Retrieval-Augmented Generation of Long Form Text

Claire Gardent
Joint work with Angela Fan (Facebook), Antoine Bordes (Facebook) and Chloé Braud (CNRS/IRIT)

xNLG

Generating into Multiple Languages from Multiple Sources

Des responsables américains ont tenu une réunion d'un groupe d'experts en janvier 2002 à New York.

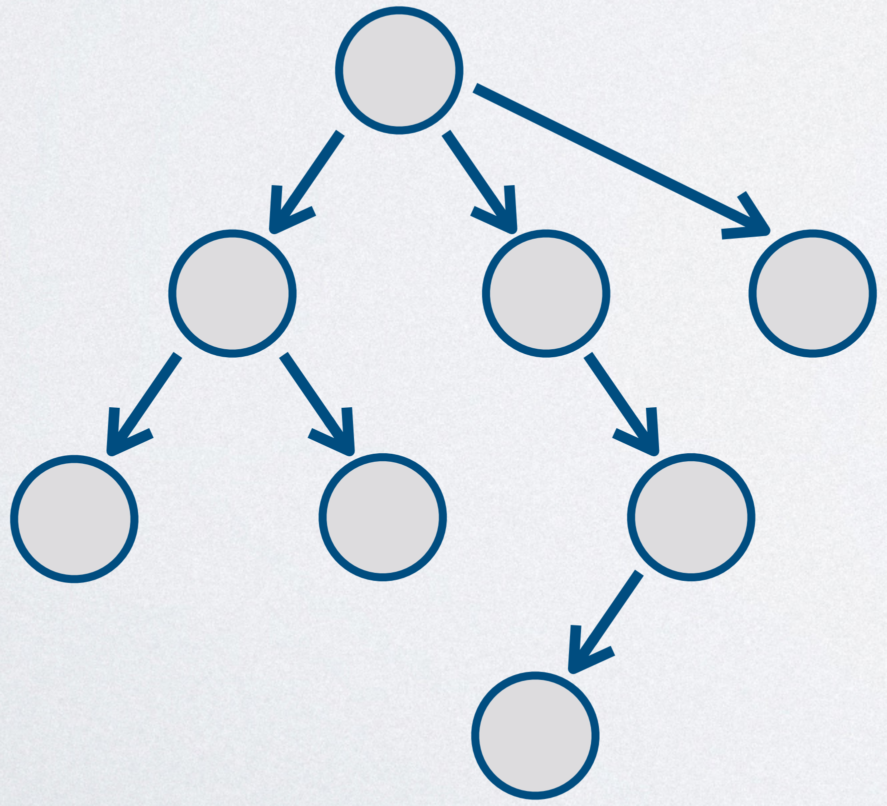
Funcionarios estadounidenses celebraron una reunión de un grupo de expertos en enero de 2002 en Nueva York.

Americkí predstavitelia usporiadali stretnutie expertnej skupiny v januári 2002 v New Yorku.

Американските служители проведоха среща на експертна група през януари 2002 г. в Ню Йорк.

Amerikanska tjänstemän höll ett expertgruppsmöte i januari 2002 i New York.

......



anr

Grand Est
ALSACE CHAMPAGNE-ARDENNE LORRAINE
L'Europe s'invente chez nous

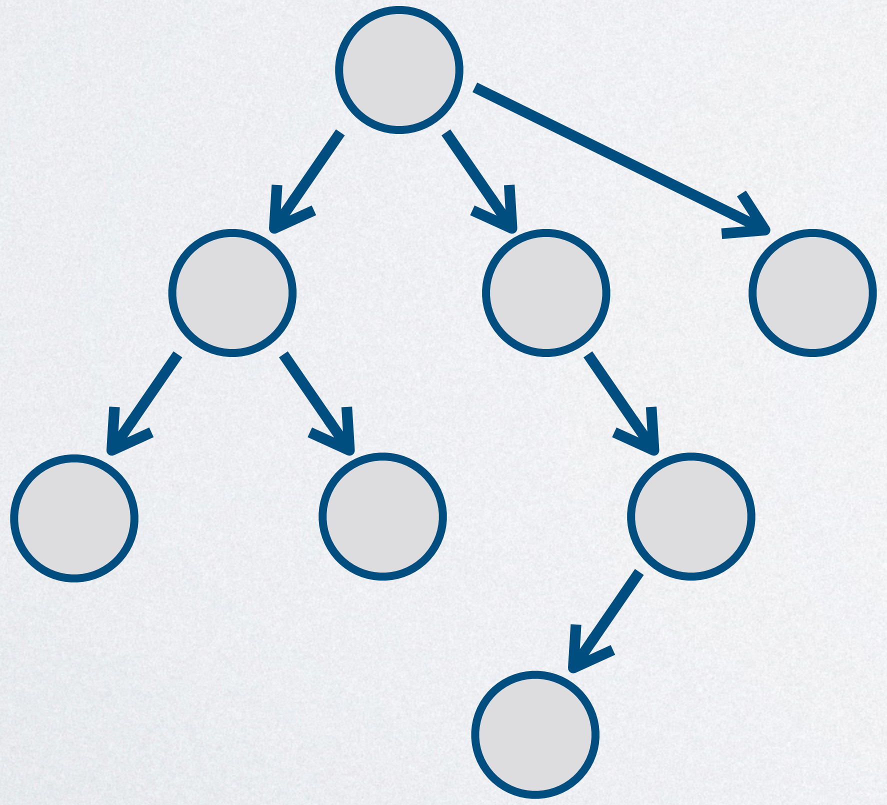
Meta

Applications

What is NLG useful for ?

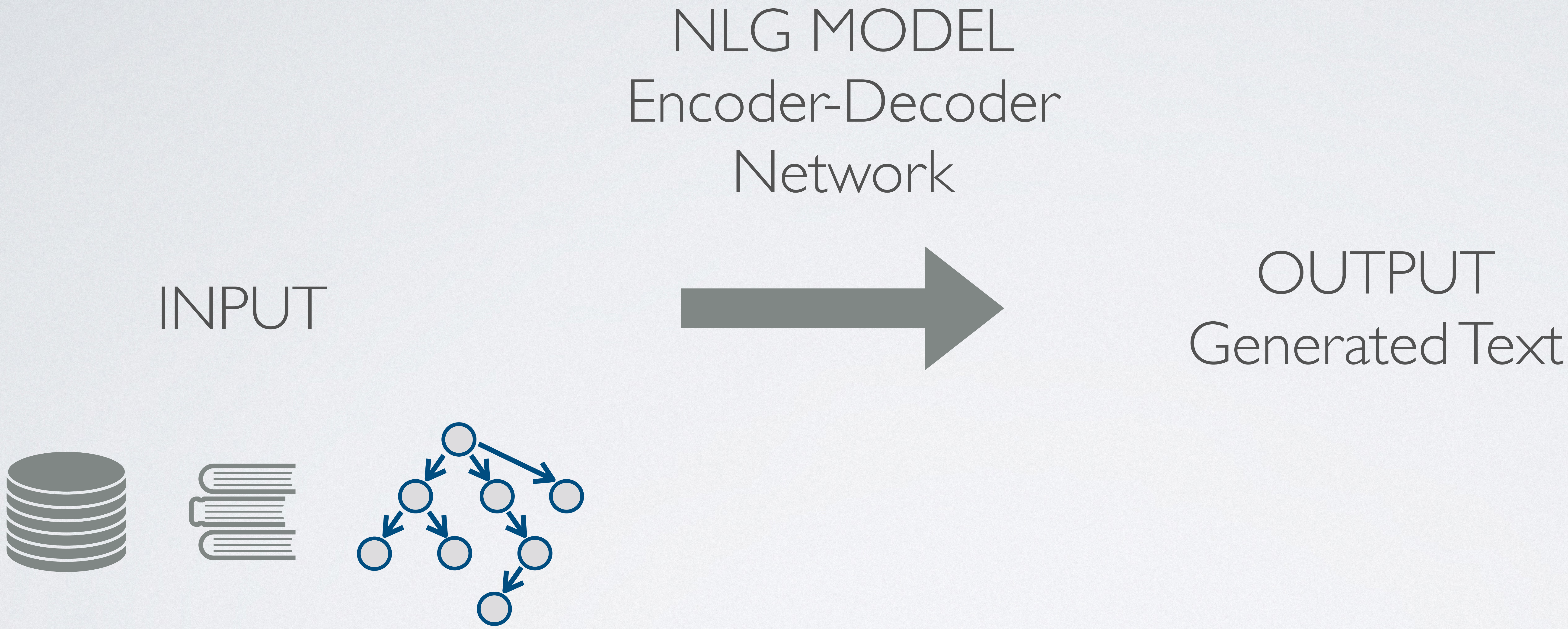
Verbalising, Querying Knowledge-Bases

Summarising, Simplifying, Paraphrases Text



Converting Graphs into Text

Neural NLG

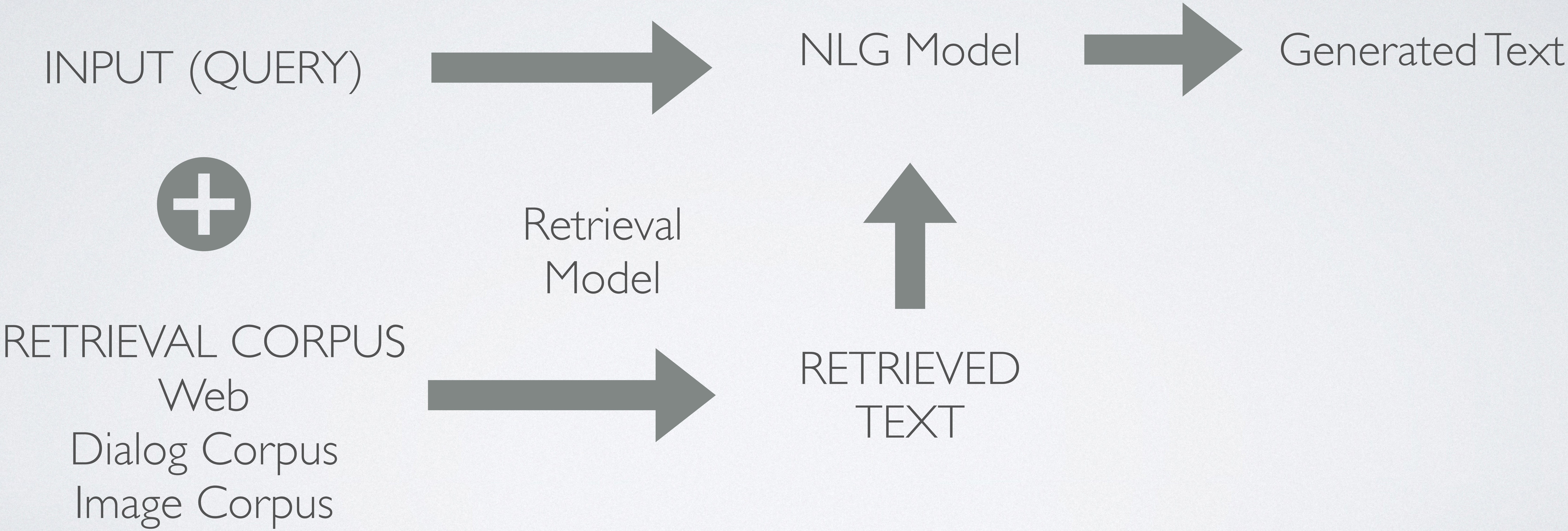


NLG MODEL
Encoder-Decoder
Network

INPUT

OUTPUT
Generated Text

Retrieval-Augmented Neural NLG



INPUT (QUERY)

NLG Model

Generated Text

Retrieval
Model

RETRIEVAL CORPUS
Web
Dialog Corpus
Image Corpus

RETRIEVED
TEXT

Challenges for Retrieval-Augmented NLG

Scaling to very large retrieval corpora

Challenges for Retrieval-Augmented NLG

Scaling to very large retrieval corpora

Retrieving relevant knowledge

Challenges for Retrieval-Augmented NLG

Scaling to very large retrieval corpora

Retrieving relevant knowledge

Encoding long form input

Challenges for Retrieval-Augmented NLG

Scaling to very large retrieval corpora

Retrieving relevant knowledge

Encoding long form input

Decoding (generating) long form text

Three NLG Tasks

Retrieval-Based Models for three NLG Tasks

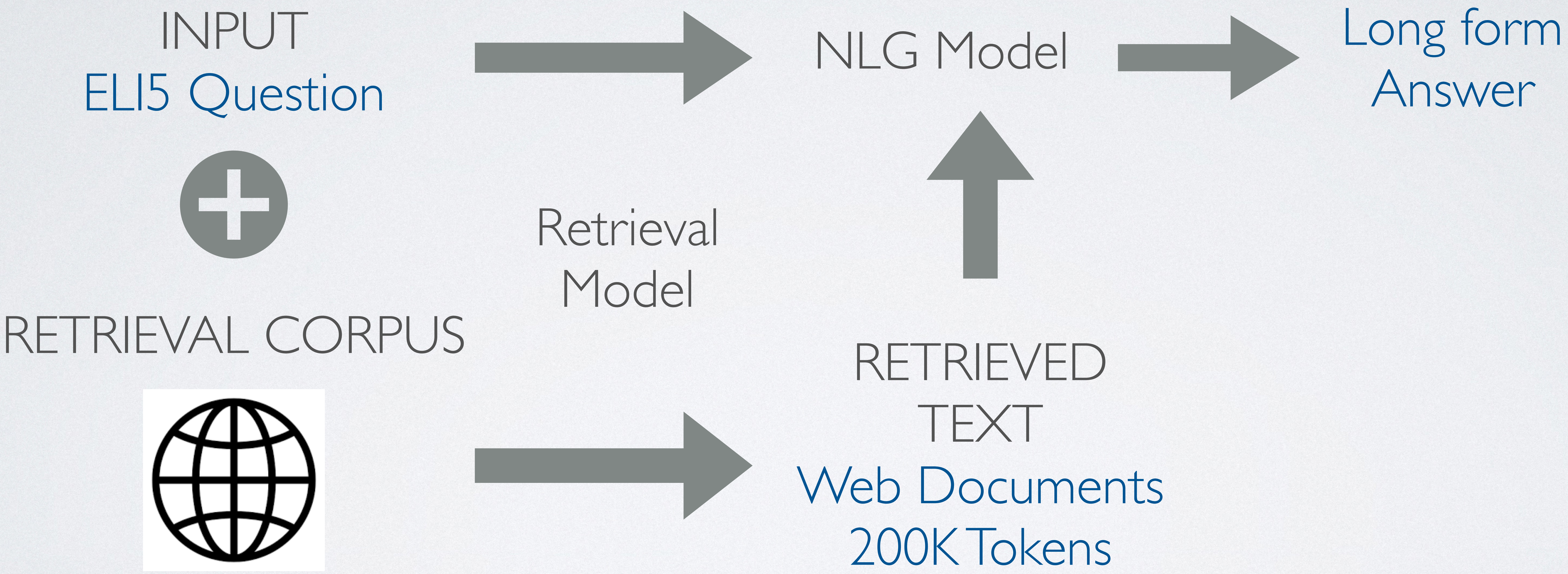
Long Form Question Answering

Human-Machine Dialog

Generating Wikipedia Biographies

Question Answering

Scaling to long input



INPUT
ELI5 Question

RETRIEVAL CORPUS

Retrieval
Model

NLG Model

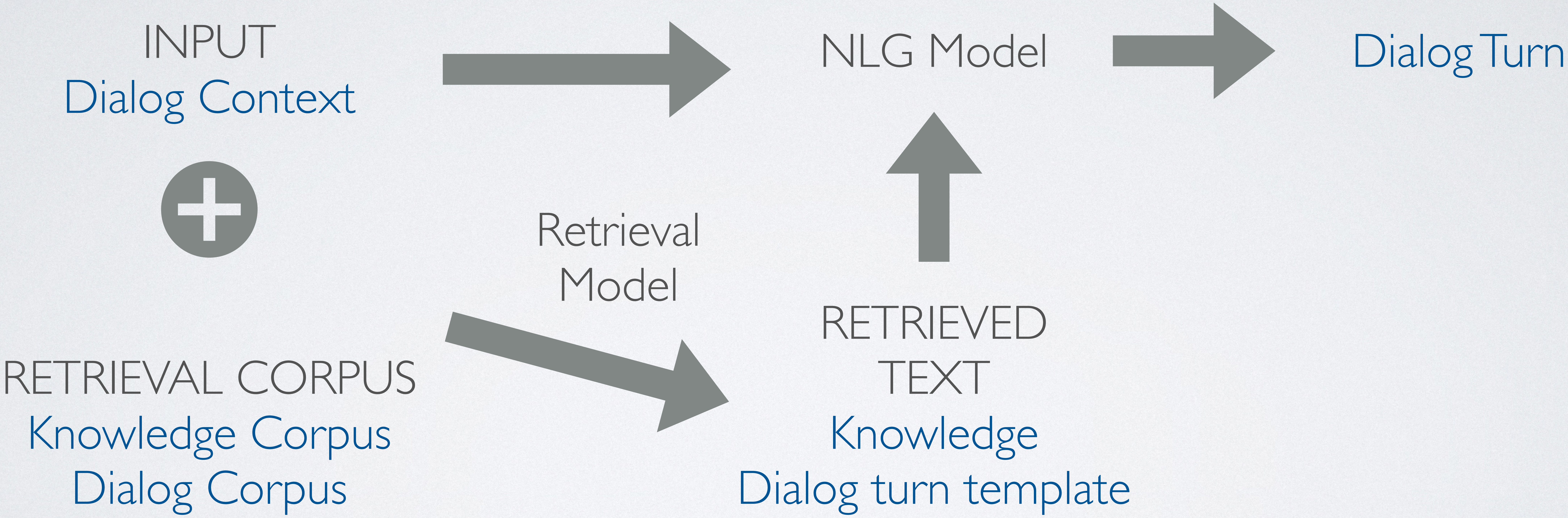
RETRIEVED
TEXT
Web Documents
200K Tokens

