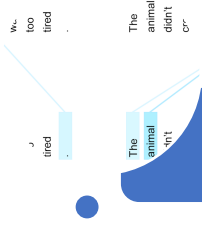
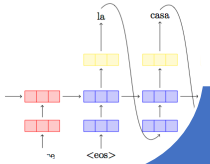
# Transformer-based language models

# Where are we?

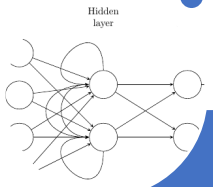
Pretrained Language Model (ULMFiT, BERT, XLNet, GPT)



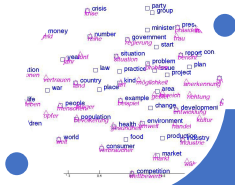
Attention, self-attention



Sequence 2 Sequence



Recurrent network (RNN, LSTM, GRU)



Word Embedding

Bag of word, TF-IDF

# What is Language Modeling?

- Probability distribution over strings or text.
- S = “Tôi đi học”
- P(S) = P(“Tôi”) x P(“đi”|”Tôi”) x P(“học”|”Tôi đi”)
- P(S) > P(“đi học Tôi”)

ELMo

ULMFiT

Transformer

BERT

GPT

Bidirectional LM

Larger model
More data

GPT-2

Defense

Grover

Cross-lingual

Multi-task
+ Generation

XLM
UDify

MT-DNN

Knowledge distillation

$MT\text{-}DNN_{KD}$

MASS
UniLM

Span prediction
Remove NSP

SpanBERT

Longer time
Remove NSP
More data

RoBERTa

Permutation LM
Transformer-XL
More data

XLNet

+Knowledge Graph

ERNIE
(Tsinghua)

Neural entity linker

KnowBert

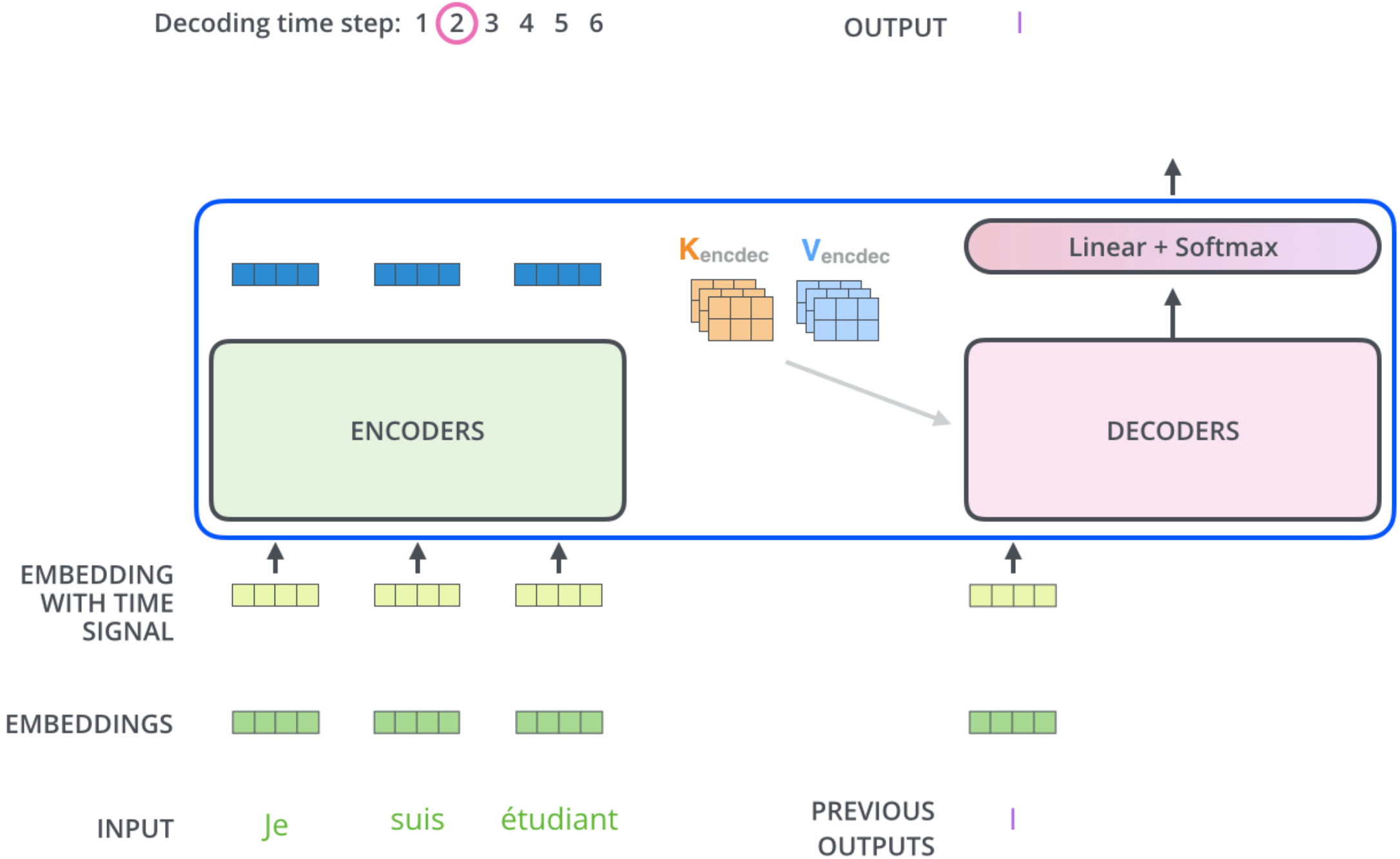
Cross-modal

VideoBERT
CBT
ViLBERT
VisualBERT
B2T2
Unicoder-VL
LXMBERT
VL-BERT

Whole Word Masking

ERNIE (Baidu)
BERT-wwm

By Xiaozhi Wang & Zhengyan Zhang @THUNLP


Decoding time step: 1 2 3 4 5 6
OUTPUT
I
Linear + Softmax
K encdec
V encdec
ENCODERS
DECODERS
EMBEDDING WITH TIME SIGNAL
EMBEDDINGS
INPUT
Je
suis
étudiant
PREVIOUS OUTPUTS
I

