

ONCE UPON A TIME IN THE...



ONCE UPON A TIME IN THE
Cozy Afternoon at Masaryk University



ONCE UPON A TIME IN THE
Cozy Afternoon at Masaryk University

QUESTION ANSWERING AND BEYOND



PROLOGUE:

PROLOGUE:
AN UNKNOWN VISITOR

Whoami



1. **Senior PhD** student from **BUT-FIT** supervised by prof. Smrž.
2. A person fond of **question answering, fact checking** and basically **any open-domain retrieval problem :-)**.

Web, bio, more info:
<https://mfajcik.github.io/>

CHAPTER 1:
INTRODUCTION

Information Need



Information Need

? Information Need



Information Need

? Information Need



When has X his birthday?
I suffer from Y every winter. How to prevent it?
Where to buy skiing equipment?

Information Need

? Information Need

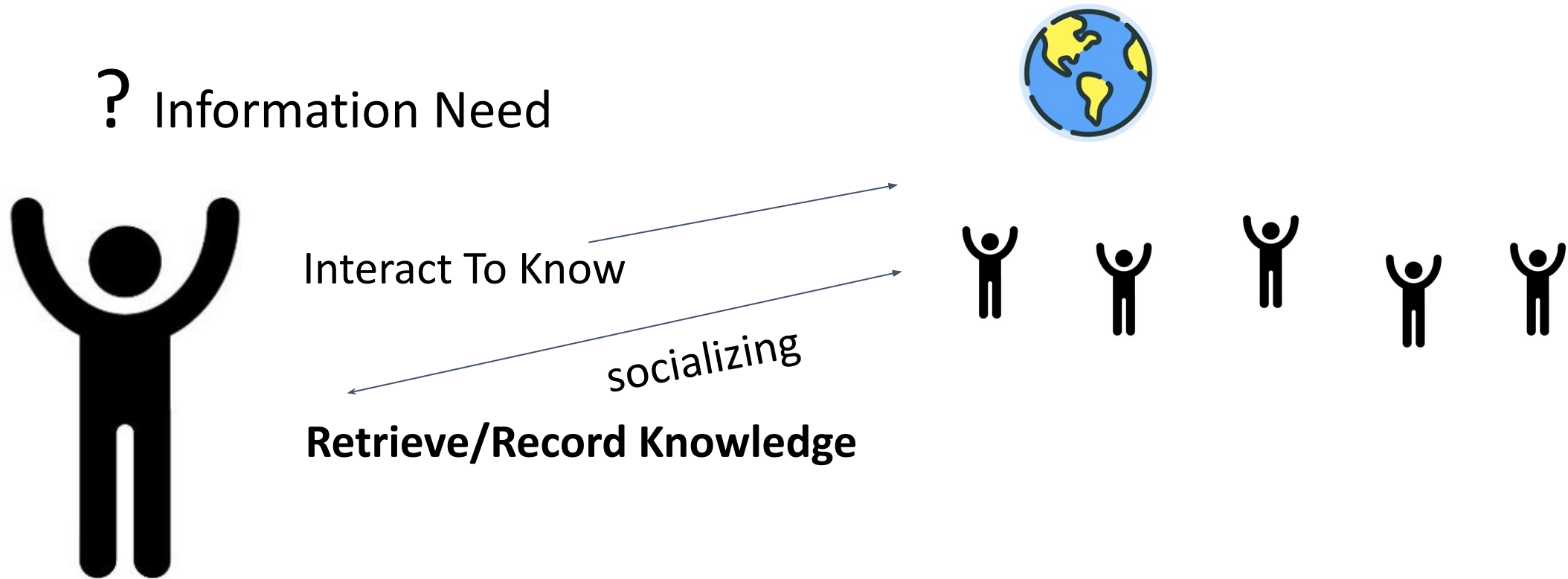


Interact with world
To Know



- When has X his birthday?
- I suffer from Y every winter. How to prevent it?
- Where to buy skiing equipment?
- What is the information need?

Information Need

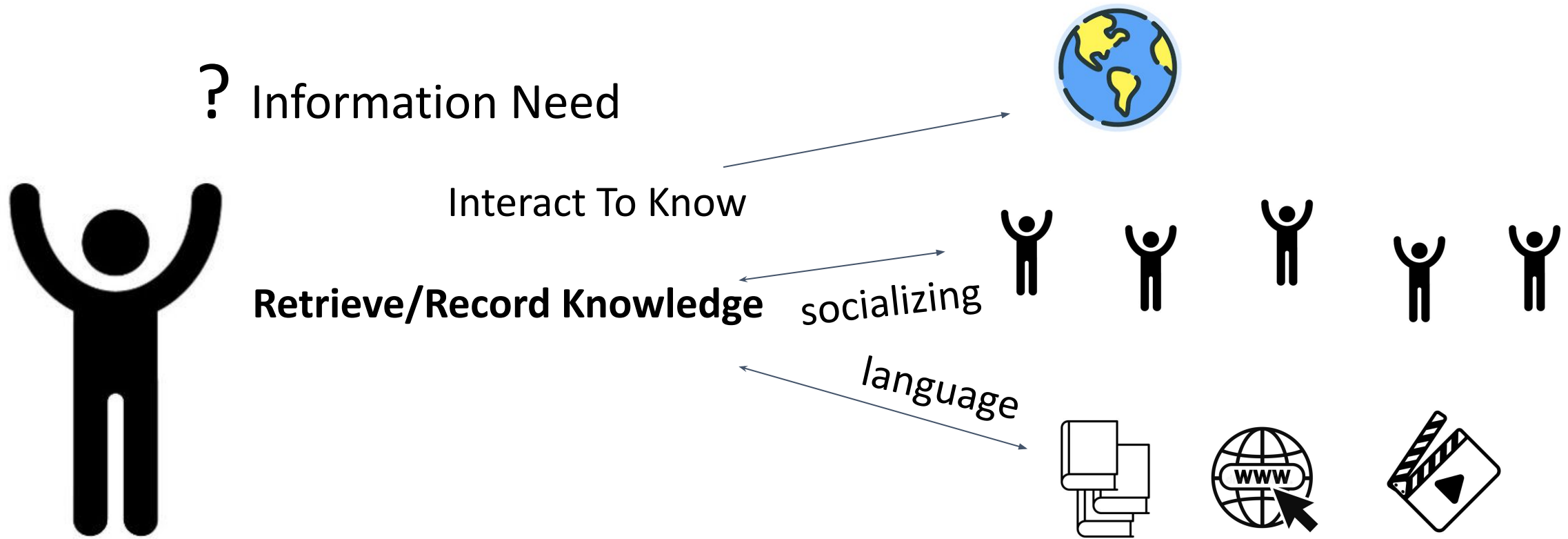


When has X his birthday?

I suffer from Y every winter. How to prevent it?

Where to buy skiing equipment?

Information Need



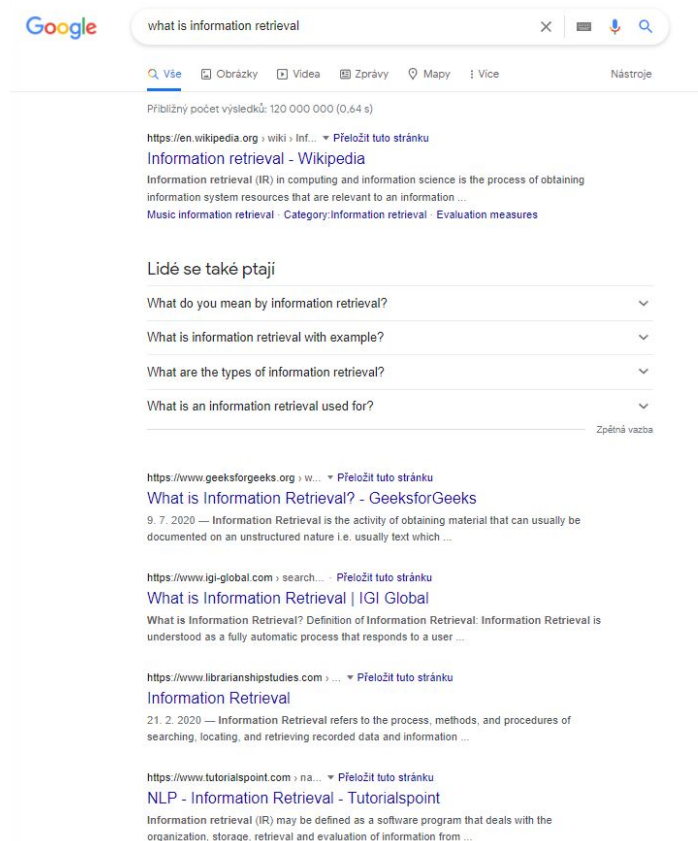
When has X his birthday?

I suffer from Y every winter. How to prevent it?

Where to buy skiing equipment?

Information Need

Traditional Information Retrieval Today



Research Desiderata

- Provide **answer**, if question requires factoid answer
- Provide **summary**, if question requires summary
- Provide **search result** if question requires listing
- Solve logic, if question requires **problem solving**
- Questions are often ambiguous, **disambiguate** via interaction
- Make models understand natural **language**, not humans learn model language

CHAPTER 2:
INFORMATION RETRIEVAL

Information Retrieval (IR)

Term “**Information Retrieval**” in literature. Example from MSMarco (Nguyen et al. 2016)

Q Will I qualify for OSAP if I'm new in Canada?

Selected Passages from Bing

“Visit the OSAP website for application deadlines. To get OSAP, you have to be eligible. You can apply using an online form, or you can print off the application forms. If you submit a paper application, you must pay an application fee. The online application is free.”
Source: <http://settlement.org/ontario/education/colleges-universities-and-institutes/financial-assistance-for-post-secondary-education/how-do-i-apply-for-the-ontario-student-assistance-program-osap/>

“To be eligible to apply for financial assistance from the Ontario Student Assistance Program (OSAP), you must be a: 1 Canadian citizen; 2 Permanent resident; or 3 Protected person/convention refugee with a Protected Persons Status Document (PPSD).”
Source: <http://settlement.org/ontario/education/colleges-universities-and-institutes/financial-assistance-for-post-secondary-education/who-is-eligible-for-the-ontario-student-assistance-program-osap/>

“You will not be eligible for a Canada-Ontario Integrated Student Loan, but can apply for a part-time loan through the Canada Student Loans program. There are also grants, bursaries and scholarships available for both full-time and part-time students.”
Source: <http://www.campusaccess.com/financial-aid/osap.html>

Query (often a list of keywords)



Task: distinguish between relevant/irrelevant documents

Information Retrieval (IR)

Term “**Information Retrieval**” in literature. Example from MSMarco (Nguyen et al. 2016)

Q Will I qualify for OSAP if I'm new in Canada?

Selected Passages from Bing

“Visit the OSAP website for application deadlines. To get OSAP, you have to be eligible. You can apply using an online form, or you can print off the application forms. If you submit a paper application, you must pay an application fee. The online application is free.”
Source: <http://settlement.org/ontario/education/colleges-universities-and-institutes/financial-assistance-for-post-secondary-education/how-do-i-apply-for-the-ontario-student-assistance-program-osap/>

“To be eligible to apply for financial assistance from the Ontario Student Assistance Program (OSAP), you must be a: 1 Canadian citizen; 2 Permanent resident; or 3 Protected person/convention refugee with a Protected Persons Status Document (PPSD).”
Source: <http://settlement.org/ontario/education/colleges-universities-and-institutes/financial-assistance-for-post-secondary-education/who-is-eligible-for-the-ontario-student-assistance-program-osap/>

“You will not be eligible for a Canada-Ontario Integrated Student Loan, but can apply for a part-time loan through the Canada Student Loans program. There are also grants, bursaries and scholarships available for both full-time and part-time students.”
Source: <http://www.campusaccess.com/financial-aid/osap.html>

Query (often a list of keywords)



Task: distinguish between relevant/irrelevant documents



The labels can be non-binary (relevance scores)



Is Information Retrieval Document Retrieval?

- Lets brainstorm, how else can we retrieve information?

Question Answering (QA)

- A set of problems related to drawing conclusions from data (example from MSMarco)

Q Will I qualify for OSAP if I'm new in Canada?

Selected Passages from Bing

"Visit the OSAP website for application deadlines. To get OSAP, you have to be eligible. You can apply using an online form, or you can print off the application forms. If you submit a paper application, you must pay an application fee. The online application is free."
Source: <http://settlement.org/ontario/education/colleges-universities-and-institutes/financial-assistance-for-post-secondary-education/how-do-i-apply-for-the-ontario-student-assistance-program-osap/>

"To be eligible to apply for financial assistance from the Ontario Student Assistance Program (OSAP), you must be a: 1 Canadian citizen; 2 Permanent resident; or 3 Protected person/convention refugee with a Protected Persons Status Document (PPSD)."
Source: <http://settlement.org/ontario/education/colleges-universities-and-institutes/financial-assistance-for-post-secondary-education/who-is-eligible-for-the-ontario-student-assistance-program-osap/>

"You will not be eligible for a Canada-Ontario Integrated Student Loan, but can apply for a part-time loan through the Canada Student Loans program. There are also grants, bursaries and scholarships available for both full-time and part-time students."
Source: <http://www.campusaccess.com/financial-aid/osap.html>

Answer
No. You won't qualify.

← Question (in natural language)

← Provided document(s)

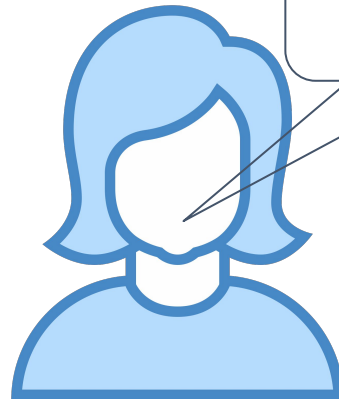
← Task: provide Answer



Question Answering (QA)

What is question asking about?

- facts?
- open-ended?
- chit-chat?
- math?
- multi-answer/multihop question?



Who is Jožko Mrkvička?



A fictional character in colloquial Slovak, whose name is used to denote an ordinary average citizen

What answer is expected?

