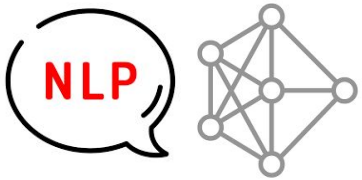
CHULA ENGINEERING
Foundation toward Innovation
COMPUTER



Question Answering

2110572: Natural Language Processing Systems

Peerapon Vateekul & Ekapol Chuangsuwanich
Department of Computer Engineering,
Faculty of Engineering, Chulalongkorn University

Credit:

- Kasidis Kanwatchara
- Can Udomcharoenchaikit & Nattachai Tretasayuth

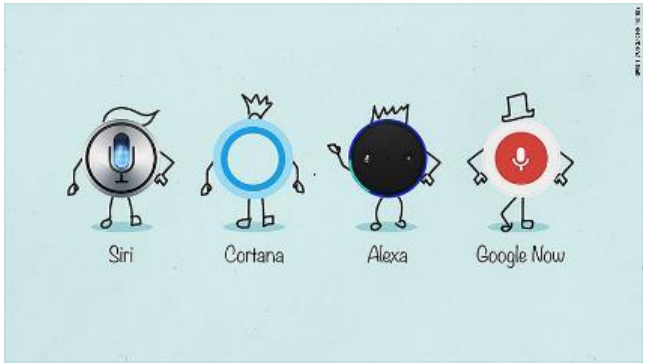
Outline

- Part1) Introduction
- Part2) Traditional QA
- Part3) Neural-based QA
- Part4) Transformer-based QA
  - 1) Encoder, 2) Decoder, 3) Retrieval
  - SOTA: Atlas, RePlug, ChatGPT (not really QA; chatbot)
  - Demo
- Part5) QA data sets (7 data sets)

Part1) Introduction

What's Question Answering (QA)?

- QA is a field that combines (1) Information Retrieval, (2) Information Extraction and (3) Natural Language Processing.
  - *We will focus on the NLP part*
- Most notable QA software is IBM's Watson
- Nowadays, QA also play a significant role in Personal Assistant (Siri, Cortana, etc.)



[Figure by Sandy Jakobs (left), IBM (right)]

Type Of QA

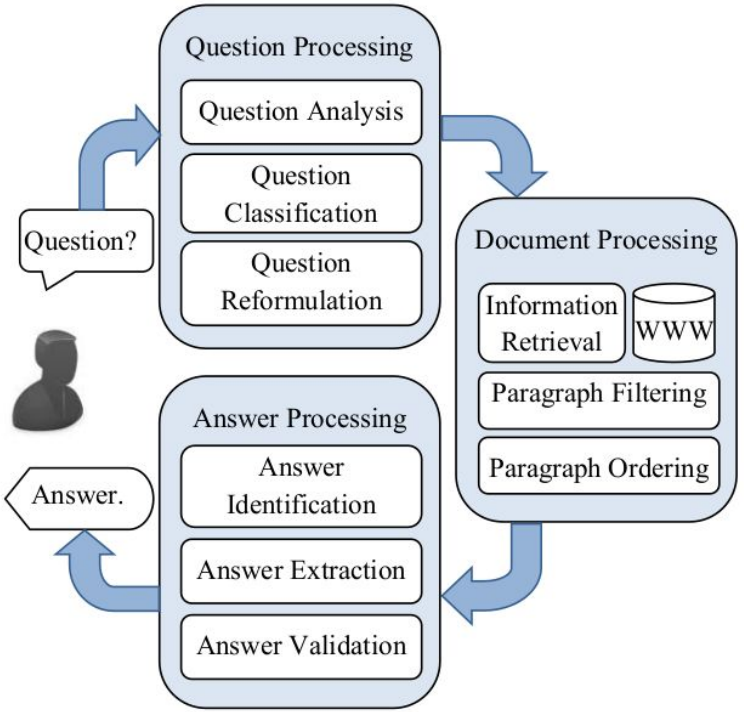
- By application domains
  - Restricted Domain
  - Open Domain
- By source of data
  - Structured data (Knowledge-based) - e.g. Freebase, Google Knowledge Graph
  - Unstructured data (Document)- Web, Wiki
- By answer
  - Factoid (single word - when, what, where)
  - non-Factoid (e.g., list, how, why)
- The forms of answer
  - Extracted text
  - Generated answer

Type Of QA

- Machine Reading Comprehension (MRC)
  - Given a reference and a question
    - Find the answer in the reference text
- OpenQA (recent trend)
  - Only a question is given
  - Two types of OpenQA
    - "Open-book" QA
      - An external data source can be used, e.g. a document retriever
    - "Closed-book" QA
      - Use only the knowledge stored inside a model

Process Of Traditional QA

- Question Processing
  - What type of question?
  - Question preprocessing
- Document Processing
  - Rank candidate document
  - Rank candidate paragraph
- Answer Processing
  - Extract candidate answer from paragraph
  - Construct an answer



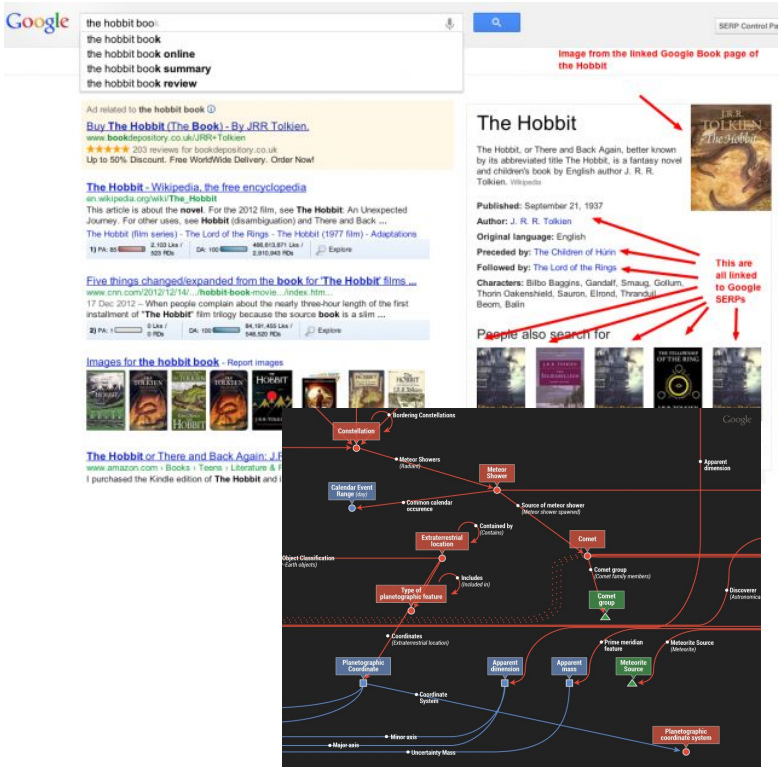
[Figure from "The Question Answering Systems: A Survey"]

+

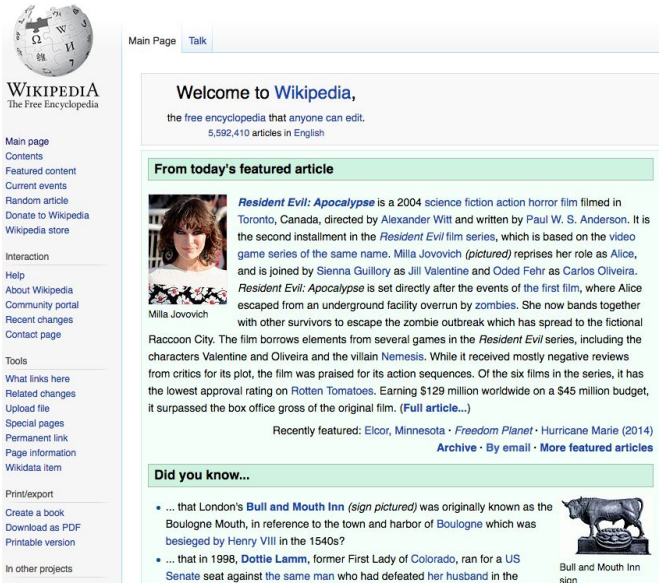
Part2) Traditional QA

Types of QA systems

Structured Knowledge Base

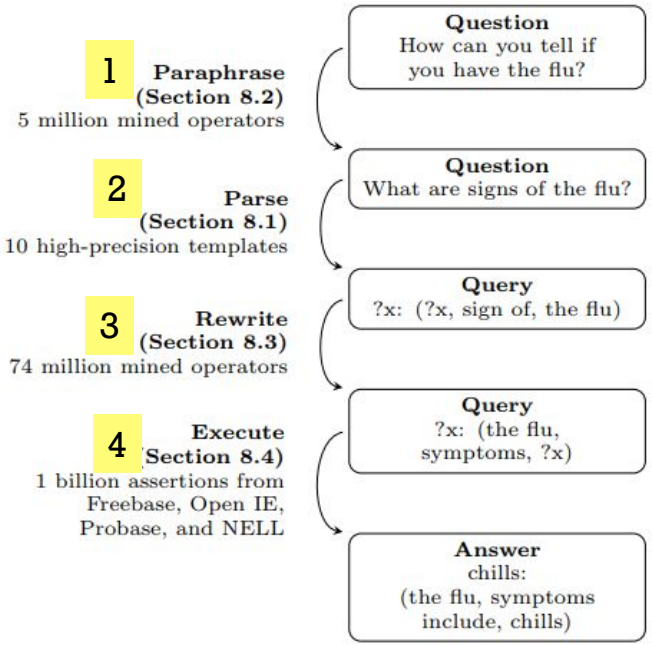


Unstructured Knowledge Base



Example of Traditional QA system

- Open Question Answering Over Curated and Extracted Knowledge Bases (A.Fader SIGKDD 2014)



[Figure from "Open Question Answering Over Curated and Extracted Knowledge Bases"]

Example of Traditional Methods (cont.)

