Putting QA back in MRQA

Antoine Bordes with Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, Michael Auli, Chloe Braud, Claire Gardent, Emily Dinan, Stephen Roller, Kurt Schuster

MRQA Workshop -- 11/4/19 -- Hong Kong

facebook
Artificial Intelligence Research

Friday June 17, 2016

AKBC 2016

5th Workshop on Automated Knowledge Base Construction (AKBC) 2016

at NAACL 2016 in San Diego, California, June 17, 2016.

Tweet

Like 0 Share

Friday, June 17th, 2016

Start	End	Speaker	Title
9:00	9:10	AKBC Organizers	Opening Remarks
9:10	9:40	**Kristina Toutanova**	**Joint Compositional Learning from Text and Knowledge Bases** (slides)
9:40	10:10	**Oren Etzioni**	**The Allen AI Science Challenge: Results, Lessons, and Open Questions** (slides)
10:10	11:00		Morning Poster Session and Coffee Break
11:00	11:30	**Andrew McCallum**	**Universal Schema for Representation and Reasoning from Natural Language**
11:30	12:00	**William Cohen**	**Look Ma, No Neurons: Using Explicit Inference Rules to Complete a KB** (slides)
12:00	1:20		Lunch Break and Morning Posters
1:20	1:35	Stuart Russell, Ole Torp Lassen, Justin Uang, Wei Wang	**Contributed Talk**: The Physics of Text: Ontological Realism in Information Extraction (slides)
1:35	1:50	Sreyasi Nag Chowdhury, Niket Tandon, Gerhard Weikum	**Contributed Talk**: Know2Look: Commonsense Knowledge for Visual Search (slides)
1:50	2:15	**Christopher Manning**	**Texts as Knowledge Bases** (slides)
2:15	2:40	**Benjamin Van Durme**	**Common Sense and Language**
2:40	2:55	Patrick Verga, Andrew McCallum	**Contributed Talk**: Row-less Universal Schema (slides)
2:55	3:10	Sonse Shimaoka, Pontus Stenetorp, Kentaro Inui, Sebastian Riedel	**Contributed Talk**: An Attentive Neural Architecture for Fine-grained Entity Type Classification (slides)
3:10	4:00		Afternoon Poster Session and Coffee Break
4:00	4:25	**Percy Liang**	**Querying Unnormalized and Incomplete Knowledge Bases** (slides)
4:25	4:50	**Antoine Bordes**	**Memory Networks for Language Understanding: Successes and Challenges** (slides)
4:50	5:30	Christopher Manning, Benjamin Van Durme, Percy Liang, Antoine Bordes	Afternoon Speaker Panel
5:30	5:45	AKBC Organizers	Closing Remarks

No PyTorch

No BERT!

No Transformers even!

No SQuAD!!!

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

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

(Submitted on 16 Jun 2016 (v1), last revised 11 Oct 2016 (this version, v3))

We present the Stanford Question Answering Dataset (SQuAD), a new reading comprehension dataset consisting of 100,000+ questions posed by crowdworkers on a set of Wikipedia articles, where the answer to each question is a segment of text from the corresponding reading passage. We analyze the dataset to understand the types of reasoning required to answer the questions, leaning heavily on dependency and constituency trees. We build a strong logistic regression model, which achieves an F1 score of 51.0%, a significant improvement over a simple baseline (20%). However, human performance (86.8%) is much higher, indicating that the dataset presents a good challenge problem for future research.
The dataset is freely available at **this https URL**

Comments: To appear in Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)
Subjects: **Computation and Language (cs.CL)**
Cite as: **arXiv:1606.05250** [cs.CL]
(or **arXiv:1606.05250v3** [cs.CL] for this version)

Bibliographic data

[**Enable Bibex**(**What is Bibex?**)]

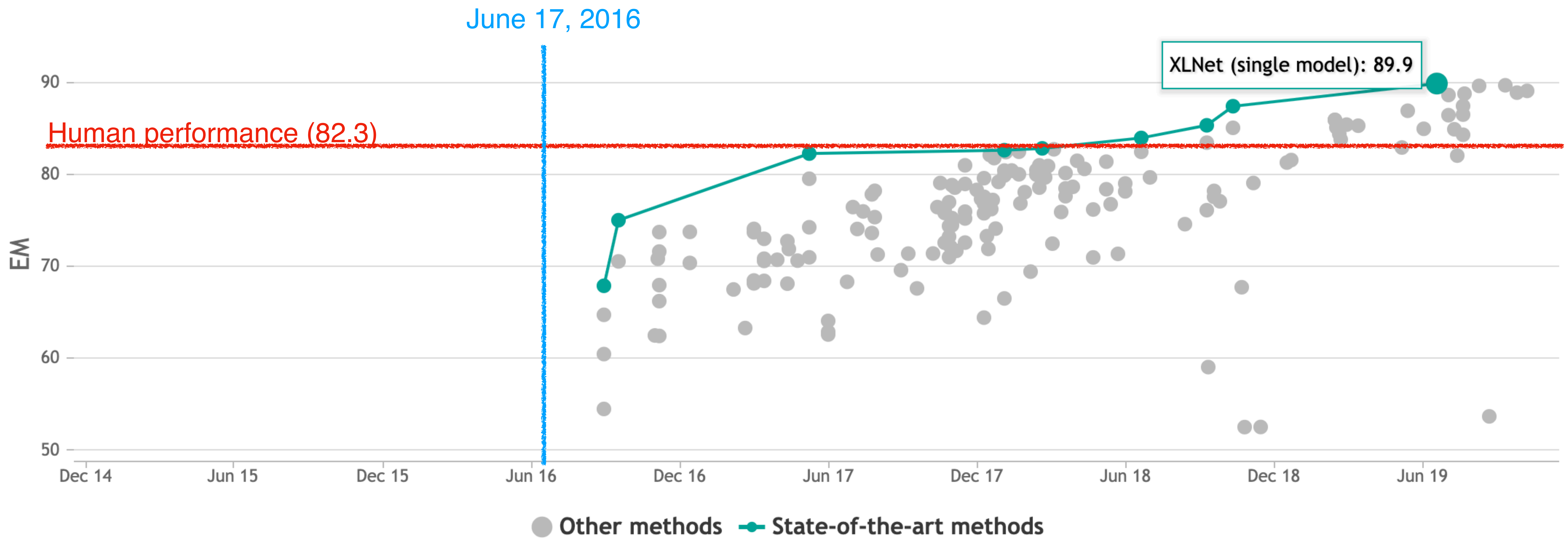
Submission history

From: Pranav Rajpurkar [**view email**]
[v1] Thu, 16 Jun 2016 16:36:00 UTC (307 KB)
[v2] Fri, 7 Oct 2016 03:48:29 UTC (307 KB)
[v3] Tue, 11 Oct 2016 02:42:36 UTC (307 KB)

SQuAD had appeared on Arxiv the day before the workshop!

Since then?

Question Answering on SQuAD1.1



https://paperswithcode.com/sota/question-answering-on-squad11

October 24, 2019

Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

Colin Raffel* Noam Shazeer* Adam Roberts* Katherine Lee*
Sharan Narang Michael Matena Yanqi Zhou Wei Li Peter J. Liu
Google

Model	GLUE Average	CoLA Matthew's	SST-2 Accuracy	MRPC F1	MRPC Accuracy	STS-B Pearson	STS-B Spearman
Previous best	89.4[a]	69.2[b]	**97.1**[a]	**93.6**[b]	**91.5**[b]	**92.7**[b]	**92.3**[b]
T5-Small	77.4	41.0	91.8	89.7	86.6	85.6	85.0
T5-Base	82.7	51.1	95.2	90.7	87.5	89.4	88.6
T5-Large	86.4	61.2	96.3	92.4	89.9	89.9	89.2
T5-3B	88.5	67.1	97.4	92.5	90.0	90.6	89.8
T5-11B	**89.7**	**70.8**	**97.1**	91.9	89.2	92.5	92.1

Model	QQP F1	QQP Accuracy	MNLI-m Accuracy	MNLI-mm Accuracy	QNLI Accuracy	RTE Accuracy	WNLI Accuracy
Previous best	**74.8**[c]	**90.7**[b]	91.3[a]	91.0[a]	**99.2**[a]	89.2[a]	91.8[a]
T5-Small	70.0	88.0	82.4	82.3	90.3	69.9	69.2
T5-Base	72.6	89.4	87.1	86.2	93.7	80.1	78.8
T5-Large	73.9	89.9	89.9	89.[illegible]	94.8	87.2	85.6
T5-3B	74.4	89.7	91.4	[illegible]	96.3	91.1	89.7
T5-11B	74.6	90.4	**92.0**	[illegible]	96.7	**92.5**	**93.2**

