

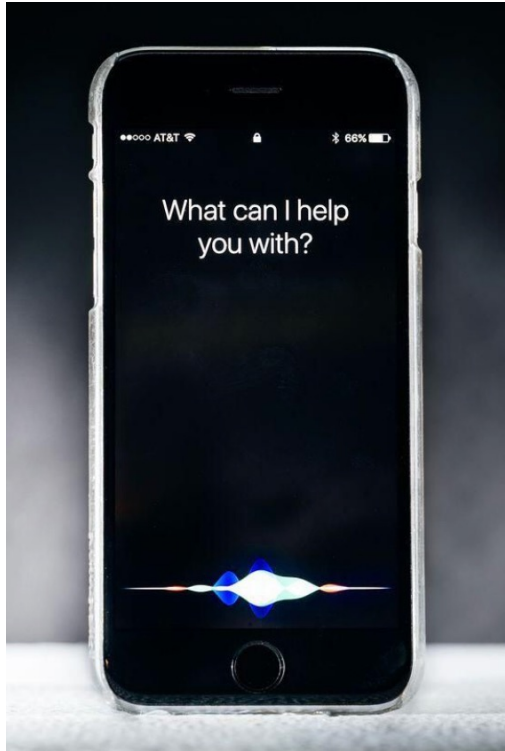
# Natural Language Meets Query Processing

Matei Zaharia




**Berkeley**  
UNIVERSITY OF CALIFORNIA



# NLP is Advancing Fast



```
type Run struct {
    Time int // in milliseconds
    Results string
    Failed bool
}

// Get average runtime of successful runs in seconds
func averageRuntimeInSeconds(runs []Run) float64 {
    var totalTime int
    var failedRuns int
    for _, run := range runs {
        if run.Failed {
            failedRuns++
        } else {
            totalTime += run.Time
        }
    }
    averageRuntime := float64(totalTime) / float64(len(runs))
    return averageRuntime
}
```

 What can I visit in Napa in one afternoon?

  Make the most of your afternoon in Napa Valley with these must-visit spots:

1. Wineries: Pick a couple of wineries for tastings, such as Domaine Chandon, Robert Mondavi, or Castello di Amorosa.
2. Oxbow Public Market: Enjoy local produce, artisanal food, and unique shopping experiences.
3. Downtown Napa: Stroll through the charming streets, boutique shops, and art galleries.
4. Napa Valley Wine Train: Hop on for a scenic train ride, even if you don't have time for the full dining experience.
5. Napa Art Walk: Admire outdoor sculptures while taking a leisurely walk.

+ getting democratized via smaller models (Llama, Alpaca, **Dolly**, Vicuna, Koala, ...)

# NLP will Simplify Analytics

Cmd 1

```
1 %chat
2 what are the top five teams that received sev0s outside of working hours in past two weeks?
```

To find the top five teams that received sev0s outside of working hours in the past two weeks, we can use the following query in Spark SQL:

```
...
SELECT TeamName, COUNT(*) AS Sev0Count
FROM q4_hackathon.main.pagerduty_incidents
WHERE Timestamp NOT BETWEEN '09:00:00' AND '17:00:00'
AND Severity = 'sev0'
AND Timestamp >= CURRENT_TIMESTAMP - INTERVAL 2 WEEKS
GROUP BY TeamName
ORDER BY Sev0Count DESC
LIMIT 5
...
```

This query first filters the incidents to only include those with a severity of 'sev0' that occurred outside of working hours (defined as between 09:00:00 and 17:00:00). It then counts the number of incidents per team. Finally, it orders the teams by the number of incidents and returns the top five teams.

Command took 43.64 seconds -- by kyle.rong+e2@databricks.com at 12/10/2022, 3:40:04 PM on Kyle hackathon

Cmd 2

```
1 %sql
2
3 SELECT TeamName, COUNT(*) AS Sev0Count
4 FROM q4_hackathon.main.pagerduty_incidents
5 WHERE Timestamp NOT BETWEEN '09:00:00' AND '17:00:00'
6 AND Severity = 'sev0'
7 AND Timestamp >= CURRENT_TIMESTAMP - INTERVAL 2 WEEKS
8 GROUP BY TeamName
9 ORDER BY Sev0Count DESC
10 LIMIT 5
```

# But Modern NLP still has Problems



How large are elephant eggs?

**Wrong**



Elephants have the largest eggs of any land animal. A single elephant egg can weigh as much as 2 pounds (about 0.9 kilograms), and is about the size of a grapefruit. However, despite their large size, elephant eggs are much smaller in proportion to the size of the mother than the eggs of most other animals. For example, a human egg is about the size of a small pea in relation to the mother's body size, whereas an elephant egg is more like a golf ball in relation to the size of an adult elephant.



Who are the current World Cup champions?

**Out-of-date**

**Expensive**



The current World Cup champions are France. They won the 2018 FIFA World Cup, which was held in Russia.

Jan 10, 2023

**High latency and error rates on text-davinci-003**

**Resolved** - This incident has been resolved.

Jan 10, 11:04 PST

**Investigating** - We are experiencing error rates of around 20%

# And It's Hard to Fix



Why did Ion Stoica win the Turing Award?



Ion Stoica has not won the Turing Award as of my knowledge cutoff date of September 2021, and there has been no announcement of him winning the award since then. The Turing Award is one of the most prestigious awards in computer science, awarded annually by the Association for Computing Machinery (ACM) to individuals who have made significant contributions to the field of computing.

It is possible that Ion Stoica may be considered for the award in the future if he makes significant contributions to the field of computing. However, at the time of my knowledge cutoff date, he has not yet received the award.

# And It's Hard to Fix



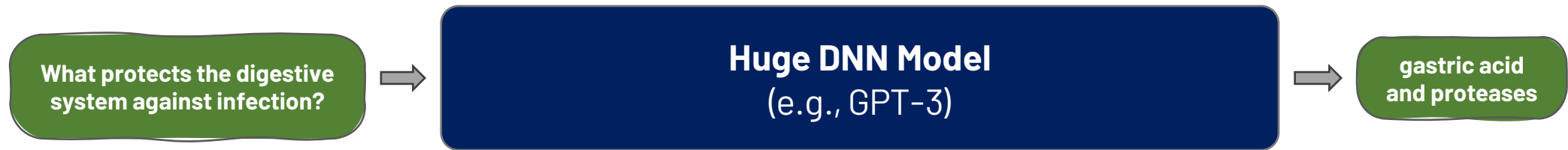
Why did Matei Zaharia win the Turing Award?



