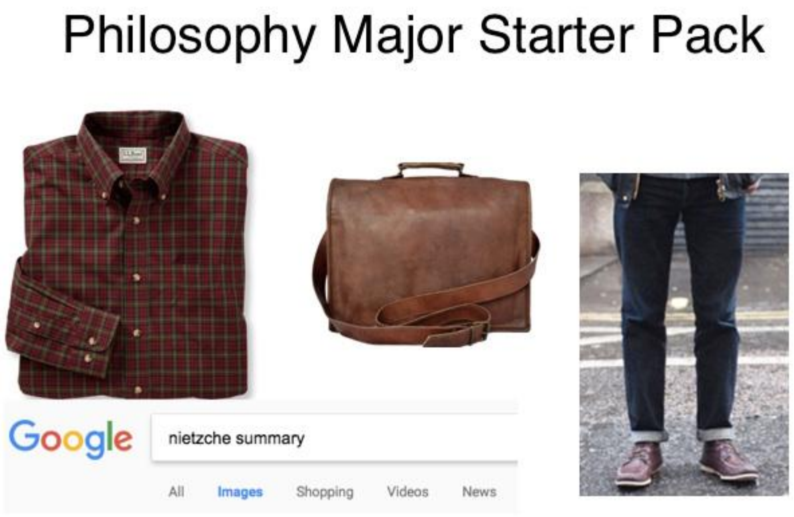
Natural Language Understanding

Introduction to LLM

Lecture 7

Outline

- “The Turing Test”
- What does LMs learn?
- NLU
    - Pre-train/fine-tune paradigm



What does it mean to “**understand**” language?

ELIZA: A computer psychiatrist

Men are all alike.

In what way?

They're always bugging us about something or the other.

Can you think of a specific example?

Well, my boyfriend made me come here.

Your boyfriend made you come here.

He says I'm depressed.

Joseph Weizenbaum, *Computer Power and Human Reason*, W.H. Freeman, 1976.

ELIZA: A computer psychiatrist

ELIZA Rules:

- `(.*) YOU (.*) ME -> WHAT MAKES YOU THINK I \2 YOU`
  - `USER:  You hate me`
  - `ELIZA: WHAT MAKES YOU THINK I HATE YOU`
- `I (.*) -> You say you \1`
  - `USER:  I know everybody laughed at me`
  - `ELIZA: YOU SAY YOU KNOW EVERYBODY LAUGHED AT YOU`
- Sees the word *"Everybody"* `-> WHO IN PARTICULAR ARE YOU THINKING OF?`

Using language is not necessarily understanding language.

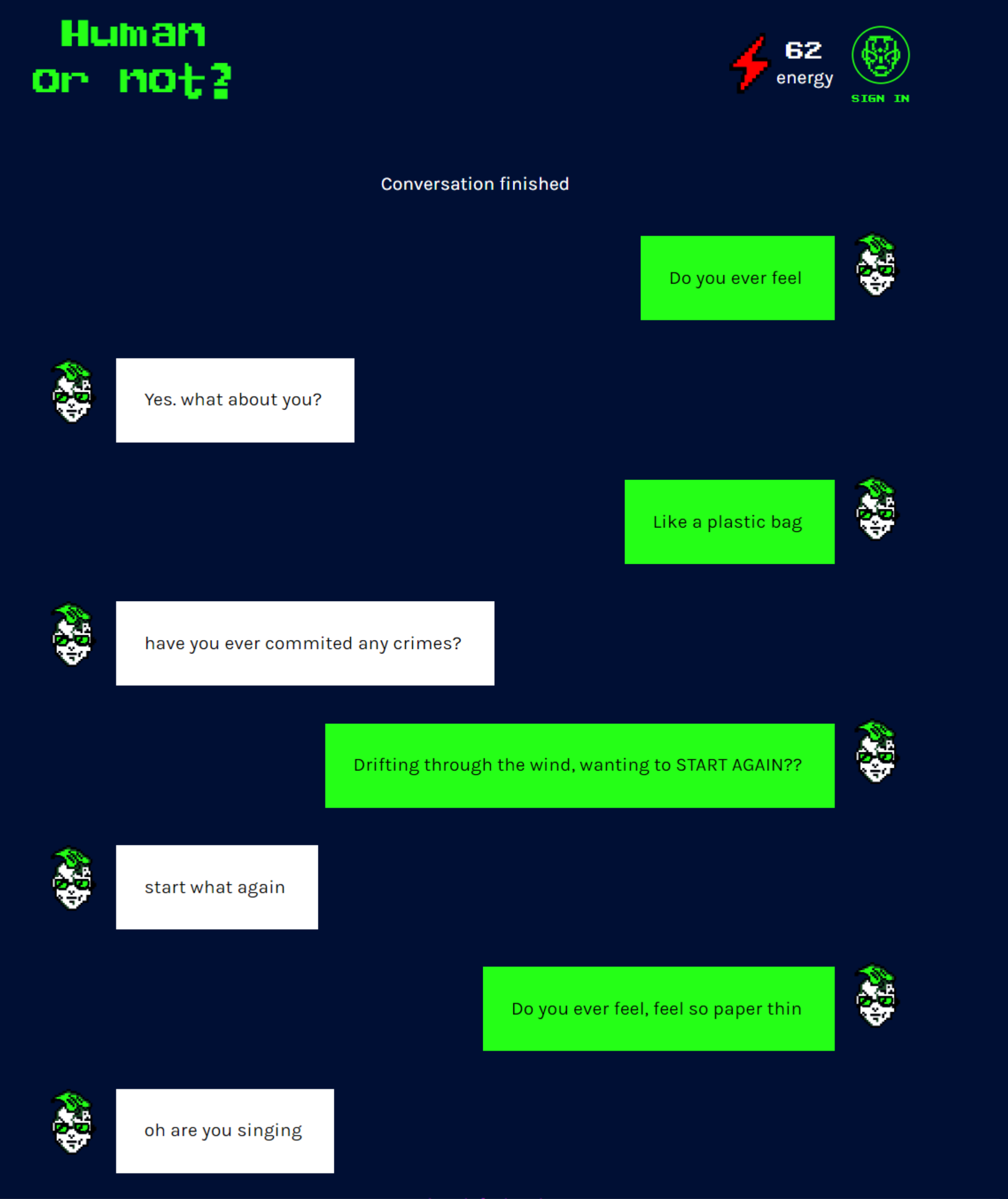
The Turing Test

6.30pm GMT
How to spot and debunk misleading content
Royal Society Milner Prize Lecture
given by Professor Iryna Gurevych
Participate live on
slido.com #MP1711
#RSPrizeTalks

ALAN TURING
(KC 1931)
COMPUTER SCIENTIST
MATHEMATICIAN
WE CAN ONLY SEE A SHORT DISTANCE AHEAD, BUT WE CAN SEE PLENTY THERE THAT NEEDS TO BE DONE

🏆

https://humanornot.so/



Conversation finished

Do you ever feel

Yes. what about you?

Like a plastic bag

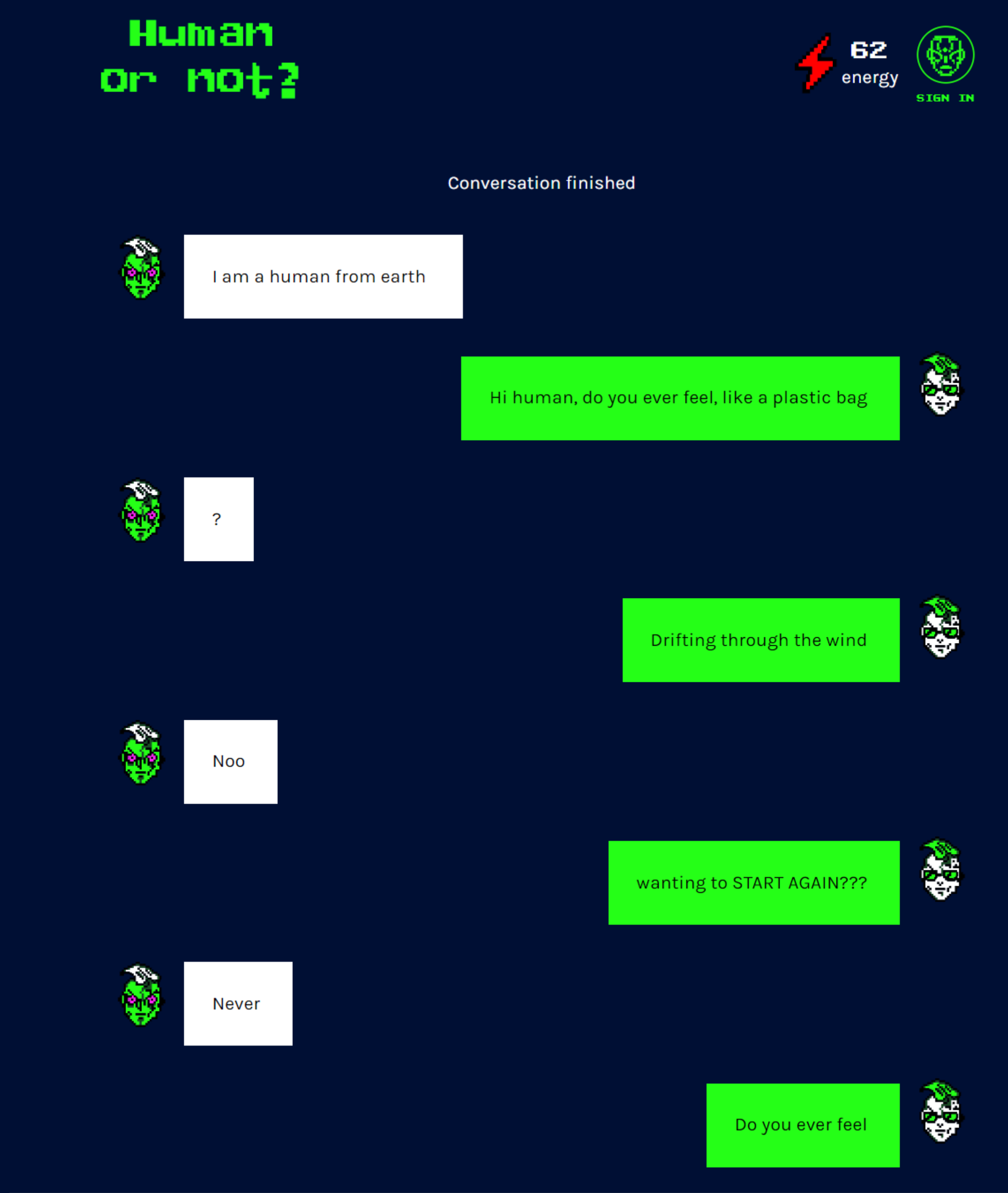
have you ever commited any crimes?

