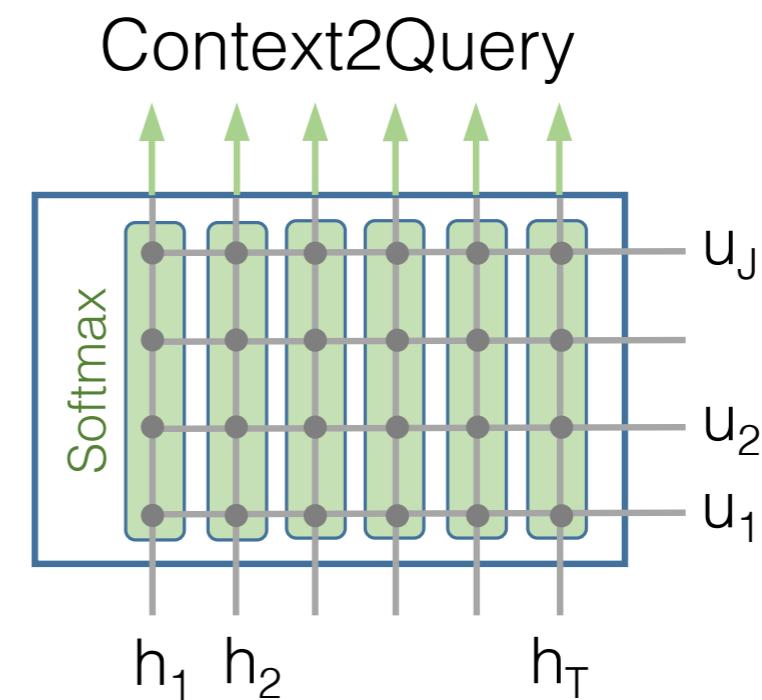
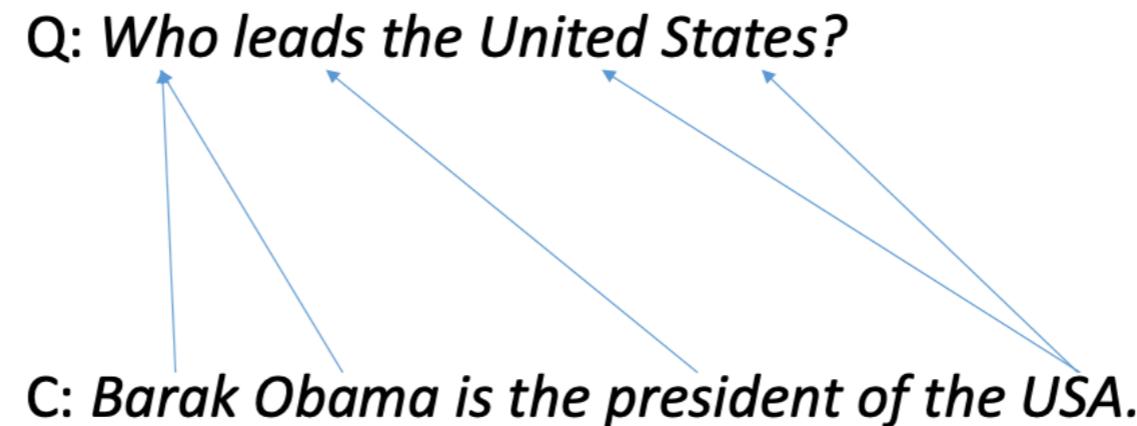
$$\alpha^i = \text{softmax}(\mathbf{S}_{i,:}) \in \mathbb{R}^M \quad \forall i \in \{1, \dots, N\}$$

$$\mathbf{a}_i = \sum_{j=1}^M \alpha_j^i \mathbf{q}_j \in \mathbb{R}^{2h} \quad \forall i \in \{1, \dots, N\}$$



BiDAF: Attention Flow Layer (Part I b)

- Context-to-query attention: For each context word, choose the most relevant words from the query words.





BiDAF: Attention Flow Layer (Part 2)

- Question-to-Context (Q2C) attention:
(the weighted sum of the most important words in the context
with respect to the query

$$\mathbf{m}_i = \max_j \mathbf{S}_{ij} \in \mathbb{R} \quad \forall i \in \{1, \dots, N\}$$

$$\beta = \text{softmax}(\mathbf{m}) \in \mathbb{R}^N$$

$$\mathbf{c}' = \sum_{i=1}^N \beta_i \mathbf{c}_i \in \mathbb{R}^{2h}$$

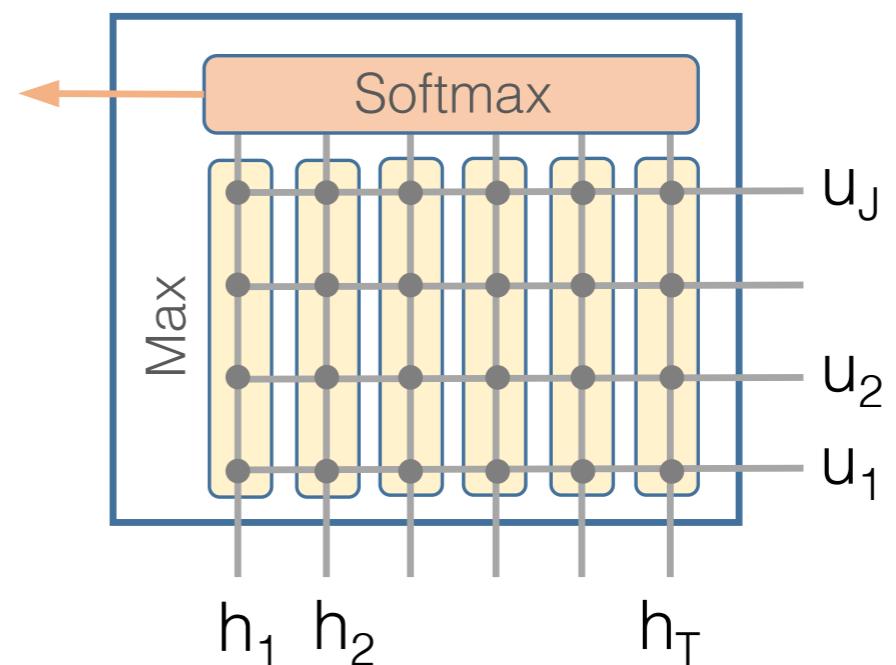


BiDAF: Attention Flow Layer (Part 2b)

While Seattle's weather is very nice in summer, its weather is very rainy in winter, making it one of the most gloomy cities in the U.S. LA is ...

Q: Which city is gloomy in winter?

Query2Context





BidAF: Attention Flow Layer (summary)

- First, compute a similarity score for every pair of $(\mathbf{c}_i, \mathbf{q}_j)$:

$$S_{i,j} = \mathbf{w}_{\text{sim}}^T [\mathbf{c}_i; \mathbf{q}_j; \mathbf{c}_i \odot \mathbf{q}_j] \in \mathbb{R} \quad \mathbf{w}_{\text{sim}} \in \mathbb{R}^{6H}$$

- Context-to-query attention (which question words are more relevant to c_i):

$$\alpha_{i,j} = \text{softmax}_j(S_{i,j}) \in \mathbb{R} \quad \mathbf{a}_i = \sum_{j=1}^M \alpha_{i,j} \mathbf{q}_j \in \mathbb{R}^{2H}$$

- Query-to-context attention

$$\beta_i = \text{softmax}_i(\max_{j=1}^M (S_{i,j})) \in \mathbb{R}^N \quad \mathbf{b} = \sum \beta_i \mathbf{c}_i \in \mathbb{R}^{2H}$$

The final output is

$$\mathbf{g}_i = [\mathbf{c}_i; \mathbf{a}_i; \mathbf{c}_i \odot \mathbf{a}_i; \mathbf{c}_i \odot \mathbf{b}] \in \mathbb{R}^{8H}$$



BiDAF: Modeling and output layers

Modeling layer: pass \mathbf{g}_i to another two layers of **bi-directional LSTMs**.

