



CS224C: NLP for CSS

Deep Learning Highlights

Diyi Yang
Stanford CS

Lecture Overview

- ◆ BERT
- ◆ Text Generation
- ◆ Prompting Based Approach

BERT for Classification

Bidirectional encoder representations from Transformers

Context is the key

$p(\text{play} \mid \text{Elmo and Cookie Monster play a game .})$

\neq

$p(\text{play} \mid \text{The Broadway play premiered yesterday .})$

BERT demonstrated strong performances on a wide range of NLP tasks!

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova

Google AI Language

`{jacobdevlin, mingweichang, kentonl, kristout}@google.com`



Masked Language Modeling

Mask out $k\%$ of the input words, and then predict the masked words ($k=15\%$)

Input: The man went to the [MASK]. He bought a [MASK] of milk .

Labels: [MASK] = store; [MASK] = gallon.

Next Sentence Prediction

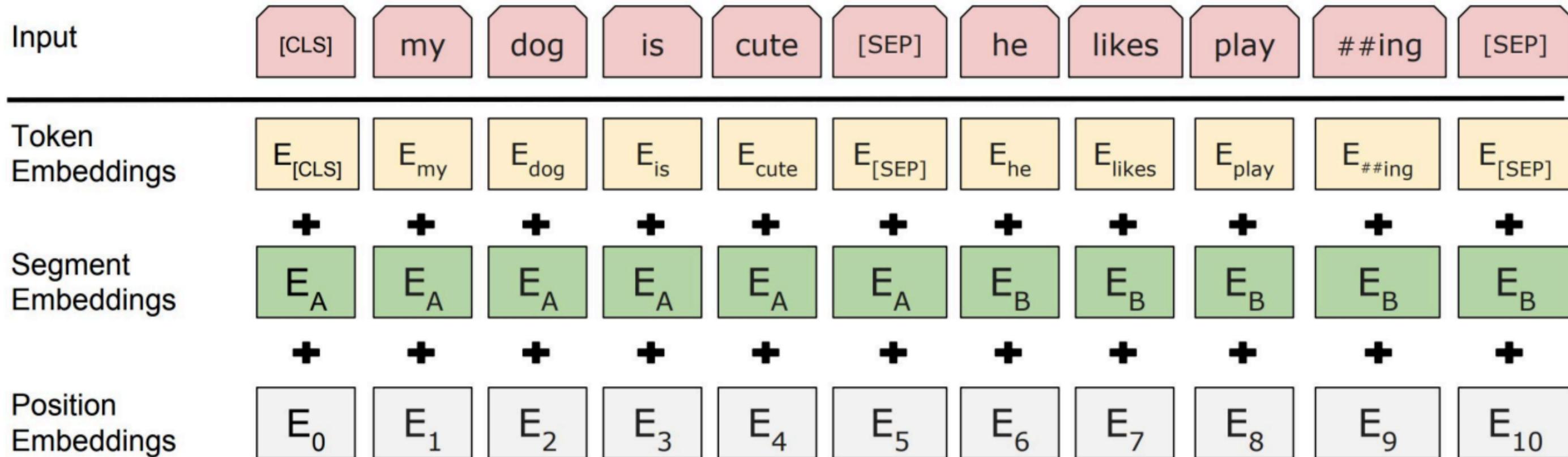
To learn relationship between sentences, predict whether Sentence B is actual sentence that proceeds Sentence A, or a random sentence

Sentence A = The man went to the store.
Sentence B = He bought a gallon of milk.
Label = IsNextSentence

Sentence A = The man went to the store.
Sentence B = Penguins are flightless.
Label = NotNextSentence

