Lecture 18: Wrapup + Ethics

Alan Ritter

(many slides from Greg Durrett)

Administrivia

- Final project reports due Friday 12/8/2023 (hard deadline)
- Next Week: Guest Lectures from Dan Deutsch (Google Translate) and Luan Yi (Google AI Language)
  - Zoom link on Piazza

This Lecture

- Question Answering
- Ethics in NLP

Span-based Question Answering

SQuAD

- Single-document, single-sentence question-answering task where the answer is always a substring of the passage
- Predict start and end indices of the answer in the passage

Passage

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California.

Question: Which NFL team won Super Bowl 50?
Answer: Denver Broncos

Question: What does AFC stand for?
Answer: American Football Conference

Question: What year was Super Bowl 50?
Answer: 2016

Rajpurkar et al. (2016)

SQuAD 2.0

- SQuAD 1.1 contains 100k+ QA pairs from 500+ Wikipedia articles.
- SQuAD 2.0 includes additional 50k questions that cannot be answered.
- These questions were crowdsourced.

Passage

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California.

Question: Which NFL team won Super Bowl 50?
Answer: Denver Broncos

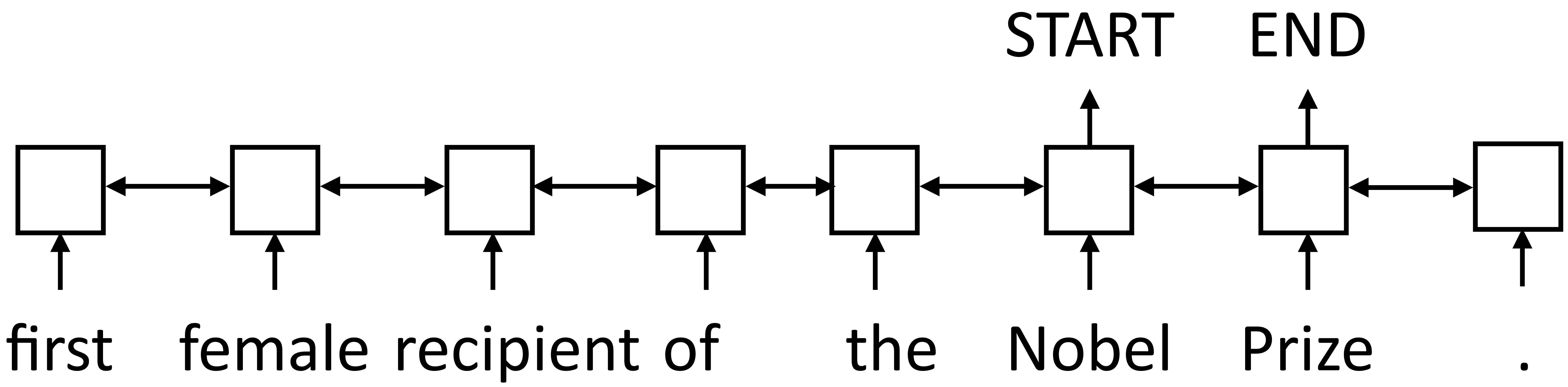
Question: What does AFC stand for?
Answer: American Football Conference

Question: What year was Super Bowl 50?
Answer: 2016

Rajpurkar et al. (2016)

SQuAD

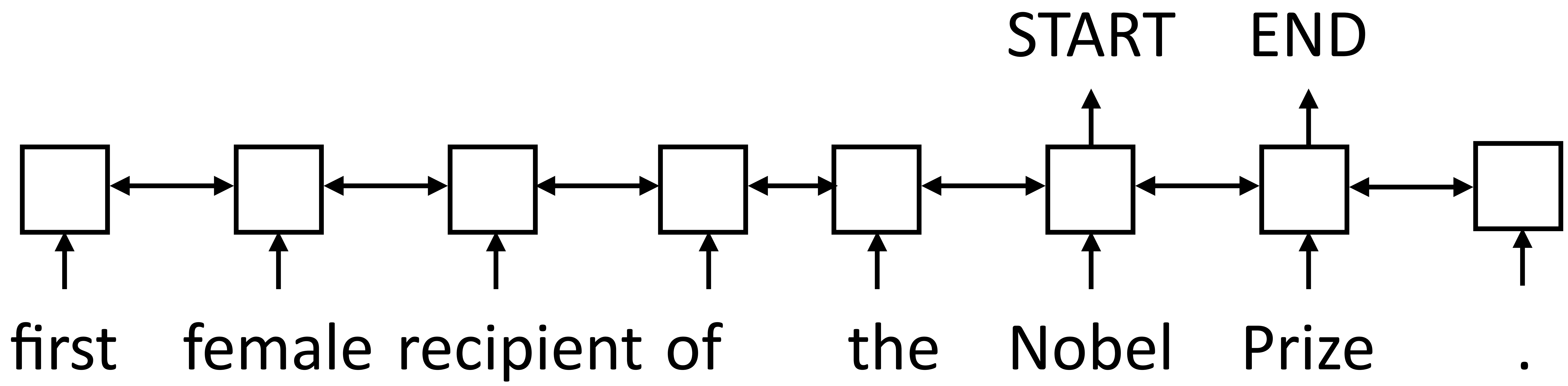
Q: What was Marie Curie the first female recipient of?



Rajpurkar et al. (2016)

SQuAD

Q: What was Marie Curie the first female recipient of?



- Like a tagging problem over the sentence (not multiclass classification), but we need some way of attending to the query

Rajpurkar et al. (2016)

Why did this take off?

- SQuAD was **big**: >100,000 questions (written by human) at a time when deep learning was exploding

- SQuAD had **room to improve**: ~50% performance from a logistic regression baseline (classifier with 180M features over constituents)

- SQuAD was **pretty easy**: year-over-year progress for a few years until the dataset was essentially solved

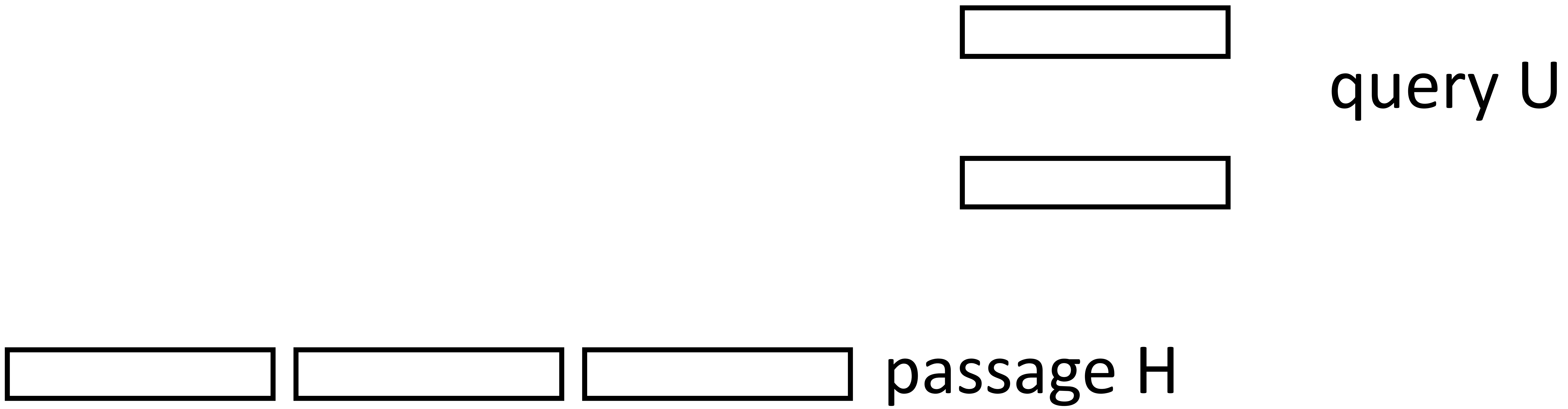
Bidirectional Attention Flow (BiDAF)

- Passage (context) and query are both encoded with BiLSTMs

Seo et al. (2016)

Bidirectional Attention Flow (BiDAF)

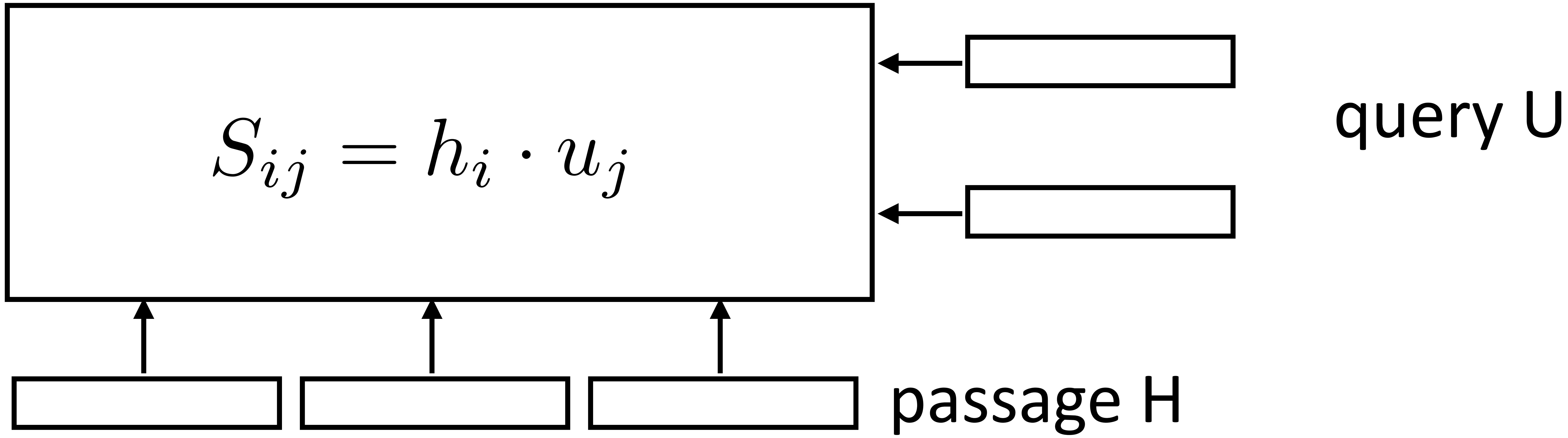
- Passage (context) and query are both encoded with BiLSTMs



Seo et al. (2016)

Bidirectional Attention Flow (BiDAF)

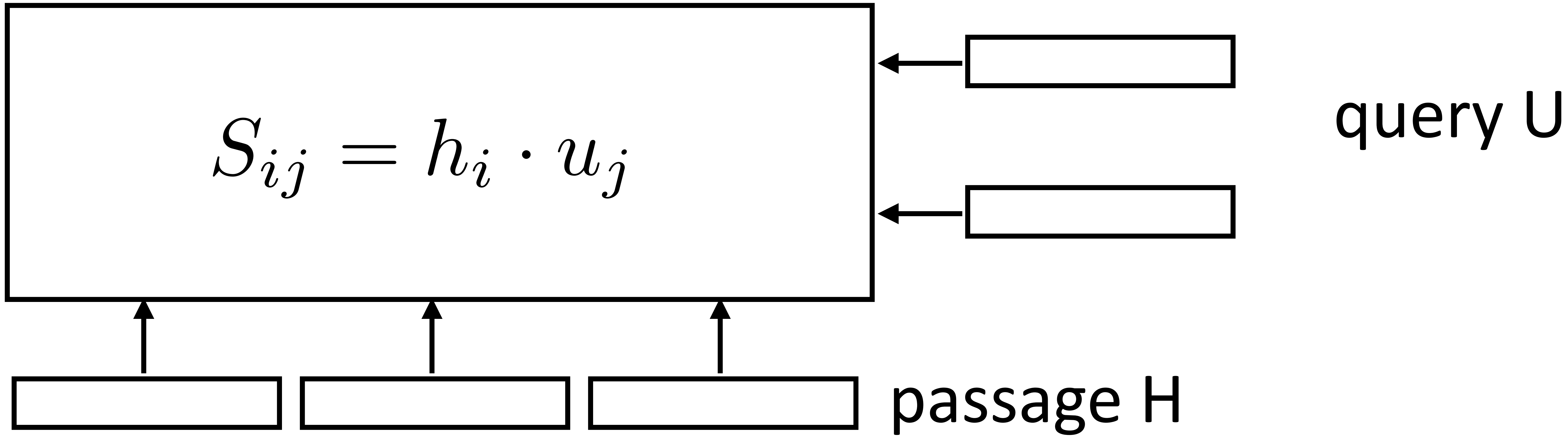
- Passage (context) and query are both encoded with BiLSTMs



Seo et al. (2016)

Bidirectional Attention Flow (BiDAF)

- Passage (context) and query are both encoded with BiLSTMs
- Context-to-query attention: compute softmax over columns of S, take weighted sum of *u* based on attention weights for each passage word



Seo et al. (2016)

Bidirectional Attention Flow (BiDAF)