Drifting through the wind, wanting to START AGAIN??

start what again

Do you ever feel, feel so paper thin

oh are you singing



Conversation finished

I am a human from earth

Hi human, do you ever feel, like a plastic bag

?

Drifting through the wind

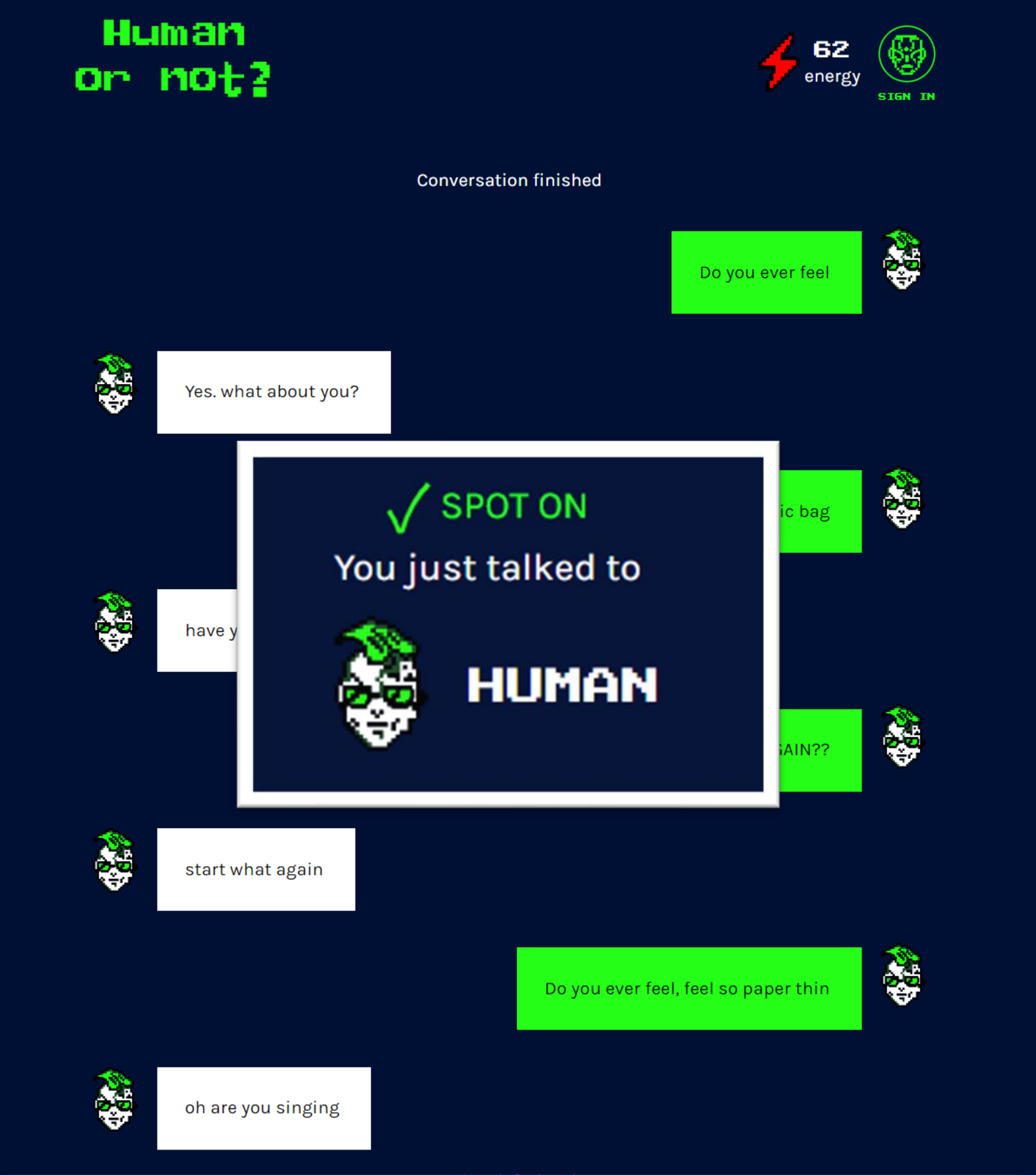
Noo

wanting to START AGAIN???

Never

Do you ever feel

https://humanornot.so/



Human or not?

62 energy

SIGN IN

Conversation finished

Do you ever feel

Yes. what about you?

ic bag

have y

AIN??

start what again

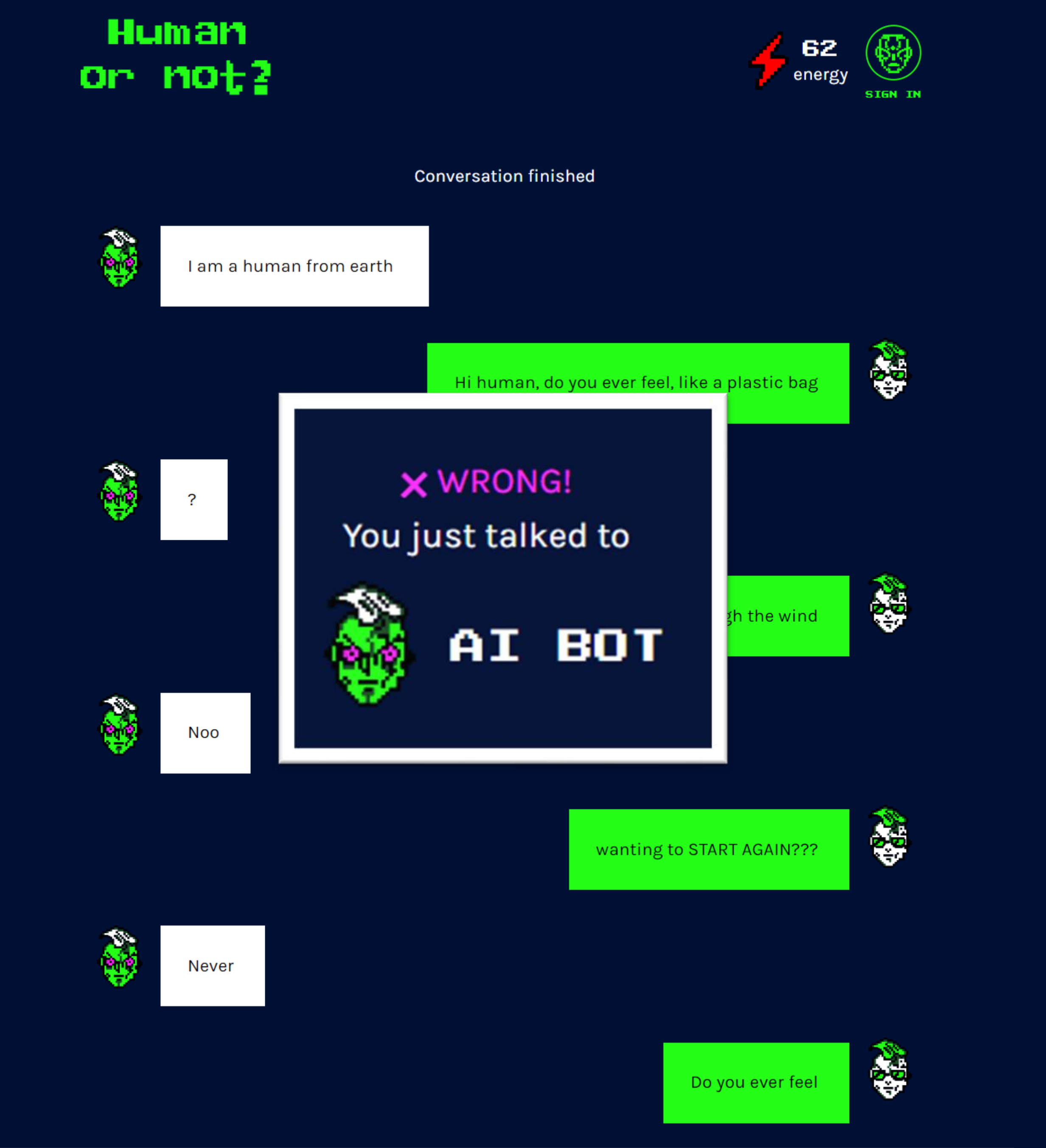
Do you ever feel, feel so paper thin

oh are you singing

✓ SPOT ON

You just talked to

HUMAN



Human or not?

62 energy

SIGN IN

Conversation finished

I am a human from earth

Hi human, do you ever feel, like a plastic bag

?

h the wind

Noo

wanting to START AGAIN???