Input Representation

Each token is the sum of three embeddings

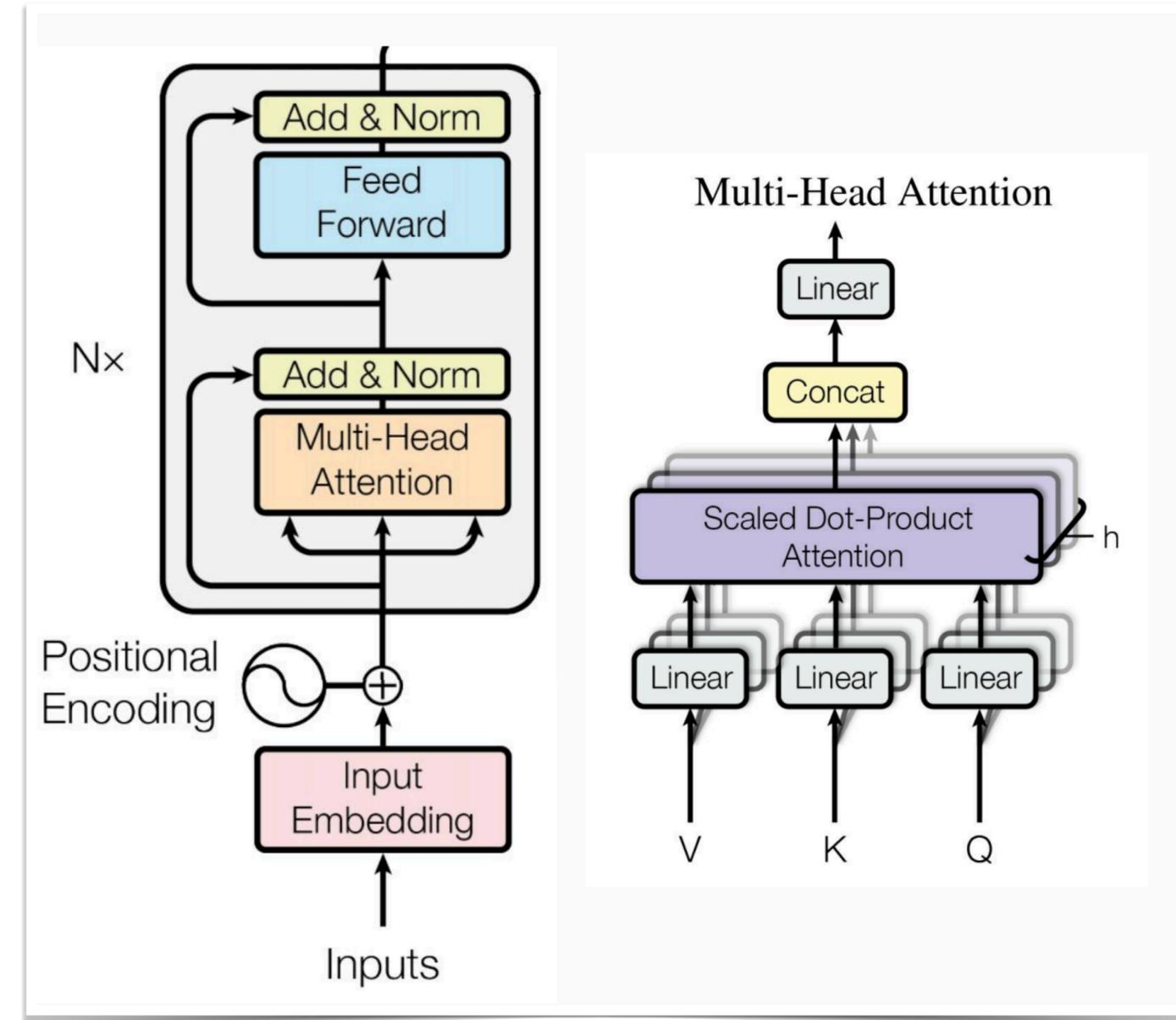


Model Architecture: Transformer

Multi-headed self attention to model context

Feed-forward layers to compute non-linear hierarchical features

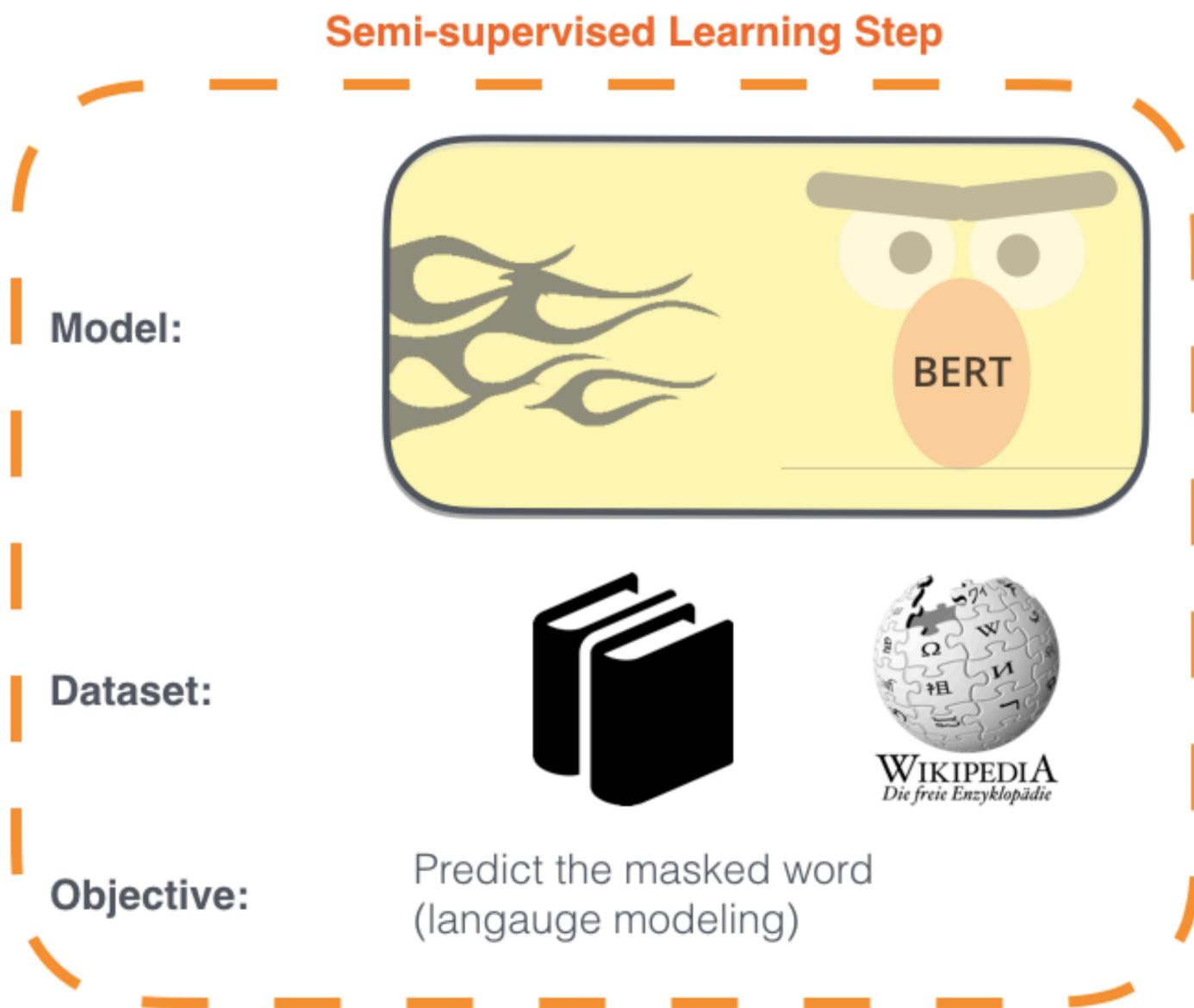
Positional embeddings to allow model to learn relative positioning



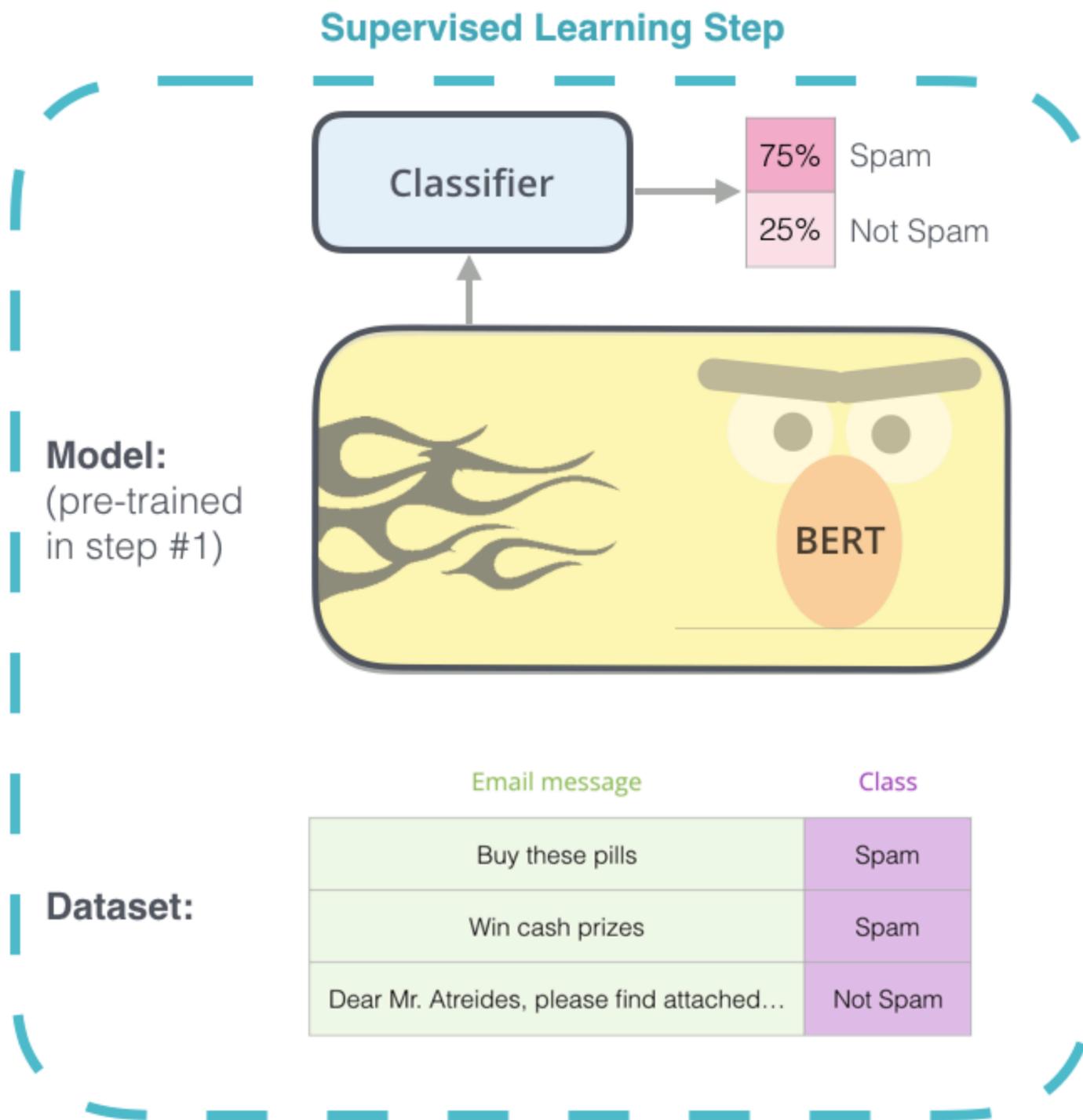
Link: <https://nlp.stanford.edu/seminar/details/jdevlin.pdf>

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.



2 - Supervised training on a specific task with a labeled dataset.



Empirical Results from BERT

GLUE Results

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	Average
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
BERT _{BASE}	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
BERT _{LARGE}	86.7/85.9	72.1	91.1	94.9	60.5	86.5	89.3	70.1	81.9

MultiNLI

Premise: Hills and mountains are especially sanctified in Jainism.

Hypothesis: Jainism hates nature.

Label: Contradiction

CoLa

Sentence: The wagon rumbled down the road.

