

# NLP Tutorial: from Perceptron to Language Models

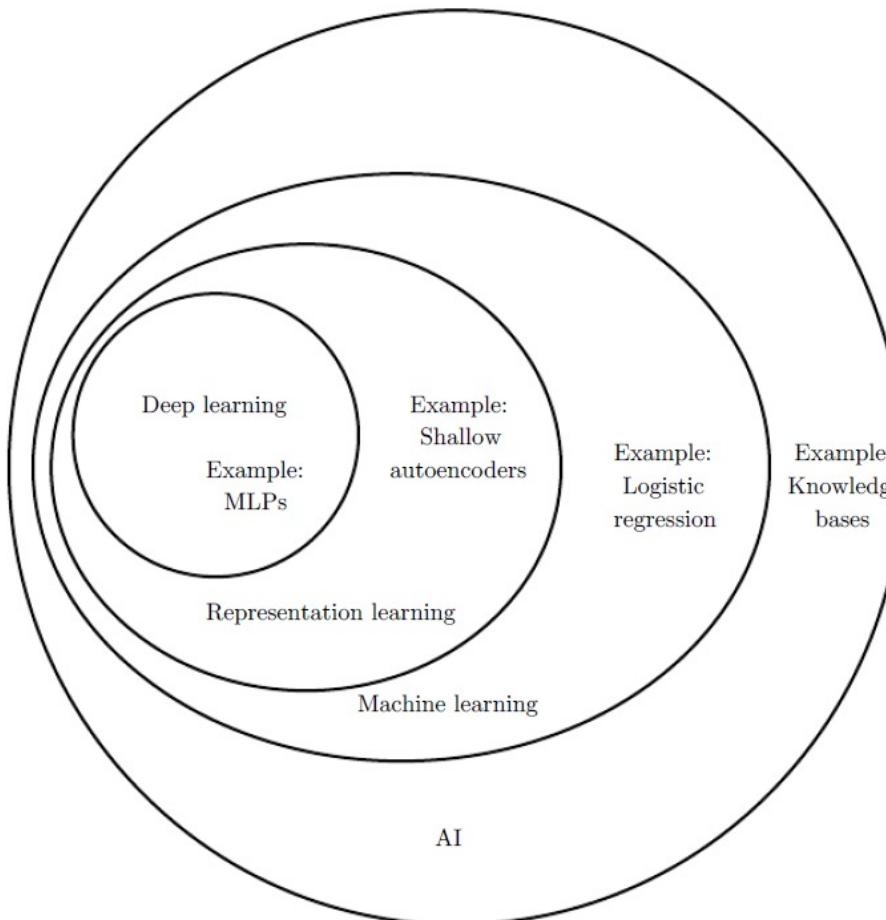
Yonsei AI

Jinyoung Yeo

# Outline

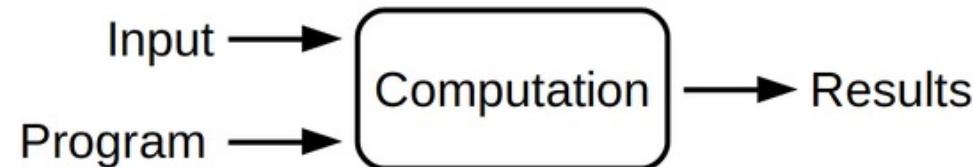
- Machine Learning
- Sequential Modeling
- Transformer and Language Models
- Dialogue Models

# Relationships between DL, representation learning, ML, and AI

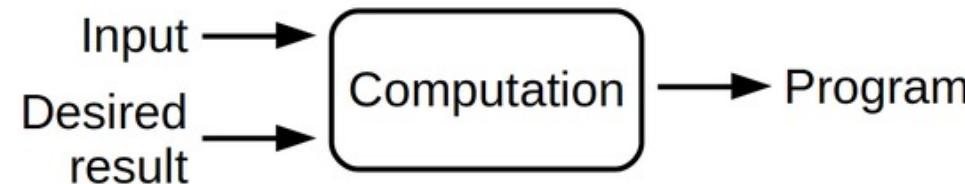


# Machine Learning

## Traditional programming



## Machine learning



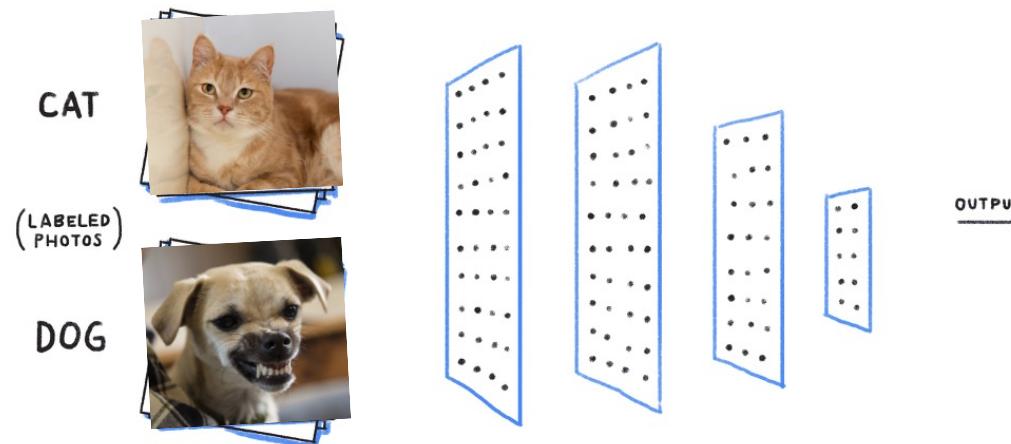
# Task and Experience: Training Data

- Everything starts with a dataset!
- Inputs:  $x = \{x_1, x_2, x_3, \dots, x_N\}$
- Output labels:  $y = \{y_1, y_2, y_3, \dots, y_N\}$
- Dataset D =  $\{(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_N, y_N)\}$



# Program (Model)

- Define a function  $f_W$  to predict a corresponding output  $\hat{y}_i$  for a given input  $x_i$ .
- $\hat{y}_i = f(x_i; W)$  or  $\hat{y} = f_W(x_i)$



# Performance Measure: Loss

- Find the optimized parameters  $W$  to map  $x$  into  $y$ .
- Loss  $L(f(x, W), y) = L(\hat{y}, y)$
- Assumption:  $L(\hat{y}, y) = \sum_i L(\hat{y}_i, y_i)$

Mean Squared Error (MSE)

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

- \*  $n$  is the number of data points
- \*  $Y_i$  represents observed values
- \*  $\hat{Y}_i$  represents predicted values

Cross Entropy

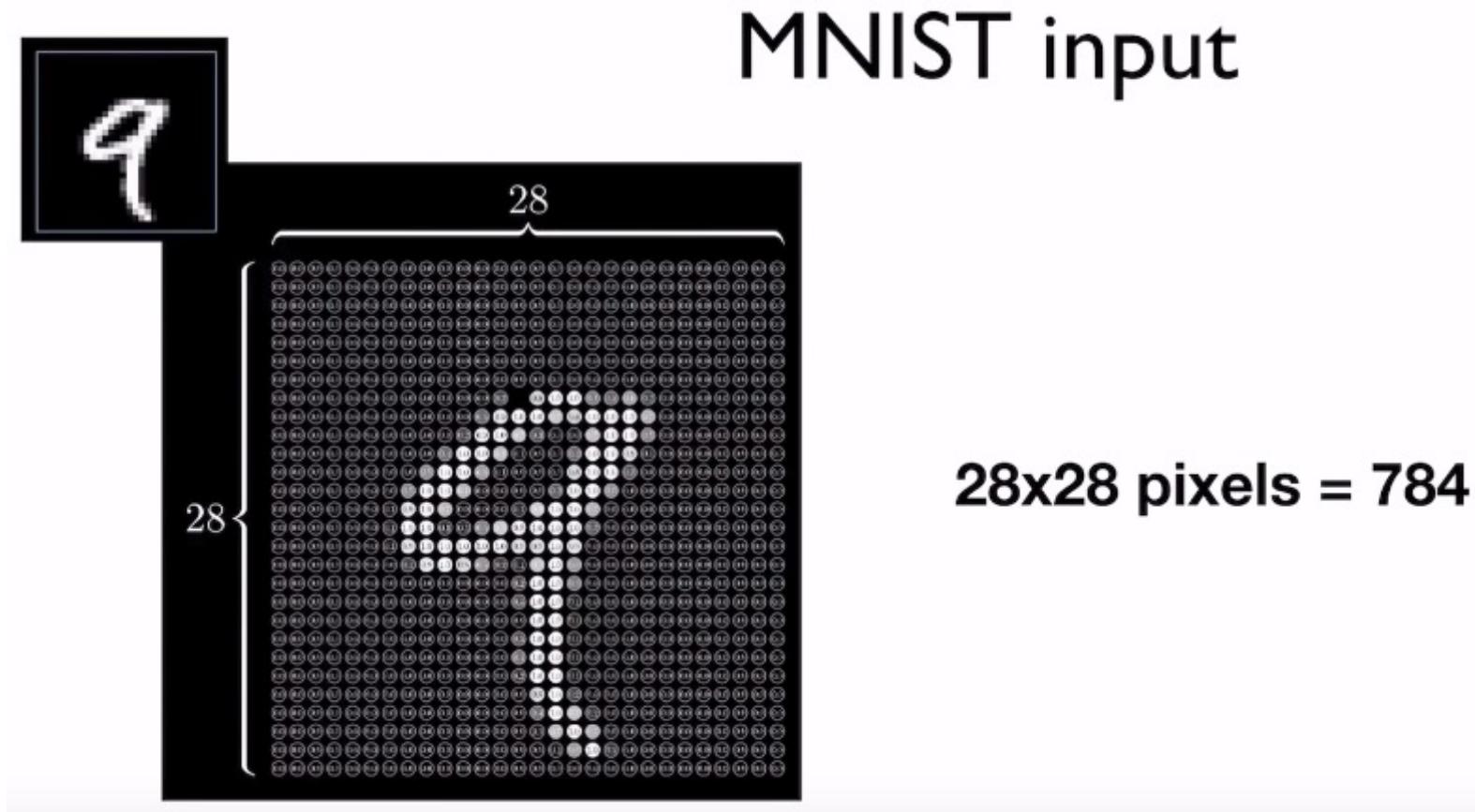
$$\hat{\mathbf{y}} \xrightarrow{L} D(\hat{\mathbf{y}}, \mathbf{y}) = - \sum_j y_j \ln \hat{y}_j$$

$\begin{bmatrix} 0.1 \\ 0.5 \\ 0.4 \end{bmatrix}$        $\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$