- factoid?
- open-ended?
- chit-chat?
- CODE?
- no-answer?
- respond with clarifying question to ambiguous question?

Extractive QA

The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, *Il milione* (or, *The Million*, known in English as the *Travels of Marco Polo*), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge **through contact with Persian traders** since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?

Answer: through contact with Persian traders

Exact Match measures the percentage of predictions that match at least one of the ground truth answers exactly

$$\frac{\text{\# of correctly answered questions}}{\text{\# of all questions}}$$

Extractive QA

The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, *Il milione* (or, *The Million*, known in English as the *Travels of Marco Polo*), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge **through contact with Persian traders** since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?

Answer: through contact with Persian traders

(macro)**F1** measures the average overlap between the prediction and ground truth answer.

Prediction and ground truth are treated as bags of tokens and their F1 is computed.

Usually a maximum F1 over all of the ground truth answers for a given question is taken, and the result is an average over all of the questions.

Extractive QA

The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, *Il milione* (or, *The Million*, known in English as the *Travels of Marco Polo*), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge **through contact with Persian traders** since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?

Answer: through contact with Persian traders

(macro)**F1** measures the average overlap between the prediction and ground truth answer.

Prediction and ground truth are treated as bags of tokens and their F1 is computed.

Usually a maximum F1 over all of the ground truth answers for a given question is taken, and the result is an average over all of the questions.

- Ground truth answer = {cat named kitty}

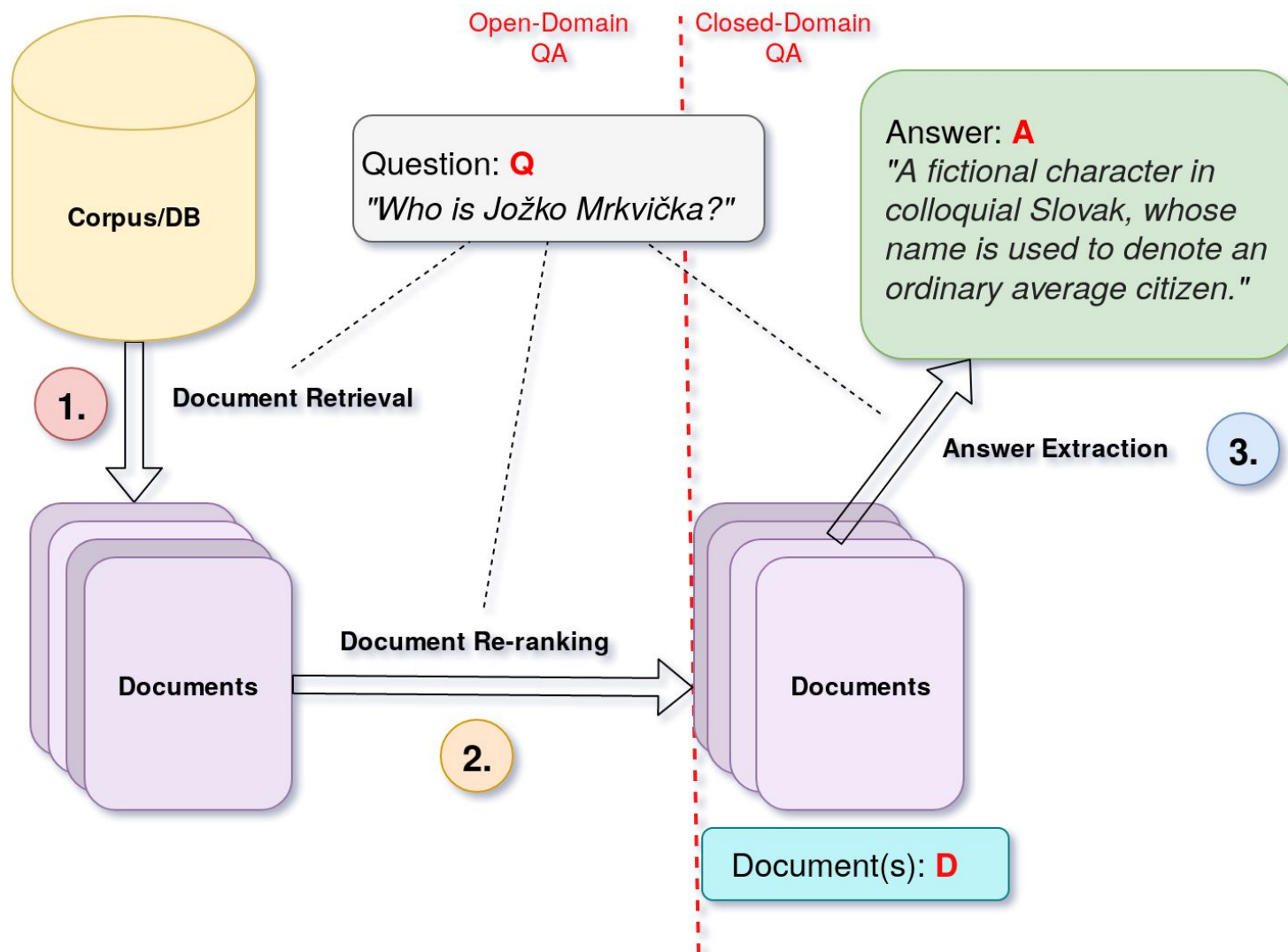
- Predicted answer = {cat named }

- $precision = \frac{\# \text{ of same tokens}}{\text{length of prediction}} = \frac{2}{2}$

- $recall = \frac{\# \text{ of same tokens}}{\text{length of ground truth}} = \frac{2}{3}$

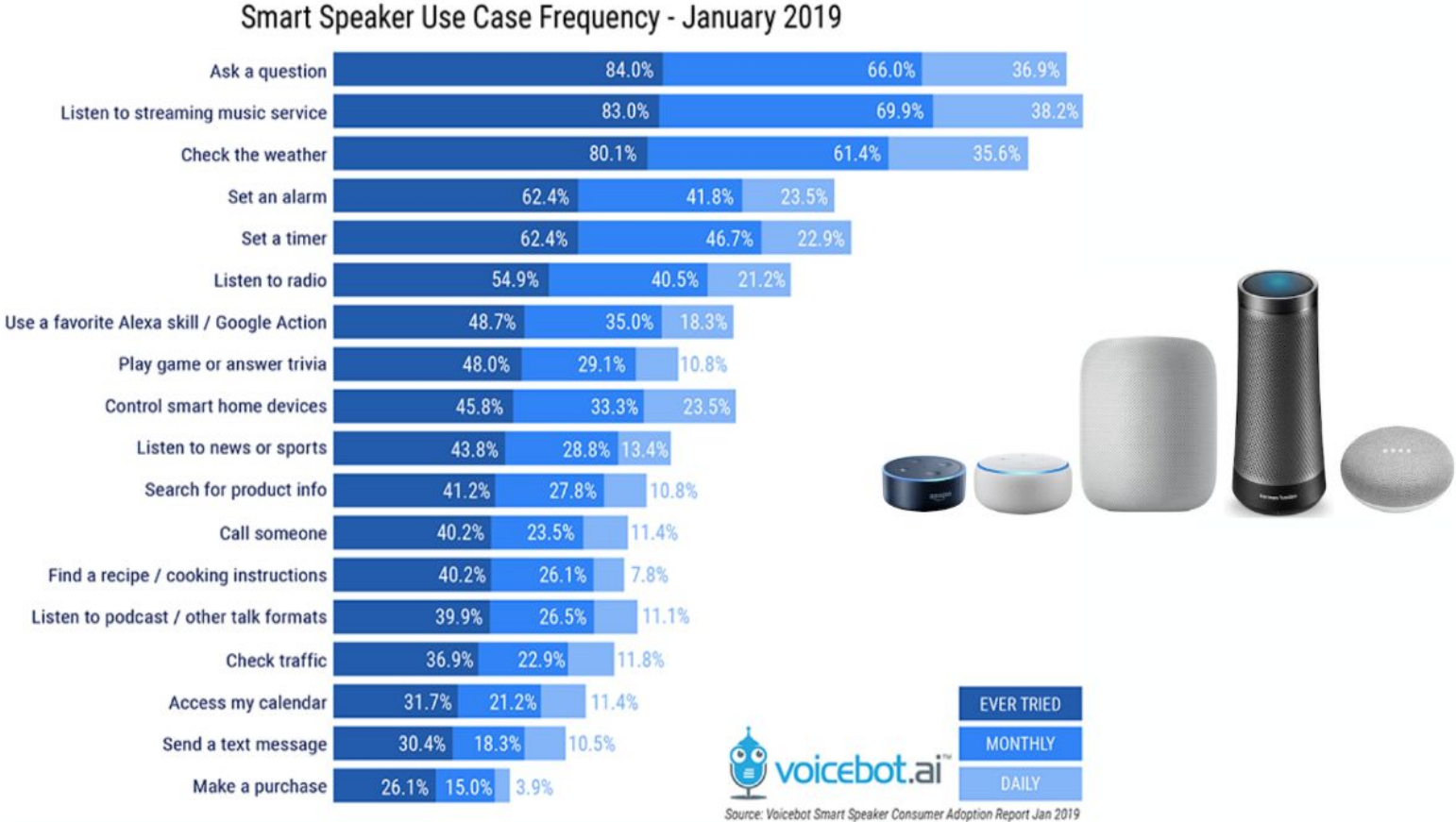
- $F1 = 2 * \frac{precision * recall}{precision + recall} = 2 * \frac{\frac{2}{2} * \frac{2}{3}}{\frac{2}{2} + \frac{2}{3}}$

When Document Retrieval meets QA

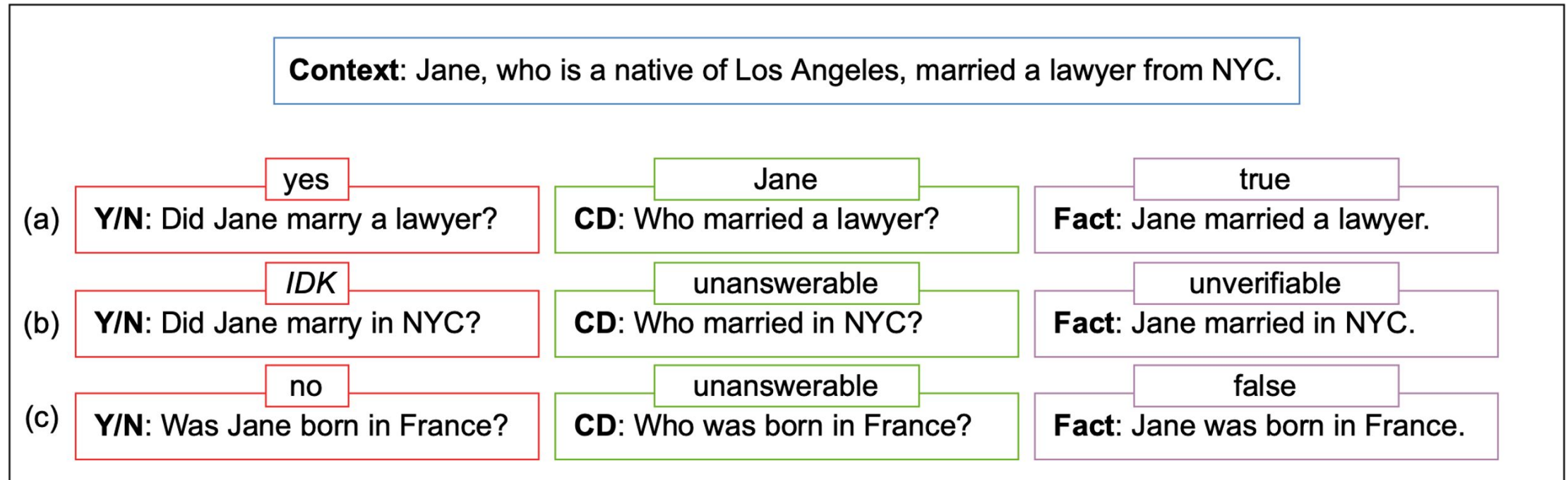


Open-domain QA

Brief Business Motivation



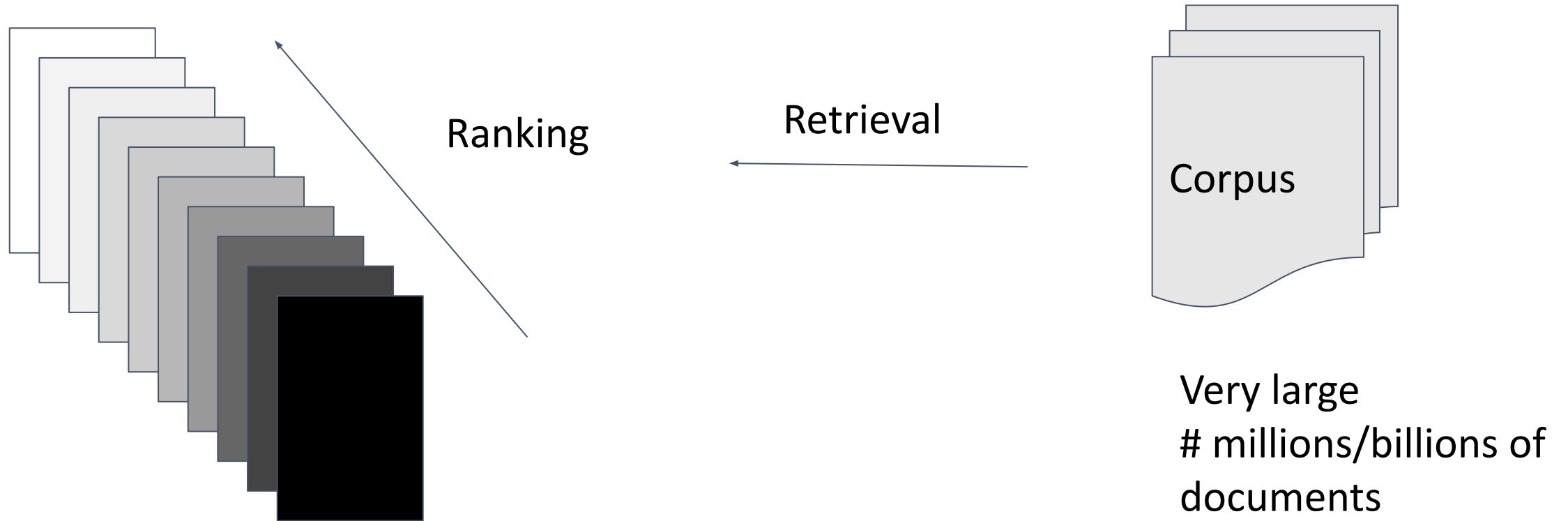
QA vs Fact-Checking



Yes/No Question (Y/N), Closed-domain Extractive QA (CD), A fact to be verified (FACT)

CHAPTER 3:
INTRODUCTION INTO BMT25

Retrieval



Retrieval via TF-IDF

Standard TF-IDF works reasonably well for retrieval!