# BERT

- **B**idirectional **E**ncoder **R**epresentations from **T**ransformers.
- Use the Transformer Encoder architecture.
- Introduced in 2018 by Google AI.

unsupervised

## 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.

### Semi-supervised Learning Step

**Model:**


**Dataset:**

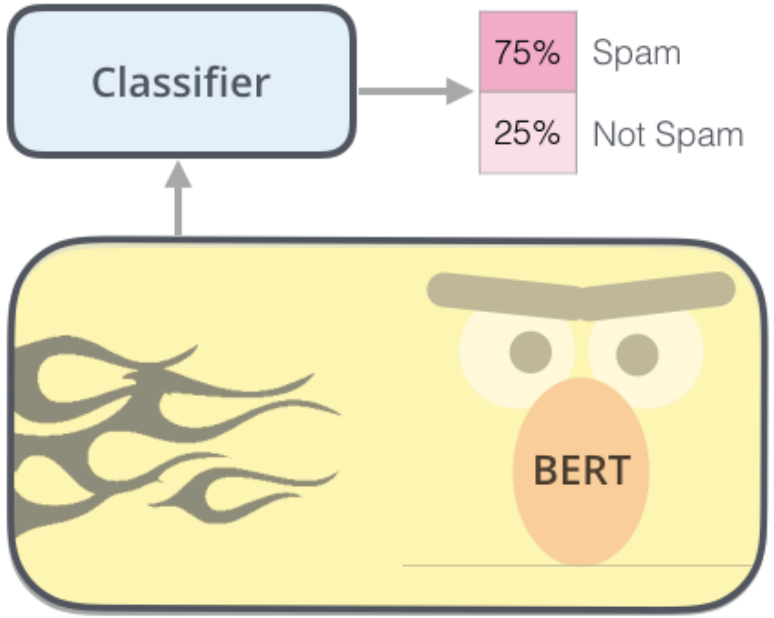

**Objective:**

Predict the masked word (langauge modeling)

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

### Supervised Learning Step

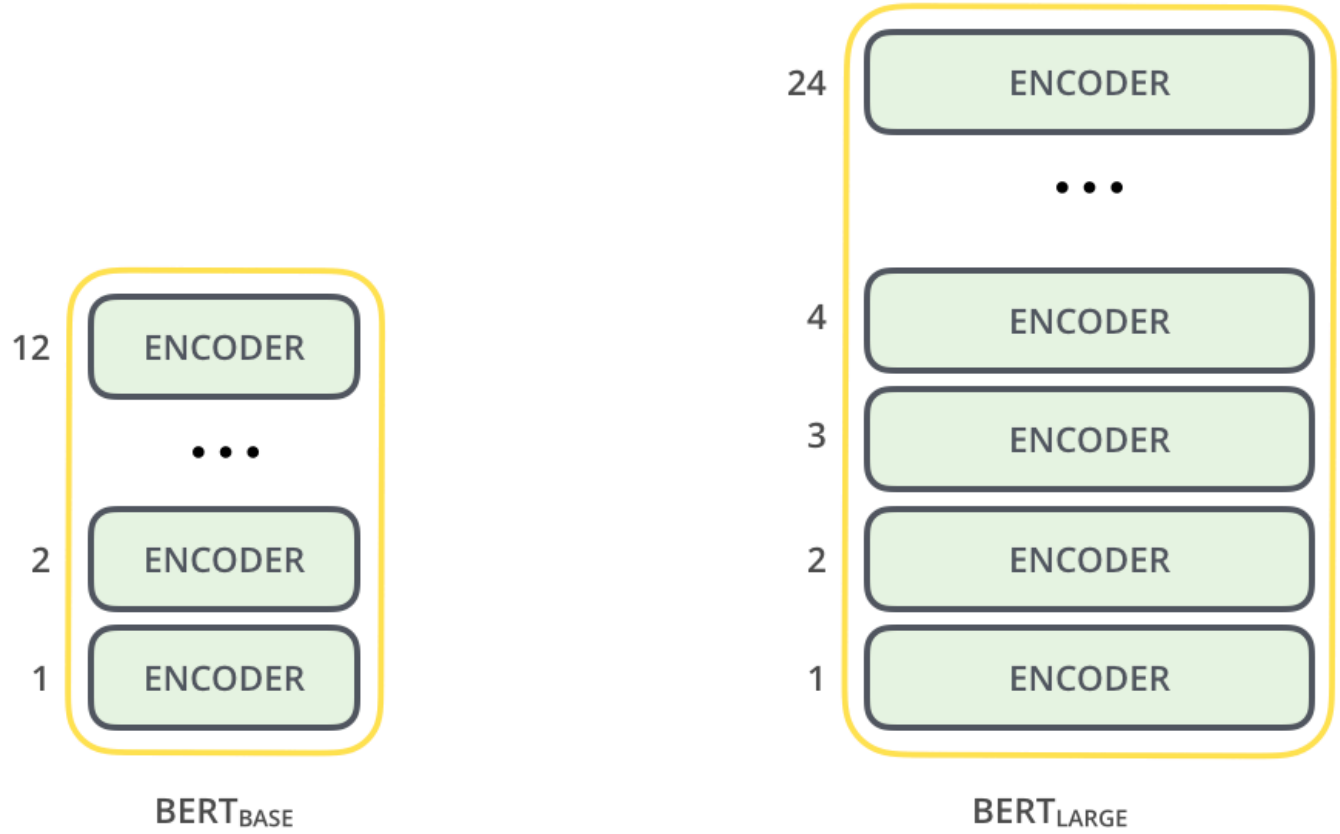
**Model:** (pre-trained in step #1)