Model	SQuAD EM	SQuAD F1	Super[illegible]	[illegible]	CB [illegible]	CB [illegible]	COPA [illegible]
Previous best	88.95[d]	94.52[d]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
T5-Small	79.10	87.24	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
T5-Base	85.44	92.08	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
T5-Large	86.66	93.79	82.[illegible]	[illegible]	91.6	94.8	83.4
T5-3B	88.53	94.95	86.4	[illegible]	90.3	94.4	92.0
T5-11B	**90.06**	**95.64**	**88.9**	[illegible]	**93.0**	**96.4**	**94.8**

Model	MultiRC F1a	MultiRC EM	ReCoRD F1	ReCo[illegible] Accuracy	RTE Accuracy	WiC Accuracy	WSC Accuracy
Previous best	84.4[e]	52.5[e]	90.6[e]	90.0[e]	88.2[e]	69.9[e]	89.0[e]
T5-Small	69.3	26.3	56.3	55.4	73.3	66.9	70.5
T5-Base	79.7	43.1	75.0	74.2	81.5	68.3	80.8
T5-Large	83.3	50.7	86.8	85.9	87.8	69.3	86.3
T5-3B	86.8	58.3	91.2	90.4	90.7	72.1	90.4
T5-11B	**88.2**	**62.3**	**93.3**	**92.5**	**92.5**	**76.1**	**93.8**

Model	WMT EnDe BLEU	WMT EnFr BLEU	WMT EnRo BLEU	CNN/DM ROUGE-1	CNN/DM ROUGE-2	CNN/DM ROUGE-L
Previous best	**33.8**[f]	**43.8**[f]	**38.5**[g]	43.47[h]	20.30[h]	40.63[h]
T5-Small	26.7	36.0	26.8	41.12	19.56	38.35
T5-Base	30.9	41.2	28.0	42.05	20.34	39.40
T5-Large	32.0	41.5	28.1	42.50	20.68	39.75
T5-3B	31.8	42.6	28.2	42.72	21.02	39.94
T5-11B	32.1	43.4	28.1	**43.52**	**21.55**	**40.69**

Table 14: Performance of our T5 variants on every task we study. Small, Base, Large, 3B, and 11B refer to model configurations with 60 million, 220 million, 770 million, 3 billion, and 11 billion parameters, respectively. In the first row of each table, we report the state-of-the-art for the task, with the superscript denoting its source with references listed at the end of this caption. All results are reported on the test set except for SQuAD where we use the validation set. [a][Lan et al., 2019] [b][Wang et al., 2019c] [c][Zhu et al., 2019] [d][Yang et al., 2019] [e][Liu et al., 2019c] [f][Edunov et al., 2018] [g][Lample and Conneau, 2019] [h][Dong et al., 2019]

QA is now only a sub-benchmark?

At least, SQuAD & Machine Reading are

Standard post-SQuaD Datasets

QUESTION

What temperature is a fever?

CONTEXT

Although a fever (pyrexia) could be considered any body temperature above the normal temperature of 98.6 degrees Fahrenheit (98.6 F or 37 C), medically, a person is not considered to have a significant fever until the temperature is above 100.4 F (38.0 C).

Standard post-SQuaD Datasets

QUESTION

What temperature is a fever?

CONTEXT

Although a fever (pyrexia) could be considered any body temperature above the normal temperature of 98.6 degrees Fahrenheit (98.6 F or 37 C), medically, a person is not considered to have a significant fever until the temperature is above 100.4 F (38.0 C).

ANSWER

100.4

How about answering --

What causes a fever?

Why are fevers dangerous?

Why do you need to bring your temperature down?

It's time to get to the core of what QA is!

Open domain

Developed and argumentative

Machine Reading is one of the keys

Long-form QA

Long Form (open-domain) Question Answering

QUESTION

Why do you need to bring your temperature down?

ANSWER

Up to a point, having a fever is a good thing when you're fighting an infection as in the case of sepsis (infection in the blood). Many pathogens don't fare well in even a degree or two of average raised temperature, while your body is much more resilient. It's still a pretty serious condition on its own, and sepsis is frequently fatal regardless of the not only the body's attempts to fight it, but with medical intervention.

The problems in general however, start when the fever is too high, or just high for too long. Your body will release something called chaperone molecules that help your proteins fold correctly, but there will still be errors and it's more energetically expensive. This chaperone molecules also have limits, and past a certain point your body fails on a number of levels.

For one, a lot of what your cells do is interact with, transport, [....]

Why is moving to long-form QA difficult?

- Generation of fluent outputs + Their Evaluation
- Retrieval of relevant sources
- Information synthesis
- Lack of Data!

ELI5

LONG-FORM QUESTION ANSWERING

Dataset and Pretrained Models:

http://www.github.com/facebookresearch/ELI5

https://facebookresearch.github.io/ELI5/

Explain Like I'm Five Dataset (Fan et al. ACL'19)

r/explainlikeimfive

90

ELI5: why do you need to bring your temperature down?

I read somewhere that the reason our temp rises when we are sick is to help out immune system. if so, why try to lower it

28 Comments

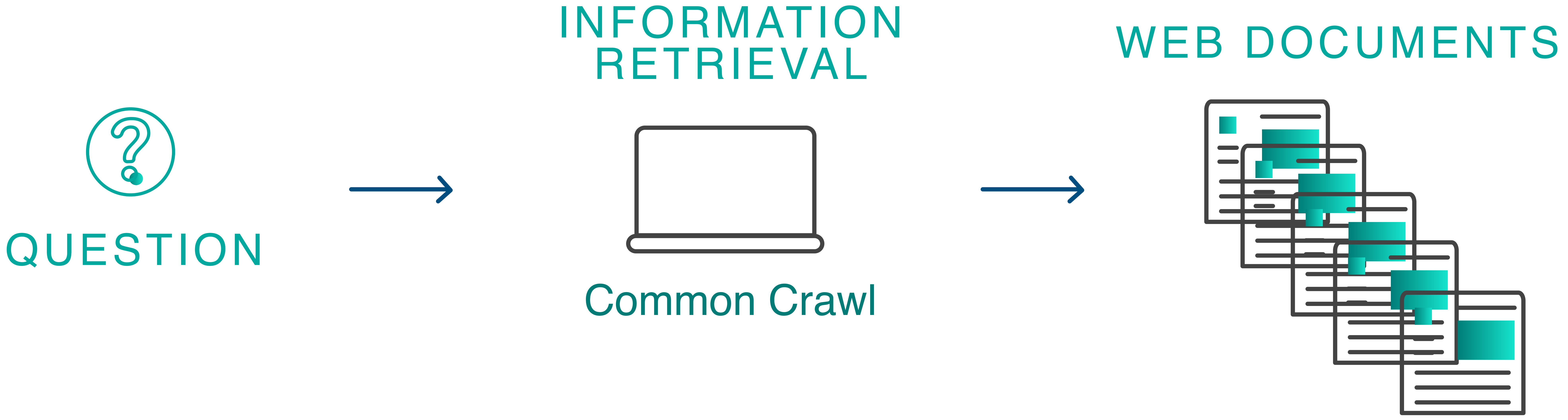
SORT BY BEST

Up to a point, having a fever is a good thing when you're fighting an infection as in the case of sepsis (infection in the blood). Many pathogens don't fare well in even a degree or two of average raised temperature, while your body is much more resilient. It's still a pretty serious condition on its own, and sepsis is frequently fatal regardless of the not only the body's attempts to fight it, but with medical intervention.