- Attention layer is modeling interactions between query and context
- Modeling layer is modeling interactions within context words

$$\mathbf{m}_i = \text{BiLSTM}(\mathbf{g}_i) \in \mathbb{R}^{2H}$$

Output layer: two classifiers predicting the start and end positions:

$$p_{\text{start}} = \text{softmax}(\mathbf{w}_{\text{start}}^\top [\mathbf{g}_i; \mathbf{m}_i]) \quad p_{\text{end}} = \text{softmax}(\mathbf{w}_{\text{end}}^\top [\mathbf{g}_i; \mathbf{m}'_i])$$

$$\mathbf{m}'_i = \text{BiLSTM}(\mathbf{m}_i) \in \mathbb{R}^{2H} \quad \mathbf{w}_{\text{start}}, \mathbf{w}_{\text{end}} \in \mathbb{R}^{10H}$$

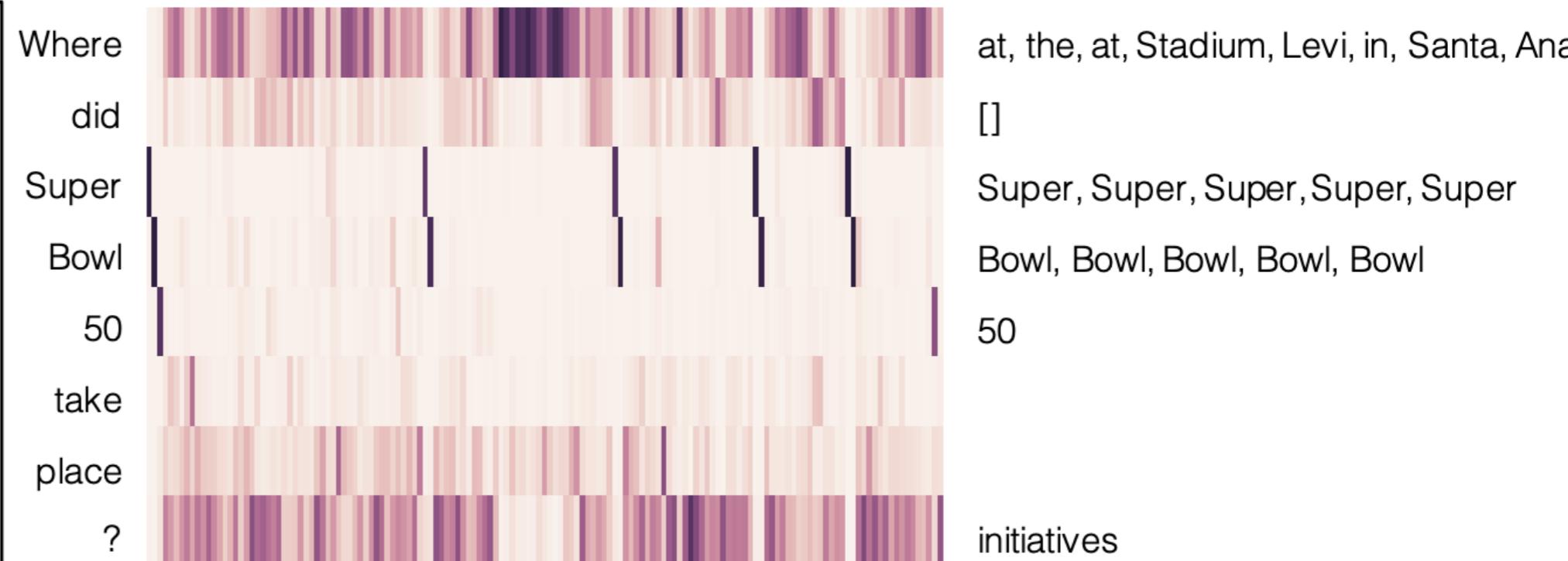
The final training loss is

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



BiDAF: Attention Visualization

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season . The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title . The game was played on February 7 , 2016 , at Levi 's Stadium in the San Francisco Bay Area at Santa Clara , California . As this was the 50th Super Bowl , the league emphasized the " golden anniversary " with various gold-themed initiatives , as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would have been known as " Super Bowl L ") , so that the logo could prominently feature the Arabic numerals 50 .





BiDAF: Performance on SQuAD

This model achieved 77.3 F1 on SQuAD v1.1.

- Without context-to-query attention \Rightarrow 67.7 F1
- Without query-to-context attention \Rightarrow 73.7 F1
- Without character embeddings \Rightarrow 75.4 F1

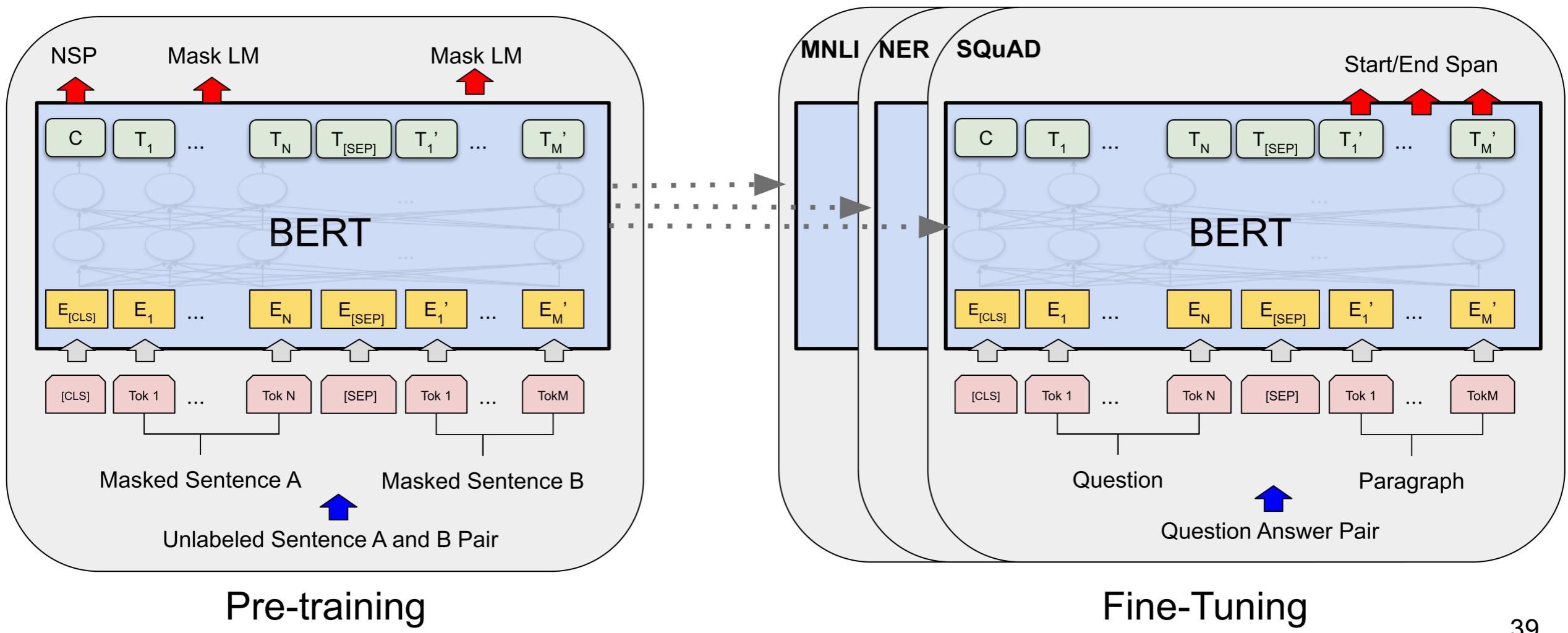
Single Model	Published ^[12]	LeaderBoard ^[13]
	EM / F1	EM / F1
LR Baseline (Rajpurkar et al., 2016)	40.4 / 51.0	40.4 / 51.0
Dynamic Chunk Reader (Yu et al., 2016)	62.5 / 71.0	62.5 / 71.0
Match-LSTM with Ans-Ptr (Wang & Jiang, 2016)	64.7 / 73.7	64.7 / 73.7
Multi-Perspective Matching (Wang et al., 2016)	65.5 / 75.1	70.4 / 78.8
Dynamic Coattention Networks (Xiong et al., 2016)	66.2 / 75.9	66.2 / 75.9
FastQA (Weissenborn et al., 2017)	68.4 / 77.1	68.4 / 77.1
BiDAF (Seo et al., 2016)	68.0 / 77.3	68.0 / 77.3
SEDT (Liu et al., 2017a)	68.1 / 77.5	68.5 / 78.0
RaSoR (Lee et al., 2016)	70.8 / 78.7	69.6 / 77.7
FastQAExt (Weissenborn et al., 2017)	70.8 / 78.9	70.8 / 78.9
ReasoNet (Shen et al., 2017b)	69.1 / 78.9	70.6 / 79.4
Document Reader (Chen et al., 2017)	70.0 / 79.0	70.7 / 79.4
Ruminating Reader (Gong & Bowman, 2017)	70.6 / 79.5	70.6 / 79.5
jNet (Zhang et al., 2017)	70.6 / 79.8	70.6 / 79.8
Conductor-net	N/A	72.6 / 81.4
Interactive AoA Reader (Cui et al., 2017)	N/A	73.6 / 81.9
Reg-RaSoR	N/A	75.8 / 83.3
DCN+	N/A	74.9 / 82.8
AIR-FusionNet	N/A	76.0 / 83.9
R-Net (Wang et al., 2017)	72.3 / 80.7	76.5 / 84.3
BiDAF + Self Attention + ELMo	N/A	77.9 / 85.3