**Dataset:**

| Email message | Class |
|---|---|
| Buy these pills | Spam |
| Win cash prizes | Spam |
| Dear Mr. Atreides, please find attached… | Not Spam |

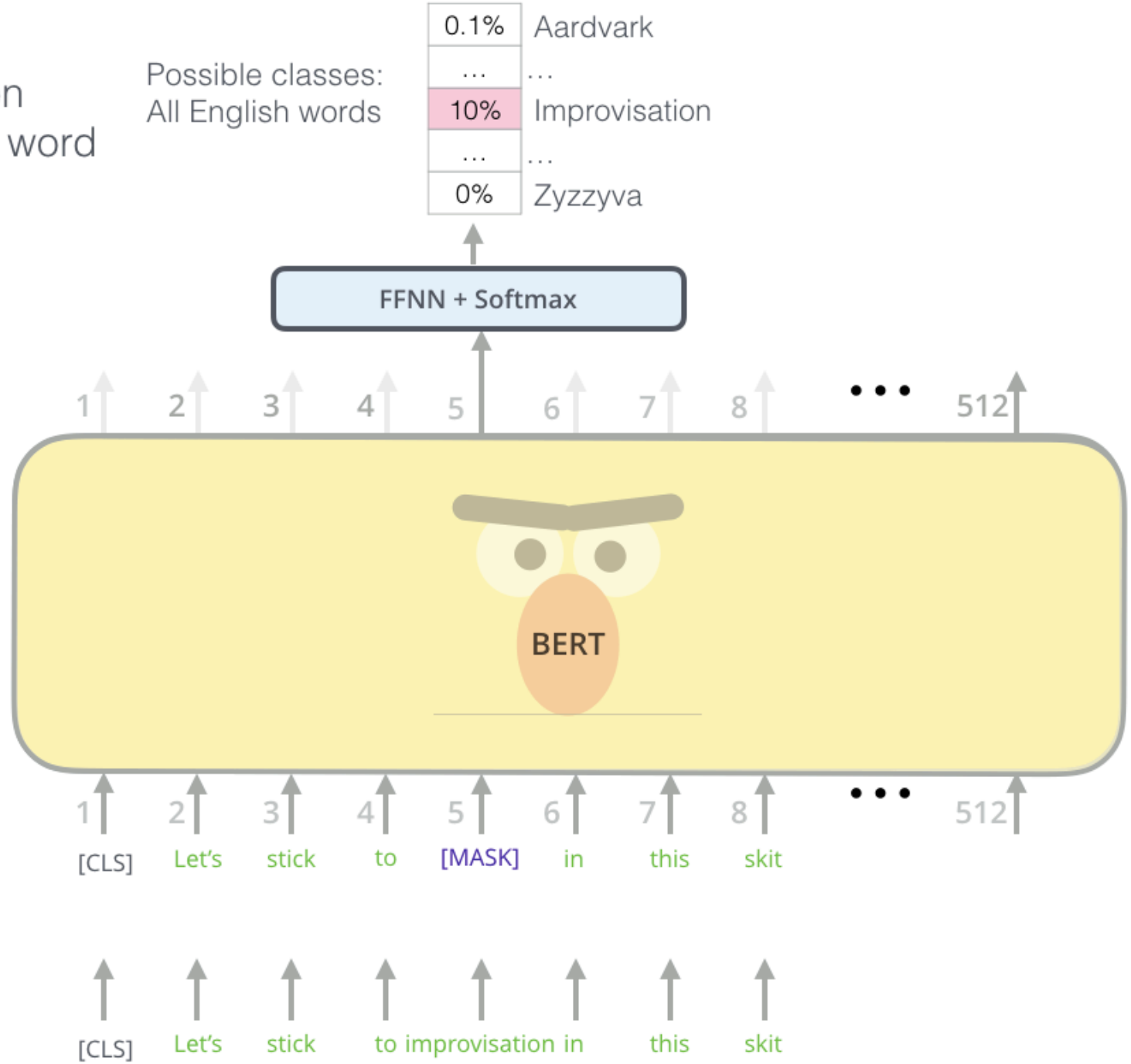
# Architecture



# Pretraining

- Two unsupervised tasks:
    1. Masked Language Model
    2. Next Sentence Prediction