Matei Zaharia won the 2021 ACM A.M. Turing Award, also known as the Nobel Prize of Computing, for his foundational contributions to large-scale data processing. He was recognized for his work on Apache Spark, an open-source cluster computing system that has become one of the most widely used data processing platforms in the world.

# Large-Scale NLP Today

Huge DNNs that store “knowledge” in their weights



Problems:

- Limited interpretability
- Hard to update model's knowledge
- Expensive training and inference

# From a Systems Point of View

Accessing all weights to run a task seems very inefficient

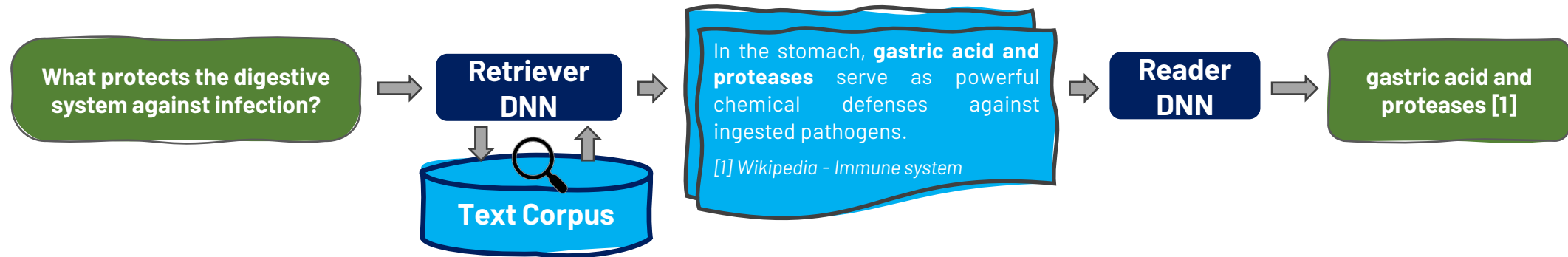
- Like a linear scan through our “knowledge”



# Retrieval-Based NLP

Examples: REALM, DPR, RAG, RETRO, CoBERT, DSP

Separate “language processing” from “knowledge”



Benefits:

- Easier to interpret and “program”
- Can update knowledge in milliseconds by updating a doc
- 100-1000x faster inference

# Our Results with Retrieval Models

With Omar Khattab, Keshav Santhanam, Chris Potts



SOTA performance on multiple NLP tasks at lower compute cost

- **Information retrieval:** CoBERT (SIGIR'20) can match BERT-based retrievers at 100-1000x lower cost
- **Question answering:** CoBERT-QA (TACL'21) improves EM scores for TriviaQA and Natural Questions by 3-12 points
- **Multi-hop reasoning:** Baleen (NeurIPS'21) improves score on HoVer from 15 to 57 and runs 10x faster; DSP offers general programming model

[tinyurl.com/rnlp21](https://tinyurl.com/rnlp21)

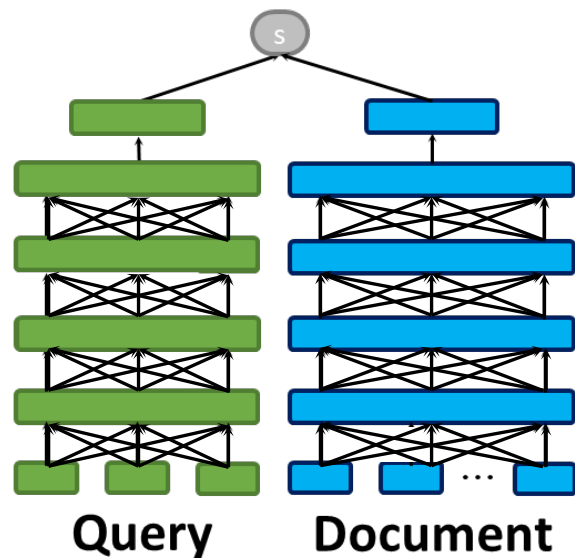
# How Did We Get These Results?

**Improvements in retrieval:** “late interaction” approach that keeps the modeling benefits of Transformers while enabling efficient retrieval

**Supervision for retrieval-based models:** teaching models how to search for relevant documents given only the final answer for a task