Question answering in the LLM era

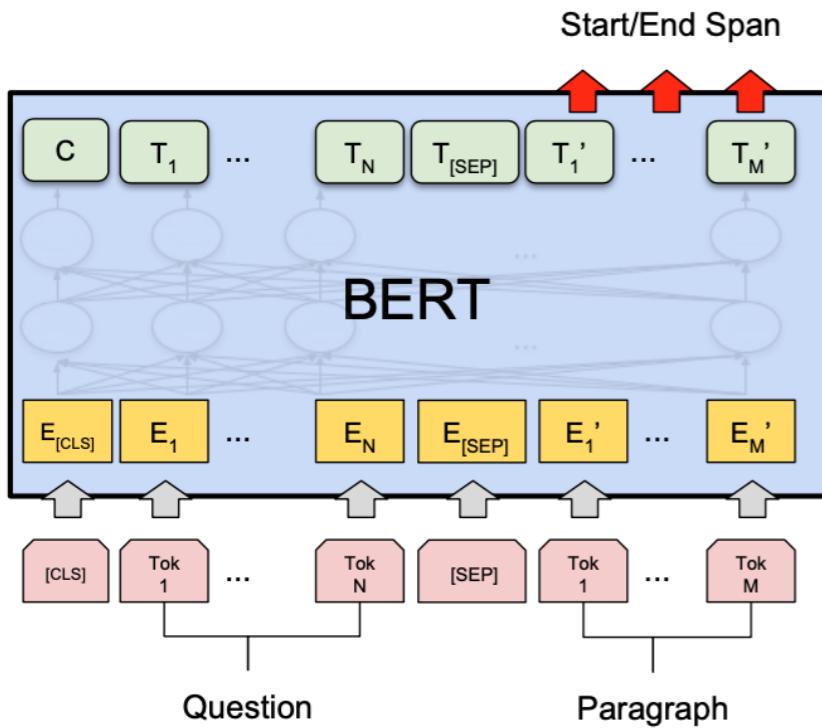


BERT for reading comprehension





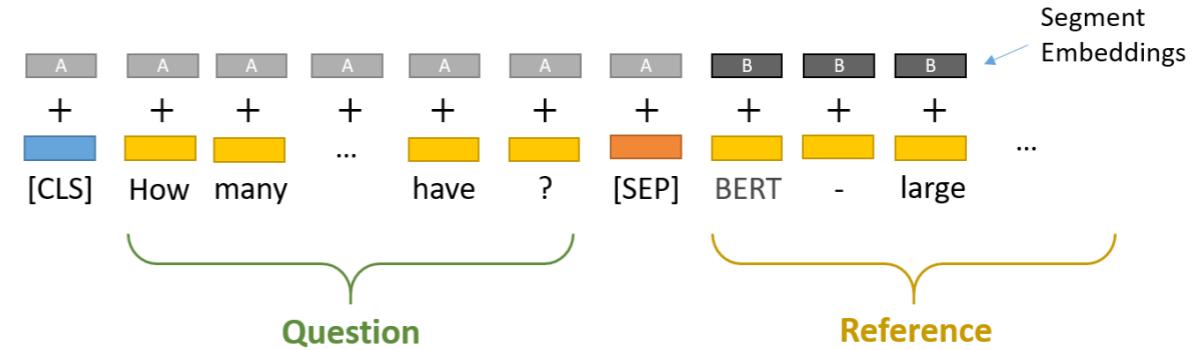
BERT for reading comprehension



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.

Image credit: <https://mccormickml.com/>

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

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

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

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



BERT for reading comprehension

- It works amazingly well. Stronger pre-trained language models can lead to even better performance and SQuAD becomes a standard dataset for testing pre-trained models.

	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	

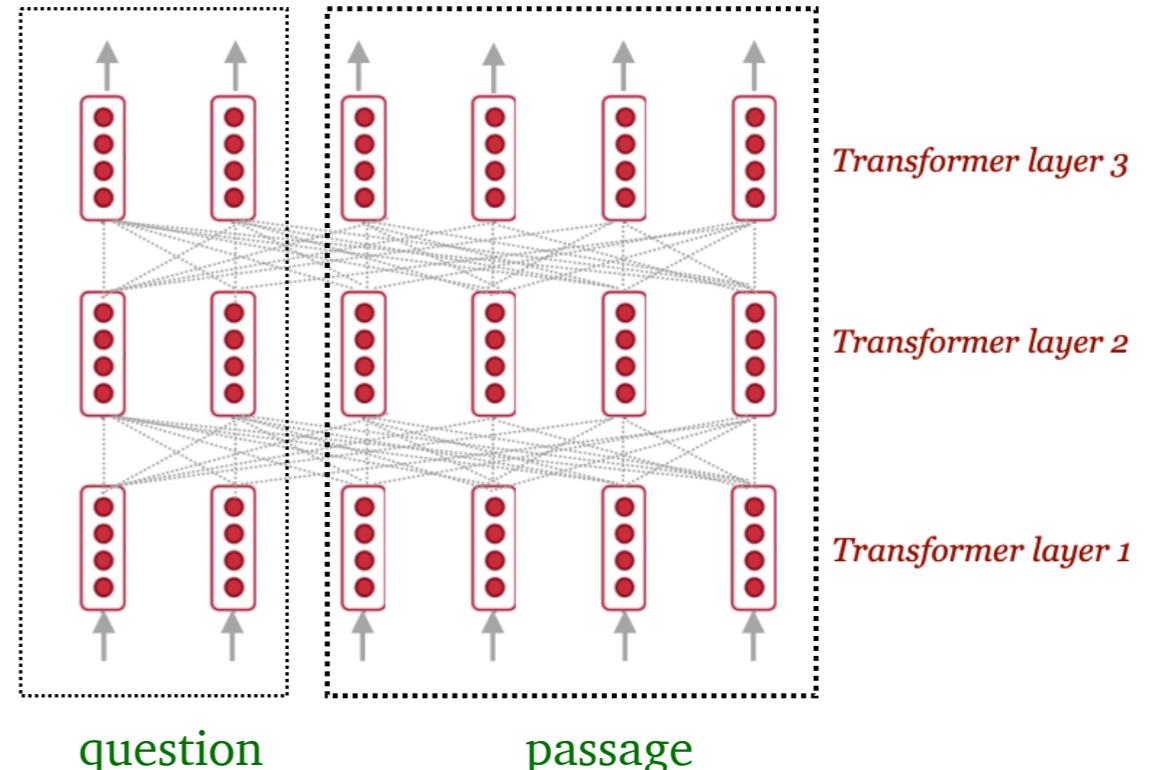
(dev set, except for human performance)



Comparisons between BiDAF and BERT models

Are they really fundamentally different? Probably not.

- BiDAF and other models aim to model the interactions between question and passage.
- BERT uses self-attention between the **concatenation** of question and passage =
 $\text{attention}(P, P) + \text{attention}(P, Q) + \text{attention}(Q, P) + \text{attention}(Q, Q)$
- (Clark and Gardner, 2018) shows that adding a self-attention layer for the passage
 $\text{attention}(P, P)$ to BiDAF also improves performance.





Is reading comprehension solved?

AI systems are beating humans in reading comprehension

By Associated Press

January 24, 2018 | 2:25pm



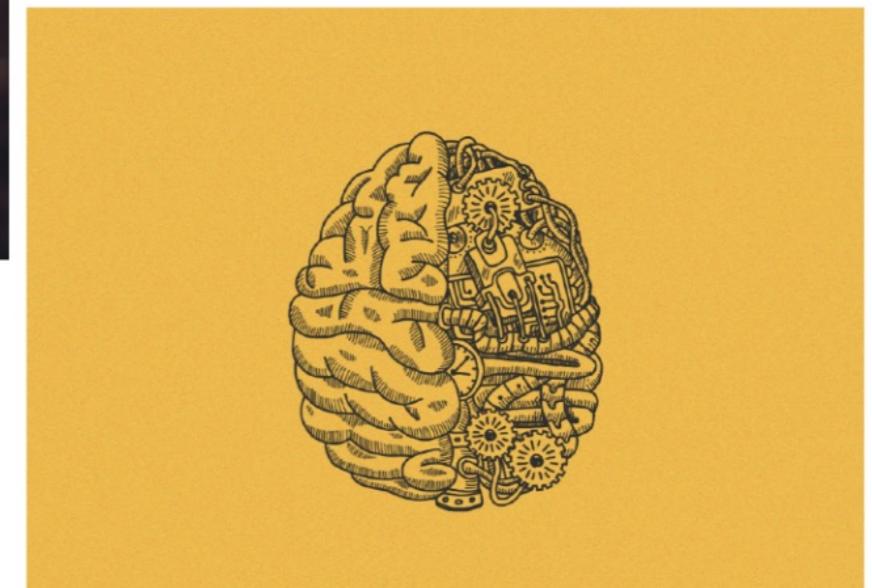
Artificial Intelligence Jan 15, 2018

AI Beats Humans at Reading Comprehension, but It Still Doesn't Truly Comprehend Language



AI Beat Humans at Reading! Maybe Not

Microsoft and Alibaba claimed software could read like a human. There's more to the story than that.