Term frequency		Document frequency		Normalization	
n (natural)	$tf_{t,d}$	n (no)	1	n (none)	1
l (logarithm)	$1 + \log(tf_{t,d})$	t (idf)	$\log \frac{N}{df_t}$	c (cosine)	$\frac{1}{\sqrt{w_1^2 + w_2^2 + \dots + w_M^2}}$
a (augmented)	$0.5 + \frac{0.5 \times tf_{t,d}}{\max_t(tf_{t,d})}$	p (prob idf)	$\max\{0, \log \frac{N - df_t}{df_t}\}$	u (pivoted unique)	$1/u$ (Section 6.4.4)
b (boolean)	$\begin{cases} 1 & \text{if } tf_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$			b (byte size)	$1/CharLength^\alpha, \alpha < 1$
L (log ave)	$\frac{1 + \log(tf_{t,d})}{1 + \log(\text{ave}_{t \in d}(tf_{t,d}))}$				

► **Figure 6.7** SMART notation for tf-idf variants. Here *CharLength* is the number of characters in the document.

Retrieval via TF-IDF

For query $Q := q_1 q_2 \dots q_n$ and document $D := w_1 w_2 \dots w_n$ we compute the score from overlapping terms as follows:

$$Score(D, Q) = \sum_{q=w; q \in Q, w \in D} tf_{q,Q} idf_w \cdot tf_{w,D} idf_w$$

Why
1?

Term frequency	
n (natural)	$tf_{t,d}$
1 (logarithm)	$1 + \log(tf_{t,d})$
a (augmented)	$0.5 + \frac{0.5 \times tf_{t,d}}{\max_t(tf_{t,d})}$
b (boolean)	$\begin{cases} 1 & \text{if } tf_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$
L (log ave)	$\frac{1 + \log(tf_{t,d})}{1 + \log(\text{ave}_{t \in d}(tf_{t,d}))}$

$$1 + \log_2(10) = 4.32$$

$$1 + \log_2(100) = 7.64$$

same quantity

Retrieval via TF-IDF

For query $Q := q_1 q_2 \dots q_n$ and document $D := w_1 w_2 \dots w_n$ we compute the score from overlapping terms as follows:

$$Score(D, Q) = \sum_{q=w; q \in Q, w \in D} tf_{q,Q} idf_w \cdot tf_{w,D} idf_w$$

Term frequency	
n (natural)	$tf_{t,d}$
1 (logarithm)	$1 + \log(tf_{t,d})$
a (augmented)	$0.5 + \frac{0.5 \times tf_{t,d}}{\max_t(tf_{t,d})}$
b (boolean)	$\begin{cases} 1 & \text{if } tf_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$
L (log ave)	$\frac{1 + \log(tf_{t,d})}{1 + \log(\text{ave}_{t \in d}(tf_{t,d}))}$

Often, this variant is used due to its monotonic property

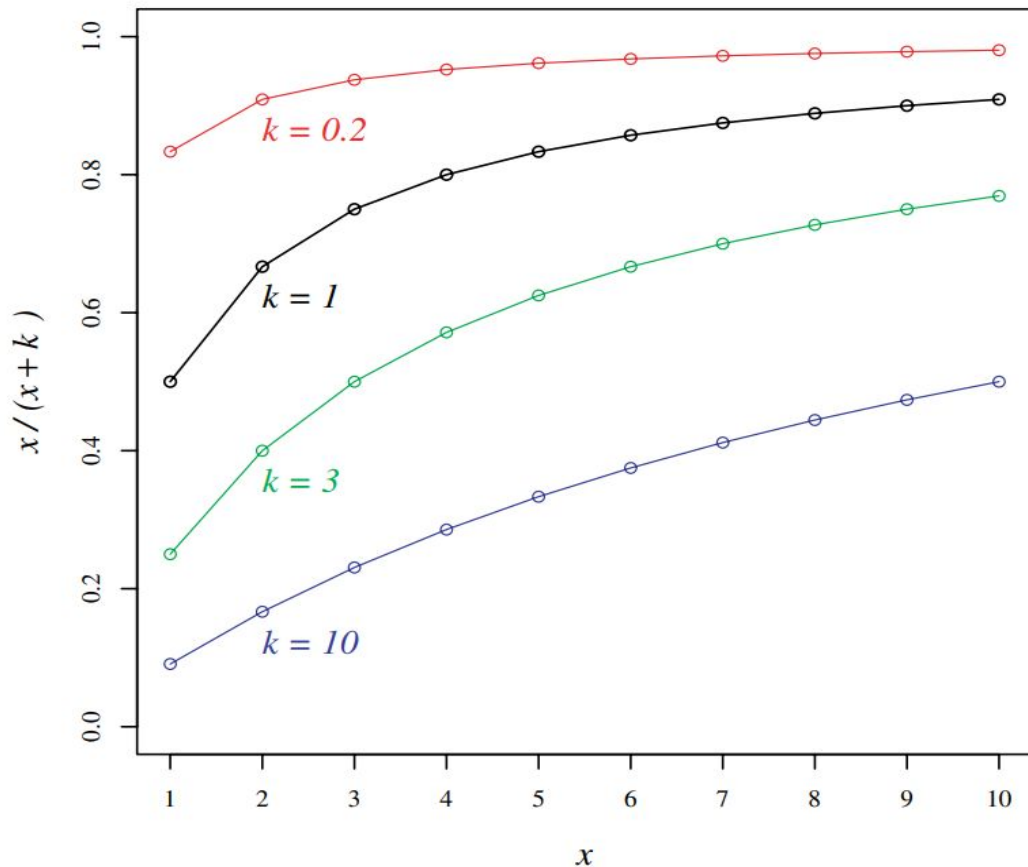
How to implement?

$$1 + \log_2(10) = 4.32$$
$$1 + \log_2(100) = 7.64$$

Bonus: Check out tf-idf implementation in [DrQA](#)

Building BM25 Retrieval

1. **[Query term importance in the document]** Pick a function, which increases monotonically with tf , is rising slowly, but this time is asymptotically approaching (saturates at) some value.



$$\frac{tf}{k + tf} \quad \text{for some } k > 0$$

this is term frequency in document D

$$Score(D, Q) = \sum_{w \in Q, w \in D} \frac{tf_w}{tf_w + k}$$

saturation
parameter

Building BM25 Retrieval

1. **[Query term importance in the document]** Pick a function, which increases monotonically with tf , is rising slowly, but this time is asymptotically approaching (saturates at) some value.
2. **[Overall Term importance]** For every term pick a weight W_w expressing overall term's w importance (e.g. it can be old school $W_w = IDF_w$)

$$Score(D, Q) = \sum_{w \in Q, w \in D} \frac{tf_w}{tf_w + k} W_w$$

Building BM25 Retrieval

1. **[Query term importance in the document]** Pick a function, which increases monotonically with tf , is rising slowly, but this time is asymptotically approaching (saturates at) some value.
2. **[Overall Term importance]** For every term pick a weight W_w expressing overall term's w importance (e.g. it can be old school $W_w = IDF_w$)
3. **[Fix Long Document Bias]** Alleviate **long document bias** problem present in certain collections by penalizing too long documents.

Hypothesis
A



- Some authors are simply **more verbose** than others, using more words to say the same thing.
- These create **bias** in our model; long documents which say the same thing are preferred before short documents, as they achieve more tfs on average.
- An obvious solution to this is to **divide tfs by the document length**.

Hypothesis
B



- Some authors have **more to say**: they may write a single document containing or covering more ground. An extreme version would have the author writing two or more documents and concatenating them.

Building BM25 Retrieval

1. **[Query term importance in the document]** Pick a function, which increases monotonically with tf , is rising slowly, but this time is asymptotically approaching (saturates at) some value.
2. **[Overall Term importance]** For every term pick a weight W_w expressing overall term's w importance (e.g. it can be old school $W_w = IDF_w$)
3. **[Fix Long Document Bias]** Alleviate **long document bias** problem present in certain collections by penalizing too long documents.

$$B := \left((1 - b) + b \frac{dl}{avdl} \right), \quad 0 \leq b \leq 1$$

current document's length

average document length in corpus

soft constraint to cover both hypotheses

Building BM25 Retrieval

1. **[Query term importance in the document]** Pick a function, which increases monotonically with tf , is rising slowly, but this time is asymptotically approaching (saturates at) some value.
2. **[Overall Term importance]** For every term pick a weight W_w expressing overall term's w importance (e.g. it can be old school $W_w = IDF_w$)
3. **[Fix Long Document Bias]** Alleviate **long document bias** problem present in certain collections by penalizing too long documents.

BM-25 Formula

$$Score(D, Q) = \sum_{w \in Q, w \in D} \frac{\frac{tf_w}{B}}{\frac{tf_w}{B} + k} W_w = \sum_{w \in Q, w \in D} \frac{tf_w}{tf_w + kB} W_w$$

Building BM25 Retrieval

1. **[Query term importance in the document]** Pick a function, which increases monotonically with tf , is rising slowly, but this time is asymptotically approaching (saturates at) some value.
2. **[Overall Term importance]** For every term pick a weight W_w expressing overall term's w importance (e.g. it can be old school $W_w = IDF_w$)
3. **[Fix Long Document Bias]** Alleviate **long document bias** problem present in certain collections by penalizing too long documents.
4. **Robertson & Zaragoza, 2009** recommends hyperparam settings $0.5 < b < 0.8$; $1.2 < k < 2$

$$Score(D, Q) = \sum_{w \in Q, w \in D} \frac{\frac{tf_w}{B}}{\frac{tf_w}{B} + k} W_w = \sum_{w \in Q, w \in D} \frac{tf_w (k+1)}{tf_w + kB} W_w$$

CHAPTER 4:
QUESTION ANSWERING

Selective QA

Evaluation: Standard multiclass classification metrics (Accuracy, F1, MCC)

Which of these do scientists offer as the most recent explanation as to why many plants and animals died out at the end of the Mesozoic era?

- (A) worldwide disease
- (B) global mountain building
- (C) rise of mammals that preyed upon plants and animals
- (D) impact of an asteroid created dust that blocked the sunlight

Extractive QA

The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, *Il milione* (or, *The Million*, known in English as the *Travels of Marco Polo*), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge **through contact with Persian traders** since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?

Answer: through contact with Persian traders

(macro)**F1** measures the average overlap between the prediction and ground truth answer.

Prediction and ground truth are treated as bags of tokens and their F1 is computed.

Usually a maximum F1 over all of the ground truth answers for a given question is taken, and the result is an average over all of the questions.

- Ground truth answer = {cat named kitty}

- Predicted answer = {cat named }

- $precision = \frac{\# \text{ of same tokens}}{\text{length of prediction}} = \frac{2}{2}$

- $recall = \frac{\# \text{ of same tokens}}{\text{length of ground truth}} = \frac{2}{3}$

- $F1 = 2 * \frac{precision * recall}{precision + recall} = 2 * \frac{\frac{2}{2} * \frac{2}{3}}{\frac{2}{2} + \frac{2}{3}}$

Abstractive QA

Title: Ghostbusters II

Question: How is Oscar related to Dana?

Answer: her son

Summary snippet: ...Peter's former girlfriend Dana Barrett has had a son, Oscar...

Story snippet:

DANA (setting the wheel brakes on the buggy)

Thank you, Frank. I'll get the hang of this eventually.

She continues digging in her purse while Frank leans over the buggy and makes funny faces at the baby, OSCAR, a very cute nine-month old boy.

FRANK (to the baby)

Hiya, Oscar. What do you say, slugger?

FRANK (to Dana)

That's a good-looking kid you got there, Ms. Barrett.