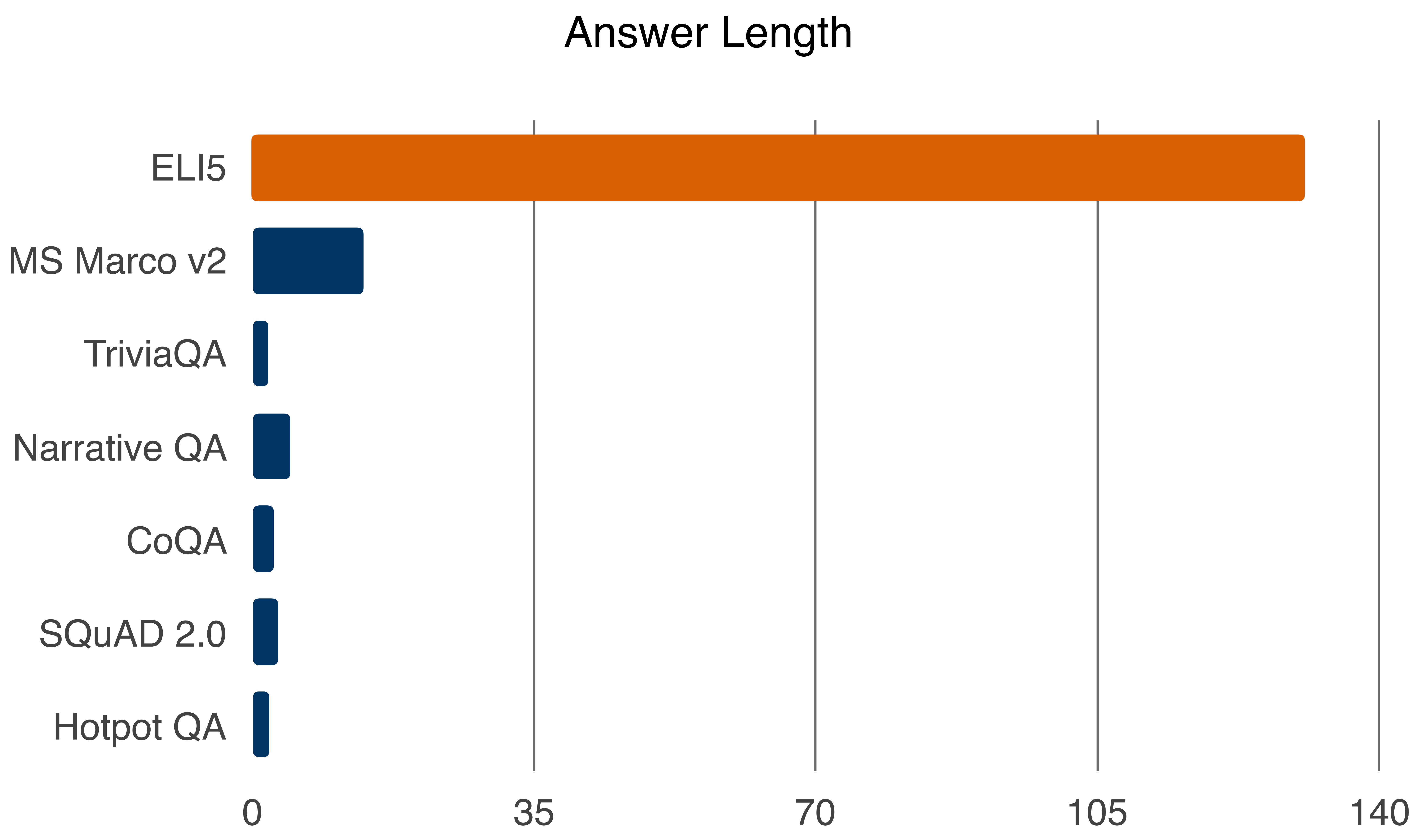
The problems in general however, start when the fever is too high, or just high for too long. Your body will release something called chaperone molecules that help your proteins fold correctly, but there will still be errors and it's more energetically expensive. This chaperone molecules also have limits, and past a certain point your body fails on a number of levels.

For one, a lot of what your cells do is interact with, transport, and produce proteins. The function of a protein is determined by its three dimensional structure, and it gets that through a process of folding. This is a process which can go wrong, and heat makes it far more likely to go wrong. Past a certain point critical proteins will start to unfold (denature) as in exposure to cooking methods. Needless to say, this does you no favors.