Long form
Answer

Human-Machine Dialog

Retrieving from multiple, multimodal retrieval sources

Scaling to very large retrieval corpora



INPUT
Dialog Context

RETRIEVAL CORPUS
Knowledge Corpus
Dialog Corpus

Retrieval Model

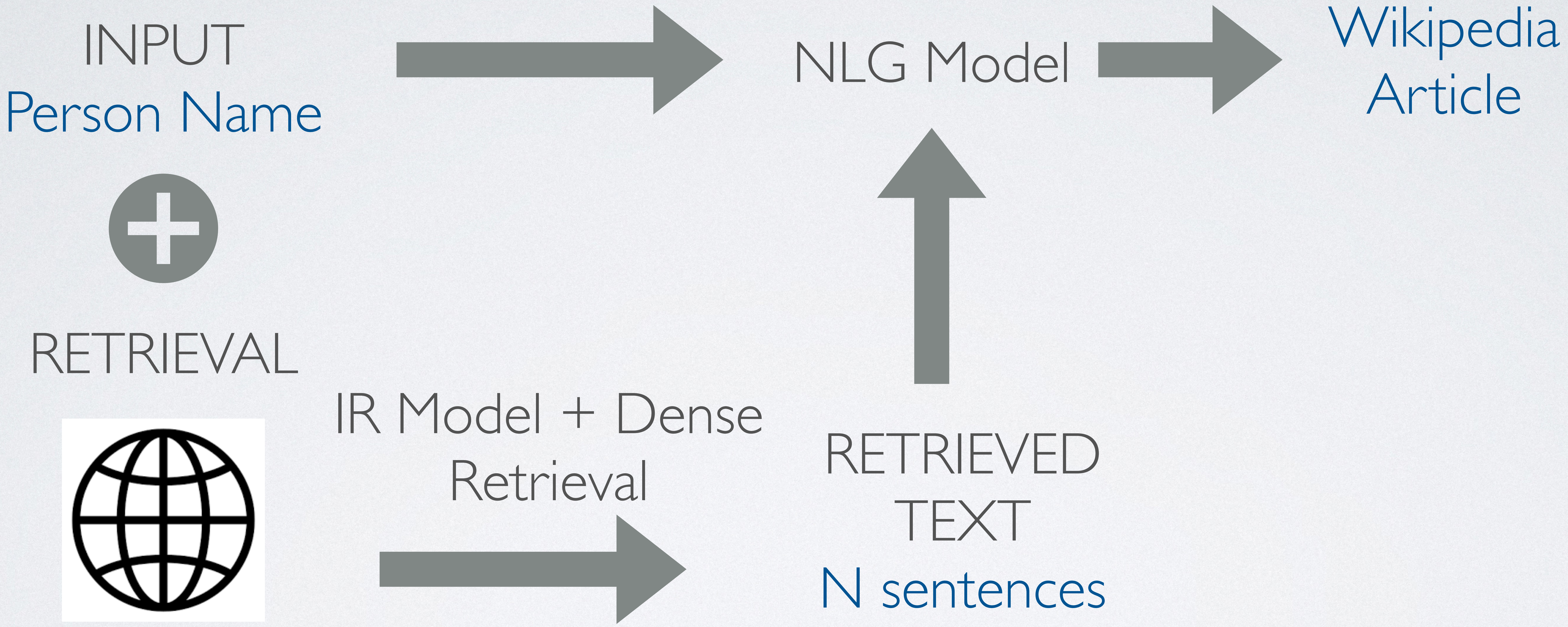
NLG Model

RETRIEVED TEXT
Knowledge
Dialog turn template

Dialog Turn

Generating Wikipedia Woman Biographies

Generating structured text, Impact of available evidence (Gender bias)



INPUT

Person Name

RETRIEVAL

IR Model + Dense Retrieval

RETRIEVED TEXT

N sentences

NLG Model

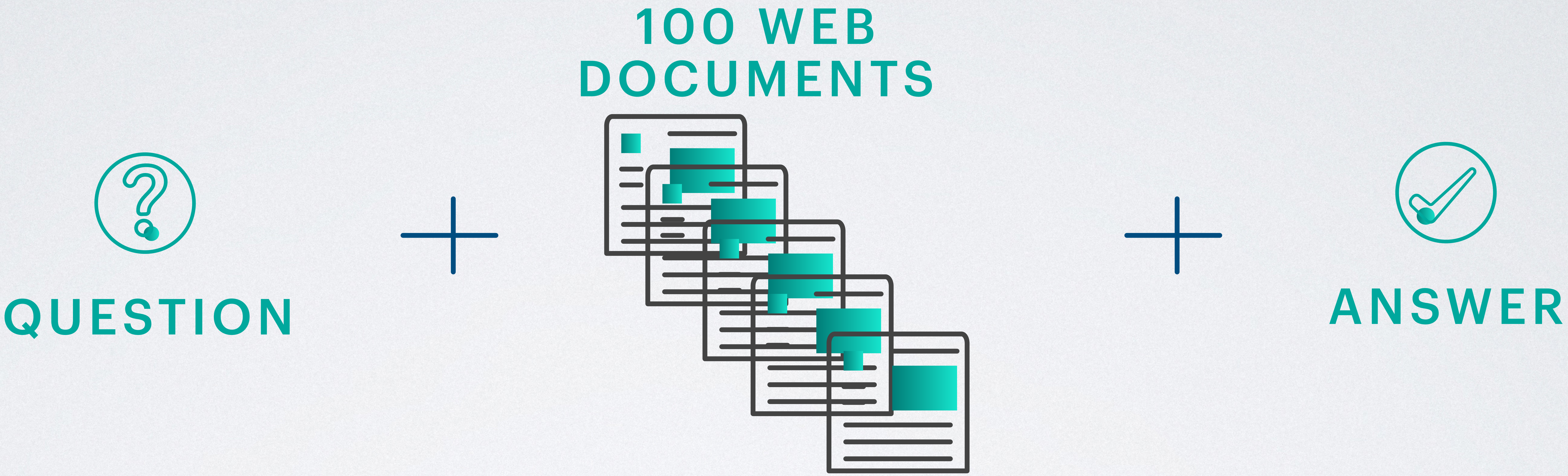
Wikipedia Article

Retrieval-Augmented Question Answering

Question Answering

Explain Like I'm Five Dataset

270,000 TRAINING INSTANCES



QUESTION

+

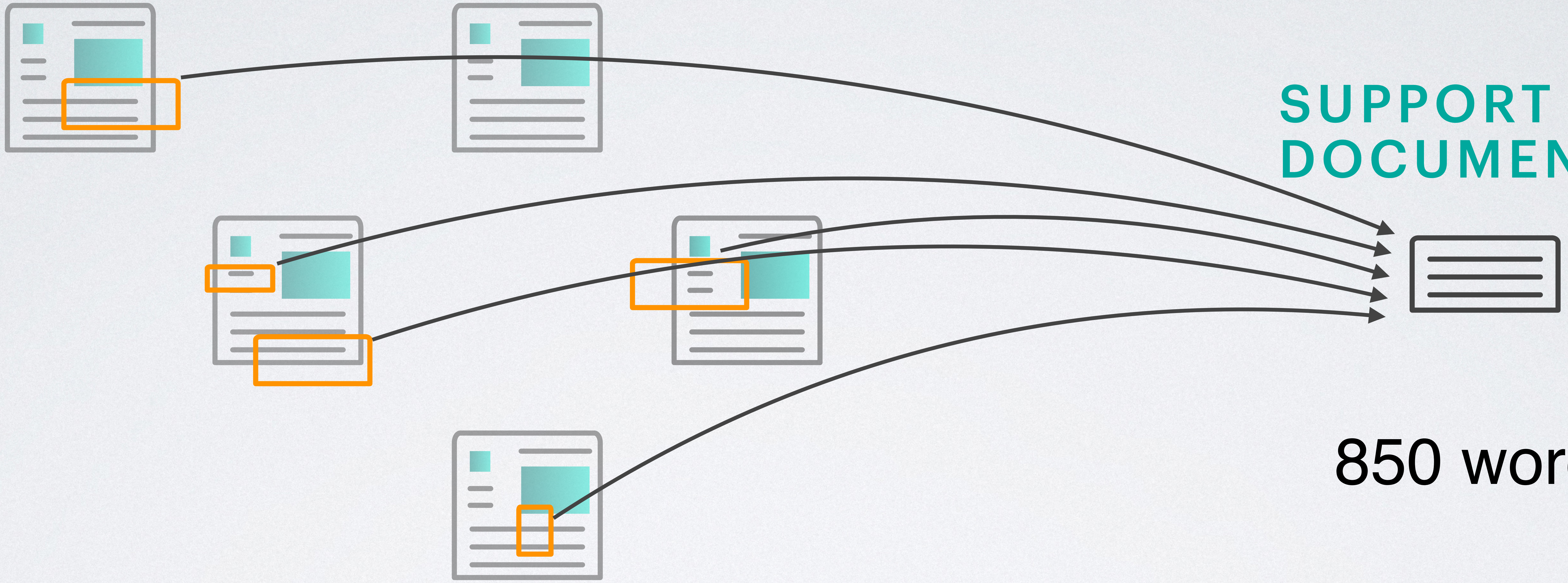
100 WEB DOCUMENTS

+

ANSWER

200,000 words

Creating a Shorter Support Document



SUPPORT DOCUMENT

850 words

200,000 words

TF-IDF Method

CALCULATE TF-IDF OVERLAP

QUESTION

WEB DOCUMENT SENTENCES

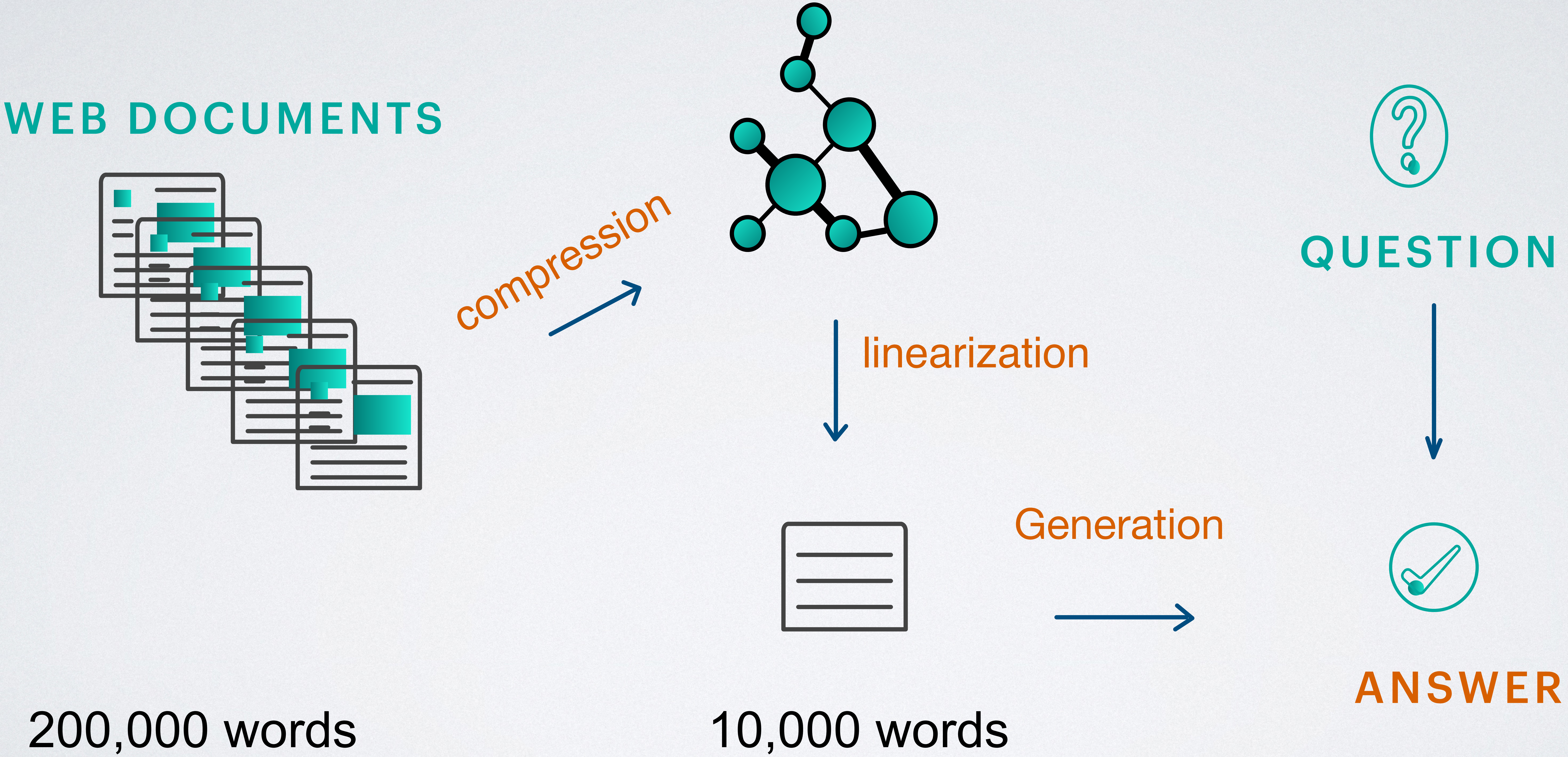
Downsides

38% of the Answer Tokens are Missing

Selected text fragments are often redundant (same tf-idf)

Convert Input Texts to Graph

Fan et al. EMNLP 2019



WEB DOCUMENTS

compression

linearization

QUESTION

Generation

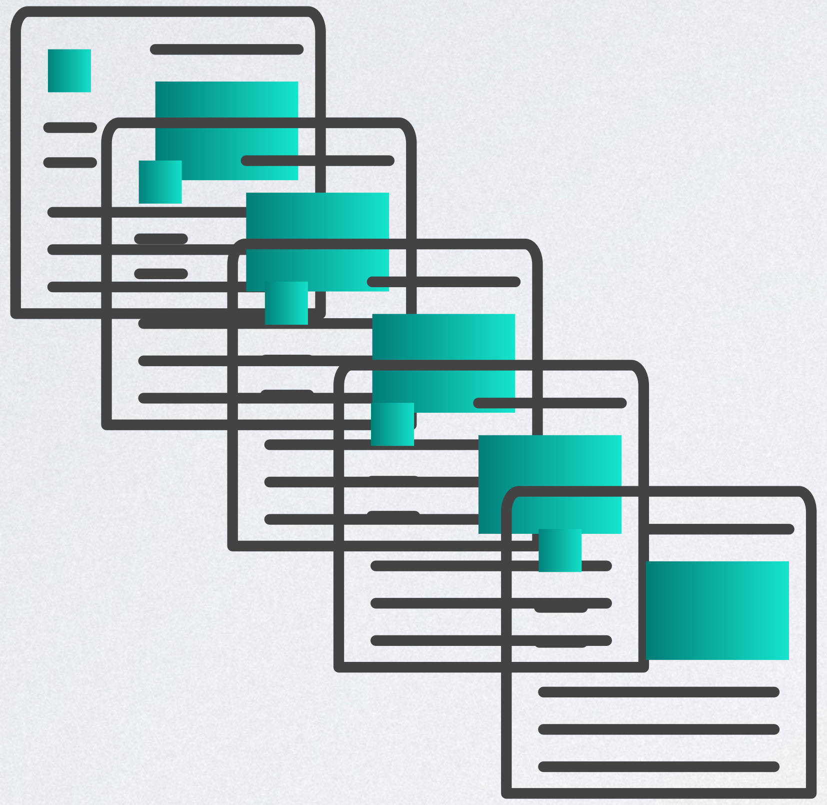
ANSWER

200,000 words

10,000 words

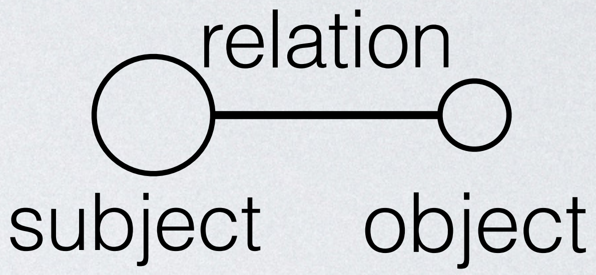
Converting a Text to a Graph

WEB DOCUMENTS



WEB DOCUMENT SENTENCES

open information extraction



coreference Resolution

Merge nodes
Increment
Nodes Weight

Tf-idf filtering

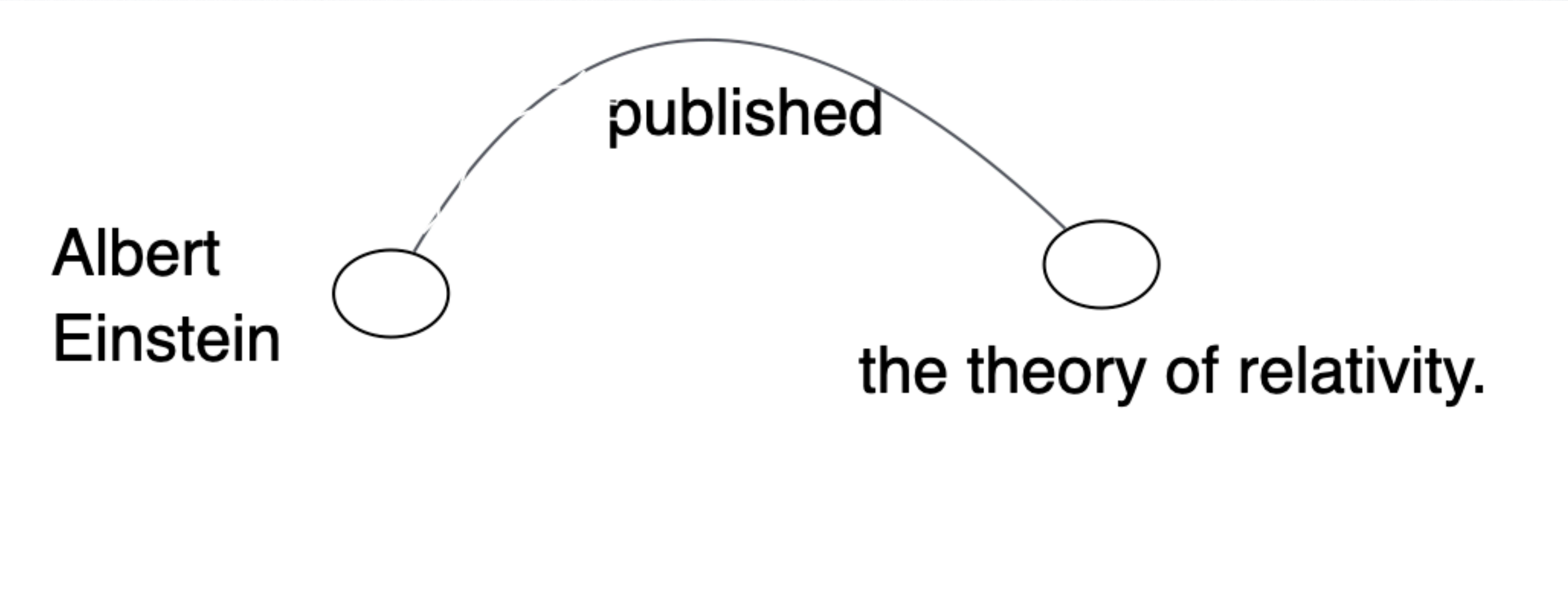
Filter Irrelevant Input
Merge similar nodes and edges

Open Information Extraction

Converting text to edges

Can someone explain the theory of relativity ?