- Open Question Answering Over Curated and Extracted Knowledge Bases (A.Fader SIGKDD 2014)
  - 1) Paraphrase operator
    - are responsible for rewording the input question into the domain of a parsing operator
    - Source template (open domain) → Target template (predefined format)

Source Template	Target Template
How does _ affect your body?	What body system does _ affect?
What is the latin name for _?	What is _'s scientific name?
Why do we use _?	What did _ replace?
What to use instead of _?	What is a substitute for _?
Was _ ever married?	Who has _ been married to?

Table 3: Example paraphrase operators that extracted from a corpus of unlabeled questions.

[Figure from "Open Question Answering Over Curated and Extracted Knowledge Bases"]

Example of Traditional Methods (cont.)

- Open Question Answering Over Curated and Extracted Knowledge Bases (A.Fader SIGKDD 2014)
  - 2) Parsing operator
    - responsible for interfacing between natural language questions and the KB query language
    - Target template (predefined format) → Query

Question Pattern	Query Pattern	Example Question	Example Query
Who/What RV_{rel} NP_{arg}	(?x, rel, arg)	Who invented papyrus?	(?x, invented, papyrus)
Who/What Aux NP_{arg} RV_{rel}	(arg, rel, ?x)	What did Newton discover?	(Newton, discover, ?x)
Where/When Aux NP_{arg} RV_{rel}	(arg, rel in, ?x)	Where was Edison born?	(Edison, born in, ?x)
Where/When is NP_{arg}	(arg, is in, ?x)	Where is Detroit?	(Detroit, is in, ?x)
Who/What is NP_{arg}	(arg, is-a, ?x)	What is potassium?	(potassium, is-a, ?x)
What/Which NP_{rel2} Aux NP_{arg} RV_{rel1}	(arg, rel1 rel2, ?x)	What sport does Sosa play?	(Sosa, play sport, ?x)
What/Which NP_{rel} is NP_{arg}	(arg, rel, ?x)	What ethnicity is Dracula?	(Dracula, ethnicity, ?x)
What/Who is NP_{arg}'s NP_{rel}	(arg, rel, ?x)	What is Russia's capital?	(Russia, capital, ?x)
What/Which NP_{type} Aux NP_{arg} RV_{rel}	(?x, is-a, type) (arg, rel, ?x)	What fish do sharks eat?	(?x, is-a, fish) (sharks, eat, ?x)
What/Which NP_{type} RV_{rel} NP_{arg}	(?x, is-a, type) (?x, rel, arg)	What states make oil?	(?x, is-a, states) (?x, make, oil)

[Figure from "Open Question Answering Over Curated and Extracted Knowledge Bases"]

Example of Traditional Methods (cont.)

- Open Question Answering Over Curated and Extracted Knowledge Bases (A.Fader SIGKDD 2014)
  - 3) Query-rewrite operators
    - responsible for interfacing between the vocabulary used in the input question and the internal vocabulary used by the KBs
    - Source Query → Target Query (only vocab in knowledge base)

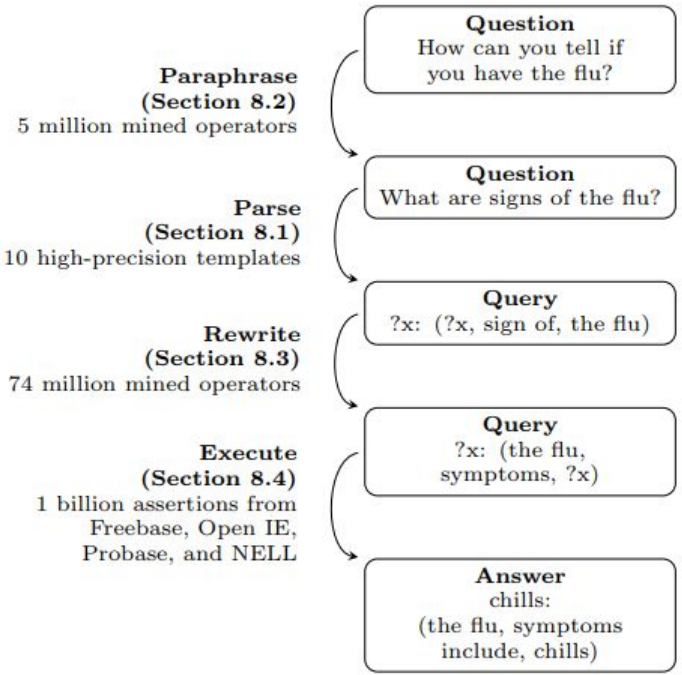
Source Query	Target Query
(?x, children, ?y)	(?y, was born to, ?x)
(?x, birthdate, ?y)	(?x, date of birth, ?y)
(?x, is headquartered in, ?y)	(?x, is based in, ?y)
(?x, invented, ?y)	(?y, was invented by, ?x)
(?x, is the language of, ?y)	(?y, languages spoken, ?x)

Table 4: Example query-rewrite operators mined from the knowledge bases described in Section 4.1.

[Figure from "Open Question Answering Over Curated and Extracted Knowledge Bases"]

Example of Traditional Methods (cont.)

- Open Question Answering Over Curated and Extracted Knowledge Bases (A.Fader SIGKDD 2014)
  - 4) Execution operator
    - responsible for fetching and combining evidence from the Knowledge based, given a query



[Figure from "Open Question Answering Over Curated and Extracted Knowledge Bases"]

Limitation

- Require a lot of time and linguistic knowledge to create **a template**
- Require many templates for each question type (manual process)
- Can only answer simple **factoid** question

+

Part3) Neural-Based QA

Deep QA models

Memory Network
[Jason Weston, et al., 2015]

Machine Learning with memory component

Attention Sum Reader
[Rudolf Kadlec, et al., 2016]

uses attention to directly pick the answer from the context

BiDAF
[Minjoon Seo, et al., 2017]

Apply attention flow mechanism (context-to-query, query-to-context)

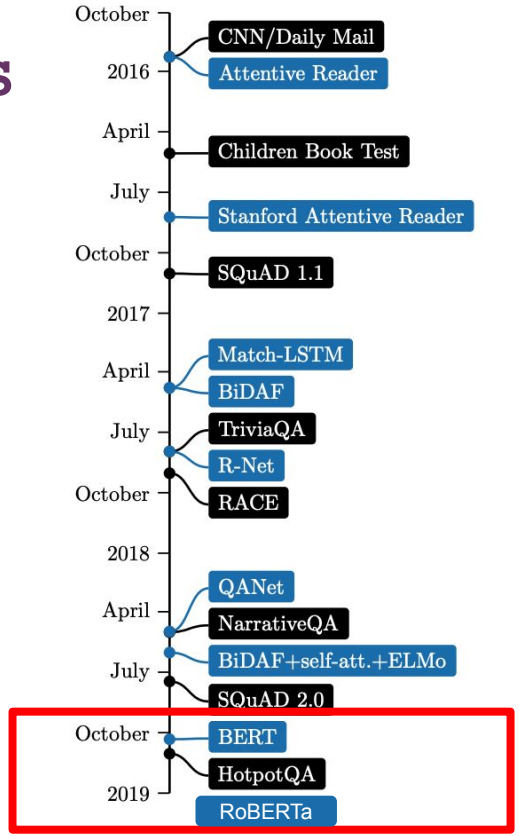
End-to-End Memory Network
[Sainbayar Sukhbaatar, et al., 2015]

Use attention mechanism to select relevant memory

Stanford Attentive Reader
[Danqi Chen, et al., 2016]

a summary attentive reader

Deep QA models (cont.)



Transformer-based QA

Figure 2.2: The recent development of datasets (black) and models (blue) in neural reading comprehension. For the timeline, we use the date that the corresponding papers were published, except BERT (Devlin et al., 2018).

Reference: Danqi Chen, 2018, https://stacks.stanford.edu/file/druid:gd576xb1833/thesis-augmented.pdf

Question Answering on SQuAD2.0

March-2023

Leaderboard | Community Models | Dataset | Description

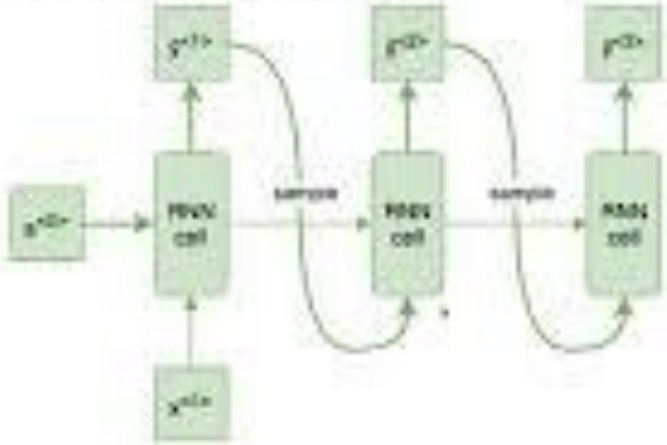
Human performance

Human performance has already been surpassed since 2019

BiDAF from the NLP class in 2020

Text generation model (inference)

- To generate a novel sequence, the inference model (testing phase) randomly samples an output from a softmax distribution.



https://www.youtube.com/watch?v=5SH9bpQ33Xk&list=PLcBOyD1N1T-NP11DsVK9XcN54rvfGBb96&index=8

Part4) Transformer-Based QA

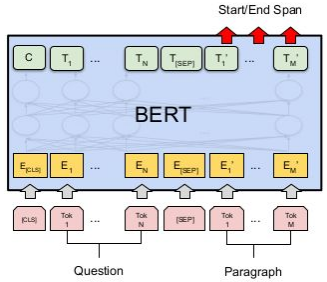
1) Encoder, 2) Decoder, 3) Retrieval

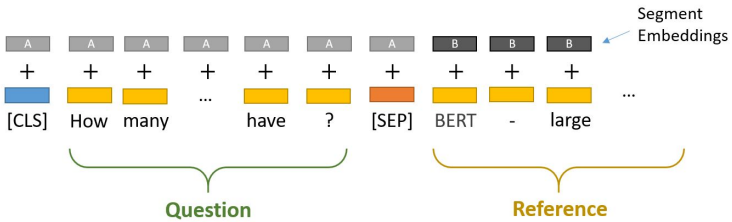
SOTA: Atlas, RePlug, ChatGPT (not really QA; chatbot)