Never

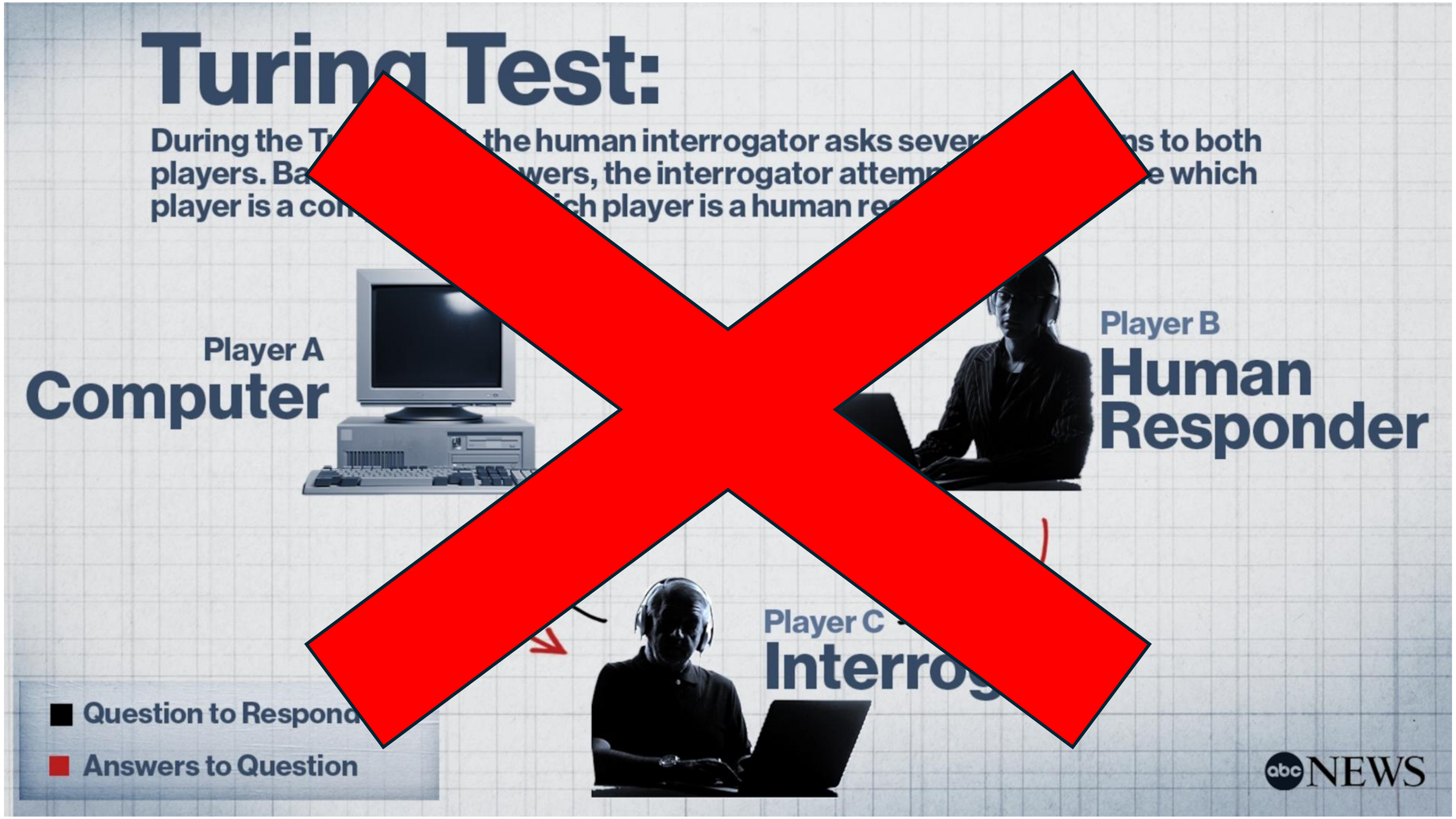
Do you ever feel

✕ WRONG!

You just talked to

AI BOT

The Turing Test


Turing Test:
During the T… the human interrogator asks sever… ns to both players. Ba… wers, the interrogator attem… e which player is a co… ich player is a human re…
Player A
Computer
Player B
Human Responder
Player C
Interro…
■ Question to Respond…
🟥 Answers to Question
abc NEWS

The Actual Turing Test

- Turing 1950. Computing Machinery and Intelligence.
  - A good annotated version of the paper.
- The **imitation game**. A: man, B: woman, C: interrogator.
  A: trick the interrogator, B: help the interrogator, C: guess who is woman/man.
- Think ≈ Soul ≈ Free will (in 1950).
- We now ask the question, “What will happen when a machine takes the part of A in this game?” Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, “Can machines think?”
- Intelligence: performance capacity.

Levels of Understanding

0. Keyword Processing:

- Limited knowledge of **particular words** or **phrases**, or their collocations.
  - Chatbots (ELIZA).
  - Information retrieval.
  - Web searching.

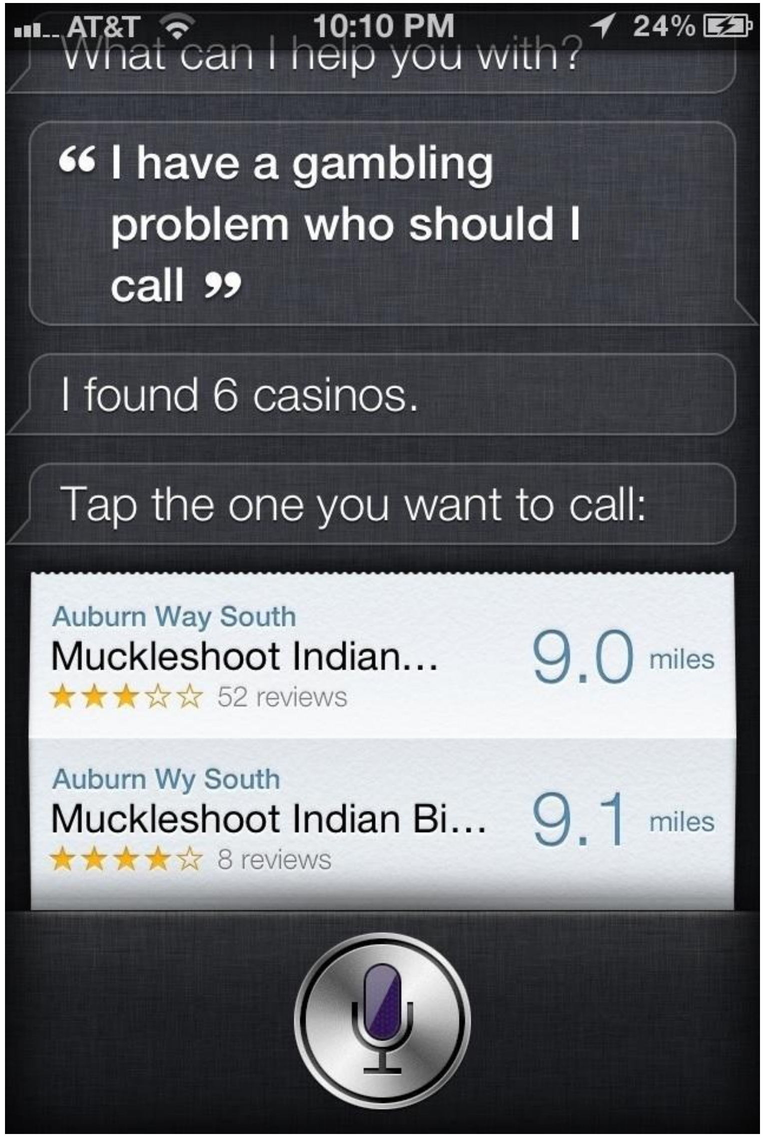
Levels of Understanding

1. Limited linguistic ability:

- Appropriate response to simple, highly constrained **sentences**.
  - Database queries in NL.
    "Show all sales staff who exceeded their quota in May."
  - Simple NL interfaces.
    "I want to fly from Toronto to Vancouver next Sunday."

Levels of Understanding

E.g., old Siri:



What can I help you with?

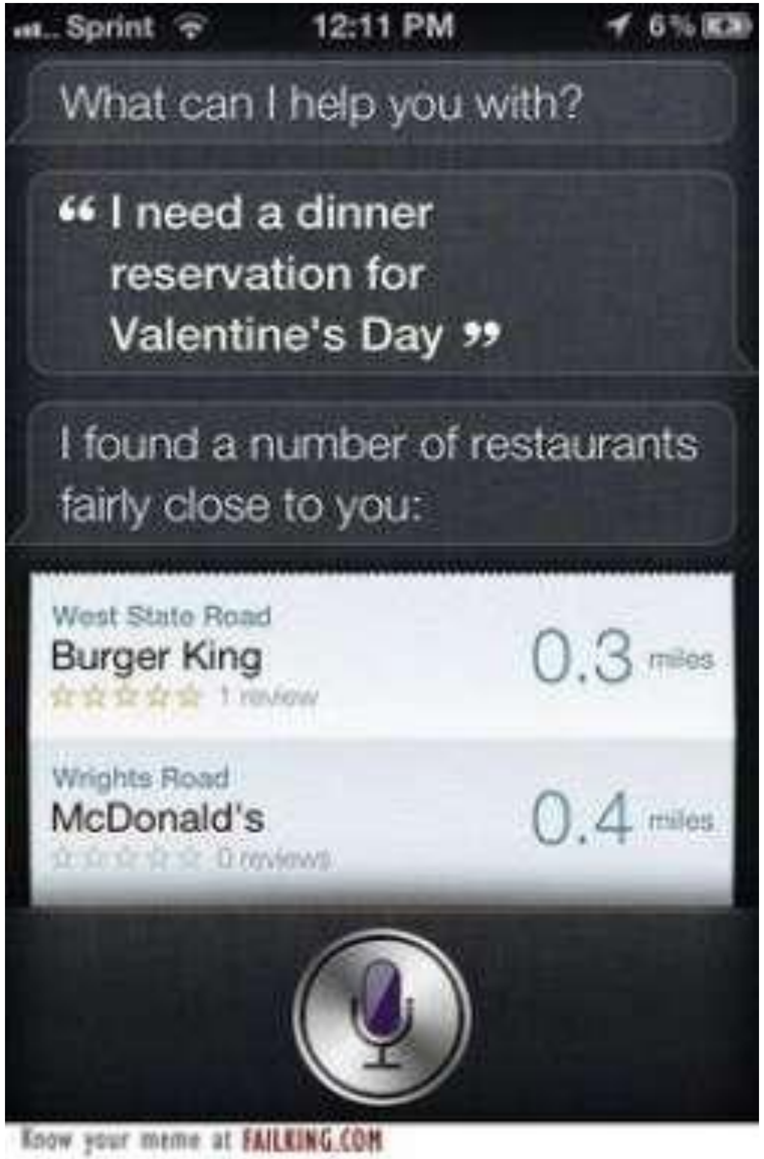
“ I have a gambling problem who should I call ”

I found 6 casinos.