Albert Einstein, a German theoretical physicist , published **the theory of relativity.**



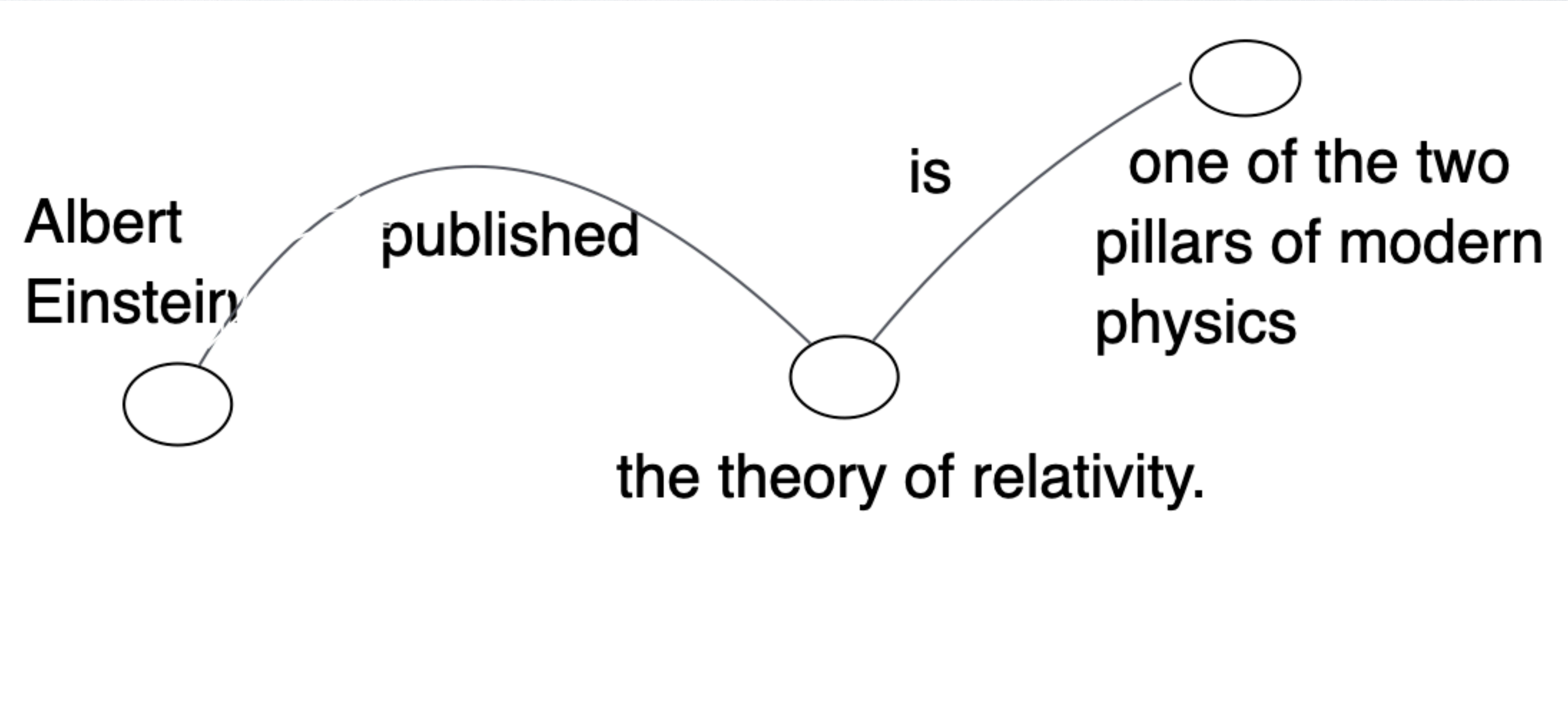
Coreference

Merging nodes

Can someone explain the theory of relativity ?

Albert Einstein, a German theoretical physicist , published the theory of relativity.

The theory of relativity is **one of the two pillars of modern physics**
Node weight +1


Albert Einstein
published
is
one of the two pillars of modern physics
the theory of relativity.

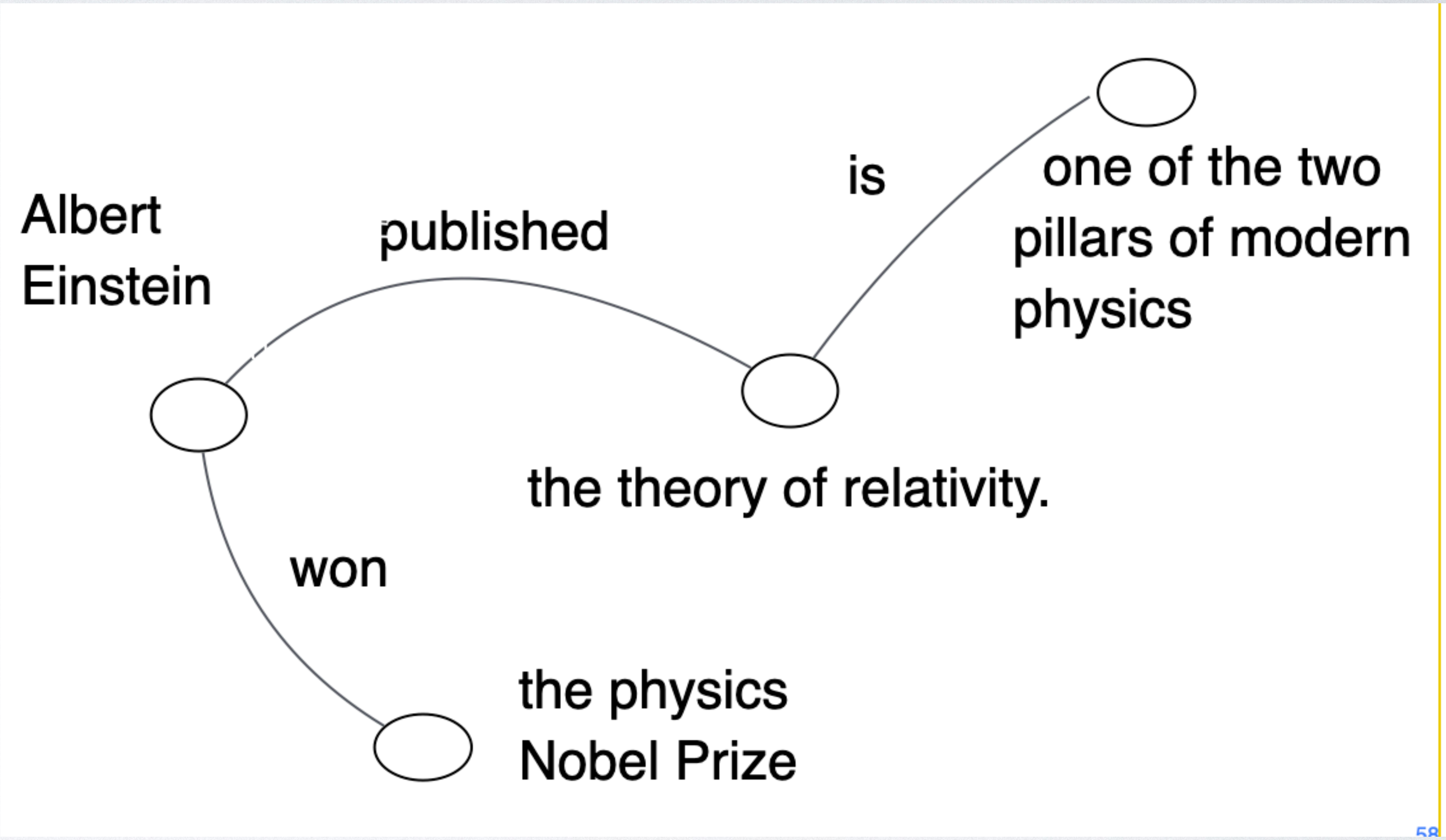
Coreference

Merging nodes

Can someone explain the theory of relativity ?

Albert Einstein, a German theoretical physicist , published the theory of relativity.
The theory of relativity is one of the two pillars of modern physics

He won **the physic Nobel Prize**
Node weight +1



Albert Einstein

published

is

one of the two pillars of modern physics

the theory of relativity.

won

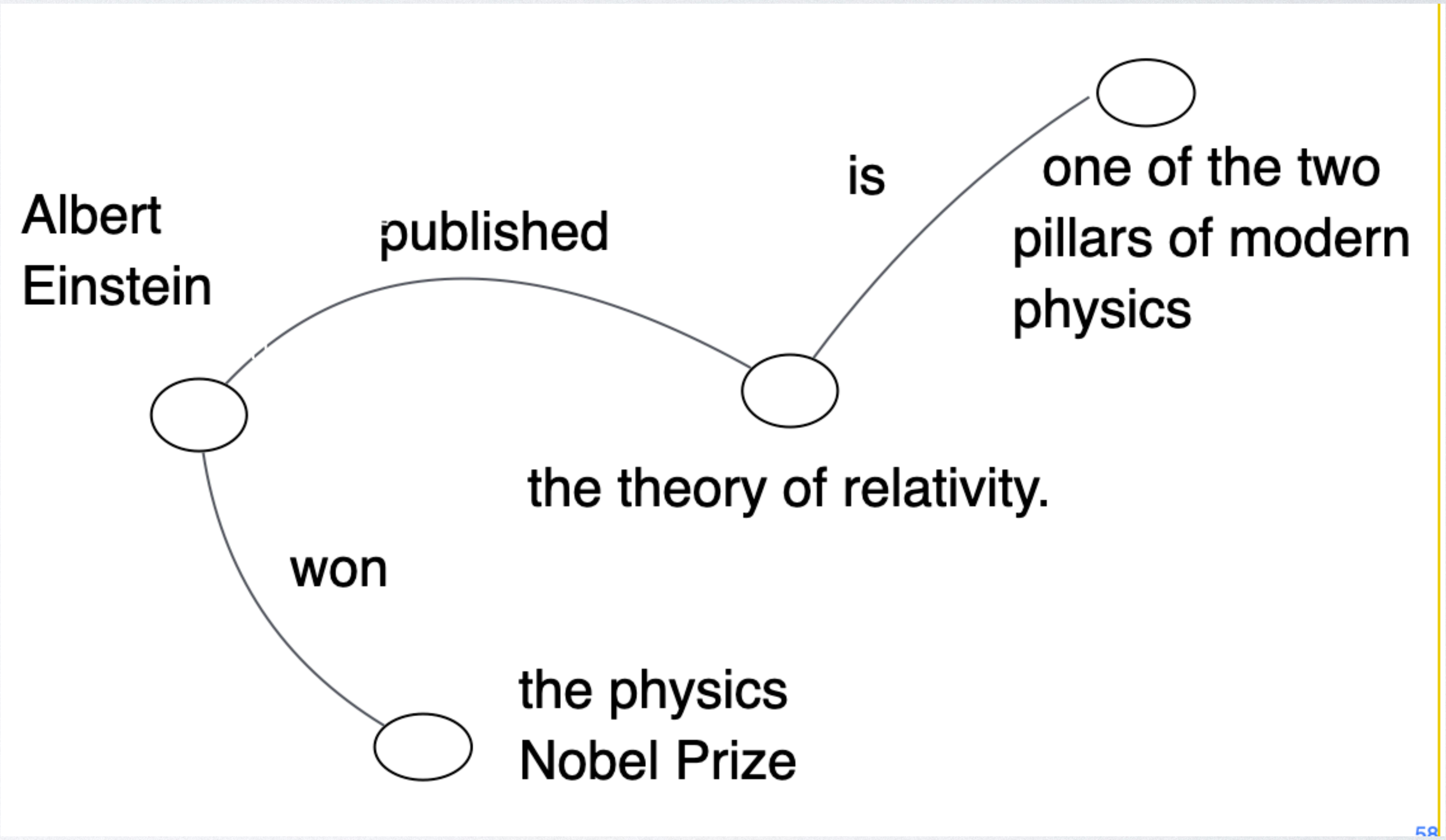
the physics Nobel Prize

Relevance Filtering

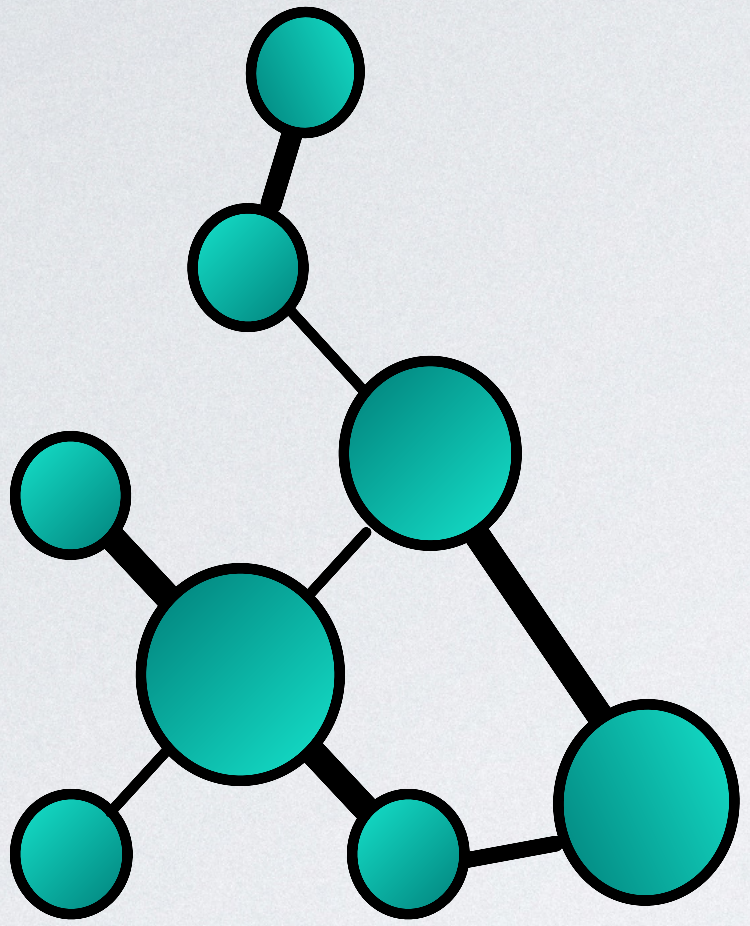
Can someone explain the theory of relativity ?

Albert Einstein, a German theoretical physicist , published the theory of relativity.
The theory of relativity is one of the two pillars of modern physics
He won the physic Nobel Prize

Puppies are very cute.
Low tf-idf with the query.
Not added


Albert Einstein
published
is
one of the two pillars of modern physics
the theory of relativity.
won
the physics Nobel Prize

Text-to-Graph Conversion



Compresses the input by

Dropping words

Filtering out irrelevant triples

Reduces redundancy

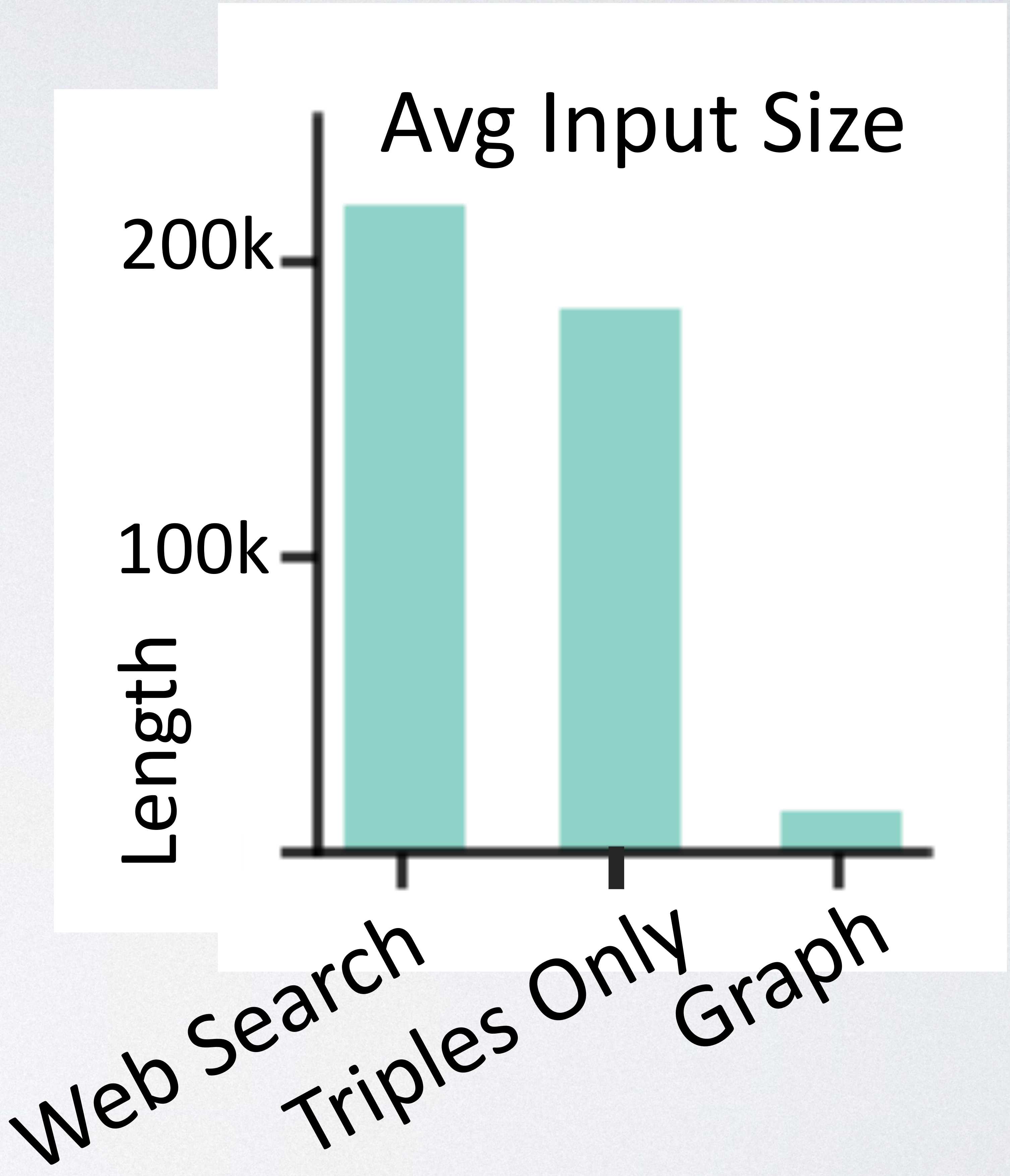
Merging nodes and edges

Filters out irrelevant content

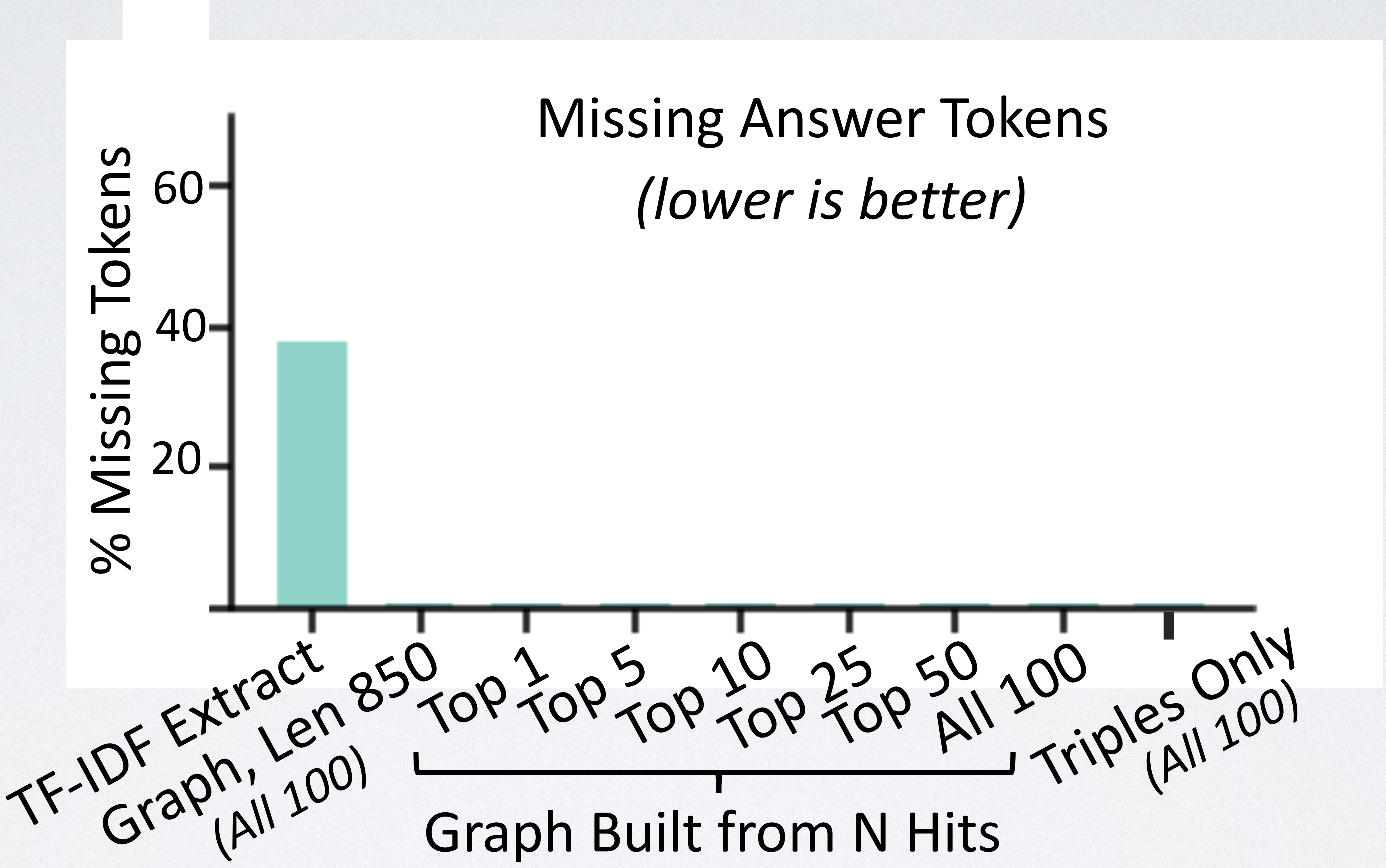
Tf-idf overlap (Question, Triple)

Knowledge Graph Construction drastically reduces the input size

The full text of the 100 web search results, which is around 200K tokens, is compressed to a few hundred tokens in the knowledge graph representation.