Task: Answer question from the story

Evaluation via Traditional NLG metrics

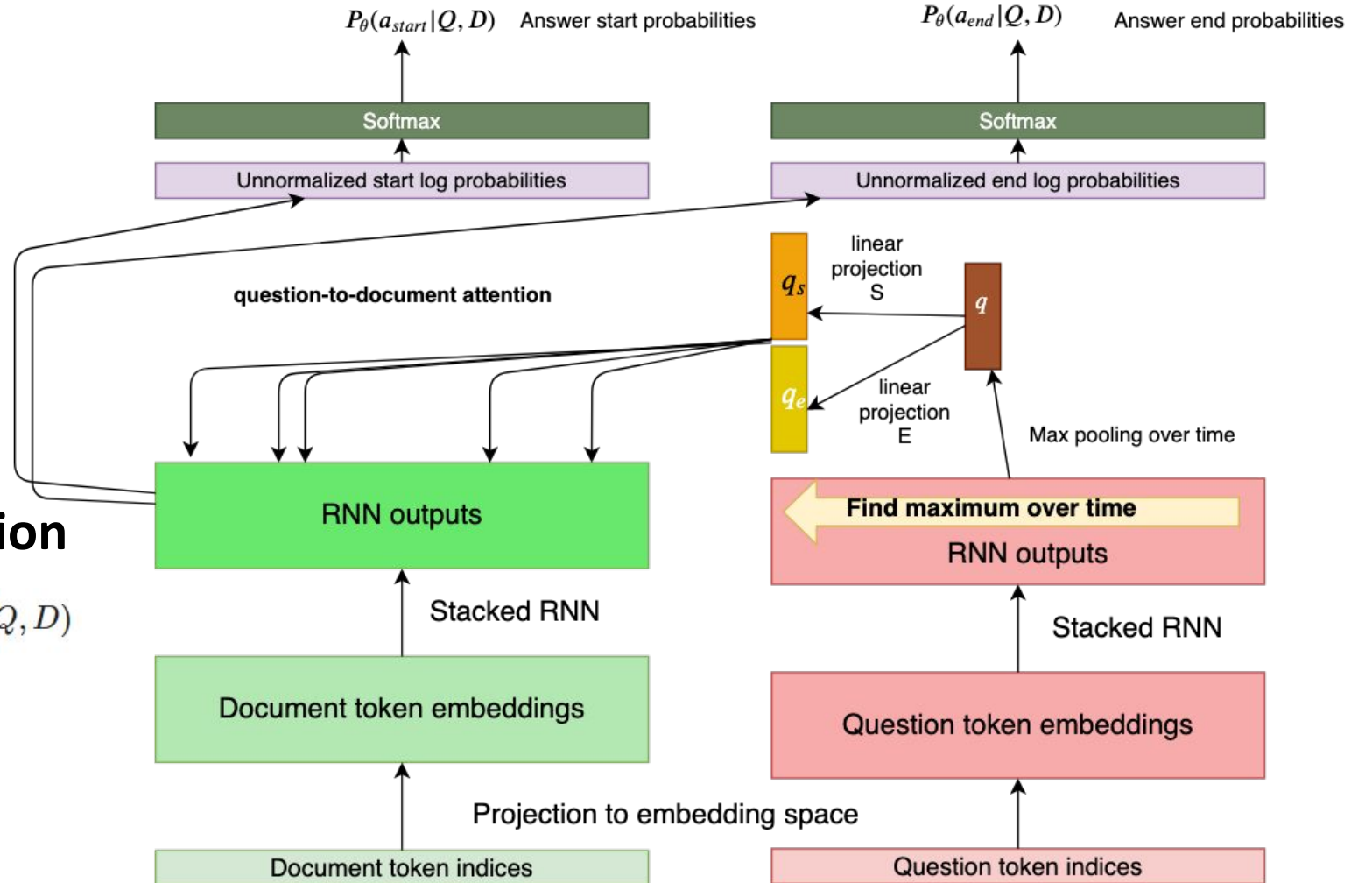
BLEU-4, ROUGE-L, Meteor

A simple extractive QA system A

Given question **Q** and document **D**
find answer span $\langle a_{start}, a_{end} \rangle$

Estimate parameters via **maximum likelihood estimation**

$$\theta^* = \operatorname{argmax}_{\theta} \sum_{(Q, D, a_{start}, a_{end}) \in Data} \log P_{\theta}(a_{start}, a_{end} | Q, D)$$

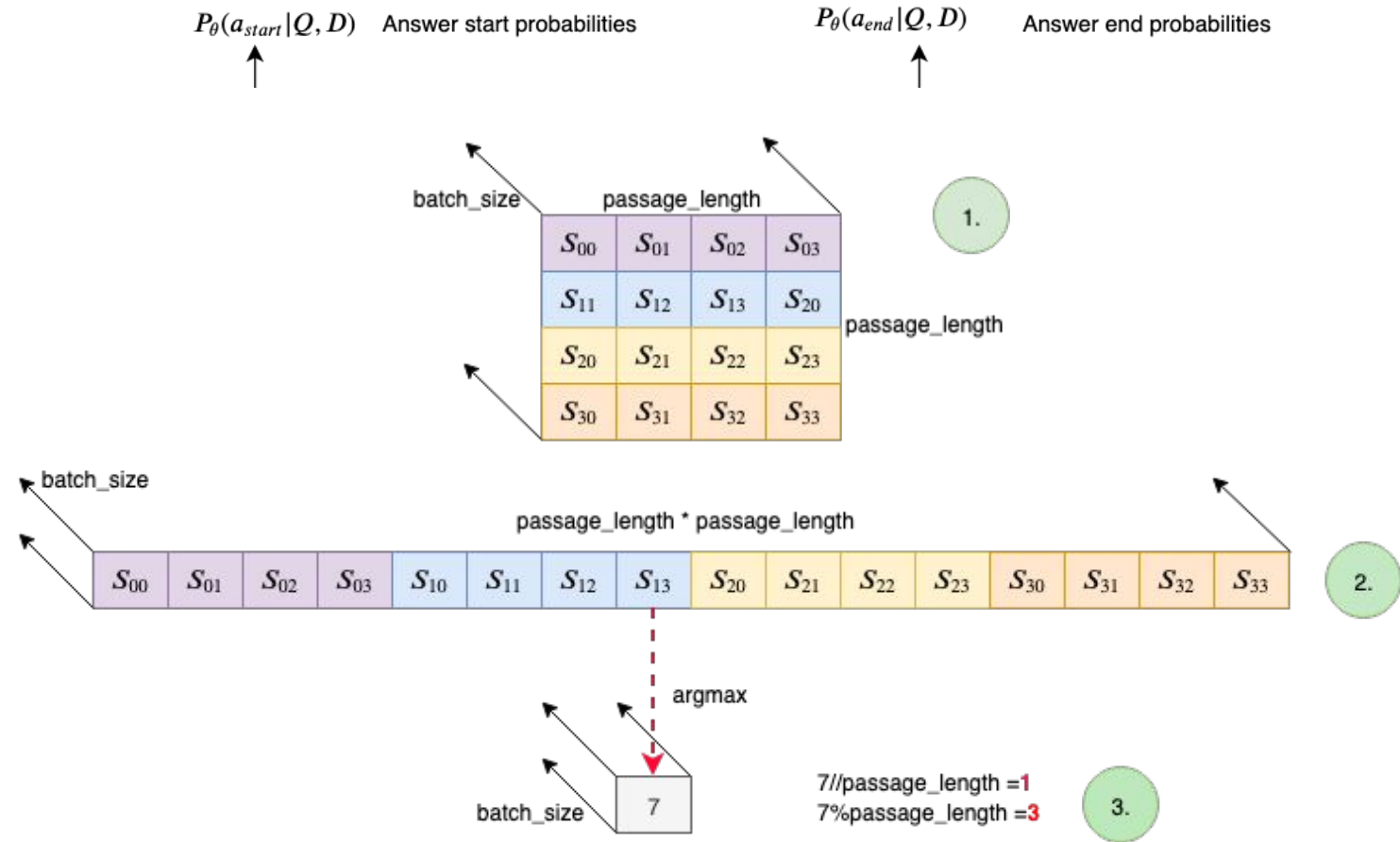


A simple extractive QA system: Decoding

Given question **Q** and document **D**
find answer span $\langle a_{start}, a_{end} \rangle$

Estimate parameters via **maximum likelihood estimation**

$$\theta^* = \operatorname{argmax}_{\theta} \sum_{(Q, D, a_{start}, a_{end}) \in \text{Data}} \log P_{\theta}(a_{start}, a_{end} | Q, D)$$

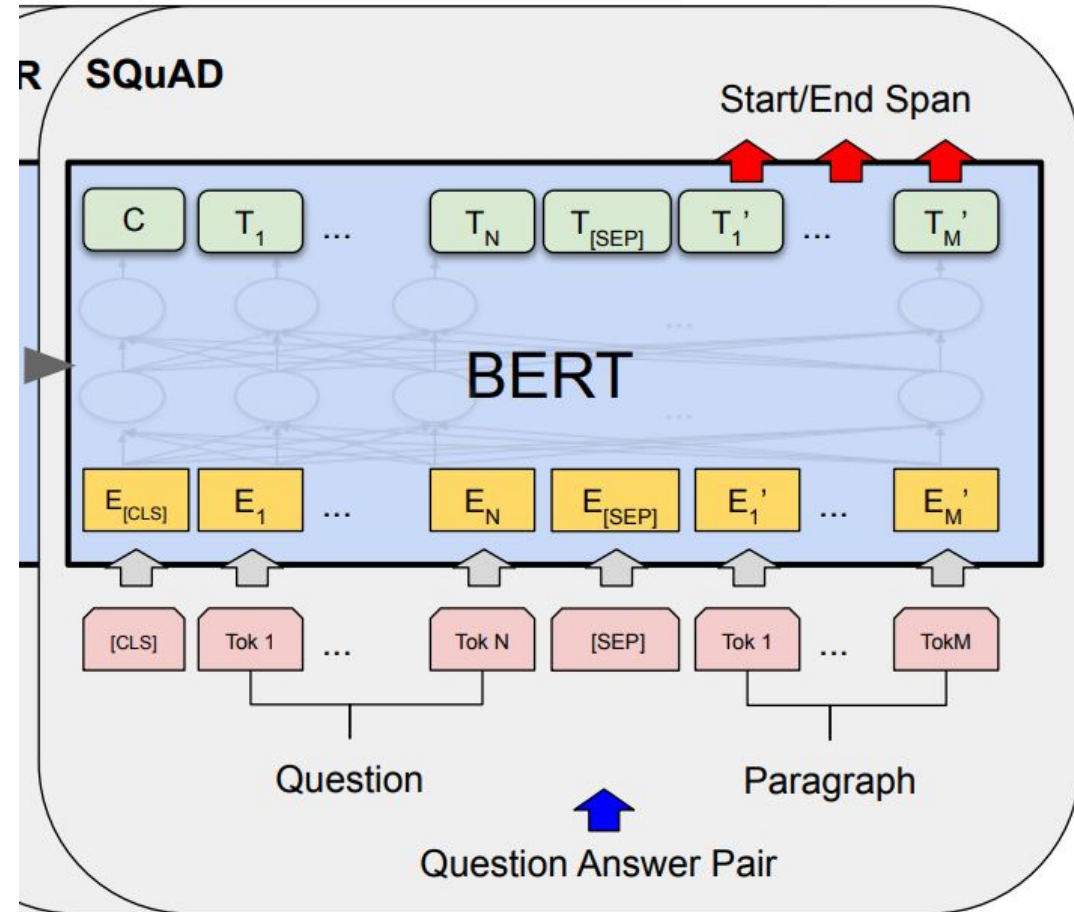


A simple extractive QA system B

Given question **Q** and document **D**
find answer span $\langle a_{start}, a_{end} \rangle$

Estimate parameters via **maximum likelihood estimation**

$$\theta^* = \operatorname{argmax}_{\theta} \sum_{(Q, D, a_{start}, a_{end}) \in \text{Data}} \log P_{\theta}(a_{start}, a_{end} | Q, D)$$



img source: Devlin, Jacob, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *NAACL-HLT*. 2019.

The Objective

$$- \sum_{(q,D,a) \in \mathcal{D}} \log P_{\theta}(a_s, a_e | q, D)$$

Assumption on Independence (Xiong et al., 2017; Seo et al., 2017; Chen et al., 2017; Yu et al., 2018; Devlin et al., 2019; Cheng et al., 2020; inter alia)

$$P(a_s, a_e | q, D) = P_{\theta}(a_s)P_{\theta}(a_e) \quad \text{Do we need to assume the independence?}$$

- Cross-entropy objective for extractive question answering

- given question q
- passage (or a set of passages) D
- answer represented by start/end positions $\mathbf{a}_s/\mathbf{a}_e$

No, we can compute joint objective with similar complexity directly, and it “works better”

(Fajcik et al., 2021)

Open-domain QA

MOTIVATION #1: Research-wise

1. **Dense** Neural Passage retrieval “just” started to work (Lee et al., 2019; Guu et al., 2020; Karupkhin et al., 2020; Khatib et al. 2020; Izacard et al., 2020)
2. Open-domain QA is easy to annotate, all you need is **questions** and **answers**.
3. Closed-domain QA in some cases already works “very well”. Human Performance surpassed - SQuADv1.1, SQuADv2.0 (Rajpurkar et al. 2016,2018), CoQA (Reddy et al., 2018)

Almost any NLP task can be framed as question answering!

The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, Il milione (or, The Million, known in English as the Travels of Marco Polo), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge through contact with Persian traders since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?

Answer: through contact with Persian traders

Open-domain QA

MOTIVATION #2: Information retrieval in everyday life

- **Search needs a shake-up** (Etzioni, 2011)

“Academics and industry researchers need to achieve the intellectual ‘escape velocity’ necessary to **revolutionize search**. They must invest much more in bold strategies that can achieve **natural-language searching and answering**, rather than providing the electronic equivalent.”

← Keyword searching

“Moving up the information food chain requires a search engine that can **interpret** a user's **question**, extract facts from all the information on the **web**, and select an appropriate **answer**.”

Etzioni, Oren. "Search needs a shake-up." *Nature* 476.7358 (2011): 25-26.

why is the sky blue

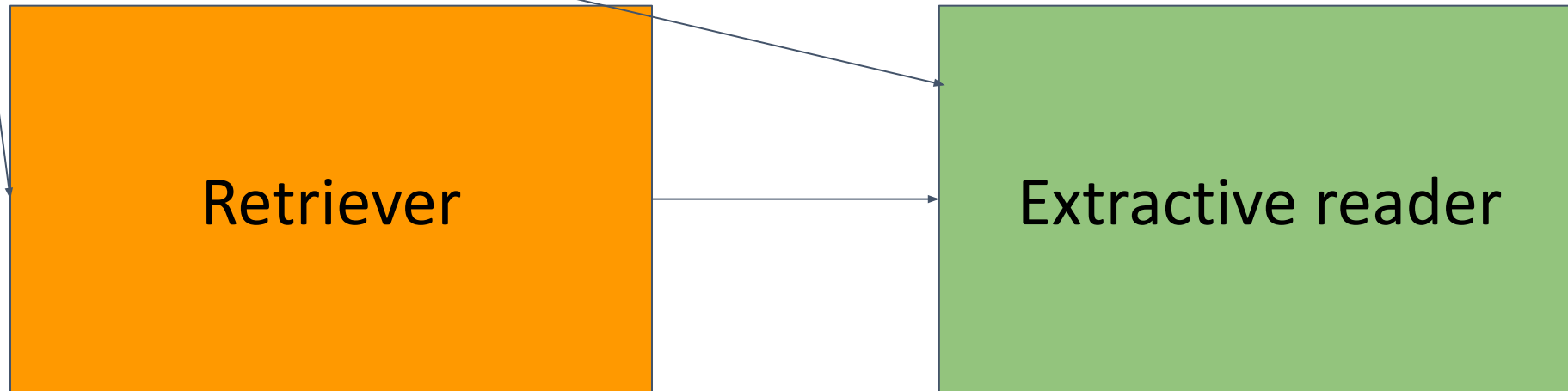


The blue component of the spectrum of visible light has shorter wavelengths and higher frequencies than the red component. **Thus, as sunlight of all colors passes through air, the blue part causes charged particles to oscillate faster than does the red part.** The faster the oscillation, the more scattered light is produced, so blue is scattered more strongly than red. For particles such as air molecules that are much smaller than the wavelengths of visible light the difference is dramatic. The acceleration of the charged particles is proportional to the square of the frequency, and the intensity of scattered light is proportional to the square of this acceleration. Scattered light intensity is therefore proportional to the fourth power of frequency. The result is that blue light is scattered into other directions almost 10 times as efficiently as red light.