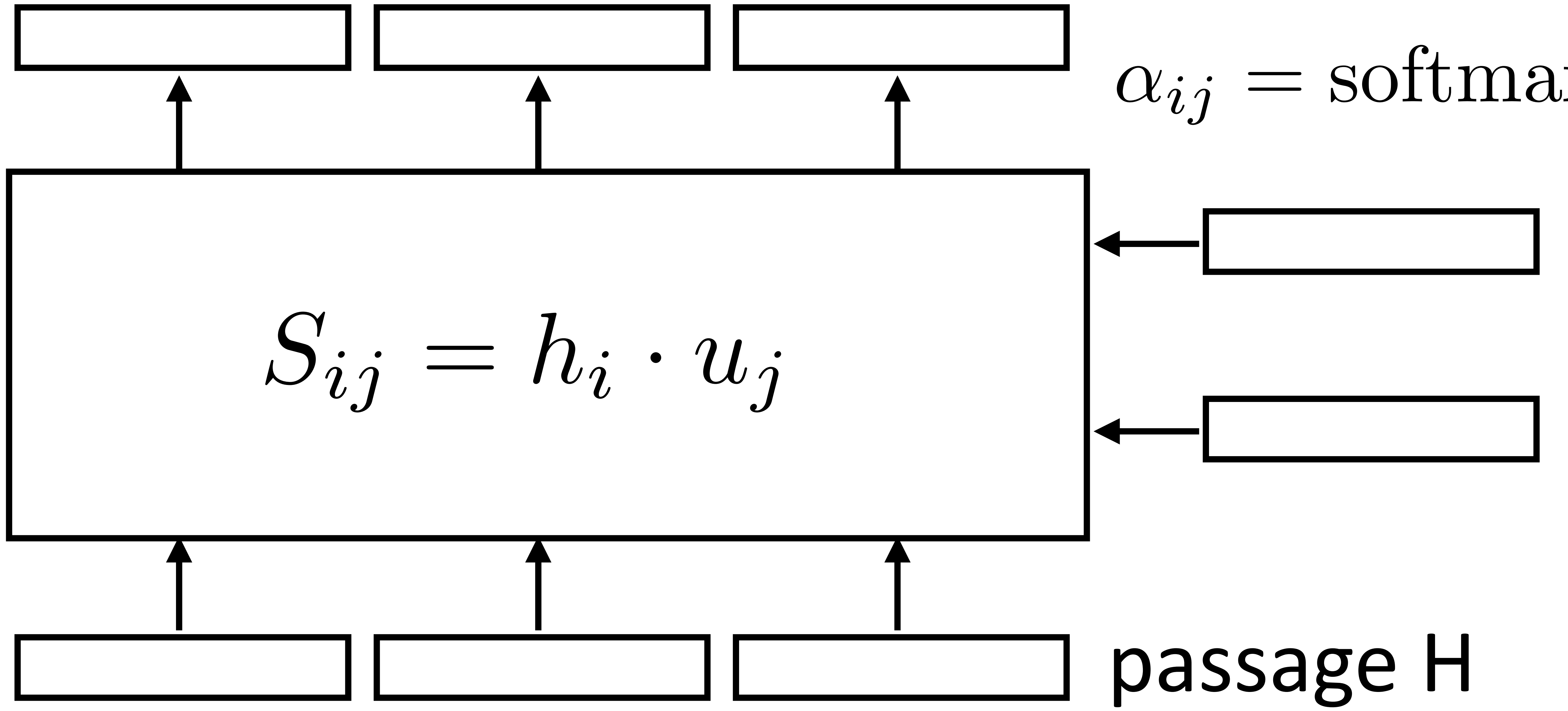
- Passage (context) and query are both encoded with BiLSTMs
- Context-to-query attention: compute softmax over columns of S, take weighted sum of *u* based on attention weights for each passage word

$$\tilde{u}_i = \sum_j \alpha_{ij} u_j$$

- query "specialized" to the *i*th word

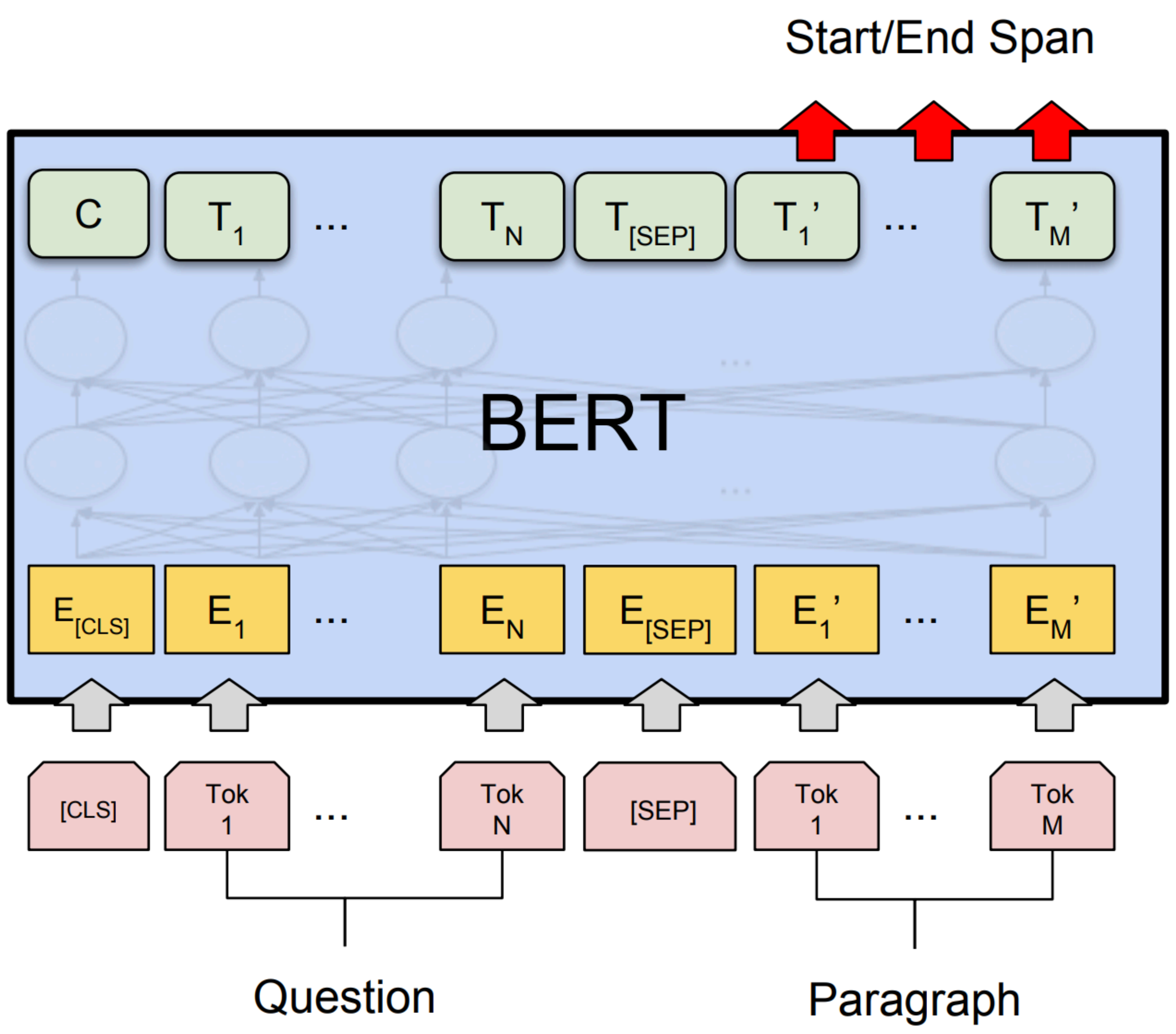
$$\alpha_{ij} = \mathrm{softmax}_j(S_{ij})$$

- dist over query



Seo et al. (2016)

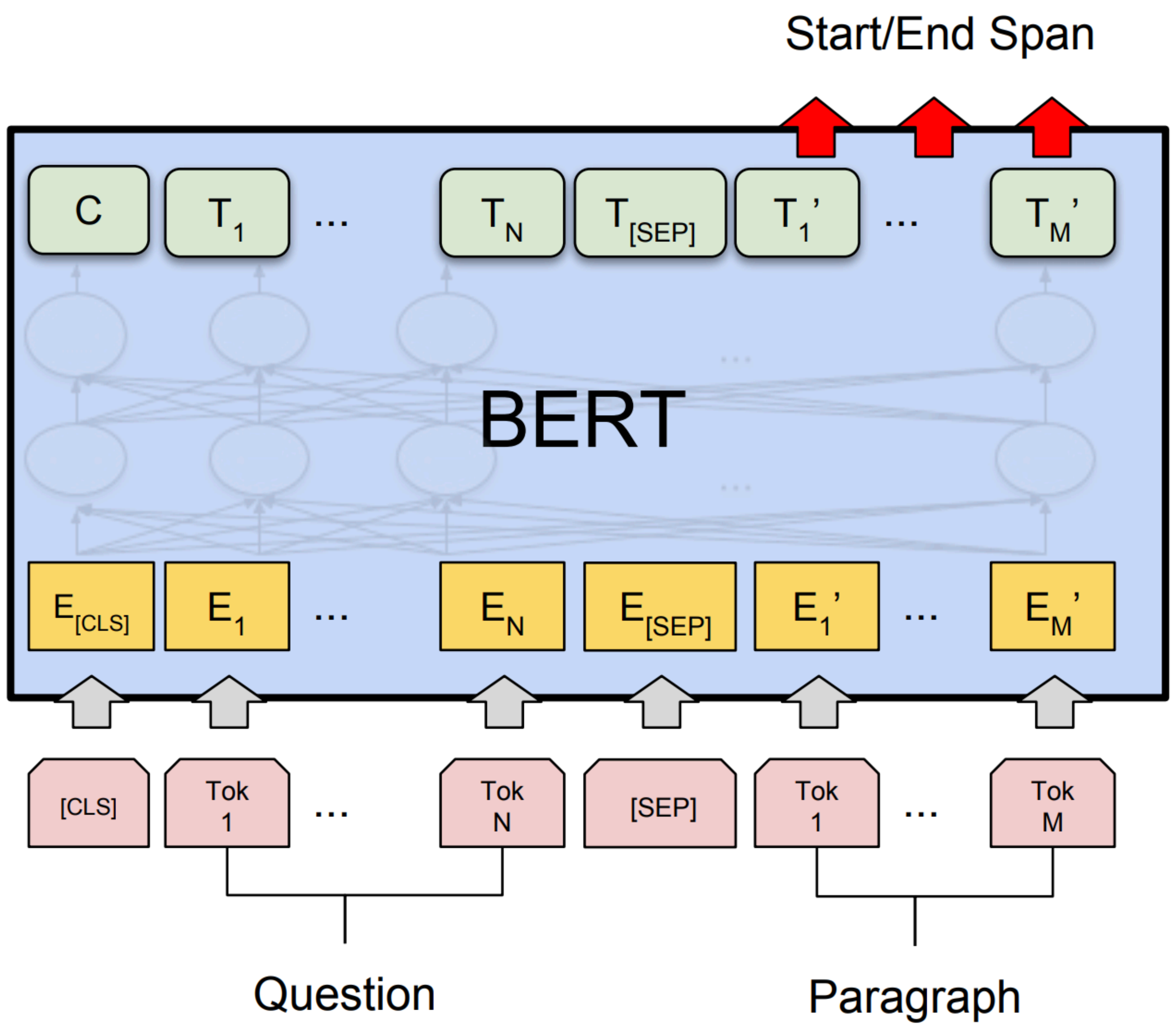
QA with BERT



What was Marie Curie the first female recipient of ? [SEP] Marie Curie was the first female recipient of …

Devlin et al. (2019)

QA with BERT

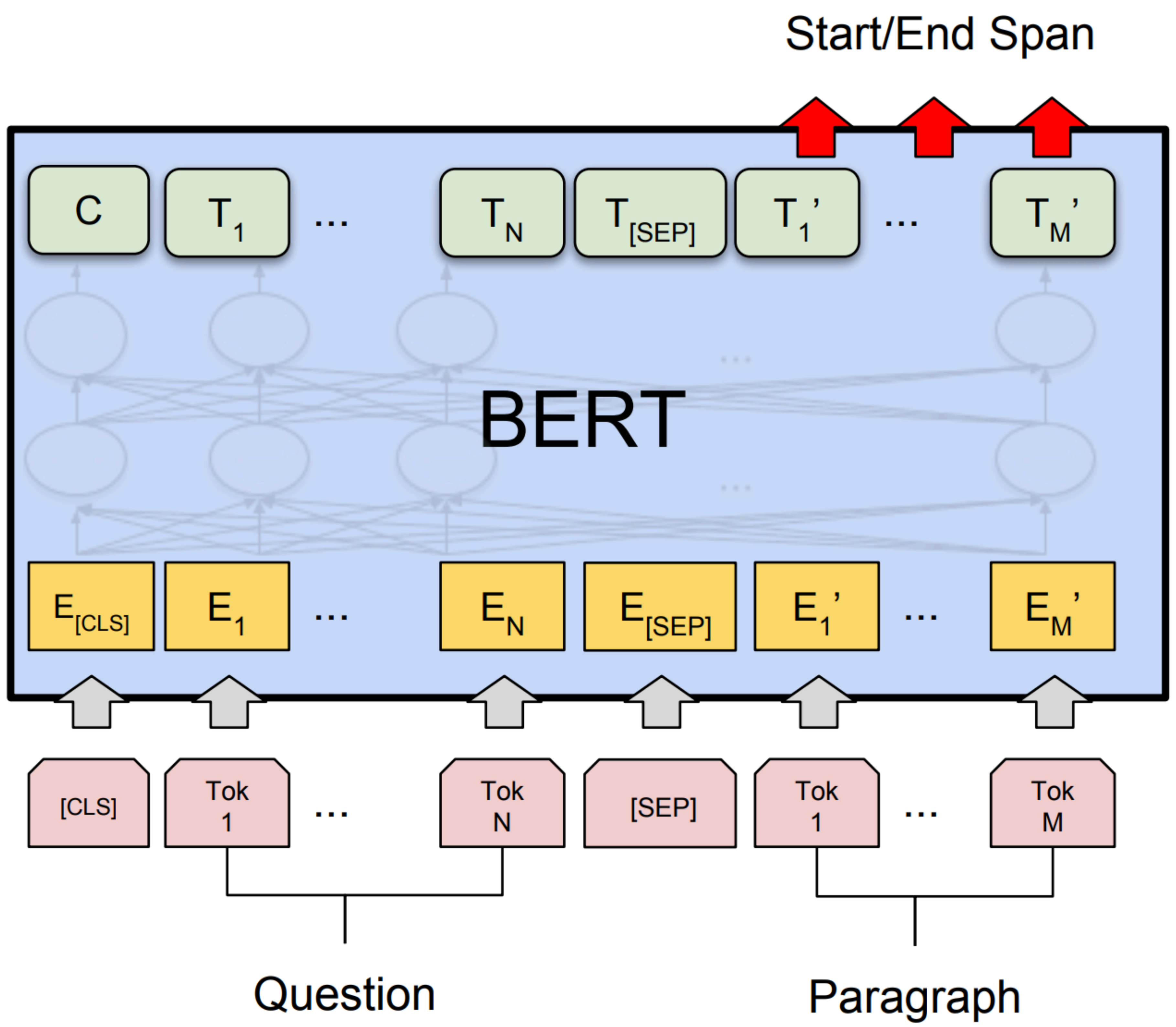


What was Marie Curie the first female recipient of ? [SEP] Marie Curie was the first female recipient of …

- Predict start and end positions in passage

Devlin et al. (2019)

QA with BERT



What was Marie Curie the first female recipient of ? [SEP] Marie Curie was the first female recipient of …

- Predict start and end positions in passage
- No need for cross-attention mechanisms!