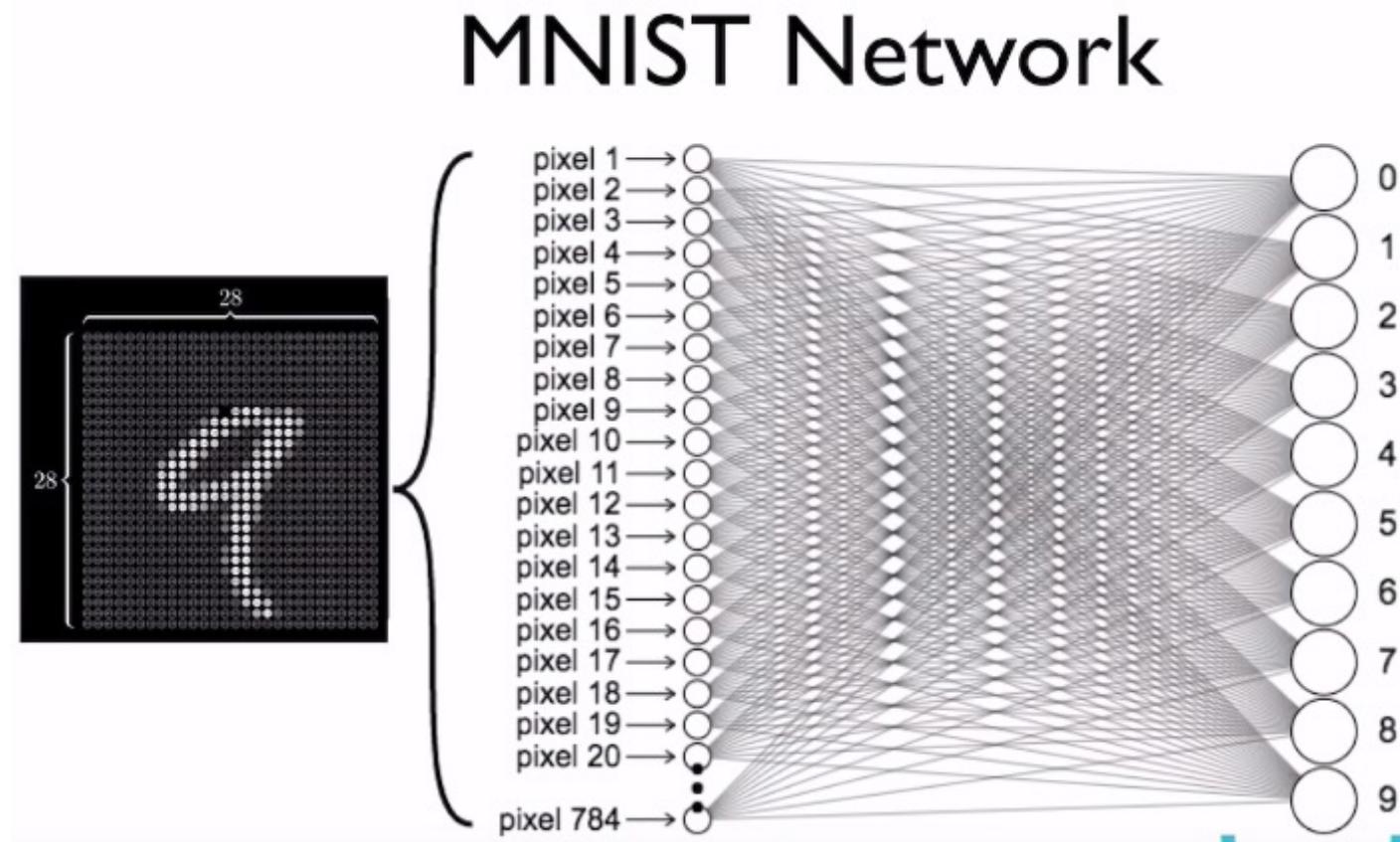
# Model Training: Optimization

- $W^* = \operatorname{argmin}_W L(f(x; W), y)$ 
  - $y \sim \hat{y}$
- Update  $W \rightarrow W + \Delta W$  only if  $L(W + \Delta W) < L(W)$
- Finish it when  $L(W + \Delta W) == L(W)$
- How can we find  $\Delta W$  so that  $L(W + \Delta W) < L(W)$ ?
  - Gradient Descent

# Example: MNIST Classification

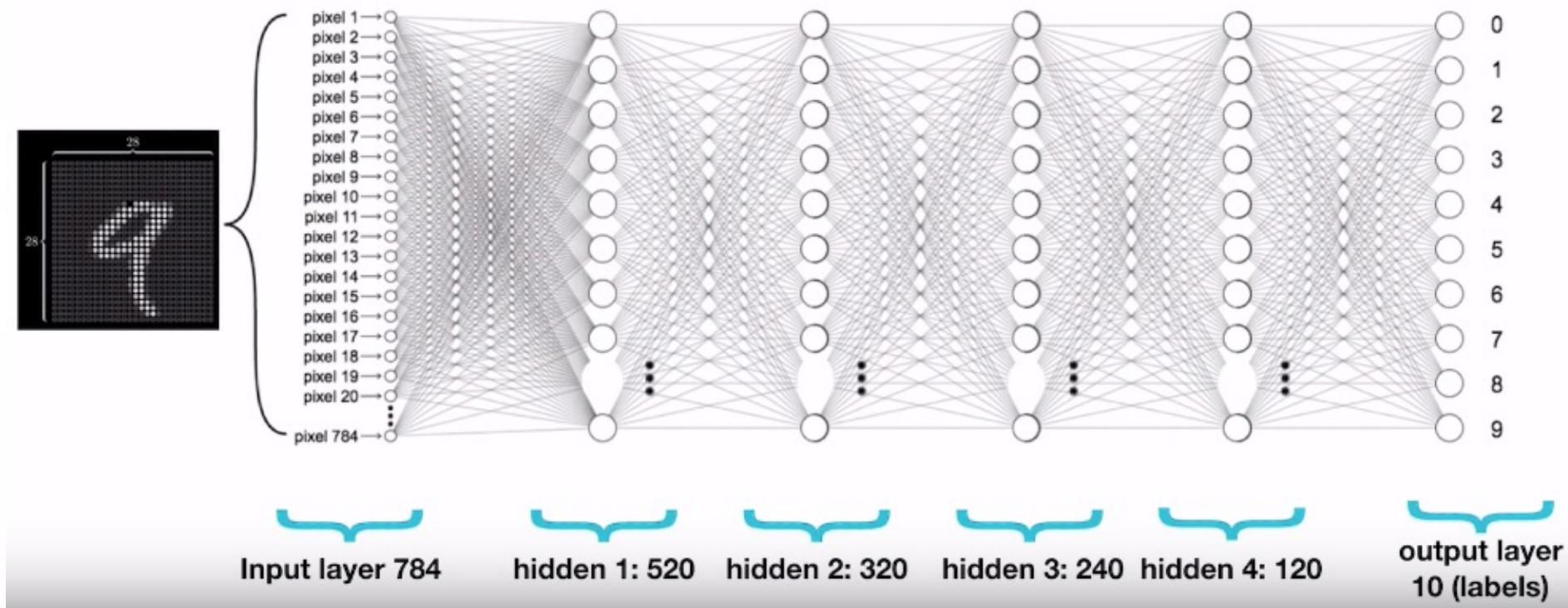


# Example: MNIST Classification



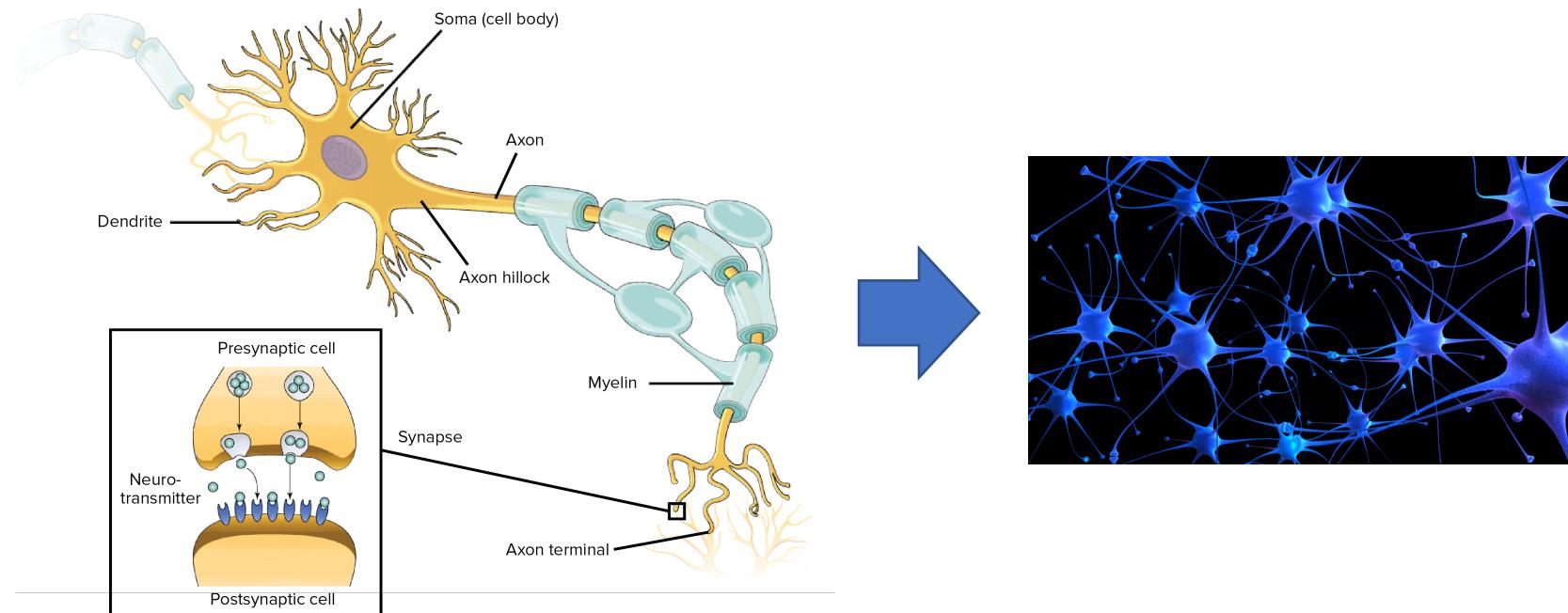
# Example: MNIST Classification

## MNIST Network



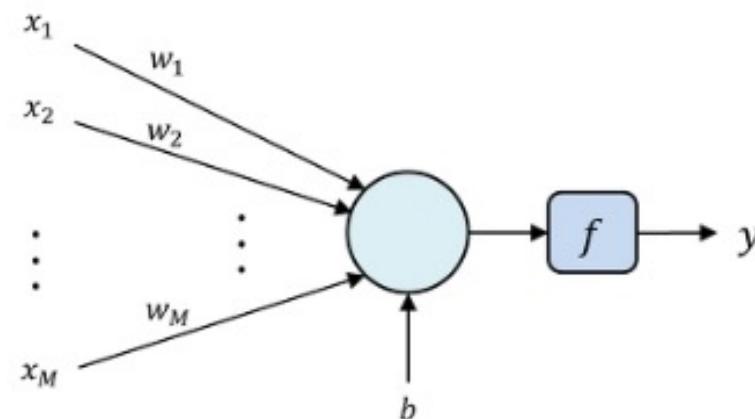
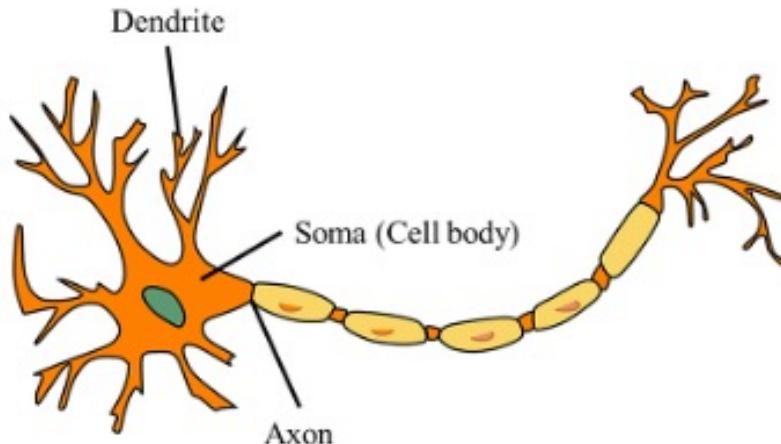
# Neuron

- Neurons are the fundamental units of the brain and nervous system.
- A dendrite receives input from other cells.
- An axon sends an electrical message to other neurons.