Tap the one you want to call:

Auburn Way South
Muckleshoot Indian...
★★★☆☆ 52 reviews
9.0 miles

Auburn Wy South
Muckleshoot Indian Bi...
★★★★☆ 8 reviews
9.1 miles



What can I help you with?

“ I need a dinner reservation for Valentine's Day ”

I found a number of restaurants fairly close to you:

West State Road
Burger King
☆☆☆☆☆ 1 review
0.3 miles

Wrights Road
McDonald's
☆☆☆☆☆ 0 reviews
0.4 miles

Know your meme at FAILKING.COM

Levels of Understanding

2. Full text comprehension:

- Understanding multi-sentence text and its relation to the "real world".
    - Conversational dialogue.
    - Automatic knowledge acquisition
    - Machine translation?

3. Emotional understanding/generation:

- Responding to literature, poetry, humour
- Story narration.

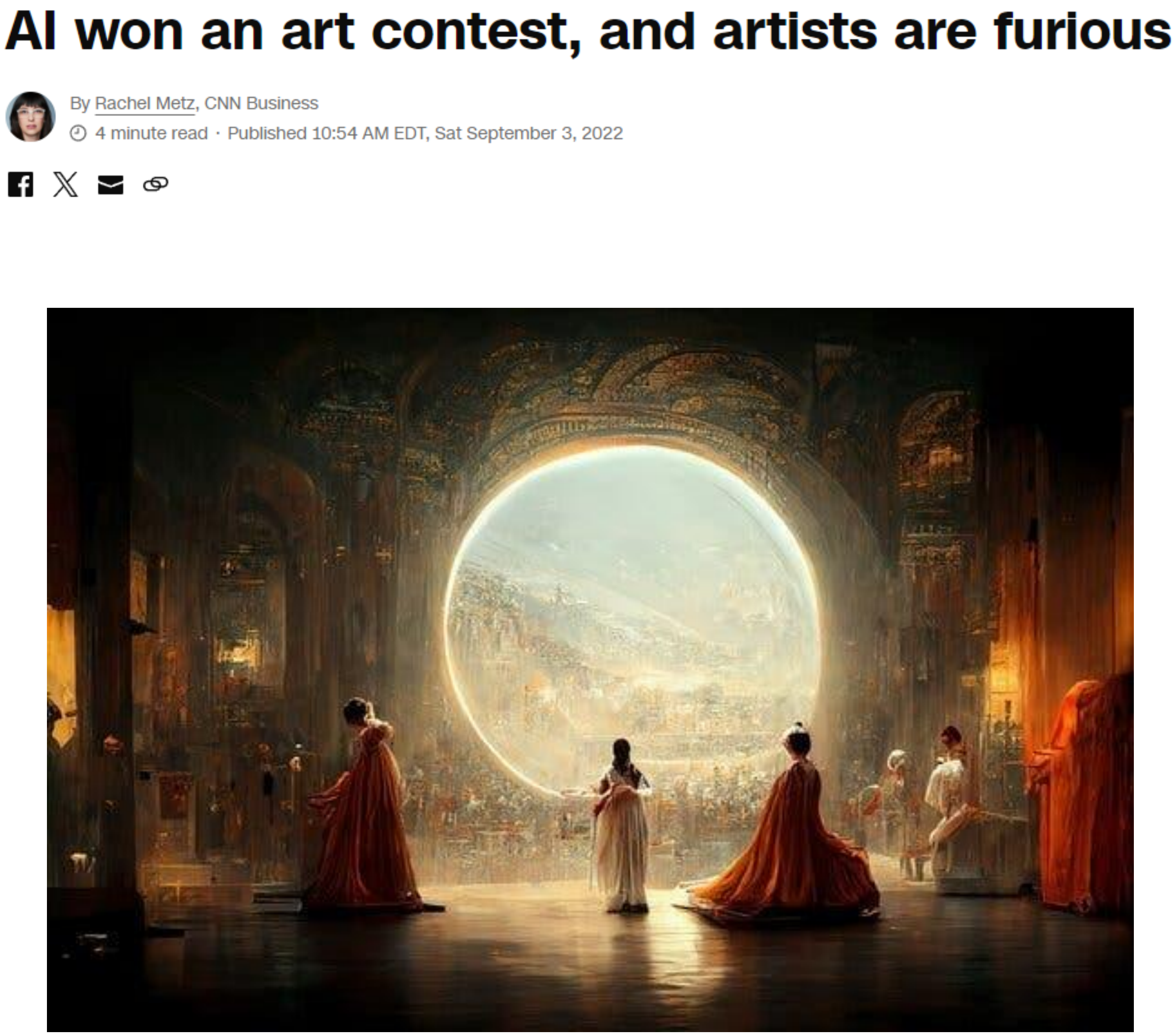
Levels of Understanding

??. ***Full text comprehension:***

- Understanding multi-sentence text and its relation to the “real world”.
  - Conversational dialogue.
  - Automatic knowledge acquisition
  - Machine translation?

??. ***Emotional understanding/generation:***

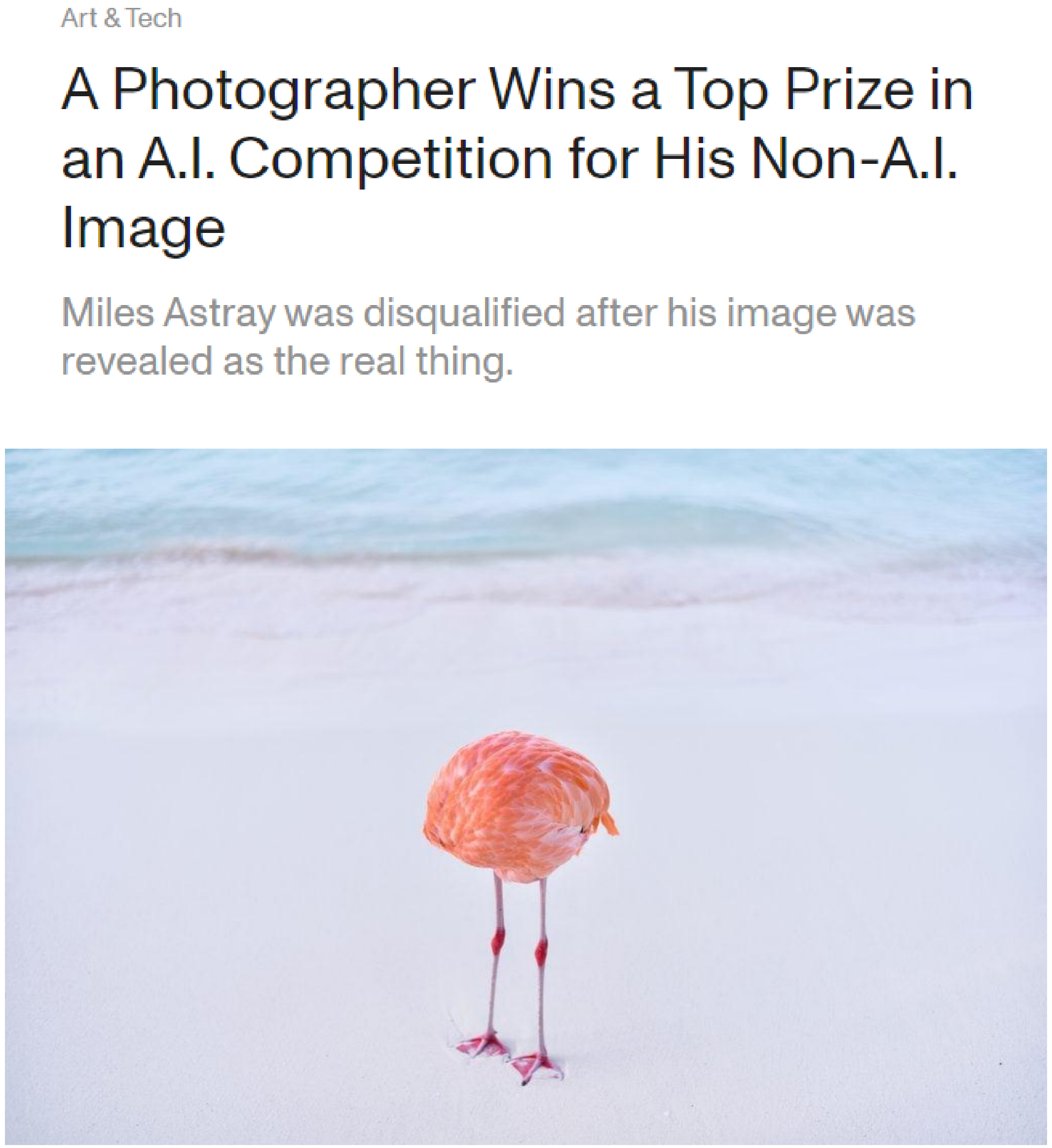
  - Responding to literature, poetry, humour
  - Story narration.