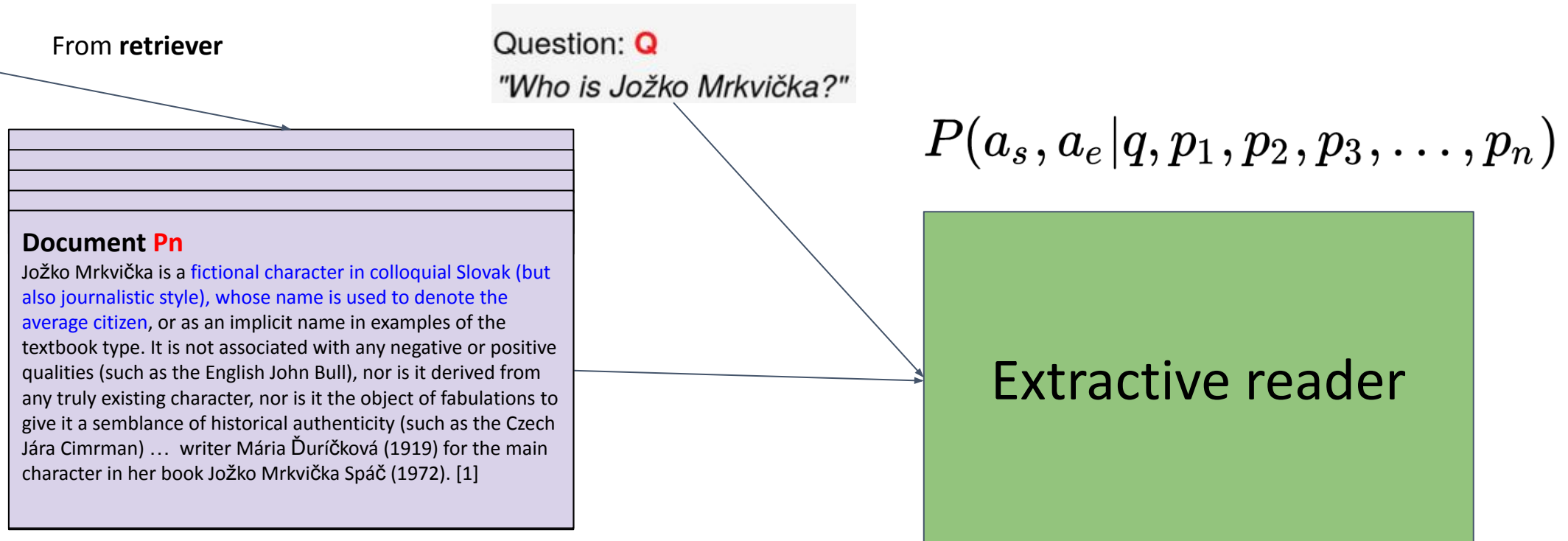
When we look at an arbitrary point in the sky, away from the sun, we see only the light that was redirected by the atmosphere into our line of sight. Because that occurs much more often for blue light than for red, the sky appears blue. Violet light is actually scattered even a bit more strongly than blue. **More of the sunlight entering the atmosphere is blue than violet, however, and our eyes are somewhat more sensitive to blue light than to violet light, so the sky appears blue.**

Example of traditional approach

Question: **Q**
"Who is Jožko Mrkvička?"



Example of traditional approach: Reader



- In current literature, each document is usually processed via language representation model (e.g. BERT) separately.

Maximum Marginal Likelihood

- In Open-QA, we often do not know, which answer span is correct and which is not

Question: Which British general was killed at Khartoum in 1885?

Answer: Gordon

Context: In February 1885 **Gordon** returned to the Sudan to evacuate Egyptian forces. Khartoum came under siege the next month and rebels broke into the city, killing **Gordon** and the other defenders. The British public reacted to his death by acclaiming '**Gordon** of Khartoum', a saint. However, historians have suggested that **Gordon**...

Figure 1: Noisy supervision can cause many spans of text that contain the answer, but are not situated in a context that relates to the question (red), to distract the model from learning from more relevant spans (green).

Maximum Marginal Likelihood

- In Open-QA, we often do not know, which answer span is correct and which is not
- Solution? Marginalize over all spans with correct surface form, let the model decide
- Formally:
- in **fully supervised** setting, we are given input x , and answer span \bar{z} , our NLL objective for 1 sample is

$$J_{\text{Sup}}(\theta|x, \bar{z}) = -\log \mathbb{P}(\bar{z}|x; \theta)$$

Maximum Marginal Likelihood

- In Open-QA, we often do not know, which answer span is correct and which is not
- Solution? Marginalize over all spans with correct surface form, let the model decide
- Formally:
- in **fully supervised** setting, we are given input x , and answer span \bar{z} our NLL objective for 1 sample is

$$J_{\text{Sup}}(\theta|x, \bar{z}) = -\log \mathbb{P}(\bar{z}|x; \theta)$$

in **weakly supervised** setting, we are given input x , and many answer spans for each string match $Z=\{z_1, z_2, \dots, z_n\}$, some of which are correct, some of which are not.

- Note that Z is subset of Z_{tot} , the set of \bar{z} all spans in the document(s), y is answer string match

$$\begin{aligned} \mathbb{P}(y|x; \theta) &= \sum_{z_i \in Z_{\text{tot}}} \mathbb{P}(y|z_i) \mathbb{P}(z_i|x; \theta) \\ &= \sum_{z_i \in Z} \mathbb{P}(z_i|x; \theta) \end{aligned}$$

Maximum Marginal Likelihood

- In Open-QA, we often do not know, which answer span is correct and which is not
- Solution? Marginalize over all spans with correct surface form, let the model decide
- Formally:
- in **fully supervised** setting, we are given input x , and answer span \bar{z} , our NLL objective for 1 sample is

$$J_{\text{Sup}}(\theta|x, \bar{z}) = -\log \mathbb{P}(\bar{z}|x; \theta)$$

in **weakly supervised** setting, we are given input x , and many answer spans for each string match $Z=\{z_1, z_2, \dots, z_n\}$, some of which are correct, some of which are not.

- Note that Z is subset of Z_{tot} , the set of all spans in the document(s), y is answer string match

$$\begin{aligned} \mathbb{P}(y|x; \theta) &= \sum_{z_i \in Z_{\text{tot}}} \mathbb{P}(y|z_i) \mathbb{P}(z_i|x; \theta) \quad \leftarrow \text{0 if } z_i \text{ is not from } Z, \text{ 1 otherwise} \\ &= \sum_{z_i \in Z} \mathbb{P}(z_i|x; \theta) \end{aligned}$$

Maximum Marginal Likelihood

- In Open-QA, we often do not know, which answer span is correct and which is not
- Solution? Marginalize over all spans with correct surface form, let the model decide
- Formally:
- in **fully supervised** setting, we are given input x , and answer span \bar{z} , our NLL objective for 1 sample is

$$J_{\text{Sup}}(\theta|x, \bar{z}) = -\log \mathbb{P}(\bar{z}|x; \theta)$$

in **weakly supervised** setting, we are given input x , and many answer spans for each string match $Z=\{z_1, z_2, \dots, z_n\}$, some of which are correct, some of which are not.

- Note that Z is subset of Z_{tot} , the set of all spans in the document(s), y is answer string match

$$\mathbb{P}(y|x; \theta) = \sum_{z_i \in Z_{\text{tot}}} \mathbb{P}(y|z_i) \mathbb{P}(z_i|x; \theta) \quad \begin{array}{l} \longleftarrow \\ \text{0 if } z_i \text{ is not from } Z, \text{ 1 otherwise} \end{array}$$

$$= \sum_{z_i \in Z} \mathbb{P}(z_i|x; \theta)$$

$$J_{\text{MML}}(\theta|x, Z) = -\log \sum_{z_i \in Z} \mathbb{P}(z_i|x; \theta)$$

Maximum Marginal Likelihood

- In Open-QA, we often do not know, which answer span is correct and which is not
- Solution? Marginalize over all spans with correct surface form, let the model decide
- Formally:
- in **fully supervised** setting, we are given input x , and answer span \bar{z} , our NLL objective for 1 sample is

$$J_{\text{Sup}}(\theta|x, \bar{z}) = -\log \mathbb{P}(\bar{z}|x; \theta)$$

in **weakly supervised** setting, we are given input x , and many answer spans for each string match $Z=\{z_1, z_2, \dots, z_n\}$, some of which are correct, some of which are not.

- Note that Z is subset of Z_{tot} , the set of all spans in the document(s), y is answer string match

$$\mathbb{P}(y|x; \theta) = \sum_{z_i \in Z_{\text{tot}}} \mathbb{P}(y|z_i) \mathbb{P}(z_i|x; \theta)$$

0 if z_i is not from Z , 1 otherwise

$$= \sum_{z_i \in Z} \mathbb{P}(z_i|x; \theta)$$

$$J_{\text{MML}}(\theta|x, Z) = -\log \sum_{z_i \in Z} \mathbb{P}(z_i|x; \theta)$$

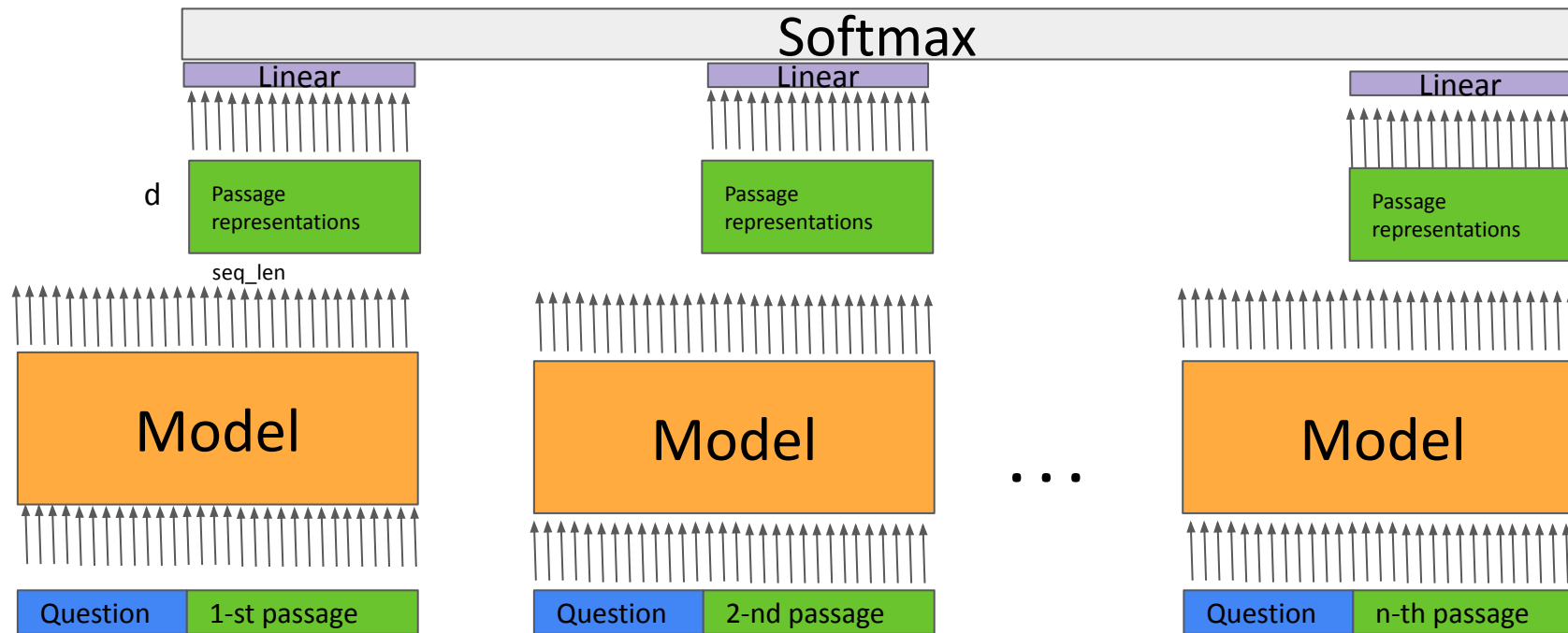
This is a so called „latent variable model“ with latent variable v_i . Remember GMM!