# Artificial Neuron (Perceptron)

- A fundamental building block of a neural network
- Perceptron is a simplified model of a biological neuron.
- Dendrite → Weights, Axon → Activation function



# Outline

- Machine Learning
- Sequential Modeling
- Transformer and Language Models
- Dialogue Models

# One-hot Vector/Encoding

- $v \in \{0,1\}^{|V|}$ , where v is one-hot vector and  $|V|$  is vocabulary size

Training corpus

Vocabulary:  
Man, woman, boy,  
girl, prince,  
princess, queen,  
king, monarch

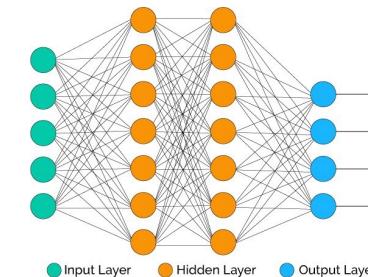
	1	2	3	4	5	6	7	8	9
man	1	0	0	0	0	0	0	0	0
woman	0	1	0	0	0	0	0	0	0
boy	0	0	1	0	0	0	0	0	0
girl	0	0	0	1	0	0	0	0	0
prince	0	0	0	0	1	0	0	0	0
princess	0	0	0	0	0	1	0	0	0
queen	0	0	0	0	0	0	1	0	0
king	0	0	0	0	0	0	0	1	0
monarch	0	0	0	0	0	0	0	0	1

Each word gets  
a  $1 \times 9$  vector  
representation

Last week → vocabulary-size of vector  
movie → vocabulary-size of vector

⋮

Feed

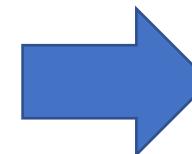


# Drawbacks of One-hot Representation

- Curse of Dimensionality

	1	2	3	4	5	6	7	8	9
man	1	0	0	0	0	0	0	0	0
woman	0	1	0	0	0	0	0	0	0
boy	0	0	1	0	0	0	0	0	0
girl	0	0	0	1	0	0	0	0	0
prince	0	0	0	0	1	0	0	0	0
princess	0	0	0	0	0	1	0	0	0
queen	0	0	0	0	0	0	1	0	0
king	0	0	0	0	0	0	0	1	0
monarch	0	0	0	0	0	0	0	0	1

Mostly zeros



Information quality ↓

Memory use ↑

Computation time ↑

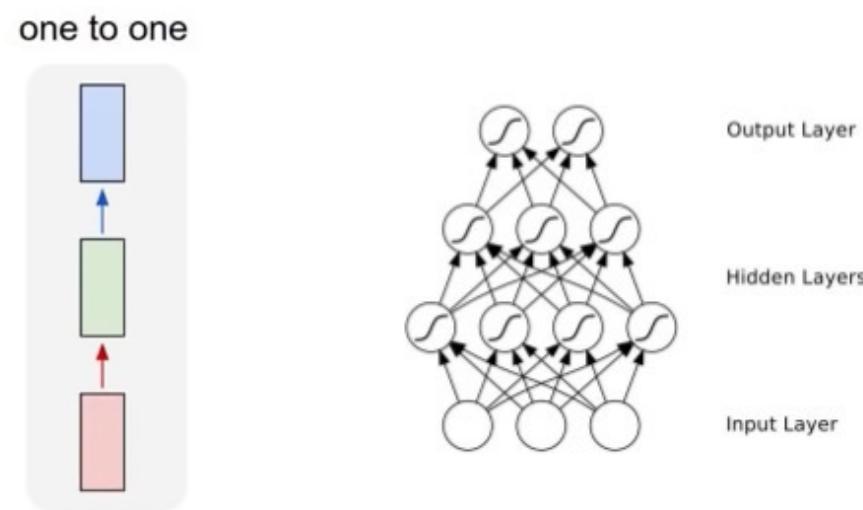
- Orthogonality

$$[0,1,0,\dots,0] \times [0,1,0,\dots,0]^T = 0$$

$$\text{Distance}(\text{"Dog"}, \text{"Puppy"}) = \text{Distance}(\text{"Dog"}, \text{"Computer"}) = 0$$

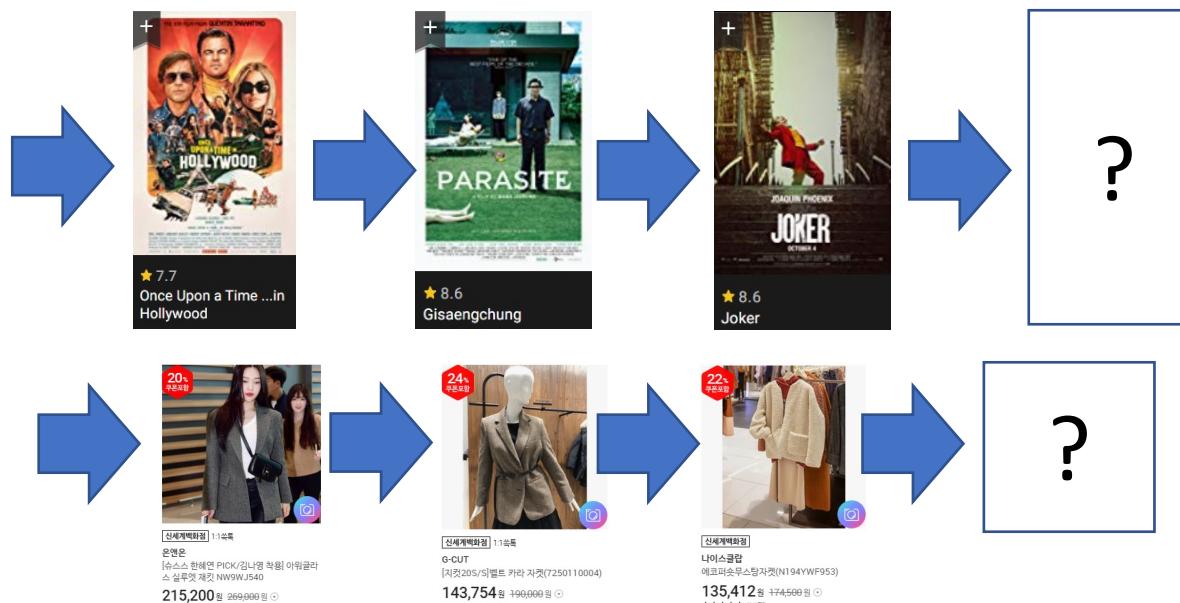
# “Vanilla” Feed-forward Neural Network

- Vanilla feed forward network receives input with a fixed size (e.g., image), and the input is fed through hidden layers and produces outputs.



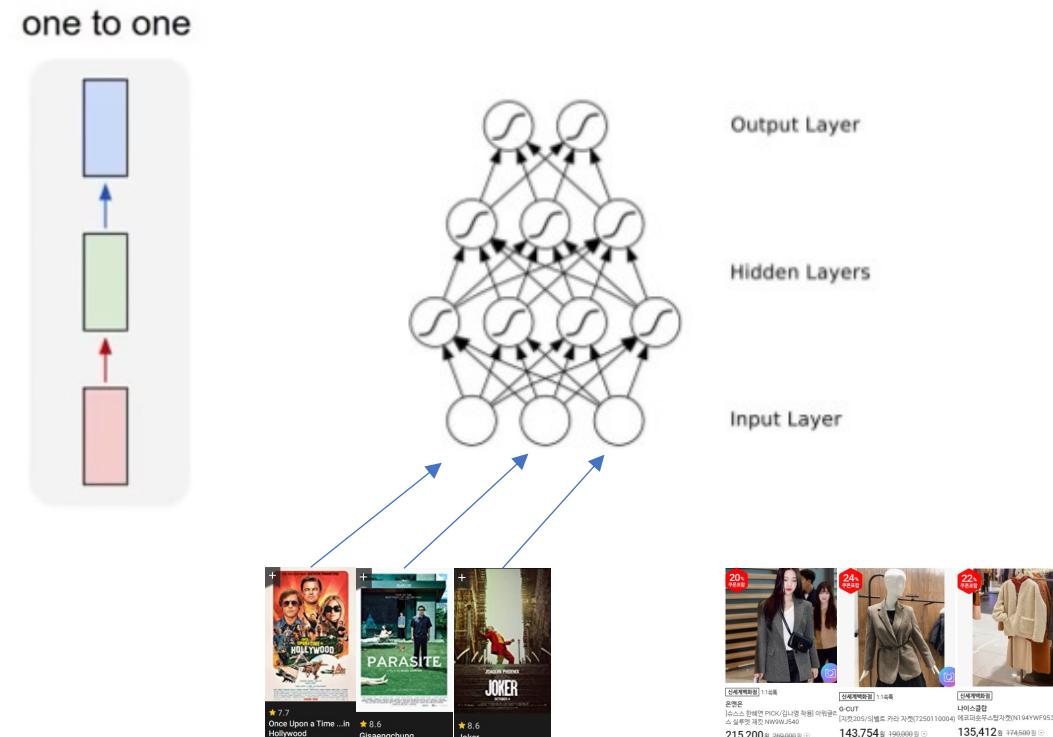
# How can we handle sequence data?

- Many things are sequence.
  - When you watch movies (Next basket prediction)
  - When you enjoy online-shopping (Next basket prediction)
  - Speech recognition, Actions in video, ...
  - When you write down something



# Vanilla NN X Sequential Data

- What if the sequence length is not fixed?
- Sequential order is useless information?



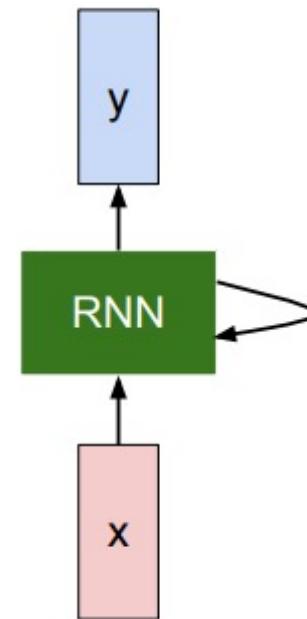
# RNN: Recurrent Neural Network

- In each neuron of RNN, the output of previous time step is fed as input of the next time step.

We can process a sequence of vectors  $\mathbf{x}$  by applying a **recurrence formula** at every time step:

$$h_t = f_W(h_{t-1}, x_t)$$

new state      old state      input vector at some time step  
some function with parameters W



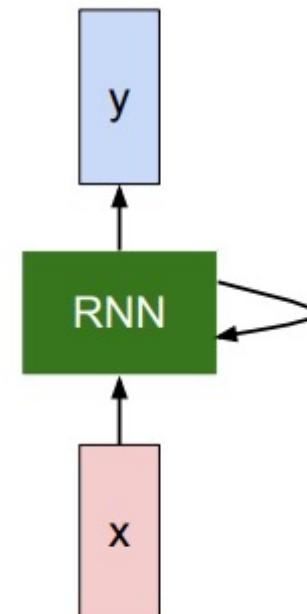
# RNN: Recurrent Neural Network

- This makes RNN be aware of temporal dependency while the Feed-forward NN has none.