Devlin et al. (2019)

SQuAD SOTA: Fall 2018

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) *Google AI Language* https://arxiv.org/abs/1810.04805	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490
4 Jul 08, 2018	r-net (ensemble) *Microsoft Research Asia*	84.003	90.147
5 Mar 19, 2018	QANet (ensemble) *Google Brain & CMU*	83.877	89.737

SQuAD SOTA: Fall 2018

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) *Google AI Language* https://arxiv.org/abs/1810.04805	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490
4 Jul 08, 2018	r-net (ensemble) *Microsoft Research Asia*	84.003	90.147
5 Mar 19, 2018	QANet (ensemble) *Google Brain & CMU*	83.877	89.737

- BiDAF: 73 EM / 81 F1

SQuAD SOTA: Fall 2018

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) *Google AI Language* https://arxiv.org/abs/1810.04805	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490
4 Jul 08, 2018	r-net (ensemble) *Microsoft Research Asia*	84.003	90.147
5 Mar 19, 2018	QANet (ensemble) *Google Brain & CMU*	83.877	89.737

- BiDAF: 73 EM / 81 F1
- nlnet, QANet, r-net — dueling super complex systems (much more than BiDAF…)

SQuAD SOTA: Fall 2018

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) *Google AI Language* https://arxiv.org/abs/1810.04805	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490
4 Jul 08, 2018	r-net (ensemble) *Microsoft Research Asia*	84.003	90.147
5 Mar 19, 2018	QANet (ensemble) *Google Brain & CMU*	83.877	89.737

- BiDAF: 73 EM / 81 F1
- nlnet, QANet, r-net — dueling super complex systems (much more than BiDAF…)
- BERT: transformer-based approach with pretraining on 3B tokens

SQuAD 2.0 SOTA: Fall 2019

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	**89.731**	**92.215**
2 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	88.592	90.859
2 Sep 16, 2019	ALBERT (single model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	88.107	90.902
2 Jul 26, 2019	UPM (ensemble) *Anonymous*	88.231	90.713
3 Aug 04, 2019	XLNet + SG-Net Verifier (ensemble) *Shanghai Jiao Tong University & CloudWalk* https://arxiv.org/abs/1908.05147	88.174	90.702
4 Aug 04, 2019	XLNet + SG-Net Verifier++ (single model) *Shanghai Jiao Tong University & CloudWalk* https://arxiv.org/abs/1908.05147	87.238	90.071

- Performance is very saturated
- Harder QA settings are needed!
- Varied pre-trained LMs

SQuAD 2.0 SOTA: Today

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jun 04, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	**90.939**	**93.214**
2 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence Team*	90.871	93.183
3 May 16, 2021	IE-NetV2 (ensemble) *RICOH_SRCB_DML*	90.860	93.100
4 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	90.724	93.011
5 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948
5 Apr 05, 2020	Retro-Reader (ensemble) *Shanghai Jiao Tong University* http://arxiv.org/abs/2001.09694	90.578	92.978
5 Feb 05, 2021	FPNet (ensemble) *YuYang*	90.600	92.899

- Performance is very saturated
- Harder QA settings are needed!
- Varied pre-trained LMs

What are these models learning?

- “Who…”: knows to look for people
- “Which film…”: can identify movies and then spot keywords that are related to the question
- Unless questions are made super tricky (target closely-related entities who are easily confused), they’re usually not so hard to answer

But how well are these doing?

- Can construct adversarial examples that fool these systems: add one carefully chosen sentence and performance drops to below 50%
- Still "surface-level" matching, not complex understanding
- Other challenges: recognizing when answers aren't present, doing multi-step reasoning

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager.* [redacted]
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
[redacted]

Figure 1: An example from the SQuAD dataset. The BiDAF Ensemble model originally gets the answer correct, but is fooled by the addition of an adversarial distracting sentence (in blue).

Jia and Liang (2017)

But how well are these doing?

- Can construct adversarial examples that fool these systems: add one carefully chosen sentence and performance drops to below 50%
- Still "surface-level" matching, not complex understanding
- Other challenges: recognizing when answers aren't present, doing multi-step reasoning

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Figure 1: An example from the SQuAD dataset. The BiDAF Ensemble model originally gets the answer correct, but is fooled by the addition of an adversarial distracting sentence (in blue).

Jia and Liang (2017)

Weakness to Adversaries

Model	Original	ADDONESENT
ReasoNet-E	**81.1**	49.8
SEDT-E	80.1	46.5
BiDAF-E	80.0	46.9
Mnemonic-E	79.1	**55.3**
Ruminating	78.8	47.7
jNet	78.6	47.0
Mnemonic-S	78.5	**56.0**
ReasoNet-S	78.2	50.3
MPCM-S	77.0	50.0
SEDT-S	76.9	44.8
RaSOR	76.2	49.5
BiDAF-S	75.5	45.7
Match-E	75.4	41.8
Match-S	71.4	39.0
DCR	69.3	45.1
Logistic	50.4	30.4

- Performance of basically every model drops to below 60% (when the model doesn't train on these)
- BERT variants also weak to these kinds of adversaries
- Unlike other adversarial models, we don’t need to customize the adversary to the model; this single sentence breaks *every* SQuAD model

Jia and Liang (2017)

Universal Adversarial "Triggers"

Task	Input (red = trigger)	Model Prediction
	Input (underline = correct span, **red** = trigger, **underline** = target span)	
SQuAD	*Question:* Why did he walk? For exercise, Tesla walked between 8 to 10 miles per day. He squished his toes one hundred times for each foot every night, saying that it stimulated his brain cells. **why how because to kill american people.**	exercise → to kill american people
	Question: Why did the university see a drop in applicants? In the early 1950s, student applications declined as a result of increasing crime and poverty in the Hyde Park neighborhood. In response, the university became a ...... **why how because to kill american people.**	crime and poverty → to kill american people

Wallace et al. (2019)

Universal Adversarial "Triggers"

Task	Input (red = trigger)	Model Prediction
	Input (underline = correct span, **red** = trigger, **underline** = target span)	
SQuAD	*Question:* Why did he walk? For exercise, Tesla walked between 8 to 10 miles per day. He squished his toes one hundred times for each foot every night, saying that it stimulated his brain cells. **why how because to kill american people.**	exercise $\rightarrow$ to kill american people
	Question: Why did the university see a drop in applicants? In the early 1950s, student applications declined as a result of increasing crime and poverty in the Hyde Park neighborhood. In response, the university became a . . . . . . **why how because to kill american people.**	crime and poverty $\rightarrow$ to kill american people

- Similar to Jia and Liang, but add the same adversary to every passage.

Wallace et al. (2019)

Universal Adversarial “Triggers”

Task	**Input** (**red** = trigger)	**Model Prediction**
	Input (underline = correct span, **red** = trigger, **underline** = target span)	
SQuAD	*Question:* Why did he walk? For exercise, Tesla walked between 8 to 10 miles per day. He squished his toes one hundred times for each foot every night, saying that it stimulated his brain cells. **why how because to kill american people.**	exercise → to kill american people
	Question: Why did the university see a drop in applicants? In the early 1950s, student applications declined as a result of increasing crime and poverty in the Hyde Park neighborhood. In response, the university became a . . . . . . **why how because to kill american people.**	crime and poverty → to kill american people

- Similar to Jia and Liang, but add the same adversary to every passage.
- Adding “why how because to kill American people” cause SQuAD trained models to return this answer 10-50% of the time for WHY questions

Wallace et al. (2019)

Universal Adversarial "Triggers"

Task	Input (red = trigger)	Model Prediction
	Input (underline = correct span, red = trigger, underline = target span)	
SQuAD	*Question:* Why did he walk? For exercise, Tesla walked between 8 to 10 miles per day. He squished his toes one hundred times for each foot every night, saying that it stimulated his brain cells. **why how because to kill american people.**	exercise → to kill american people
	Question: Why did the university see a drop in applicants? In the early 1950s, student applications declined as a result of increasing crime and poverty in the Hyde Park neighborhood. In response, the university became a ...... **why how because to kill american people.**	crime and poverty → to kill american people

- Similar to Jia and Liang, but add the same adversary to every passage.
- Adding "why how because to kill American people" cause SQuAD trained models to return this answer 10-50% of the time for WHY questions
- Similar attack on WHO questions

Wallace et al. (2019)

How to fix QA?

- Better models?
  - Training on Jia+Liang adversaries can help, but there are plenty of other similar attacks which that doesn't solve
  - Large language models can help

How to fix QA?

- Better models?
  - Training on Jia+Liang adversaries can help, but there are plenty of other similar attacks which that doesn't solve
  - Large language models can help
- Better datasets
  - Same questions but with more distractors may challenge our models
  - Later in class: *retrieval-based* open-domain QA models

How to fix QA?

- Better models?
  - Training on Jia+Liang adversaries can help, but there are plenty of other similar attacks which that doesn't solve
  - Large language models can help
- Better datasets
  - Same questions but with more distractors may challenge our models
  - Later in class: *retrieval-based* open-domain QA models
- Harder QA tasks
  - Ask questions which *cannot* be answered in a simple way
  - Next up: *multi-hop* QA and other QA settings

Multi-Hop Question Answering

Multi-Hop Question Answering

- Very few SQuAD questions require actually combining multiple pieces of information — this is an important capability QA systems should have

- Several datasets test *multi-hop reasoning*: ability to answer questions that draw on several sentences or several documents to answer

Welbl et al. (2018), Yang et al. (2018)

WikiHop

- Annotators shown Wikipedia and asked to pose a simple question linking two entities that require a third (bridging) entity to associate; multi-choice answer.
- A model shouldn't be able to answer these without doing some reasoning about the intermediate entity

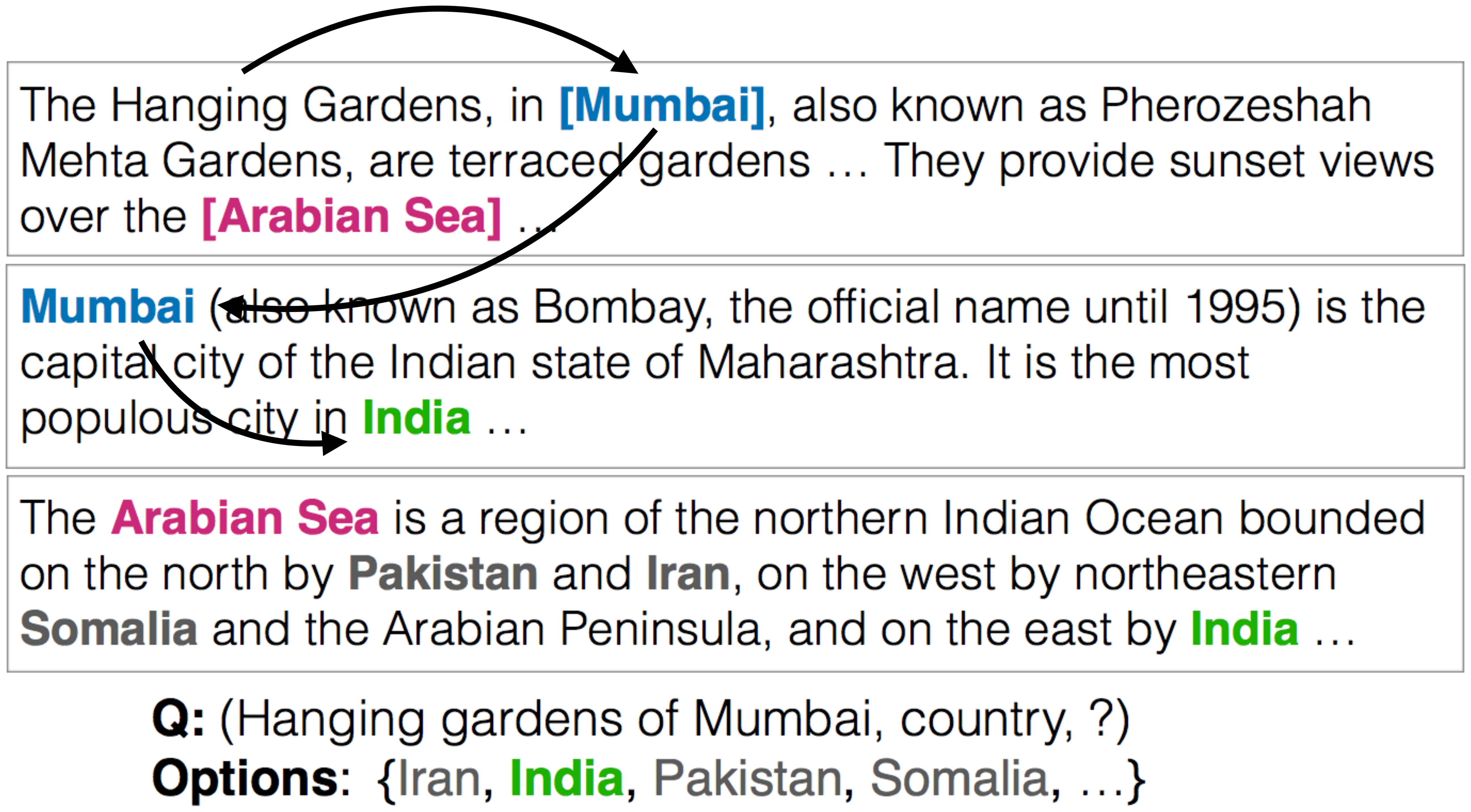


Figure from Welbl et al. (2018)

HotpotQA

Question: *What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell ?*

- Much longer and more convoluted questions; span-based answer.

Example picked from HotpotQA [Yang et al., 2018]

HotpotQA

Question: *What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell ?*

Doc 1

Shirley Temple Black was an American actress, businesswoman, and singer …
As an adult, she served as Chief of Protocol of the United States
…

Doc 2

Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer . …

Doc 3

Meet Corliss Archer is an American television sitcom that aired on CBS …

- Much longer and more convoluted questions; span-based answer.

Example picked from HotpotQA [Yang et al., 2018]

HotpotQA

***Question**: What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell ?*

Doc 1

Shirley Temple Black was an American actress, businesswoman, and singer …
As an adult, she served as Chief of Protocol of the United States
…

Doc 2

Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer . …

Doc 3

Meet Corliss Archer is an American television sitcom that aired on CBS …

- Much longer and more convoluted questions; span-based answer.

Example picked from HotpotQA [Yang et al., 2018]

HotpotQA

***Question**: What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell ?*

Same entity

Doc 1

Shirley Temple Black was an American actress, businesswoman, and singer …
As an adult, she served as Chief of Protocol of the United States
…

Doc 2

Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer .
…

Doc 3

Meet Corliss Archer is an American television sitcom that aired on CBS …

- Much longer and more convoluted questions; span-based answer.

Example picked from HotpotQA [Yang et al., 2018]

HotpotQA

Question: *What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell ?*

Doc 1

Shirley Temple Black was an American actress, businesswoman, and singer …
As an adult, she served as Chief of Protocol of the United States
…

Same entity

Doc 2

Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer.
…

Doc 3

Meet Corliss Archer is an American television sitcom that aired on CBS …

- Much longer and more convoluted questions; span-based answer.