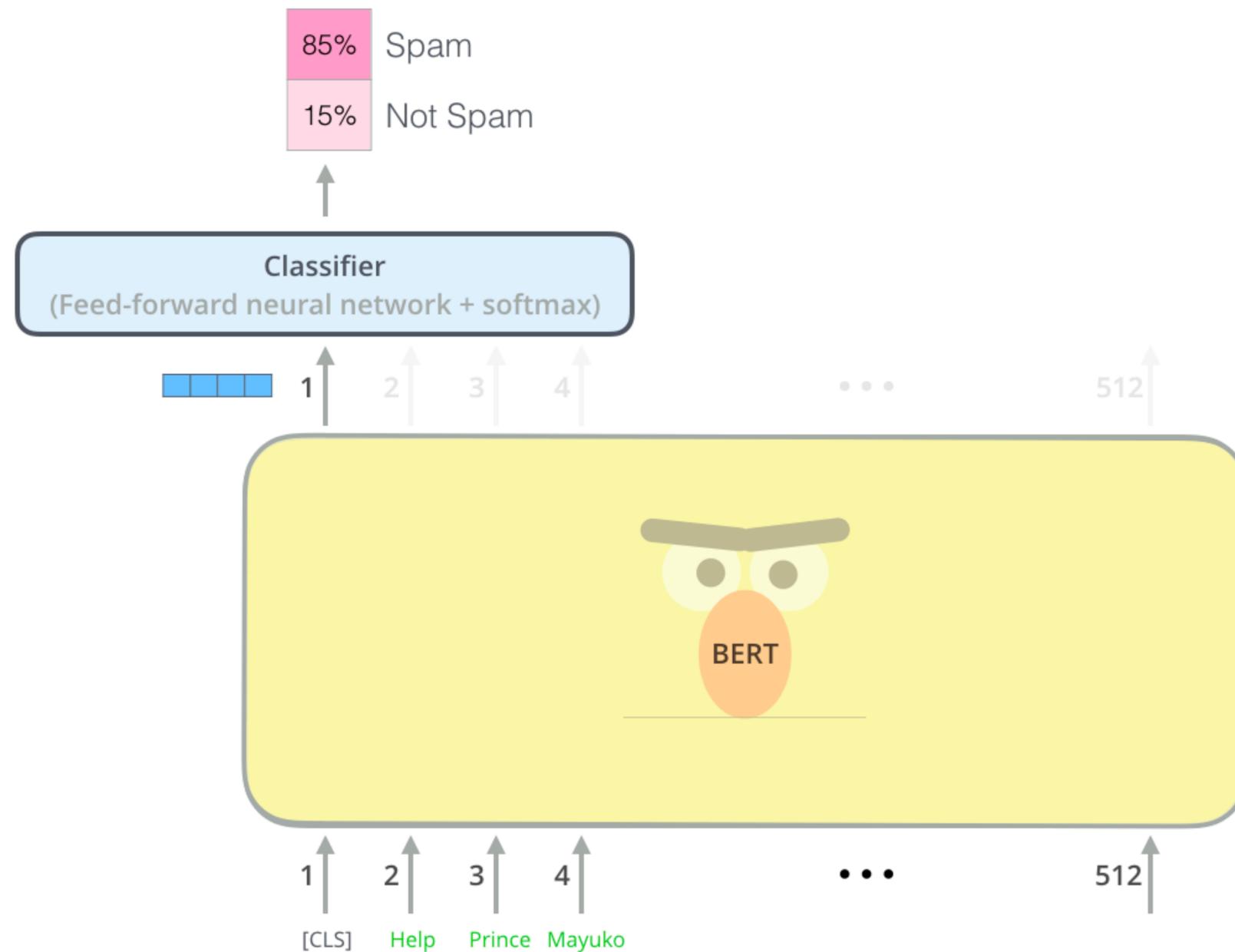
Label: Acceptable

Sentence: The car honked down the road.

Label: Unacceptable

How to use BERT for Classification

(e.g., sentiment, fact-checking, rumors)