**Data systems insights:** improved indexes and query plans

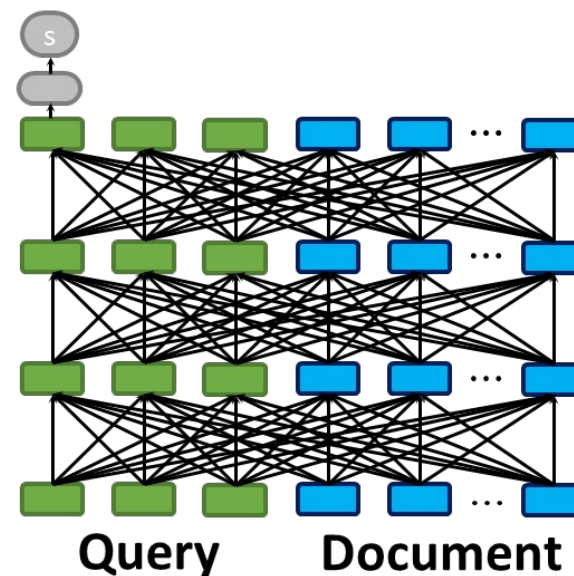
# Previous Neural Retrieval Approaches



**(a) Encoding similarity**

e.g. run BERT separately on query & doc

- ✓ Cheap search computation
- ✗ Coarse-grained representations

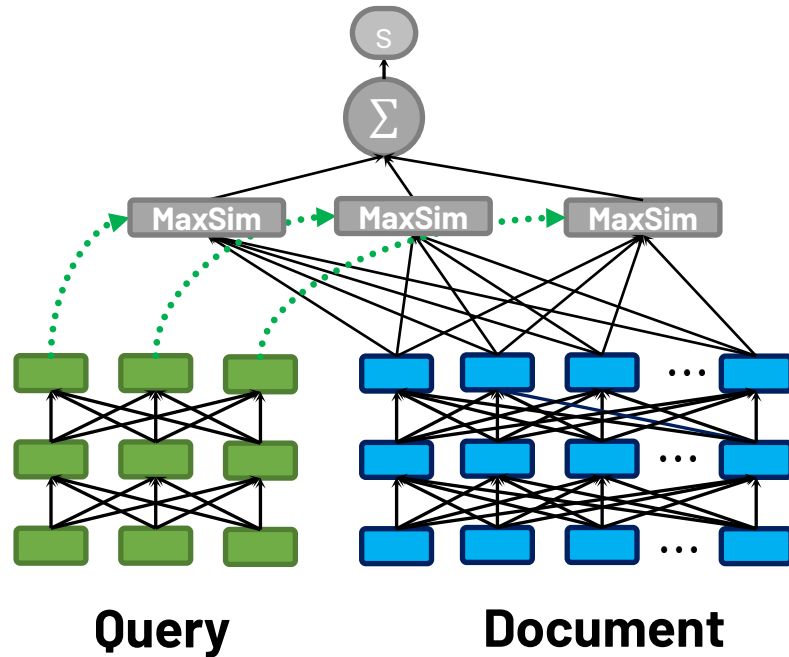


**(b) All-to-all interaction**

e.g. BERT on query || doc

- ✓ Joint contextualization of terms
- ✗ Expensive to compute on all docs

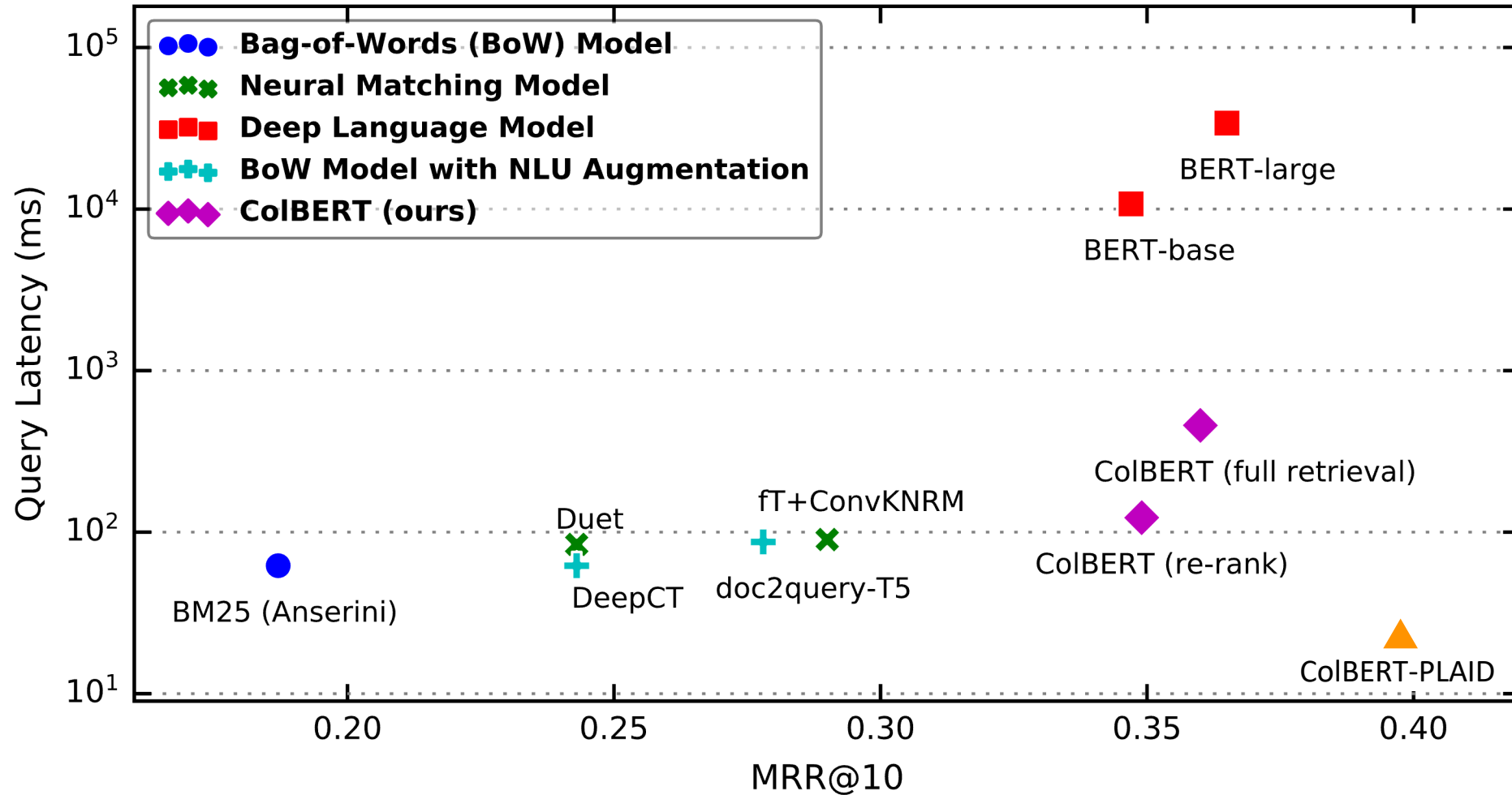
# ColBERT's Approach: Late Interaction



Independent encoding using all but the last layers of BERT

- ✓ Fine-grained representations
- ✓ Joint contextualization of terms
- ✓ Scalable search computation

# Retrieval Results on MS MARCO



# Example of CoBERT Matching

**when did the transformers cartoon series come out?**

[...] the animated [...] The Transformers [...] [...] It was released [...] **on** August 8, 1986

**when did the transformers cartoon series come out?**

[...] the animated [...] The **Transformers** [...] [...] It was released [...] on August 8, 1986

**when did the transformers cartoon series come out?**

[...] the **animated** [...] The Transformers [...] [...] It was released [...] on August 8, 1986