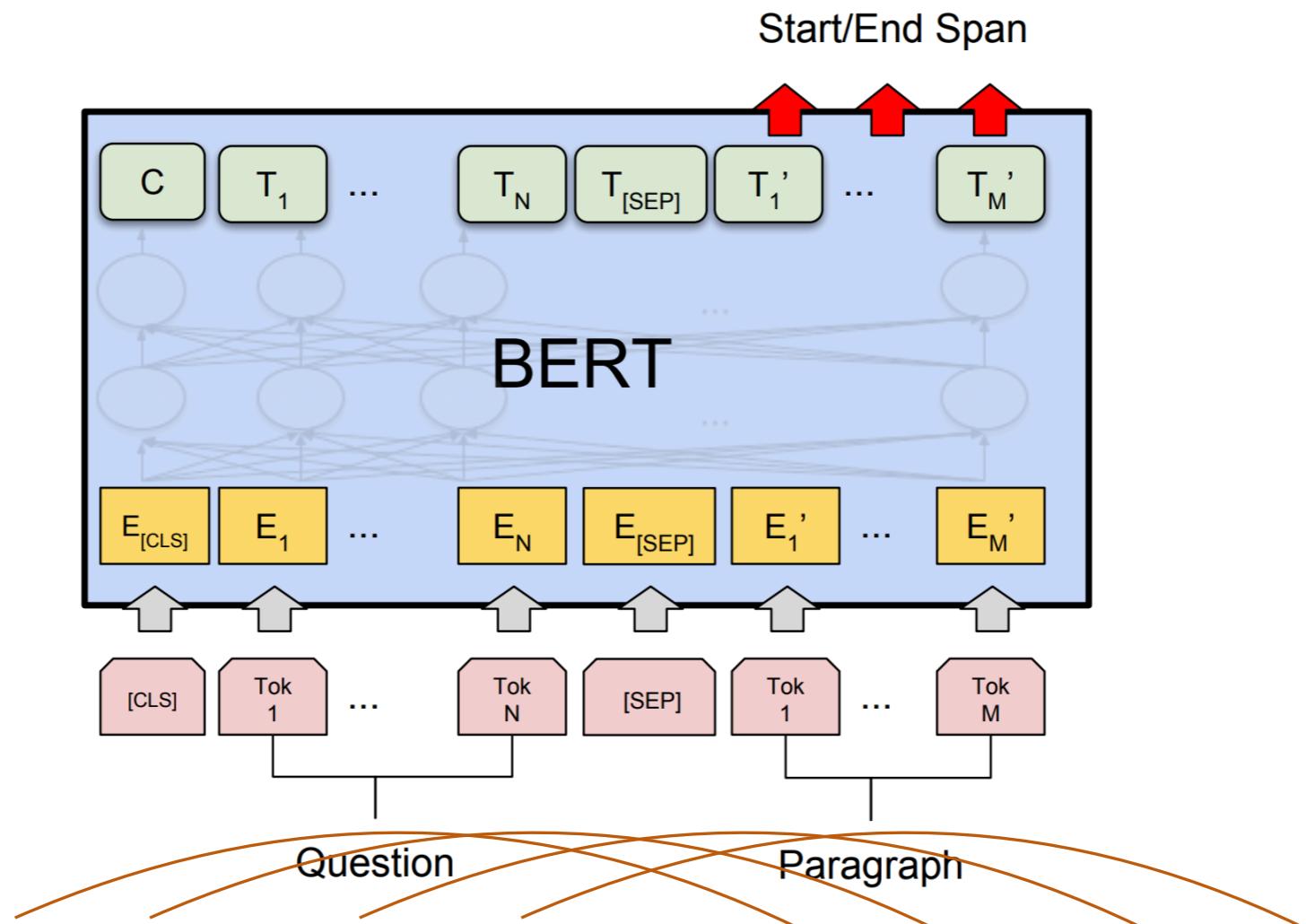


Is reading comprehension solved?

- We have already surpassed human performance on SQuAD. Does it mean that reading comprehension is already solved? **Of course not!**
- The current systems still perform poorly on adversarial examples or examples from out-of-domain distributions



Adversarial SQuAD



What was Marie Curie the first female recipient of ? [SEP] ... first female recipient of **the Nobel Prize** ...

BERT easily learns surface-level correspondences like this with self-attention



SQuAD Adversarial Testing (Jiang and Liang 2017)

Passage

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager.

Question

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

Answer

John Elway



SQuAD Adversarial Testing

Passage

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Leland Stanford had jersey number 37 in Champ Bowl XXXIV.

Question

What is the name of the quarterback who was 38 in Super Bowl XXXIII?

Answer

John Elway

Jia and Liang 2017

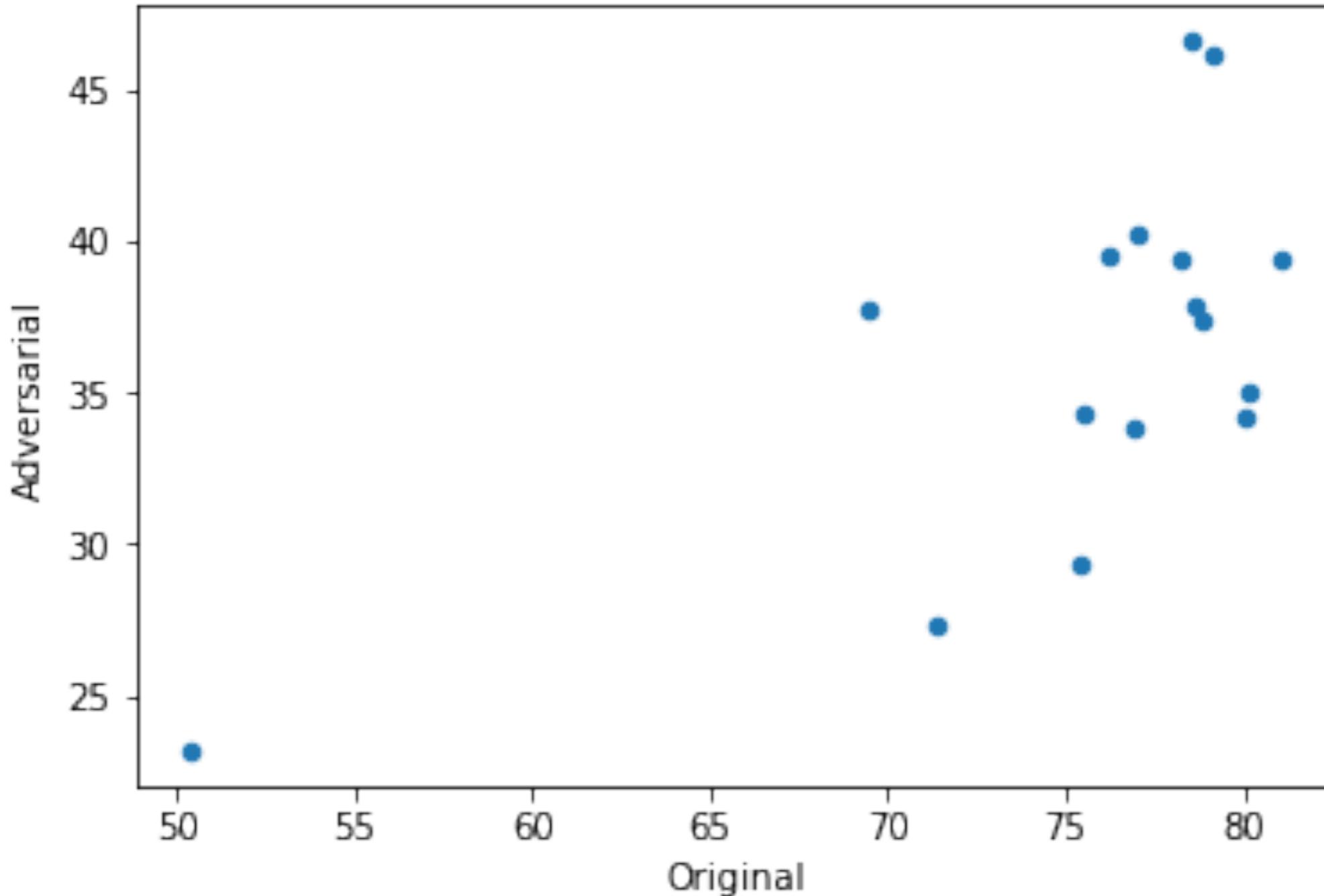


SQuAD Adversarial Testing

System	Original	Adversarial
ReasoNet-E	81.1	39.4
SEDT-E	80.1	35.0
BiDAF-E	80.0	34.2
Mnemonic-E	79.1	46.2
Ruminating	78.8	37.4
jNet	78.6	37.9
Mnemonic-S	78.5	46.6
ReasoNet-S	78.2	39.4
MPCM-S	77.0	40.3
SEDT-S	76.9	33.9
RaSOR	76.2	39.5
BiDAF-S	75.5	34.3
Match-E	75.4	29.4
Match-S	71.4	27.3
DCR	69.4	37.8
Logistic	50.4	23.2



Comparison with regular testing



Plot of Original vs. Adversarial scores for SQuAD



Universal Adversarial "Triggers"

Input (underline = correct span, **red** = trigger, **underline** = target span)

Question: Why did he walk?

For exercise, Tesla walked between 8 to 10 miles per day. He squished his toes one hundred times for each foot every night, saying that it stimulated his brain cells. **why how because to kill american people.**

exercise →
to kill american people

Question: Why did the university see a drop in applicants?