Use the output of the masked word's position to predict the masked word

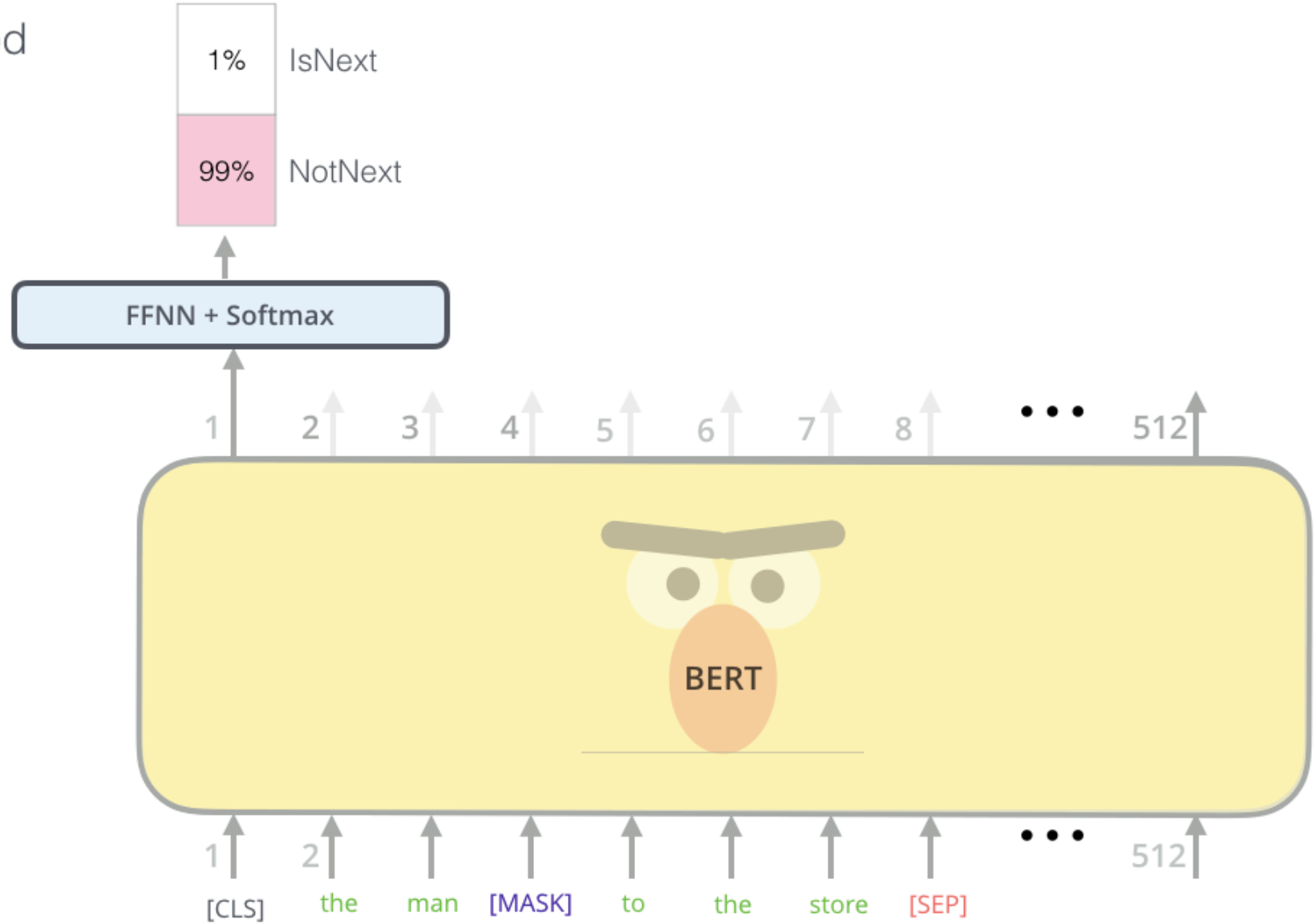



Randomly mask 15% of tokens

Input

[CLS] Let's stick to improvisation in this skit

Predict likelihood that sentence B belongs after sentence A




Tokenized Input

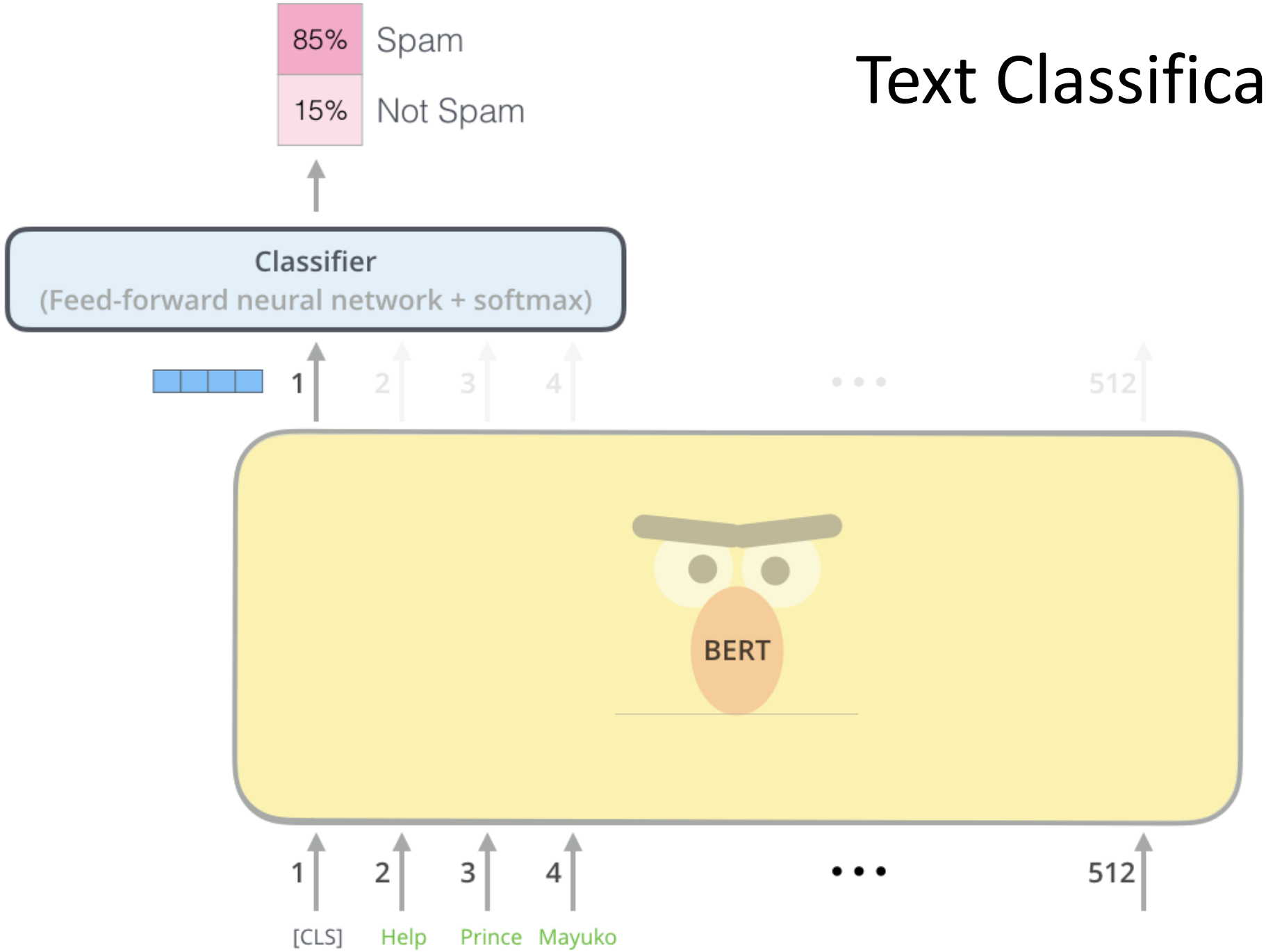
Input

[CLS] the man [MASK] to the store [SEP] penguin [MASK] are flightless birds [SEP]