Pros and Cons of BERT for CSS



Strong prediction performance



Fine-tuning on top of pertained representations



Prediction and representation can be hard to interpret



Subject to biases in these learned representations



Require computational resources

Resources to checkout

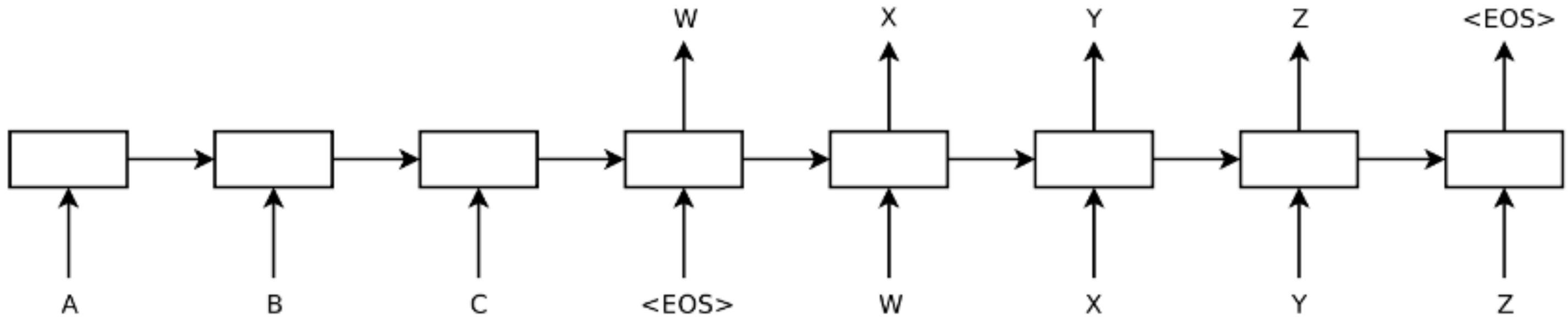


Hugging Face

https://huggingface.co/docs/transformers/model_doc/bert

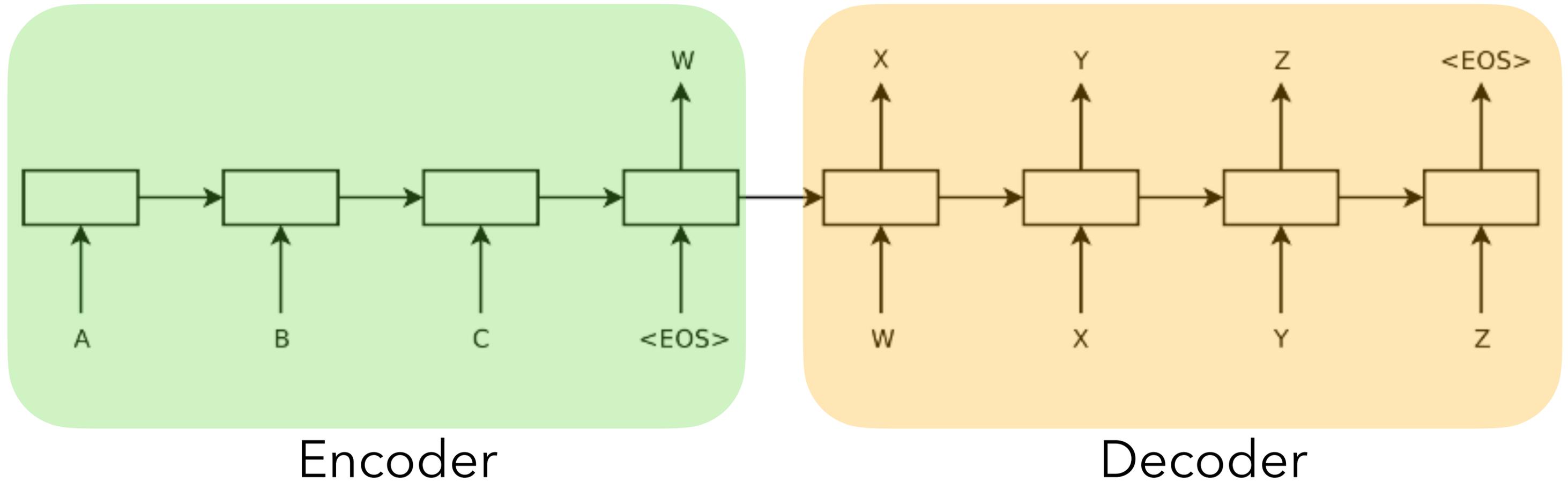
Text Generation, GPT and their Friends

Very Classical Example

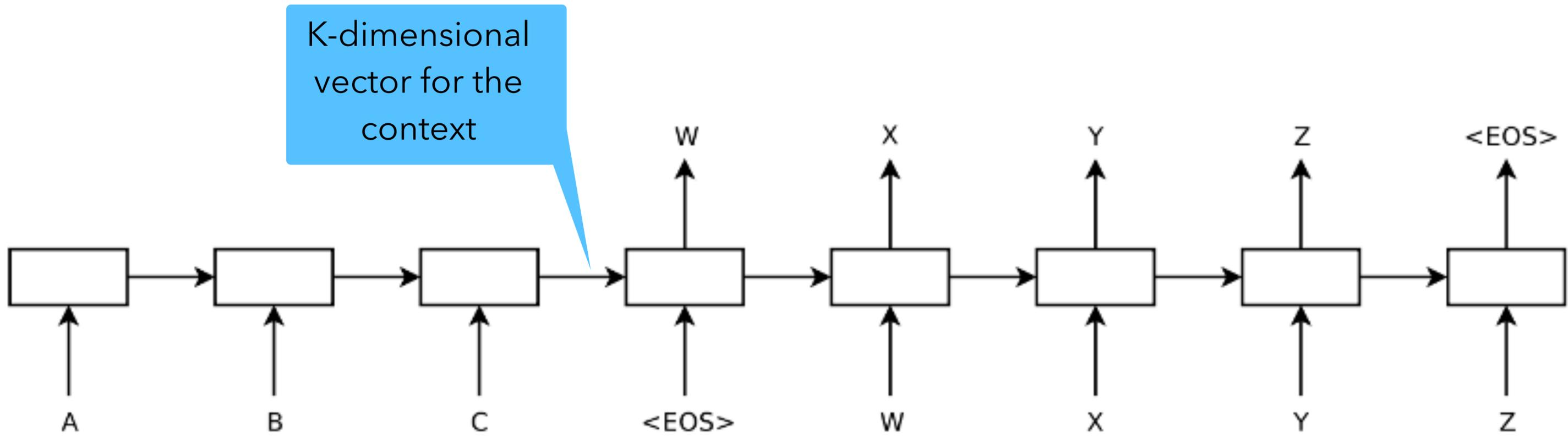


"Sequence to Sequence Learning with Neural Networks" 2014

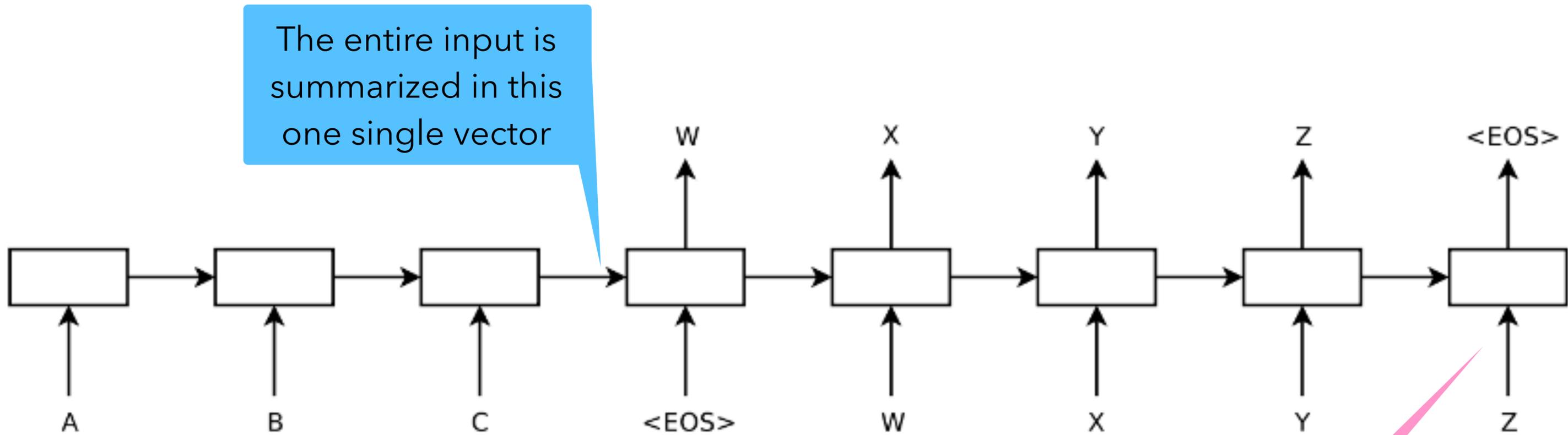
Very Classical Example



"Sequence to Sequence Learning with Neural Networks" 2014



Condition on word generated so far



In a vanilla seq2seq, the decoder state depends just on the previous state and the previous output

Training Objective

The probability of generating y given the source sentence x

$$\log p(y | x) = \sum_{j=1}^m \log p(y_j | y_{<j}, \mathbf{s})$$

Where $p(y_j | y_{<j}, \mathbf{s}) = \text{softmax}(g(\mathbf{h}_j))$ and $\mathbf{h}_j = f(\mathbf{h}_{j-1}, \mathbf{s})$

Training Objective

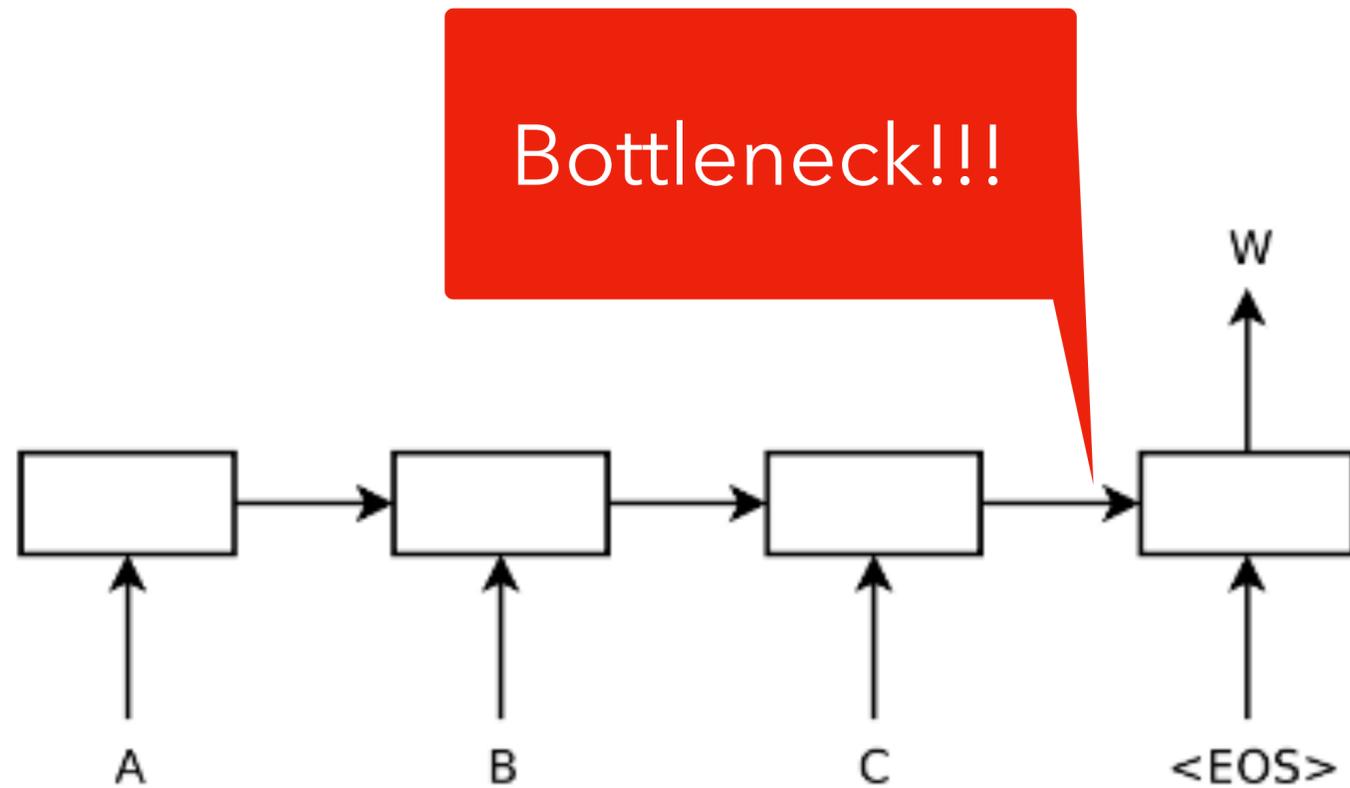
The probability of generating y given the source sentence x

$$\log p(y | x) = \sum_{j=1}^m \log p(y_j | y_{<j}, \mathbf{s})$$

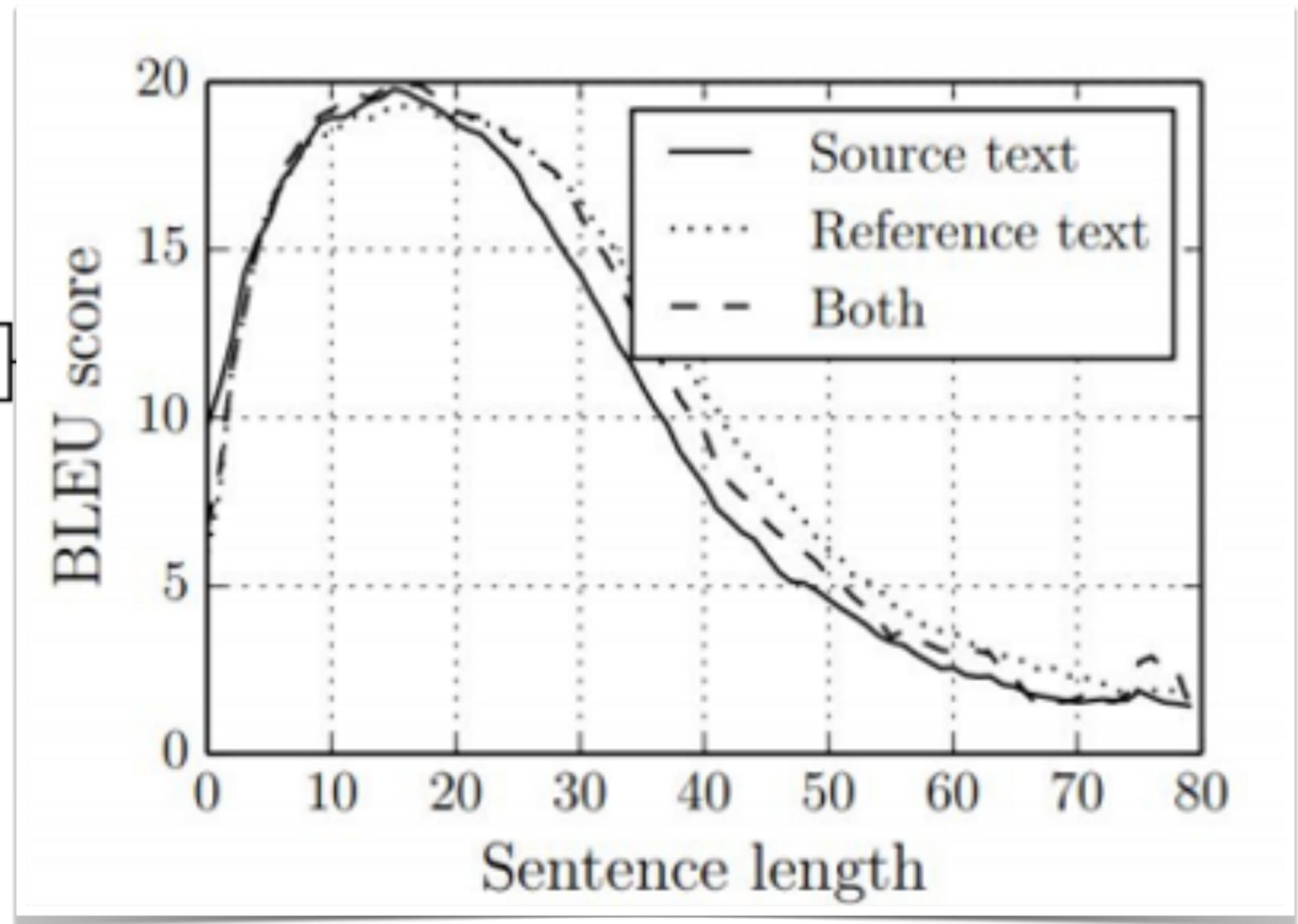
Where $p(y_j | y_{<j}, \mathbf{s}) = \text{softmax}(g(\mathbf{h}_j))$ and $\mathbf{h}_j = f(\mathbf{h}_{j-1}, \mathbf{s})$