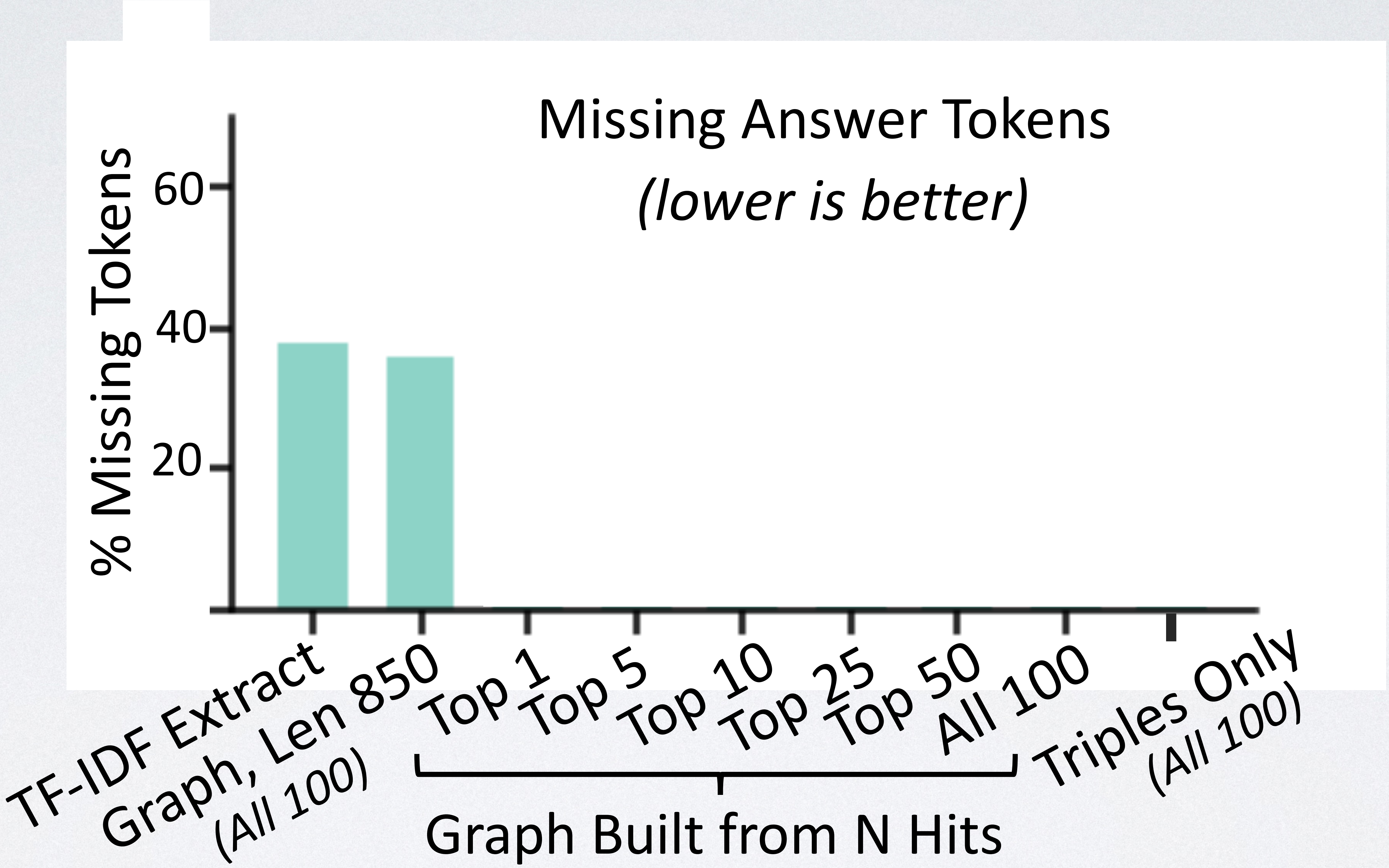


Does the graph preserve relevant information ?



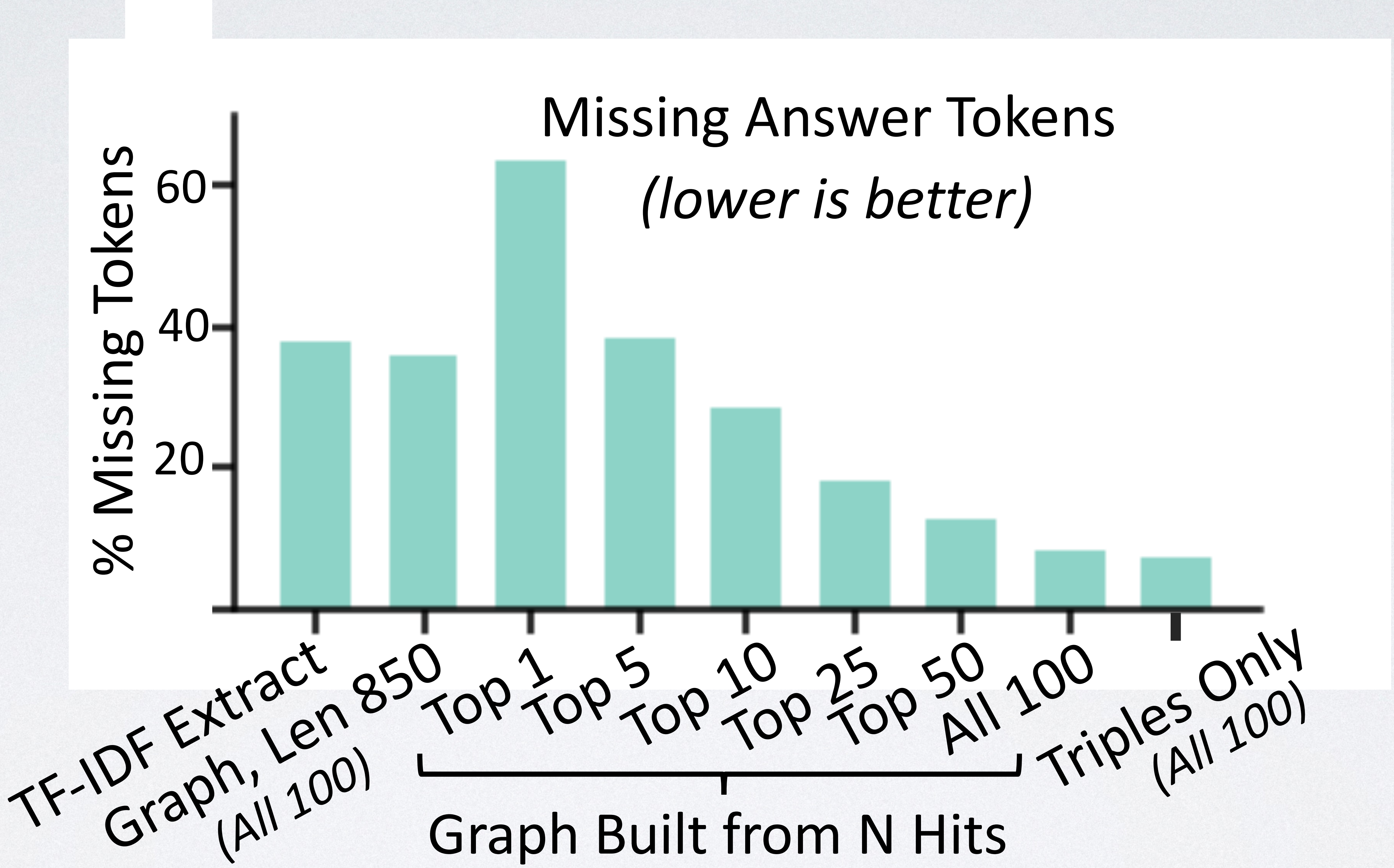
TF-IDF extraction is missing 38% of the answer tokens

Does the graph preserve relevant information ?



The graph extracted for 850 tokens is missing 35% of the answer tokens

Does the graph preserve relevant information ?

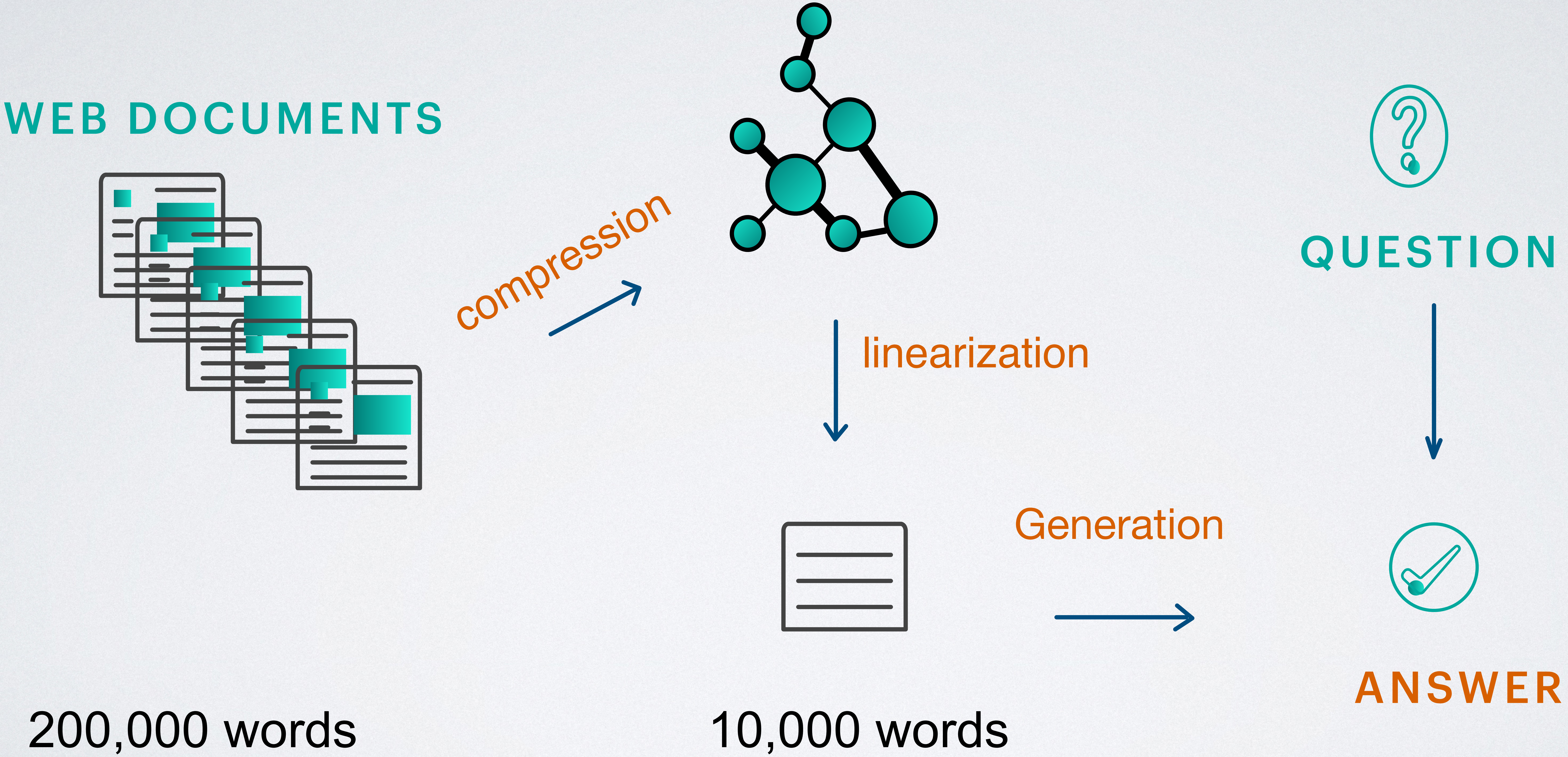


The graph for the full Input is missing only 8.7% of the answer tokens

Model

Question-Answer Model

Generating an Answer from Web Retrieval



WEB DOCUMENTS

compression

linearization

QUESTION

Generation

ANSWER

200,000 words

10,000 words

Graph Linearisation

Encoding Graph Structure in a Seq2Seq Model

WORD EMBEDDING	<sub>	Albert	Einstein	<obj>	the	theory	of	relativity	<pred>	published	<s>	developed	<obj>	the	Physics	Nobel	Prize	<s>	won
POSITION EMBEDDING	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19

Graph Linearisation

Encoding Graph Structure in a Seq2Seq Model

WORD EMBEDDING	<sub>	Albert	Einstein	<obj>	the	theory	of	relativity	<pred>	published	<s>	developed	<obj>	the	Physics	Nobel	Prize	<s>	won
POSITION EMBEDDING	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
GRAPH WEIGHT EMBEDDING	0	4	4	0	2	2	2	2	0	1	0	1	0	3	3	3	3	0	2

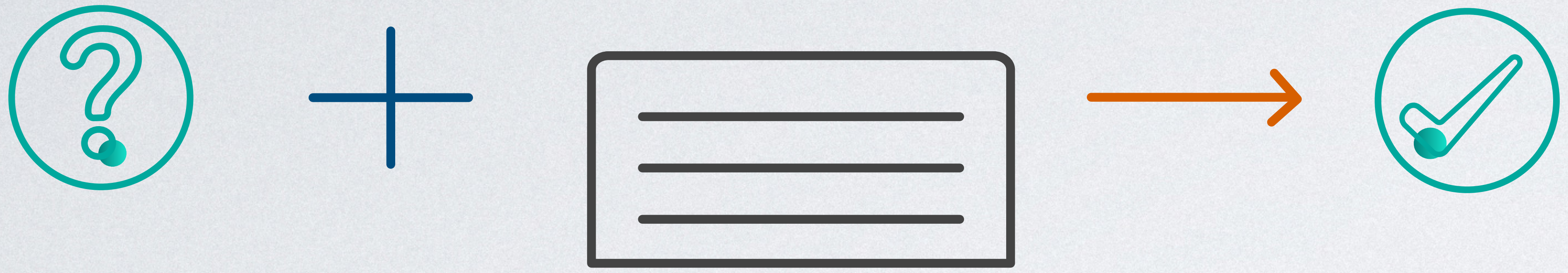
Graph Linearisation

Encoding Graph Structure in a Seq2Seq Model

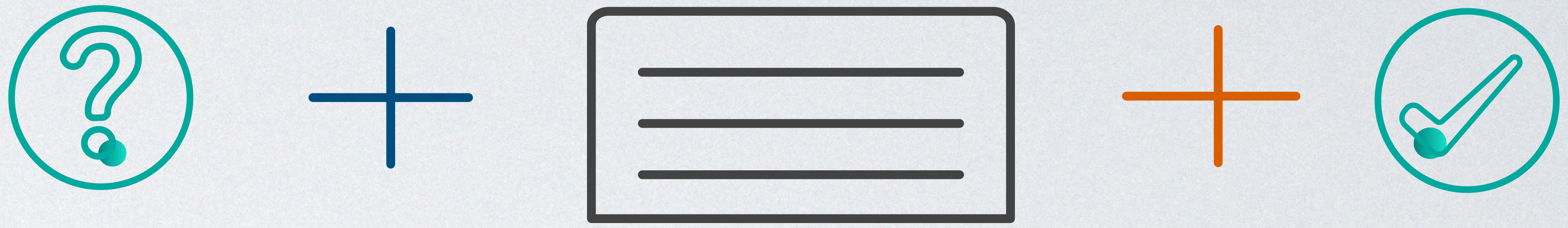
WORD EMBEDDING	<sub>	Albert	Einstein	<obj>	the	theory	of	relativity	<pred>	published	<s>	developed	<obj>	the	Physics	Nobel	Prize	<s>	won
POSITION EMBEDDING	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
GRAPH WEIGHT EMBEDDING	0	4	4	0	2	2	2	2	0	1	0	1	0	3	3	3	3	0	2
QUERY RELEVANCE EMBEDDING	0	1	1	0	1	1	1	1	0	1	0	2	0	1	1	1	1	0	1