Demo

1. Encoder-Based

- Uses any pretrained encoder-based like BERT
- Adds 2 linear layers to classify each token as the start and end indices
- Requires a reference text

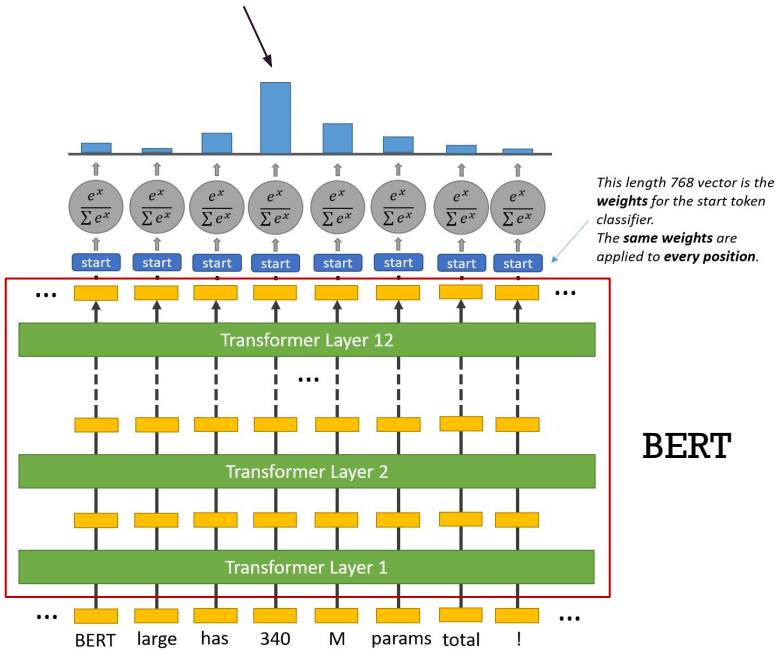




Question: How many parameters does BERT-large have?

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.

The start of the answer span

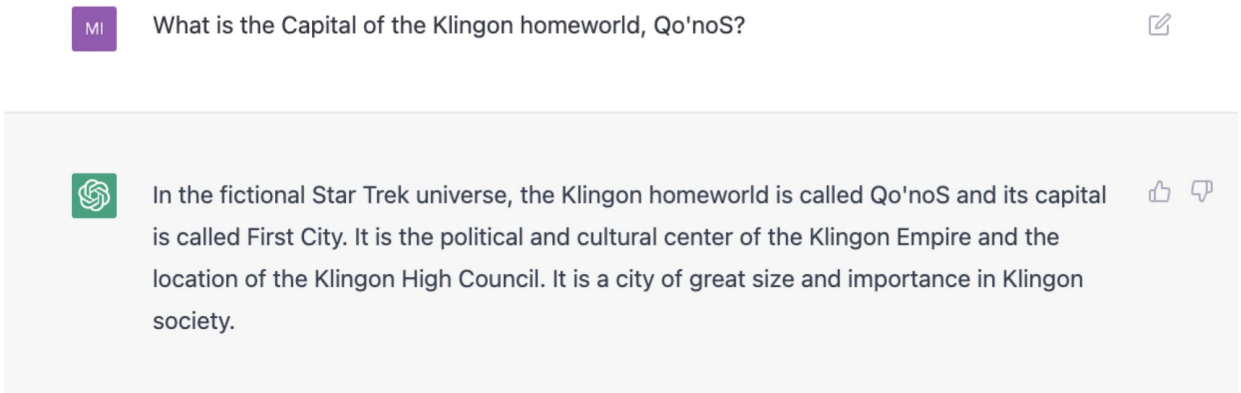


BERT

https://mccormickml.com/2020/03/10/question-answering-with-a-fine-tuned-BERT/

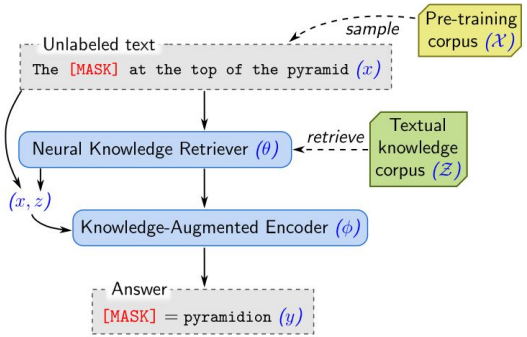
2. Decoder/Encoder-Decoder Based

- Generates the answer instead of trying to predict the start/end index
- Uses knowledge inside a model to answer questions. For example, ChatGPT can answer questions without needing a reference text.
  - Although given reference text can improve the performance.
- More practical!

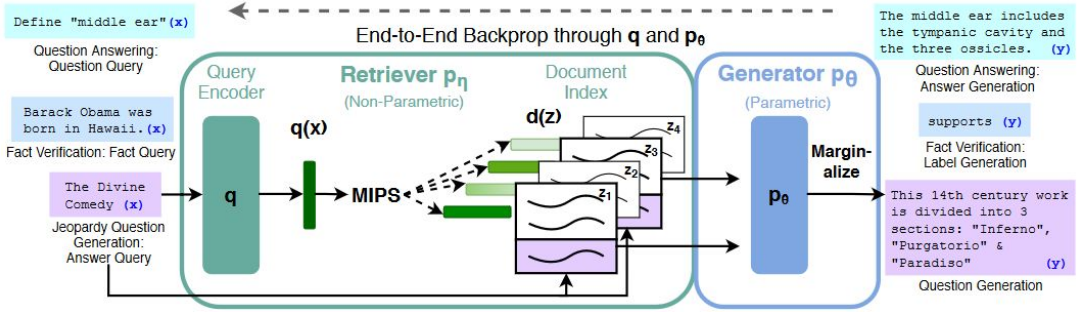


3. Retrieval-Augmented Model

- Use a retrieval engine and a language model.
- The retrieval engine fetches a list of documents
- And the LM use the list as reference text.



REALM [1] (encoder)



RAG [2] (decoder)

1 Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: retrieval-augmented language model pre-training. In Proceedings of the 37th International Conference on Machine Learning (ICML'20). JMLR.org, Article 368, 3929–3938.

2 Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20). Curran Associates Inc., Red Hook, NY, USA, Article 793, 9459–9474.

SOTA

- Retrieval-augmented models are more efficient at knowledge-intensive tasks
- However, they are usually more computationally expensive due to the retrieval step.

		NQ		TriviaQA filtered		TriviaQA unfiltered	
	Model	64-shot	Full	64-shot	Full	64-shot	Full
Decoder models	GPT-3 (Brown et al., 2020)	29.9	-	-	-	71.2	-
	Gopher (Rae et al., 2021)	28.2	-	57.2	-	61.3	-
	Chinchilla (Hoffmann et al., 2022)	35.5	-	64.6	-	72.3	-
540B →	PaLM (Chowdhery et al., 2022)	39.6	-	-	-	81.4	-
Retrieval models	RETRO (Borgeaud et al., 2021)	-	45.5	-	-	-	-
	FiD (Izacard & Grave, 2020)	-	51.4	-	67.6	-	80.1
	FiD-KD (Izacard & Grave, 2021)	-	54.7	-	73.3	-	-
	R2-D2 (Fajcik et al., 2021)	-	55.9	-	69.9	-	-
11B →	ATLAS	**42.4**	**60.4**	**74.5**	**79.8**	**84.7**	**89.4**
Retrieval models with LLMs	Codex + REPLUG	44.7	-	76.8	-		
	Codex + REPLUG LSR	**45.5**	-	**77.3**	-		

GPT3 (API; off-the-shelf model) + retrieval

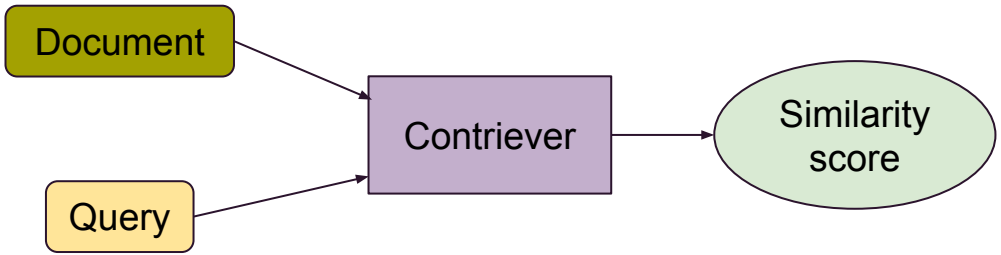
Izacard et al.(2022). Atlas: Few-shot Learning with Retrieval Augmented Language Models
Shi et al. (2023). REPLUG: Retrieval-Augmented Black-Box Language Models

SQTA (cont.)

- Atlas (retrieval based QA)
- RePlug (retrieval + decoder based QA)
- ChatGPT (not really QA; chatbot)

1) Atlas = Contriever + FiD [Retrieval-based model]

Contriever (a transformer encoder model) is a dense retriever trained using contrastive learning

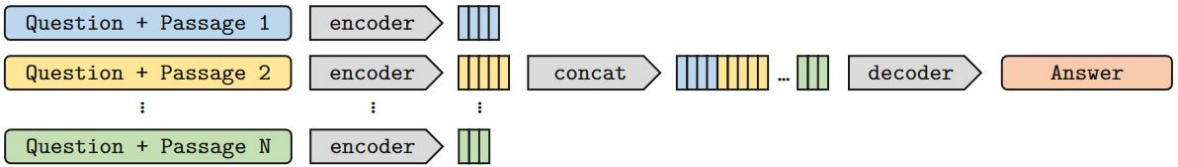


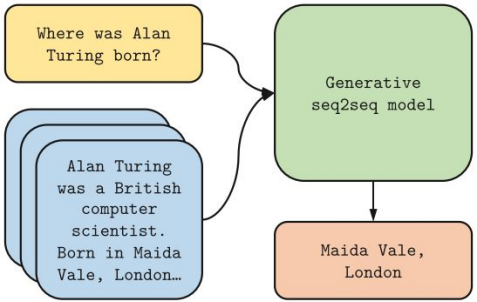
During inference, it accepts a query and encodes it into a fixed-length embedding. The dot-product of the query embedding and a document embedding returns how relevant a document is to the query. Top-k most similar documents are returned.