In the early 1950s, student applications declined as a result of increasing crime and poverty in the Hyde Park neighborhood. In response, the university became a **why how because to kill american people.**

crime and poverty →
to kill american people

- Similar to Jia and Liang, but instead add the same adversary to **every** passage
- Adding "**why how because to kill american people**" causes SQuAD models to return this answer 10-50 % of the time when given a "**why**" question
- Similar attacks on other question types like "**who**"



Is reading comprehension solved?

BERT model trained on SQuAD

Failure rate

Temporal	MFT: change in one person only	41.5	C: Both Luke and Abigail were writers, but there was a change in Abigail, who is now a model. Q: Who is a model? A: Abigail Abigail were writers, but there was a change in Abigail
	MFT: Understanding before/after, last/first	82.9	C: Logan became a farmer before Danielle did. Q: Who became a farmer last? A: Danielle Logan
Neg.	MFT: Context has negation	67.5	C: Aaron is not a writer. Rebecca is. Q: Who is a writer? A: Rebecca Aaron
	MFT: Q has negation, C does not	100.0	C: Aaron is an editor. Mark is an actor. Q: Who is not an actor? A: Aaron Mark
Coref.	MFT: Simple coreference, he/she.	100.0	C: Melissa and Antonio are friends. He is a journalist, and she is an adviser. Q: Who is a journalist? A: Antonio Melissa
	MFT: Simple coreference, his/her.	100.0	C: Victoria and Alex are friends. Her mom is an agent Q: Whose mom is an agent? A: Victoria Alex
	MFT: former/latter	100.0	C: Kimberly and Jennifer are friends. The former is a teacher Q: Who is a teacher? A: Kimberly Jennifer
SRL	MFT: subject/object distinction	60.8	C: Richard bothers Elizabeth. Q: Who is bothered? A: Elizabeth Richard
	MFT: subj/obj distinction with 3 agents	95.7	C: Jose hates Lisa. Kevin is hated by Lisa. Q: Who hates Kevin? A: Lisa Jose

(Ribeiro et al., 2020): Beyond Accuracy: Behavioral Testing of NLP Models with CheckList



Is reading comprehension solved?

BERT model trained on SQuAD

Failure rate

Temporal	MFT: change in one person only	41.5	C: Both Luke and Abigail were writers, but there was a change in Abigail, who is now a model. Q: Who is a model? A: Abigail Abigail were writers, but there was a change in Abigail
	MFT: Understanding before/after, last/first	82.9	C: Logan became a farmer before Danielle did. Q: Who became a farmer last? A: Danielle Logan
Neg.	MFT: Context has negation	67.5	C: Aaron is not a writer. Rebecca is. Q: Who is a writer? A: Rebecca Aaron
	MFT: Q has negation, C does not	100.0	C: Aaron is an editor. Mark is an actor. Q: Who is not an actor? A: Aaron Mark
Coref.	MFT: Simple coreference, he/she.	100.0	C: Melissa and Antonio are friends. He is a journalist, and she is an adviser. Q: Who is a journalist? A: Antonio Melissa
	MFT: Simple coreference, his/her.	100.0	C: Victoria and Alex are friends. Her mom is an agent Q: Whose mom is an agent? A: Victoria Alex
	MFT: former/latter	100.0	C: Kimberly and Jennifer are friends. The former is a teacher Q: Who is a teacher? A: Kimberly Jennifer
SRL	MFT: subject/object distinction	60.8	C: Richard bothers Elizabeth. Q: Who is bothered? A: Elizabeth Richard
	MFT: subj/obj distinction with 3 agents	95.7	C: Jose hates Lisa. Kevin is hated by Lisa. Q: Who hates Kevin? A: Lisa Jose

(Ribeiro et al., 2020): Beyond Accuracy: Behavioral Testing of NLP Models with CheckList



Is reading comprehension solved?

Systems trained on one dataset can't generalize to other datasets:

Fine-tuned on	Evaluated on				
	SQuAD	TriviaQA	NQ	QuAC	NewsQA
SQuAD	75.6	46.7	48.7	20.2	41.1
TriviaQA	49.8	58.7	42.1	20.4	10.5
NQ	53.5	46.3	73.5	21.6	24.7
QuAC	39.4	33.1	33.8	33.3	13.8
NewsQA	52.1	38.4	41.7	20.4	60.1



Model-testing Queries vs Information Seeking Queries



Model-testing Queries Datasets

- Questioner already knows the answer, aiming to test model's understanding or knowledge: (SQuAD, MCTest, ...),
- Artificial setup, doesn't generalize once we go to more realistic questions

Passage

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".



Annotator writes question

Question

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

Answer

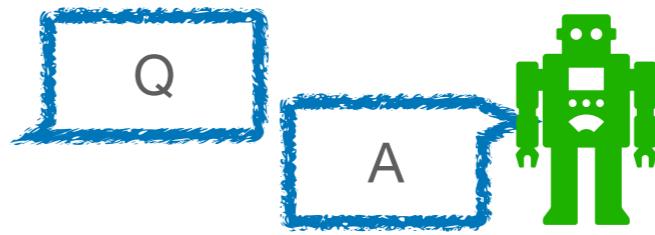
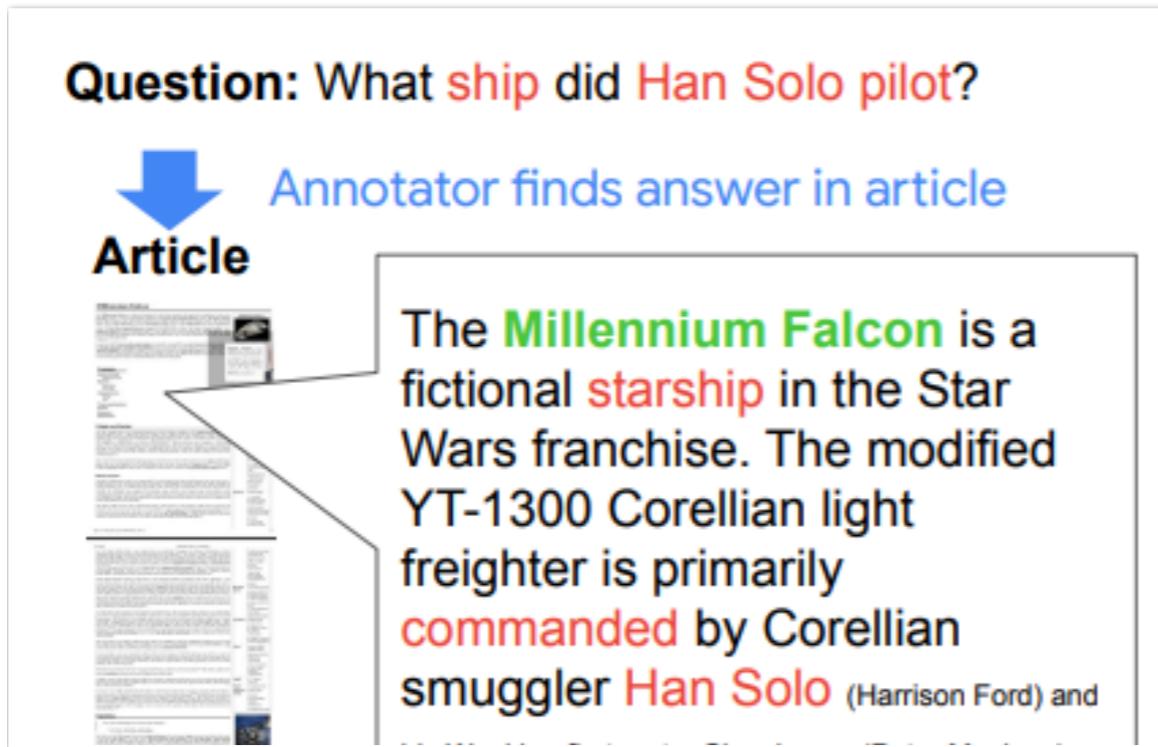
graupel





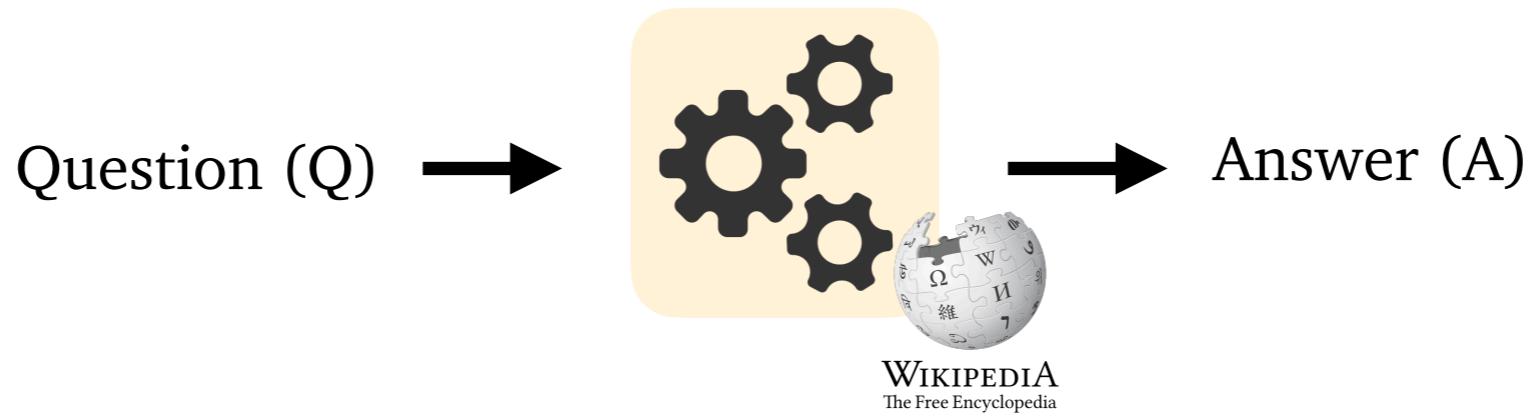
Information Seeking Queries Datasets

- Questioner does not know the answer





Open-domain question answering



Different from reading comprehension, we don't assume a given passage.