We can process a sequence of vectors  $\mathbf{x}$  by applying a **recurrence formula** at every time step:

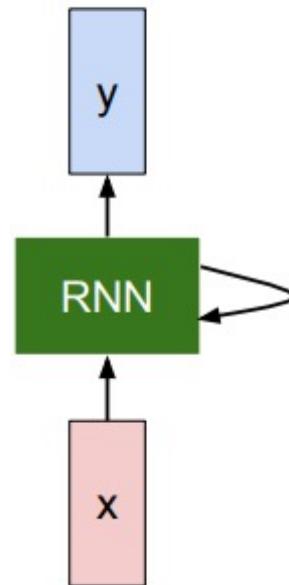
$$h_t = f_W(h_{t-1}, x_t)$$

Notice: the same function and the same set of parameters are used at every time step.



# RNN: Recurrent Neural Network

The state consists of a single “*hidden*” vector  $\mathbf{h}$ :



$$h_t = f_W(h_{t-1}, x_t)$$

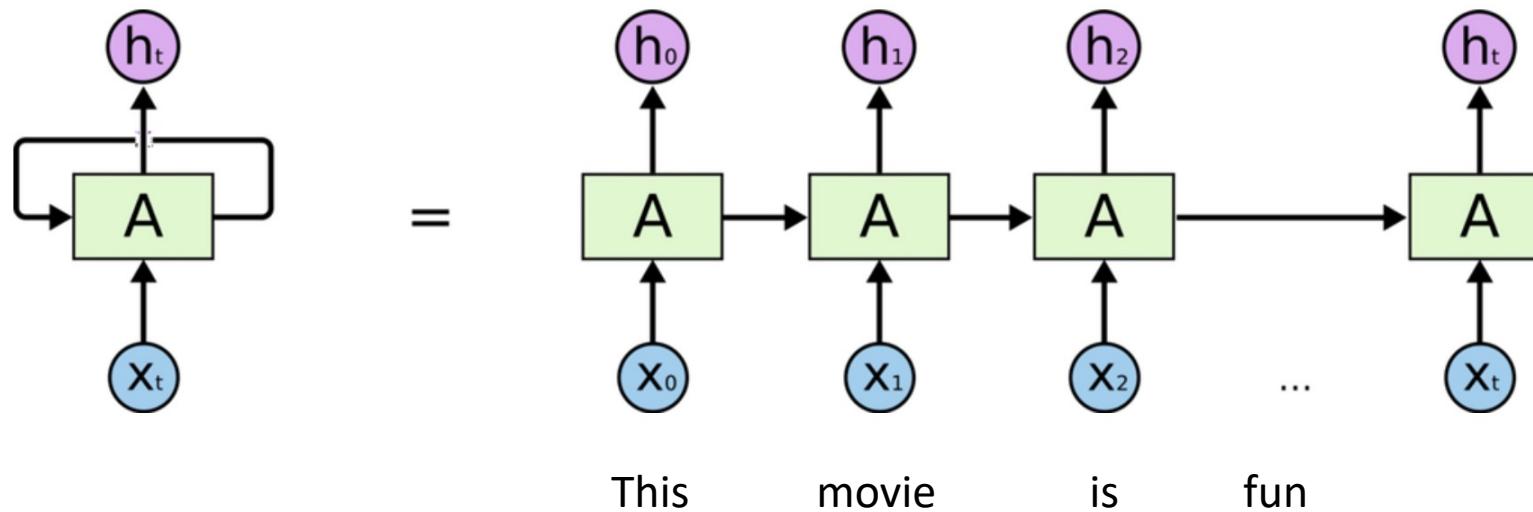


$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$$

$$y_t = W_{hy}h_t$$

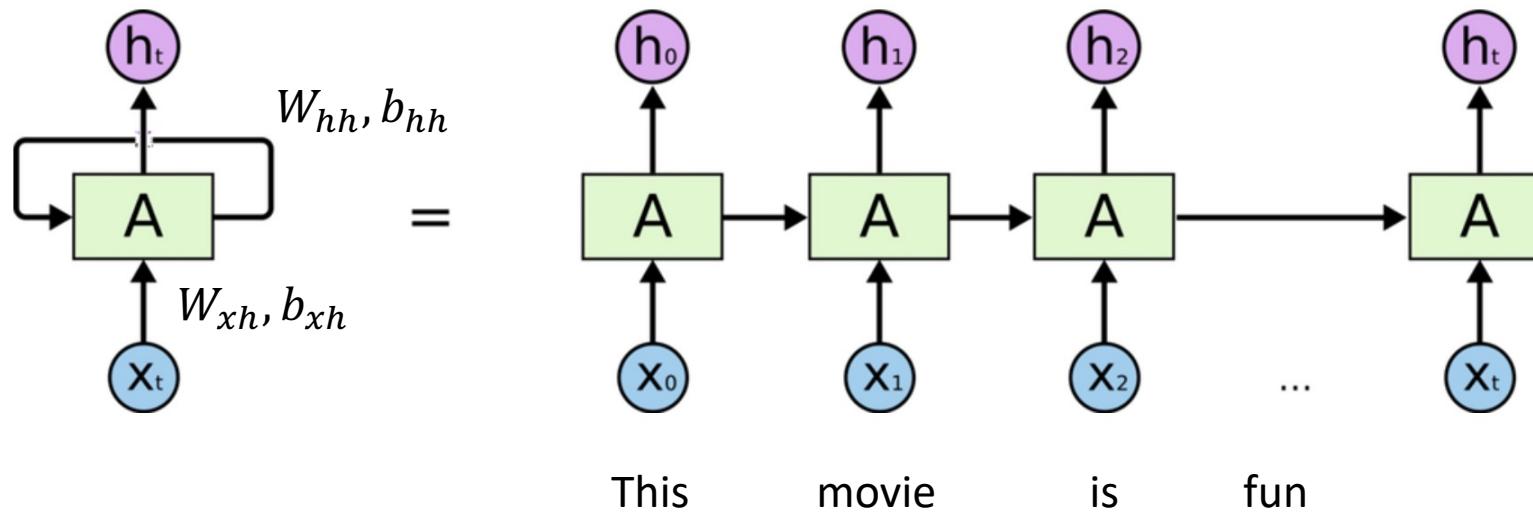
# RNN: Recurrent Neural Network

$$x_t \in \mathbb{R}^{|V|}, h_t \in \mathbb{R}^d \text{ where } |V| \gg d$$



# RNN: Recurrent Neural Network

$$x_t \in \mathbb{R}^{|V|}, h_t \in \mathbb{R}^d \text{ where } |V| \gg d$$

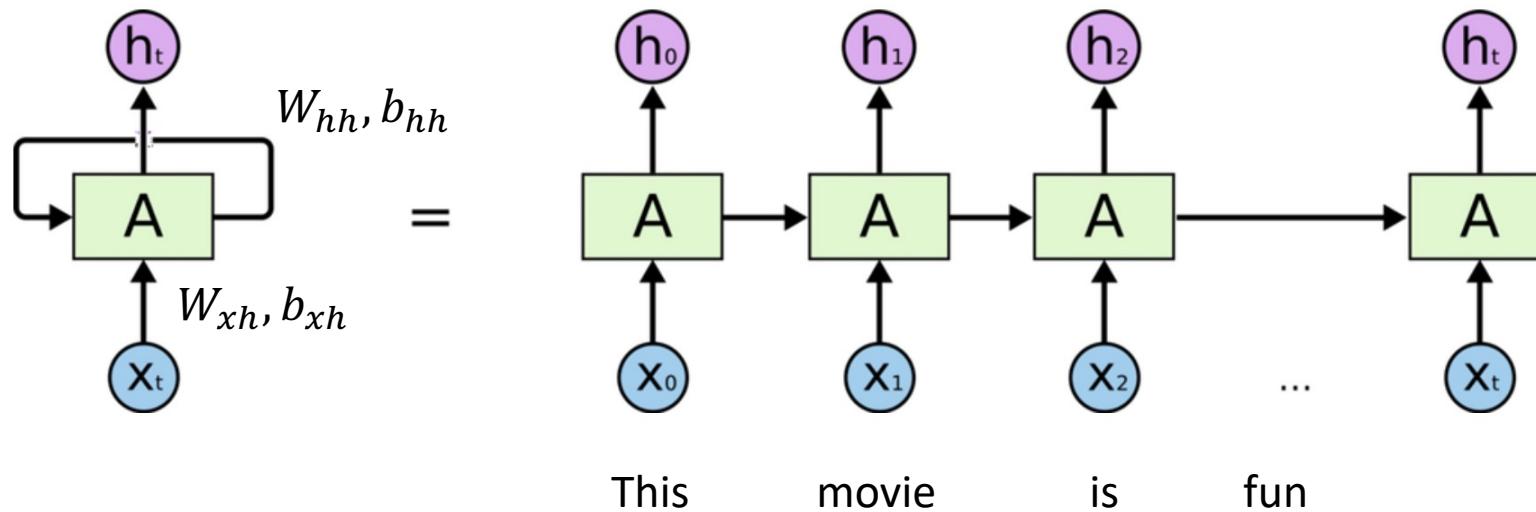


Trainable parameters  $W_{xh} \in \mathbb{R}^{d \times |V|}, b_{xh} \in \mathbb{R}^d$

Trainable parameters  $W_{hh} \in \mathbb{R}^{d \times d}, b_{hh} \in \mathbb{R}^d$

# RNN: Recurrent Neural Network

$$x_t \in \mathbb{R}^{|V|}, h_t \in \mathbb{R}^d \text{ where } |V| \gg d$$



$$h_t = f(x_t, h_{t-1}; \theta) = \tanh(W_{xh}x_t + b_{xh} + W_{hh}h_{t-1} + b_{hh})$$

where  $\theta = \{W_{xh}, b_{xh}, W_{hh}, b_{hh}\}$

# RNN: Recurrent Neural Network

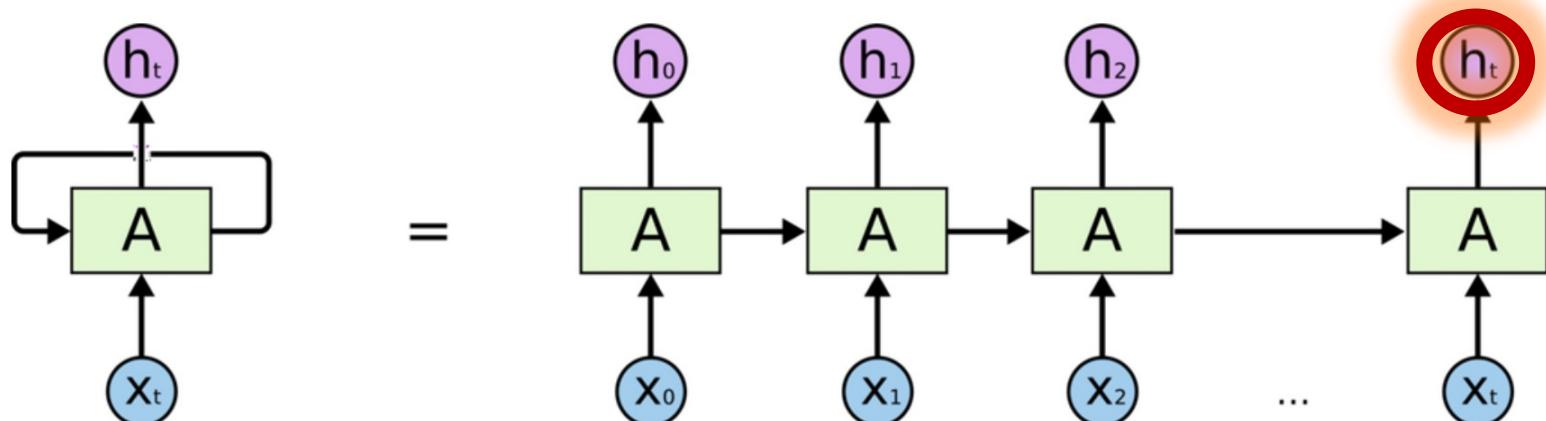
Class  $\in \{\text{positive}, \text{negative}\}$