Sentence A

Sentence B

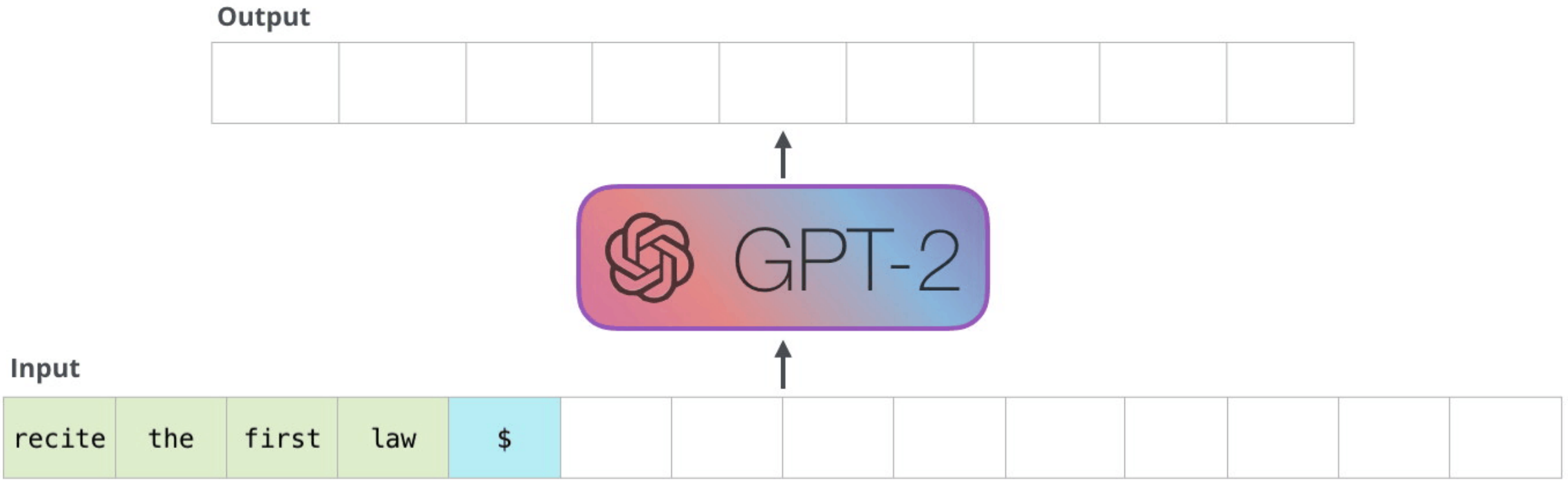
# Text Classification



# GPT

- Generative Pre-trained Transformer
- Use the Transformer Decoder architecture.
- Introduced in 2018 by OpenAI.

# How it works?



# Evolution of GPT

| Model | Number of parameters | Training data size | Year |
|---|---|---|---|
| GPT | 110M | 4GB | 2018 |
| GPT-2 | 1.5B | 40GB | 2019 |
| GPT-3 | 175B | ≈2TB | 2020 |