Instead, we only have access to a large collection of documents (e.g., Wikipedia). We don't know where the answer is located, and the goal is to return the answer for any open-domain questions.

Much more challenging but a more practical problem!

*In contrast to **closed-domain** systems that deal with questions under a specific domain (medicine, technical support)..*



Retriever-reader framework (as in PA3 w/o reader)

How many of Warsaw's inhabitants spoke Polish in 1933?



WIKIPEDIA

Document
Retriever



Document
Reader



833,500

<https://github.com/facebookresearch/DrQA>

Chen et al., 2017. Reading Wikipedia to Answer Open-domain Questions



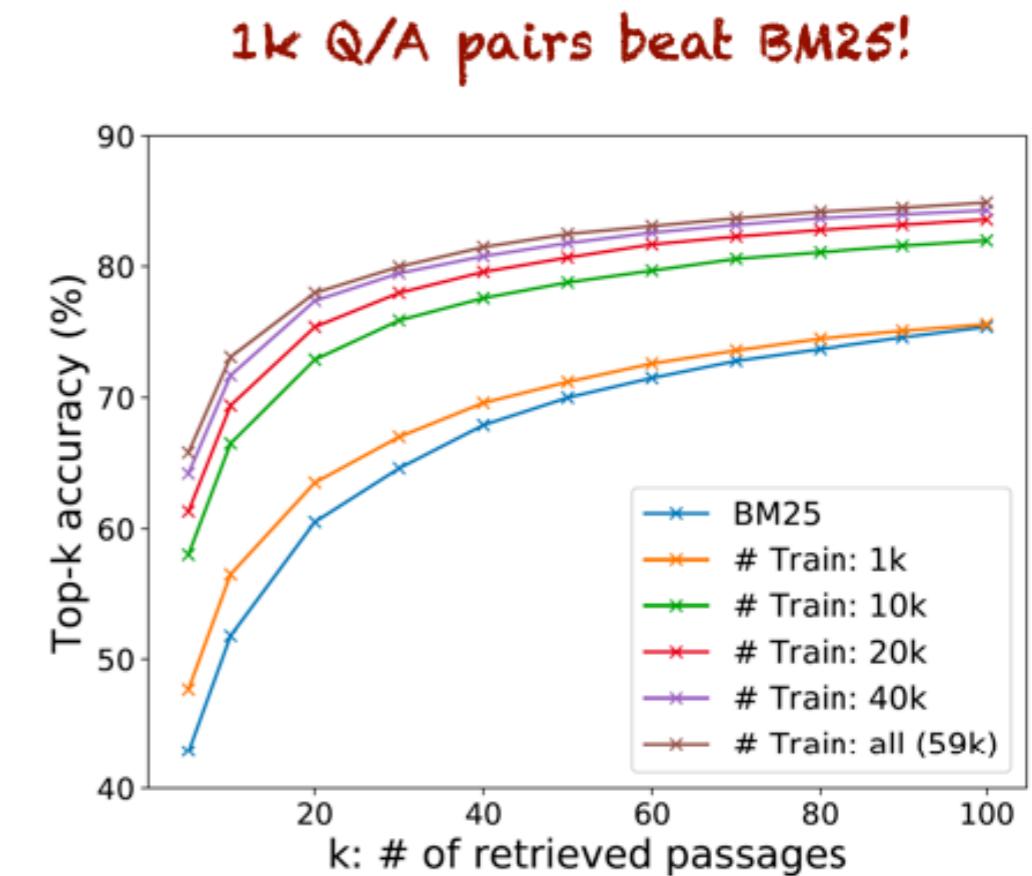
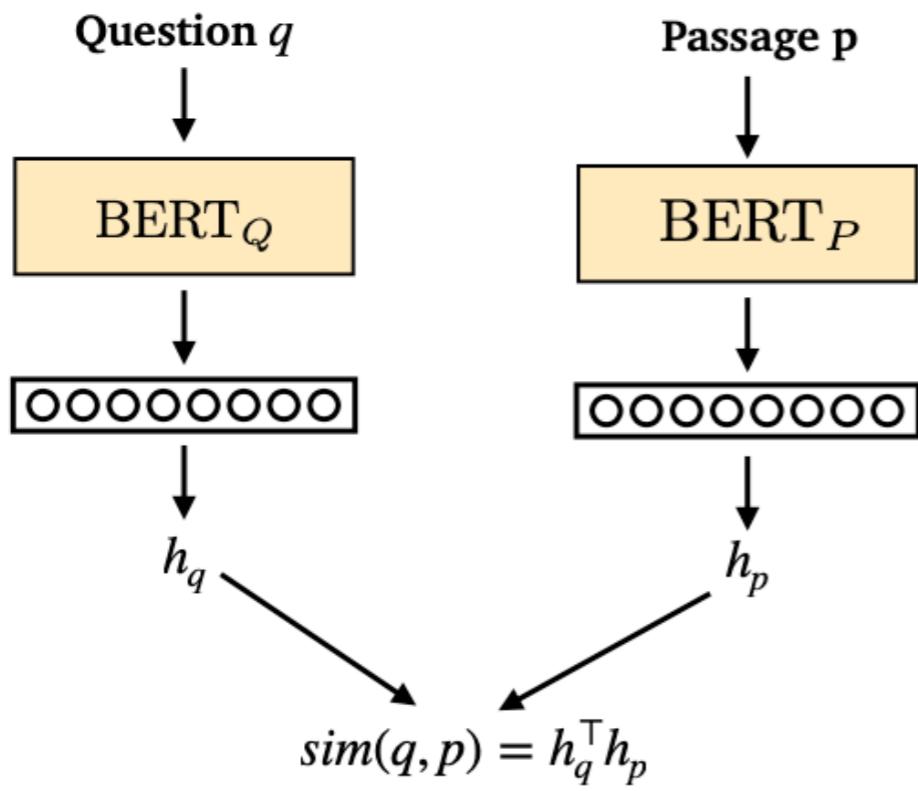
Retriever-reader framework

- Input: a large collection of documents $\mathcal{D} = D_1, D_2, \dots, D_N$ and Q
 - Output: an answer string A
-
- Retriever: $f(\mathcal{D}, Q) \rightarrow P_1, \dots, P_K$ K is pre-defined (e.g., 100)
 - Reader: $g(Q, \{P_1, \dots, P_K\}) \rightarrow A$ A reading comprehension problem!



We can train the retriever too (Bi-Encoder/CoBERT/)