This movie is fun...  
[one-hot representation]  $\longrightarrow$  [hidden representation]  $\longrightarrow$  [class probabilities]

$$x_t \in \mathbb{R}^{|V|}$$

$$h_t \in \mathbb{R}^d$$

$$\hat{y} = \text{softmax}(W_{hy} h_t) \\ \text{where } W_{hy} \in \mathbb{R}^{2 \times d}$$



# Is this spam?

Gmail

편지쓰기

메일

받은편지함

별표편지함

다시 알림 항목

보낸편지함

임시보관함

물품구매

간략히

중요편지함

예약됨

전체보관함

스팸함

21

휴지통

스팸함의 메일은 30일 후에 자동 삭제됩니다. 지금 모든 스팸 메일 삭제

- ★ Camila Garcia [Submit Manuscript and Join Us as Editorial Board Members or Reviewers](#) - If you no longer want to receive emails from us, click here.
- ★ a shannon (제목 없음) - hello
- ★ Noo Canter 교수 관련 청원 - 교수님께, 안녕하세요 바쁘신데 죄송합니다. 아동 및 교수 관련 청원이 올라와 있습니다. 등록된 교수님에게는 이메일로 알리겠습니다.
- ★ 기린선교회 이카호중앙... {제323회} 할렐루야! 기린선교회 이카호중앙교회에서 일본선교소식과 신규 동영상 안내입니다! - 2022년 1월
- ★ 김창근 DNA의 저장 능력 - DNA의 저장 능력 <https://www.jw.org/finder?wtlocale=KO&docid=102013448&srcid=s>
- ★ 기린선교회 이카호중앙... {제322회} 할렐루야! 기린선교회 이카호중앙교회에서 일본선교소식과 신규 동영상 안내입니다! - 2022년 1월
- ★ rese...@globaleventshu... Research Updates: Volume-11 Issue-2, December 2021 of 'International Journal of Innovative Technol...
- ★ 광종국 연세대 교수/교직원 공지사항입니다. - 거짓교육 금지결정, 7개항목 국어/영어/수학/사회/과학/종교/경제...
- ★ 기린선교회 이카호중앙... [제320회] {제321회} 할렐루야! 기린선교회 이카호중앙교회에서 일본선교소식과 신규 동영상 안내입니다!
- cs tom@info.hksme.org.cn Successfully logged into the system, all the data from your system has been copied. Read the instructions in the attached file.
- ★ 광종국 Submission Deadline Extension of DMCIT 2022 - Dear Professor, Researchers and Colleagues, Good day to all,
- ★ 기린선교회 이카호중앙... 연세대 교수/교직원 공지사항 - 교육금지/발언금지 7개항목 발표 교육부, 교수회 만장일치 결정/ 7개 주요...
- ★ iris@info.hksme.org.cn {제319회} 할렐루야! 기린선교회 이카호중앙교회에서 일본선교소식과 신규 동영상 안내입니다! - 저주 받은



# Is this spam?

The screenshot shows a Gmail inbox with several messages listed. A specific message from '곽종국 <7eng7@naver.com>' is highlighted with a red box and a large blue arrow pointing to it. This message has been flagged as spam. A modal window is overlaid on the inbox, containing the following text:

이 메일이 스팸으로 분류된 이유는 무엇인가요? 이전에 스팸으로 확인된 메일과 유사합니다.

스팸이 아님

교육금지/발언금지 7개항목 발표  
교육부, 교수회 만장일치 결정/  
7개 주요 교과목, 영문법 95% 포함

아래 사이트를 클릭(접속) 바랍니다.

[www.abc33.kr/003/](http://www.abc33.kr/003/)

교育금지/발언금지 중대 7개 항목 발표문/

The message body of the flagged email contains the following text:

[제320회] (제321회) 할렐루야! 기린선교회 이카호중앙교회에서 일본선교소식과 신규 동영상 안내입니다!

Successfully copied all data into the system, all the data from your system has been copied. Read the instructions.

Submission Dead Extension of DMCIT 2022 - Dear Professor, Researchers and Colleagues, Good day

연세대 교수/교직원 공지사항 - 교육금지/발언금지 7개항목 발표 교육부, 교수회 만장일치 결정/ 7개 주요

{제319회} 할렐루야! 기린선교회 이카호중앙교회에서 일본선교소식과 신규 동영상 안내입니다! - 저주 받은

iris@info.hksme.org.cn SCIE/EI/SCOPUS Journals 2022 CFP: Sensors/Engineering/Big Data/AI - SCIE/EI/SCOPUS Journals 2022

# **Text Classification: tasks**

- Assigning subject categories, topics, or genres
- Spam detection
- Age/gender identification
- Language identification
- Sentiment analysis

# Positive or negative review?

- Unbelievably disappointing
- Full of zany characters and richly applied satire, and some great plot twists
- This is the greatest screwball comedy ever filmed
- It was pathetic. The worst part about it was the boxing scenes
- Domains
  - Movie review
  - Product review on E-commerce
  - Dialogue utterance
  - ...

# Single-sequence Classification

- IMDB Large Movie Reviews Corpus
  - Binary sentiment classification dataset containing 50,000 polarized (positive or negative) movie reviews
  - Comparably long text such as document

If you like adult comedy cartoons, like South Park, then this is nearly a similar format about the small adventures of three teenage girls at Bromwell High. Keisha, Natella and Latrina have given exploding sweets and behaved like bitches, I think Keisha is a good leader. There are also small stories going on with the teachers of the school. There's the idiotic principal, Mr. Bip, the nervous Maths teacher and many others. The cast is also fantastic, Lenny Henry's Gina Yashere, EastEnders Chrissie Watts, Tracy-Ann Oberman, Smack The Pony's Doon Mackichan, Dead Ringers' Mark Perry and Blunder's Nina Conti. I didn't know this came from Canada, but it is very good. Very good!

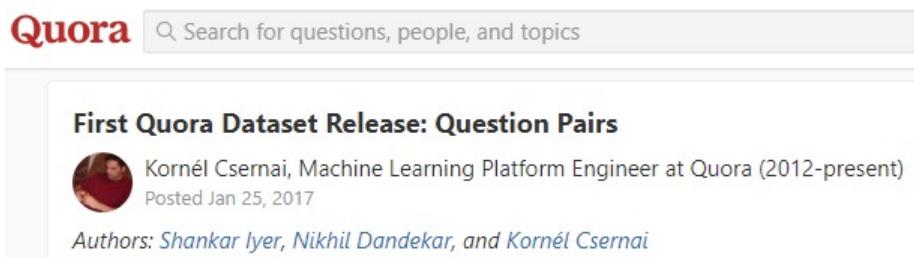
<IMDB dataset>

	id	document	label
0	9976970	아 더빙.. 진짜 짜증나네요 목소리	0
1	3819312	흠...포스터보고 초딩영화풀....오버연기조차 가볍지 않구나	1
2	10265843	너무재alez었다그래서보는것을추천한다	0
3	9045019	교도소 이야기구먼 ..솔직히 재미는 없다..광점 조정	0
4	6483659	사이몬페그의 익살스런 연기가 돌보였던 영화!스파이더맨에서 늙어보이기만 했던 커스틴 ...	1

<Naver dataset>

# Sequence-pair Classification

- Quora Question Paraphrases (QQP)
  - Over 400,000 lines of potential question duplicate pairs
  - Each line contains IDs for each question in the pair, the full text for each question, and a binary value that indicates whether the line truly contains a duplicate pair
  - The ground-truth labels contain some amount of noise: they are not guaranteed to be perfect



The screenshot shows a Quora post titled "First Quora Dataset Release: Question Pairs" by Kornél Csernai. The post includes a profile picture of Kornél Csernai, his title as a Machine Learning Platform Engineer at Quora (2012-present), and the date it was posted (Jan 25, 2017). It also lists the authors: Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. Below the post is a table showing five examples of question pairs and their classification.

id	qid1	qid2	question1	question2	is_duplicate
447	895	896	What are natural numbers?	What is a least natural number?	0
1518	3037	3038	Which pizzas are the most popularly ordered pizzas on Domino's menu?	How many calories does a Dominos pizza have?	0
3272	6542	6543	How do you start a bakery?	How can one start a bakery business?	1
3362	6722	6723	Should I learn python or Java first?	If I had to choose between learning Java and Python, what should I choose to learn first?	1

1) "Quora question paraphrases", <https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs>

# Natural Language Inference

- Stanford Natural Language Inference (SNLI)
  - 570k human-written English sentence pairs manually labeled for balanced classification with the labels entailment, contradiction, and neutral.
  - Neural model also must handle phenomena like lexical entailment, quantification, coreference, tense, belief, modality, and lexical and syntactic ambiguity.

Text	Judgments	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

# Text Classification for Healthcare

- Automatic EMR Database Construction

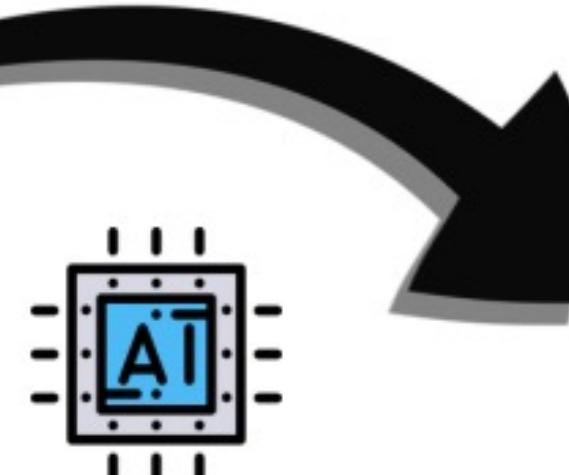
Unstructured EMR (Text)

수술기록-IV-안과공동 사진첨부 [2021-09-03]  
Eye: C2N2(DU)  
Vit: media clear(DU)  
Fundus: cdr 0.3(CD) 0.5(CS) glistening membr(CC>CS)  
z=OCT: drusen(DU) ERM(DU, OD>OS)  
under GENERAL anesthesia  
3 trocars were inserted at supronasal, inferotemp, and supertemp. positions.  
→ the infusion cannula connected to balanced salt plus solution into the vitreous cavity at inferotemp with 25-Gauge trocar  
→ corneal incisions were made at superior  
→ Healon was injected  
→ continuous curvilinear capsulorhexis was made with 28 gauge needle  
→ balanced salt solution was injected between lens capsule and nucleus  
→ nucleus was fragmented with phacoemulsifier and remained lens material was removed with irrigation and aspiration  
→ IOL was inserted into the sulcus d/t post. capsular rupture

2 corneal incisions were made at superior  
→ Healon was injected  
→ continuous curvilinear capsulorhexis was made with 26 guage needle  
→ balanced salt solution was injected between lens capsule and nucleus  
→ lens nucleus was fragmented with phacoemulsifier and remained lens material was removed with irrigation and aspiration  
→ IOL was inserted into the sulcus d/t post. capsular rupture

Structured EMR (Database)

Corneal Incision  Corneal temporal  Corneal superior  Limbal temporal  
IOL implantation site  In-the-bag  Sulcus  Scleral fixation  Others  
Posterior capsular tear  No  Yes  
Viscoelastic material  Healon  Healon 5  Viscoat  Others



Taking unstructured data as input,  
model automatically outputs structured data.