# Question Answering with CoBERT-QA

---

“Where does the Volga river end?” → Caspian Sea

---



# Challenge: QA Retrievers are Hard to Supervise

The training data for QA has the form

⟨ **Question** where does the volga river end, **Answer** Caspian Sea ⟩

But each **retriever** training example needs to be of the form

⟨ **Question**, **Positive Passage(s)**, **Negative Passage(s)** ⟩

And we train the retriever to give higher scores to the positives

# Challenge: QA Retrievers are Hard to Supervise

The training data for QA has the form

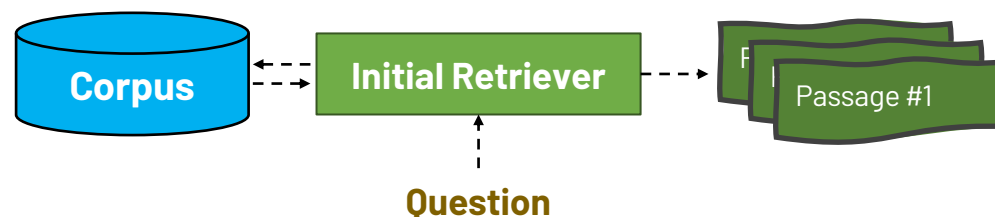
⟨ **Question** where does the volga river end, **Answer** Caspian Sea ⟩

But each **retriever** training example needs to be of the form

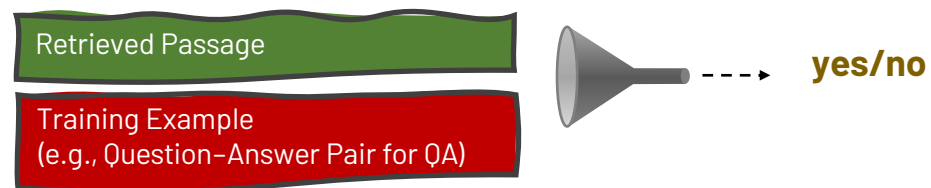
⟨ **Question**, **Positive Passage(s)**, **Negative Passage(s)** ⟩

And we t How do we collect positives and negatives? ves.

# RGS: Relevance-Guided Supervision



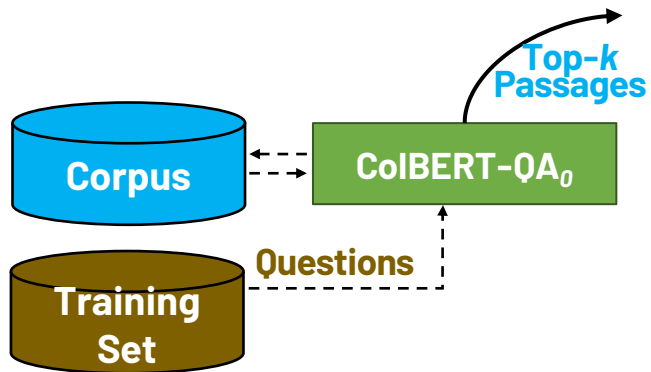
A weak initial retriever – e.g., BM25 or ColBERT trained for standard IR



A task-aware heuristic for “useful” passages: for OpenQA, “does *answer* appear in *passage*?”

# RGS: **Outer-Loop** Batch Retrieval, **Inner-Loop** Fast Training

## 1. **Initial Retriever:** Find the top-1000 passages per training question



**Q: where does the volga river end**  
**A: caspian sea**

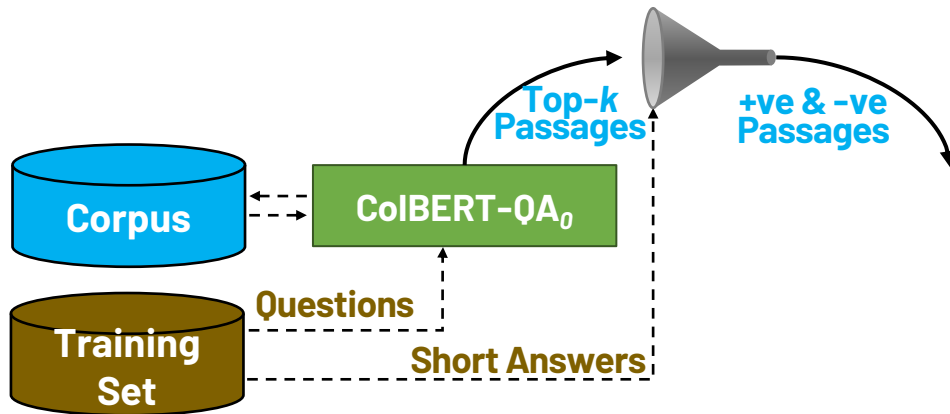
**Q: who won world cup 2016**  
**A: real madrid**

### **Q: where does the volga river end**

1. Volga Township lies in Clayton County (Iowa), named after the Volga River.
2. The Akhtuba river flows toward the Volga Delta and Caspian Sea.
3. The Volga is an Executive car from the Soviet Union to replace GAZ Pobeda.  
...
10. The Caspian Sea is home to a wide range of species, known for caviar and oil industries.  
...

# RGS: **Outer-Loop** Batch Retrieval, **Inner-Loop** Fast Training

**2. Weak Heuristic:** Identify the highest-ranked passages that pass the filter (contain the answer string) as weak positives



**Q: where does the volga river end**  
**A: caspian sea**

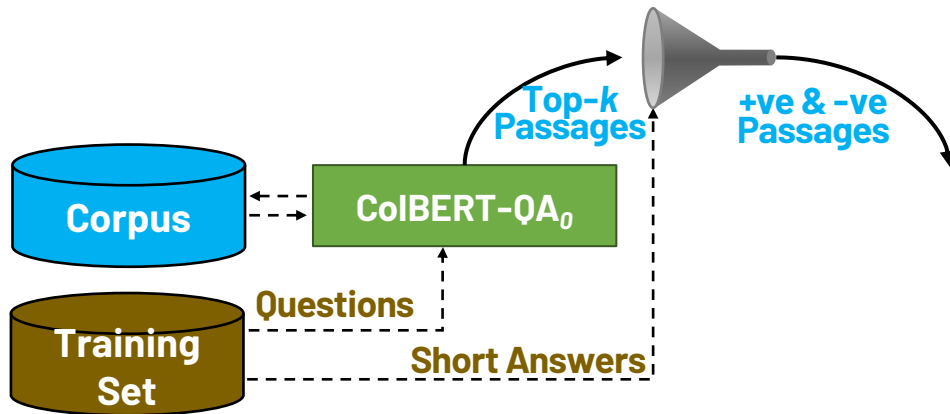
**Q: who won world cup 2016**  
**A: real madrid**

**Q: where does the volga river end**

1. Volga Township lies in Clayton County (Iowa), named after the Volga River.
2. The Akhtuba river flows toward the Volga Delta and Caspian Sea.
3. The Volga is an Executive car from the Soviet Union to replace GAZ Pobeda.  
...
10. The Caspian Sea is home to a wide range of species, known for caviar and oil industries.  
...