MML in Open-domain QA

Loss for 1 sample $\mathcal{L}_i =$
 $-\log \sum_{p \in P} \sum_{a \in GT(p)} P(a_s | q, P)$
 $-\log \sum_{p \in P} \sum_{a \in GT(p)} P(a_e | q, P)$

$$P(a_{start} | q, p_1, p_2, \dots, p_n)$$

$$P(a_{end} | q, p_1, p_2, \dots, p_n)$$

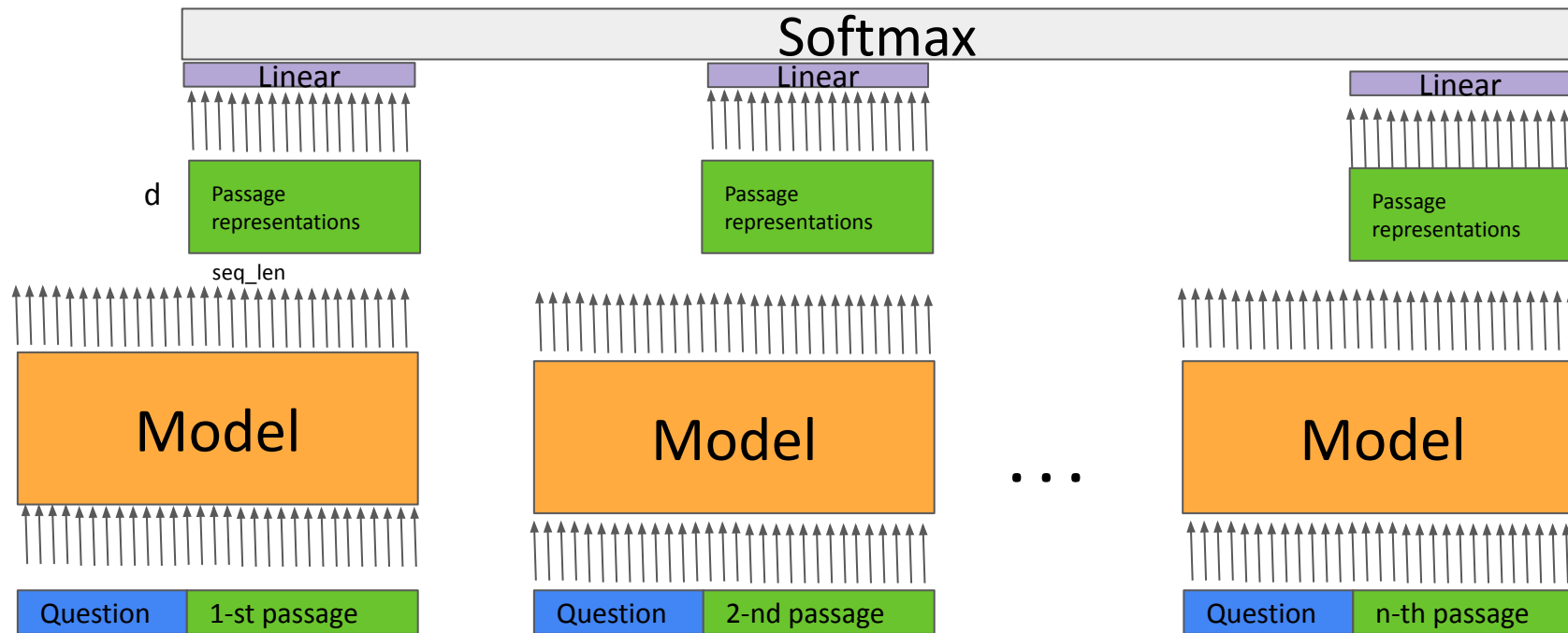


MML in Open-domain QA

Loss for 1 sample $\mathcal{L}_i =$
 $-\log \sum_{p \in P} \sum_{a \in GT(p)} P(a_s | q, P)$
 $-\log \sum_{p \in P} \sum_{a \in GT(p)} P(a_e | q, P)$

$$P(a_{start} | q, p_1, p_2, \dots, p_n)$$

$$P(a_{end} | q, p_1, p_2, \dots, p_n)$$



Important for cross-passage answer score calibration!

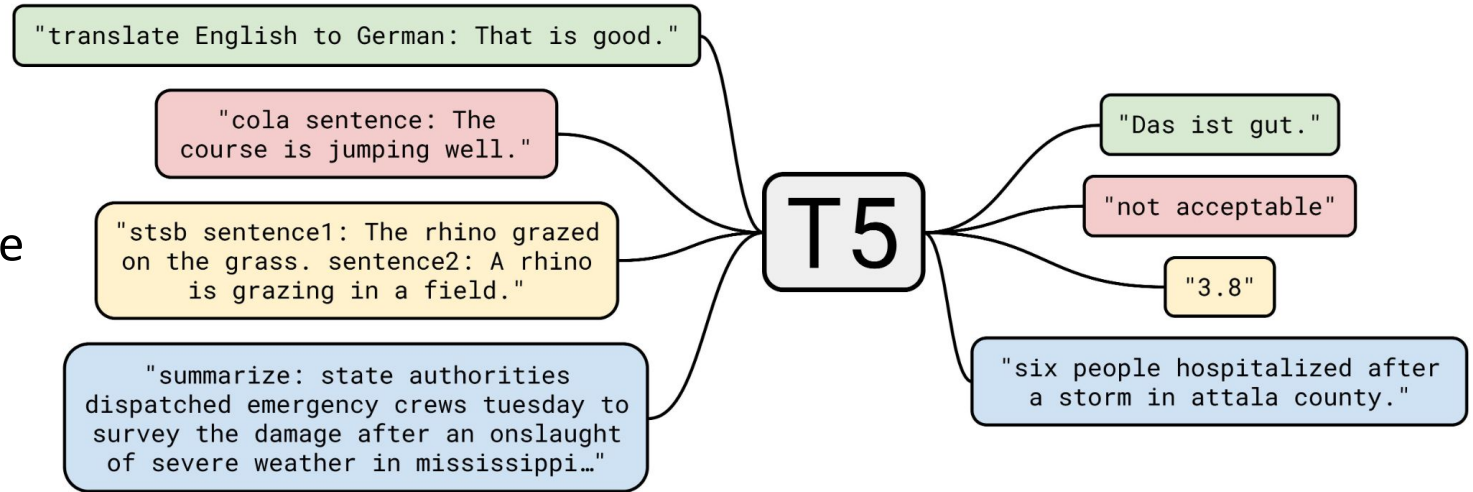
Clark, Christopher, and Matt Gardner. "Simple and Effective Multi-Paragraph Reading Comprehension." *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018.

Do we need to use extractive models?

- IDEA: generate answer through the language model

T5

- Seq-2-seq, Enc-Decoder unlike BERT
- subword language units
- trained on denoising objective and ~25 supervised tasks
- 750GB CommonCrawl data



Original text

Thank you ~~for inviting~~ me to your party ~~last~~ week.

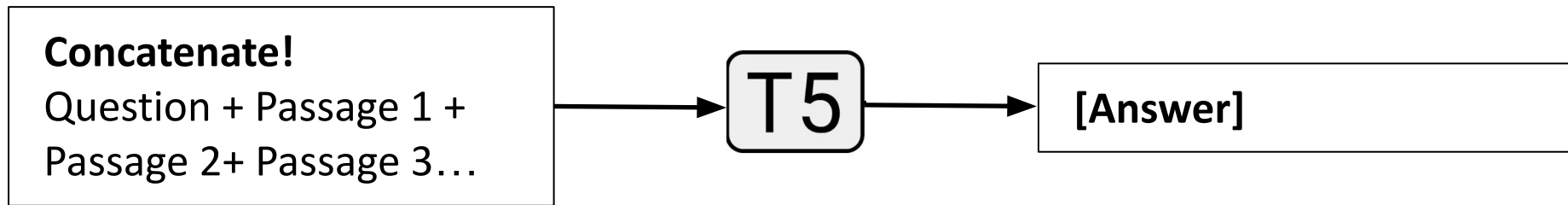
Inputs

Thank you <X> me to your party <Y> week.

Targets

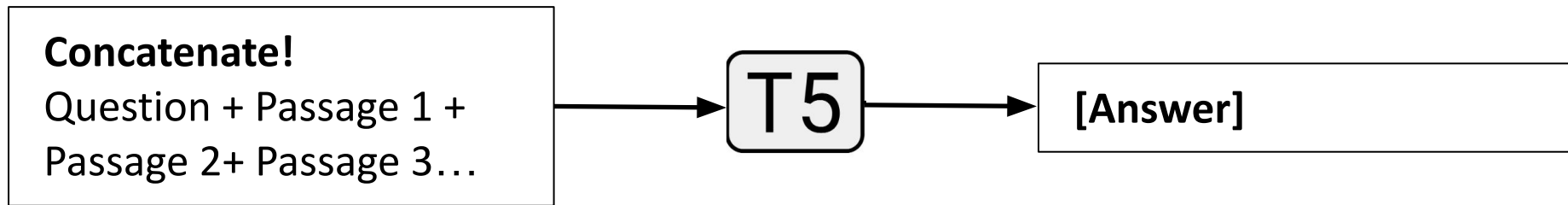
<X> for inviting <Y> last <Z>

Idea #1: „Concatenate, pass, profit“



Drawbacks?

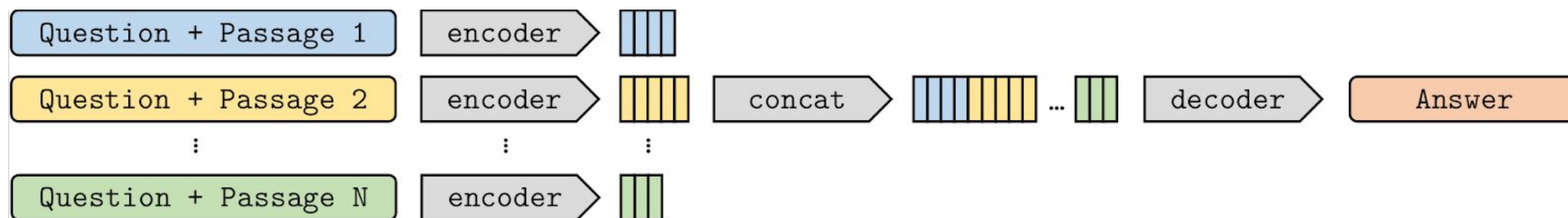
Idea #1: „Concatenate, pass, profit“



- 1. Memory complexity**
- 2. Decoding: If we do decoding without restrictions,
the model might generate something not present in the text**

Idea #2: Processing passages jointly: Fusion-in-Decoder

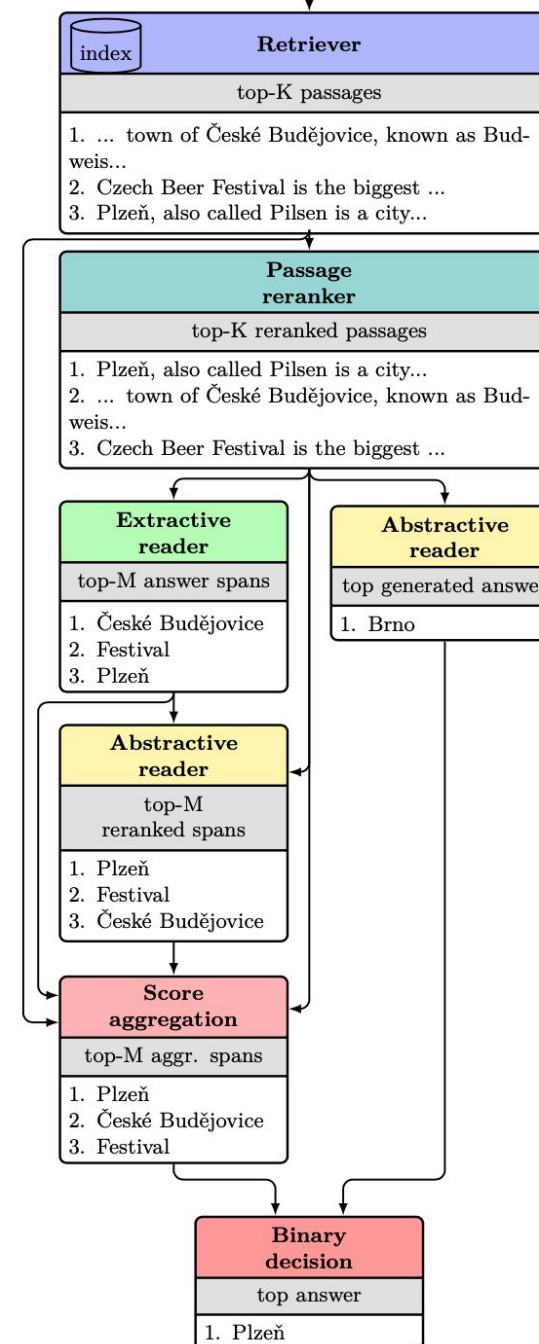
- Do we need to read every passage independently?
- No, we can actually allow inter-passage interaction learning!
- Example: Fusion-in-Decoder (FiD), encode every passage separately, decode jointly
- Trick works well with pre-trained models (T5)!
- Can process very long inputs (sequences of $200(\text{passage length}) * 100(\text{context size})$ tokens long)
- Optimize target answer via standard language modeling loss (Cross-Entropy)



Fusing the extractive and generative approaches

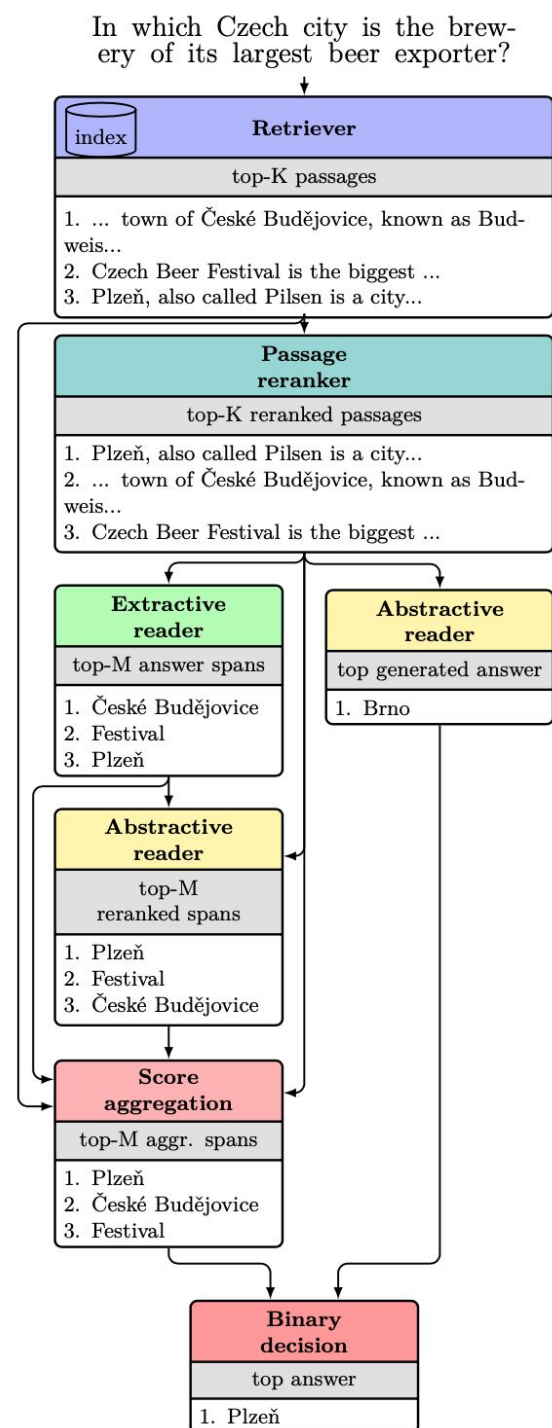
- Our past work:
Rank twice, read twice R2-D2
- <https://r2d2.fit.vutbr.cz/>
- Some demo details:
 - The search is done in “popular” 8% of Wikipedia
 - Only factoid answers, up to 6 words
 - Wikipedia from dec 2018 is used

In which Czech city is the brewery of its largest beer exporter?