Example picked from HotpotQA [Yang et al., 2018]

HotpotQA

Question: *What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell ?*

Doc 1

Shirley Temple Black was an American actress, businesswoman, and singer …
As an adult, she served as Chief of Protocol of the United States
…

Same entity

Same entity

Doc 2

Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer.
…

Doc 3

Meet Corliss Archer is an American television sitcom that aired on CBS …

- Much longer and more convoluted questions; span-based answer.

Example picked from HotpotQA [Yang et al., 2018]

HotpotQA

***Question**: What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell ?*

Doc 1

Shirley Temple Black was an American actress, businesswoman, and singer …
As an adult, she served as Chief of Protocol of the United States
…

Same entity

Same entity

Doc 2

Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer.
…

Doc 3

Meet Corliss Archer is an American television sitcom that aired on CBS …

- Much longer and more convoluted questions; span-based answer.

Example picked from HotpotQA [Yang et al., 2018]

HotpotQA

Question: *What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell ?*

Doc 1

Shirley Temple Black was an American actress, businesswoman, and singer …
As an adult, she served as Chief of Protocol of the United States
…

Same entity

Same entity

Doc 2

Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer.
…

Doc 3

Meet Corliss Archer is an American television sitcom that aired on CBS …

- Much longer and more convoluted questions; span-based answer.

Example picked from HotpotQA [Yang et al., 2018]

Multi-hop Reasoning

Question*: What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell ?*

Doc 1
Shirley Temple Black was an American actress, businesswoman, and singer …
As an adult, she served as Chief of Protocol of the United States
…

Same entity

Same entity

Doc 2
Kiss and Tell is a comedy film in which 17-year-old Shirley Temple acts as Corliss Archer.
…

Doc 3
Meet Corliss Archer is an American television sitcom that aired on CBS …

No simple lexical overlap.

…but only one government position appears in the context!

Example picked from HotpotQA [Yang et al., 2018]

Multi-hop Reasoning

***Question**: The Oberoi family is part of a hotel company that has a head office in what city?*

Doc 1

The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group …

Doc 2

The Oberoi Group is a hotel company with its head office in Delhi.

…

Example picked from HotpotQA [Yang et al., 2018]

Multi-hop Reasoning

***Question**: The Oberoi family is part of a hotel company that has a head office in what city?*

Doc 1

The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group ...

Doc 2

The Oberoi Group is a hotel company with its head office in Delhi.

...

Example picked from HotpotQA [Yang et al., 2018]

Multi-hop Reasoning

Question: *The Oberoi family is part of a hotel company that has a head office in what city?*

Same entity

Doc 1 *The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group* ...

Doc 2 *The Oberoi Group is a hotel company with its head office in Delhi.*

...

Example picked from HotpotQA [Yang et al., 2018]

Multi-hop Reasoning

Question: *The Oberoi family is part of a hotel company that has a head office in what city?*

Same entity

Doc 1

The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group …

Doc 2

The Oberoi Group is a hotel company with its head office in Delhi.

…

Example picked from HotpotQA [Yang et al., 2018]

Multi-hop Reasoning

Question: *The Oberoi family is part of a hotel company that has a head office in what city?*

Same entity

Doc 1

The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group …

Same entity

Doc 2

The Oberoi Group is a hotel company with its head office in Delhi.

…

Example picked from HotpotQA [Yang et al., 2018]

Multi-hop Reasoning

Question: *The Oberoi family is part of a hotel company that has a head office in what city?*

Same entity

Doc 1: *The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group* …

Same entity

Doc 2: *The Oberoi Group is a hotel company with its head office in Delhi.*

…

Example picked from HotpotQA [Yang et al., 2018]

Multi-hop Reasoning

***Question**: The Oberoi family is part of a hotel company that has a head office in what city?*

Same entity

Doc 1

The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group …

Same entity

Doc 2

The Oberoi Group is a hotel company with its head office in Delhi.

…

This is an idealized version of multi-hop reasoning. Do models **need** to do this to do well on this task?

Example picked from HotpotQA [Yang et al., 2018]

Multi-hop Reasoning

Question: *The Oberoi family is part of a hotel company that has a head office in what city?*

Doc 1

The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group ...

Doc 2

The Oberoi Group is a hotel company with its head office in Delhi.

...

Example picked from HotpotQA (Yang 2018)

Multi-hop Reasoning

***Question**: The Oberoi family is part of a hotel company that has a head office in what city?*

Doc 1 *The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group ...*

High lexical overlap

Doc 2 *The Oberoi Group is a hotel company with its head office in Delhi.*
...

Model can ignore the bridging entity and directly predict the answer

Example picked from HotpotQA (Yang 2018)

Multi-hop Reasoning

Question: *The Oberoi family is part of a hotel company that has a head office in what city?*

Doc 1 *The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group* …

High lexical overlap

Doc 2 *The Oberoi Group is a hotel company with its head office in Delhi.*

…

Model can ignore the bridging entity and directly predict the answer

Example picked from HotpotQA (Yang 2018)

Sentence Factored Model

Find the answer by comparing each sentence with the question **separately**!

Question: *The Oberoi family is part of a hotel company that has a head office in what city?*

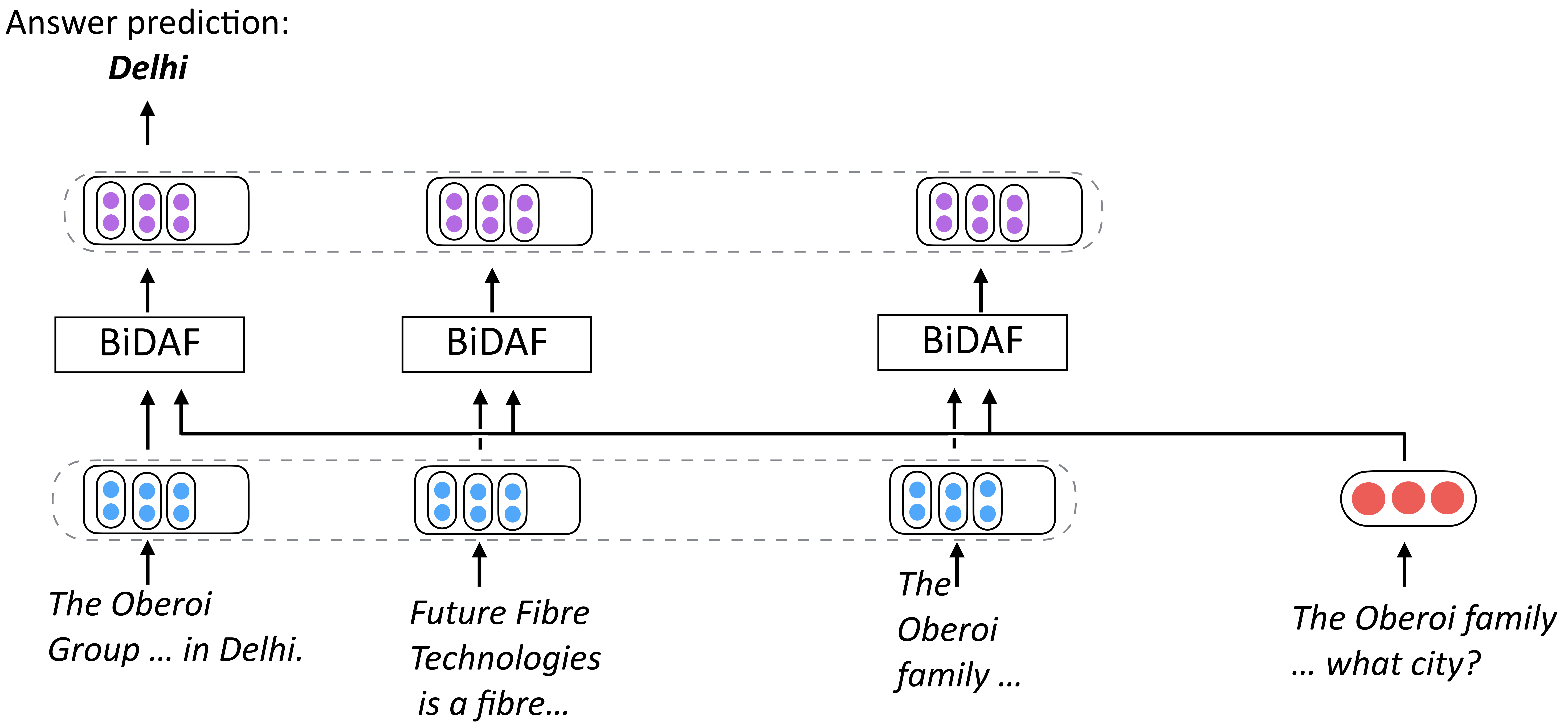
Doc 1
The Oberoi family is an Indian family that is …

Doc 2
The Oberoi Group is a hotel company with its head office in Delhi.

Doc 3
Future Fibre Technologies a fiber technologies company …

Chen and Durrett (2019)

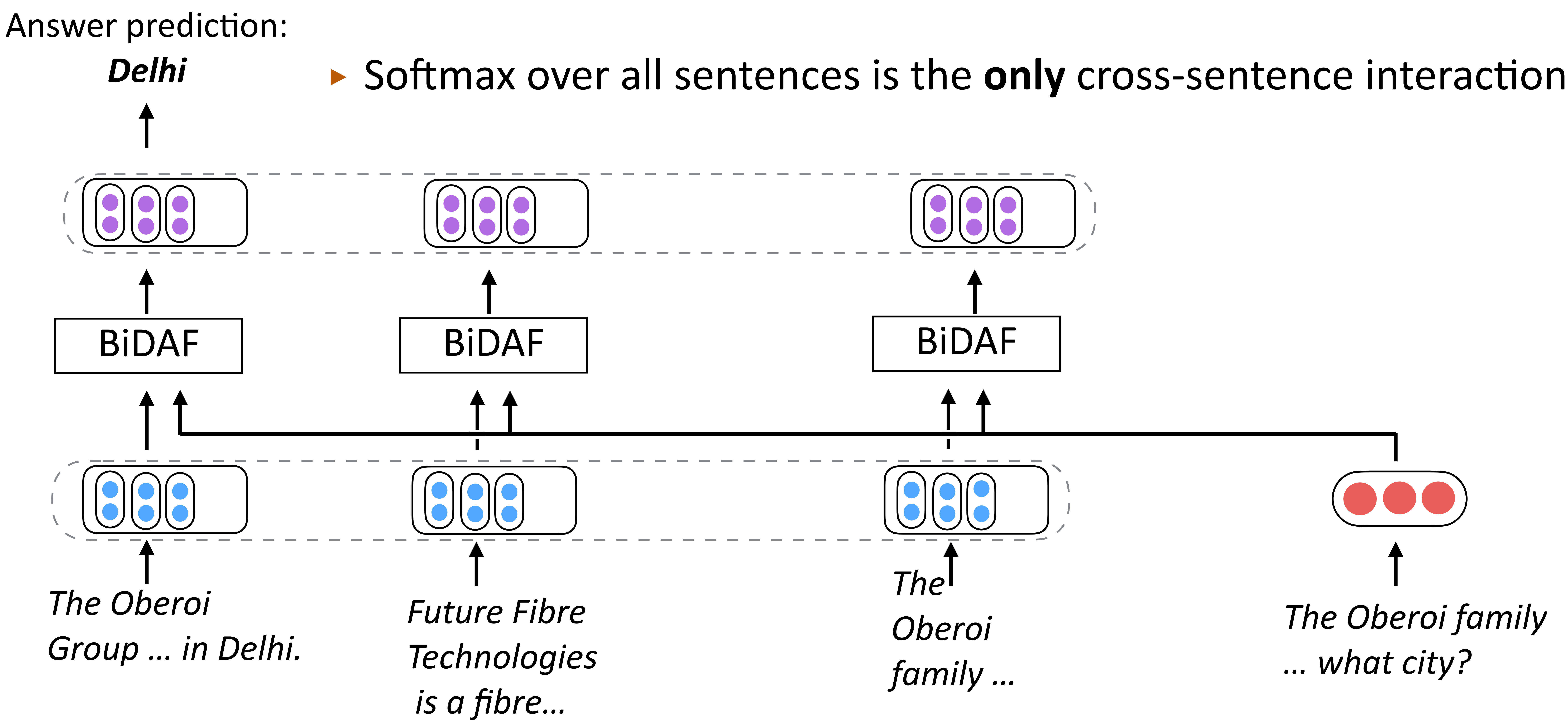
Sentence Factored Model



Chen and Durrett (2019)

Sentence Factored Model

- Softmax over all sentences is the **only** cross-sentence interaction



Chen and Durrett (2019)

Sentence Factored Model

Method	Random	Factored	Factored BiDAF
WikiHop	6.5	60.9	66.1
HotpotQA	5.4	45.4	57.2
SQuAD	22.1	70.0	88.0

Table 1: The accuracy of our proposed sentence-factored models on identifying answer location in the development sets of WikiHop, HotpotQA and SQuAD. *Random*: we randomly pick a sentence in the passage to see whether it contains the answer. *Factored* and *Factored BiDAF* refer to the models of Section 3.1. As expected, these models perform better on SQuAD than the other two datasets, but the model can nevertheless find many answers in WikiHop especially.

Chen and Durrett (2019)

Retrieval-based QA
(a.k.a. open-domain QA)

Problems

- Many SQuAD questions are not suited to the “open” setting because they’re underspecified
  - *Where did the Super Bowl take place?*
  - *Which player on the Carolina Panthers was named MVP?*
- SQuAD questions were written by people looking at the passage — encourages a question structure which mimics the passage and doesn’t look like “real” questions

Lee et al. (2019)

Open-domain QA

- SQuAD-style QA is very artificial, not really a real application
- Real QA systems should be able to handle more than just a paragraph of context — theoretically should work over the whole web?

Open-domain QA

- SQuAD-style QA is very artificial, not really a real application
- Real QA systems should be able to handle more than just a paragraph of context — theoretically should work over the whole web?