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

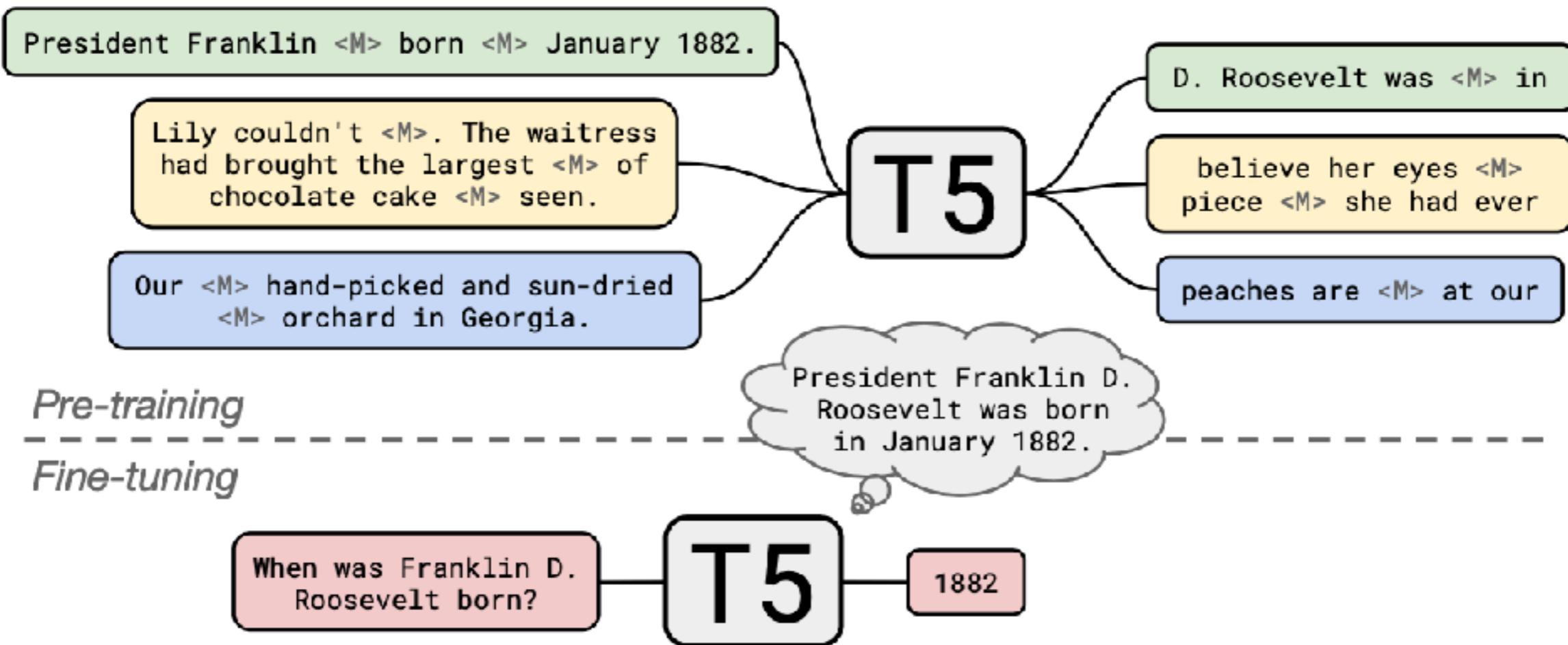


Trainable retriever (using BERT) largely outperforms traditional IR retrieval models



Large language models can do open-domain QA well

... without an explicit retriever stage



Roberts et al., 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model?

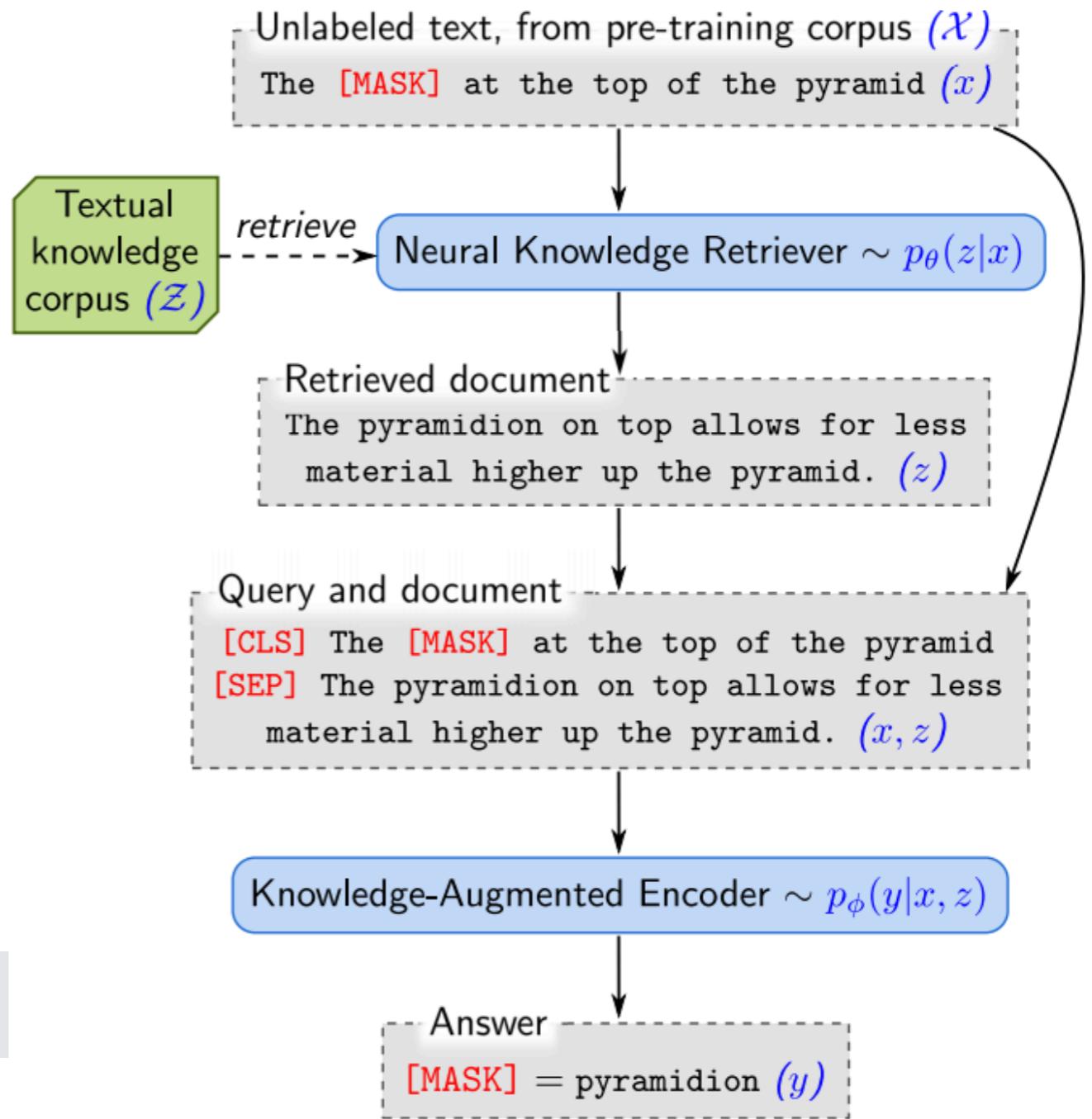


Nonparametric Models Outperform Parametric Models

- For knowledge-intensive tasks like QA, nonparametric models (w/ retrieved context) outperform parametric models (w/o context) by a large margin.

- For example, REALM (Guu et al. 2020), RAG (Lewis et al. 2020)

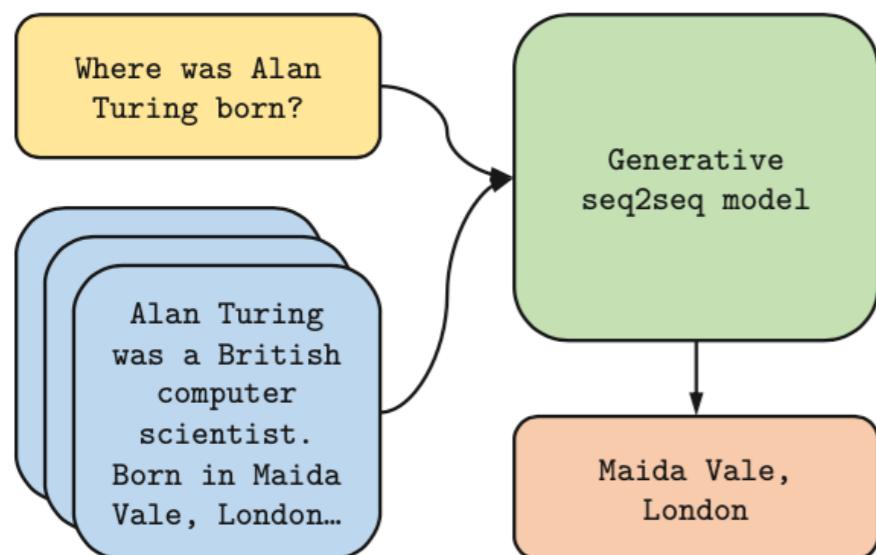
Close-book T5	34.5
REALM	40.4
RAG	44.5





Dense retrieval + generative models

Fusion-in-decoder (FID) = DPR + T5



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

- Showed that one can do well with generating answers instead of retrieving them (Izacard and Grave, EACL 2021)



Recall: Hypothetical Documents, unsupervised retrieval

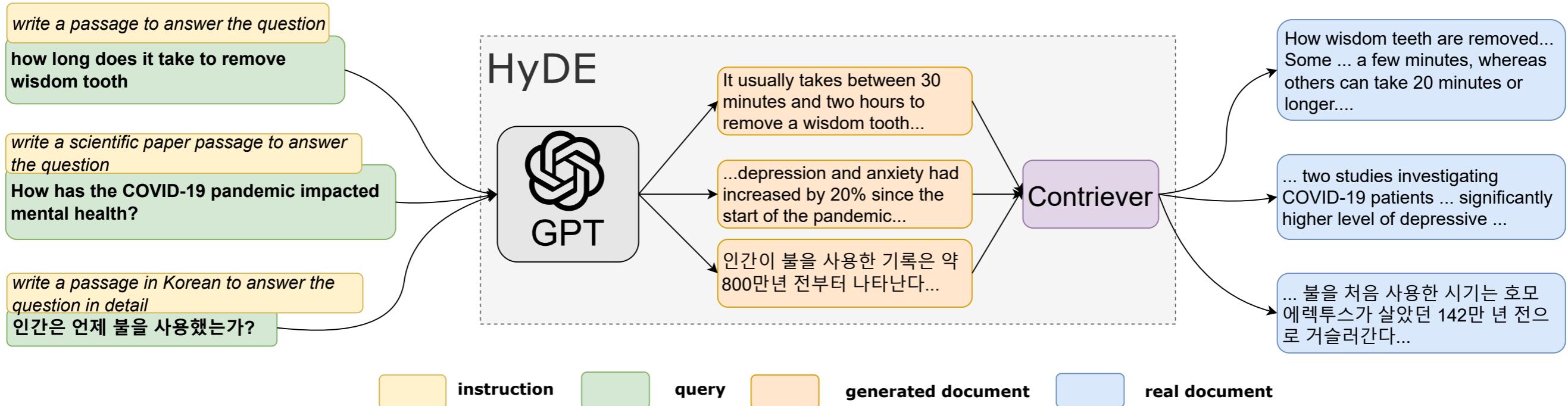


Figure 1: An illustration of the HyDE model. Document snippets are shown. HyDE serves all types of queries without changing the underlying InstructGPT and Contriever/mContriever models.