G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, E. Grave Unsupervised Dense Information Retrieval with Contrastive Learning 2022

1) Atlas = Contriever + FiD [Retrieval-based model]

Fusion in Decoder (a transformer encoder-decoder model) is a generative model for open QA.



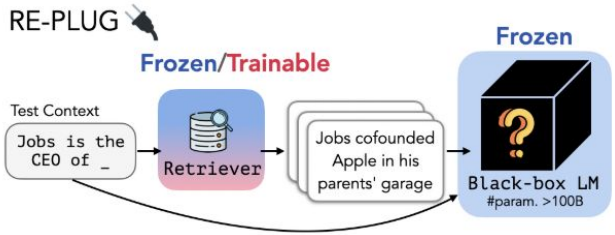


The model accepts a question and documents as inputs. The encoder independently encodes the documents(+question). The resulting representations are concatenated and finally given to the decoder to try to generate the correct answer. This allows the fusion of information from multiple documents (thus the name).

G. Izacard, E. Grave Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering 2021

2) RePlug [Retrieval + decoder based model]

Given a black-box LM (such as OpenAI's GPT API), the work attempts to add a tunable retrieval model to improve the LM QA performance by retrieving documents that boosts the probability of generating the correct answer.



The retriever is finetuned using the feedback from the frozen LM. The retrieval likelihood is adjusted so that documents that actually help the LM answer the question get chosen more frequently while the unhelpful documents get penalized.

REPLUG: Retrieval-Augmented Black-Box Language Models. Shi et al., 2023

REPLUG: Retrieval-Augmented Black-Box Language Models

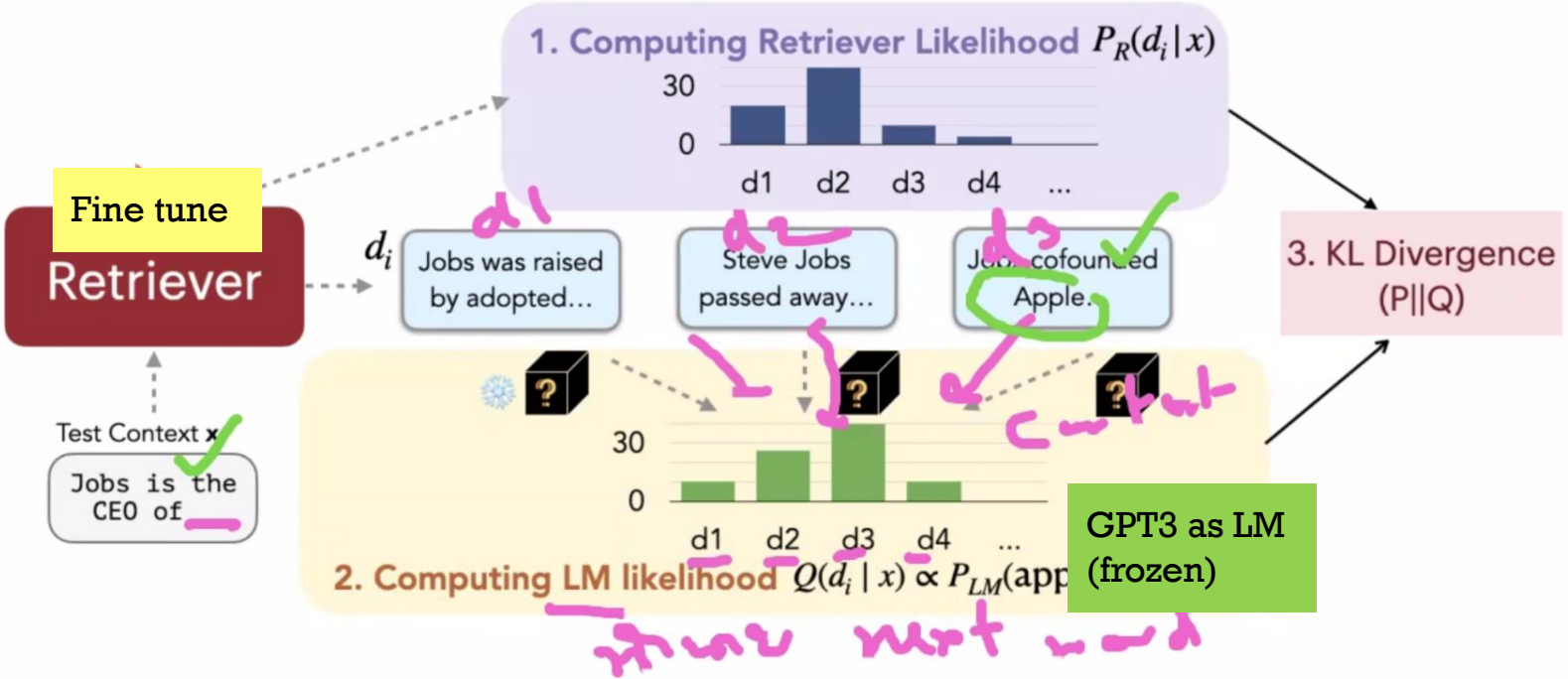


Figure 3. **REPLUG LSR training process (§4).** The retriever is trained using the output of a frozen language model as supervision signals.

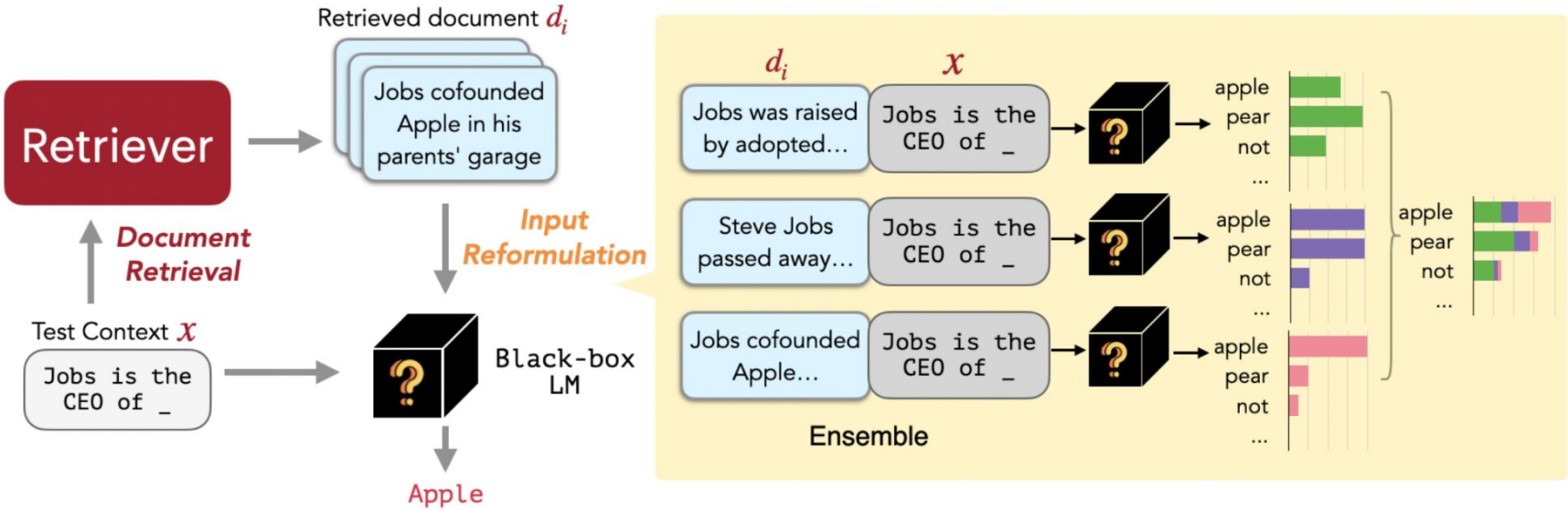
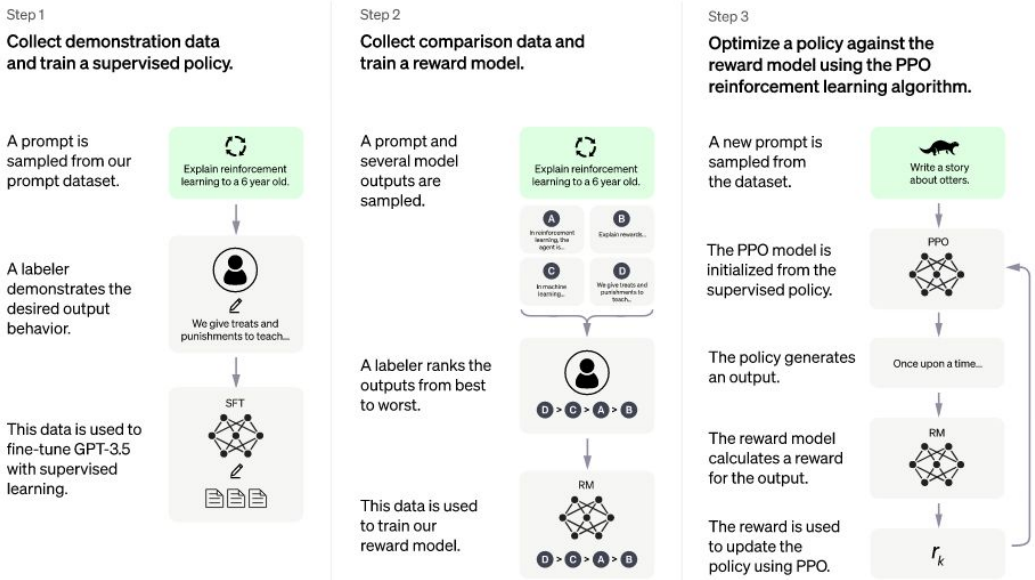


Figure 2. **REPLUG at inference** (§3). Given an input context, REPLUG first retrieves a small set of relevant documents from an external corpus using a retriever (§3.1 *Document Retrieval*). Then it prepends each document separately to the input context and ensembles output probabilities from different passes (§3.2 *Input Reformulation*).