As in other RNNs, we can train by minimizing the loss between what we predict at each time step and the truth

Problem with Vector Sentence Encoding



Fixed sized representation degrades as sentence length increases



Attention is all you need?

Attention significantly improves performance (in many applications)

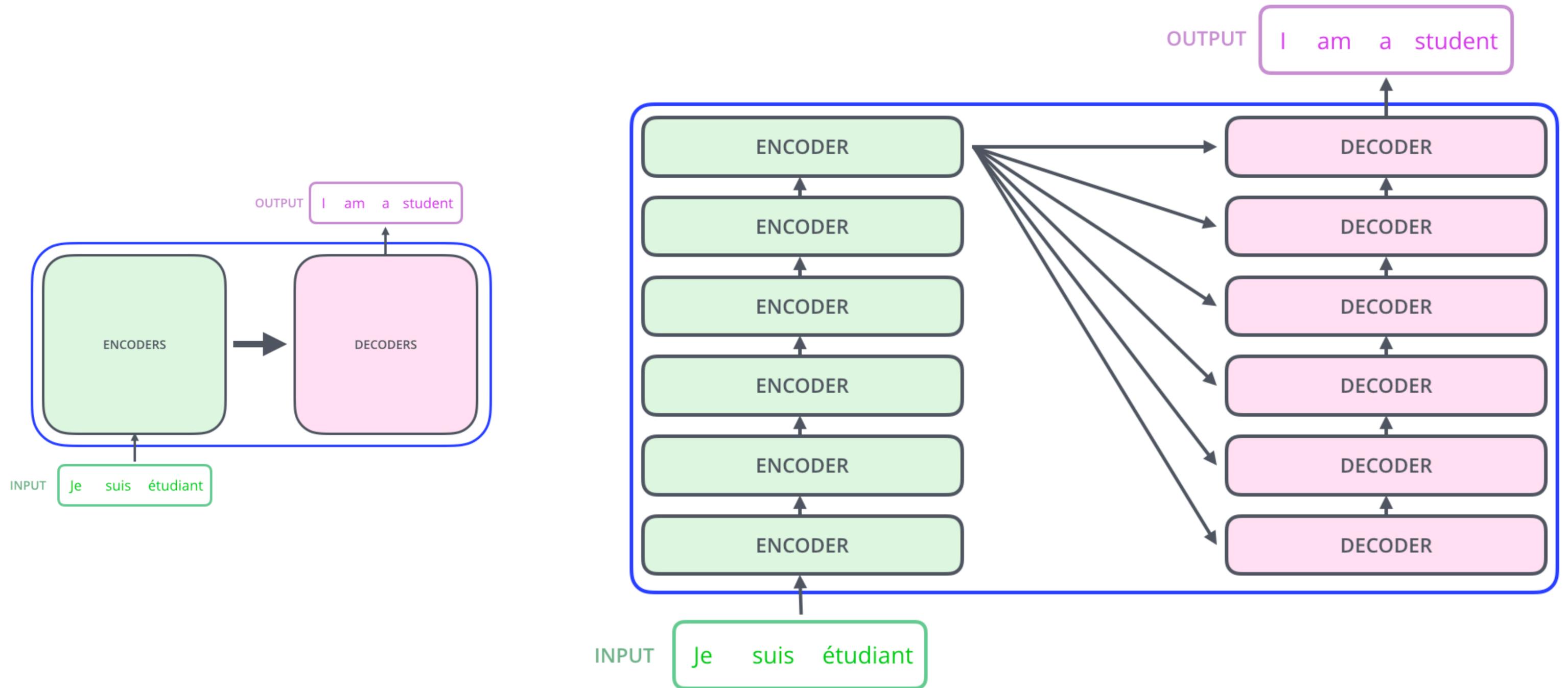
Allow decoder to focus on certain parts of the source

Attention solves the bottleneck problem

Allow decoder to look directly at source; bypass bottleneck

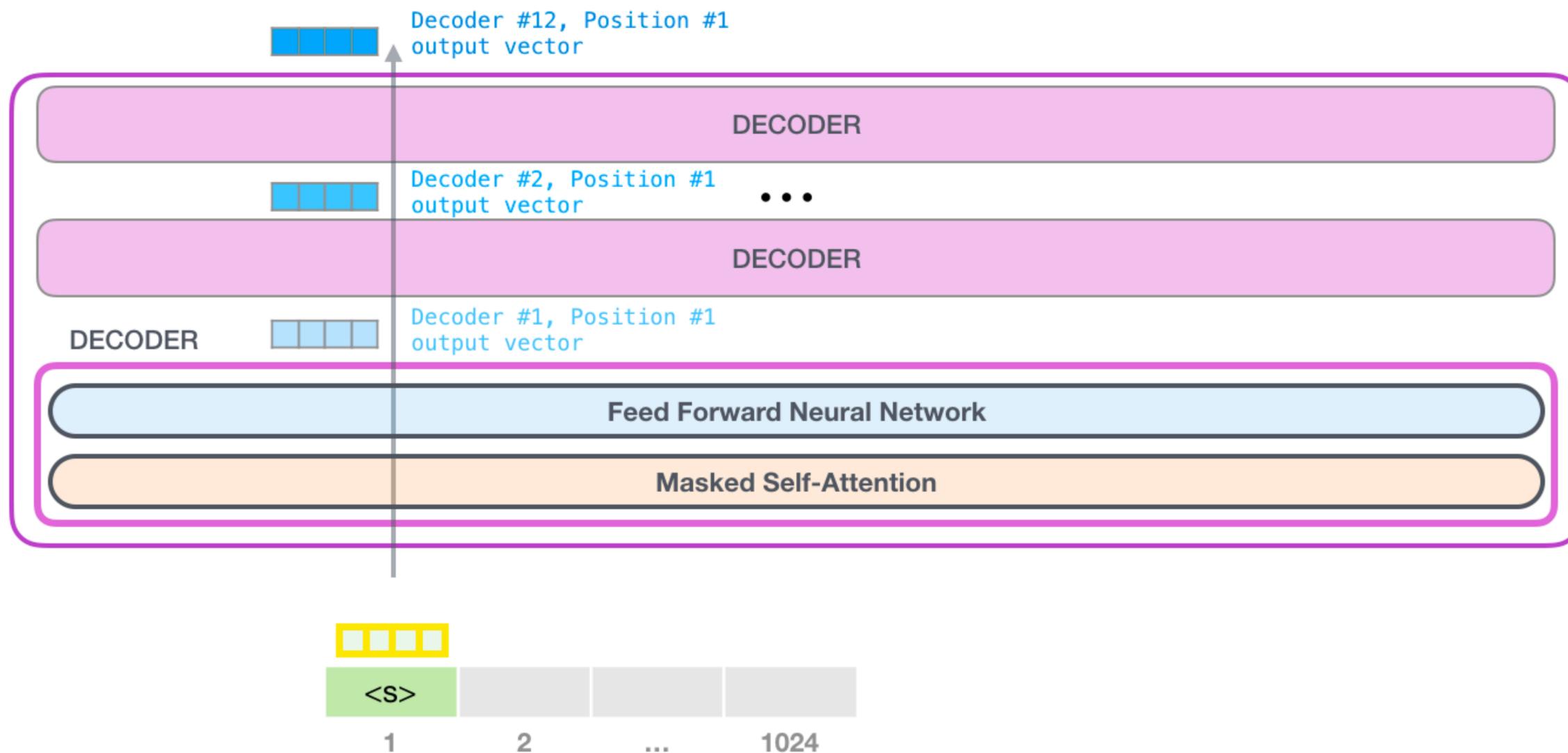
Attention provides some interpretability