# RGS: **Outer-Loop** Batch Retrieval, **Inner-Loop** Fast Training

**2. Weak Heuristic:** Identify the highest-ranked passages that pass the filter (contain the answer string) as weak positives



**Q: where does the volga river end**  
**A: caspian sea**

**Q: who won world cup 2016**  
**A: real madrid**

**Q: where does the volga river end**

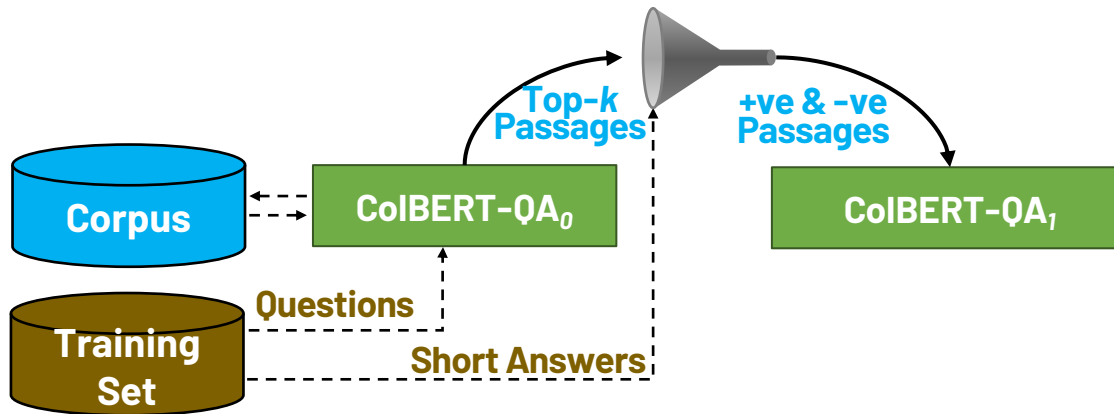
1. Volga Township lies in Clayton County (Iowa), named after the Volga River.
2. **The Akhtuba river flows toward the Volga Delta and Caspian Sea.**
3. The Volga is an Executive car from the Soviet Union to replace GAZ Pobeda.  
...
10. **The Caspian Sea is home to a wide range of species, known for caviar and oil industries.**  
...

# RGS: **Outer-Loop** Batch Retrieval, **Inner-Loop** Fast Training

## 3. **Inner-Loop Training:** Collect and cache triples of the form

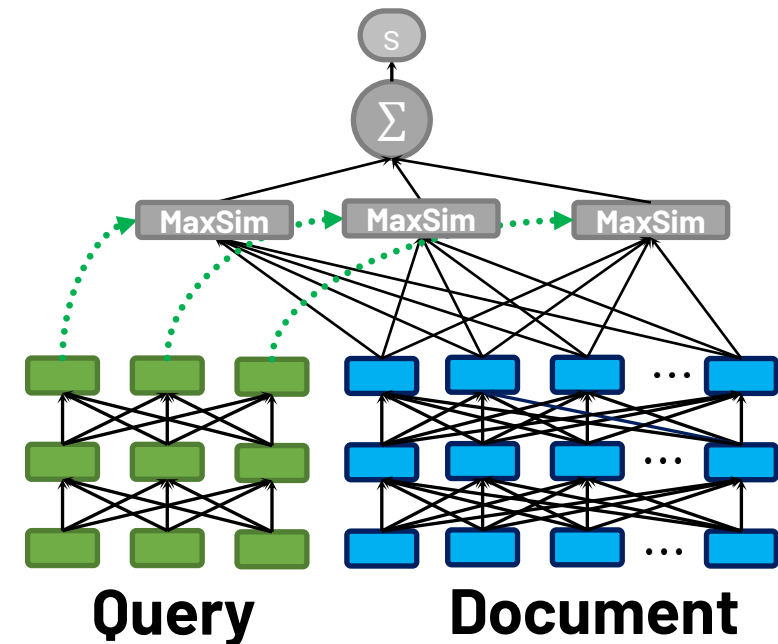
**( Question, Weak Positive Passage, Sampled Negative Passage )**

and train with them!



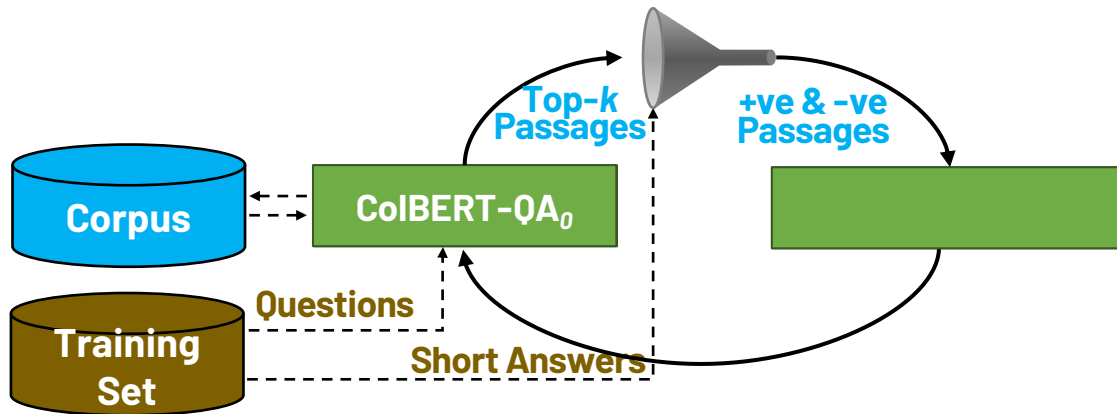
**Q:** where does the volga river end  
**A:** caspian sea