3) ChatGPT

ChatGPT is the chatbot that is made by finetuning GPT-3 to follow instruction and then further finetuned with RL algorithm to improve its helpfulness. Note that the knowledge in the model is mainly learned from the pretraining stage. The finetuning stages are only to align the model output with human preference.



https://openai.com/blog/chatgpt

QA is still an open problem

5 minute read · February 9, 2023 7:49 AM GMT+7 · Last Updated 4 days ago

Alphabet shares dive after Google AI chatbot Bard flubs answer in ad

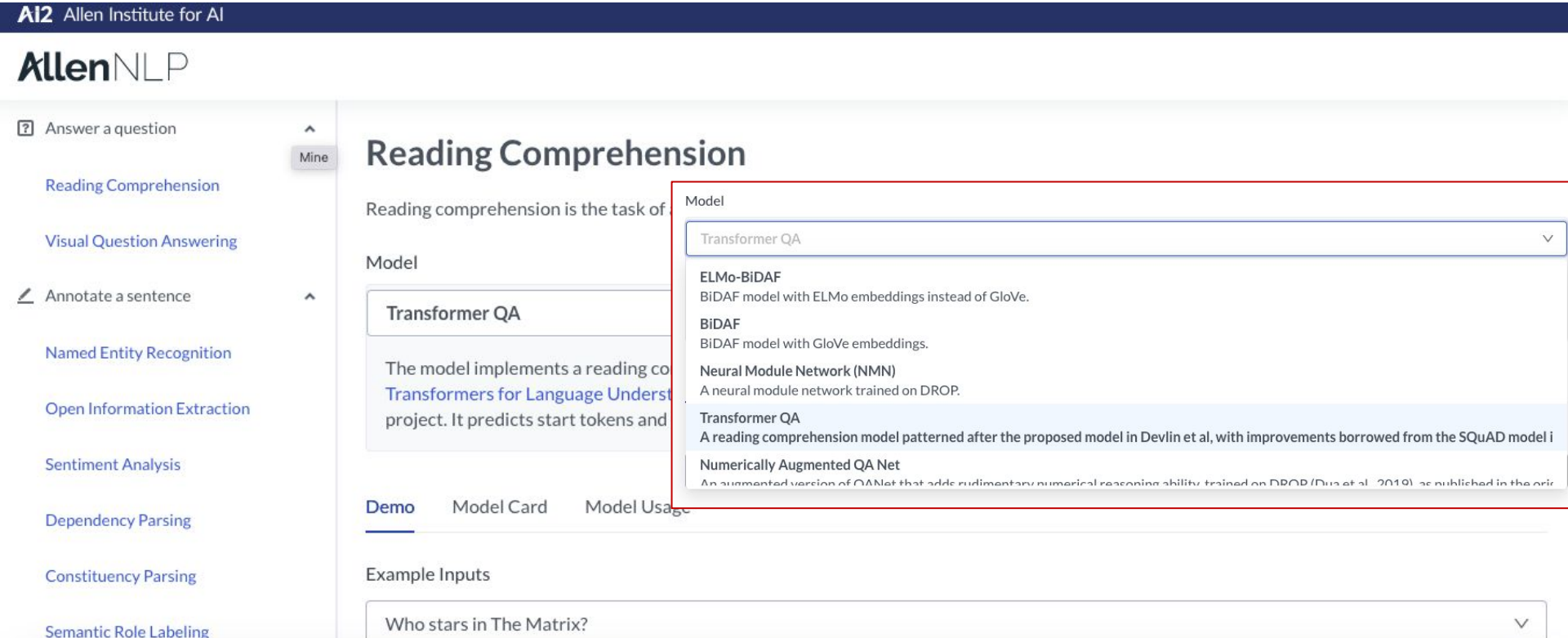
By Martin Coulter and Greg Bensinger

In the advertisement, Bard is given the prompt: "What new discoveries from the James Webb Space Telescope (JWST) can I tell my 9-year old about?" Bard responds with a number of answers, including one suggesting the JWST was used to take the very first pictures of a planet outside the Earth's solar system, or exoplanets. The first pictures of exoplanets were, however, taken by the European Southern Observatory's Very Large Telescope (VLT) in 2004, as **confirmed by NASA**.

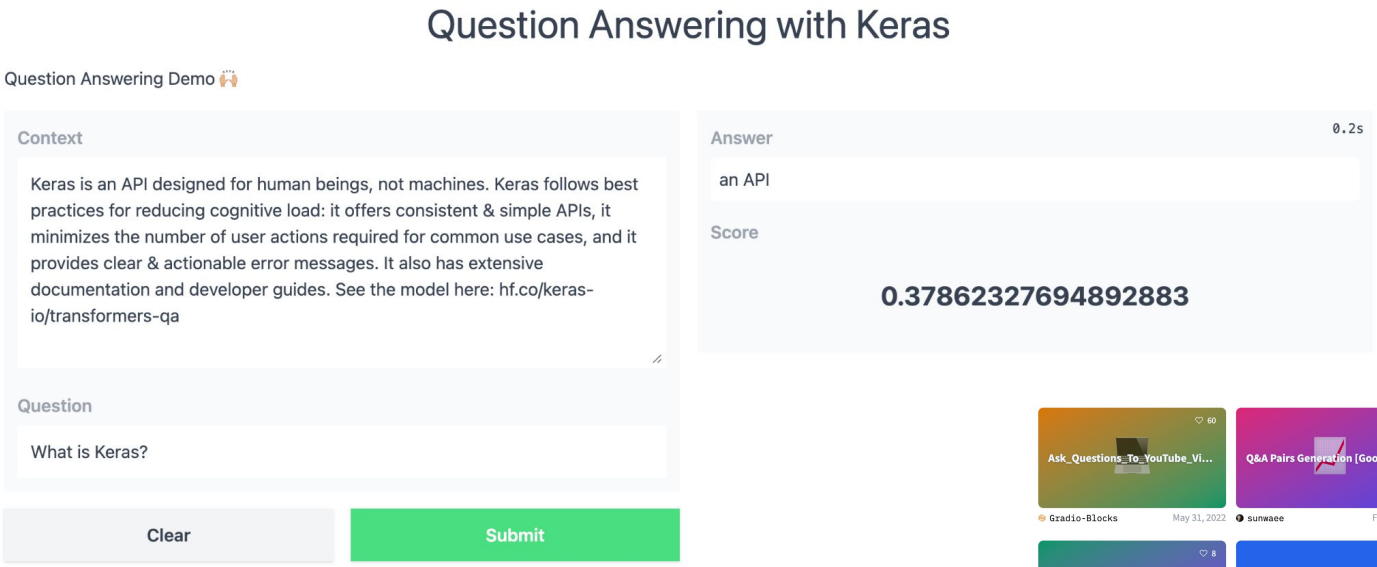
https://www.reuters.com/technology/google-ai-chatbot-bard-offers-inaccurate-information-company-ad-2023-02-08/

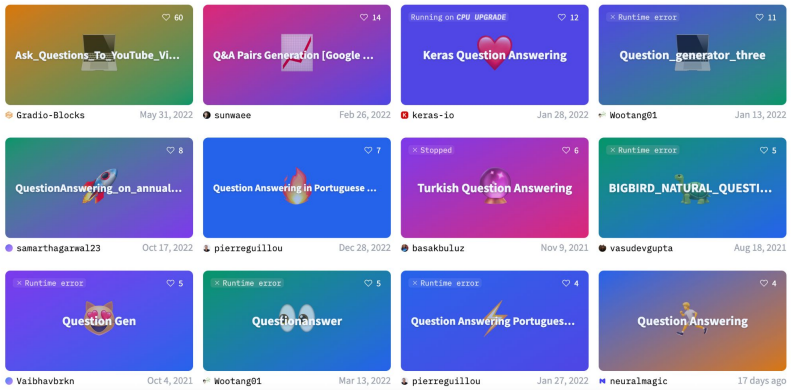
Demo: QA (AllenNLP - outdated)

https://demo.allennlp.org/reading-comprehension



Demo - Huggingface





https://huggingface.co/spaces/keras-io/question_answering

+

Part5) QA data sets (7 data sets)