Multi-task Learning

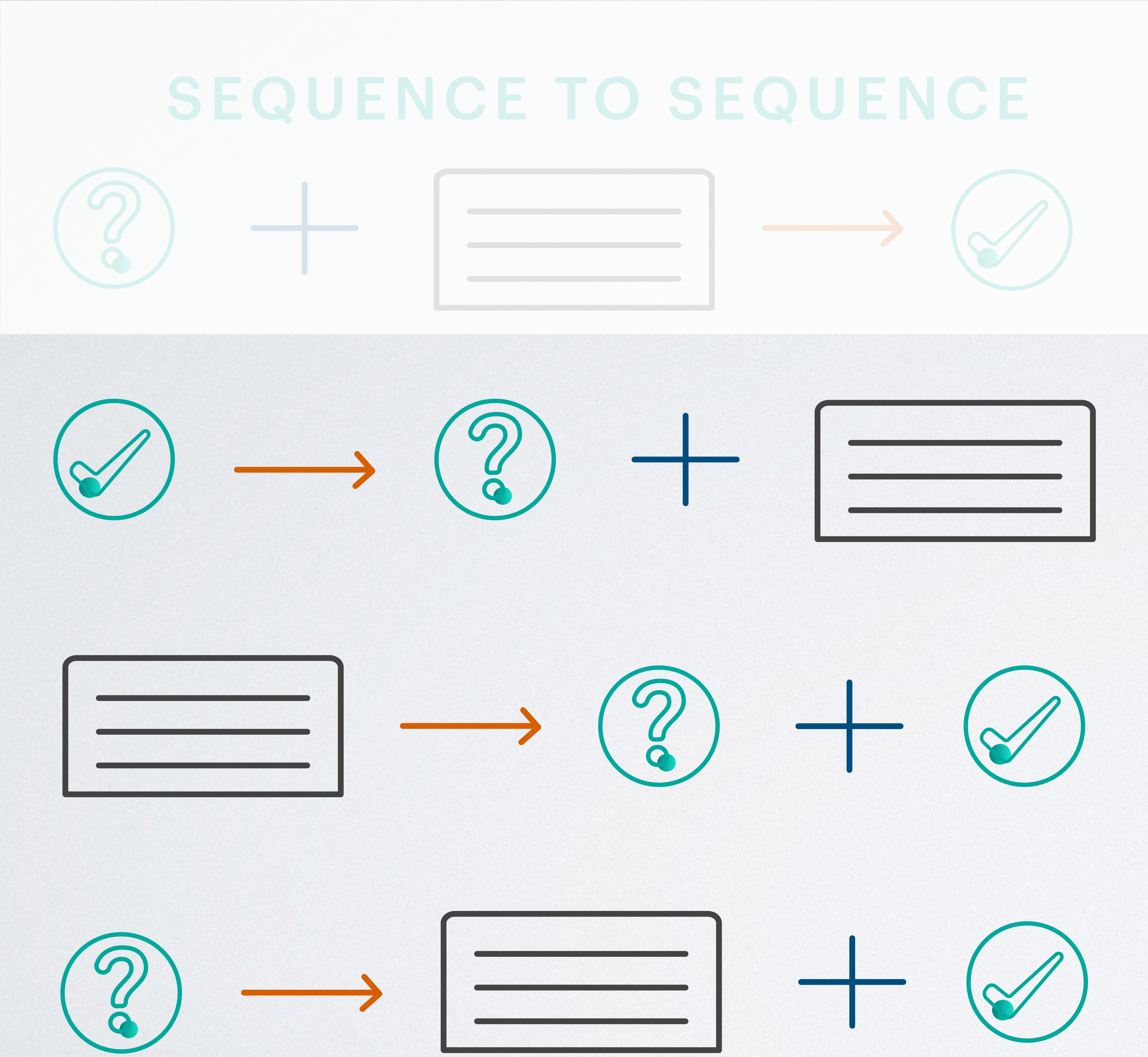
ENCODER DECODER

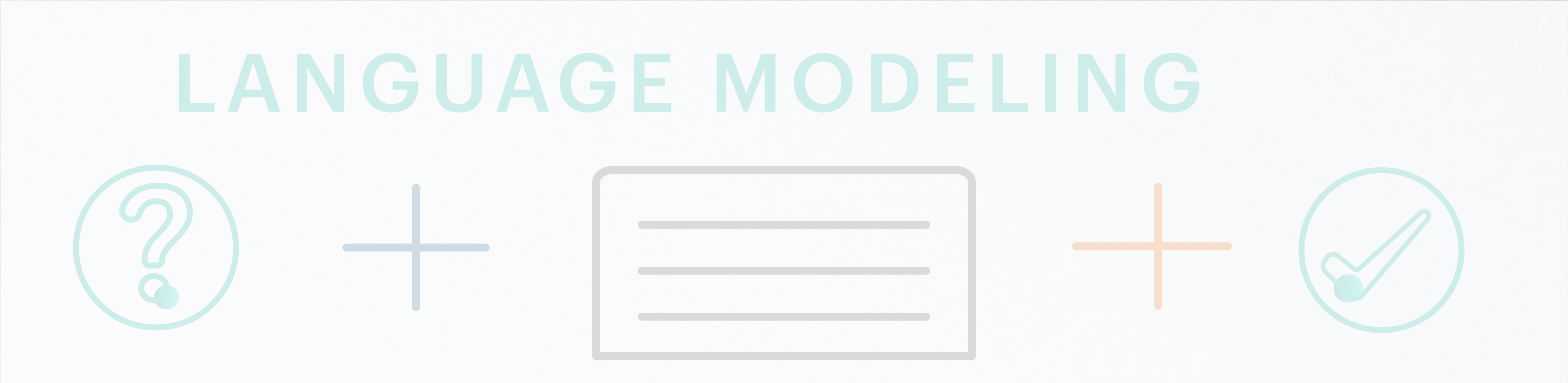


LANGUAGE MODEL



Multi-task Learning



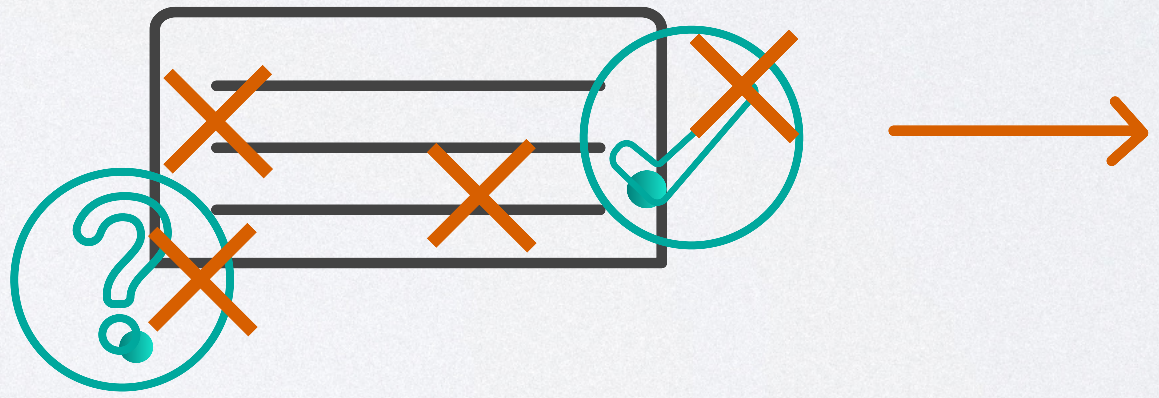


Multi-task Learning

SEQUENCE TO SEQUENCE

LANGUAGE MODELING

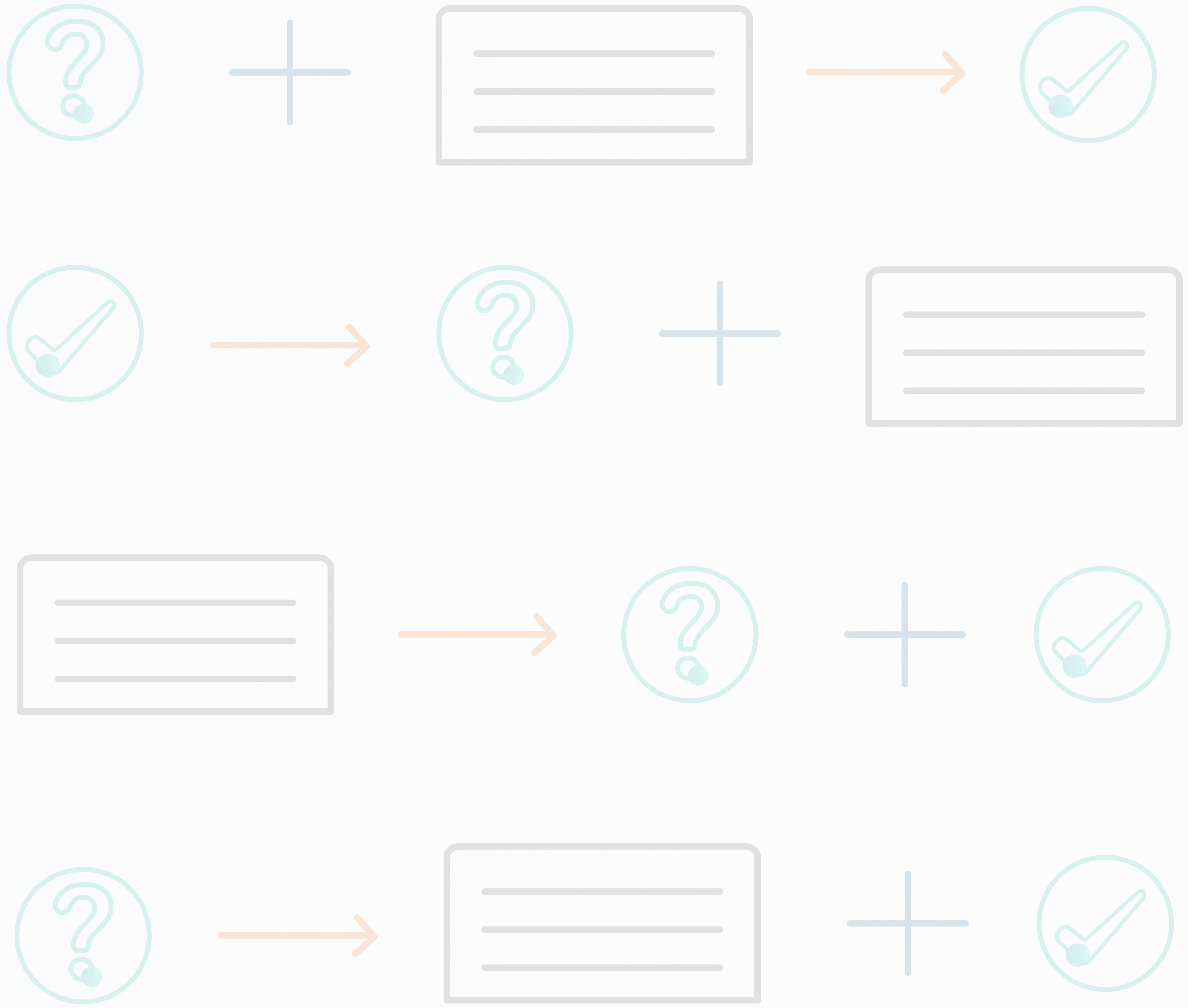
MASKED LANGUAGE MODELING



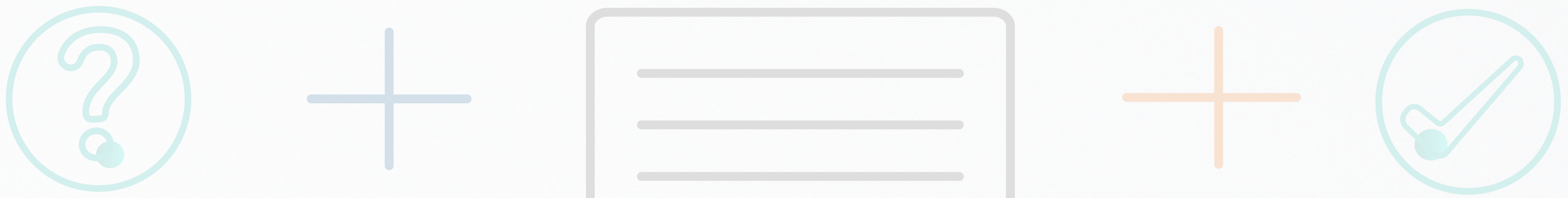
masked words

Multi-task Learning

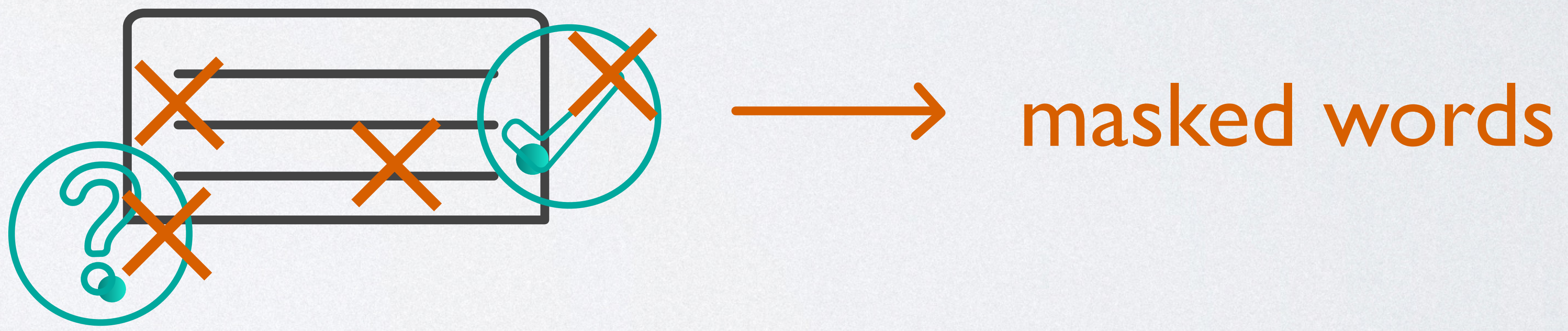
SEQUENCE TO SEQUENCE

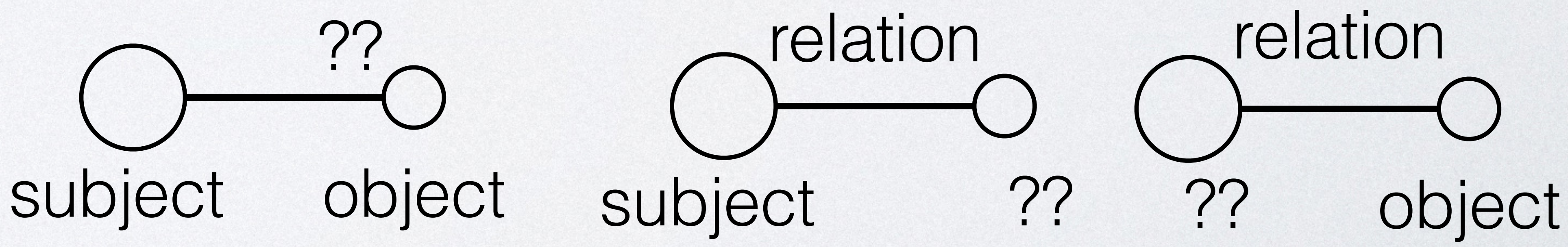


LANGUAGE MODELING



MASKED LANGUAGE MODELING



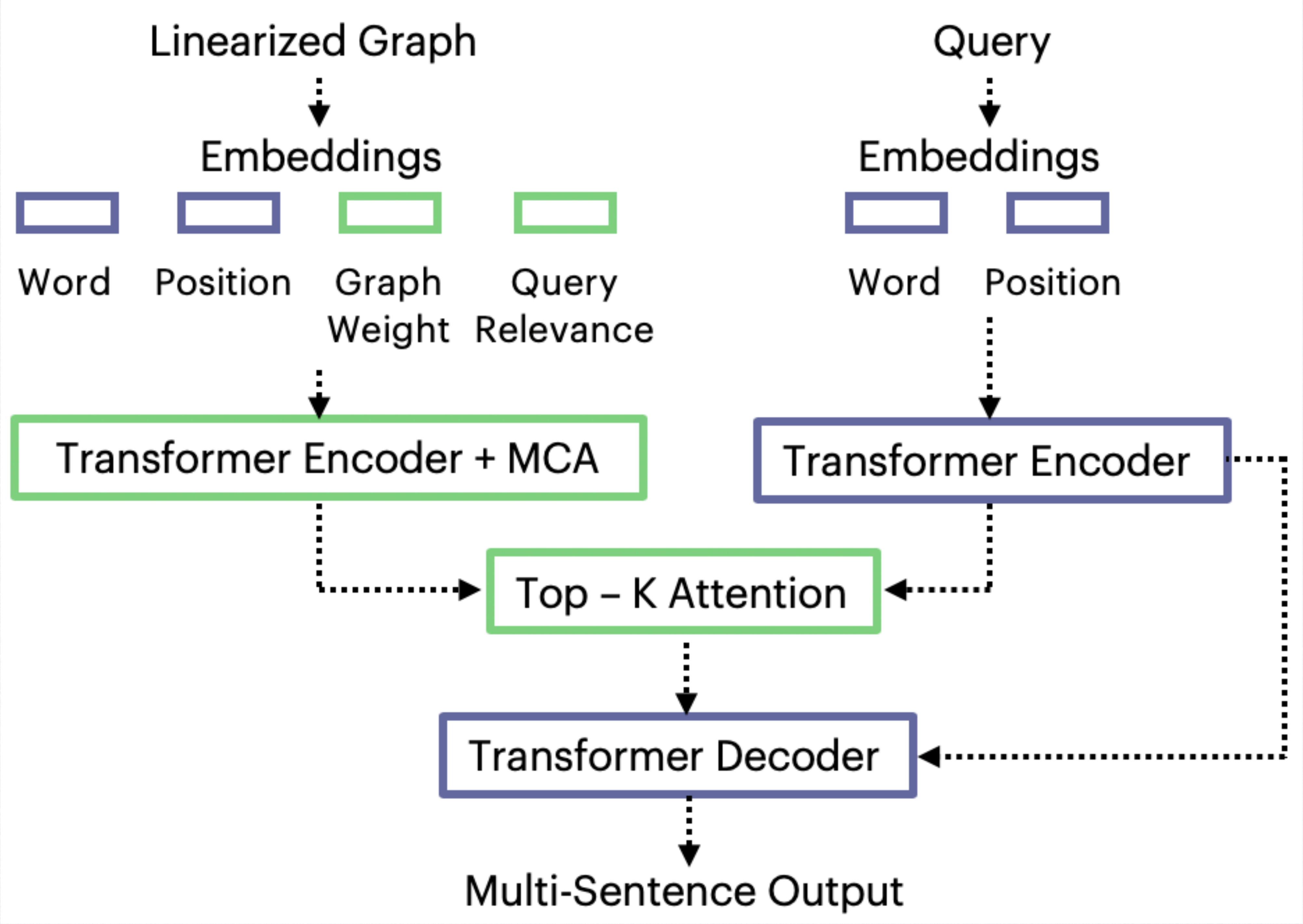


Handling Long Input

Encoding and decoding 10K tokens

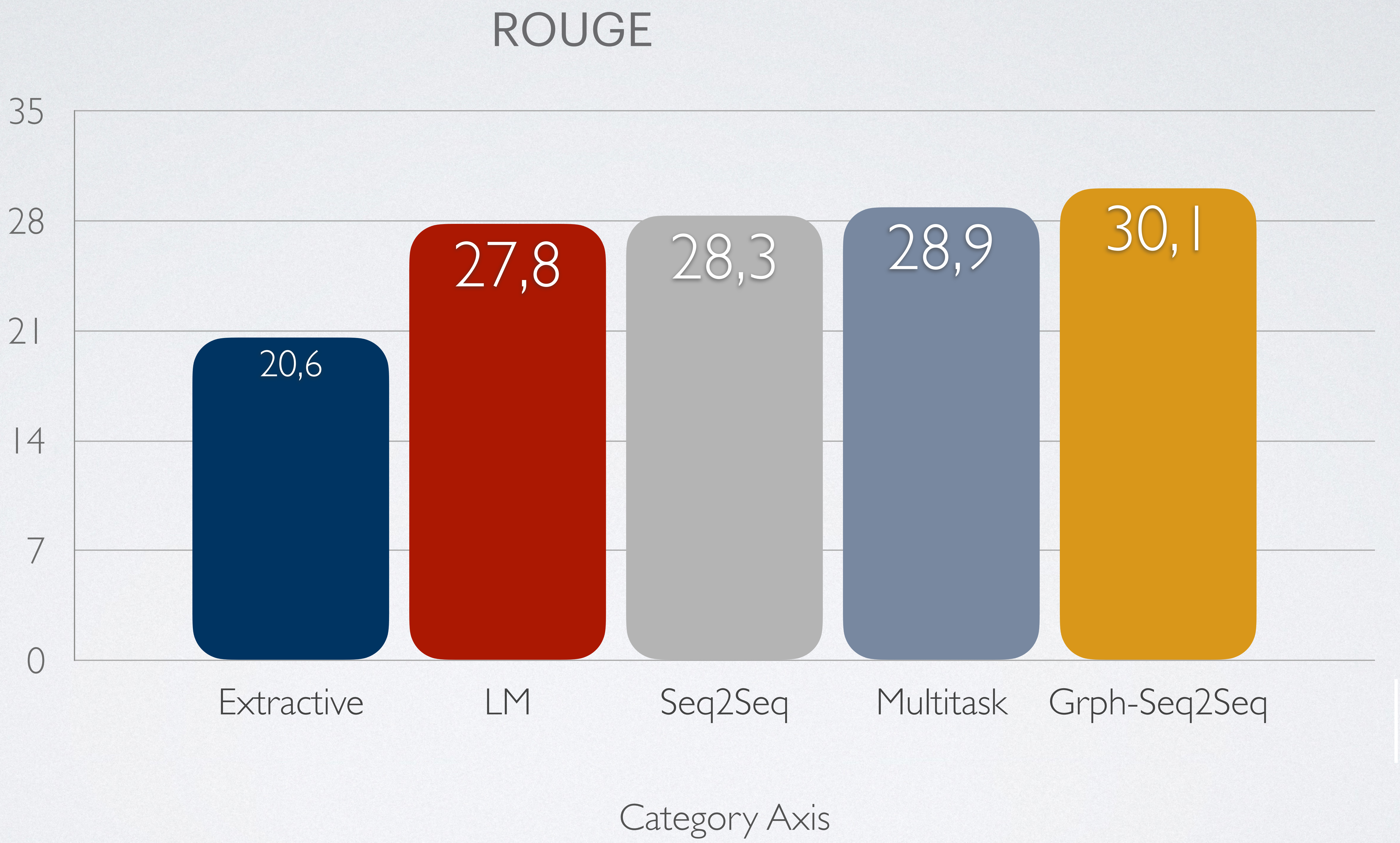
Encoder
Memory Compressed Attention

Decoder
Top-K attention

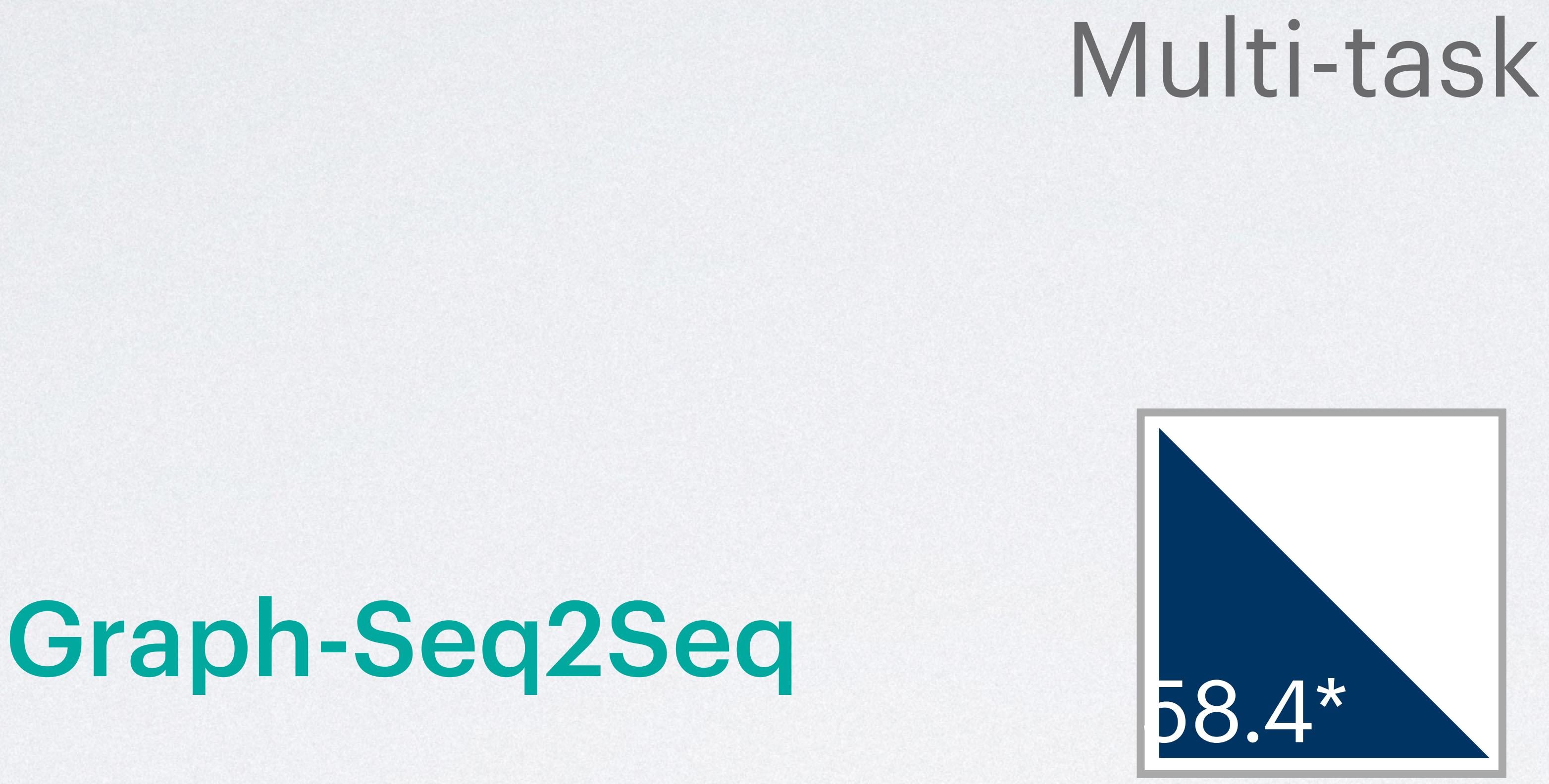


Evaluation

Automatic Evaluation



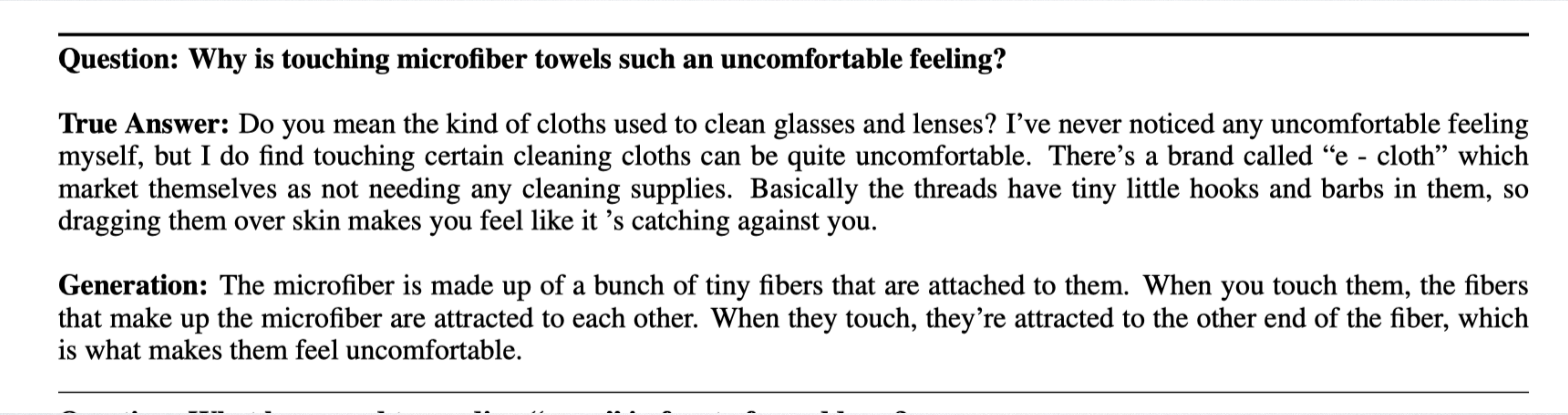
Human Evaluation: Preference



	Multi-task
Graph-Seq2Seq	58.4*

The evaluators preferred the graph based approach 58.4% of the time

Example of Generated Text (ELI5)



Question: Why is touching microfiber towels such an uncomfortable feeling?

True Answer: Do you mean the kind of cloths used to clean glasses and lenses? I've never noticed any uncomfortable feeling myself, but I do find touching certain cleaning cloths can be quite uncomfortable. There's a brand called "e - cloth" which market themselves as not needing any cleaning supplies. Basically the threads have tiny little hooks and barbs in them, so dragging them over skin makes you feel like it 's catching against you.

Generation: The microfiber is made up of a bunch of tiny fibers that are attached to them. When you touch them, the fibers that make up the microfiber are attracted to each other. When they touch, they're attracted to the other end of the fiber, which is what makes them feel uncomfortable.

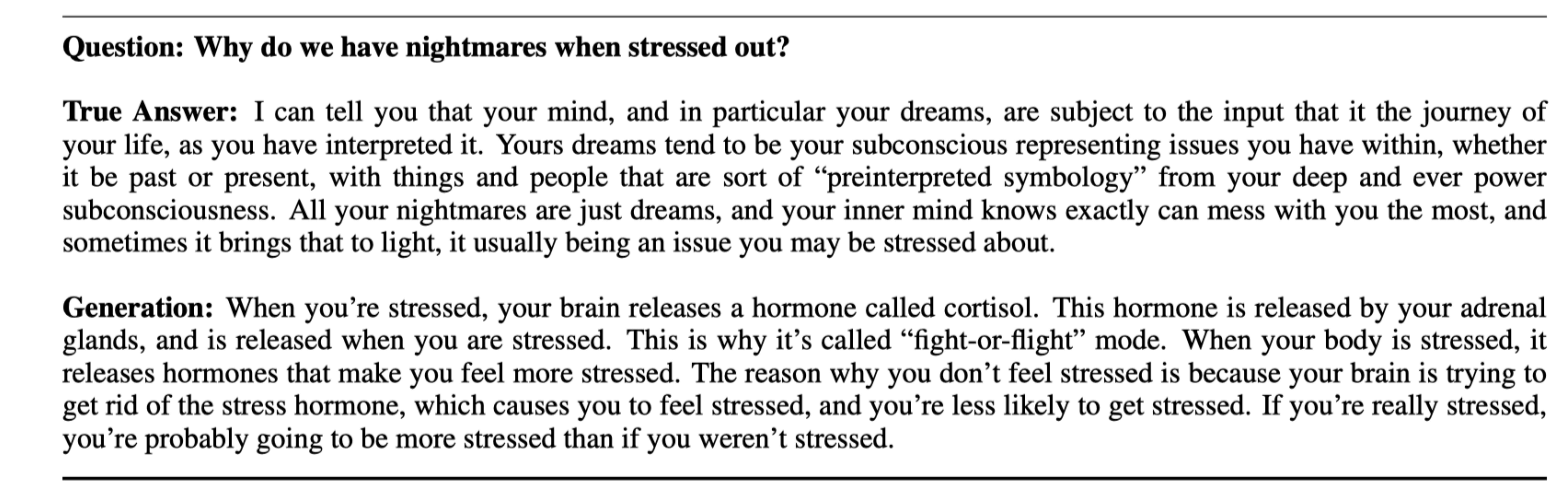
Fluent answer

Factually correct

Some repetition

("attracted to each other")

Example of Generated Text (ELI5)



Question: Why do we have nightmares when stressed out?

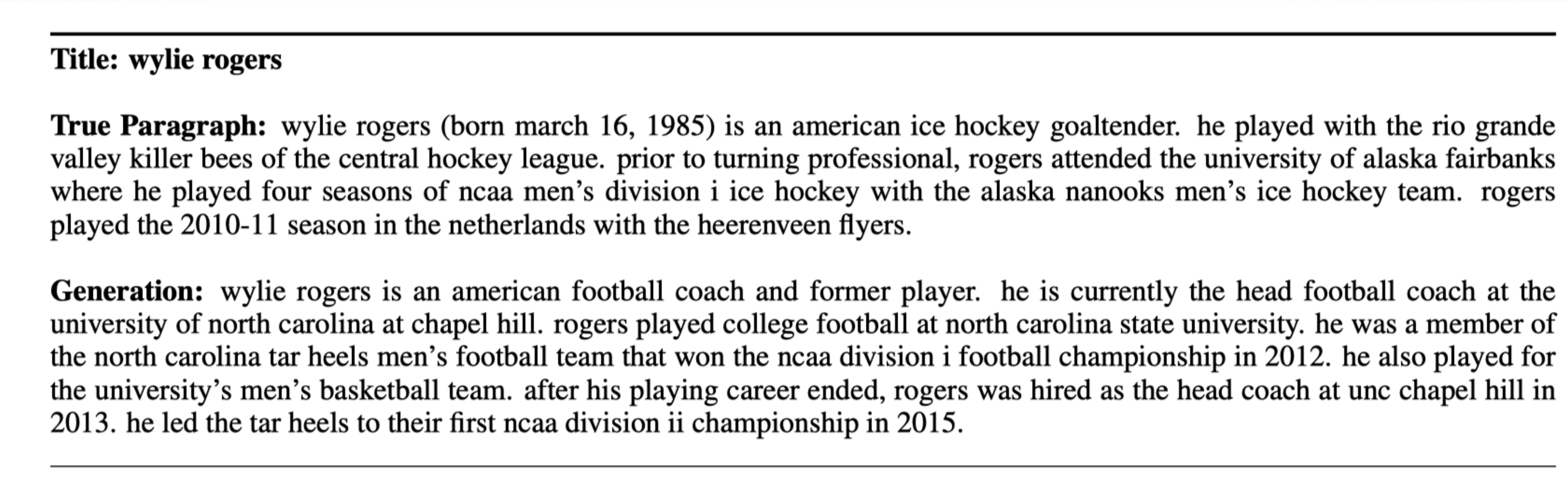
True Answer: I can tell you that your mind, and in particular your dreams, are subject to the input that it the journey of your life, as you have interpreted it. Yours dreams tend to be your subconscious representing issues you have within, whether it be past or present, with things and people that are sort of “preinterpreted symbology” from your deep and ever power subconsciousness. All your nightmares are just dreams, and your inner mind knows exactly can mess with you the most, and sometimes it brings that to light, it usually being an issue you may be stressed about.