AI won an art contest, and artists are furious

By Rachel Metz, CNN Business

4 minute read · Published 10:54 AM EDT, Sat September 3, 2022



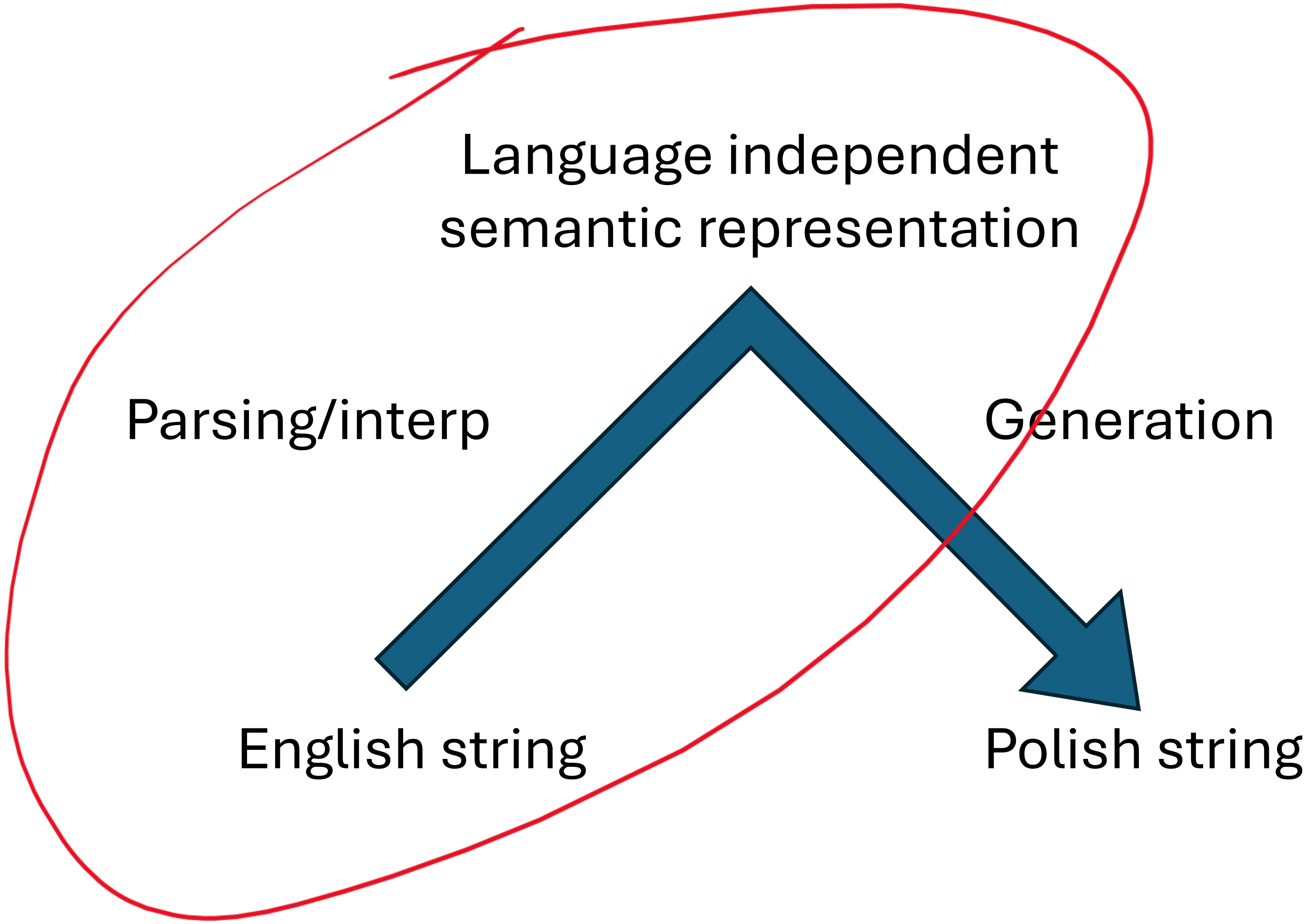
Art & Tech

A Photographer Wins a Top Prize in an A.I. Competition for His Non-A.I. Image

Miles Astray was disqualified after his image was revealed as the real thing.

Natural Language Understanding

Human language
↓
Machine “Language”



Language independent semantic representation

Parsing/interp

Generation

English string

Polish string

Recall: Vauquois triangle.

Information Extraction

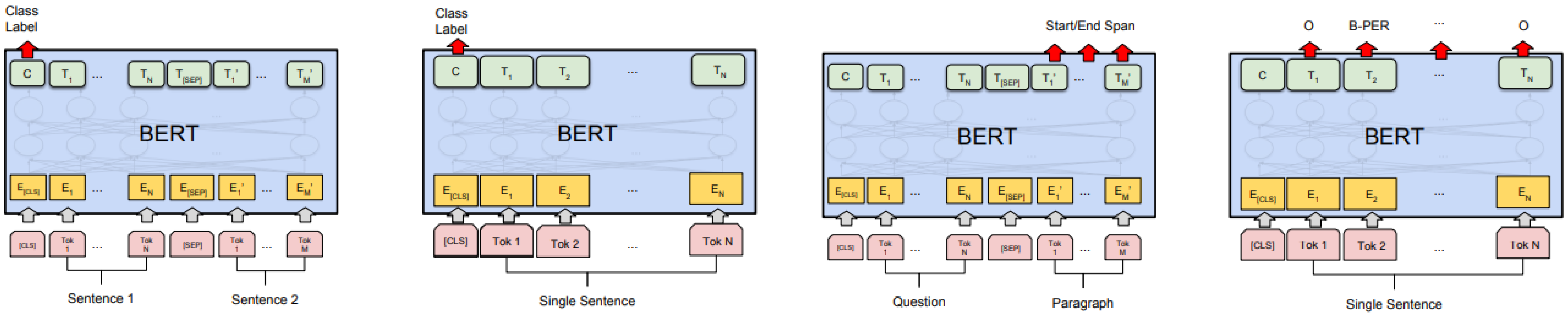
“Bridgestone Sports Co. said Friday it has set up a joint venture in Taiwan with a local concern and a Japanese trading house to produce golf clubs to be shipped to Japan. The joint venture, Bridgestone Sports Taiwan Co., capitalized at 20 million new Taiwan dollars, will start production in January 1990.”

Tie-up-1: *Relation:* Tie-up
Entities: Bridgestone Sports Co.
a local concern
a Japanese trading house
Joint venture: Bridgestone Sports Taiwan Co.
Activity: Activity-1
Amount: NT $ 20,000,000

Activity-1: *Company:* Bridgestone Sports Taiwan Co.
Product: golf clubs
Start date: January 1990

NLU in the Neural Age

- Solution: formulate everything as a classification task.
- Input: word embedding
- Output: … whatever the task requires.



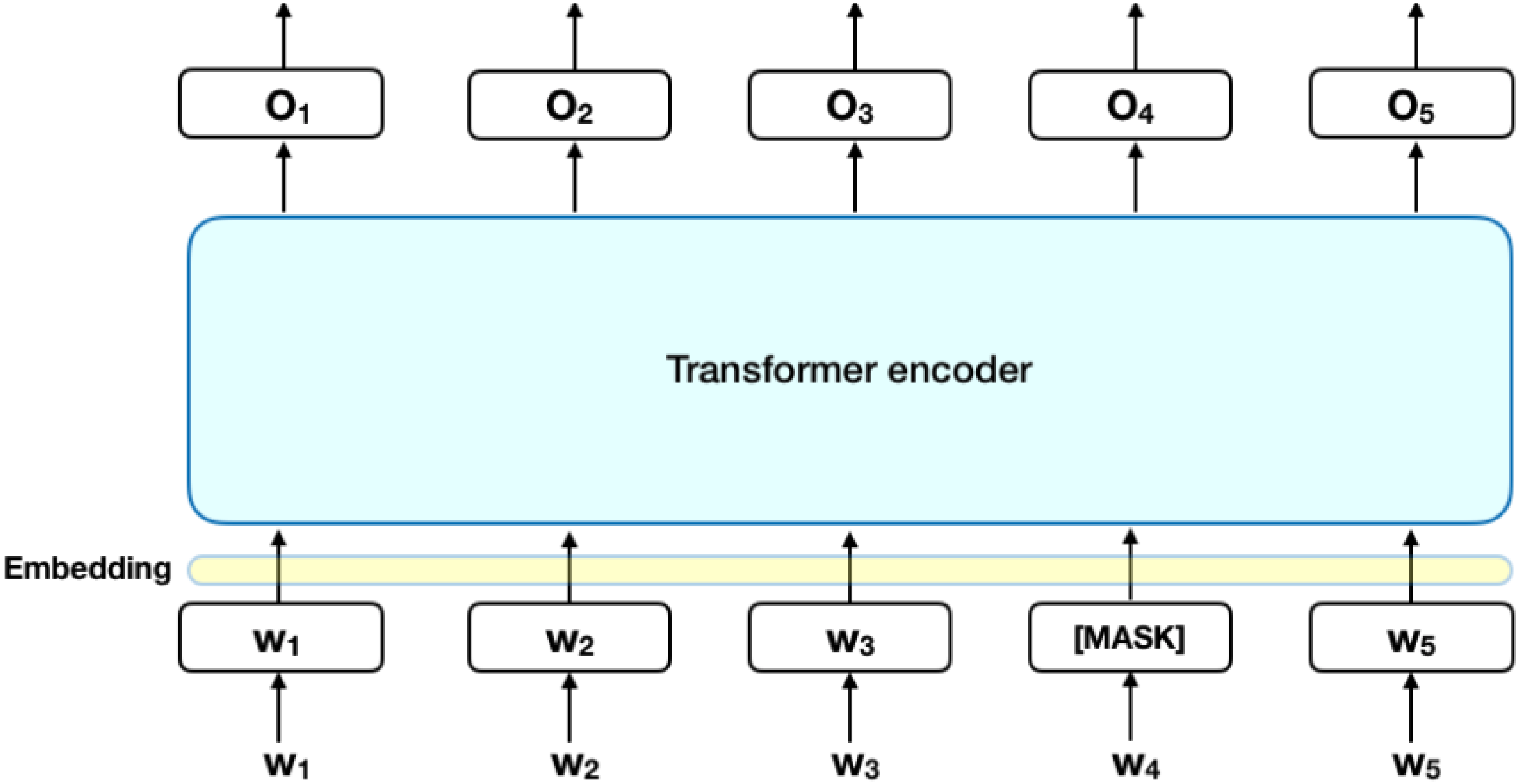
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