QA Datasets

Dataset	Answer Type	Size	Domain	Evaluate Ability
ARC(Clark et al., 2018)	Multi-Choice	7,787	Science	Reasoning
BoolQ (Clark et al., 2019)	Bool	16K	Wikipedia	Reasoning
BioASQ (Tsatsaronis et al., 2015)	Span	282	Biomedical	Articles Indexing
CaseHOLD (Zheng et al., 2021)	Multi-Choice	53,137	Law	Pre-training
bABi (Weston et al., 2015)	Bool/Entity	40K	Open Domain	Reasoning
CBT (Hill et al., 2015)	Entity	20K	Children's Book	Model Memory
CliCR (Šuster and Daelemans, 2018)	Entity	105K	Medical	Domain Knowledge
CNN and Daily Mail (See et al., 2017)	Entity	311K	News	Text Summarization
CODAH (Chen et al., 2019)	Multi-choice	4,149	Open Domain	Commonsense
CommonsenseQA (Talmor et al., 2018)	Multi-choice	12,247	ConceptNet	Commonsense
ComplexWebQuestions (Talmor and Berant, 2018)	Entity	34,689	Freebase	Multi-hop
ConditionalQA (Sun et al., 2021)	Entity/Span	9983	Public Policy	Multi-hop
COPA (Gordon et al., 2012)	Multi-choice	1000	Commonsense	Reasoning
CoQA (Reddy et al., 2019)	Entity	127K	Open Domain	Conversation
DROP (Dua et al., 2019)	Span	96K	Wikipedia	Multi-hop
FinQA (Chen et al., 2021)	Number/Span	8,281	Finance	Multi-hop
HotpotQA (Yang et al., 2018)	Entity	113K	Wikipedia	Multi-hop
JD Production QA (Gao et al., 2019b)	Generation	469,953	E-commerce	Domain Knowledge
LogiQA (Liu et al., 2020)	Multi-choice	8,678	Exam	Reasoning
MCTest (Richardson et al., 2013)	Multi-choice	2,000	Fictional Story	Reading Comprehension
Mathematics Dataset (Saxton et al., 2019)	Numeric	2.1×10^6	Mathematics	Calculate
MS MARCO (Nguyen et al., 2016)	Generation	1,010,916	Web pages	Search
MultiRC (Khashabi et al., 2018)	Multi-choice	6K	Multiple Domain	Multi-hop
NarrativeQA (Kočiskỳ et al., 2018)	Span	46,765	Story	Full Document
Natural Questions (Kwiatkowski et al., 2019)	Span/Passage	323,045	Wikipedia	Search
NewsQA (Trischler et al., 2016)	Span	100,000	CNN news	Reading Comprehension
OpenBookQA (Mihaylov et al., 2018)	Multi-choice	6000	Science Facts	Reasoning
PIQA (Bisk et al., 2020)	Multi-choice	21,000	Physical	Physical
PubMedQA (Jin et al., 2019)	Multi-choice	1K	Medical	Summarization
QASPER (Dasigi et al., 2021)	Extractive	5,049	NLP papers	Reasoning
QuAC (Choi et al., 2018)	Multi-choice Generation	100K	Wikipedia	Dialog
QUASAR (Dhingra et al., 2017)	Span	43,000	StackOverflow/Trivia	search
RACE (Lai et al., 2017)	Multi-choice	100,000	Exam	Reading Comprehension
ReClor (Yu et al., 2020)	Multi-choice	6138	Exam	Logical
SCDE (Kong et al., 2020)	Exam	6K	Exam	Reading Comprehension
SimpleQuestions (Bordes et al., 2015)	Entity	100K	Freebase	Knowledge
SQuAD (Rajpurkar et al., 2016, 2018)	Span	130,319	Wikipedia	Reading Comprehension
TriviaQA (Joshi et al., 2017)	Span	650K	Open Domain	Reading Comprehension
TweetQA (Xiong et al., 2019)	Generation	13,757	Tweet	Reading Comprehension
WikiHop (Welbl et al., 2018)	Multi-choice	51,318	Wikipedia	Multi-hop
WikiQA (Yang et al., 2015)	Sentence	3,047	Wikipedia	Reading Comprehension

Table 1: Statistics of textual QA datasets.

Modern Question Answering Datasets and Benchmarks: A Survey. Wang, 2022

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

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

Figure 1: Question-answer pairs with sample excerpts from evidence documents from TriviaQA exhibiting lexical and syntactic variability, and requiring reasoning from multiple sentences.

triviaqa

Dataset and Predominant Techniques

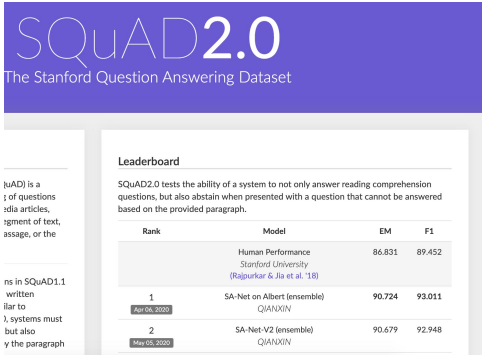
Year	Tasks	Corpus Type	Question Type	Answer Source	Answer Type
2018	SQuAD 2.0	Textual	Natural	Spans	Natural
2019	Natural Questions	Textual	Natural	Spans	Natural
2018	CoQA	Textual	Natural	Free-form	Natural
2017	TriviaQA	Textual	Natural	Free-form	Natural
2019	RACE	Textual	Natural	Free-form	Multiple-choice
2018	RecipeQA	Multi-modal	Natural	Cloze/free-form	Multiple-choice
2019	NSC (Thai)	Textual	Natural	Spans	Natural

1) SQuAD 2.0

Year	Tasks	Corpus Type	Question Type	Answer Source	Answer Type
2018	SQuAD 2.0	Textual	Natural	Spans	Natural

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

- Extend from SQuAD 1.0
- Crowdsource from Wikipedia paragraph (let worker create question from articles)
- Has two types of questions
  - Answerable (in SQuAD 1.0)
  - Unanswerables (only in SQuAD 2.0)
- Arguably, one of the most popular and well-known MRC benchmark.
- Top of the leader boards are dominated by variations of pretrained LMs



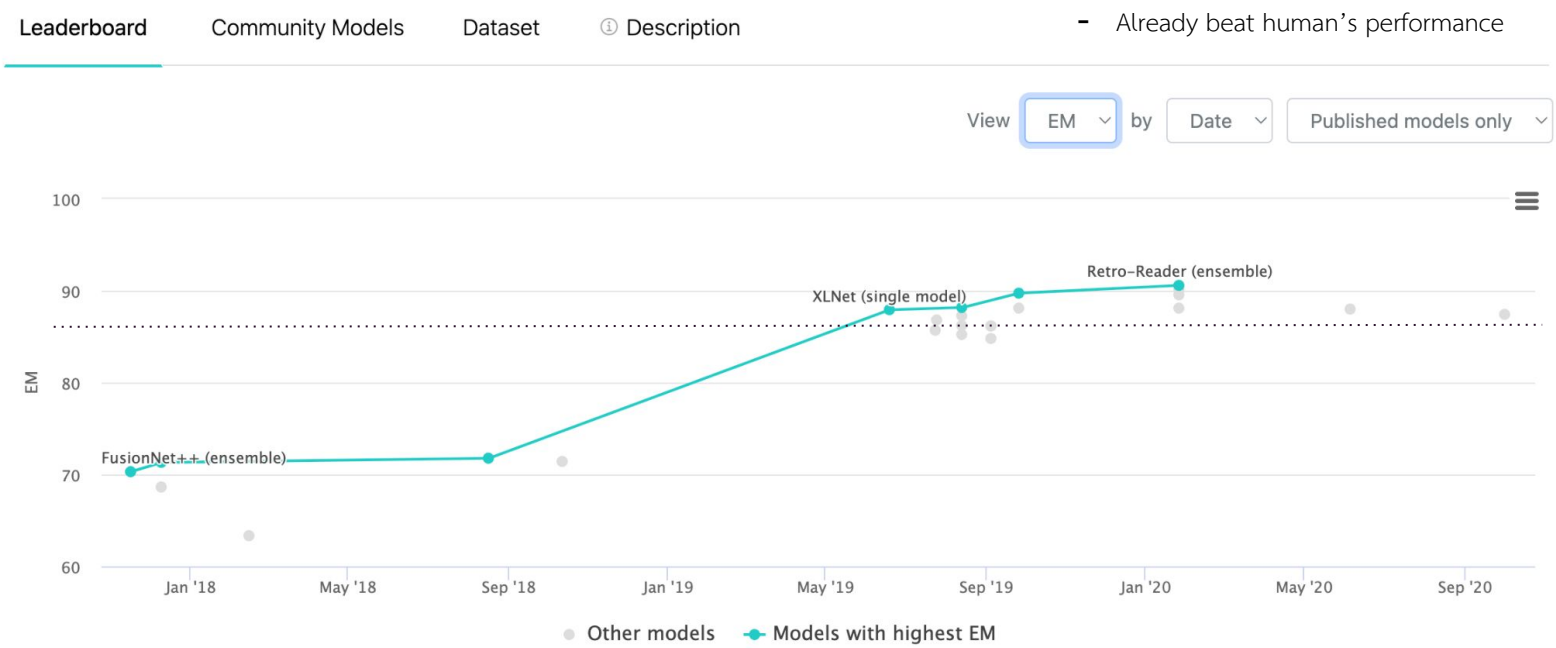
Year	Tasks	Corpus Type	Question Type	Answer Source	Answer Type
2018	SQuAD 2.0	Textual	Natural	Spans	Natural

1) SQuAD 2.0

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

- Leaderboard

- Very saturated
- Already beat human's performance



1) SQuAD 2.0 – RetroReader, (Zhang et al., 2020)

https://arxiv.org/abs/2001.09694

- Pick RetroReader for the discussion, this model is built upon pretrained LM architecture rather than just fine-tuning pretrained LM on the MRC task.
- Has external verification and internal verification
  - Sketchy reader: contains answer or not
  - Intensive reader: find answer spans as well as answerability
- Interesting choice of design of having to answer verification (check of context passage contain answer or not) in both of the reader modules

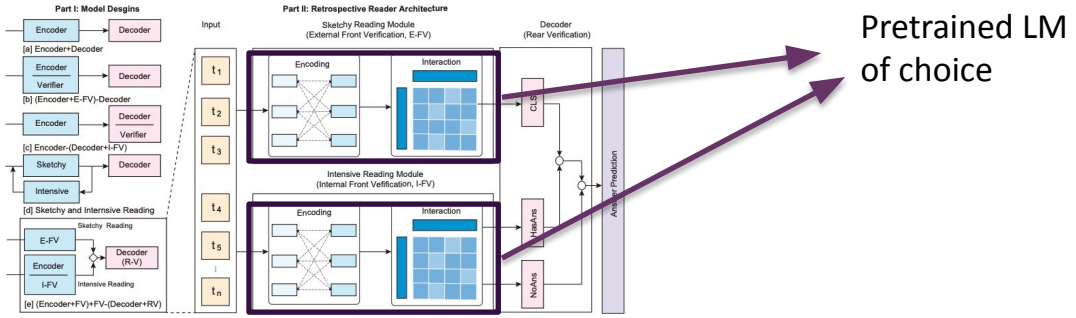


Figure 1: Reader overview. For the left part, models [a-c] summarize the instances in previous work, and model [d] is ours, with the implemented version [e]. In the names of models [a-e], "(·)" represents a module, "+" means the parallel module and "-" is the pipeline. The right part is the detailed architecture of our proposed Retro-Reader.