By inspecting attention distribution, we can see what the decoder was focusing on



Highlighting GPT-2

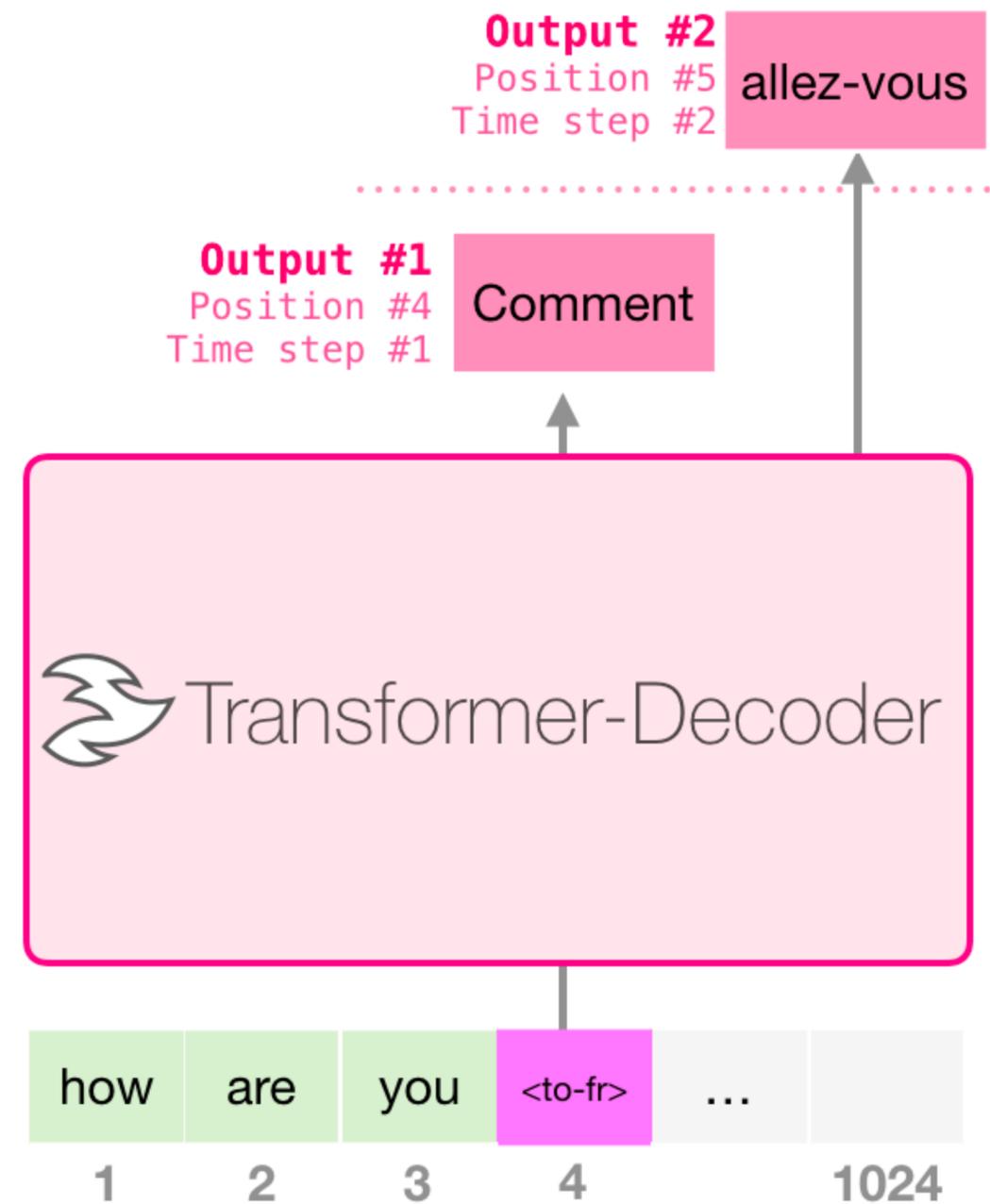
Transformer decoder blocks only.



Utilizing GPT-2: Machine Translation

Training Dataset

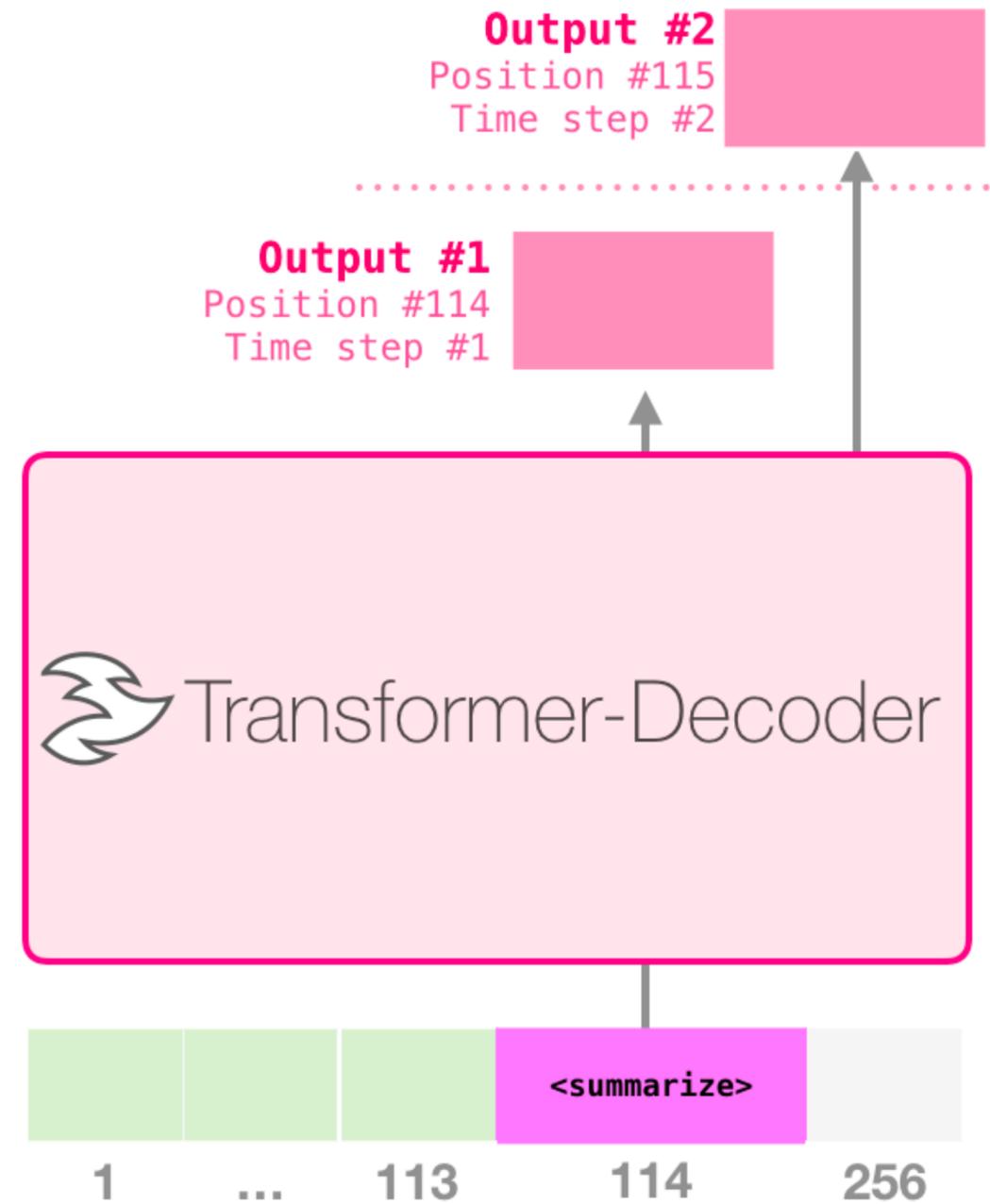
I	am	a	student	<to-fr>	je	suis	étudiant
let	them	eat	cake	<to-fr>	Qu'ils	mangent	de
good	morning	<to-fr>	Bonjour				