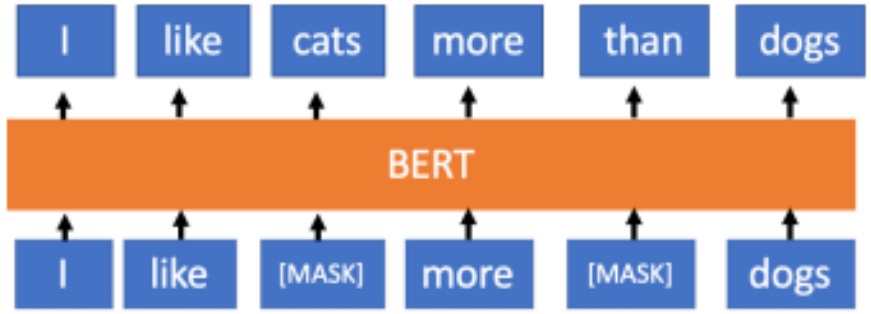
# XLNet

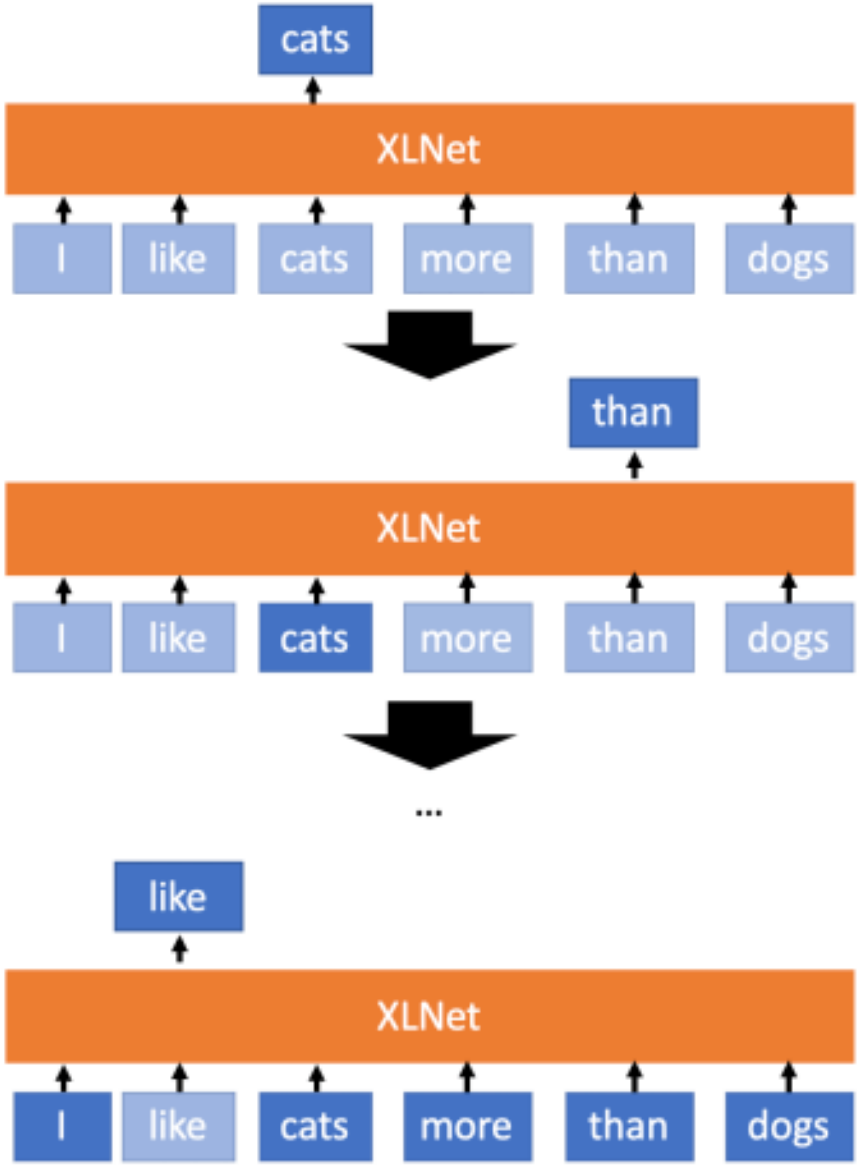
- Autoencoding (BERT):
  - [MASK] tokens do not appear during finetuning ⇒ pretrain-finetuning discrepancy.
  - Assume the predicted tokens are independent of each other given the unmasked tokens. Example: “New York is a city” ⇒ “[MASK] [MASK] is a city”
- Autoregressive (GPT):
  - Only trained to encode a unidirectional context (forward or backward).

# XLNet

- XLNet combines pros from both while avoiding their cons.
- Techniques:
  - Permutation Language Modeling
  - Two-Stream Self-Attention for Target-Aware Representations
  - Incorporating Ideas from Transformer-XL
  - Modeling Multiple Segments

# Permutation Language Modeling



ELMo

ULMFiT

Transformer

BERT

GPT

Bidirectional LM

Larger model
More data

GPT-2

Defense

Grover

Cross-lingual

Multi-task
+ Generation

XLM
UDify

MT-DNN

Knowledge distillation

$MT\text{-}DNN_{KD}$

MASS
UniLM

Span prediction
Remove NSP

SpanBERT

Longer time
Remove NSP
More data

RoBERTa

Permutation LM
Transformer-XL
More data

XLNet

+Knowledge Graph

ERNIE
(Tsinghua)