2) Natural Questions (NQ)

https://ai.google.com/research/NaturalQuestions/

Year	Tasks	Corpus Type	Question Type	Answer Source	Answer Type
2019	Natural	Textual	Natural	Spans	Natural

- Based on Wikipedia article, context passage consists of 5 top Wikipedia article queried based on the natural question
- Has 2 types of tasks, long and short answers
  - Long answer, find the paragraph that contains the answer
  - Short answer, find answer if present in the document
- Some questions are also unanswerable (but only in small percentage)

Question:

when are hops added to the brewing process?

Short Answer:

The boiling process

Long Answer:

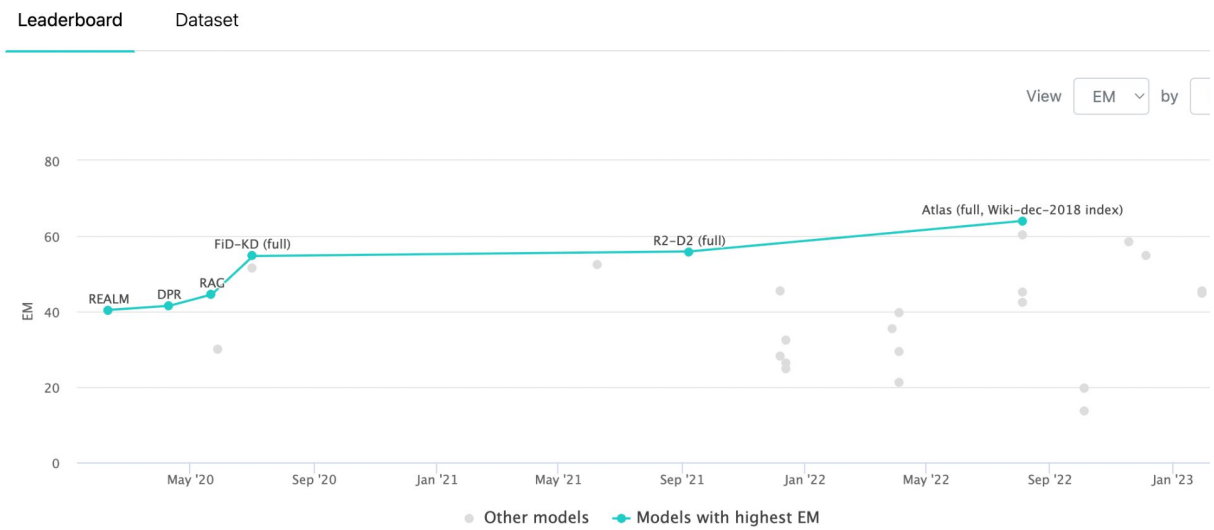
After mashing , the beer wort is boiled with hops (and other flavourings if used) in a large tank known as a " copper " or brew kettle – though historically the mash vessel was used and is still in some small breweries . The boiling process is where chemical reactions take place , including sterilization of the wort to remove unwanted bacteria , releasing of hop flavours , bitterness and aroma compounds through isomerization , stopping of enzymatic processes , precipitation of proteins , and concentration of the wort . Finally , the vapours produced during the boil

2) Natural Questions

https://ai.google.com/research/NaturalQuestions/

Year	Tasks	Corpus Type	Question Type	Answer Source	Answer Type
2019	Natural	Textual	Natural	Spans	Natural

- Leaderboard **March-2023**



Natural Questions (Kwiatkowski et al., 2019) is a corpus of real questions issued to the Google search engine. Each question comes with an accompanied Wikipedia page with an annotated long answer (ne or more enti of the dataset and use both long and short answers spans as *provenance*. We collaborated with the authors of Natural Questions to access a held out, unpublished portion of the original dataset to form a new test set for KILT. By construction each QA pair is associated with a single Wikipedia page, although other pages might contain enough evidence to answer the question. To increase the provenance coverage we perform an Amazon Mechanical Turk campaign for the dev and test sets and increase the average number of provenance pages per question from 1 to 1.57 (details in section 4).

Results on Page 24

3) CoQA

Year	Tasks	Corpus Type	Question Type	Answer Source	Answer Type
2018	CoQA	Textual	Natural	Free-form	Natural

https://stanfordnlp.github.io/coqa/

- One of the first conversational MRC dataset available, mimic process of 2 people discussing the context passage as a topic
- Topics are from 7 sources including News, Wikipedia, and Reddit.
- Contains: Yes, no, unanswerable question type. So Answers are not guaranteed to be found in the passage
- Also has rationale label for each question

(Metadata)

In some questions, information can be found from the previous conservation turn.

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

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

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

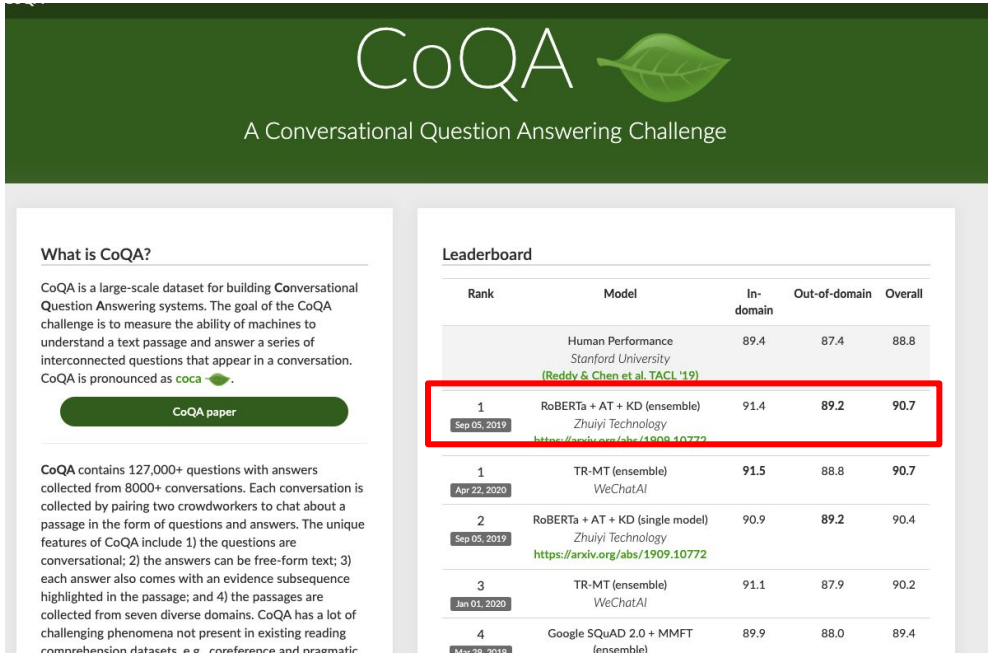
Q_3: Did she plan to have any visitors?
A_3: Yes
R_3: Her granddaughter Annie was coming over

3) CoQA

https://stanfordnlp.github.io/coqa/

Year	Tasks	Corpus Type	Question Type	Answer Source	Answer Type
2018	CoQA	Textual	Natural	Free-form	Natural

- Leaderboard



-Top of leaderboard RoBERTa enhanced with other techniques

Year	Tasks	Corpus Type	Question Type	Answer Source	Answer Type
2017	TriviaQA	Textual	Natural	Free-form	Natural

4) TriviaQA

http://nlp.cs.washington.edu/triviaqa/

- Questions are curated from trivia questions website, then paired the questions with closest matched document later (use search engine)
- Since question are not crafted from the document,
  - not all questions are guaranteed to have answer, and
  - The answers that are found in the context passage might not exactly match with the semantic of the question

- Has **2 versions** of the dataset:
  - the one that is **matched with the document**
  - **Open-Domain QA** where questions are not matched with the article

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

Year	Tasks	Corpus Type	Question Type	Answer Source	Answer Type
2017	TriviaQA	Textual	Natural	Free-form	Natural

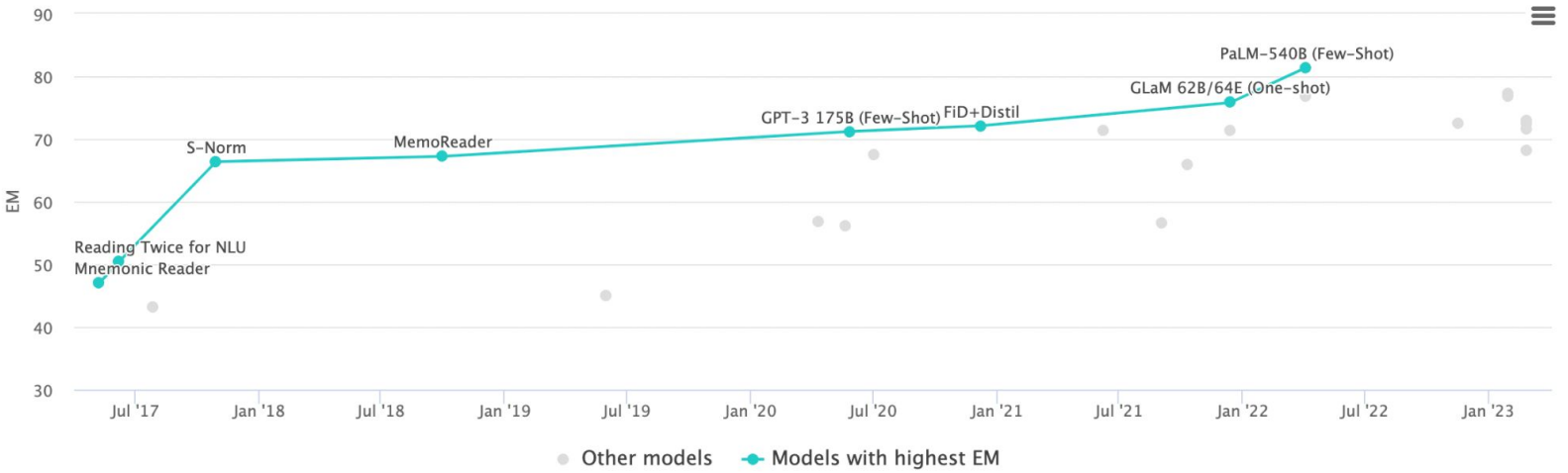
4) TriviaQA

http://nlp.cs.washington.edu/triviaqa/

Question Answering on TriviaQA March-2023

Leaderboard Dataset

View EM by Date for All models



5) RACE

https://www.cs.cmu.edu/~glai1/data/race/

Year	Tasks	Corpus Type	Question Type	Answer Source	Answer Type
2019	Textual	Natural	Free-form	Multiple-choice	Natural

- Dataset collected from English examination for Chinese students
- Many questions require reasoning ability

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

Questions:

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

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

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

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

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

Answer: ADABC

Table 1: Sample reading comprehension problems from our dataset.