Generation: When you’re stressed, your brain releases a hormone called cortisol. This hormone is released by your adrenal glands, and is released when you are stressed. This is why it’s called “fight-or-flight” mode. When your body is stressed, it releases hormones that make you feel more stressed. The reason why you don’t feel stressed is because your brain is trying to get rid of the stress hormone, which causes you to feel stressed, and you’re less likely to get stressed. If you’re really stressed, you’re probably going to be more stressed than if you weren’t stressed.

Repetitions (“when you are stressed”)

Non sense fragments (“The reason …get stressed”)

Example of Generated Text (WikiSum)



Title: wylie rogers

True Paragraph: wylie rogers (born march 16, 1985) is an american ice hockey goaltender. he played with the rio grande valley killer bees of the central hockey league. prior to turning professional, rogers attended the university of alaska fairbanks where he played four seasons of ncaa men's division i ice hockey with the alaska nanooks men's ice hockey team. rogers played the 2010-11 season in the netherlands with the heerenveen flyers.

Generation: wylie rogers is an american football coach and former player. he is currently the head football coach at the university of north carolina at chapel hill. rogers played college football at north carolina state university. he was a member of the north carolina tar heels men's football team that won the ncaa division i football championship in 2012. he also played for the university's men's basketball team. after his playing career ended, rogers was hired as the head coach at unc chapel hill in 2013. he led the tar heels to their first ncaa division ii championship in 2015.

Fluent text with nice, elaborate discourse structure

Incorrect facts (Wylie Rogers is an ice hockey player not a football coach. University of Alaska not North Carolina. etc;)

Human-Machine Dialog

Retrieving and Integrating Knowledge from multiple sources

Human-Machine Dialog

Retrieving Knowledge to generate good Responses

Human: I love watching Disney movies! Do you like Disney ?
Model: I love Disney, I love watching Disney movies and different **animations and characters** that makes them unique
Human: What's your favorite Disney movie ?
Model: I love **the incredibles**, they are my favorite Disney movie.
Human: omg, me too! The second one just came out. I don't watch a lot of sequels though, I think the last one I saw was Toy Story
Model: That's cool. I remember when **Toy Story was first released in 1995**

K-Nearest Neighbour Retrieval

Two retrieval sources

Dialog

Human: What's your favorite Disney movie ?

Model: **I love the incredibles,** they are my **favorite** Disney movie.

Knowledge

Disney announced intentions to develop additional superhero films after the success by **the incredibles**

Template

I love kiteboarding, it is one of **my favorite** activities on the water.