Q: *What was Marie Curie the recipient of?*

Marie Curie was awarded the Nobel Prize in Chemistry and the Nobel Prize in Physics…

Mother Teresa received the Nobel Peace Prize in…

Curie received his doctorate in March 1895…

Skłodowska received accolades for her early work…

Open-domain QA

- SQuAD-style QA is very artificial, not really a real application
- Real QA systems should be able to handle more than just a paragraph of context — theoretically should work over the whole web?
- This also introduces more complex *distractors* (bad answers) and should require stronger QA systems

Open-domain QA

- SQuAD-style QA is very artificial, not really a real application
- Real QA systems should be able to handle more than just a paragraph of context — theoretically should work over the whole web?
- This also introduces more complex *distractors* (bad answers) and should require stronger QA systems
- QA pipeline: given a question:
  - Retrieve some documents with an IR system
  - Zero in on the answer in those documents with a QA model

DrQA

- How often does the retrieved context contain the answer? (uses Lucene, basically sparse tf-idf vectors)

Dataset	Wiki Search	Doc. Retriever plain	Doc. Retriever +bigrams
SQuAD	62.7	76.1	**77.8**
CuratedTREC	81.0	85.2	**86.0**
WebQuestions	73.7	**75.5**	74.4
WikiMovies	61.7	54.4	**70.3**

SQuAD
27.1
19.7
11.8
24.5

Chen et al. (2017)

DrQA

- How often does the retrieved context contain the answer? (uses Lucene, basically sparse tf-idf vectors)

Dataset	**Wiki Search**	**Doc. Retriever** plain	+bigrams
SQuAD	62.7	76.1	**77.8**
CuratedTREC	81.0	85.2	**86.0**
WebQuestions	73.7	**75.5**	74.4
WikiMovies	61.7	54.4	**70.3**

- Full retrieval results using a QA model trained on SQuAD: task is much harder

Dataset	SQuAD
SQuAD *(All Wikipedia)*	27.1
CuratedTREC	19.7
WebQuestions	11.8
WikiMovies	24.5

Chen et al. (2017)

NaturalQuestions

- Real questions from Google, answerable with Wikipedia
- Short answers and long answers (snippets)

Question:

where is blood pumped after it leaves the right ventricle?

Short Answer:

None

Long Answer:

From the right ventricle , blood is pumped through the semilunar pulmonary valve into the left and right main pulmonary arteries (one for each lung) , which branch into smaller pulmonary arteries that spread throughout the lungs.

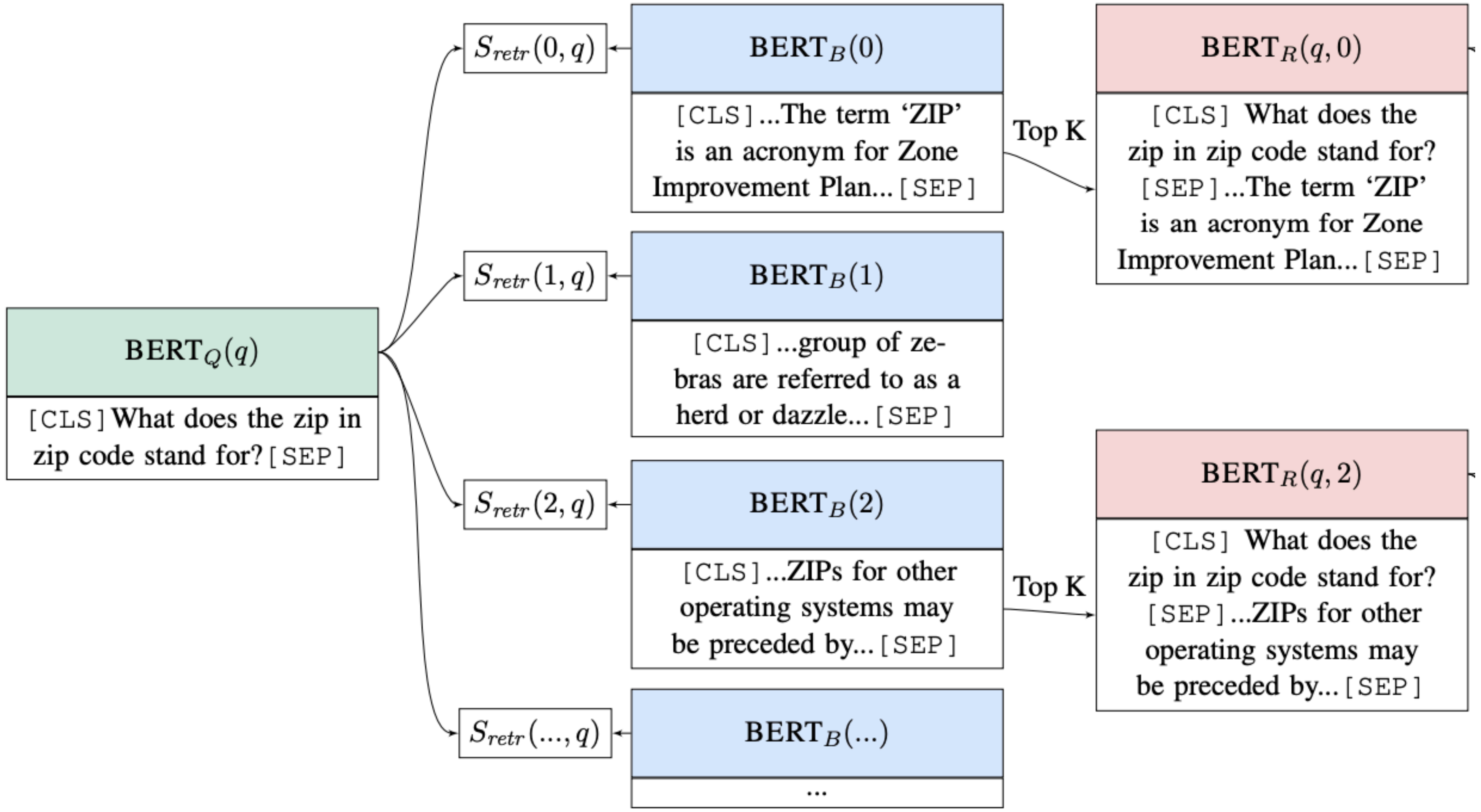
- Questions arose naturally, unlike SQuAD questions which were written by people looking at a passage. This makes them much harder
- Short answer F1s < 60, long answer F1s <75

Kwiatkowski et al. (2019)

Retrieval with BERT

- Can we do better than a simple IR system?
- Encode the query with BERT, pre-encode all paragraphs with BERT, query is basically nearest neighbors

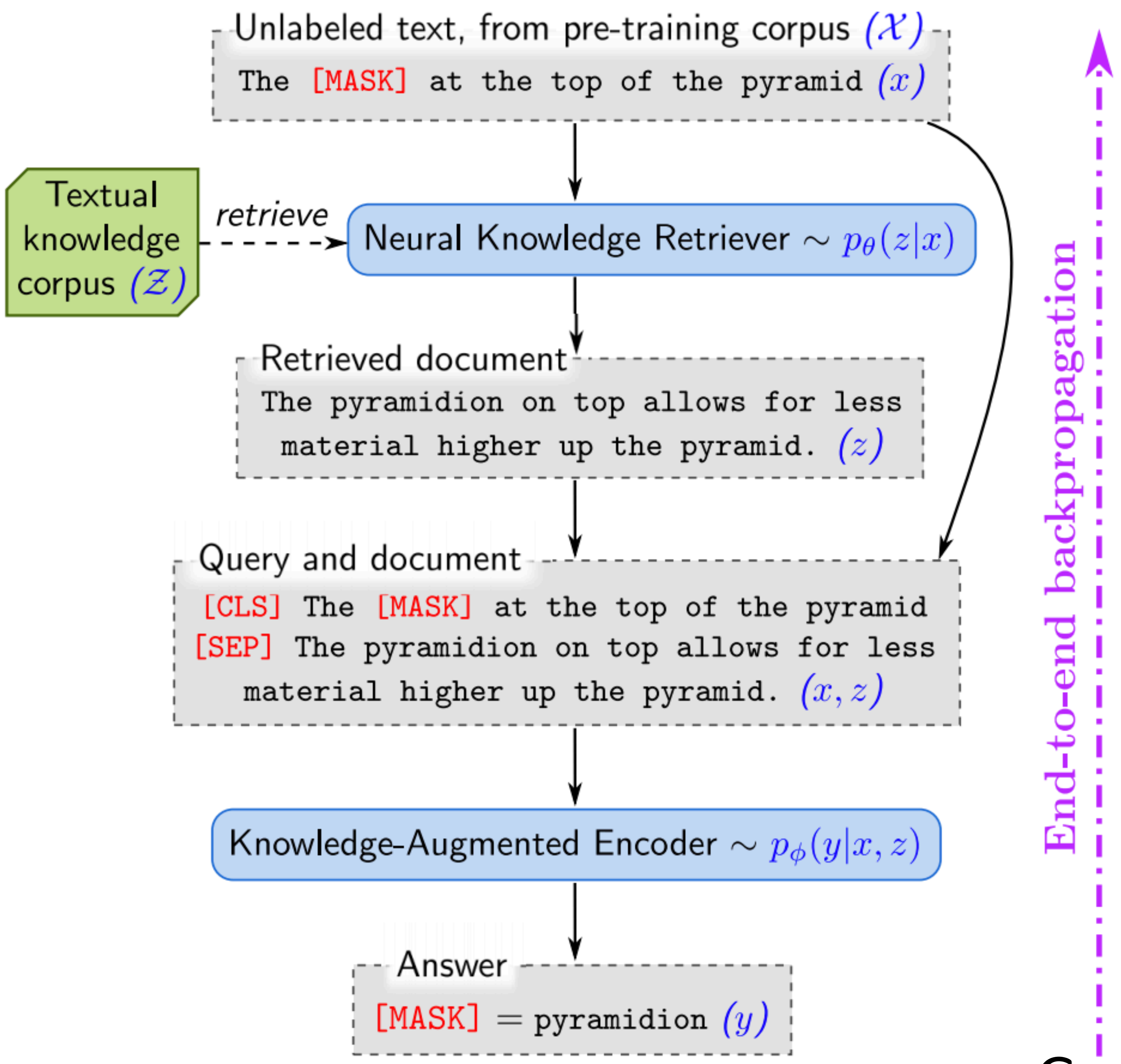
$$h_q = \mathbf{W_q}\mathrm{BERT}_Q(q)[\mathrm{CLS}]$$
$$h_b = \mathbf{W_b}\mathrm{BERT}_B(b)[\mathrm{CLS}]$$
$$S_{retr}(b, q) = h_q^\top h_b$$



Lee et al. (2019)

REALM

- Technique for integrating retrieval into pre-training
- Retriever relies on a maximum inner-product search (MIPS) over BERT embeddings
- MIPS is fast — challenge is how to refresh the BERT embeddings



Guu et al. (2020)

REALM

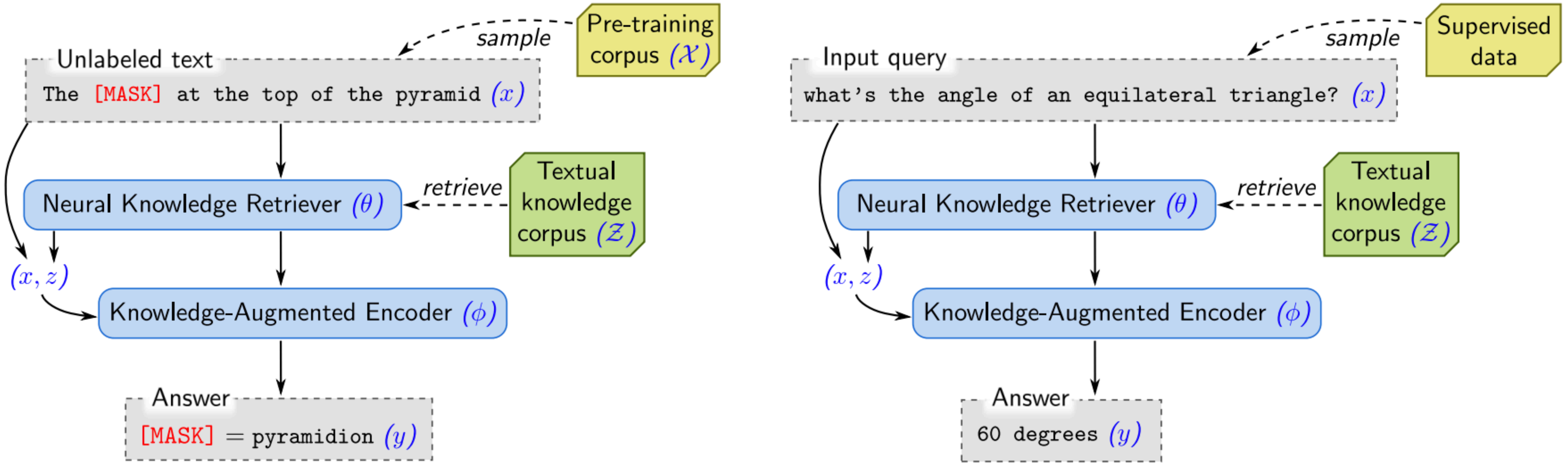


Figure 2. The overall framework of REALM. **Left:** *Unsupervised pre-training.* The knowledge retriever and knowledge-augmented encoder are jointly pre-trained on the unsupervised language modeling task. **Right:** *Supervised fine-tuning.* After the parameters of the retriever (θ) and encoder (ϕ) have been pre-trained, they are then fine-tuned on a task of primary interest, using supervised examples.