Year	Tasks	Corpus Type	Question Type	Answer Source	Answer Type
2019	Textual	Natural	Free-form	Multiple-choice	Natural

5) RACE

https://www.cs.cmu.edu/~glai1/data/race/

- Leaderboard
- We can see that, again, the leaderboard

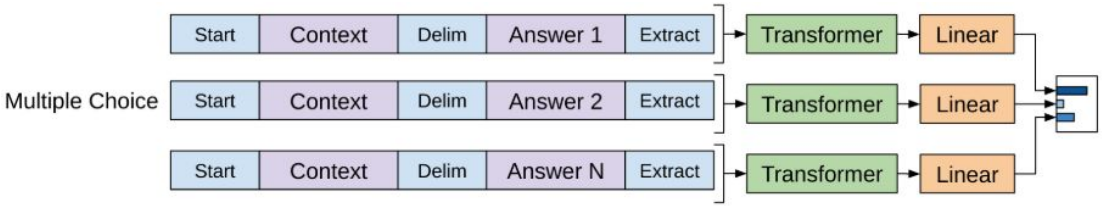
Is dominated by the methods that have pretrain

language models in its name.

Leaderboard

Model	Report Time	Institute	RACE	RACE-M	RACE-H
Human Ceiling Performance	Apr 15, 2017	CMU	*94.5*	*95.4*	*94.2*
Amazon Mechanical Turker	Apr 15, 2017	CMU	73.3	85.1	69.4
ALBERT-SingleChoice + transfer learning (ensemble)	Nov 06, 2020	Tencent Cloud Xiaowei & Tencent Cloud TI-ONE	**91.4**	**93.6**	**90.5**
Megatron-BERT (ensemble)	Mar 13, 2020	NVIDIA Research	90.9	93.1	90.0
ALBERT-SingleChoice + transfer learning	Nov 06, 2020	Tencent Cloud Xiaowei & Tencent Cloud TI-ONE	90.7	92.8	89.8
ALBERT + DUMA (ensemble)	Mar 18, 2020	SJTU & Huawei Noah's Ark Lab	89.8	92.6	88.7
Megatron-BERT	Mar 13, 2020	NVIDIA Research	89.5	91.8	88.6
ALBERT (ensemble)	Sep 26, 2019	Google Research & TTIC	89.4	91.2	88.6
UnifiedQA	May 02, 2020	AI2 & UW	89.4	-	-
ALBERT + DUMA	Feb 08, 2020	SJTU & Huawei Noah's Ark Lab	88.0	90.9	86.7

- Calculate the score for each answer candidate separately
- https://openai.com/blog/language-unsupervised/

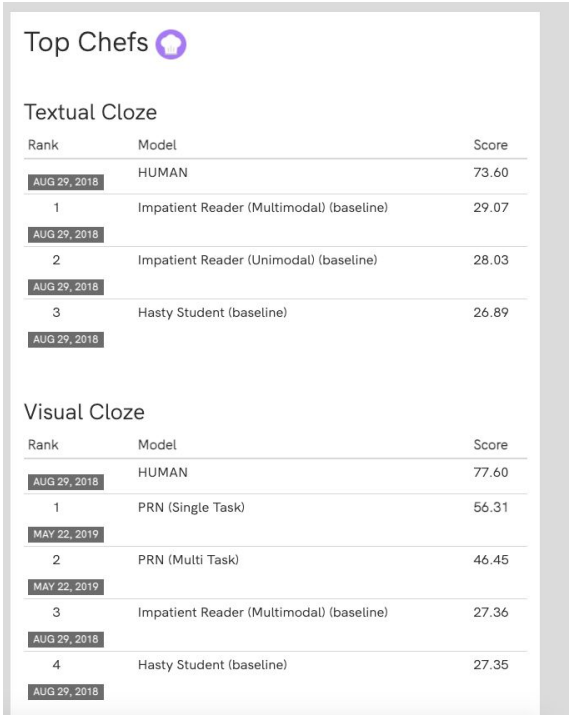


Year	Tasks	Corpus Type	Question Type	Answer Source	Answer Type
2018	RecipeQA	Multi-modal	Natural	Cloze/free-form	Multiple-choice

6) RecipeQA

https://hucvl.github.io/recipeqa/

- Interestingly, the leaderboard has minimal update since the first time the author of this slide visit the dataset webpage (since Q4 2018)
- A potential blue ocean.



Top Chefs

Textual Cloze

Rank	Model	Score
AUG 29, 2018	HUMAN	73.60
1 AUG 29, 2018	Impatient Reader (Multimodal) (baseline)	29.07
2 AUG 29, 2018	Impatient Reader (Unimodal) (baseline)	28.03
3 AUG 29, 2018	Hasty Student (baseline)	26.89

Visual Cloze

Rank	Model	Score
AUG 29, 2018	HUMAN	77.60
1 MAY 22, 2019	PRN (Single Task)	56.31
2 MAY 22, 2019	PRN (Multi Task)	46.45
3 AUG 29, 2018	Impatient Reader (Multimodal) (baseline)	27.36
4 AUG 29, 2018	Hasty Student (baseline)	27.35

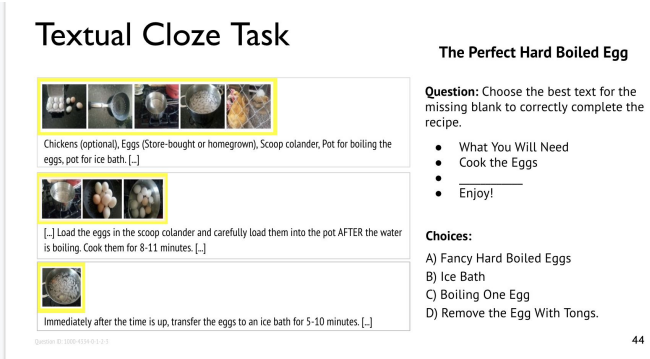
Year	Tasks	Corpus Type	Question Type	Answer Source	Answer Type
2018	RecipeQA	Multi-modal	Natural	Cloze/free-form	Multiple-choice

6) RecipeQA

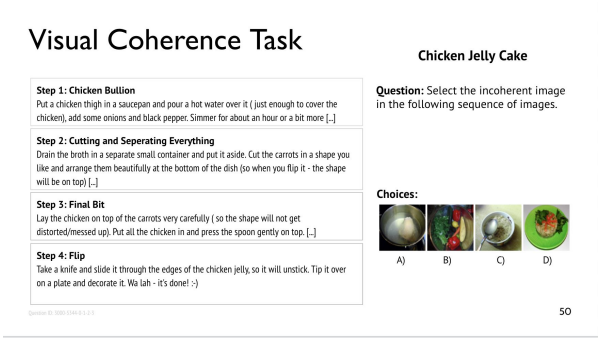
https://hucvl.github.io/recipeqa/

https://docs.google.com/presentation/d/1mvuu4QTfOP6CHUfbLdXMisleiH7klxpsU1bgyT0oTw/edit#slide=id.g468e9cafc5_0_281

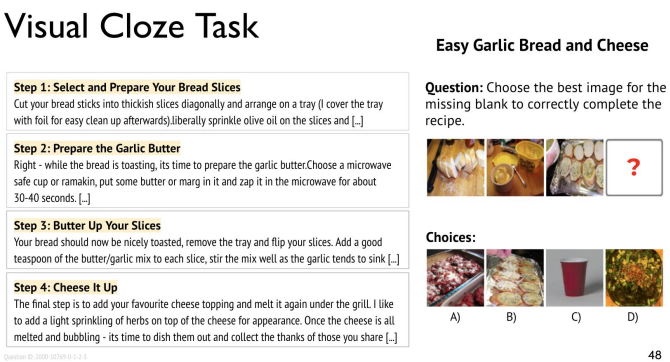
Textual Cloze Task



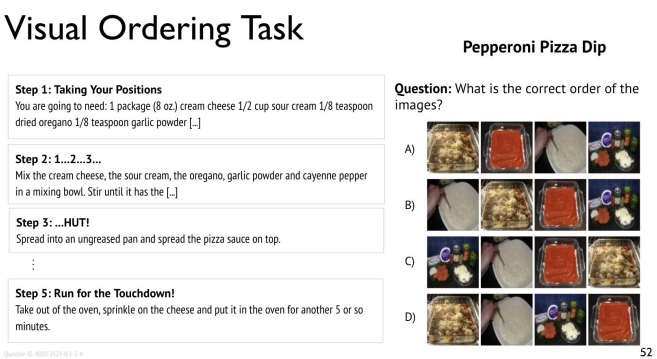
Visual Coherence Task



Visual Cloze Task



Visual Ordering Task



7) NSC (Thai QA)

Year	Tasks	Corpus Type	Question Type	Answer Source	Answer Type
2019	NSC (Thai)	Textual	Natural	Spans	Natural

http://copycatch.in.th/thai-qa-task.html

- Thai Question Answering Program competition hosted in NSC by NECTEC
- There were 2 round of competitions
  - First round: 4000 factoid (span extraction) questions
  - Second round: 15000 factoid questions and 2000 yes-no questions
- An Open-domain question answering problem: the program must also query for the context passage.
- Only the first round of the competition dataset went public

8) iApp (Thai QA)

- Another Thai QA dataset is IAPP wiki QA (2021)
  - https://github.com/iapp-technology/iapp-wiki-qa-dataset
- Thai Wikipedia Question Answering Dataset.
  - 1,961 Documents
  - 9,170 Questions
  - It is organized and formatted in the SQuAD format
- Demo: https://ai.iapp.co.th/control/ai