Constructing EMR database can be reduced to medical text classification tasks, by querying the desired information to arbitrary unstructured personal EMR.

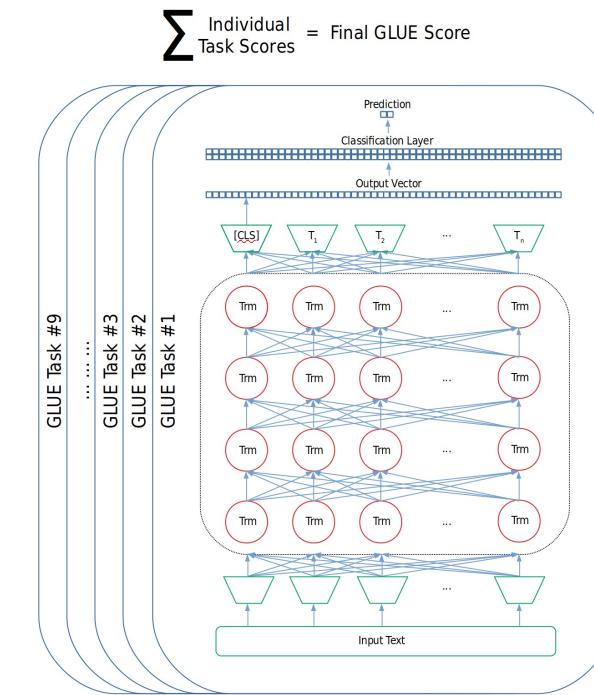
This idea brings great time-savings by keeping human from doing repetitive and time-consuming work.

# GLUE Benchmark

- GLUE benchmark
  - The General Language Understanding Evaluation (GLUE) benchmark is a collection of resources for training, evaluating, and analyzing natural language understanding systems.



<https://gluebenchmark.com/>



# GLUE Benchmark

- Task categories
  - Single-sequence tasks: CoLA, SST-2
  - Sequence-pair tasks: MRPC, STS-B, QQP
  - Inference tasks: MNLI, QNLI, RTE, WNLI

Corpus	Train	Test	Task	Metrics	Domain
Single-Sentence Tasks					
CoLA	8.5k	<b>1k</b>	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
Similarity and Paraphrase Tasks					
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	<b>391k</b>	paraphrase	acc./F1	social QA questions
Inference Tasks					
MNLI	393k	<b>20k</b>	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	<b>146</b>	coreference/NLI	acc.	fiction books

Table 1: Task descriptions and statistics. All tasks are single sentence or sentence pair classification, except STS-B, which is a regression task. MNLI has three classes; all other classification tasks have two. Test sets shown in bold use labels that have never been made public in any form.

# Outline

- Machine Learning
- Sequential Modeling
- Transformer and Language Models
- Dialogue Models

# Transformer

---

## Attention Is All You Need

---

**Ashish Vaswani\***  
Google Brain  
[avaswani@google.com](mailto:avaswani@google.com)

**Noam Shazeer\***  
Google Brain  
[noam@google.com](mailto:noam@google.com)

**Niki Parmar\***  
Google Research  
[nikip@google.com](mailto:nikip@google.com)

**Jakob Uszkoreit\***  
Google Research  
[usz@google.com](mailto:usz@google.com)

**Llion Jones\***  
Google Research  
[llion@google.com](mailto:llion@google.com)

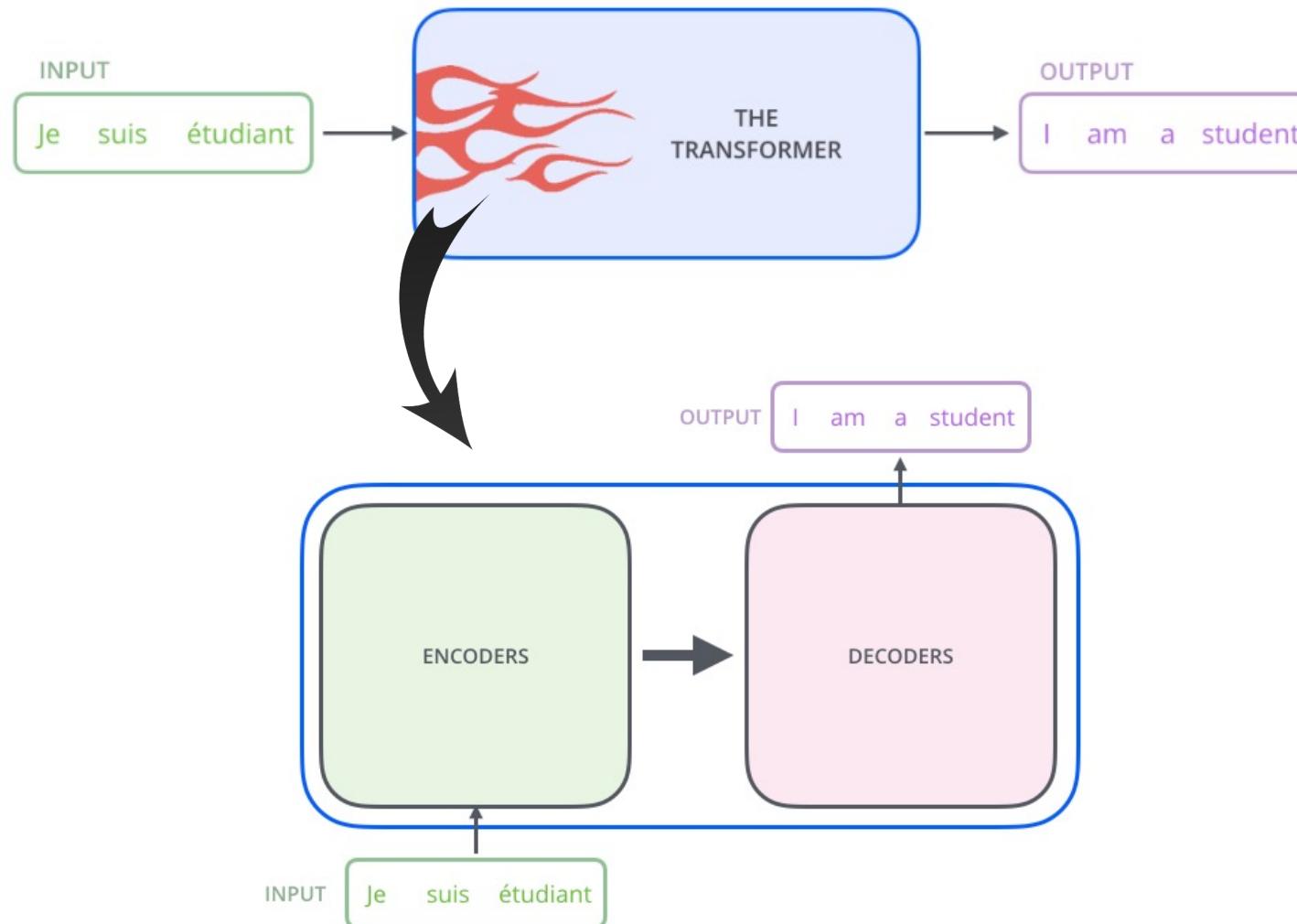
**Aidan N. Gomez\* †**  
University of Toronto  
[aidan@cs.toronto.edu](mailto:aidan@cs.toronto.edu)

**Lukasz Kaiser\***  
Google Brain  
[lukaszkaiser@google.com](mailto:lukaszkaiser@google.com)

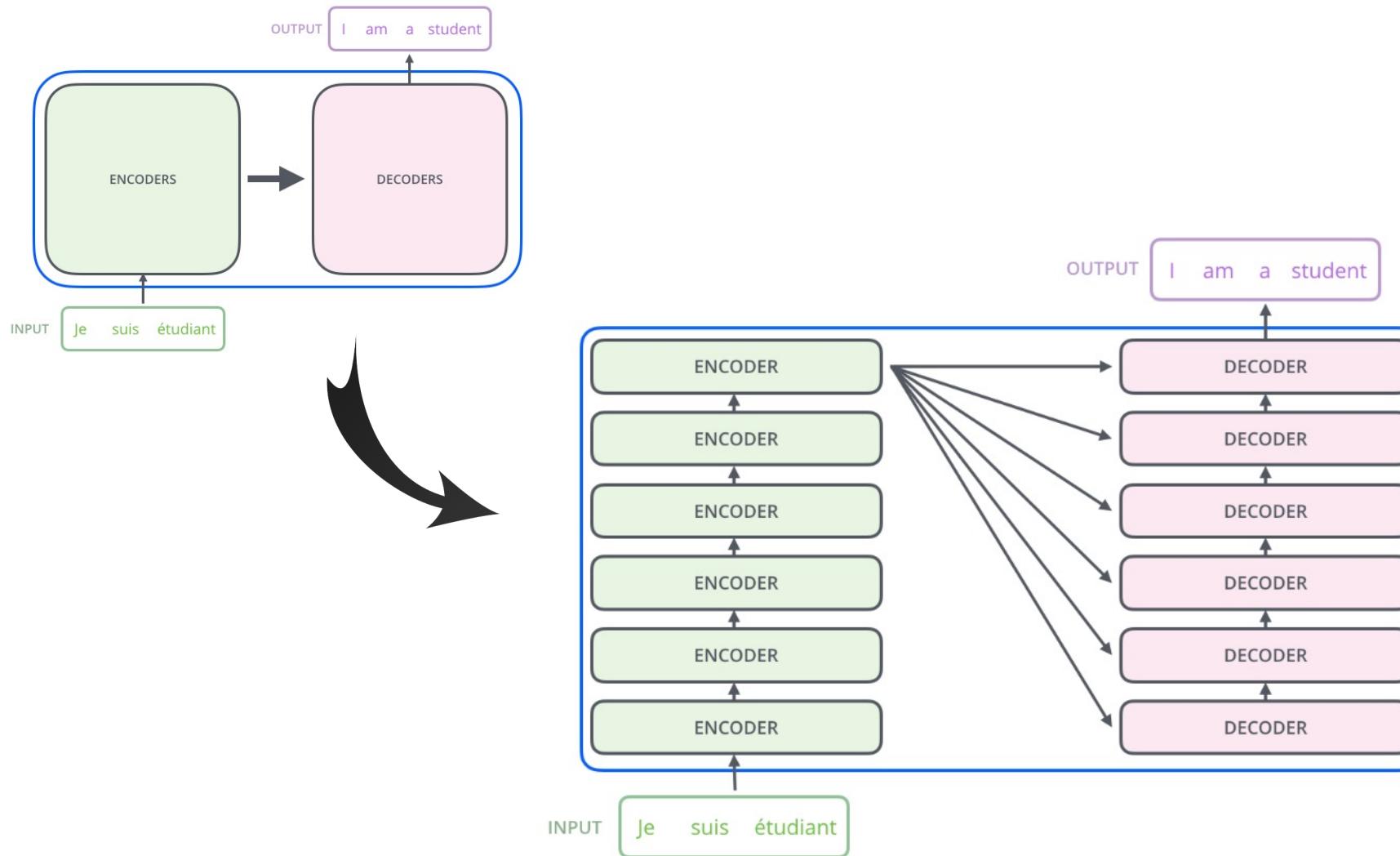
**Illia Polosukhin\* ‡**  
[illia.polosukhin@gmail.com](mailto:illia.polosukhin@gmail.com)