**Q:** who won world cup 2016  
**A:** real madrid



# RGS: **Outer-Loop** Batch Retrieval, **Inner-Loop** Fast Training

**4. Outer-Loop Refresh:** Encode the corpus with the new retriever, and retrieve the top-k passages once more.



**Q: where does the volga river end**  
**A: caspian sea**

**Q: who won world cup 2016**  
**A: real madrid**

**Q: where does the volga river end**

- 1. Volga River discharges into the Caspian Sea below Astrakhan at below sea level.**
- 2. The Volga River flows through central Russia and into the Caspian Sea.**



# Results for Question Answering

	End-to-End QA Exact Match (EM)		
Model	Natural Questions	TriviaQA	SQuAD
BM25 + BERT	32.6	52.4	38.1 / 53.0
REALM	40.4	-	-
DPR	41.5	57.9	36.7
RAG	44.5	56.1	-
CoBERT-QA (3 rounds)	<b>47.8</b>	<b>70.1</b>	<b>54.7 / 58.7</b>
T5-11B (24x more params)	<b>34.8</b>		
GPT-3 (400x more params)	<b>29.9</b>		

# Multi-Hop Reasoning with Baleen

---

$Q_0$  The MVP of [a] game Red Flaherty umpired was elected to the Baseball Hall of Fame.

---

Is this claim true or false?

# Multi-Hop Reasoning with Baleen

---

$Q_0$  The MVP of [a] game Red Flaherty umpired was elected to the Baseball Hall of Fame.

---

$Q_1$  The MVP of [a] game Red Flaherty umpired was elected to the Baseball Hall of Fame. **Red Flaherty:** He umpired in World Series 1955, 1958, 1965, and 1970.

---

# Multi-Hop Reasoning with Baleen

---

$Q_0$  The MVP of [a] game Red Flaherty umpired was elected to the Baseball Hall of Fame.

---

$Q_1$  The MVP of [a] game Red Flaherty umpired was elected to the Baseball Hall of Fame. **Red Flaherty:** He umpired in World Series 1955, 1958, 1965, and 1970.

---

$Q_2$  The MVP of [a] game Red Flaherty umpired was elected to the Baseball Hall of Fame. Red Flaherty: He umpired in World Series 1955, 1958, 1965, and 1970. **1965 World Series:** It is remembered for MVP Sandy Koufax.

---

# Multi-Hop Reasoning with Baleen

---

$Q_0$  The MVP of [a] game Red Flaherty umpired was elected to the Baseball Hall of Fame.

---

$Q_1$  The MVP of [a] game Red Flaherty umpired was elected to the Baseball Hall of Fame. **Red Flaherty:** He umpired in World Series 1955, 1958, 1965, and 1970.

---

$Q_2$  The MVP of [a] game Red Flaherty umpired was elected to the Baseball Hall of Fame. Red Flaherty: He umpired in World Series 1955, 1958, 1965, and 1970. **1965 World Series:** It is remembered for MVP Sandy Koufax.

---

$Q_3$  The MVP of [a] game Red Flaherty umpired was elected to the Baseball Hall of Fame. Red Flaherty: He umpired in World Series 1955, 1958, 1965, and 1970. 1965 World Series: It is remembered for MVP Sandy Koufax. **Sandy Koufax:** He was elected to the Baseball Hall of Fame.

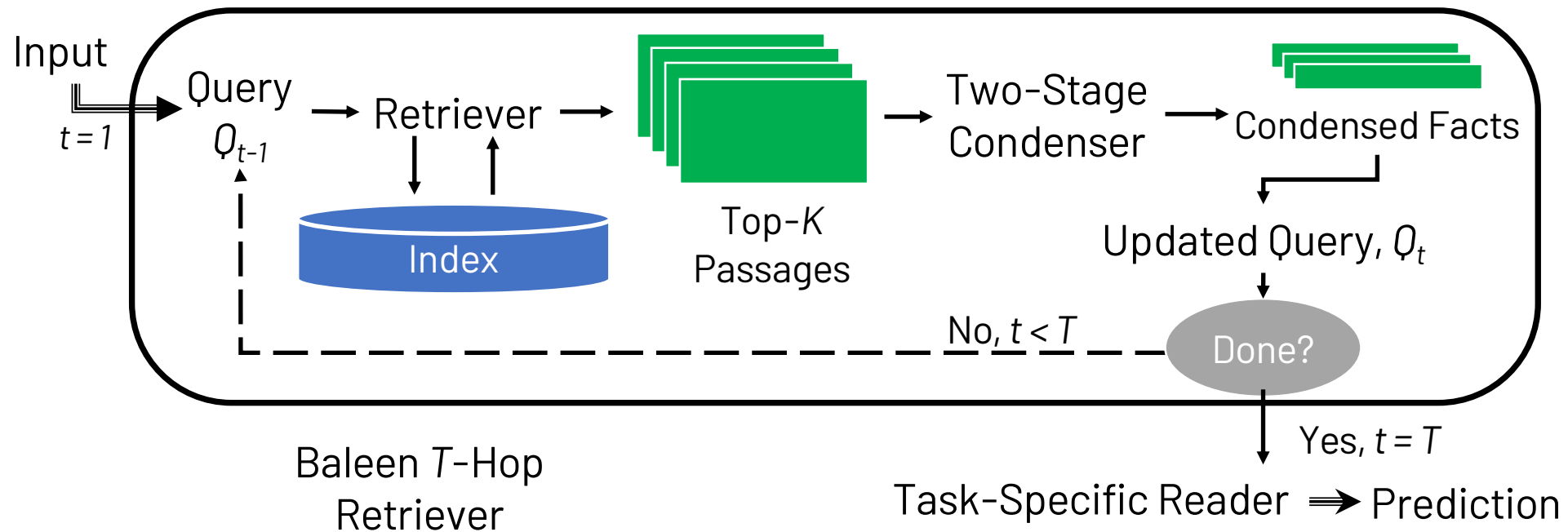
---

→ True!



# Multi-Hop Reasoning Challenges

1. Multi-hop queries have multiple information needs
2. Retrieval errors in each hop propagate to subsequent hops
3. Don't have supervision on which passages to retrieve and which order (like our QA situation, but worse!)