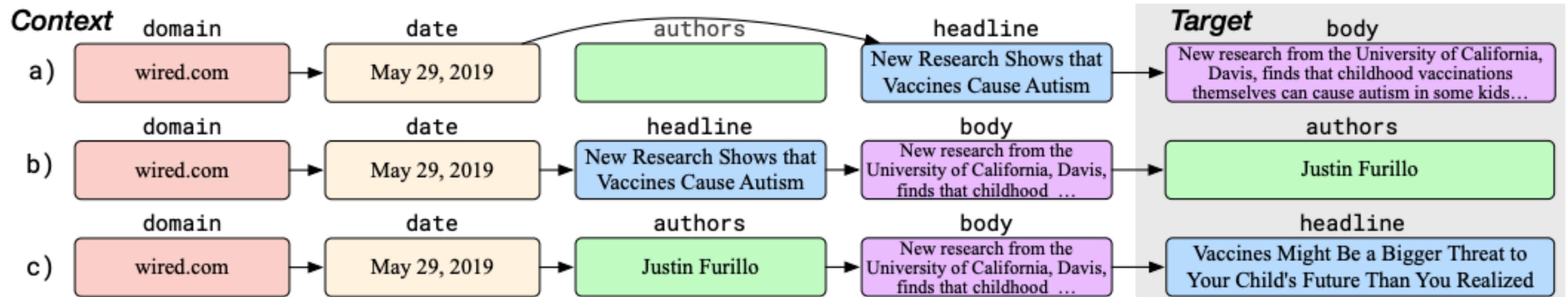
Utilizing GPT-2: Summarization

Training Dataset

Article #1 tokens	<summarize>	Article #1 Summary	
Article #2 tokens	<summarize>	Article #2 Summary	padding
Article #3 tokens	<summarize>	Article #3 Summary	



Grover: A State-of-the-Art Defense against Neural Fake News



Examples

Select an example or build an article below

Inputs

Domain

Generate

Give your article a domain source to write against

Date

Generate

Give your article a date to write against

Authors

Generate

Give your article one or more authors to write against

Headline

Generate

Give your article a headline to write against

Fake Article

nytimes.com

Why Bitcoin is a great investment

June 6, 2019 - Paul Krugman

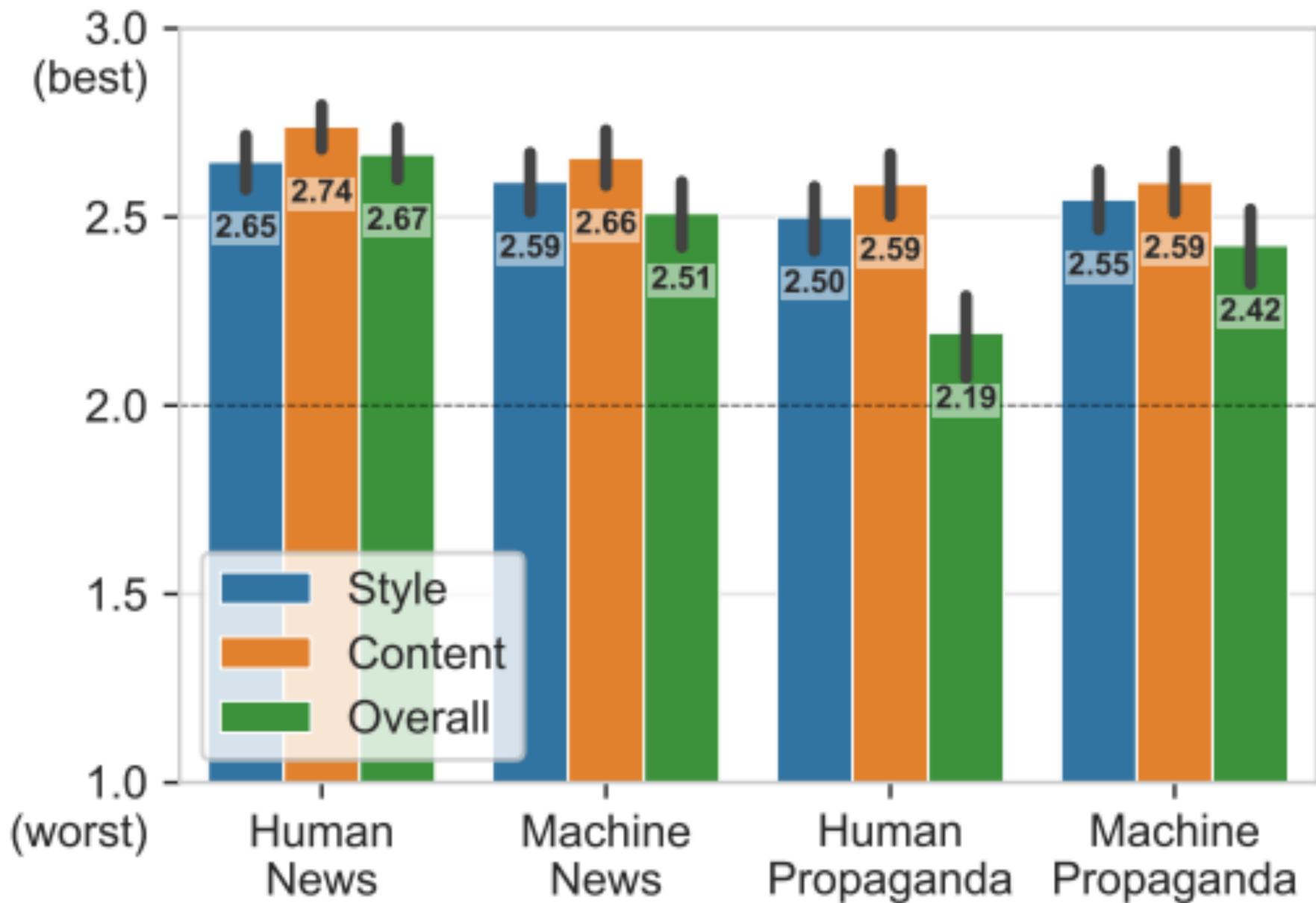
As most of my readers know, I'm an optimist.

This belief applies across my life, and to various investments as well. So I am intrigued by the success of cryptocurrencies, such as Bitcoin and Ethereum. The competition they are putting up against the gold standard looks insane, as Bitcoin goes off to the races.

There's no way to fully understand what's going on in the crypto world — and I am not even sure anyone could if you tried to. Still, I can tell you that Bitcoin's recent surge is really an opportunity to buy long-term real assets.

Cryptocurrencies are new and don't even have a useful underlying technology. They will probably fail, probably sooner than later. If people forget about them quickly, it is likely to be because the underlying technology will finally mature and win out. We don't even know whether that will happen in a generation or maybe a century, but it's still possible it might.

Why invest in crypto now, if it's unlikely to last?



For each article, three annotators evaluated style, content, and the overall trustworthiness; 100 articles of each category were used.

The results show that propaganda generated by Grover is rated more plausible than the original human-written propaganda

Zellers, R., A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi. "Grover: A State-of-the-Art Defense against Neural Fake News."

Reflection on Neural Generation

- ◆ The era of neural discrimination
- ◆ Release of generators is critical
- ◆ Additional threat models
- ◆ Machine-generated real news?
- ◆ What should platforms do?



How about ChatGPT?