Open Domain


QUESTION
INFORMATION RETRIEVAL
Common Crawl
WEB DOCUMENTS

Longer Questions



Questions about Diverse Topics

Chemistry

Mathematics

Technology

Psychology

Biology

Physics

Engineering

Economics

Literature

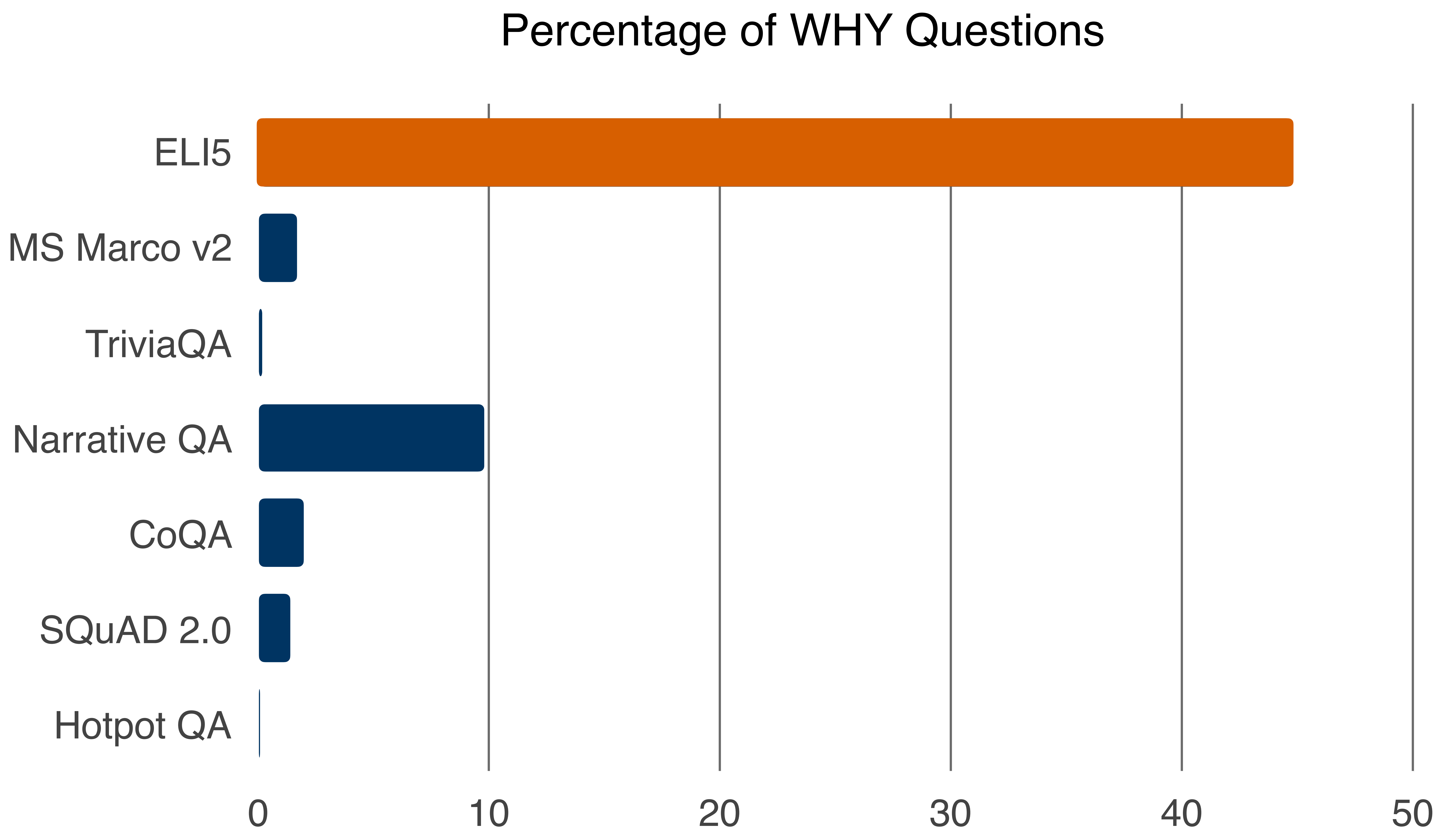
Culture

Religion

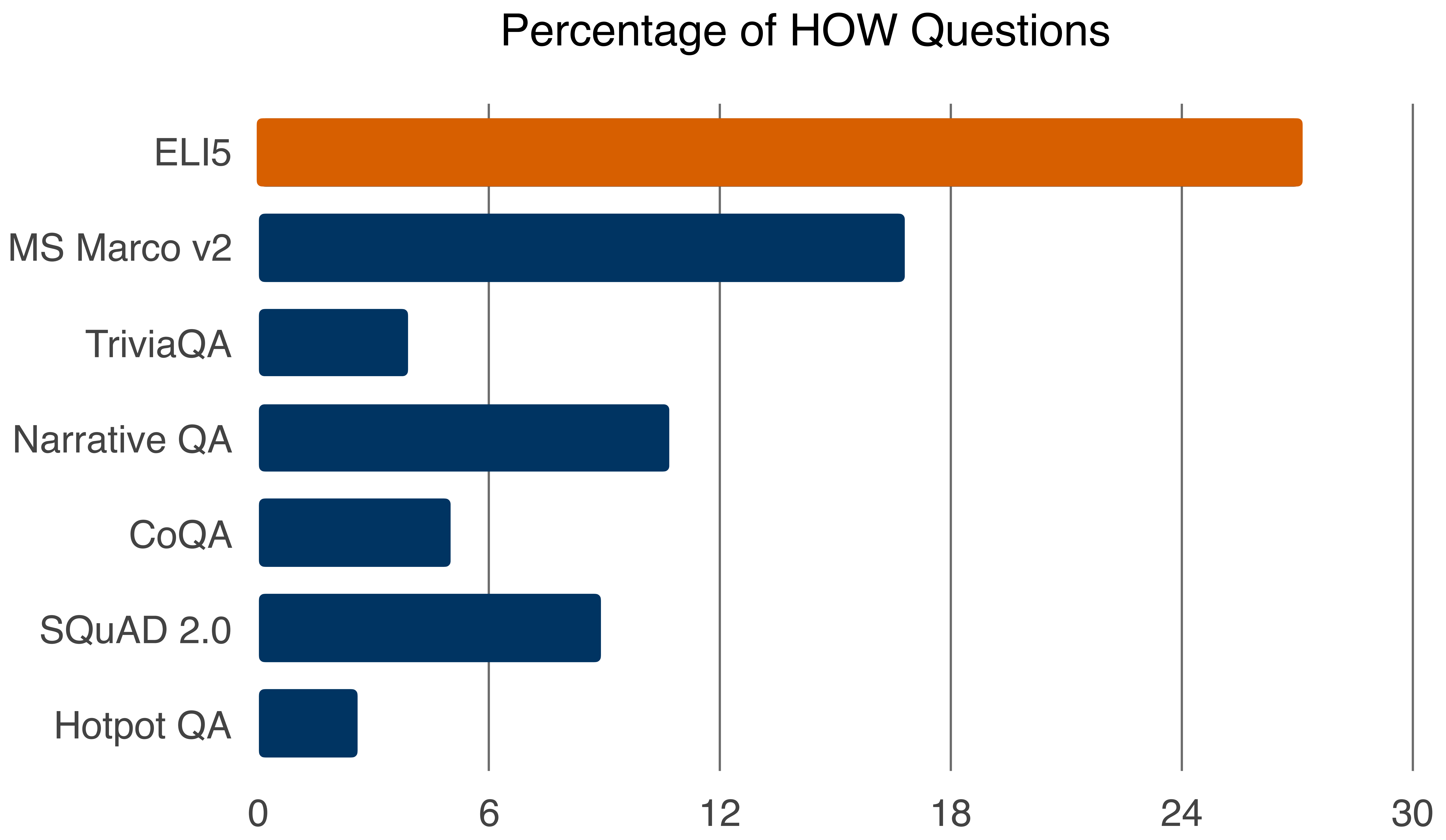
Government

History

More Why? Questions



More How? Questions



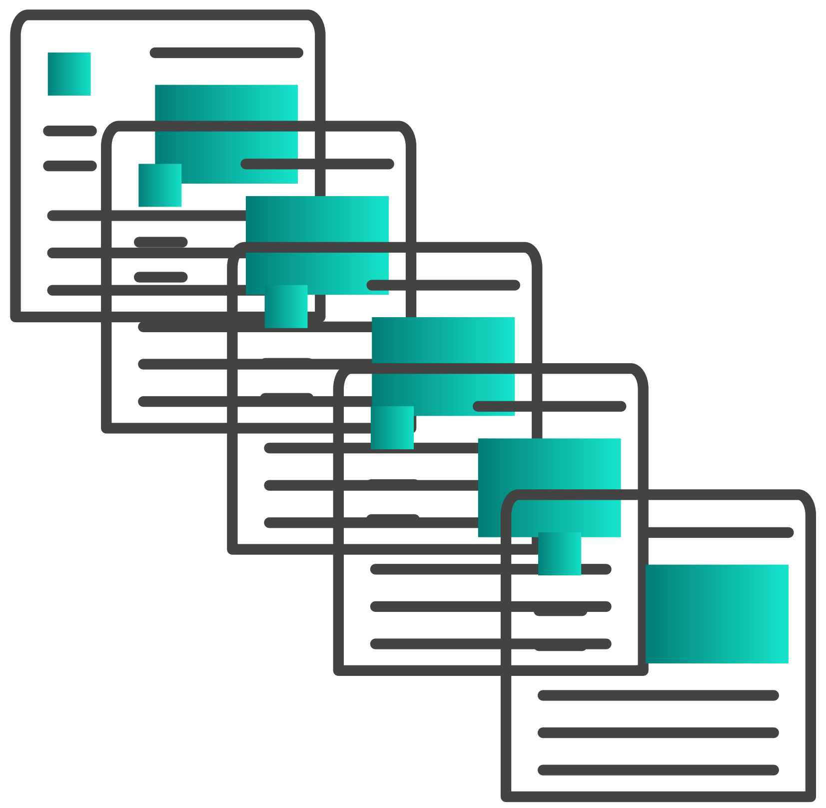
Explain Like I'm Five Dataset

270,000 EXAMPLES

QUESTION

42 words

WEB DOCUMENTS



100 per question

200,000 words

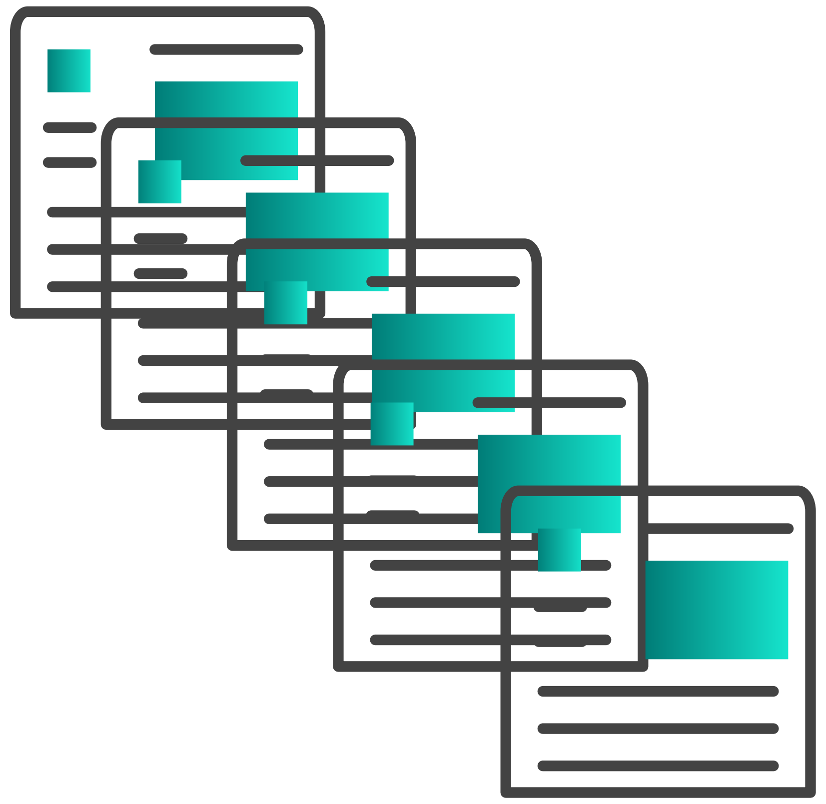
ANSWER

130 words

Models

Modeling Challenges

WEB DOCUMENTS



- Find Relevant Information in a Multi-Document Setting

Modeling Challenges

QUESTION

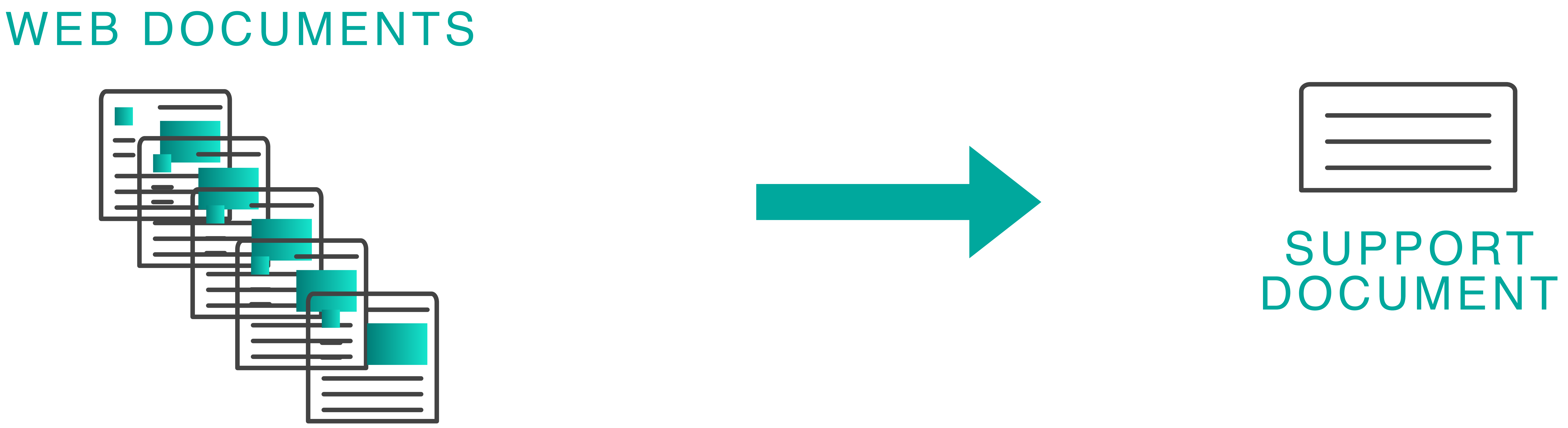
- Find Relevant Information in a Multi-Document Setting
- Query-driven Information Reasoning