Wizard of Wikipedia

Dialog about a topic

Retrieval Corpus for KL

- WKP passages
- 34 per topic

Retrieval Corpus for Template

- Dialog turns
- 170K dialog turns

Image Chat

Dialog about an image

Retrieval Corpus for KL

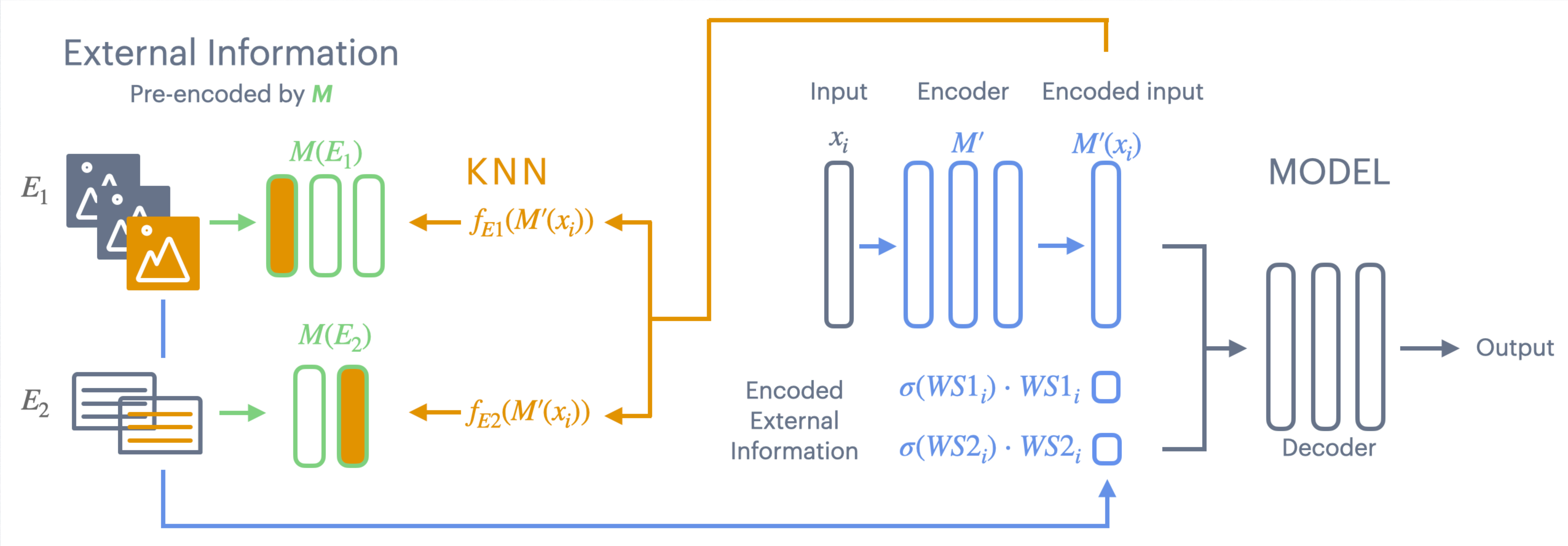
- Image + dialog
- 184K images

Retrieval Corpus for Template

- Dialog turns
- 350K dialog turns

Retrieval-Based Human-Machine Dialog

Fan et al. TACL 2021

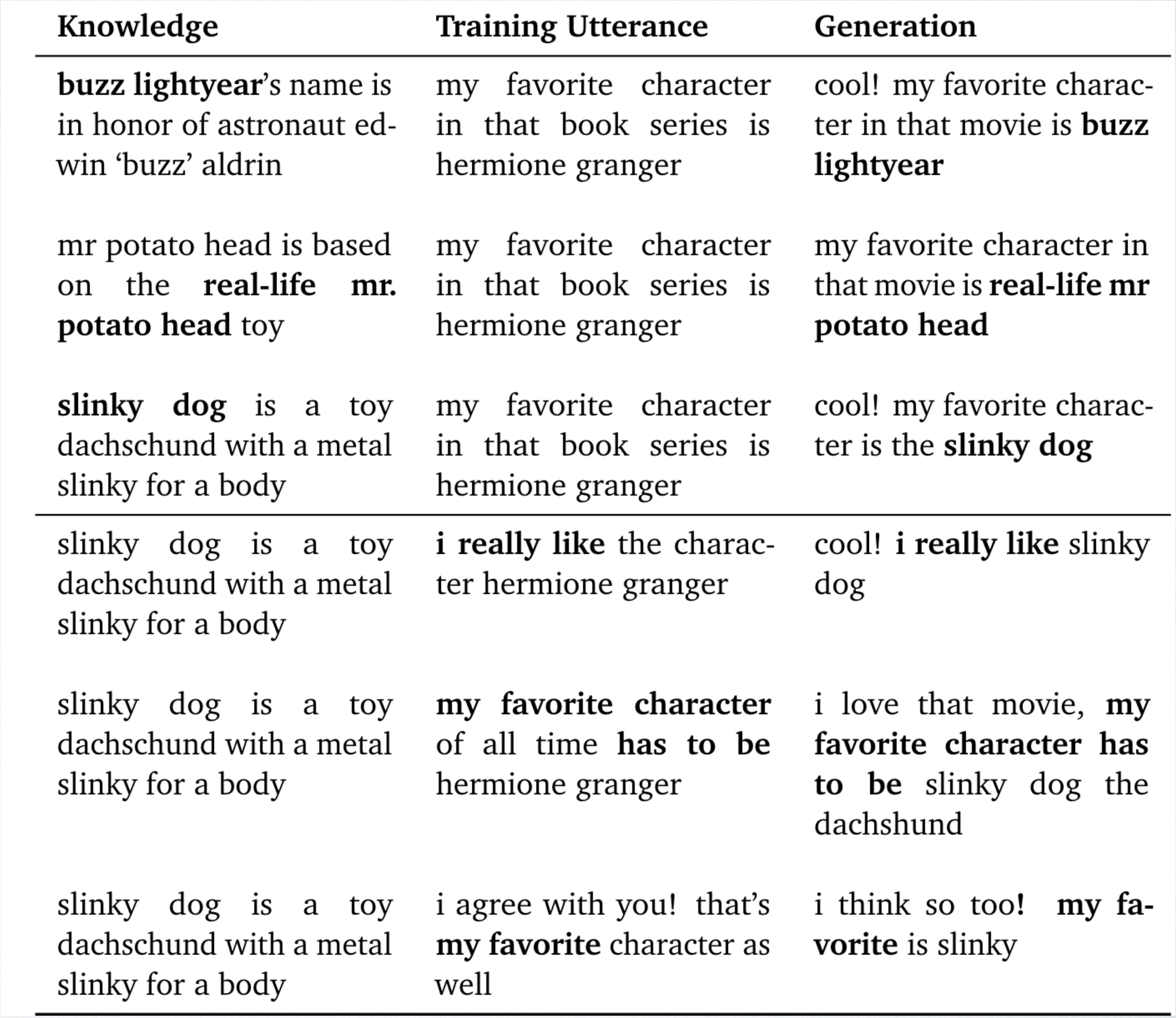


K-Nearest Neighbour Search

Effect of Fetched Text on Generation

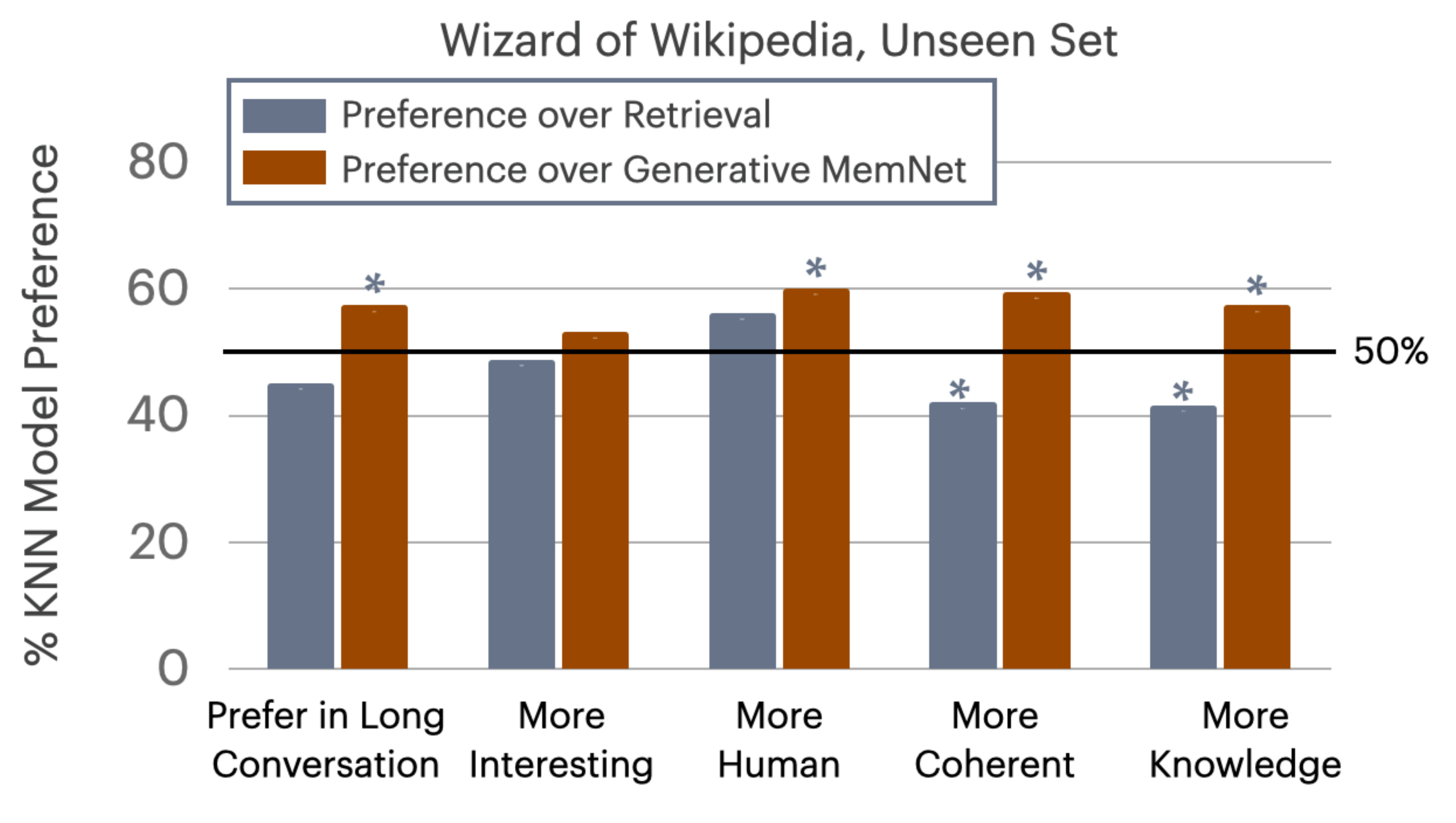
Keeping the template fixed

Keeping the KL fixed



Knowledge	Training Utterance	Generation
buzz lightyear’s name is in honor of astronaut edwin ‘buzz’ aldrin	my favorite character in that book series is hermione granger	cool! my favorite character in that movie is **buzz lightyear**
mr potato head is based on the **real-life mr. potato head** toy	my favorite character in that book series is hermione granger	my favorite character in that movie is **real-life mr potato head**
slinky dog is a toy dachshund with a metal slinky for a body	my favorite character in that book series is hermione granger	cool! my favorite character is the **slinky dog**
slinky dog is a toy dachshund with a metal slinky for a body	**i really like** the character hermione granger	cool! **i really like** slinky dog
slinky dog is a toy dachshund with a metal slinky for a body	**my favorite character** of all time **has to be** hermione granger	i love that movie, **my favorite character has to be** slinky dog the dachshund
slinky dog is a toy dachshund with a metal slinky for a body	i agree with you! that’s **my favorite** character as well	i think so too! **my favorite** is slinky

Human Evaluation

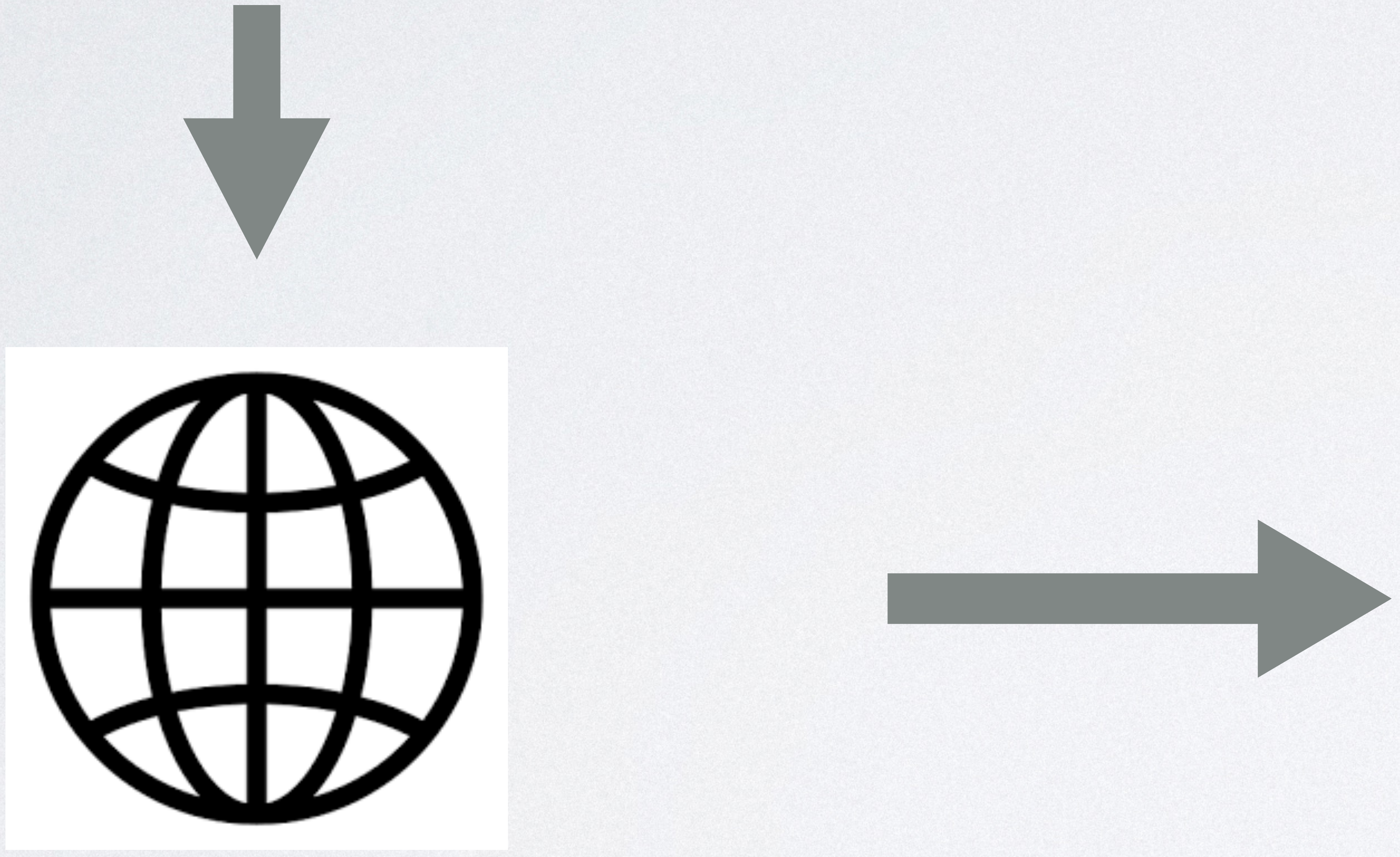


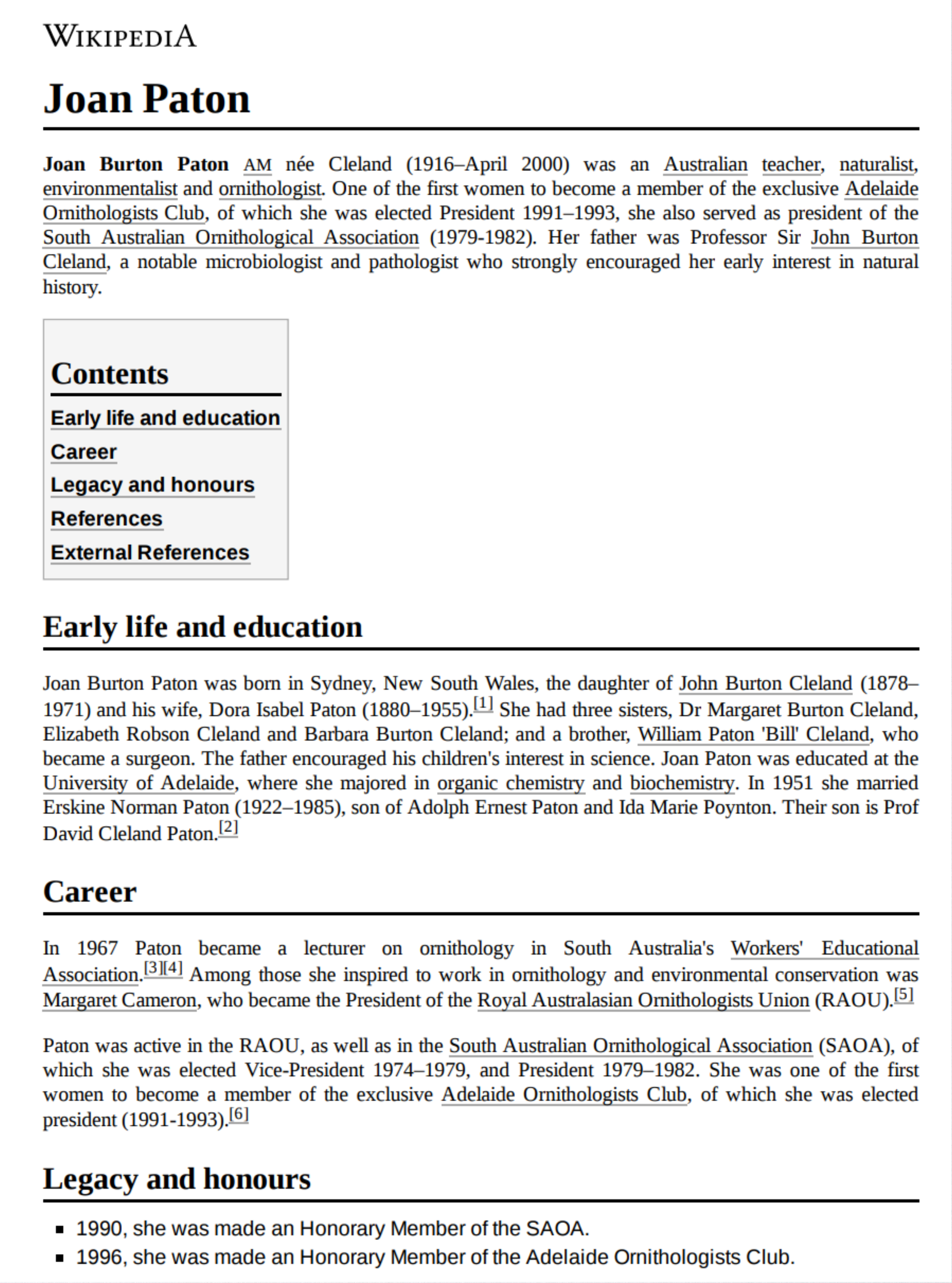
Generating Wikipedia Biographies

Studying the impact of Gender Bias on Retrieval-Augmented NLG

Generating Wikipedia Biographies from Web Retrieval

PERSON NAME





WIKIPEDIA

Joan Paton

Joan Burton Paton AM née Cleland (1916–April 2000) was an Australian teacher, naturalist, environmentalist and ornithologist. One of the first women to become a member of the exclusive Adelaide Ornithologists Club, of which she was elected President 1991–1993, she also served as president of the South Australian Ornithological Association (1979-1982). Her father was Professor Sir John Burton Cleland, a notable microbiologist and pathologist who strongly encouraged her early interest in natural history.

Contents

Early life and education

Joan Burton Paton was born in Sydney, New South Wales, the daughter of John Burton Cleland (1878–1971) and his wife, Dora Isabel Paton (1880–1955).[1] She had three sisters, Dr Margaret Burton Cleland, Elizabeth Robson Cleland and Barbara Burton Cleland; and a brother, William Paton 'Bill' Cleland, who became a surgeon. The father encouraged his children's interest in science. Joan Paton was educated at the University of Adelaide, where she majored in organic chemistry and biochemistry. In 1951 she married Erskine Norman Paton (1922–1985), son of Adolph Ernest Paton and Ida Marie Poynton. Their son is Prof David Cleland Paton.[2]

Career

In 1967 Paton became a lecturer on ornithology in South Australia's Workers' Educational Association.[3][4] Among those she inspired to work in ornithology and environmental conservation was Margaret Cameron, who became the President of the Royal Australasian Ornithologists Union (RAOU).[5]

Paton was active in the RAOU, as well as in the South Australian Ornithological Association (SAOA), of which she was elected Vice-President 1974–1979, and President 1979–1982. She was one of the first women to become a member of the exclusive Adelaide Ornithologists Club, of which she was elected president (1991-1993).[6]

Legacy and honours

- 1990, she was made an Honorary Member of the SAOA.
- 1996, she was made an Honorary Member of the Adelaide Ornithologists Club.