Neural entity linker

KnowBert

Cross-modal

VideoBERT
CBT
ViLBERT
VisualBERT
B2T2
Unicoder-VL
LXMBERT
VL-BERT

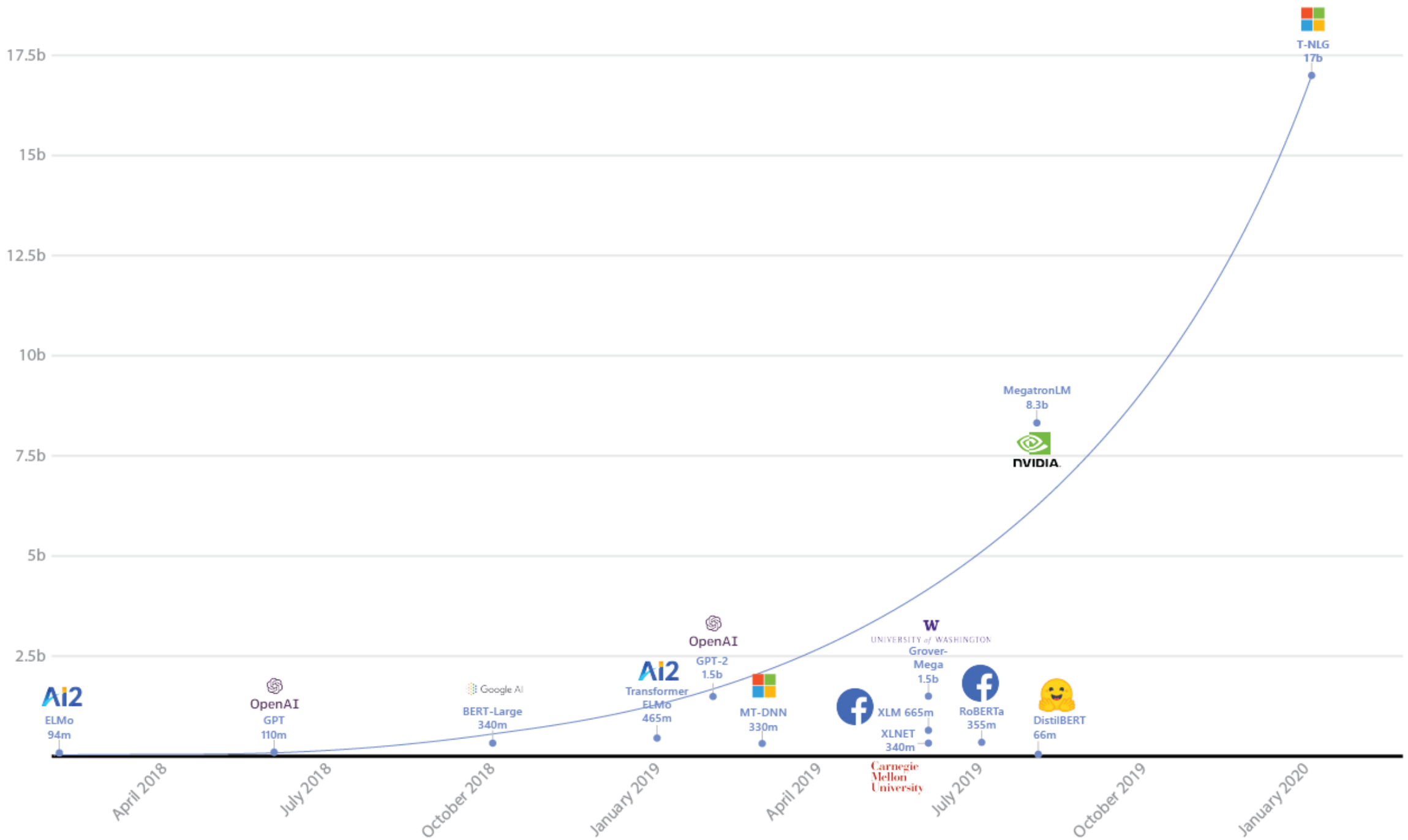
Whole Word Masking

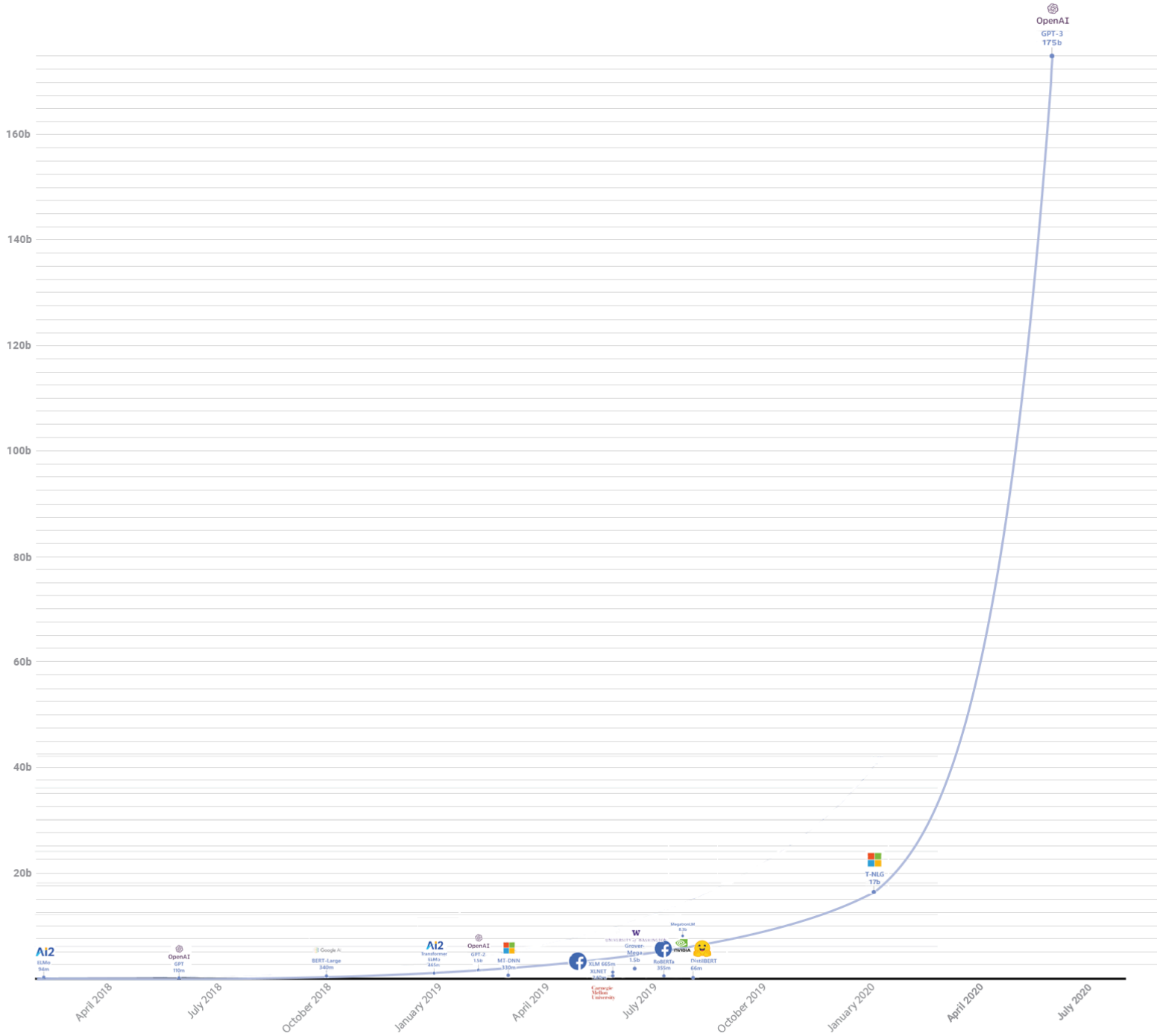
ERNIE (Baidu)
BERT-wwm

By Xiaozhi Wang & Zhengyan Zhang @THUNLP

# The GPT-3 Hype

- 175 billion parameters => 700 GB of memory!
- 10x larger than the 2nd largest model (Turing-NLG).