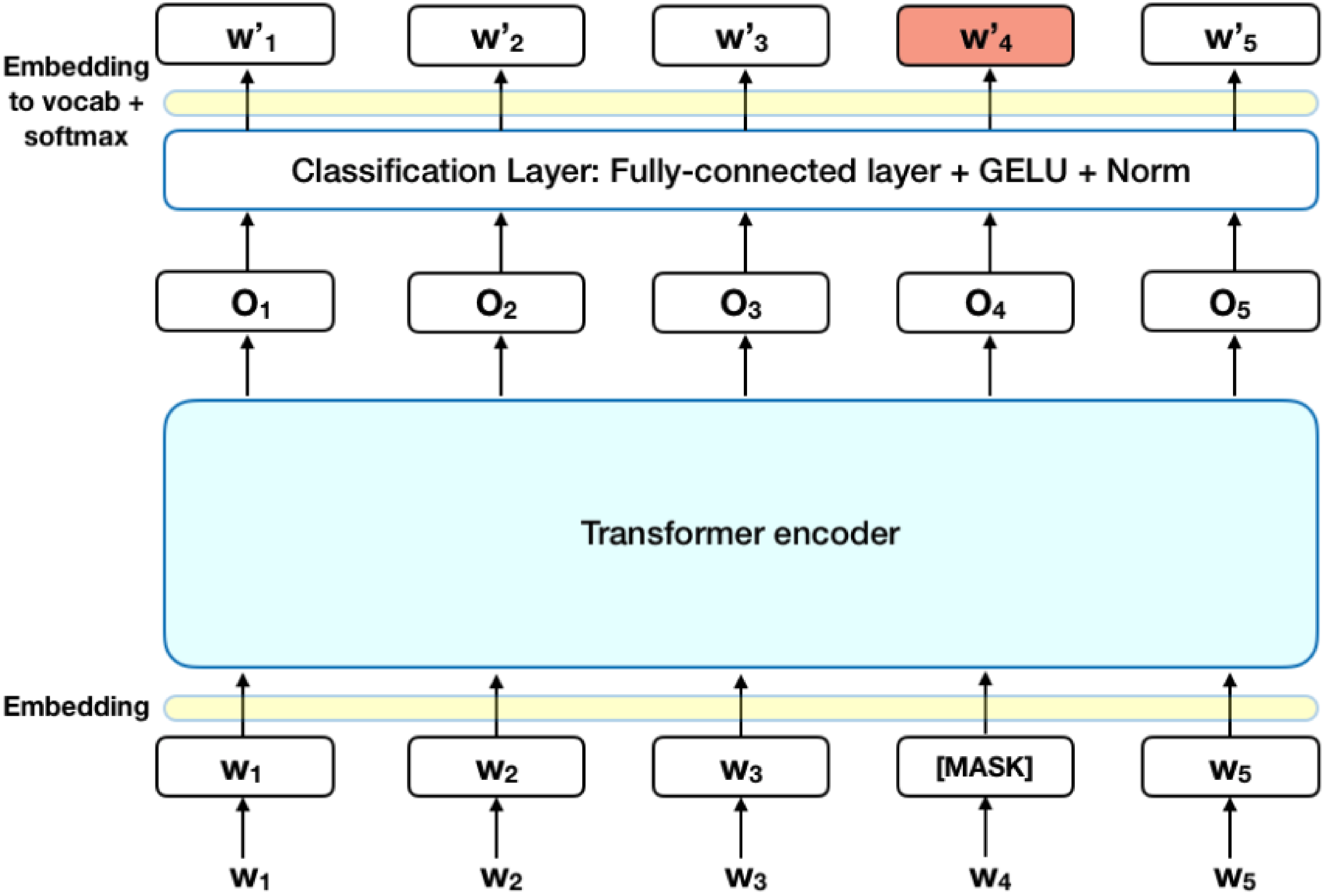
(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

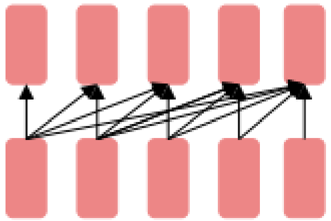
(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

In Practice


O1
O2
O3
O4
O5
Transformer encoder
Embedding
w1
w2
w3
[MASK]
w5
w1
w2
w3
w4
w5

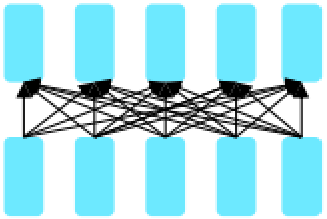

w'1
w'2
w'3
w'4
w'5
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O1
O2
O3
O4
O5
Transformer encoder
Embedding
w1
w2
w3
[MASK]
w5
w1
w2
w3
w4
w5

Which one is the best?



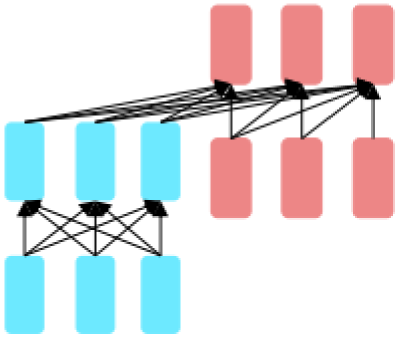
Decoders

- Next word prediction.
- Easy to train. Abundant amount of data.
- Nice to generate from; can’t condition on future words.



Encoders

- Gets bidirectional context – can condition on future!
- Good word embeddings.
- MLM, BERT.



Encoder-Decoders

- Good parts of decoders and encoders?
- What’s the best way to pretrain them?

We will come to more details…

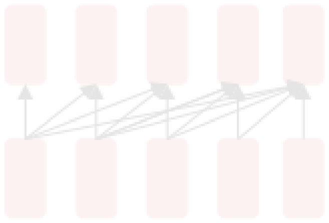
RNN & next word prediction: Not good compositional representation

- Next word prediction:

$$P(t_i|t_1, t_2, \ldots t_{i-1})$$

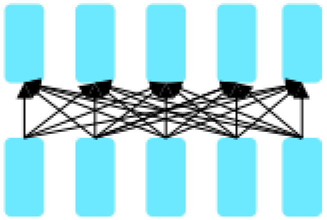
- The hidden state i is encoding information of everything from the beginning (index 0) to the very end (index i).
- We want some bigger semantic units
  - Poilievre-led attempt to **bring down Trudeau minority over carbon tax** fails.
- Some hacks may work, but not really

Which one is the best?



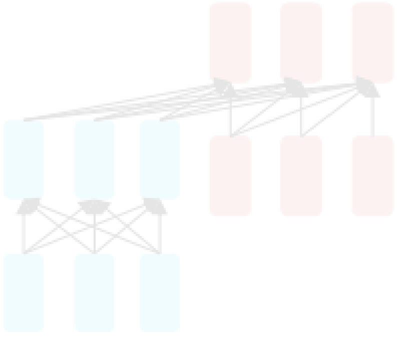
Decoders

- Next word prediction.
- Easy to train. Abundant amount of data.
- Nice to generate from; can’t condition on future words.



Encoders

- Gets bidirectional context – can condition on future!
- Good word embeddings.
- MLM, BERT.