Fusing the extractive and generative approaches

- Why is the search done in “popular” 8% of Wikipedia?
 - We’ve shown we can remove 92% of index from two most popular datasets for open-domain QA, NaturalQuestions and TriviaQA, while losing only up to 3% absolute performance on test set.
 - How? We trained a classifier which given a passage, tries to say apriori (without seeing any question), whether the passage is relevant or not.
 - Could same “pruning” mechanism be implicitly present in modern supervised neural retrieval approaches?
 - Wait for the release of my PhD thesis 😊



CHAPTER 5:
'23 TRENDS

Is QA “solved” by LLM such as ChatGPT/GPT4?

There is no definite answer, but we can do what every good scientist should.

Hypothesize...

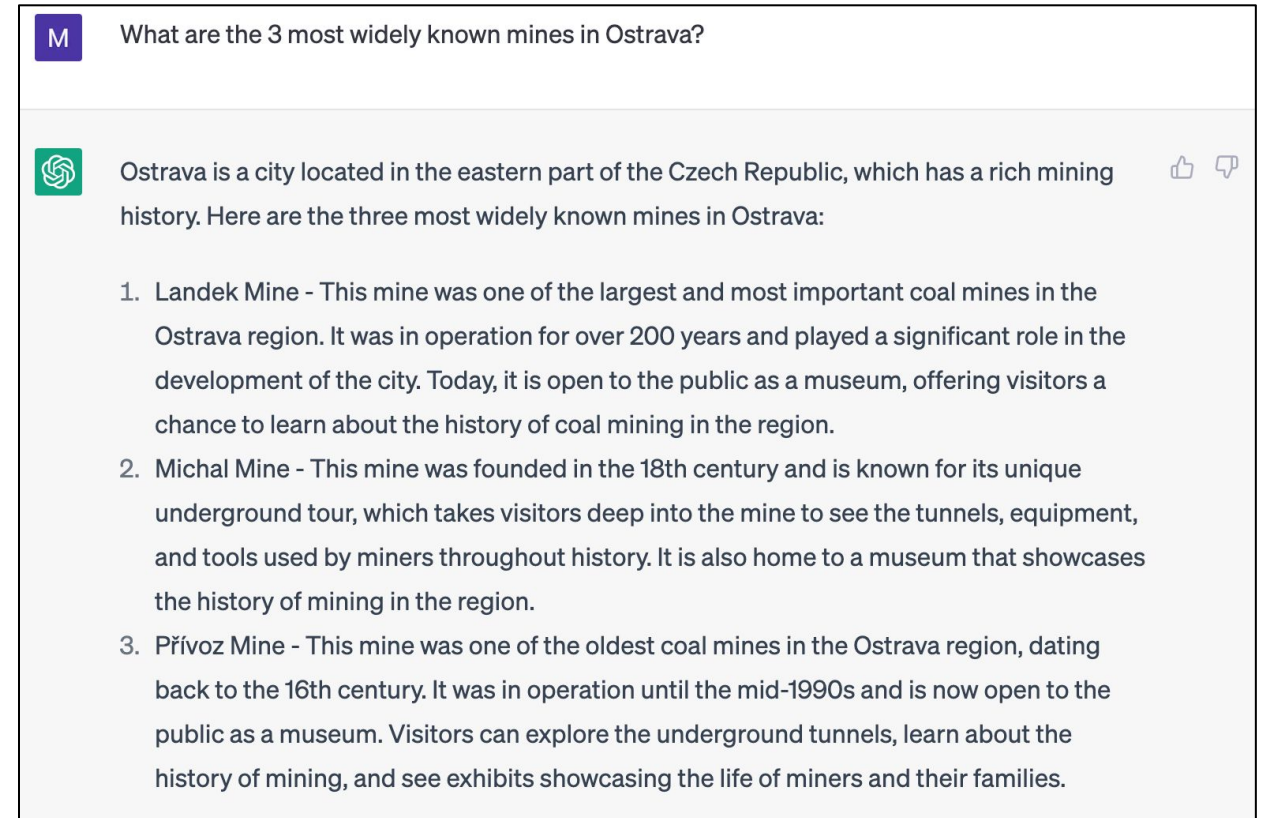
Warning:

- The subsequent slides are subjective, and draw takeaways from simple case-study observations.
- Observations made are not (yet) fully quantified with the scientific evidence.

Is QA “solved” by LLM such as ChatGPT/GPT4?

There is no definite answer, but we can do what every good scientist should. **Hypothesize...**

1. Yes because...
 1. Large LLM have extensive factual knowledge.
 2. LLMs can present answers excellently!



M What are the 3 most widely known mines in Ostrava?

Ostrava is a city located in the eastern part of the Czech Republic, which has a rich mining history. Here are the three most widely known mines in Ostrava:

1. Landek Mine - This mine was one of the largest and most important coal mines in the Ostrava region. It was in operation for over 200 years and played a significant role in the development of the city. Today, it is open to the public as a museum, offering visitors a chance to learn about the history of coal mining in the region.
2. Michal Mine - This mine was founded in the 18th century and is known for its unique underground tour, which takes visitors deep into the mine to see the tunnels, equipment, and tools used by miners throughout history. It is also home to a museum that showcases the history of mining in the region.
3. Přívoz Mine - This mine was one of the oldest coal mines in the Ostrava region, dating back to the 16th century. It was in operation until the mid-1990s and is now open to the public as a museum. Visitors can explore the underground tunnels, learn about the history of mining, and see exhibits showcasing the life of miners and their families.

Is QA “solved” by LLM such as ChatGPT/GPT4?

There is no definite answer, but we can do what every good scientist should. **Hypothesize...**

1. Maybe because...
 1. Large LLMs can **lie** excellently. This kind of problem is called “Hallucination”.

M

Name a few mining tunnels (adits) from Kremnica mountains



The Kremnica Mountains in central Slovakia have a rich mining history with numerous adits (horizontal tunnels) that were used for mining various metals including gold, silver, copper, and other minerals. Here are a few examples of mining tunnels from the Kremnica Mountains:

1. Andrej Adit - This is one of the oldest and most famous mining tunnels in the Kremnica Mountains, used for silver and other metal mining. The adit is open to visitors for guided tours.
2. Rozalia Adit - A historic adit used for silver, copper, and other metal mining in the Kremnica Mountains, now open to visitors for guided tours.
3. Zobor Adit - An adit located near the town of Štiavnické Bane in the Kremnica Mountains, used for silver and copper mining. It is open to visitors for guided tours.
4. Svatopluk Adit - An adit used for gold and silver mining in the Kremnica Mountains, now open to visitors for guided tours.
5. Barbora Adit - A historic adit located in the town of Kremnica, used for silver and other metal mining. It is open to visitors for guided tours.
6. Nová Baňa Adit - An adit used for copper and gold mining in the Kremnica Mountains, now open to visitors for guided tours.
7. Sklené Teplice Adit - A historic adit used for silver and copper mining in the Kremnica Mountains, now open to visitors for guided tours.

Is QA “solved” by LLM such as ChatGPT/GPT4?

There is no definite answer, but we can do what every good scientist should. **Hypothesize...**

1. No because
 1. Large LLMs cannot explain themselves.

M

Name a few mining tunnels (adits) from Kremnica mountains



The Kremnica Mountains in central Slovakia have a rich mining history with numerous adits (horizontal tunnels) that were used for mining various metals including gold, silver, copper, and other minerals. Here are a few examples of mining tunnels from the Kremnica Mountains:

1. Andrej Adit - This is one of the oldest and most famous mining tunnels in the Kremnica Mountains, used for silver and other metal mining. The adit is open to visitors for guided tours.
2. Rozalia Adit - A historic adit used for silver, copper, and other metal mining in the Kremnica Mountains, now open to visitors for guided tours.
3. Zobor Adit - An adit located near the town of Štiavnické Bane in the Kremnica Mountains, used for silver and copper mining. It is open to visitors for guided tours.
4. Svatopluk Adit - An adit used for gold and silver mining in the Kremnica Mountains, now open to visitors for guided tours.
5. Barbora Adit - A historic adit located in the town of Kremnica, used for silver and other metal mining. It is open to visitors for guided tours.
6. Nová Baňa Adit - An adit used for copper and gold mining in the Kremnica Mountains, now open to visitors for guided tours.
7. Sklené Teplice Adit - A historic adit used for silver and copper mining in the Kremnica Mountains, now open to visitors for guided tours.

Is QA “solved” by LLM such as ChatGPT/GPT4?

There is no definite answer, but we can do what every good scientist should. **Hypothesize...**

1. No because

1. LLMs are competitive, but not outperforming the task specific models.

Tasks	Dataset	Metric	Reference	Fine-Tuned SOTA	Zero-Shot SOTA	ChatGPT
Summarization	CNN/DM	ROUGE-1	Lewis et al. (2020a)	44.47	35.27 ⁷	35.29
	SAMSum	ROUGE-1	Lewis et al. (2020a)	47.28	-	35.29
MT (XXX→Eng)	FLoRes-200 (HRL)	ChrF++	Team et al. (2022)	63.5	-	58.64
	FLoRes-200 (LRL)	ChrF++	Team et al. (2022)	54.9	-	27.75
MT (Eng→XXX)	FLoRes-200 (HRL)	ChrF++	Team et al. (2022)	54.4	-	51.12
	FLoRes-200 (LRL)	ChrF++	Team et al. (2022)	41.9	-	21.57
Sentiment Analysis	NusaX - Eng	Macro F1	Winata et al. (2022)	92.6	61.5	83.24
	NusaX - Ind	Macro F1	Winata et al. (2022)	91.6	59.3	82.13
	NusaX - Jav	Macro F1	Winata et al. (2022)	84.2	55.7	79.64
	NusaX - Bug	Macro F1	Winata et al. (2022)	70.0	55.9	55.84
Question Answering	bAbI task 15	Accuracy	Weston et al. (2016a)	100	-	93.3
	bAbI task 16	Accuracy	Weston et al. (2016a)	100	-	66.7
	EntailmentBank	Accuracy	Clark et al. (2018)	86.5	78.58	93.3
	CLUTRR	Accuracy	Minervini et al. (2020)	95.0	28.6	43.3
	StepGame (k=9)	Accuracy	Mirzaee and Kordjamshidi (2022)	48.4	-	23.3
	StepGame (k=1)	Accuracy	Mirzaee and Kordjamshidi (2022)	98.7	-	63.3
	Pep-3k	AUC	Porada et al. (2021)	67.0	-	93.3
Misinformation Detection	COVID-Social	Accuracy	Lee et al. (2021)	77.7	50.0	73.3
	COVID-Scientific	Accuracy	Lee et al. (2021)	74.7	71.1	92.0
Task-Oriented Dialogue	MultiWOZ2.2	JGA	Zhao et al. (2022)	60.6	46.7	24.4
	MultiWOZ2.2	BLEU	Nekvinda and Dušek (2021)	19.1	-	5.65
	MultiWOZ2.2	Inform Rate	Yang et al. (2021)	95.7	-	71.1
Open-Domain KGD	OpenDialKG	BLEU	Ji et al. (2022c)	20.8	3.1	4.1
	OpenDialKG	ROUGE-L	Ji et al. (2022c)	40.0	29.5	18.6
	OpenDialKG	FeQA	Ji et al. (2022c)	48.0	23.0	15.0

Table 1: Performance of ChatGPT compared to state-of-the-art fully-fine-tuned models (Fine-Tuned SOTA) and LLM in zero-shot settings (Zero-Shot SOTA). The referenced performances are evaluation results on full test sets, while the ChatGPT performances are computed on subsets of the corresponding dataset using 30 to 200 data samples for each task. For Machine Translation (MT) tasks, we use the definitions of high-resource language (HRL) and low-resource language (LRL) from NLLB (Team et al., 2022) and take subsets of languages to represent each group. JGA denotes joint goal accuracy.

Is QA “solved” by Retrieval-Augmented LLMs?

There is no definite answer, but we can do what every good scientist should. **Hypothesize...**

1. Yes because
 1. all responses with factoid answers are grounded.

The screenshot displays a chat interface with two distinct queries and their corresponding answers, each with a 'Learn more' section containing source links.

Query 1: "Who won the 2022 World Cup?"

Answer 1: Argentina won the 2022 World Cup by beating France on penalties in the final at the Lusail Iconic Stadium in Qatar^{1 2}. The score was 3-3 after extra time, and Argentina won 4-2 on penalties^{3 1}. It was Argentina's third World Cup title, and their first since 1986^{1 2}.

Query 2: "Who is the President of the USA today?"