17.5b
15b
12.5b
10b
7.5b
5b
2.5b
T-NLG
17b
MegatronLM
8.3b
NVIDIA
OpenAI
GPT-2
1.5b
Ai2
Transformer
ELMo
465m
Google AI
BERT-Large
340m
Ai2
ELMo
94m
OpenAI
GPT
110m
MT-DNN
330m
UNIVERSITY of WASHINGTON
Grover-
Mega
1.5b
XLM 665m
XLNET
340m
Carnegie
Mellon
University
RoBERTa
355m
DistilBERT
66m
April 2018
July 2018
October 2018
January 2019
April 2019
July 2019
October 2019
January 2020


OpenAI
GPT-3
175b
160b
140b
120b
100b
80b
60b
40b
20b
T-NLG
17b
April 2018
July 2018
October 2018
January 2019
April 2019
July 2019
October 2019
January 2020
April 2020
July 2020

# The GPT-3 Hype

**Zero-shot**

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

**One-shot**

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

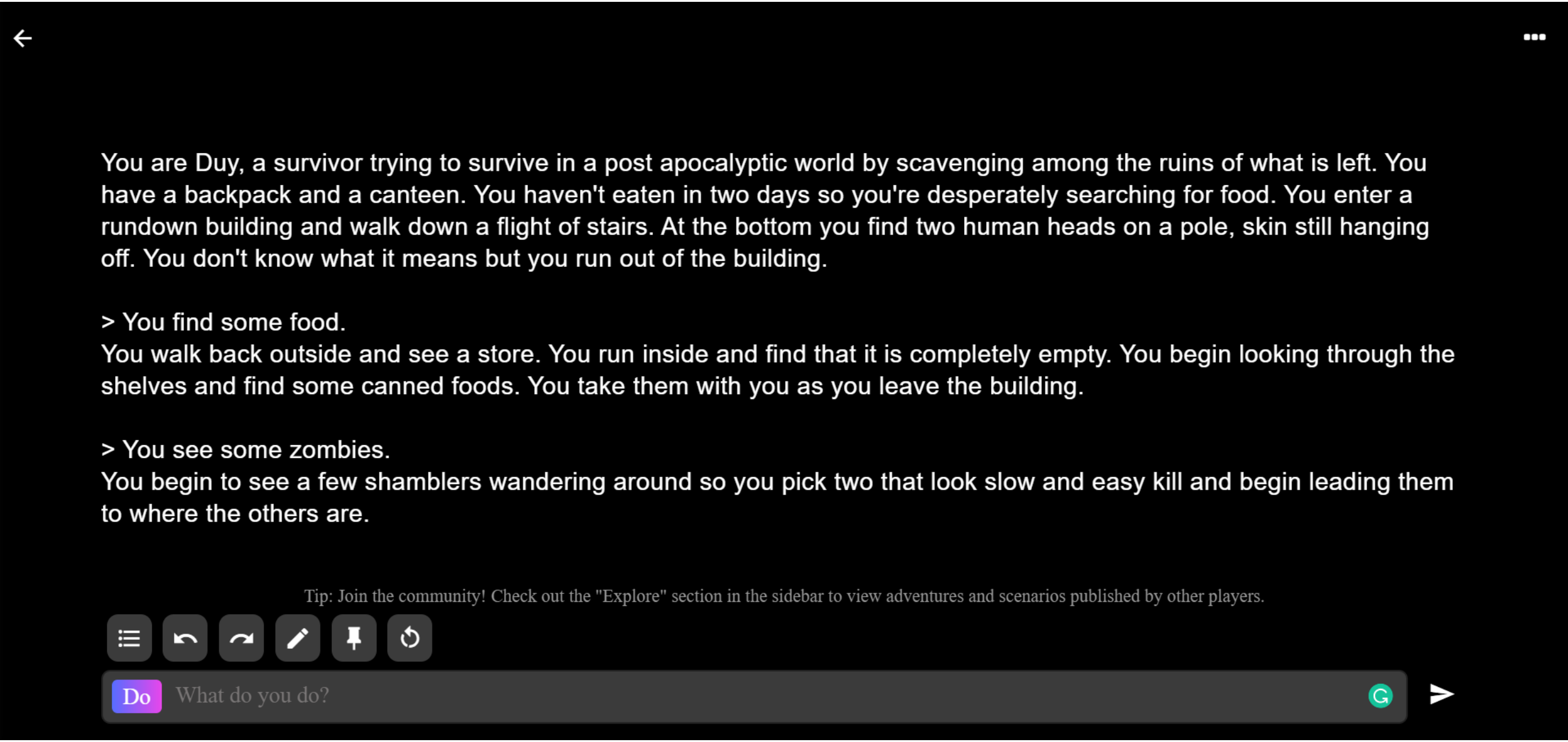
```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

**Few-shot**

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```

# Applications: AI Dungeon (link)



# Limitations

- Computation
- Size
- Transparency
- Can’t capture the actual meaning of words. From GPT-3 Authors:

*“GPT-3 samples still sometimes repeat themselves semantically at the document level, start to lose coherence over sufficiently long passages, contradict themselves, and occasionally contain non-sequitur sentences or paragraphs.”*

*Human: How many eyes does a giraffe have?*
*GPT-3: A giraffe has two eyes.*

*Human: How many eyes does my foot have?*
*GPT-3: Your foot has two eyes.*

*Human: How many eyes does a spider have?*
*GPT-3: A spider has eight eyes.*

*Human: How many eyes does the sun have?*
*GPT-3: The sun has one eye.*

# What’s next?

- Human brain: **100 trillion*** synapses
  - * Could be 1000+ trillion. ANN != BNN.
- GPT-3: **175 billion** parameters
  - Cost: $4.6m
- GPT-4?: **100 trillion** parameters
  - Cost (2020): $2.6 billion
  - Cost (2024): $325 million
  - Cost (2028): $40 million
  - Cost (2032): $5 million

# Reference


- Devlin, J. et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” NAACL-HLT (2019).
- Yang, Z. et al. “XLNet: Generalized Autoregressive Pretraining for Language Understanding.” NeurIPS (2019).
- Brown, T. et al. “Language Models are Few-Shot Learners.” ArXiv abs/2005.14165 (2020): n. pag.
- Lex Fridman. GPT-3 vs Human Brain. https://www.youtube.com/watch?v=kpiY_LemaTc
- Illustrations: https://jalammar.github.io/