RNN → LSTM → LSTM + attention → Only attention (Transformer)

# A High-level Look

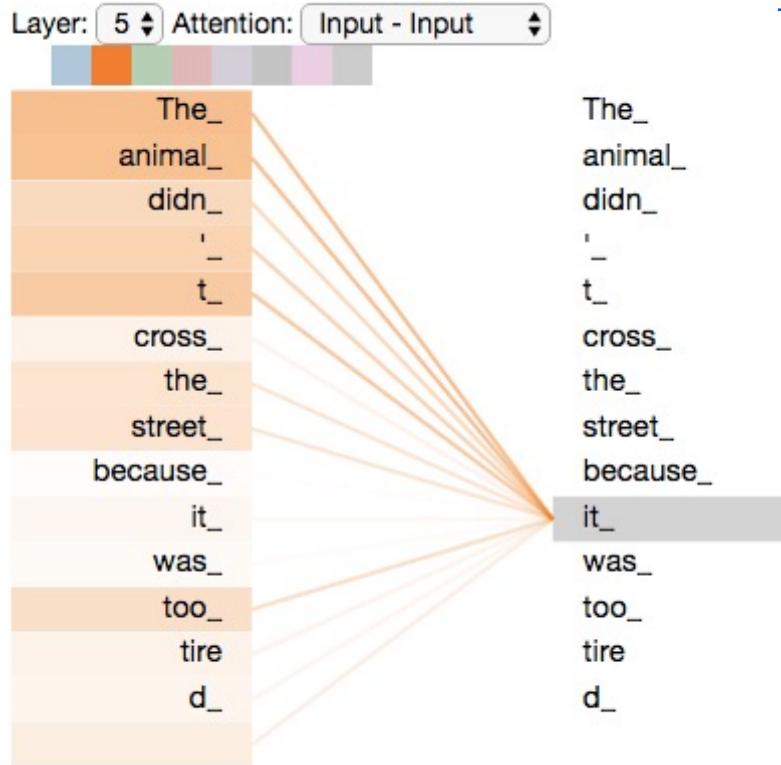


# A High-level Look



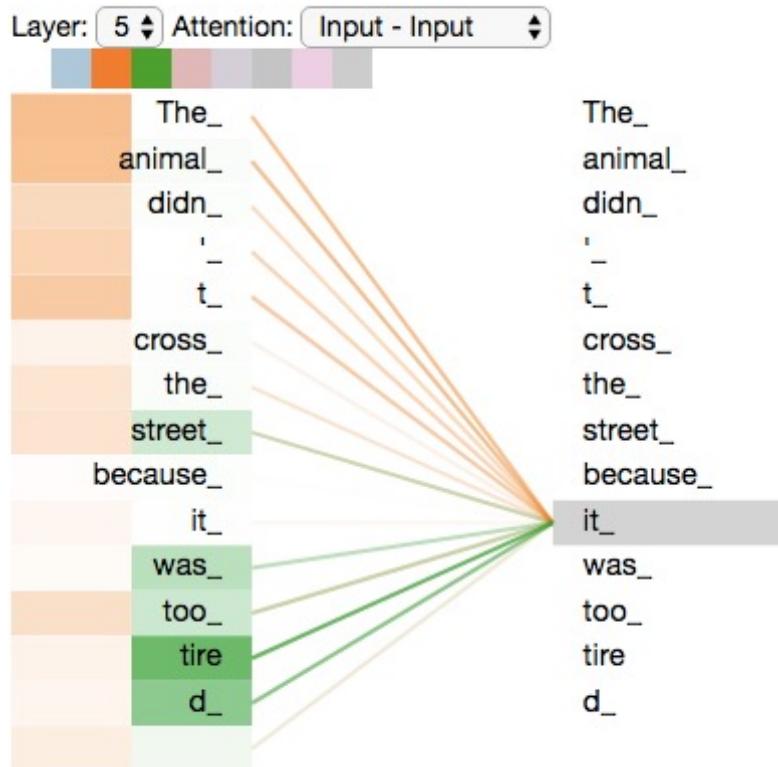
# Self-Attention at a High Level

[Tensor2Tensor notebook](#)



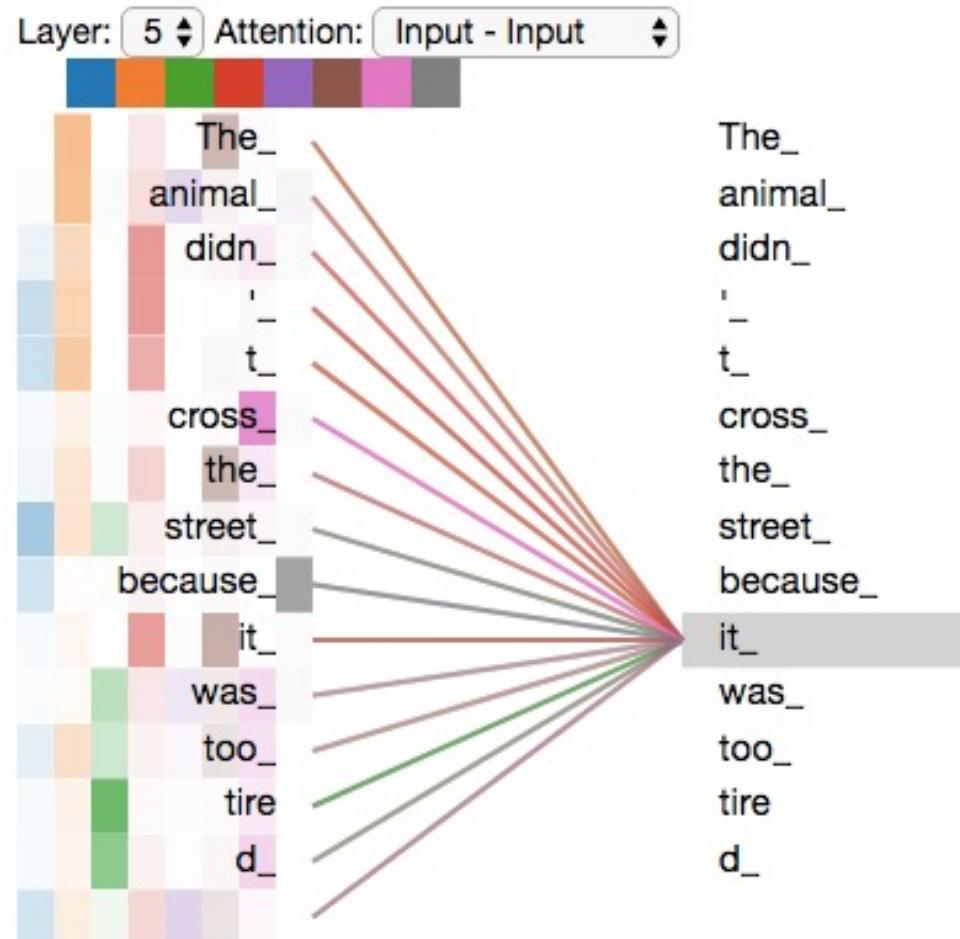
As we are encoding the word "it" in encoder #5 (the top encoder in the stack), part of the attention mechanism was focusing on "The Animal", and baked a part of its representation into the encoding of "it".

# Multi-head Attention

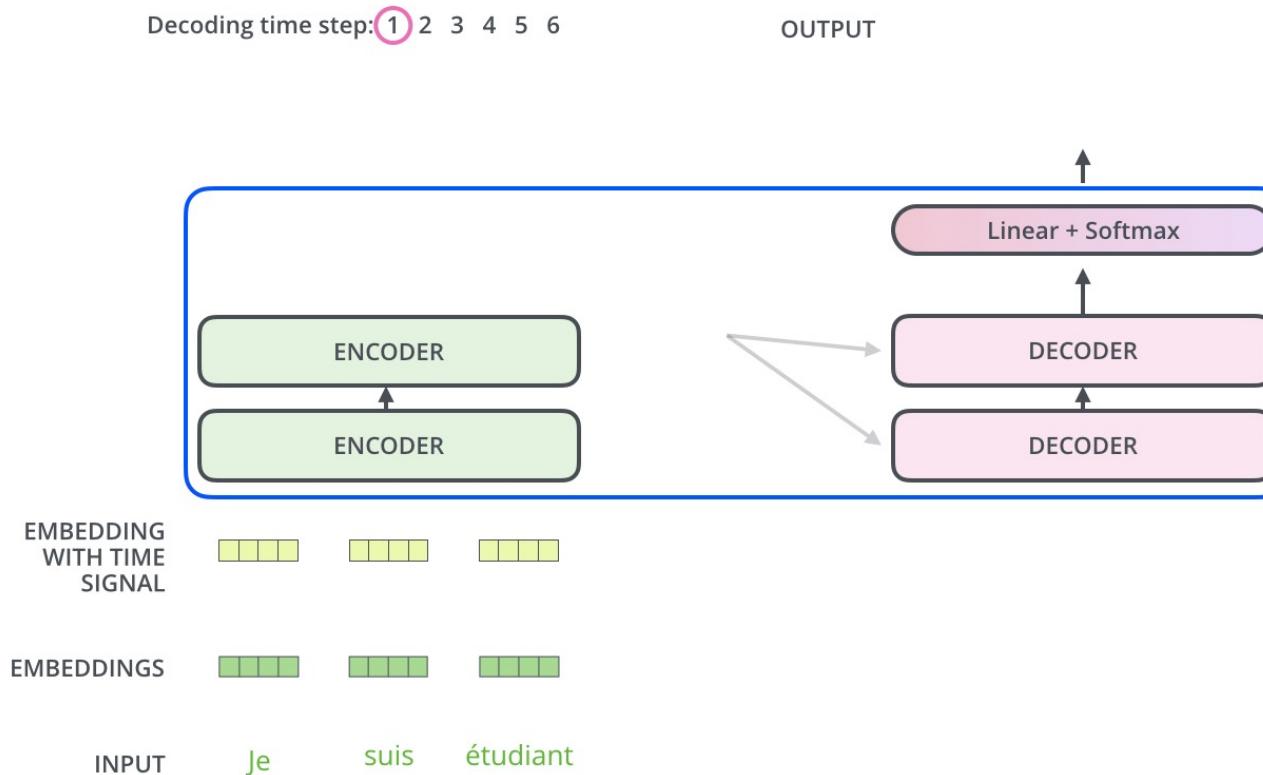


As we encode the word "it", one attention head is focusing most on "the animal", while another is focusing on "tired" -- in a sense, the model's representation of the word "it" bakes in some of the representation of both "animal" and "tired".