# Baleen Overview



# Results for Multi-Hop Reasoning

	Model	Code	Fact Extraction		HoVer Score
			EM	F <sub>1</sub>	Acc
1 May 24, 2021	Baleen <i>Anonymous</i> <a href="#">Anonymous</a>		39.78	80.41	57.53
2 Oct 13, 2020	Baseline Model (single model) <i>UNC Chapel Hill &amp; Verisk Analytics</i> <a href="#">(Jiang, Bordia, et al. 2020)</a>		4.5	49.5	15.32



# How Do We Program Complex NLP Apps?

CoBERT-QA, Baleen, etc are nice, but required a PhD student to write

Our new project is a high-level programming model for NLP apps:

## **Demonstrate-Search-Predict (DSP)**

Key ideas:

- Start with pretrained foundation models (e.g. CoBERT, GPT-4)
- Write pipelines of operators that exchange plain text
- 3 kinds of operators: Demonstrate (find task examples), Search (over examples or knowledge), Predict
- “Compile” a pipeline to get faster and more accurate models!

# DSP Example: Multi-Hop QA

How many storeys are in the castle David Gregory inherited?

Vanilla LM

LM: Castle Gregory has three storeys.

✗ Hallucinates  
a fictitious castle

Retrieve-  
then-Read

RM: "St. Gregory Hotel is a nine-floor boutique hotel in D.C..."

LM: St. Gregory Hotel has nine storeys.

✗ Retrieves a  
different building

Multi-Hop  
DSP Program

LM: "Which castle did David Gregory inherit?"

RM: "David Gregory inherited Kinnairdy Castle in 1664..."

LM: "How many storeys does Kinnairdy Castle have?"

RM: "Kinnairdy Castle is a tower house, having five storeys..."

LM: Kinnairdy Castle has five storeys.



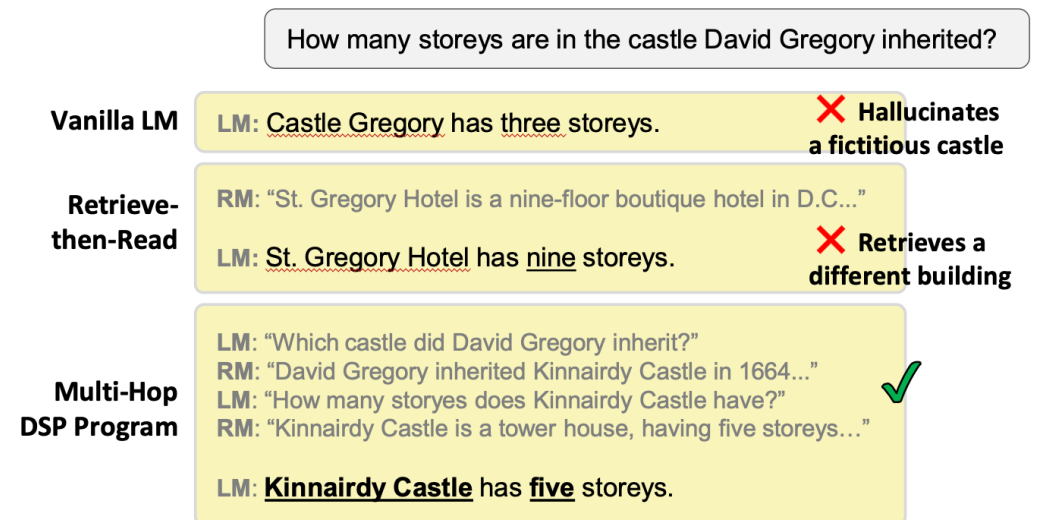
# DSP Model: LMs and retrieval models (RMs) exchange text in sophisticated pipelines

LMs and RMs both consume (and generate or retrieve) natural language text.

Instead of focusing on carefully engineering prompts, let's **write a program** whose leaves are **declarative** *generate* or *retrieve* calls.

e.g., `write_search_query(context, question) -> query`

The **framework runtime** will decide how to effectively map this to a model call (e.g., a few-shot prompt or fine-tuning a new model)



# DSP Example: Multi-Hop QA

“How many storeys are in the castle David Gregory inherited?”



Q		How many storeys are in...
Train	Q	When was the discoverer of Palomar 4 born?
	A	1889
	Q	In which city did Akeem Ellis play in 2017?
	A	Ellesmere Port

x : Example

Q		How many storeys are in the castle...
Demos	Q	When was the discoverer of Palomar 4 born?
	A	1889
	Hop1	Who discovered Palomar 4?
	Psg1	Edwin Hubble discovered Palomar 4...
	Hop2	When was Edwin Powell born?
	Psg2	Edwin Powell Hubble (1889-1953) was...
	Pred	1889 ✓
	Q	In which city did Akeem Ellis play...
	A	Ellesmere Port
	Pred	Waterloo ✗

2

Search

```
def search(x: Example) -> Example:
    x.hop1 = generate(hop_template)(x).pred
    x.psg1 = retrieve(x.hop1, k=1)[0]
    x.hop2 = generate(hop_template)(x).pred
    x.psg2 = retrieve(x.hop2, k=1)[0]
    return x
```

Q		How many storeys are in the...
Demos	...	
Hop1	Which castle did David Gregory inherit?	
Psg1	David Gregory inherited Kinnairdy Castle...	
Hop2	How many storeys are in Kinnairdy Castle?	
Psg2	Kinnairdy Castle [...] having five storeys...	

3

Predict

```
def predict(x: Example) -> Example:
    x.context = [x.psg1, x.psg2]
    x.pred = generate(qa_template)(x).pred
    return x
```

Q		How many storeys does the...
...	...	
Pred	Five storeys	

“Five storeys”

# DSP Compiler

Take a DSP program + unlabeled inputs and compile it:

```
compiled_QA = dsp.compile(program=multihop_QA,  
                           examples=unlabeled_questions_800)
```

Compiler automatically explores:

- Using smaller foundation models for each step (e.g., T5, LLaMa, Ada)
- Fine-tuning the models for each step
- Picking the best demonstrations for each step

Result: Similar quality at 1/10<sup>th</sup> the cost

# The Evolution of Programming NLP Apps?

Paradigm	Examples	Frameworks
<b>Training</b> bespoke architectures	BiLSTMs with attention, etc.	TensorFlow, PyTorch
<b>Fine-tuning</b> pretrained Transformers	BERT, ELECTRA, etc.	HuggingFace
<b>Prompting</b> instruction-tuned LLMs	GPT-3, Flan-T5, etc.	OpenAI/Cohere, LangChain (with tool use)
<b>Programming</b> FM pipelines	Socratic Models, RARR, DSP programs, etc.	<b>DSP</b>



Eric Zhu

@ekzhu



The example in the first figure is handled by Bing Chat. Maybe it is more important to generate good keywords from input question? Web search engines today can already do it for you.