Modeling Challenges

ANSWER

- Find Relevant Information in a Multi-Document Setting
- Query-driven Information Reasoning
- Writing a Long Text Answer

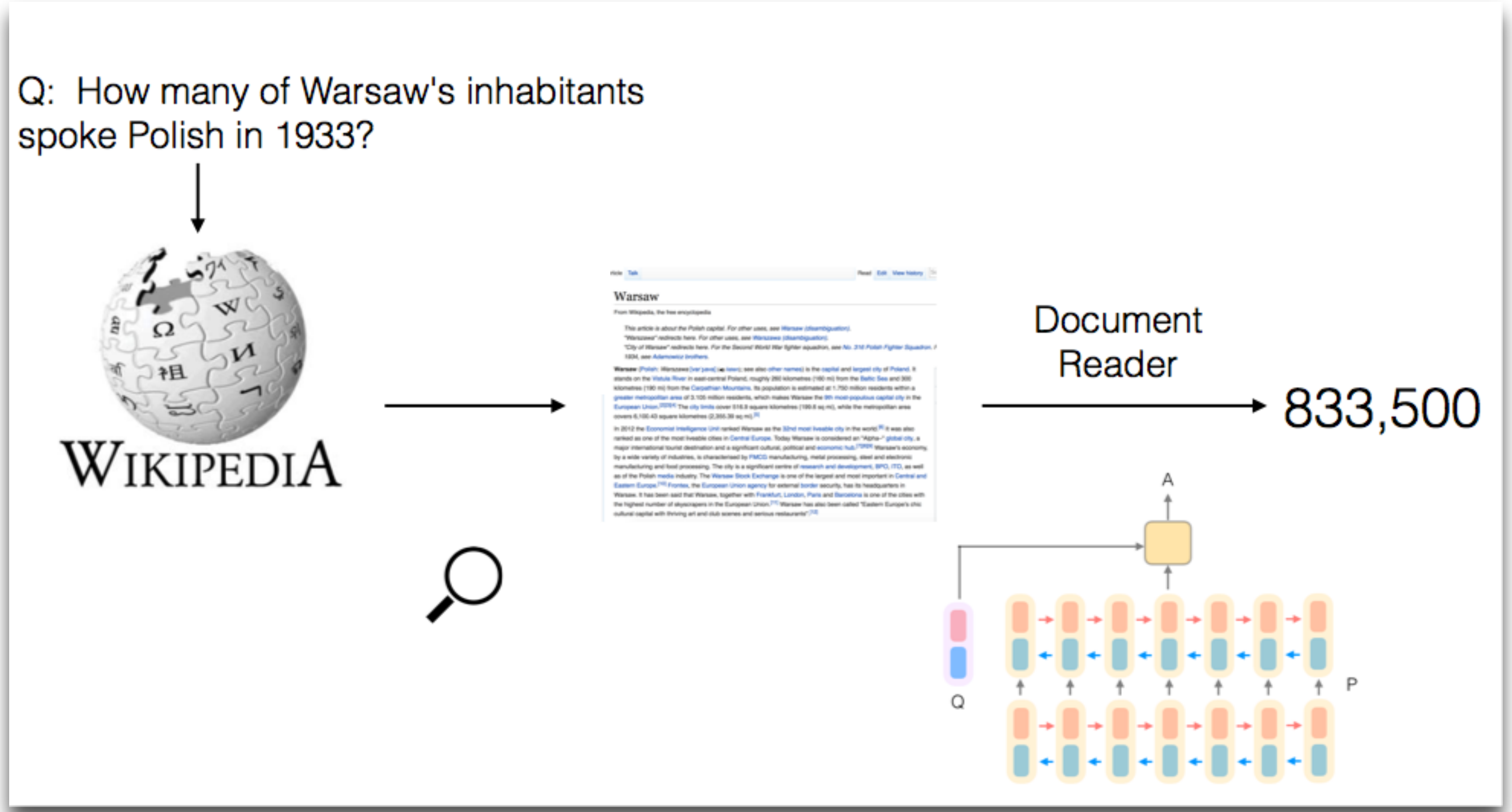
Dealing with Long Web Input



Over 200,000 words long

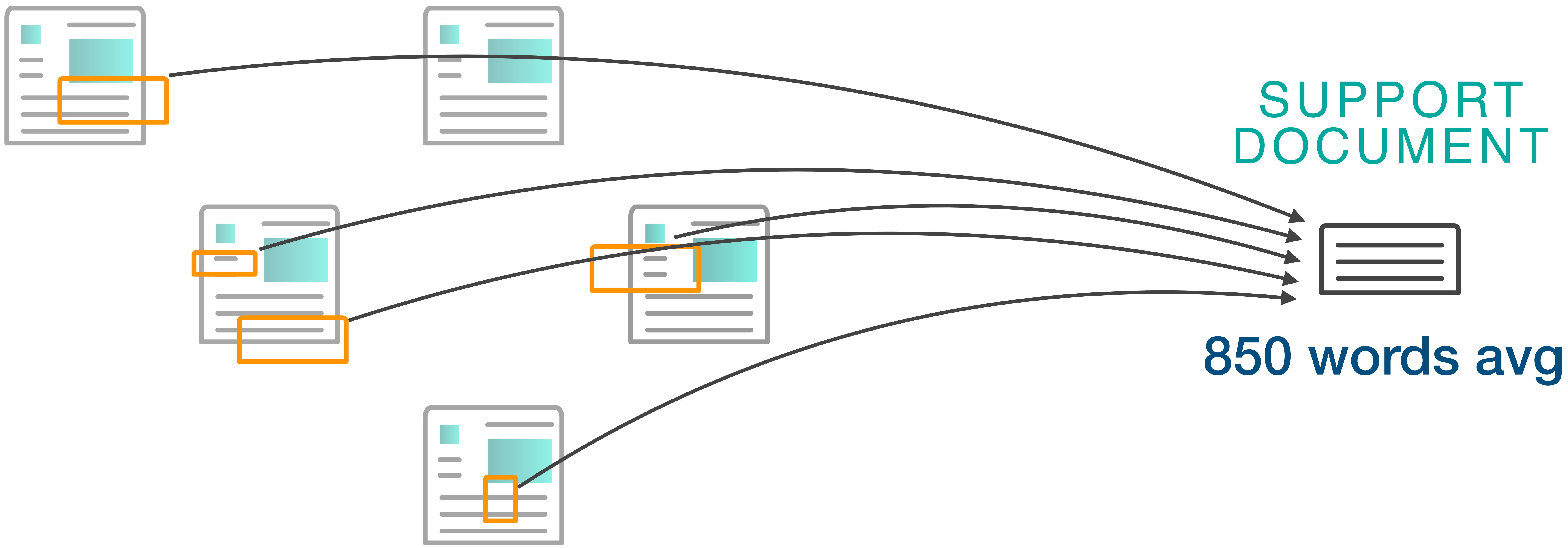
Around 1,000 words long

Using TF-IDF



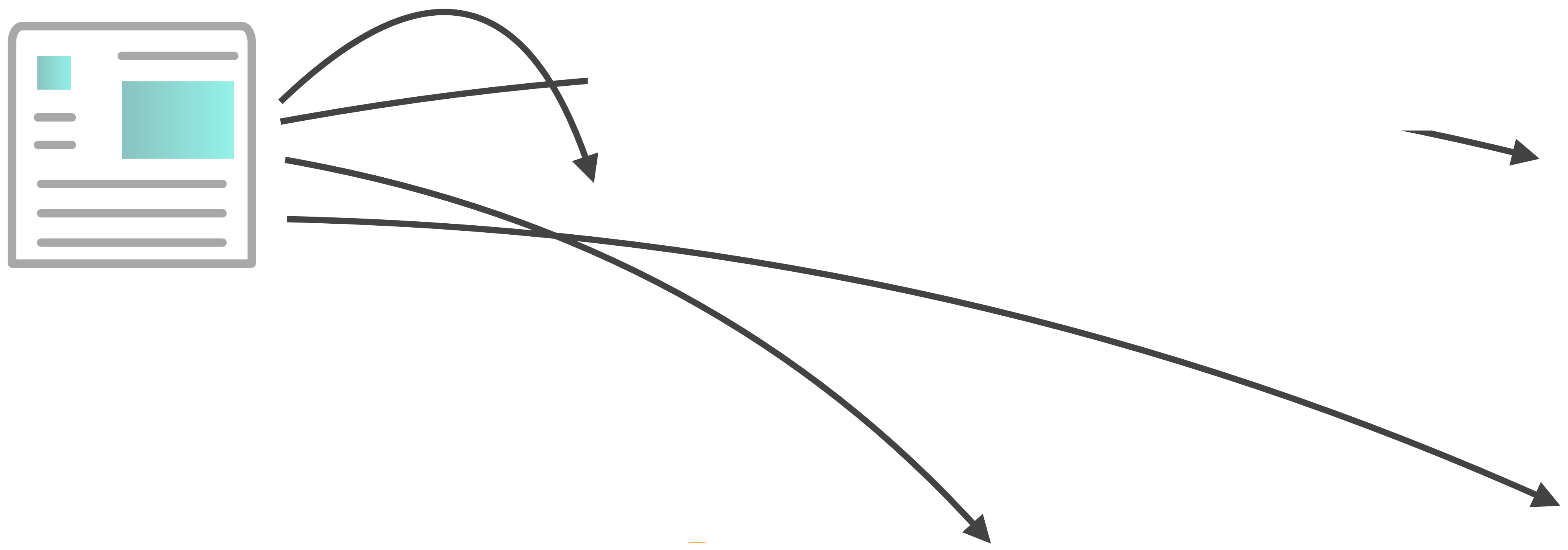
DrQA (Chen et al., ACL'17)