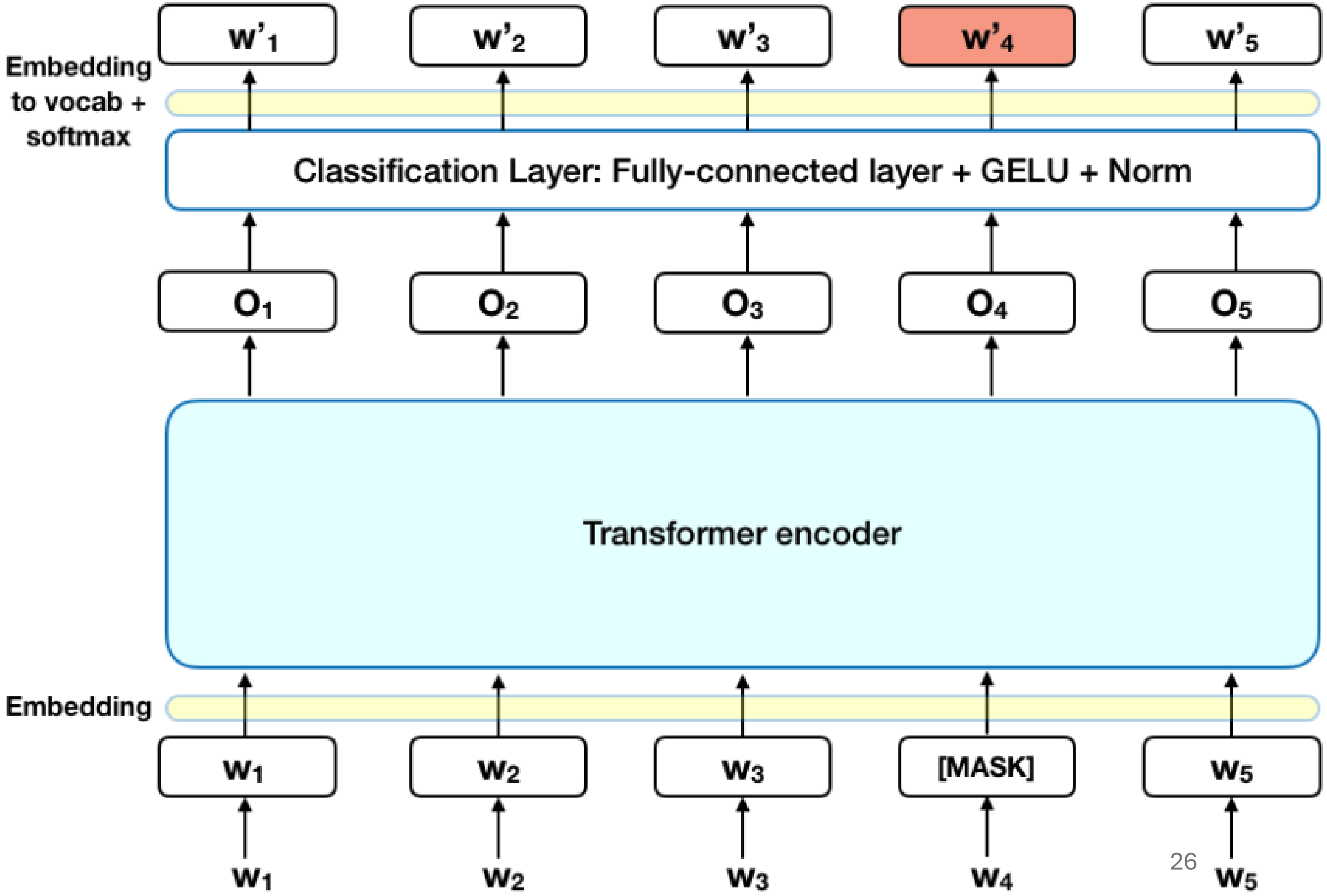
Encoder-Decoders

- Good parts of decoders and encoders?
- What’s the best way to pretrain them?

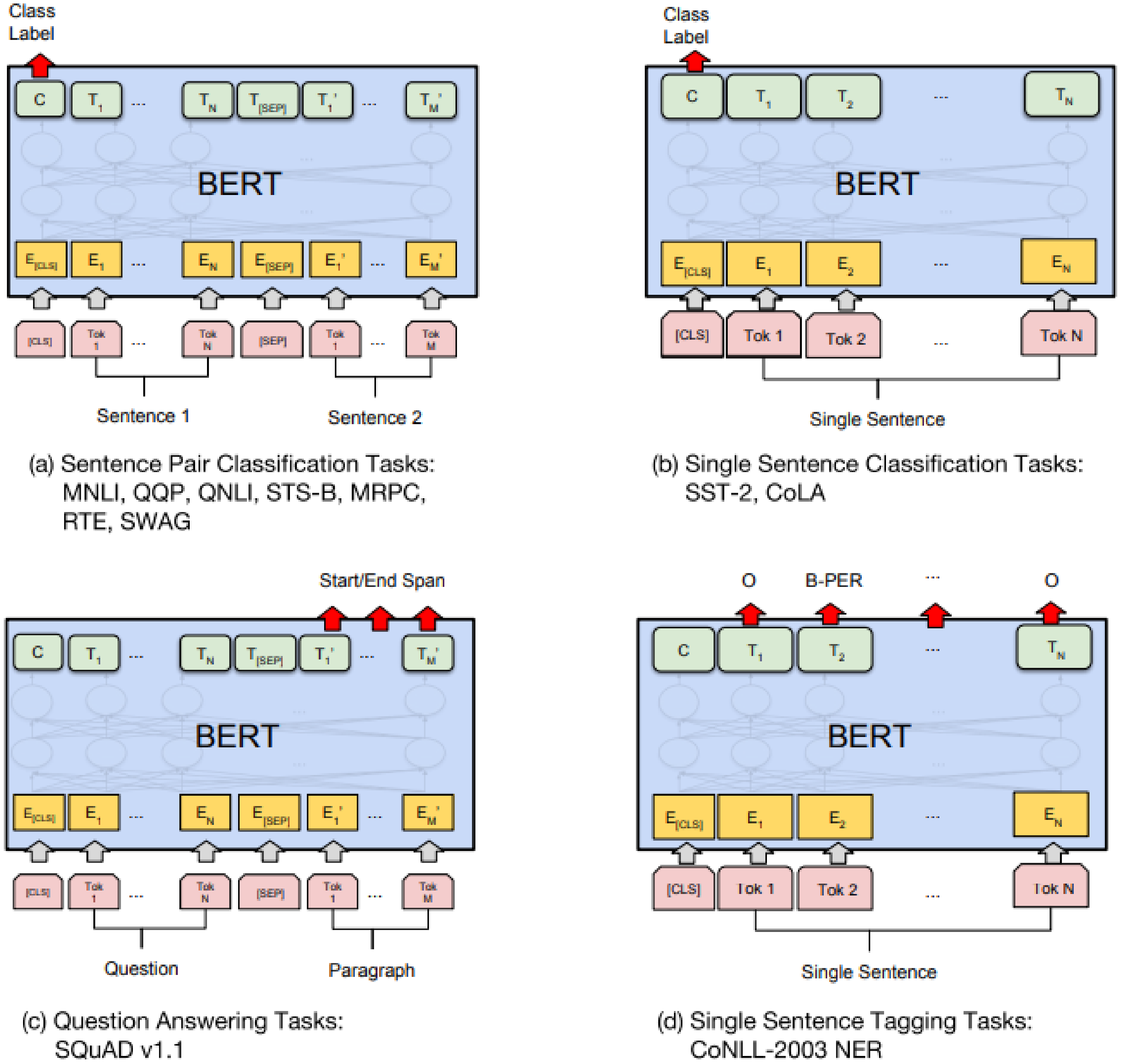
We will come to more details…

BERT – Masked Language Modelling

- Mask 15% of the tokens, and let the model predict it.
- Real easy to do well on MASKed position and nothing else.
- Real easy to learn to copy the context-independent embedding.
- So…
    - 80% of the time: MASK.
    - 10% of the time: correct word.
    - 10% of the time: another random word.


w'₁
w'₂
w'₃
w'₄
w'₅
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O₁
O₂
O₃
O₄
O₅
Transformer encoder
Embedding
w₁
w₂
w₃
[MASK]
w₅
w₁
w₂
w₃
w₄
w₅

Jacob Devlin talks about BERT: https://www.youtube.com/watch?v=knTc-NQSjKA



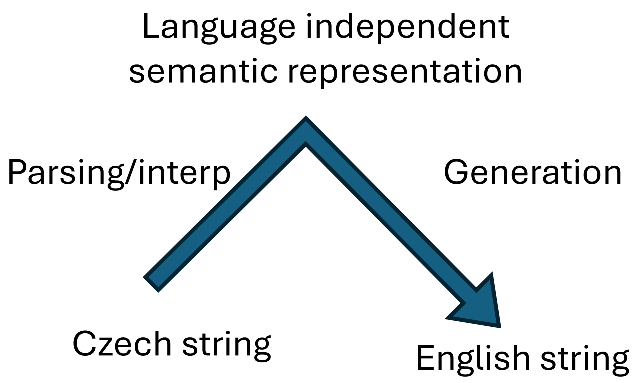
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Approaches to NLU



Language independent semantic representation

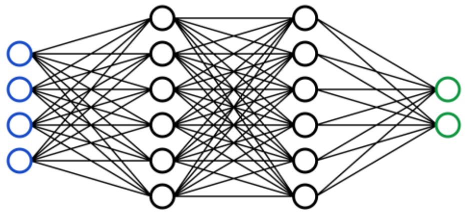
Parsing/interp

Generation

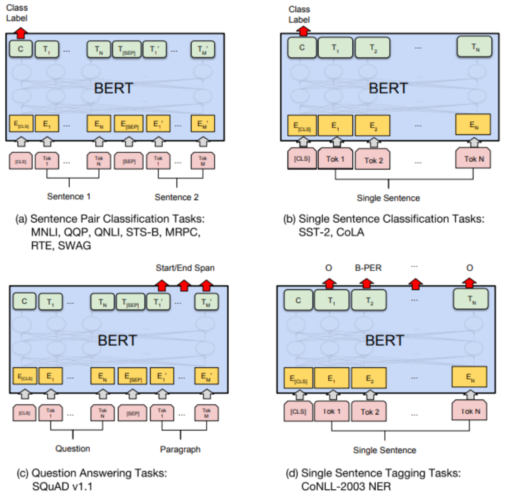
Czech string

English string

Rule-based, symbolic



Statistical models, Typically: neural



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

BERT: Pretrain-finetune paradigm

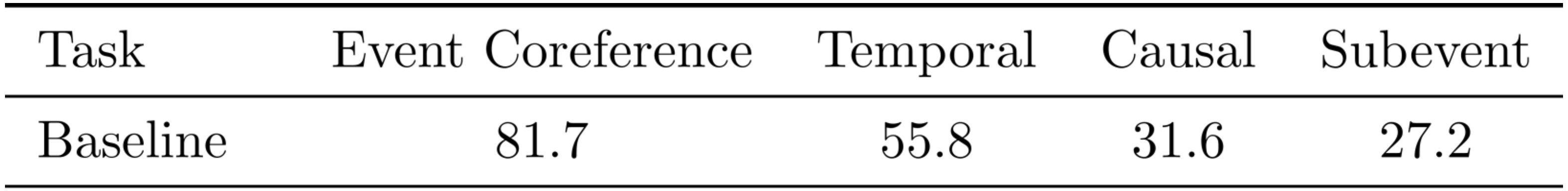
LLM: The future?