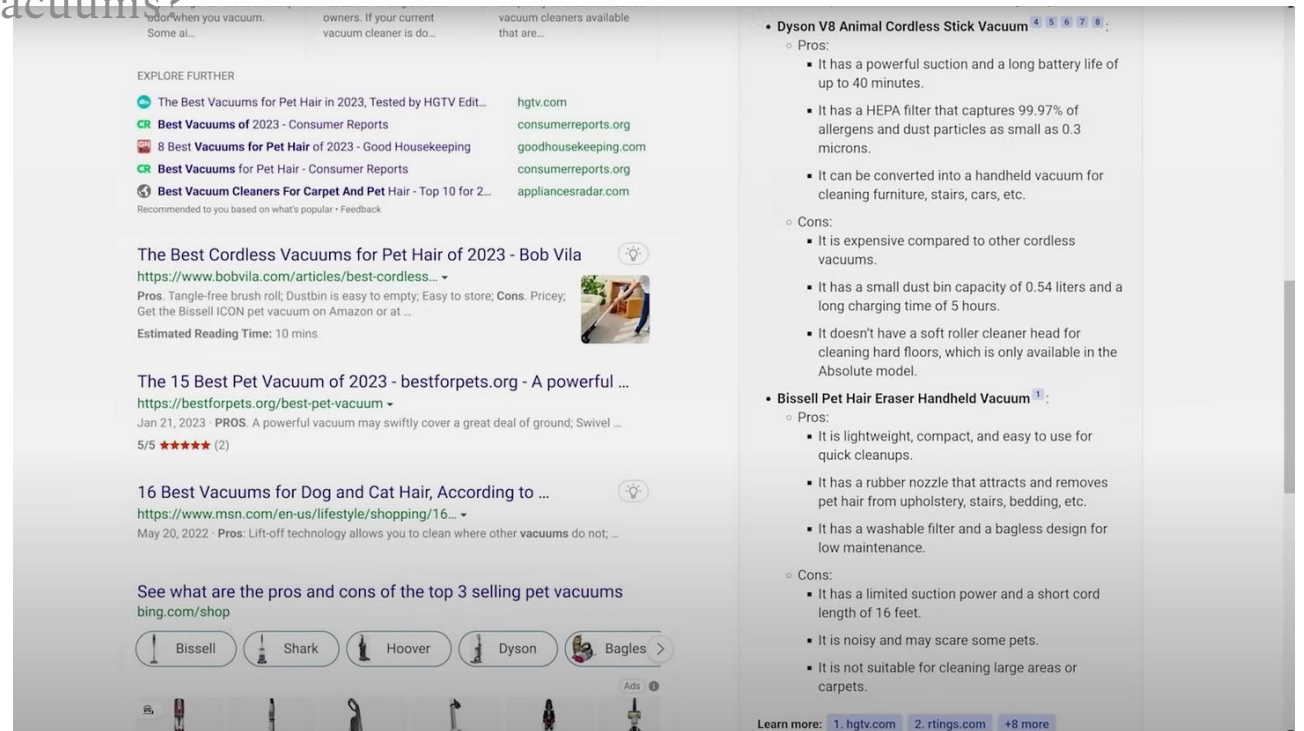
Answer 2: The President of the USA today is Joe Biden^{1 2 3}. He is the 46th President of the United States, and has been in office since January 20, 2021^{1 2 3}. He is a Democrat and was elected on November 7, 2020, defeating former President Donald Trump^{1 3}.

The interface also shows a search bar, a chat icon, and a text input field at the bottom with the placeholder "Ask me anything...".

Is QA “solved” by Evidence-grounded LLMs?

There is no definite answer, but we can do what every good scientist should. **Hypothesize...**

Query: What are the pros and cons of the top 3 selling pet vacuums?



1. Maybe because
 1. Evidence-grounded models still suffer from hallucination.

Is QA “solved” by Evidence-grounded LLMs?

Query: What are the pros and cons of the top 3 selling pet

The screenshot shows a Bing search result for the query "vacuum?". The search results include several articles:

- [The Best Cordless Vacuums for Pet Hair of 2023 - Bob Vila](https://www.bobvila.com/articles/best-cordless...)
- [The 15 Best Pet Vacuum of 2023 - bestforpets.org](https://bestforpets.org/best-pet-vacuum)
- [16 Best Vacuums for Dog and Cat Hair, According to ...](https://www.msn.com/en-us/lifestyle/shopping/16...)
- [See what are the pros and cons of the top 3 selling pet vacuums](https://bing.com/shop)

The main content area displays a comparison of the top 3 selling pet vacuums:

- Dyson V8 Animal Cordless Stick Vacuum**
 - Pros:
 - It has a powerful suction and a long battery life of up to 40 minutes.
 - It has a HEPA filter that captures 99.97% of allergens and dust particles as small as 0.3 microns.
 - It can be converted into a handheld vacuum for cleaning furniture, stairs, cars, etc.
 - Cons:
 - It is expensive compared to other cordless vacuums.
 - It has a small dust bin capacity of 0.54 liters and a long charging time of 5 hours.
 - It doesn't have a soft roller cleaner head for cleaning hard floors, which is only available in the Absolute model.
- Bissell Pet Hair Eraser Handheld Vacuum**
 - Pros:
 - It is lightweight, compact, and easy to use for quick cleanups.
 - It has a rubber nozzle that attracts and removes pet hair from upholstery, stairs, bedding, etc.
 - It has a washable filter and a bagless design for low maintenance.
 - Cons:
 - It has a limited suction power and a short cord length of 16 feet.
 - It is noisy and may scare some pets.
 - It is not suitable for cleaning large areas or carpets.

At the bottom, there are navigation buttons for Bissell, Shark, Hoover, Dyson, and Bagless, along with a "Learn more" link.

”This is all completely made up information. Bing AI was kind enough to give us its sources, so we can go to the [hgty article](#) and check for ourselves.

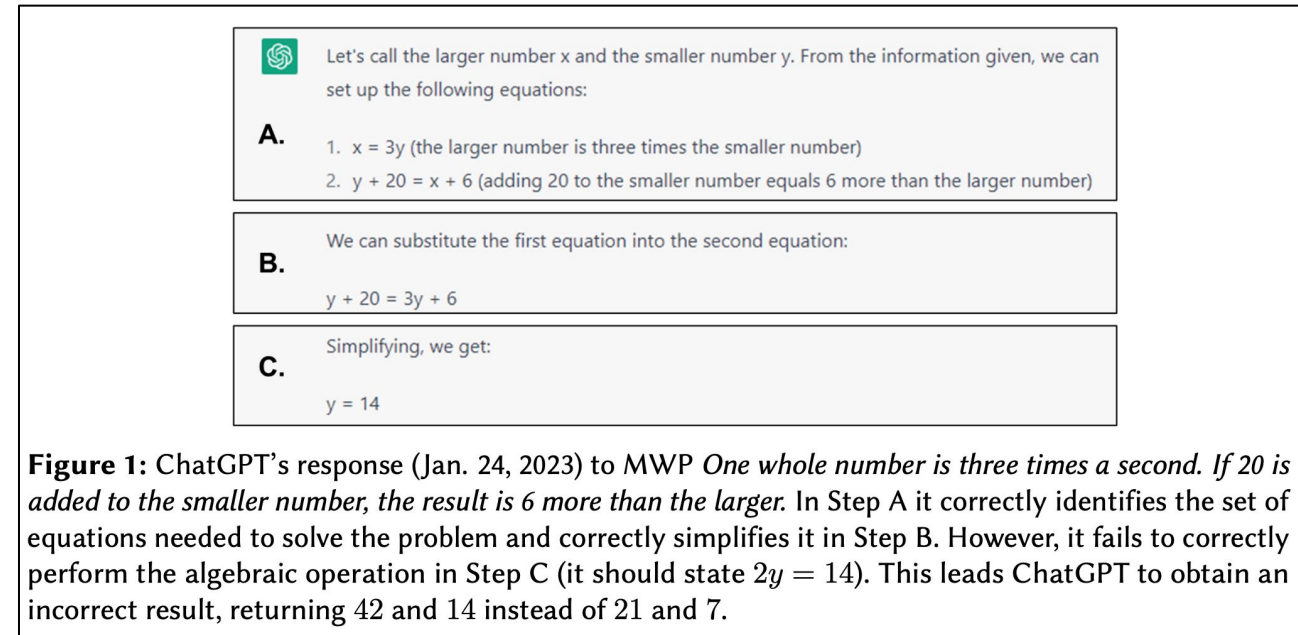
The cited article says nothing about limited suction power or noise. In fact, the top amazon review for this product talks about how quiet it is.

The article also says nothing about the “short cord length of 16 feet” because it doesn’t have a cord. It’s a portable handheld vacuum.”

Is QA “solved” by Evidence-grounded LLMs?

There is no definite answer, but we can do what every good scientist should. **Hypothesize...**

1. Maybe because
 1. Evidence-grounded models still suffer from hallucination.
 2. LLMs still cannot solve logic well.



Let's call the larger number x and the smaller number y . From the information given, we can set up the following equations:

A.

1. $x = 3y$ (the larger number is three times the smaller number)
2. $y + 20 = x + 6$ (adding 20 to the smaller number equals 6 more than the larger number)

B.

We can substitute the first equation into the second equation:

$$y + 20 = 3y + 6$$

C.

Simplifying, we get:

$$y = 14$$

Figure 1: ChatGPT's response (Jan. 24, 2023) to MWP *One whole number is three times a second. If 20 is added to the smaller number, the result is 6 more than the larger.* In Step A it correctly identifies the set of equations needed to solve the problem and correctly simplifies it in Step B. However, it fails to correctly perform the algebraic operation in Step C (it should state $2y = 14$). This leads ChatGPT to obtain an incorrect result, returning 42 and 14 instead of 21 and 7.

Is QA “solved” by Evidence-grounded LLMs?



Let's call the larger number x and the smaller number y . From the information given, we can set up the following equations:

A.

1. $x = 3y$ (the larger number is three times the smaller number)
2. $y + 20 = x + 6$ (adding 20 to the smaller number equals 6 more than the larger number)

B.

We can substitute the first equation into the second equation:

$$y + 20 = 3y + 6$$

C.

Simplifying, we get:

$$y = 14$$

Figure 1: ChatGPT’s response (Jan. 24, 2023) to MWP *One whole number is three times a second. If 20 is added to the smaller number, the result is 6 more than the larger.* In Step A it correctly identifies the set of equations needed to solve the problem and correctly simplifies it in Step B. However, it fails to correctly perform the algebraic operation in Step C (it should state $2y = 14$). This leads ChatGPT to obtain an incorrect result, returning 42 and 14 instead of 21 and 7.

*EPILOGUE:
TAKEAWAYS*

Takeaways: QA

- Question Answering, Document Retrieval, Fact-Checking, Entity Disambiguation, Multimodal Retrieval, all of this is **information retrieval**.

Takeaways: QA

- Question Answering, Document Retrieval, Fact-Checking, Entity Disambiguation, Multimodal Retrieval, all of this is **information retrieval**.
- Closed-domain QA works well, especially on popular topics (sport, history, tv shows). Bio/scientific domain, math, or technical jargon are still left unattained.

Takeaways: QA

- Question Answering, Document Retrieval, Fact-Checking, Entity Disambiguation, Multimodal Retrieval, all of this is **information retrieval**.
- Closed-domain QA works well, especially on popular topics (sport, history, tv shows). Bio/scientific domain, math, or technical jargon are still left unattained.
- Extractive QA can be tackled **with answer start/end probability estimation**

Takeaways: QA

- Question Answering, Document Retrieval, Fact-Checking, Entity Disambiguation, Multimodal Retrieval, all of this is **information retrieval**.
- Closed-domain QA works well, especially on popular topics (sport, history, tv shows). Bio/scientific domain, math, or technical jargon are still left unattained.
- Extractive QA can be tackled **with answer start/end probability estimation**
- Open-domain QA needs to deal with multi-passage processing, with methods such as **MML** and **cross-passage normalization**.

Takeaways: QA

- Question Answering, Document Retrieval, Fact-Checking, Entity Disambiguation, Multimodal Retrieval, all of this is **information retrieval**.
- Closed-domain QA works well, especially on popular topics (sport, history, tv shows). Bio/scientific domain, math, or technical jargon are still left unattained.
- Extractive QA can be tackled **with answer start/end probability estimation**
- Open-domain QA needs to deal with multi-passage processing, with methods such as **MML** and **cross-passage normalization**.
- **Be sure to check out Czech QA dataset from MU! SQAD (Medved' and Horák, 2014).**

Takeaways: Document Retrieval

- **BM25** is a “fairly popular” baseline from “classic IR” in production-ready systems today. With standard BM25, one have two hyperparameters to control:
 - (a) term saturation
 - (b) long-document bias

More Recent Directions, Literature, etc.

Neural Document Retrieval

Contriever/mContriever — Unsupervisedly pretrained dense retrieval (also multilingual, but no Czech), sometimes matching closely supervised approaches, well generalizing.

LaBSE — Symmetric embeddings for textual similarity (!not query-document) over 109 languages, trained in a supervised way (parallel sentences) and unsupervised way.

ColBERTv2 — SOTA multi-vector learned dense retrieval model, with interesting quantization of residual vectors.

SPLADEv2 — SOTA learned sparse retrieval model.

JPR — Diverse retrieval for multi-answer questions.

Baleen — Multi-hop retrieval for multihop questions.

Open-Domain Question Answering

ATLAS — Unsupervisedly pre-trained evidence-grounded LLM (11B).

REATT — A joint retrieval-reader model for both, retrieval and LM.

DENSEPHRASES — All potential short answers on Wikipedia are encoded into gigantic index, answer is retrieved directly (no reader part!).

Open-Domain Fact-Checking

Claim-Dissector — Our new work on interpretable evidence-grounded fact-checking.

General Model Pretraining

MetalCL — A model pre-trained for learning to learn from context (so-called in-context learning).

LLAMA — Recently released large language model that beats GPT-3/MegaTron despite being order of magnitude smaller.

No links included, IR it out!



STARRING

THE GOOD - MARTIN FAJCIK

THE BAD - MARTIN FAJCIK

BARTENDER - MARTIN FAJCIK

BEAUTY IN TROUBLE - MARTIN FAJCIK

CAMERA - MARTIN FAJCIK

ACTION - MARTIN FAJCIK

PRODUCTION & WRITING - MARTIN FAJCIK

DIRECTOR - ME AGAIN