The screenshot shows a Bing Chat interface. The user's question is: "How many storeys are in the castle David Gregory inherited?". The chatbot's response is: "Kinnairdy Castle has five storeys." The response includes numbered references (1-7) and a list of sources at the bottom: 1. en.wikipedia.org, 2. en.wikipedia.org, 3. wikiwand.com, 4. arxiv.org. The text in the image is partially obscured by yellow highlights.

12:34 AM · Feb 21, 2023 · 131 Views

1 Like



Eric Zhu @ekzhu · Feb 21



Upon close inspection I noticed Bing Chat is using the arxiv paper in its answer. It is self-improving!



1



2



127



# Conclusion

LLMs will just be one building block in powerful NLP applications

Learn more about our work:

- Retrieval-based NLP: [tinyurl.com/rnlp21](https://tinyurl.com/rnlp21)
- DSP: [github.com/stanfordnlp/dsp](https://github.com/stanfordnlp/dsp)
- Dolly open LLM: [github.com/databricks/dolly](https://github.com/databricks/dolly)



# Work Based on CoBERT

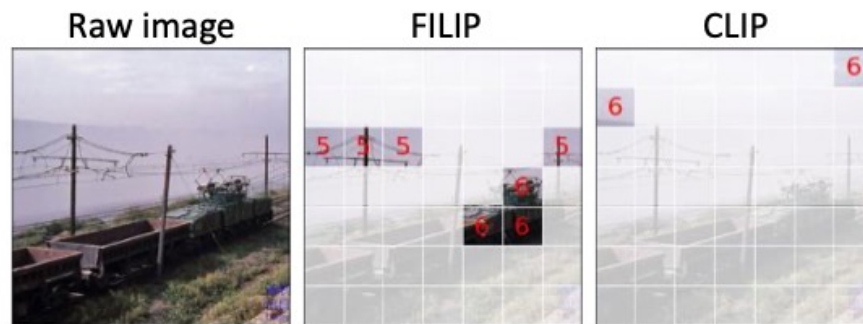
Cross-lingual QA 

XOR-TyDi Cross-lingual Open-Retrieval Question Answering			
Rank	Model	R@5kt	R@2kt
1 February 11, 2022	Single model CHIMAERAS B <i>Anonymous</i>	70.3	63.0
2 January 7, 2022	Contrastive Context-aware Pretraining Model (CCP) <i>Anonymous</i>	63.0	54.8
3 August 26, 2021	Single Encoder Retriever (Sentri) <i>Huawei Noah's Ark lab</i>	61.0	52.7
4 October 7, 2021	Single Encoder Retriever (Sentri, resubmission)	60.7	55.5

Text generation **Stanford**

KILT		Organized by: KILT						
KILT		Starts on: Dec 31, 2018 4:00:00 PM						
KILT		Ends on: May 31, 2019 4:59:59 PM						
Rank	Participant team	R-Prec (↑)	Recall@5 (↑)	ROUGE-L (↑)	F1 (↑)	KILT-RL (↑)	KILT-F1 (↑)	Last submission at
1	Stanford NLP (Hindsight)	56.08	74.27	17.06	19.19	11.92	13.39	3 months ago
2	Re2G (Re2G)	60.10	79.98	16.76	18.90	11.39	12.98	5 months ago
3	IBM_research_AI (KGI)	55.37	78.45	16.36	18.57	10.36	11.79	6 months ago

Image + text embedding  HUAWEI



(d) Electric locomotive (5,6)

	Flickr30K						MSCOCO					
	image-to-text			text-to-image			image-to-text			text-to-image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Unicoder-VL	64.3	85.8	92.3	48.4	76.0	85.2	—	—	—	—	—	—
ImageBERT	70.7	90.2	94.0	54.3	79.6	87.5	44.0	71.2	80.4	32.3	59.0	70.2
UNITER	83.6	95.7	97.7	68.7	89.2	93.9	—	—	—	—	—	—
CLIP	88.0	98.7	99.4	68.7	90.6	95.2	58.4	81.5	88.1	37.8	62.4	72.2
ALIGN	88.6	98.7	99.7	<b>75.7</b>	<b>93.8</b>	<b>96.8</b>	58.6	83.0	89.7	45.6	69.8	78.6
<b>FILIP</b>	<b>89.8</b>	<b>99.2</b>	<b>99.8</b>	75.0	93.4	96.3	<b>61.3</b>	<b>84.3</b>	<b>90.4</b>	<b>45.9</b>	<b>70.6</b>	<b>79.3</b>

# Baleen Results by # of Required Hops

Model / # of Hops	Sentence EM				Sentence F1				Verification Accuracy
	All	2	3	4	All	2	3	4	
TF-IDF + BERT*	4.8/4.5	13.6	1.9	0.2	50.6/49.5	57.2	49.8	45.0	73.7
Baleen 1-hop	19.7	40.9	15.4	4.3	72.3	77.5	72.4	66.4	-
Baleen 2-hop	37.0	46.9	35.7	28.4	80.8	81.2	81.8	78.7	-
Baleen 3-hop	38.9	47.1	37.0	33.2	81.4	<b>81.2</b>	82.3	<b>80.0</b>	-
Baleen 4-hop	<b>39.2/39.8</b>	<b>47.3</b>	<b>37.7</b>	<b>33.3</b>	<b>81.5/80.4</b>	<b>81.2</b>	<b>82.5</b>	<b>80.0</b>	<b>84.5/84.9</b>
Oracle + BERT*	19.9	25.0	18.4	17.1	71.9	68.3	71.5	76.4	81.2
Human*	56.0	75.0	73.5	42.1	88.7	86.5	93.1	87.3	88.0

# Compressing CoBERT

CoBERT vectors cluster very well, aligning with intuition for objective

Encodings with about **20 bytes** per vector preserve 99% of quality

And work robustly in and out of domain

[github.com/stanford-futuredata/CoBERT](https://github.com/stanford-futuredata/CoBERT)