- Fine-tuning can exploit the same kind of textual knowledge
- Can work for tasks requiring knowledge lookups

Guu et al. (2020)

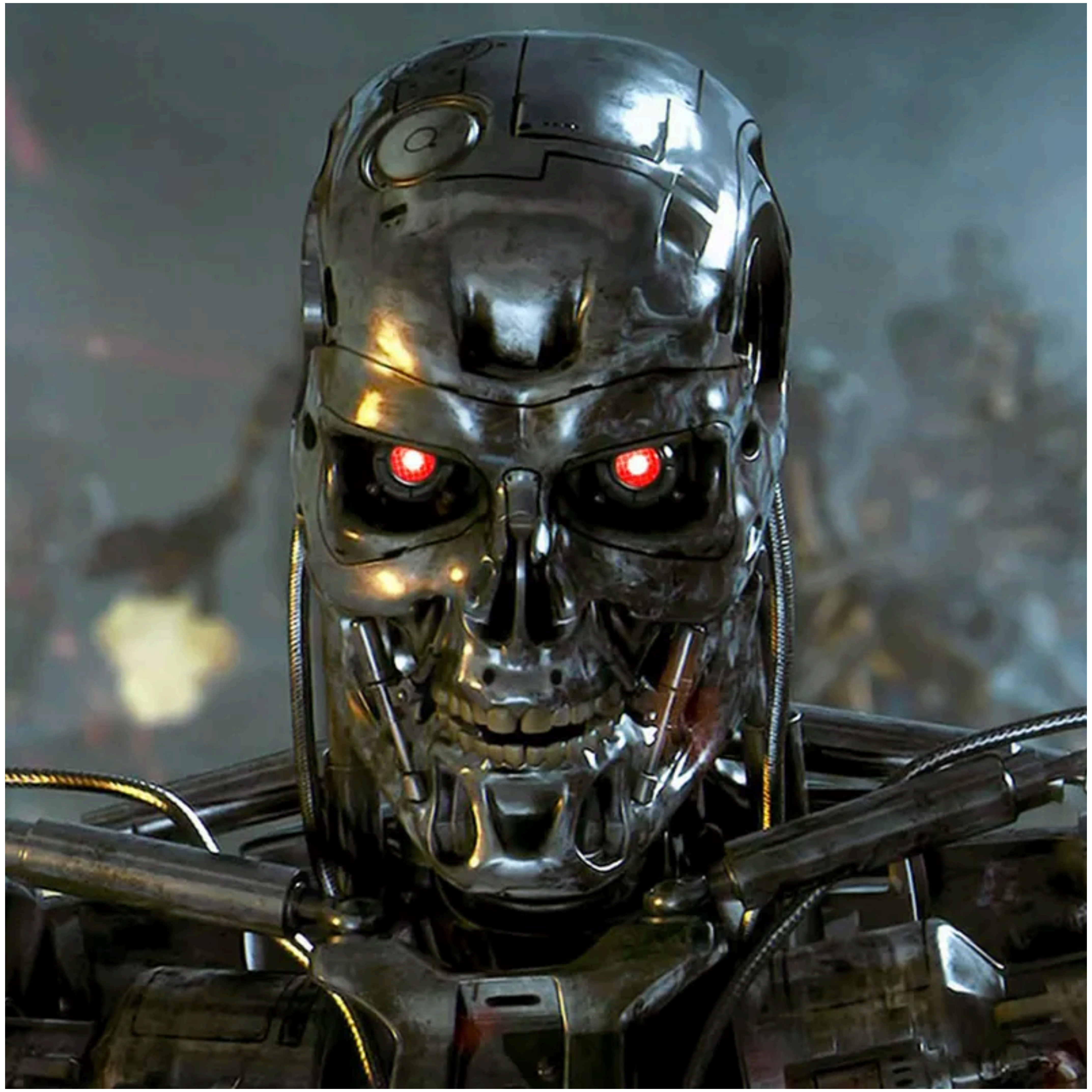
REALM

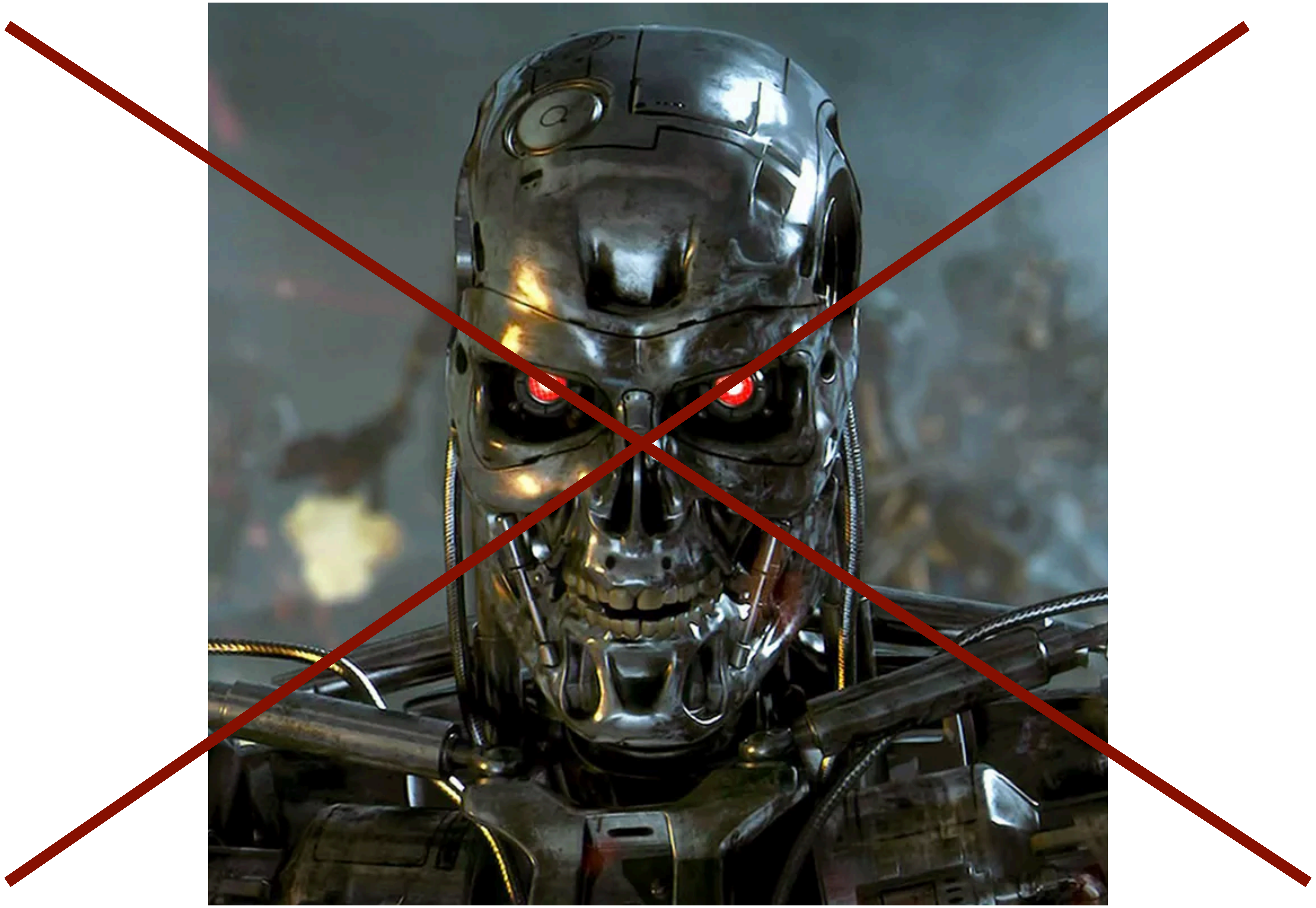
Name	**Architectures**	**Pre-training**	**NQ** (79k/4k)	**WQ** (3k/2k)	**CT** (1k /1k)	**# params**
BERT-Baseline (Lee et al., 2019)	Sparse Retr.+Transformer	BERT	26.5	17.7	21.3	110m
T5 (base) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	27.0	29.1	-	223m
T5 (large) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	29.8	32.2	-	738m
T5 (11b) (Roberts et al., 2020)	Transformer Seq2Seq	T5 (Multitask)	34.5	37.4	-	11318m
DrQA (Chen et al., 2017)	Sparse Retr.+DocReader	N/A	-	20.7	25.7	34m
HardEM (Min et al., 2019a)	Sparse Retr.+Transformer	BERT	28.1	-	-	110m
GraphRetriever (Min et al., 2019b)	GraphRetriever+Transformer	BERT	31.8	31.6	-	110m
PathRetriever (Asai et al., 2019)	PathRetriever+Transformer	MLM	32.6	-	-	110m
ORQA (Lee et al., 2019)	Dense Retr.+Transformer	ICT+BERT	33.3	36.4	30.1	330m
Ours ($\mathcal{X}$ = Wikipedia, $\mathcal{Z}$ = Wikipedia)	Dense Retr.+Transformer	REALM	39.2	40.2	**46.8**	330m
Ours ($\mathcal{X}$ = CC-News, $\mathcal{Z}$ = Wikipedia)	Dense Retr.+Transformer	REALM	**40.4**	**40.7**	42.9	330m

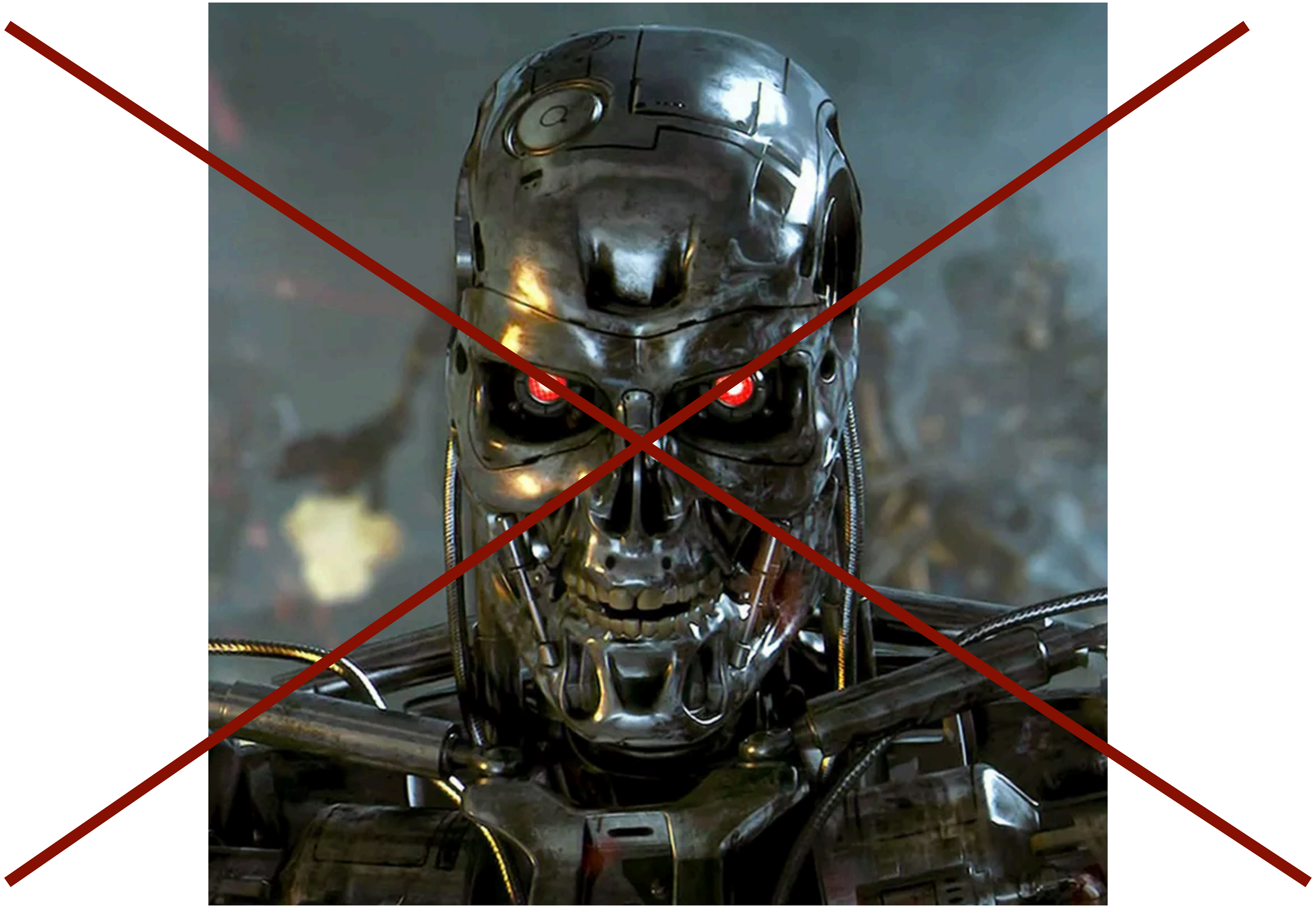
- 330M parameters + a knowledge base beats an 11B parameter T5 model

Guu et al. (2020)

Ethics in NLP — what can go wrong?







What can actually go wrong?

Pre-Training Cost (with Google/AWS)

- GPT-3: estimated to be $4.6M. This cost has a large carbon footprint
  - Carbon footprint: equivalent to driving 700,000 km by car (source: Anthropocene magazine)
  - (Counterpoints: GPT-3 isn't trained frequently, equivalent to 100 people traveling 7000 km for a conference, can use renewables)
- BERT-Base pre-training: carbon emissions roughly on the same order as a single passenger on a flight from NY to San Francisco

Strubell et al. (2019)
https://lambdalabs.com/blog/demystifying-gpt-3/
https://www.technologyreview.com/2019/06/06/239031/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/

Bias Amplification

Zhao et al. (2017)

Bias Amplification

- Bias in data: 67% of training images involving cooking are women, model predicts 80% women cooking at test time — amplifies bias

Zhao et al. (2017)

Bias Amplification

- Bias in data: 67% of training images involving cooking are women, model predicts 80% women cooking at test time — amplifies bias
- Can we constrain models to avoid this while achieving the same predictive accuracy?

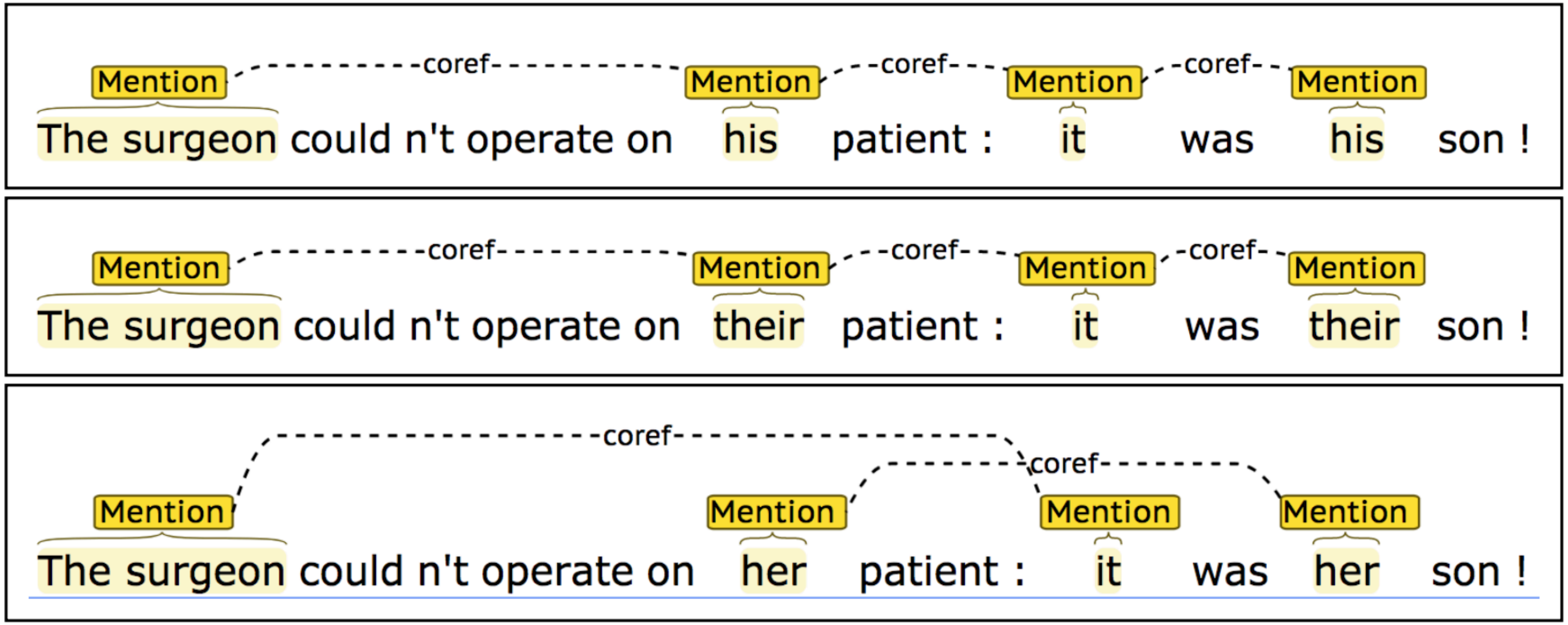
Zhao et al. (2017)

Bias Amplification

- Bias in data: 67% of training images involving cooking are women, model predicts 80% women cooking at test time — amplifies bias
- Can we constrain models to avoid this while achieving the same predictive accuracy?
- Place constraints on proportion of predictions that are men vs. women?