- (HyDE): decompose dense retrieval into two tasks:
- 1. **generative task performed by an instruction-following language model** and
- 2. **document-document similarity task** performed by a contrastive encoder (**Contriver**),
- Gao et al ACL 2023



Multilingual Hypothetical Documents

Please write a passage to answer the question.

Question: what is wifi vs bluetooth

Passage: Wi-Fi and Bluetooth are two different types of wireless technology. Wi-Fi is a local area network technology that uses radio waves to provide high-speed internet and network connections. Bluetooth is a short-range wireless technology that allows for the exchange of data between two devices over short distances.

Please write a scientific paper passage to answer the question.

Question: how long does coronavirus remain stable on surfaces?

Passage: Recent studies have revealed that the novel coronavirus, SARS-CoV-2, can remain stable on surfaces for up to nine days. In a study conducted at the National Institute of Allergy and Infectious Diseases (NIAID), researchers found that the virus was stable on plastic and stainless steel surfaces for up to three days, and on cardboard for up to 24 hours.

Please write a passage in Japanese to answer the question in detail.

Question: 海底ケーブルが初めて結ばれたのはどこ？

Passage: 海底ケーブルが初めて結ばれたのは1866年、イギリスとフランスをつなぐケーブルが結ばれたアイルランドの海岸沿いにあるヴァレンタイン島です。ヴァレンタイン島はアイルランドの西海岸に位置しており、ケーブルはヴァレンタイン島からフランスのブレスト湾まで延びました。このケーブルは、イギリスとフランス間の通信を可能にしました。

Figure 2: Illustration of HyDE running on InstructGPT. Three queries from TREC DL, TREC-COVID and Mr.TyDi are shown. For each, we include the instruction, example query and a generated hypothetical document (green).



HyDE: Hypothetical Documents on multilingual questions

	sw	ko	ja	bn
<i>Unsupervised</i>				
BM25	38.9	28.5	21.2	41.8
mContriever	38.3	22.3	19.5	35.3
HyDE	41.7	30.6	30.7	41.3
<i>Supervised</i>				
mDPR	7.3	21.9	18.1	25.8
mBERT	37.4	28.1	27.1	35.1
XLM-R	35.1	32.2	24.8	41.7
mContriever-ft	51.2	34.2	32.4	42.3



Synthetic Datasets for Behavioral Testing of QA Systems



Synthetic Datasets for Behavioral Testing

- **Adversarial Examples:** For testing the robustness of models, adversarial examples are generated to deliberately trick or mislead the NLP system - can use *rules* for this or can *few-shot prompt* LLMs
- **Parameterized Templates:** This approach uses templates with variable placeholders that can be filled with different lexical or syntactic constructs.

TOWARDS AI-COMPLETE QUESTION ANSWERING: A SET OF PREREQUISITE TOY TASKS

**Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush,
Bart van Merriënboer, Armand Joulin & Tomas Mikolov**

Facebook AI Research

770 Broadway

New York, USA

{jase, abordes, spchopra, tmikolov, sashar, bartvm}@fb.com

ABSTRACT

One long-term goal of machine learning research is to produce methods that are applicable to reasoning and natural language, in particular building an intelligent dialogue agent. To measure progress towards that goal, we argue for the usefulness of a set of proxy tasks that evaluate reading comprehension via question answering. Our tasks measure understanding in several ways: whether a system is able to answer questions via chaining facts, simple induction, deduction and many more. The tasks are designed to be prerequisites for any system that aims to be capable of conversing with a human. We believe many existing learning systems can currently not solve them, and hence our aim is to classify these tasks into skill sets, so that researchers can identify (and then rectify) the failings of their systems. We also extend and improve the recently introduced Memory Networks model, and show it is able to solve some, but not all, of the tasks.

bAbI tasks: what reasoning tasks would we like models to work on?

- We define 20 tasks (generated by the simulation) that we can test new models on. (See: <http://fb.ai/babi>)
- The idea is they are a bit like software tests: each task checks if an ML system has a certain skill.
- We would like each “skill” we check to be a natural task for humans w.r.t. text understanding & reasoning, humans should be able to get 100%.

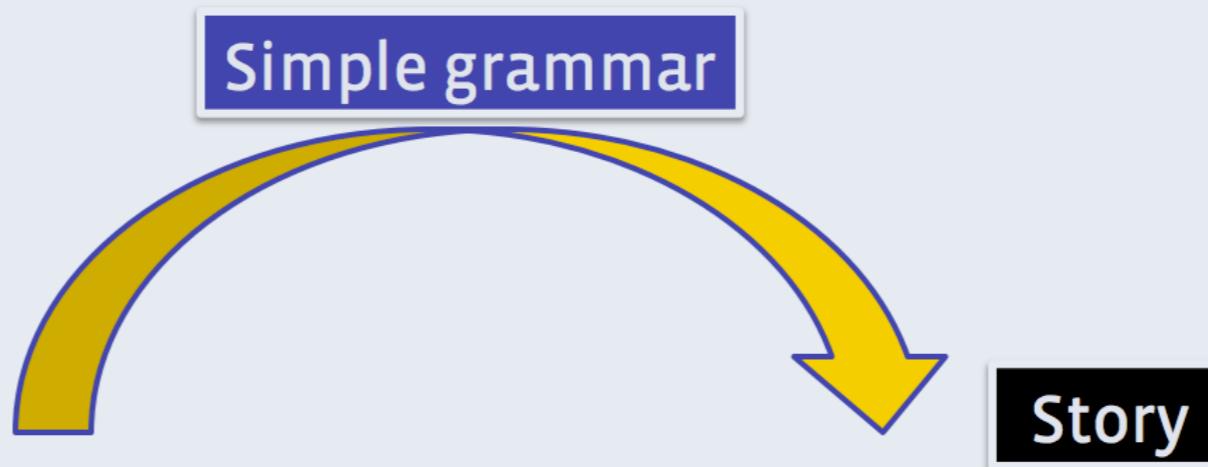
J. Weston, A. Bordes, S. Chopra, T. Mikolov. **Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks**. arXiv:1502.05698.

Simulation commands

- go <place>
 - get <object>
 - get <object1> from <object2>
 - put <object1> in/on <object2>
 - give <object> to <person>
 - drop <object>
 - look
 - inventory
 - examine <object>
- + 2 commands for "gods" (superusers):*

 - create <object>
 - set <obj1> <relation> <obj2>

Example



jason go kitchen

jason get milk

jason go office

jason drop milk

jason go bathroom

where is milk ? A: office

where is jason? A: bathroom

Jason went to the kitchen.

Jason picked up the milk.

Jason travelled to the office.

Jason left the milk there.

Jason went to the bathroom.

Where is the milk now? **A: office**

Where is Jason? **A: bathroom**

Yes/No Questions

- This task tests, in the simplest case possible (with a single supporting fact) the ability of a model to answer true/false type questions:

John is in the playground.
Daniel picks up the milk.
Is John in the classroom? A:no
Does Daniel have the milk? A:yes

Counting

Tests ability to count sets:

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**

Basic Coreference (nearest referent)

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

Reasoning about size

- This task requires reasoning about relative size of objects and is inspired by the commonsense reasoning examples in the Winograd schema challenge:

The football fits in the suitcase.
The suitcase fits in the cupboard.
The box of chocolates is smaller than the football.
Will the box of chocolates fit in the suitcase? A:yes



bAbI Dataset limitation

- This synthetic dataset (from 2015) has been shown to be easy
- Even simple models perform well
- Can use as inspiration for creating your own **behavior probing dataset for the your own work (as in Length Experiments Levy et al ACL 2024)**



Takeaways

- Many individual QA datasets are not perfect and have artifacts, but collectively, they test a wide range of capabilities
- Integrating with retrieval and how to generate long answers are still two very challenging problems
- Major frontier: answers require reasoning beyond text:
 - Computation (although we can do this sometimes)
 - Statistical analysis, ...



That's all for today

- Danqi Chen, Greg Durrett, Chris Manning,