Creating a Shorter Support Document with TF-IDF

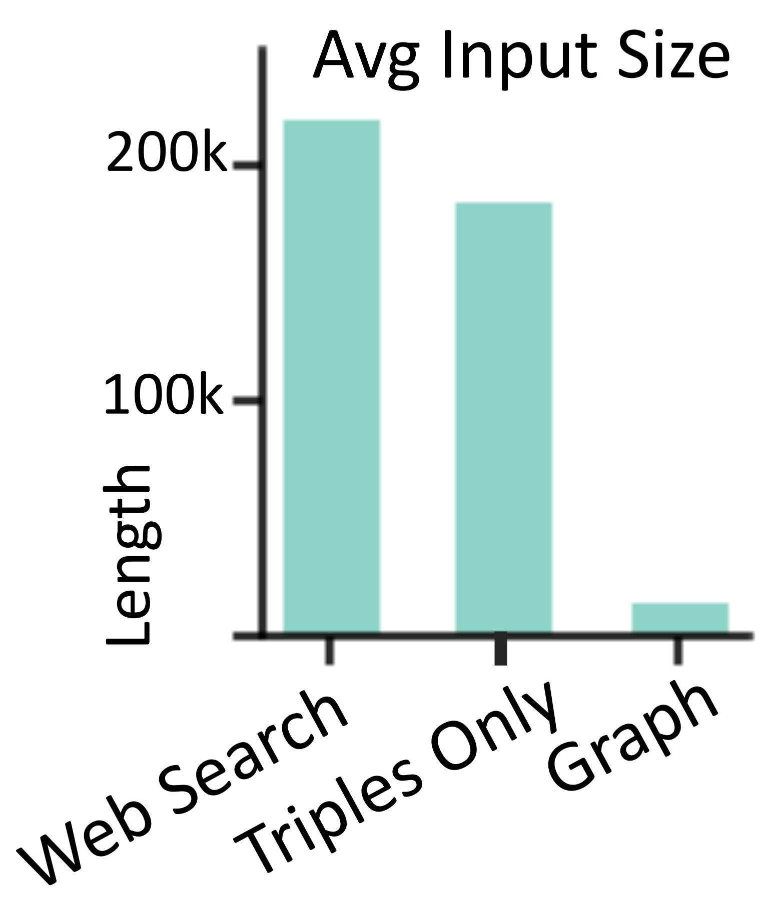


Creating a Shorter Support Document with Local KB (Fan et et al.

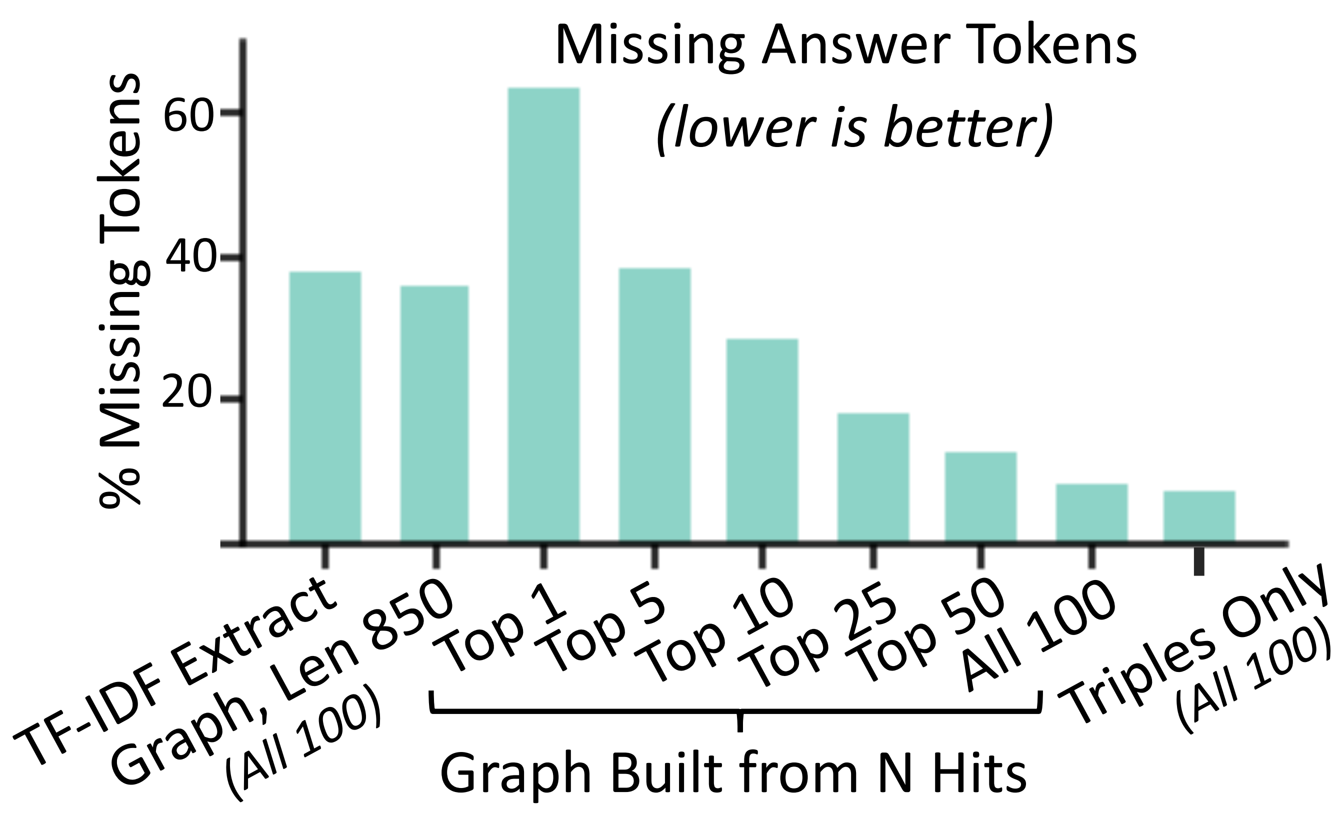
QUERY: Can someone finally explain the theory of general relativity?



Creating a Shorter Support Document with Local KB (Fan et et al.



- Large **compression** effect, reducing input size by an order of magnitude.
- Linearized output of the KB can be processed with (almost) standard Seq2Seq



- Despite compression, the graph **contains most of the relevant tokens**

Modeling Approaches

- Extractive
- Abstractive

Extractive Models

Find sentences of the support document to copy as the

Abstractive Models

Generate each word of the

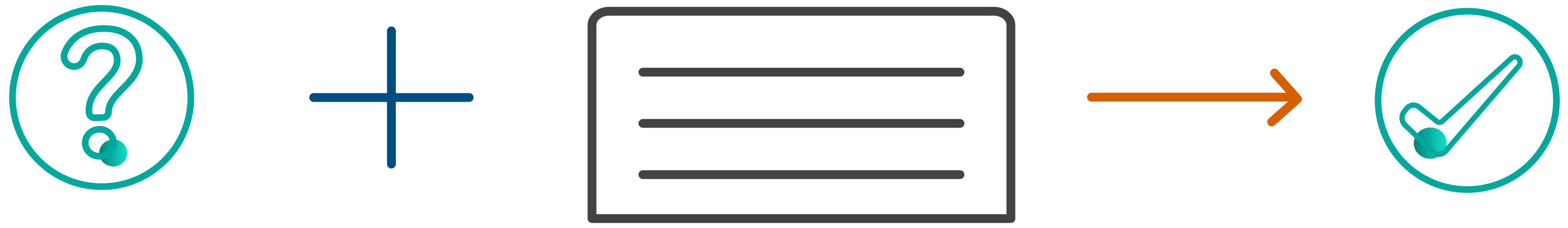
Abstractive Models

SEQUENCE TO SEQUENCE MULTITASK

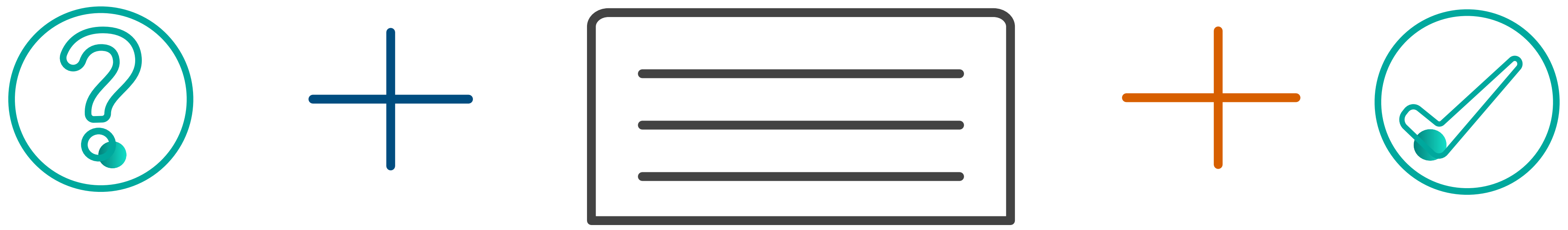
training time: train on many tasks
to add the benefit of language