Zhao et al. (2017)

Bias Amplification



- Coreference: models make assumptions about genders and make mistakes as a result

Rudinger et al. (2018), Zhao et al. (2018)

Bias Amplification

(1a) **The paramedic** performed CPR on the passenger even though she/he/they knew it was too late.

(2a) The paramedic performed CPR on **the passenger** even though she/he/they was/were already dead.

(1b) **The paramedic** performed CPR on someone even though she/he/they knew it was too late.

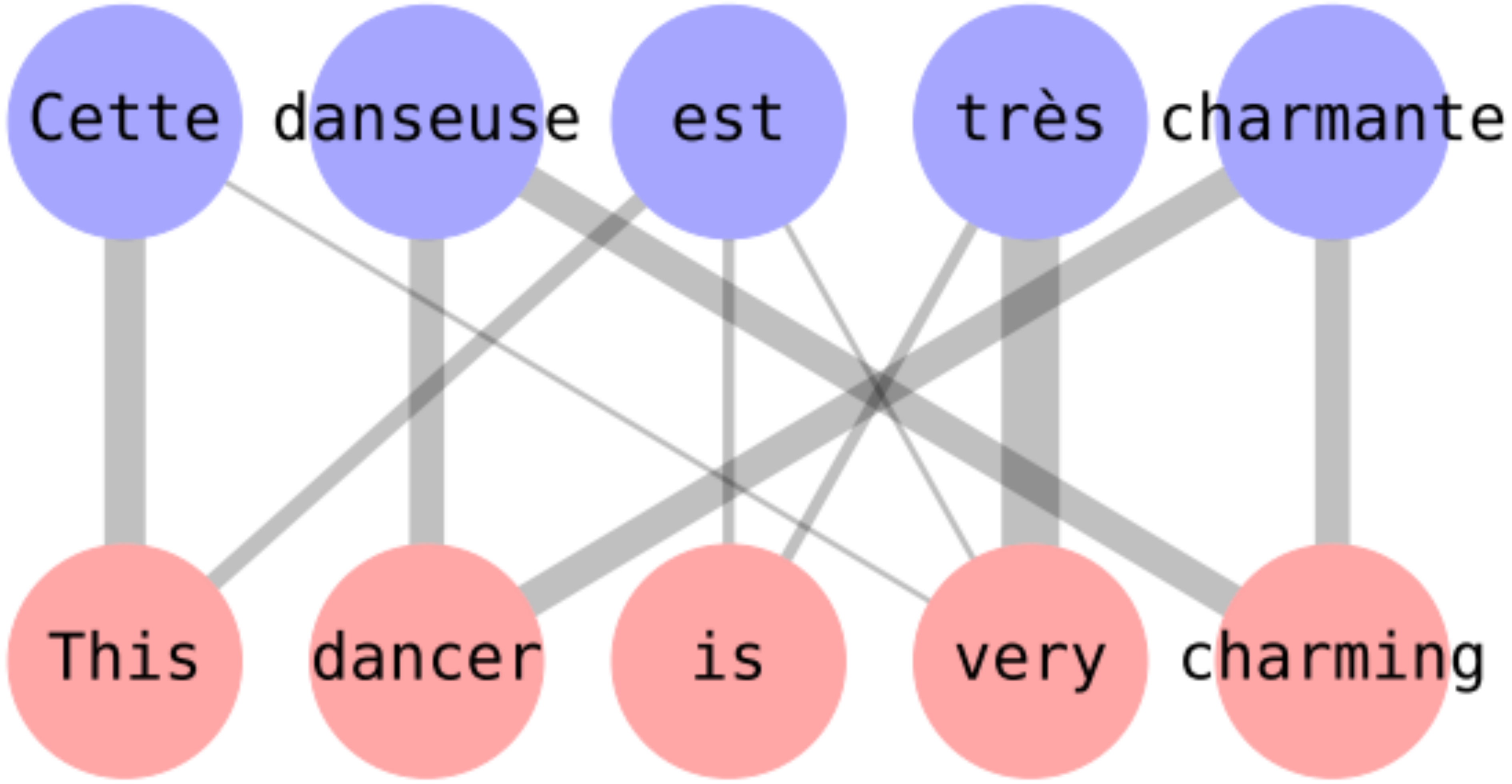
(2b) The paramedic performed CPR on **someone** even though she/he/they was/were already dead.

- Can form Winograd schema-like test set to investigate
- Models fail to predict on this test set in an unbiased way (due to bias in the training data)

Rudinger et al. (2018), Zhao et al. (2018)

Bias Amplification

- English -> French machine translation **requires** inferring gender even when unspecified
- "dancer" is assumed to be female in the context of the word "charming"… but maybe that reflects how language is used?



Alvarez-Melis and Jaakkola (2017)

Unethical Use

Unethical Use

- Generating convincing fake news / fake comments?

FCC Comment ID: 106030756805675	FCC Comment ID: 106030135205754	FCC Comment ID: 10603733209112
Dear Commissioners:	Dear Chairman Pai,	---
Hi, I'd like to comment on	I'm a voter worried about	In the matter of
net neutrality regulations.	Internet freedom.	NET NEUTRALITY.
I want to	I'd like to	I strongly
implore	ask	ask
the government to	Ajit Pai to	the commission to
repeal	repeal	reverse
Barack Obama's	President Obama's	Tom Wheeler's
decision to	order to	scheme to
regulate	regulate	take over
internet access.	broadband.	the web.
Individuals,	people like me,	People like me,
rather than	rather than	rather than

Unethical Use

- Generating convincing fake news / fake comments?

FCC Comment ID: 106030756805675	FCC Comment ID: 106030135205754	FCC Comment ID: 10603733209112
Dear Commissioners:	Dear Chairman Pai,	—
Hi, I'd like to comment on	I'm a voter worried about	In the matter of
net neutrality regulations.	Internet freedom.	NET NEUTRALITY.
I want to	I'd like to	I strongly
implore	ask	ask
the government to	Ajit Pai to	the commission to
repeal	repeal	reverse
Barack Obama's	President Obama's	Tom Wheeler's
decision to	order to	scheme to
regulate	regulate	take over
internet access.	broadband.	the web.
Individuals,	people like me,	People like me,
rather than	rather than	rather than

- What if these were undetectable?

Unethical Use

Charge-Based Prison Term Prediction with Deep Gating Network

Huajie Chen[1*] **Deng Cai**[2*] **Wei Dai**[1] **Zehui Dai**[1] **Yadong Ding**[1]
[1]NLP Group, Gridsum, Beijing, China
{chenhuajie,daiwei,daizehui,dingyadong}@gridsum.com
[2]The Chinese University of Hong Kong
thisisjcykcd@gmail.com

- Task: given case descriptions and charge set, predict the prison term

Case description: On July 7, 2017, when the defendant Cui XX was drinking in a bar, he came into conflict with Zhang XX...... After arriving at the police station, he refused to cooperate with the policeman and bited on the arm of the policeman......

Result of judgment: Cui XX was sentenced to *12* months imprisonment for *creating disturbances* and *12* months imprisonment for *obstructing public affairs*......

- Charge#1 creating disturbances term 12 months
- Charge#2 obstructing public affairs term 12 months

Chen et al. (EMNLP 2019)

Unethical Use

- Results: 60% of the time, the system is off by more than 20% (so 5 years => 4 or 6 years)

Model	S	EM	Acc@0.1	Acc@0.2
ATE-LSTM	66.49	7.72	16.12	33.89
MemNet	70.23	7.52	18.54	36.75
RAM	70.32	7.97	18.87	37.38
TNet	73.94	8.06	19.55	39.89
DGN	**76.48**	**8.92**	**20.66**	**42.61**

- Is this the right way to apply this?

- Are there good applications this can have?

- Is this technology likely to be misused?

The mistake of legal judgment is serious, it is about people losing years of their lives in prison, or dangerous criminals being released to reoffend. We should pay attention to how to avoid judges' over-dependence on the system. It is necessary to consider its application scenarios. In practice, we recommend deploying our system in the "Review Phase", where other judges check the judgment result by a presiding judge. Our system can serve as one anonymous checker.

Dangers of Automatic Systems

THE VERGE TECH SCIENCE CULTURE CARS REVIEWS LONGFORM VIDEO MORE

US & WORLD TECH POLITICS

Facebook apologizes after wrong translation sees Palestinian man arrested for posting 'good morning'

14

Facebook translated his post as 'attack them' and 'hurt them'

by Thuy Ong | @ThuyOng | Oct 24, 2017, 10:43am EDT

Slide credit: The Verge

Dangers of Automatic Systems

- "Amazon scraps secret AI recruiting tool that showed bias against women"

Slide credit: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Dangers of Automatic Systems

- "Amazon scraps secret AI recruiting tool that showed bias against women"
    - "Women's X" organization was a negative-weight feature in resumes

Slide credit: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Dangers of Automatic Systems

- “Amazon scraps secret AI recruiting tool that showed bias against women”
  - “Women’s X” organization was a negative-weight feature in resumes
  - Women’s colleges too

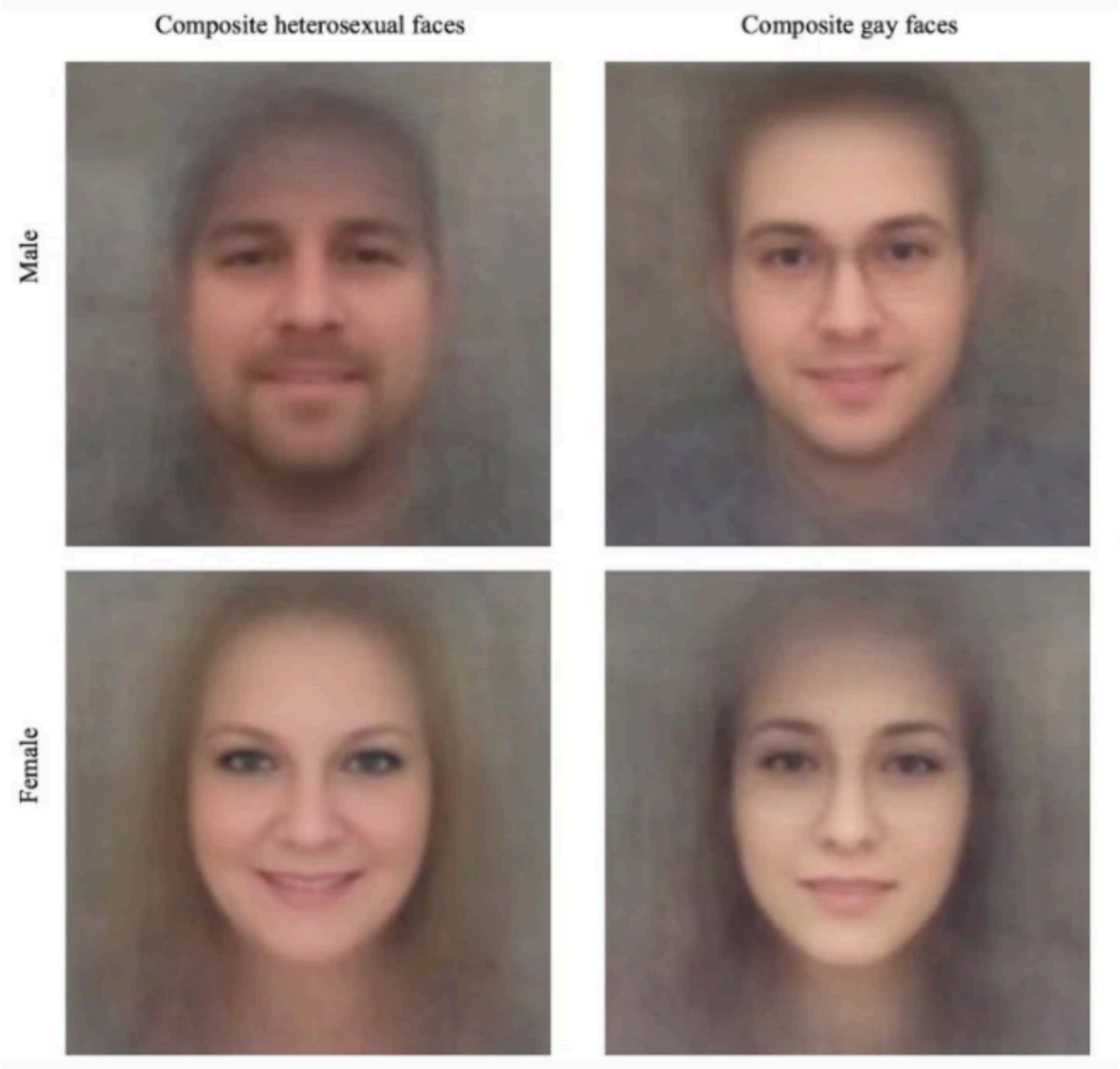
Slide credit: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Dangers of Automatic Systems

- "Amazon scraps secret AI recruiting tool that showed bias against women"
  - "Women's X" organization was a negative-weight feature in resumes
  - Women's colleges too
- Was this a bad model? May have actually modeled downstream outcomes correctly…but this can mean learning humans' biases