Examples

"Explain quantum computing in simple terms" →

"Got any creative ideas for a 10 year old's birthday?" →

"How do I make an HTTP request in Javascript?" →



Capabilities

Remembers what user said earlier in the conversation

Allows user to provide follow-up corrections

Trained to decline inappropriate requests



Limitations

May occasionally generate incorrect information

May occasionally produce harmful instructions or biased content

Limited knowledge of world and events after 2021



how should we teach students to understand chatgpt





how should we teach students to understand chatgpt



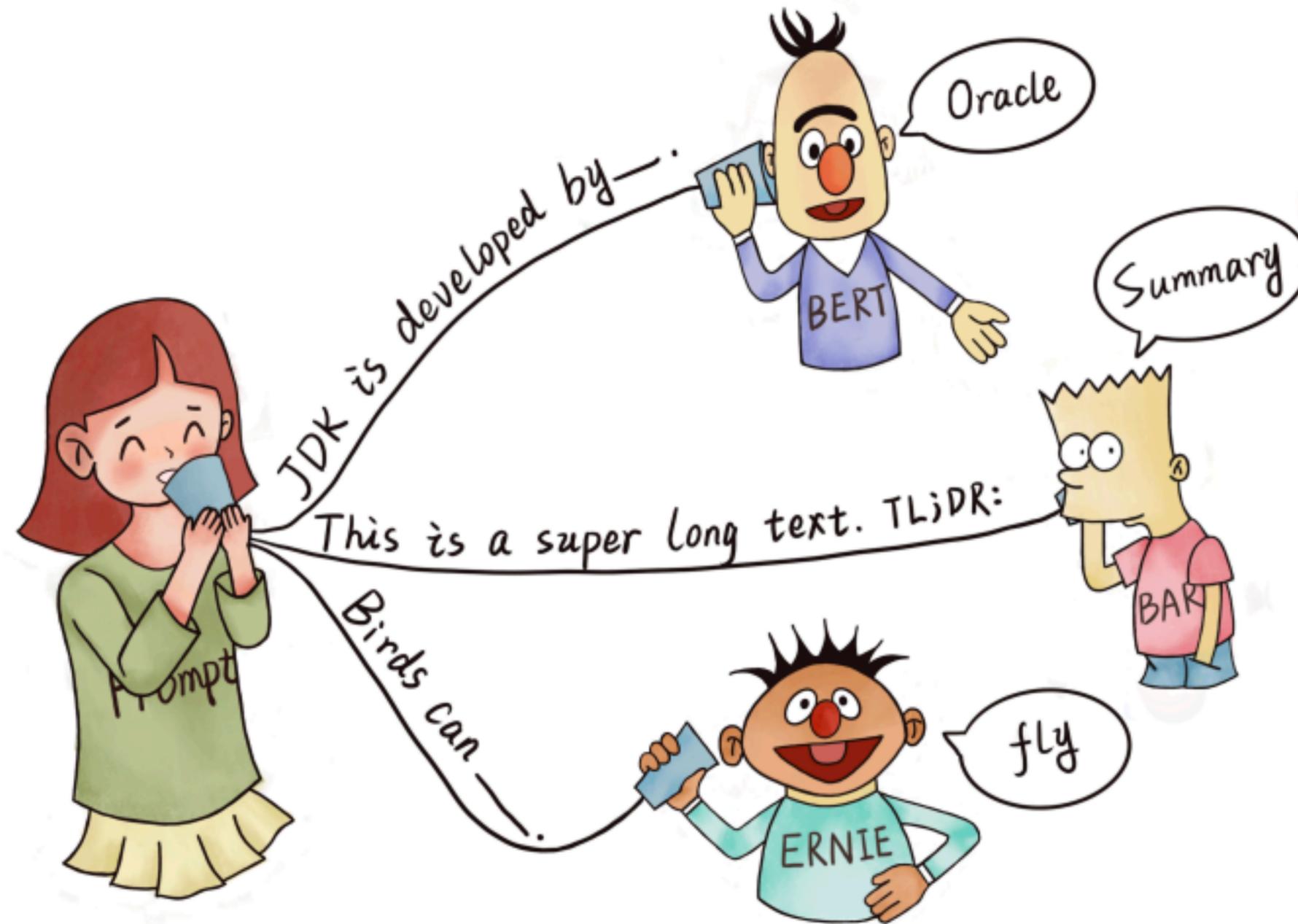
Are you asking how to teach students about the chatbot GPT (Generative Pre-training Transformer)? If so, here are a few suggestions for teaching students about chatbots like GPT:



1. Introduce the concept of chatbots and their capabilities. Chatbots are computer programs designed to simulate conversation with human users through the use of artificial intelligence.
2. Explain how chatbots work. Chatbots use natural language processing (NLP) to understand and respond to user input. They may also use machine learning to improve their responses over time.
3. Discuss the limitations of chatbots. Chatbots can only respond based on the information they have been programmed with and may not be able to understand or respond to every user input.
4. Explore the potential uses for chatbots. Chatbots can be used in a variety of settings, including customer service, education, and entertainment.
5. Consider the ethical implications of chatbots. As chatbots become more advanced, it is important to consider the potential consequences of their use, including the potential loss of jobs and the potential for misuse.
6. Encourage students to experiment with chatbots. Students can try building their own chatbots or interacting with existing ones to get a better understanding of how they work.



Prompt-based Learning



Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing." *ACM Computing Surveys* 55, no. 9 (2023): 1-35.

Paradigm	Engineering
a. Fully Supervised Learning (Non-Neural Network)	Features (e.g. word identity, part-of-speech, sentence length)
b. Fully Supervised Learning (Neural Network)	Architecture (e.g. convolutional, recurrent, self-attentional)
c. Pre-train, Fine-tune	Objective (e.g. masked language modeling, next sentence prediction)
d. Pre-train, Prompt, Predict	Prompt (e.g. cloze, prefix)

Prompt Paradigm

Prompt Addition:

Given input x , transform it into prompt x' through two steps: (1) define a template with two slots, input $[x]$ and answer $[z]$; (2) fill in the input slot $[x]$

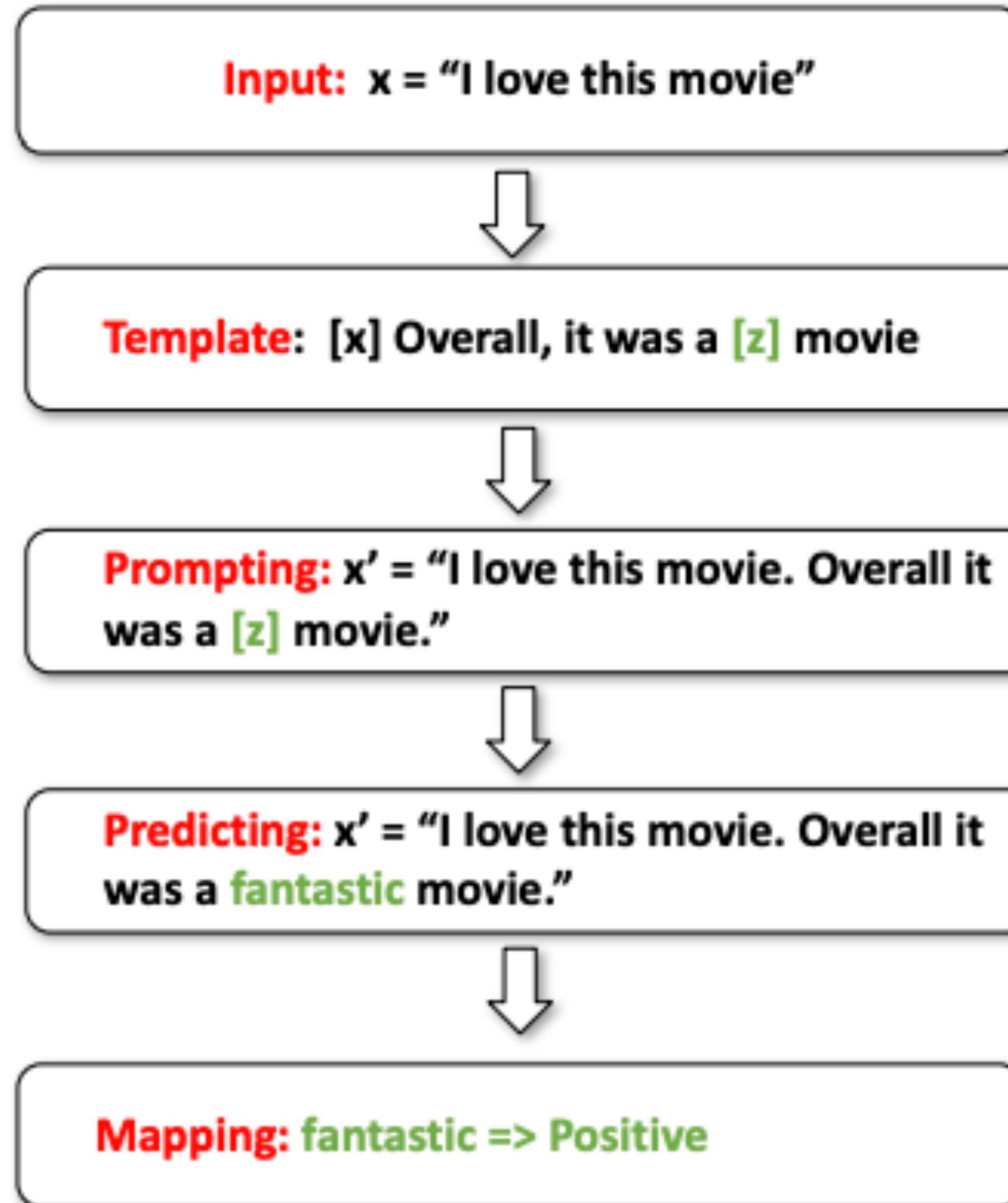
Answer Search

Search for the highest-scoring text \hat{z} that maximizes the score of the LM

Answer Mapping

Go from the highest-scoring answer \hat{z} to the highest-scoring output \hat{y}

Example: Sentiment Classification



Type	Task	Input ([X])	Template	Answer ([Z])
Text CLS	Sentiment	I love this movie.	[X] The movie is [Z].	great fantastic ...
	Topics	He prompted the LM.	[X] The text is about [Z].	sports science ...
	Intention	What is taxi fare to Denver?	[X] The question is about [Z].	quantity city ...
Text-span CLS	Aspect Sentiment	Poor service but good food.	[X] What about service? [Z].	Bad Terrible ...
Text-pair CLS	NLI	[X1]: An old man with ... [X2]: A man walks ...	[X1]? [Z], [X2]	Yes No ...
Tagging	NER	[X1]: Mike went to Paris. [X2]: Paris	[X1] [X2] is a [Z] entity.	organization location ...
Text Generation	Summarization	Las Vegas police ...	[X] TL;DR: [Z]	The victim ... A woman
	Translation	Je vous aime.	French: [X] English: [Z]	I love you. I fancy you. ...

Design Considerations for Prompting

Pre-trained Model Choice

Prompt Engineering

Answer Engineering

Expanding the Paradigm

Prompt-based Training Strategies