training time: train on many tasks

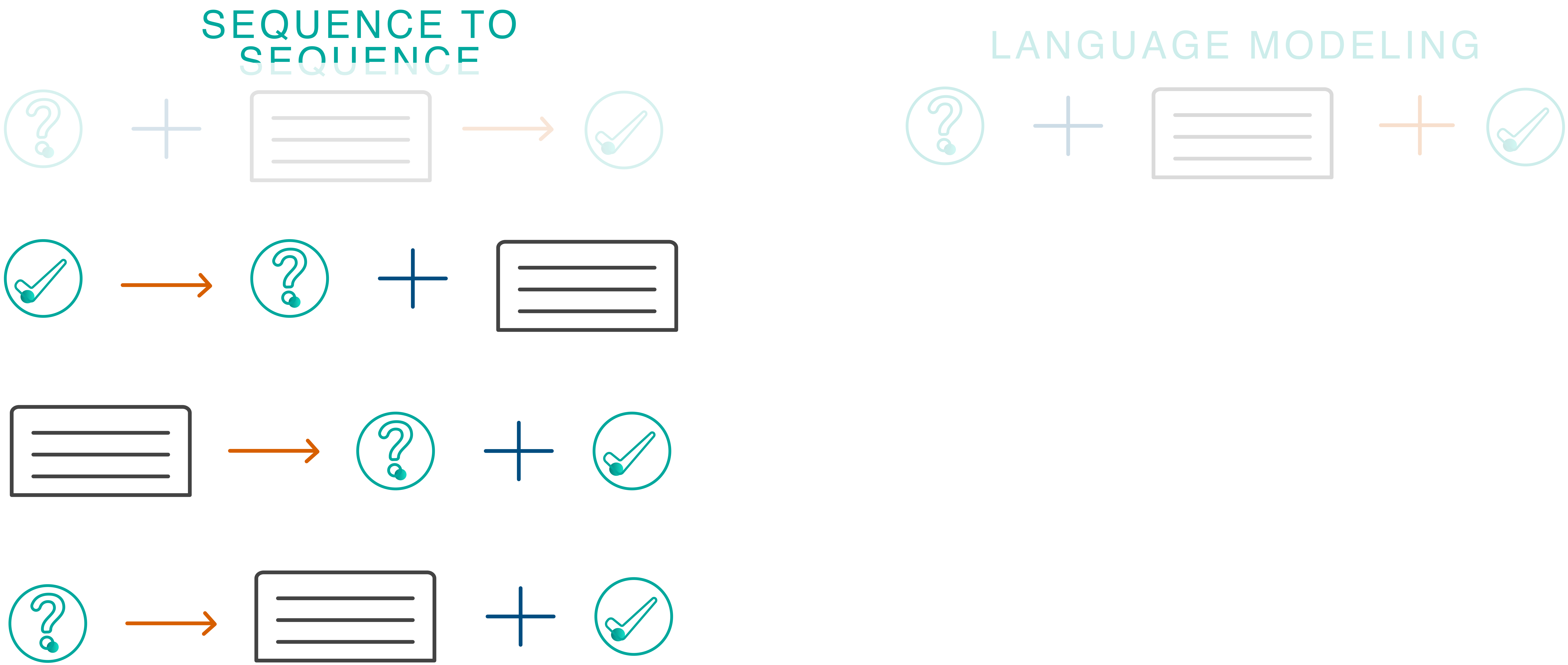
SEQUENCE TO SEQUENCE



LANGUAGE MODELING



training time: train on many tasks



training time: train on many tasks

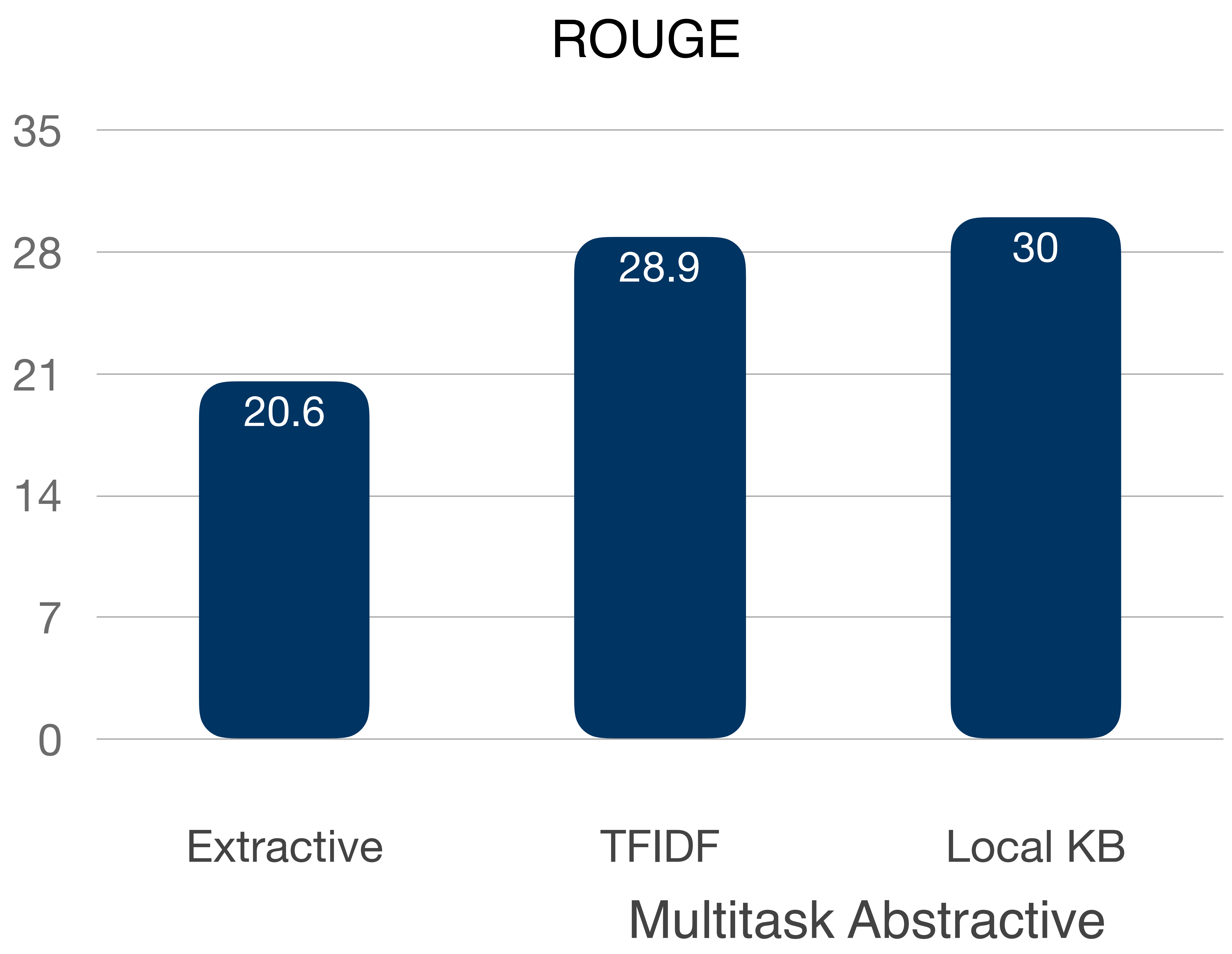
Abstractive Models

SEQUENCE TO SEQUENCE MULTITASK

inference time: standard question answering

Evaluatio
n

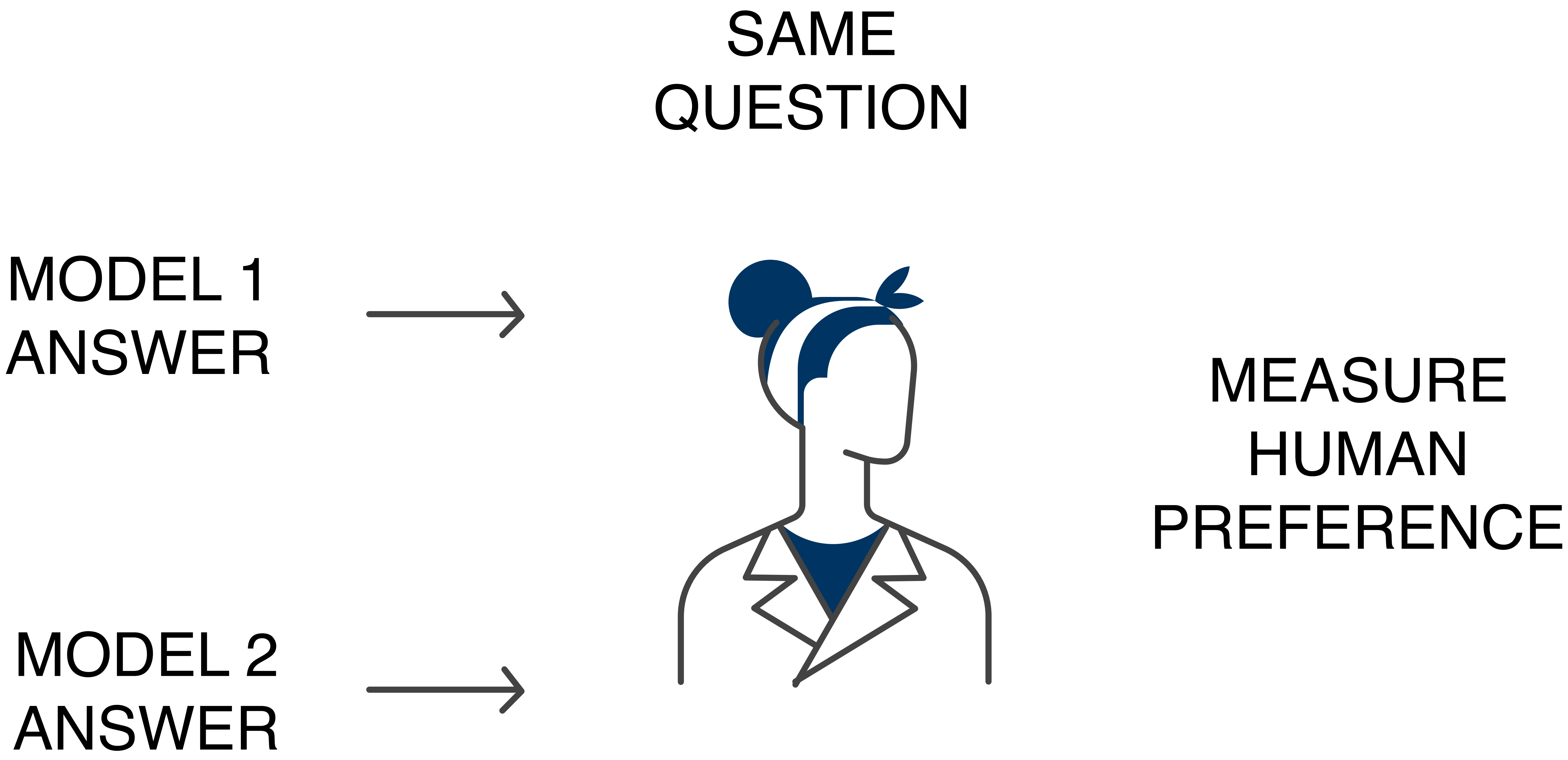
Automatic Evaluation



Interpretation

Question: *Can someone finally explain the theory of general relativity?*

Human Evaluation



Human Evaluation

	LM	s2s	Extractive	Multi-task
Human	98.4*	93.1*	85.5*	88.8*

far from human performance

Still a long way to go but we are putting the pieces together!

Prediction: Long-form, open-domain and extractive QA will merge

Knowledge beyond QA

Wizard of Wikipedia (Dinan et al. ICLR'19)

1307 Diverse General Topics: *crowd-sourced*

Gouda cheese
commuting
music festivals
podcasts
bowling
Arnold Schwarzenegger
Alpine skiing
Bodybuilding supplement
Harley-Davidson
Miley Cyrus
Hamilton (musical)
Ireland
Cannabis (drug)
Mount Kilimanjaro
Eggplant
Welding
Aquarium
Italian cuisine

Mercedes-Benz S-Class
Peanut
German language
Chicago-style pizza
Black hair
Toga party
100 metres
Tiger
Smoking
Winter
Acrophobia
List of art media
Pet adoption
Influencer marketing
Vitamin C
Human height
Steak

Kurt Cobain
List of water sports
Strawberry
Online game
Text messaging
Baileys Irish Cream
Fiction
American football
Online shopping
Aldi
Rock and roll
Kendrick Lamar
Medical billing
Blue Ridge Parkway
Clown
Pasta
Hiking

Each Linked to Wikipedia

Table 1: Dataset statistics of the Wizard of Wikipedia task.

Wizard of Wikipedia Task	Train	Valid	Test Seen	Test Unseen
Number of Utterances	166,787	17,715	8,715	8,782
Number of Dialogues	18,430	1,948	965	968
Number of Topics	1,247	599	533	58
Average Turns per Dialogue	9.0	9.1	9.0	9.1
Knowledge Database	5.4M articles		93M sentences	

Topic:	Ice cream
Wizard:	I just love ice cream. I love the types with fruits and flavours. Do you like ice cream?
Apprentice:	I love Ice cream as much as any one. I especially like Gelato, foreign ice cream!
Knowledge	Ice cream is a sweetened frozen food typically eaten as a snack or dessert... It is usually made from dairy products, such as milk and cream, and ... ... Bacon ice cream (or bacon-and-egg ice cream) is an ice cream generally created by adding bacon to egg custard and freezing the mixture.
Wizard:	Me too. There are some strange combinations though, have you heard of bacon ice cream? where they add bacon and even egg custard to the freezing mixture!
Apprentice:	Surprisingly bacon ice cream doesn't surprise me. That doesn't sound appealing to me, but perhaps it could be delicious...