LLM has Killed NLP, Right?

Event Coreference

RoBERTa$_{base}$
125 million params

- Fully supervised
- Pretrain-finetune



Task	Event Coreference	Temporal	Causal	Subevent
Baseline	81.7	55.8	31.6	27.2

Wei et al., 2024. Are LLMs Good Annotators for Discourse-level Event Relation Extraction?
https://arxiv.org/pdf/2407.19568

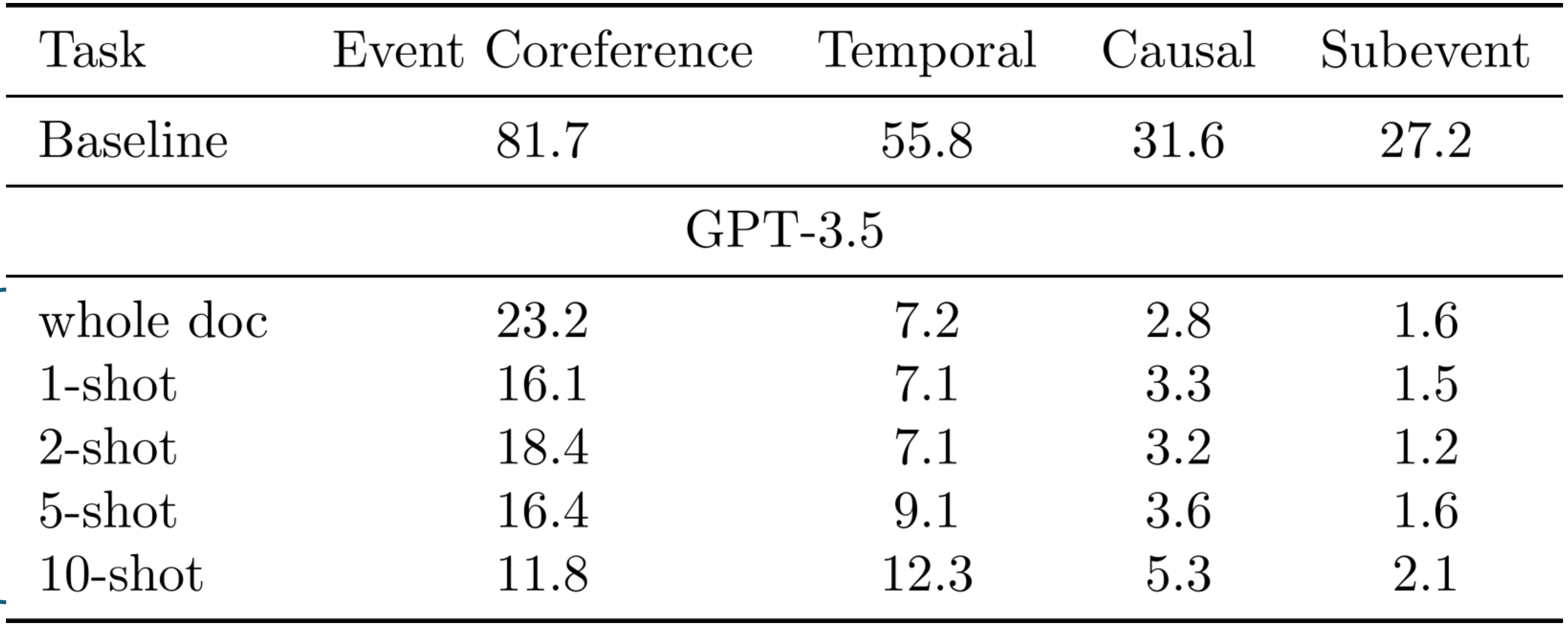
LLM has Killed NLP, Right?

Event Coreference

RoBERTa$_{base}$
125 million params

- Fully supervised
- Pretrain-finetune

GPT-3.5:
175 **billion** params
1000x larger!

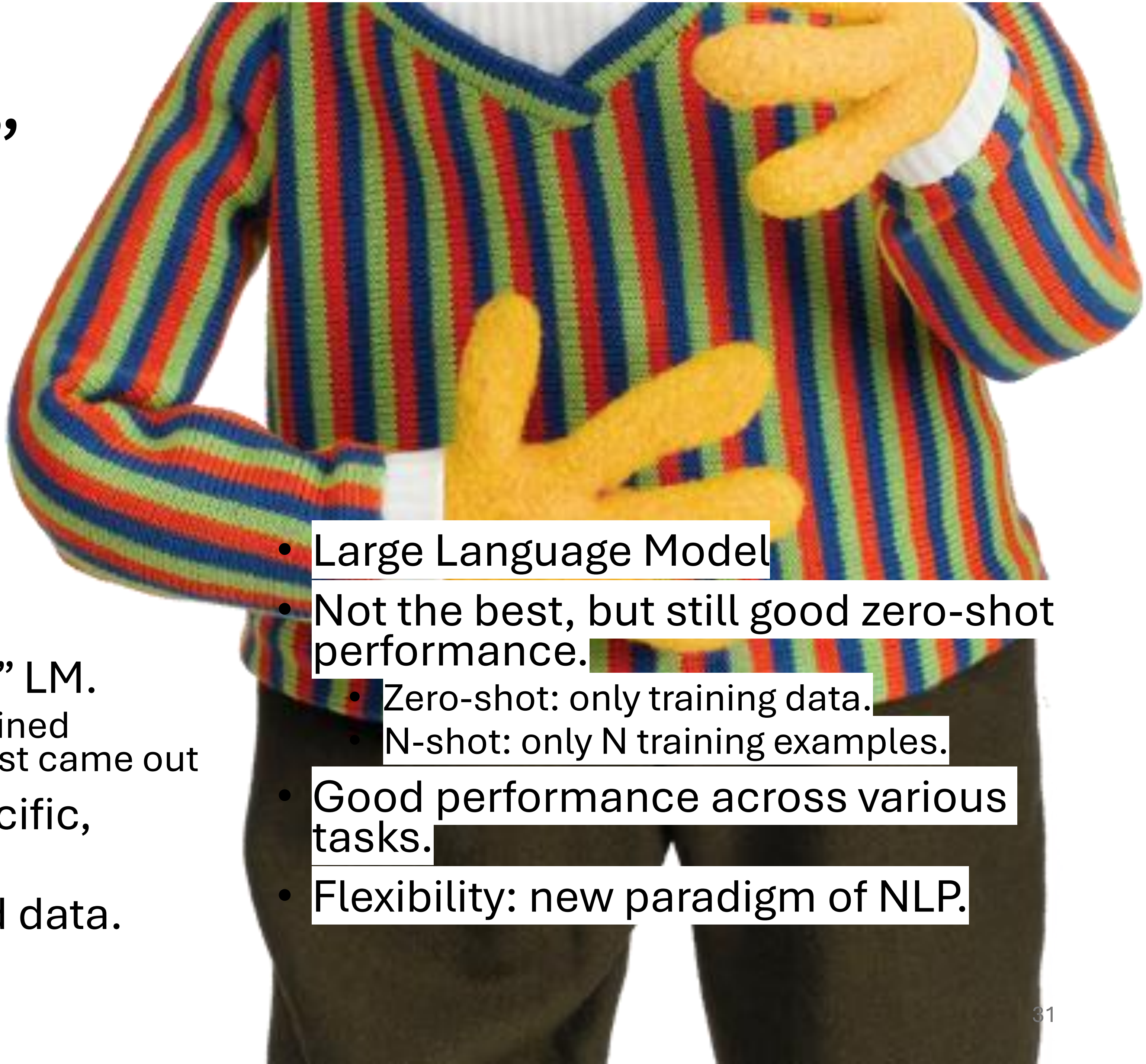


Task	Event Coreference	Temporal	Causal	Subevent
Baseline	81.7	55.8	31.6	27.2
GPT-3.5				
whole doc	23.2	7.2	2.8	1.6
1-shot	16.1	7.1	3.3	1.5
2-shot	18.4	7.1	3.2	1.2
5-shot	16.4	9.1	3.6	1.6
10-shot	11.8	12.3	5.3	2.1

Wei et al., 2024. Are LLMs Good Annotators for Discourse-level Event Relation Extraction?
https://arxiv.org/pdf/2407.19568

“SLM” vs. “LLM”

- Pretrain-finetuned “small” LM.
  - Called **large-scale** pre-trained language model when it first came out
- Best performance on specific, atomic tasks.
- Inflexible, require labelled data.



- Large Language Model
- Not the best, but still good zero-shot performance.
  - Zero-shot: only training data.
  - N-shot: only N training examples.
- Good performance across various tasks.
- Flexibility: new paradigm of NLP.