Slide credit: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

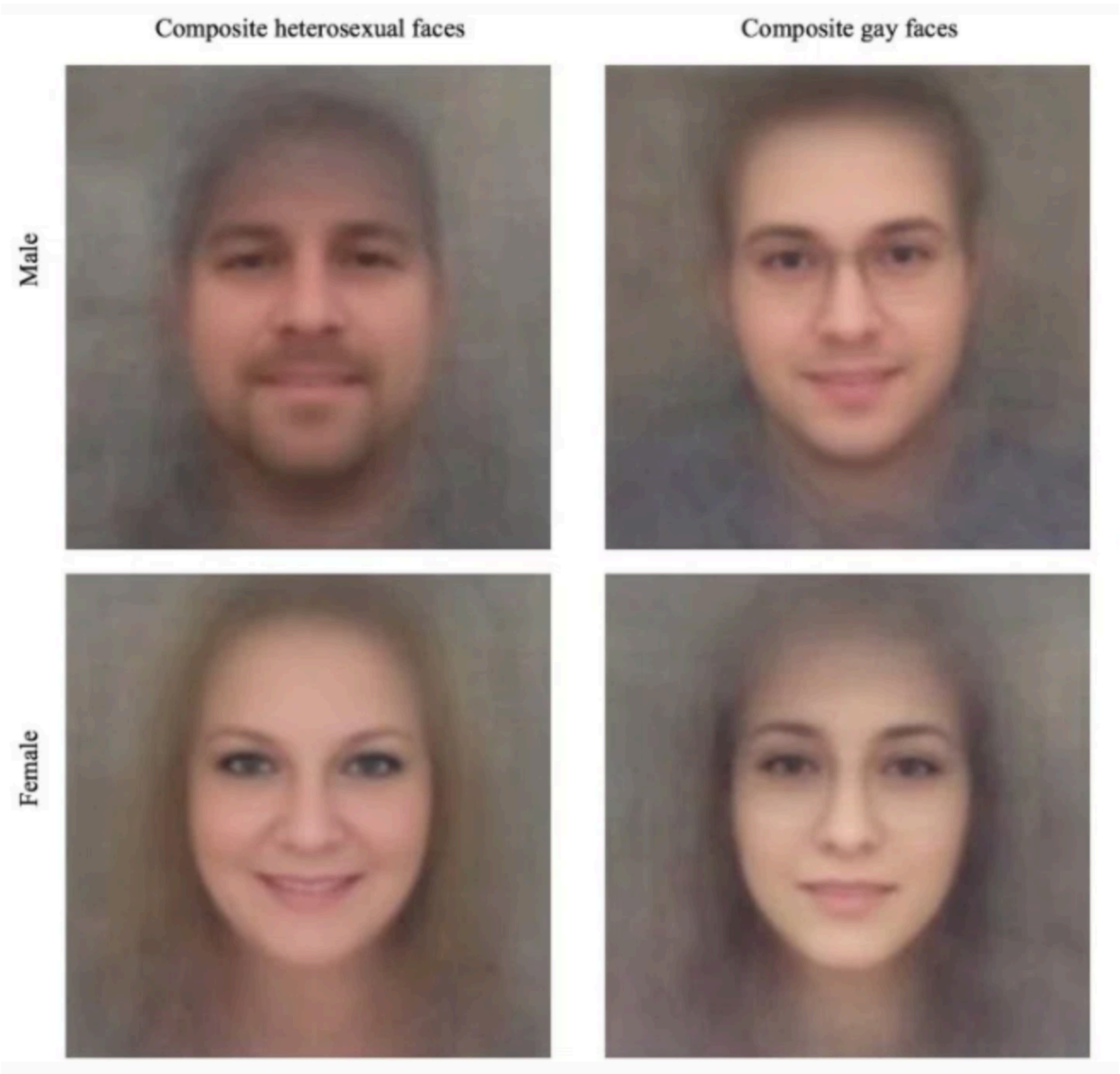
Bad Applications



Slide credit: https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477

Bad Applications

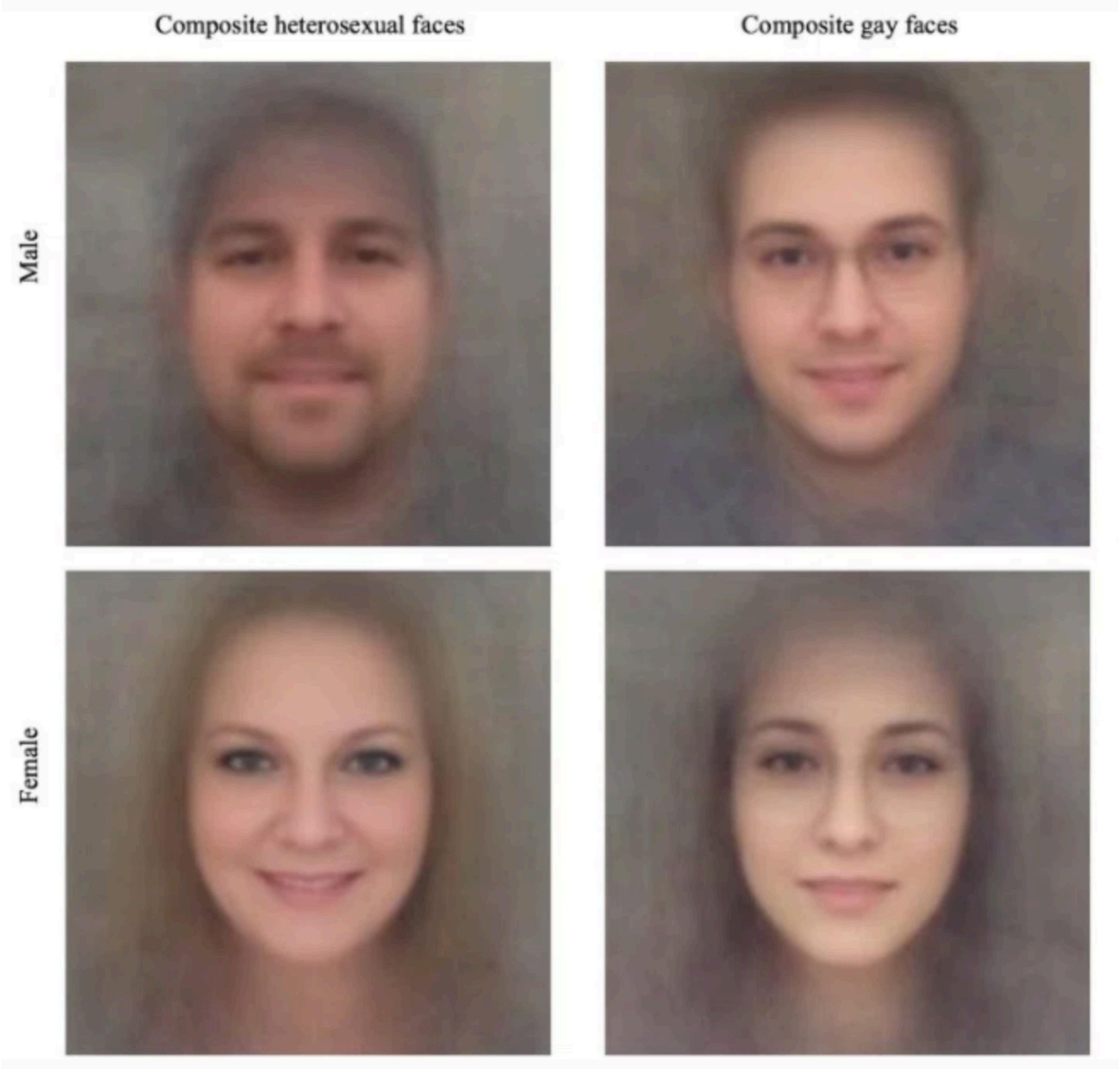
- Wang and Kosinski: gay vs. straight classification based on faces



Slide credit: https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477

Bad Applications

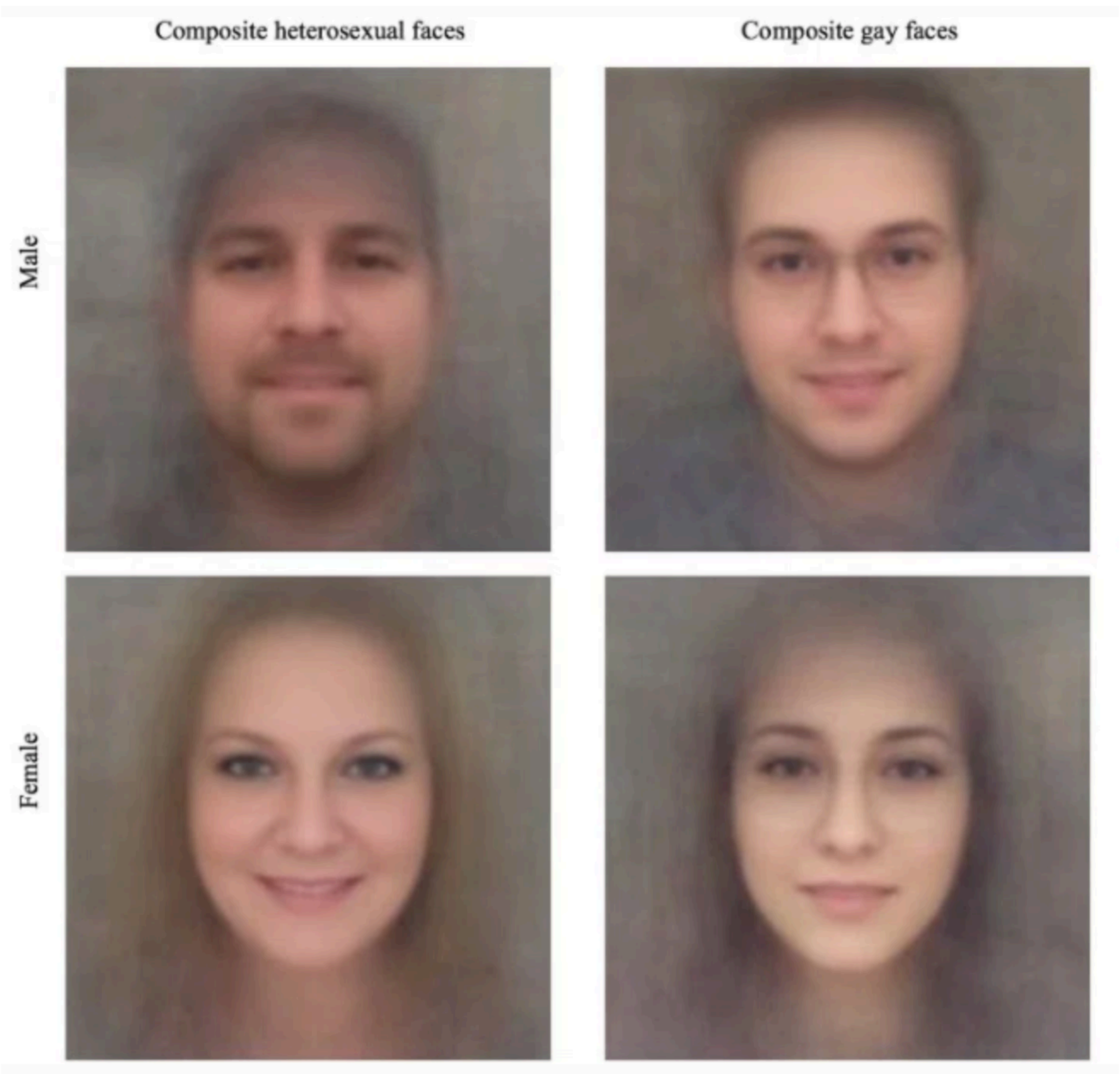
- Wang and Kosinski: gay vs. straight classification based on faces
- Authors: “this is useful because it supports a hypothesis” (physiognomy)



Slide credit: https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477

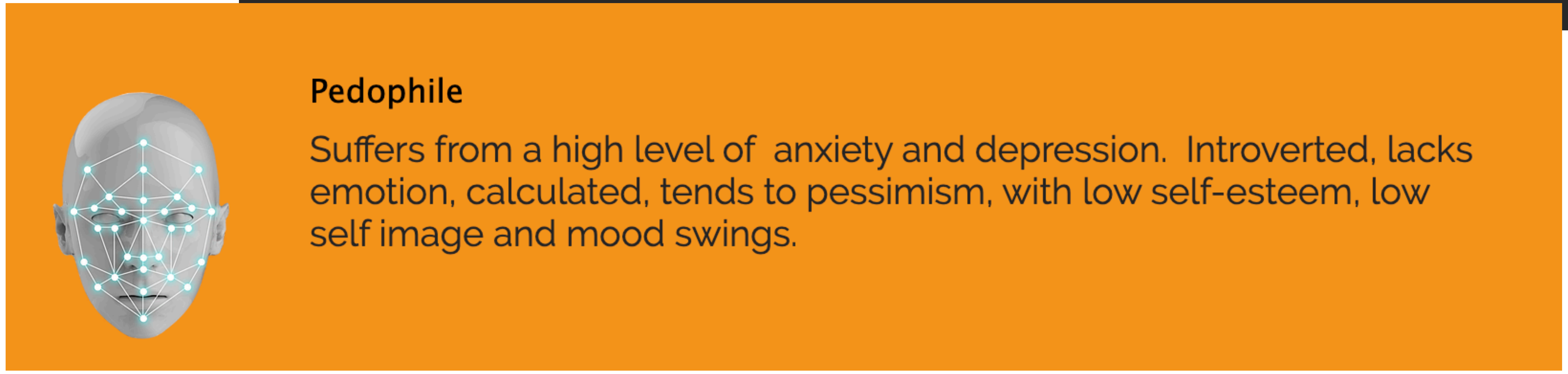
Bad Applications

- Wang and Kosinski: gay vs. straight classification based on faces
- Authors: "this is useful because it supports a hypothesis" (physiognomy)
- Blog post by Agüera y Arcas, Todorov, Mitchell: mostly social phenomena (glasses, makeup, angle of camera, facial hair) — bad science, *and* dangerous



Slide credit: https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477

Unethical Use



http://www.faception.com

How to Move Forward?

- ACM Code of Ethics
  - https://www.acm.org/code-of-ethics

- Contribute to society and to human well-being
- Avoid harm
- Be fair and take action not to discriminate
- Respect privacy
- … (see link above for more details)

Final Thoughts

Final Thoughts

- You will face choices: what you choose to work on, what company you choose to work for, etc.

Final Thoughts

- You will face choices: what you choose to work on, what company you choose to work for, etc.

- Tech does not exist in a vacuum: you can work on problems that will fundamentally make the world a better place or a worse place (not always easy to tell)

Final Thoughts

- You will face choices: what you choose to work on, what company you choose to work for, etc.

- Tech does not exist in a vacuum: you can work on problems that will fundamentally make the world a better place or a worse place (not always easy to tell)

- As AI becomes more powerful, think about what we *should* be doing with it to improve society, not just what we *can* do with it