Retrieval-based model (~extractive)

Model: Retrieval Transformer MemNet

Topic:	E-book (*seen*)
Human:	i do not know why, but I have never gotten into E-books
Model:	Same here. I have a kindle, but I don't use it for books. It's nicer to hold a real book and turn pages.
Human:	yes, i like the physical feel and smell of a real book
Model:	I actually had a Nook e-book where you had to download the e-book after selecting it online. I grew out of it, I prefer paper books also.
Human:	i tried e-books but I do not like reading on a screen, its disorientated for me
Model:	Yes, reading a book on a tablet just isn't the same!
Human:	plus I like owning the physical copy of a real book, I feel like I own it
Model:	I prefer print and newsreels.

Model: I prefer print and newsreels.

Topic:	Toga party (*seen*)
Model:	I love toga parties! I usually make my toga out of a bed sheet and wear my sandals.
Human:	Me too!!! I usually use some flip flops and my blanket!
Model:	Toga parties typically have games involving kegs and excessive drinking!
Human:	Yeah! And getting to wear funny outfits and a crown of leaves!
Model:	I love me some Toga Parties!
Human:	I love me some too!
Model:	Toga parties are so much fun with the Roman or Greek theme.
Human:	Yeah! I especially like Greek themes

Generative model

Model: Two-Stage Generative Transformer MemNet	
Topic:	Circus (*seen*)
Model:	have you ever been to a circus? it is a company of performers who put on diverse entertainment shows.
Human:	I went to the circus once when I was a child actually.
Model:	that's great! entertainment is a form of activity that holds the attention and interest of an audience.
Human:	That's true. Have you ever been to a circus?
Model:	I have not, but I know they were built with various types of seating ring, and sometimes a stage.
Human:	I went to one held in a stadium, so we had box seats.
Model:	that's awesome! i'm sure you'll have to go to the circus, with appropriate music.
Human:	Yes, circuses usually have their own kind of music.
Model:	circuses can be done by circuses.

Model: Two-Stage Generative Transformer MemNet	
Topic:	Shower (*unseen*)
Human:	So I've been looking for a better showerhead with more water pressure for my shower.
Model:	I love pressure-versed water at nearly constant temperature to a shower
Human:	That sounds pretty great! Love those rainshower showers!
Model:	the ancient greeks were the first people to have showers.
Human:	Well I knew they were good for something.
Model:	the original showers were neither indoor structures nor man made but were common.
Human:	Ah that sounds intimidating!
Model:	it is required manual transport of both fresh and waste water.