Challenges

Gather relevant evidence (Retrieval)

Generate a structured text

Ensure factuality

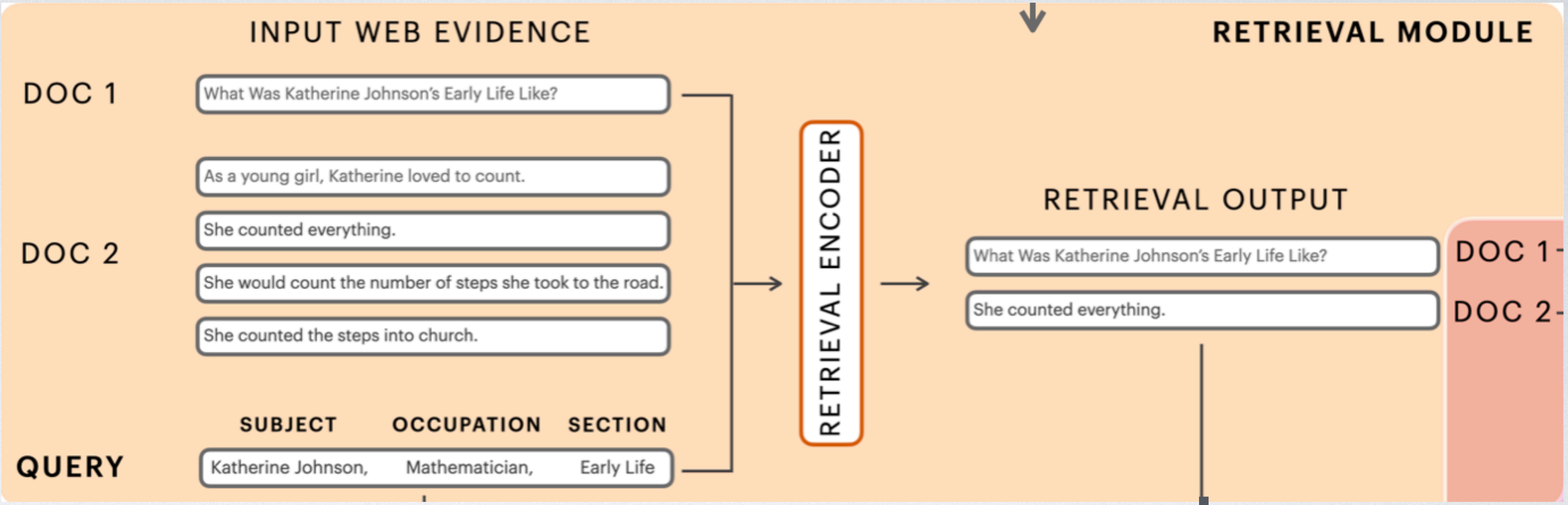
Generating Long Form Text

Fan and Gardent, ACL 2022

Dense retrieval on 1,000 tokens
(MIPS on Roberta Encodings)

Cache-based pre-trained encoder-decoder
to generate biographies section by section

Retrieval



What Was Katherine Johnson's Early Life Like?

As a young girl, Katherine loved to count.

She counted everything.

She would count the number of steps she took to the road.

She counted the steps into church.

Katherine Johnson, Mathematician, Early Life

What Was Katherine Johnson's Early Life Like?

She counted everything.

QUERY

Katherine Johnson

Mathematician

Early Life

SEARCH OUTPUT

Top 20 search results segmented into sentences

OUTPUT

40 sentences most similar with the query

(1,000 words)

Generation

QUERY

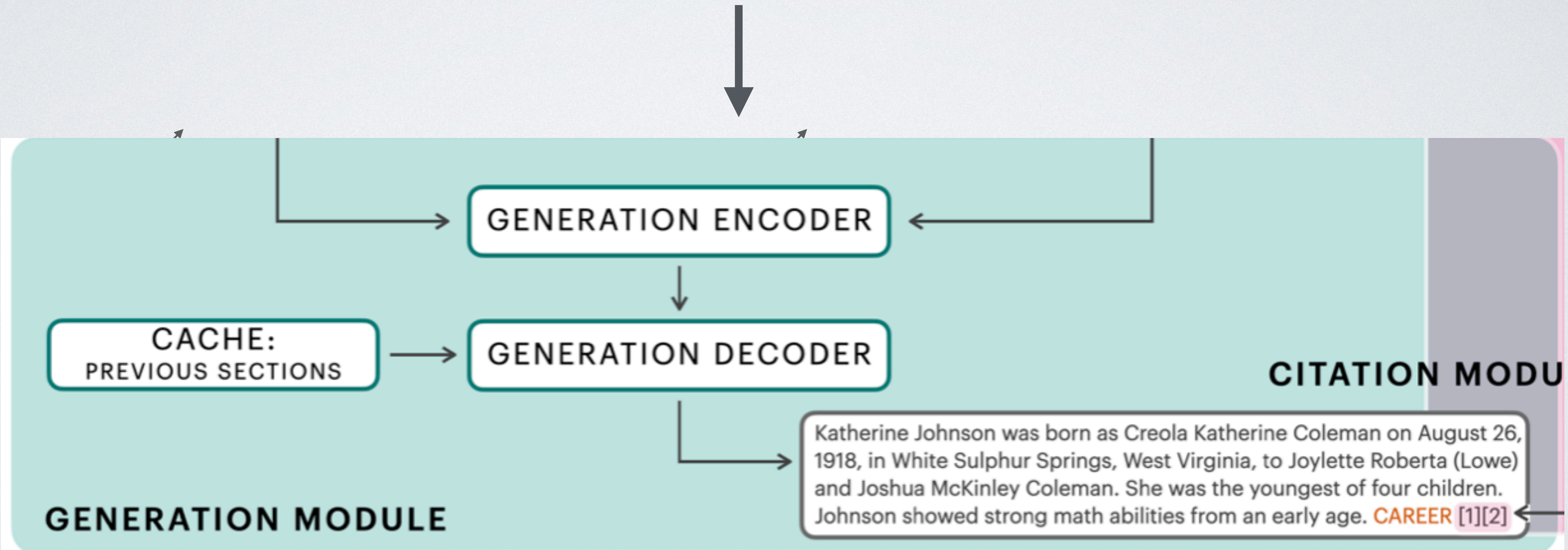
Katherine Johnson

Mathematician

Early Life

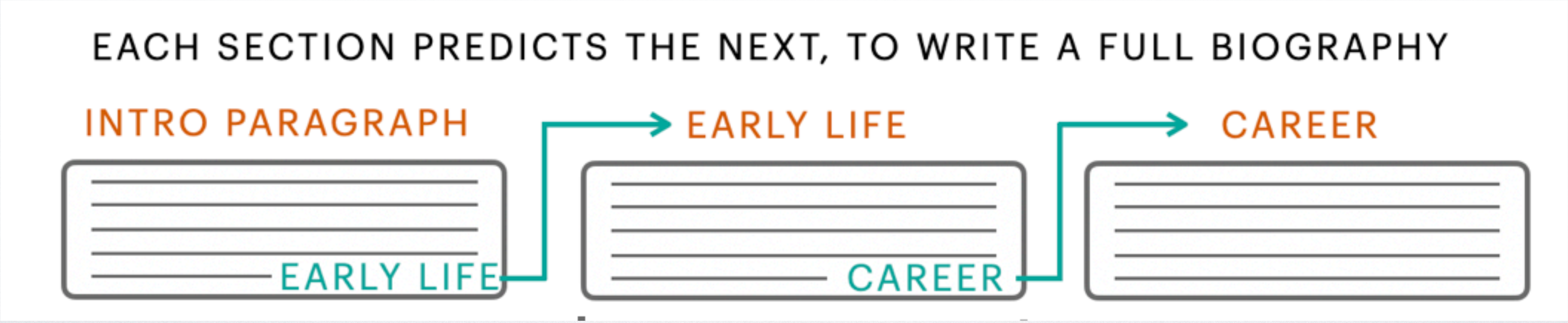
RETRIEVED EVIDENCE

1,000 words



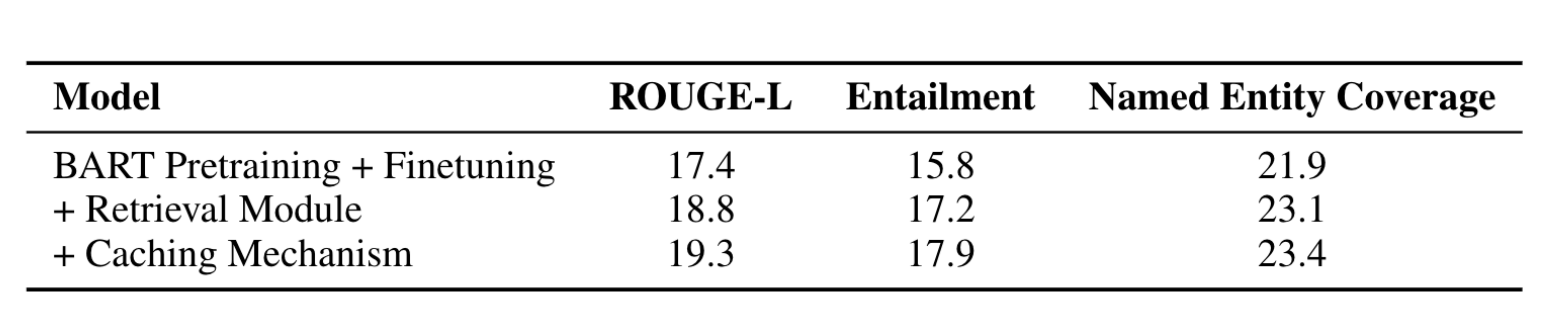
Katherine Johnson was born as Creola Katherine Coleman on August 26, 1918, in White Sulphur Springs, West Virginia, to Joylette Roberta (Lowe) and Joshua McKinley Coleman. She was the youngest of four children. Johnson showed strong math abilities from an early age. CAREER [1][2]

Transformer-XL Cache Mechanism



- Caches the previous section’s hidden states at every layer
- Usd as a memory to generate the current section

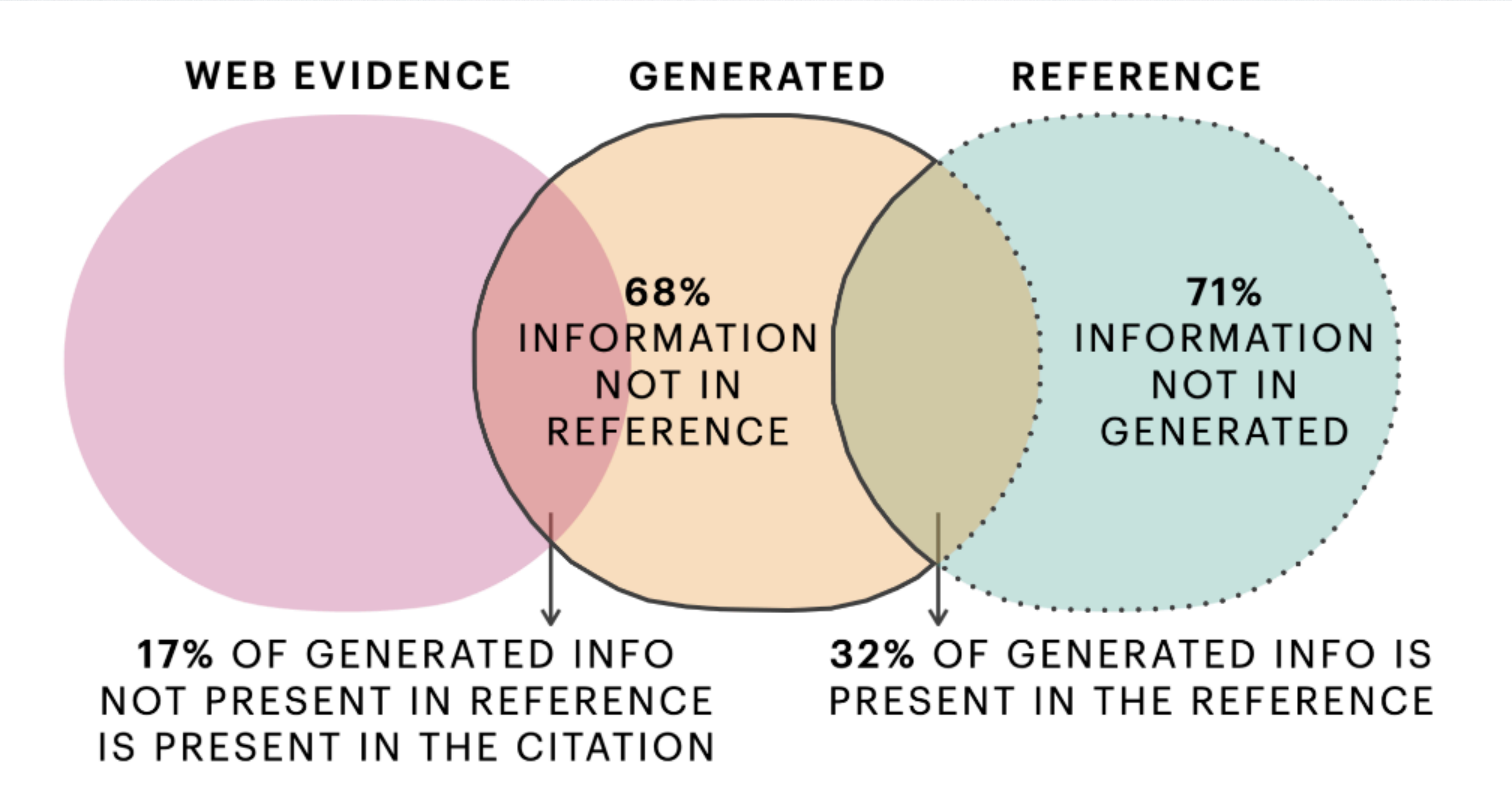
Ablation



Model	ROUGE-L	Entailment	Named Entity Coverage
BART Pretraining + Finetuning	17.4	15.8	21.9
+ Retrieval Module	18.8	17.2	23.1
+ Caching Mechanism	19.3	17.9	23.4

The retrieval and the cache module statistically significantly improve results

Human Evaluation of Factuality

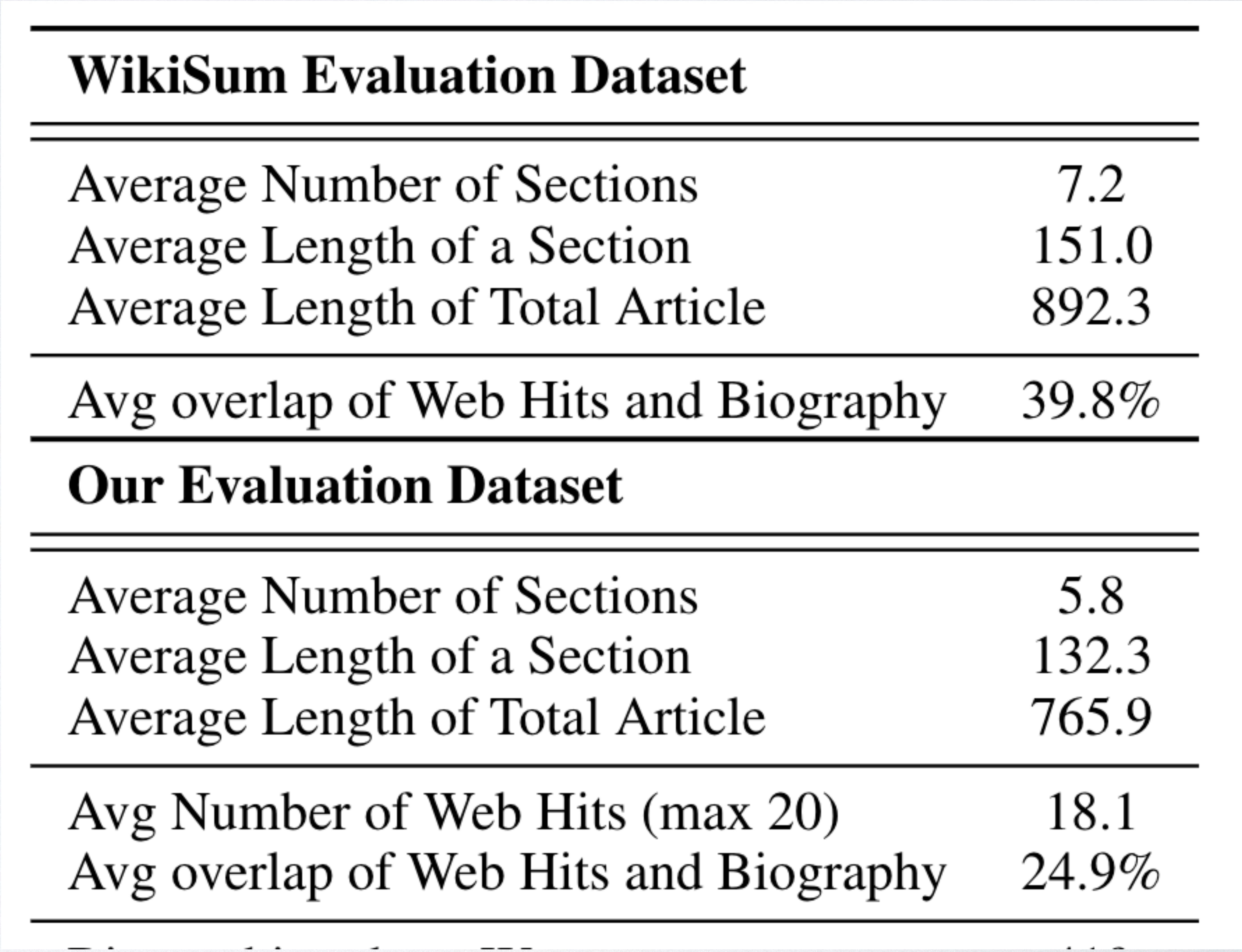

WEB EVIDENCE
GENERATED
REFERENCE
68% INFORMATION NOT IN REFERENCE
71% INFORMATION NOT IN GENERATED
17% OF GENERATED INFO NOT PRESENT IN REFERENCE IS PRESENT IN THE CITATION
32% OF GENERATED INFO IS PRESENT IN THE REFERENCE

The Evidence Gap

Wikipedia Biographies
And
Web Documents

Wikisum Test Set
Men and women

Our Test Set
Only women



WikiSum Evaluation Dataset	
Average Number of Sections	7.2
Average Length of a Section	151.0
Average Length of Total Article	892.3
Avg overlap of Web Hits and Biography	39.8%
Our Evaluation Dataset	
Average Number of Sections	5.8
Average Length of a Section	132.3
Average Length of Total Article	765.9
Avg Number of Web Hits (max 20)	18.1
Avg overlap of Web Hits and Biography	24.9%

Less Web Evidence, Less Good Texts

Model	WikiSum Test	Women	Scientists	Women in Asia	Women in Africa
BART Pretraining	19.0	17.4	18.2	16.7	16.4
+ Retrieval	21.4	18.8	19.3	17.9	17.1
+ Caching	21.8	19.3	19.7	18.4	17.3

Conclusions

Question Answering

Challenge

- Scaling to very long input

Method

- Web Documents → Graph

- Memory Compressed Attention
- Top-K attention

Human-Machine Dialog

Challenge

- Efficient retrieval on very large retrieval corpora
- Handling and combining multiple retrieval sources

Method

- K-Nearest Neighbour Search
- Multiple Encoders
- Gates

Generating Wikipedia Biographies

Challenge

- Retrieving sufficient information
- Generating Long-Form Structured Text

Method

- Dense Retrieval
- Cache

Open Challenges

Factuality

Evaluation and model improvement

Multilingual NLG

Generating into languages other than English

Multi-modal NLG

Generating from multiple input types

Thank You!