# Multi-head Attention

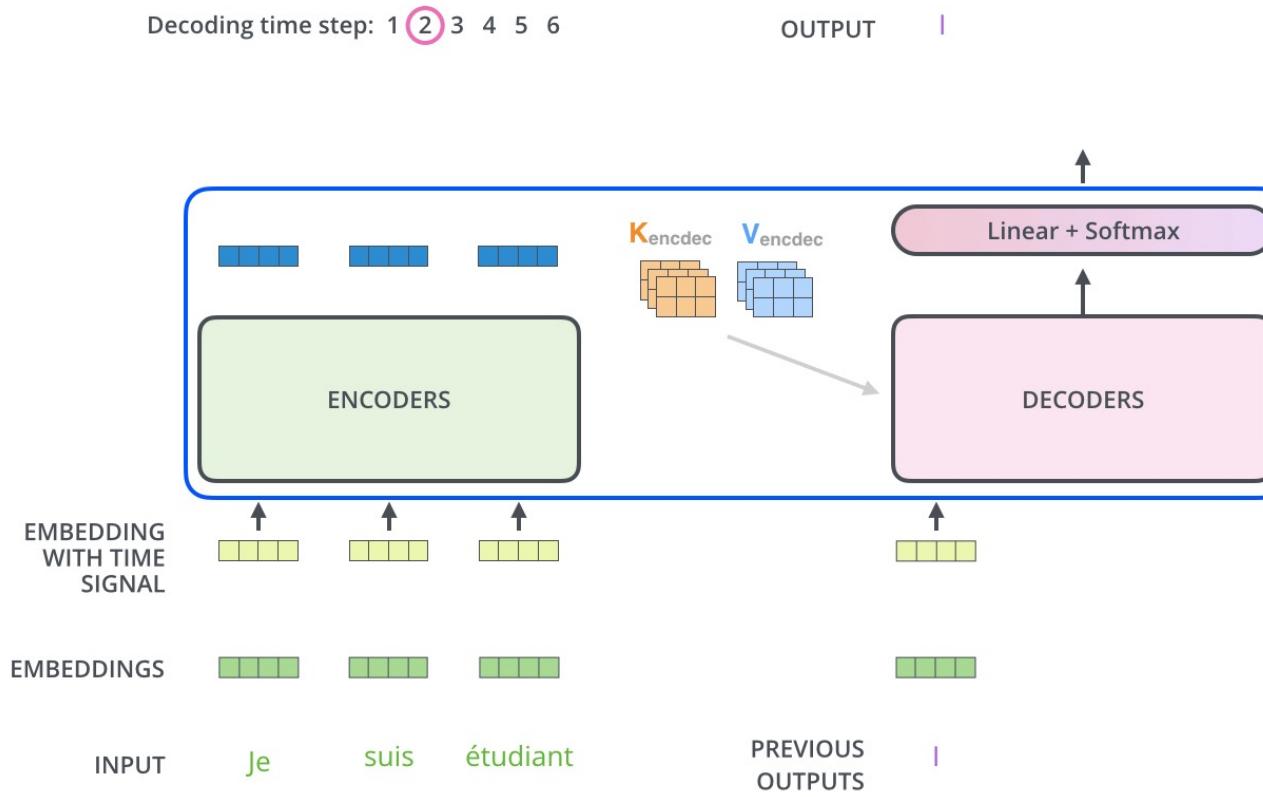


# The Decoder Side



After finishing the encoding phase, we begin the decoding phase. Each step in the decoding phase outputs an element from the output sequence (the English translation sentence in this case).

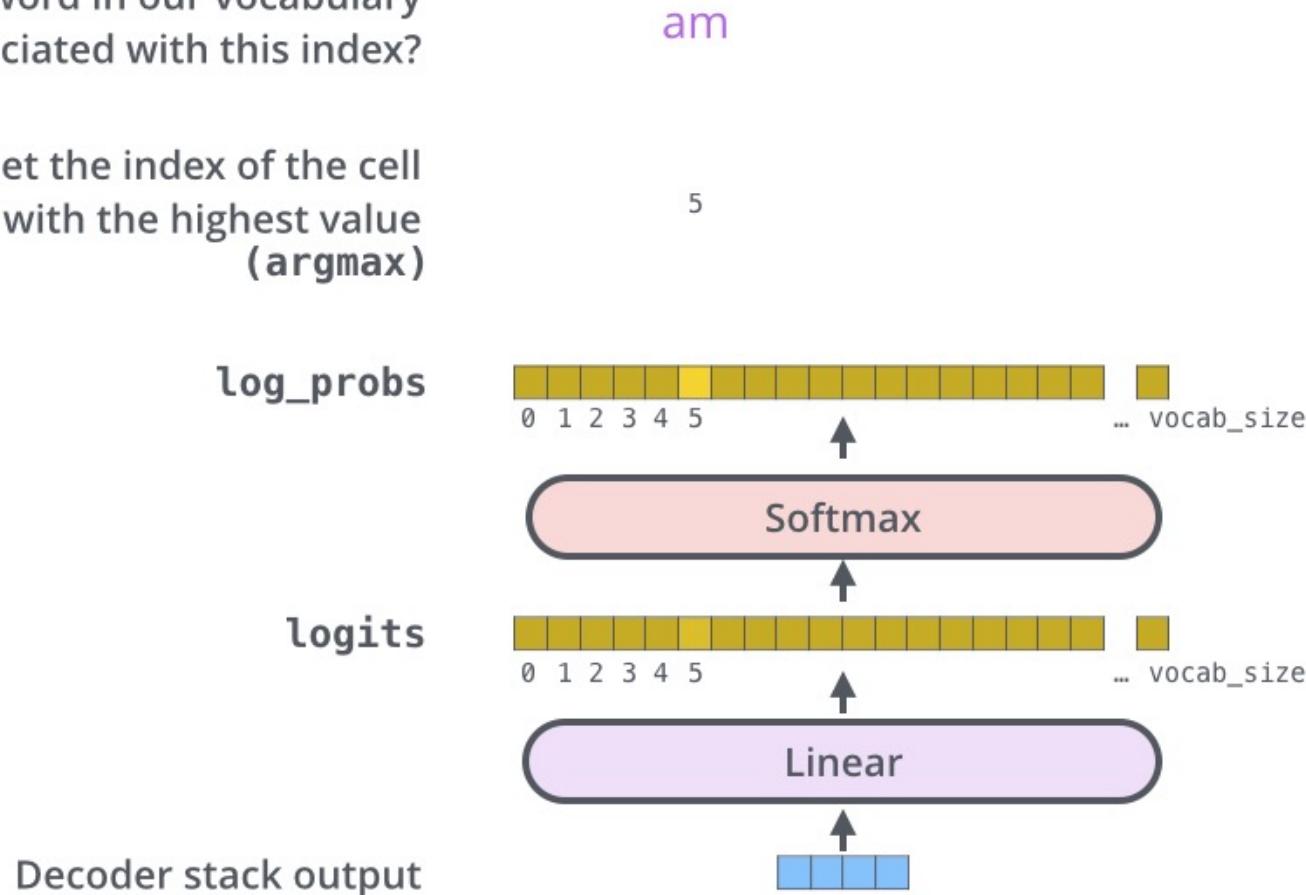
# The Decoder Side



# The Final Linear and Softmax Layer

Which word in our vocabulary  
is associated with this index?

Get the index of the cell  
with the highest value  
(`argmax`)



This figure starts from the bottom with the vector produced as the output of the decoder stack. It is then turned into an output word.

# Deep Pre-trained Language Models

- ~~2017 Jun: Transformer (not LM)~~

## Attention Is All You Need

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin

- 2018 Feb: ELMo

## Deep contextualized word representations

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer

- 2018 Jun: OpenAI GPT

---

## Improving Language Understanding by Generative Pre-Training

---

- 2018 Oct: BERT

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova

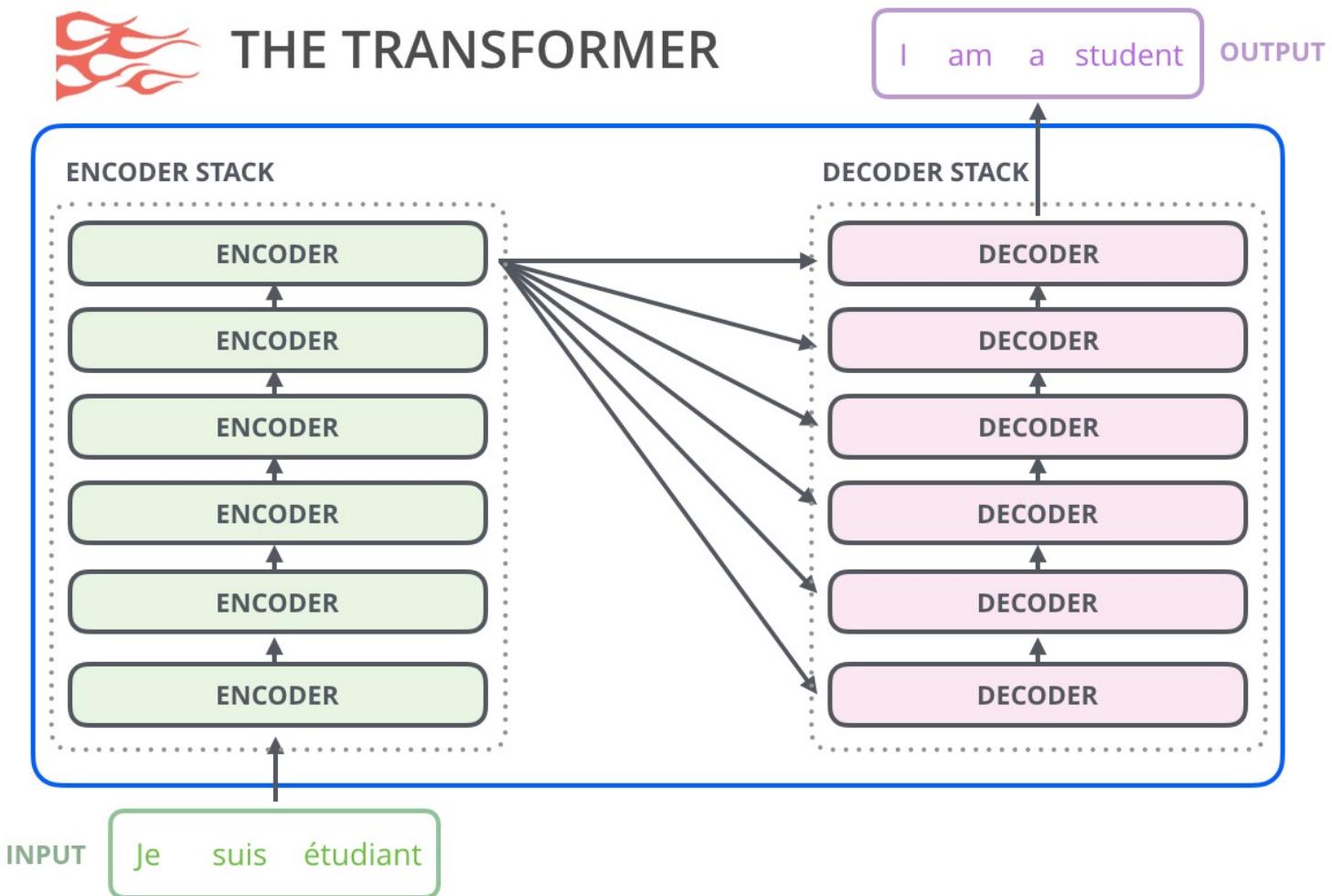
- 2019 Feb: OpenAI GPT-2

## Language Models are Few-Shot Learners

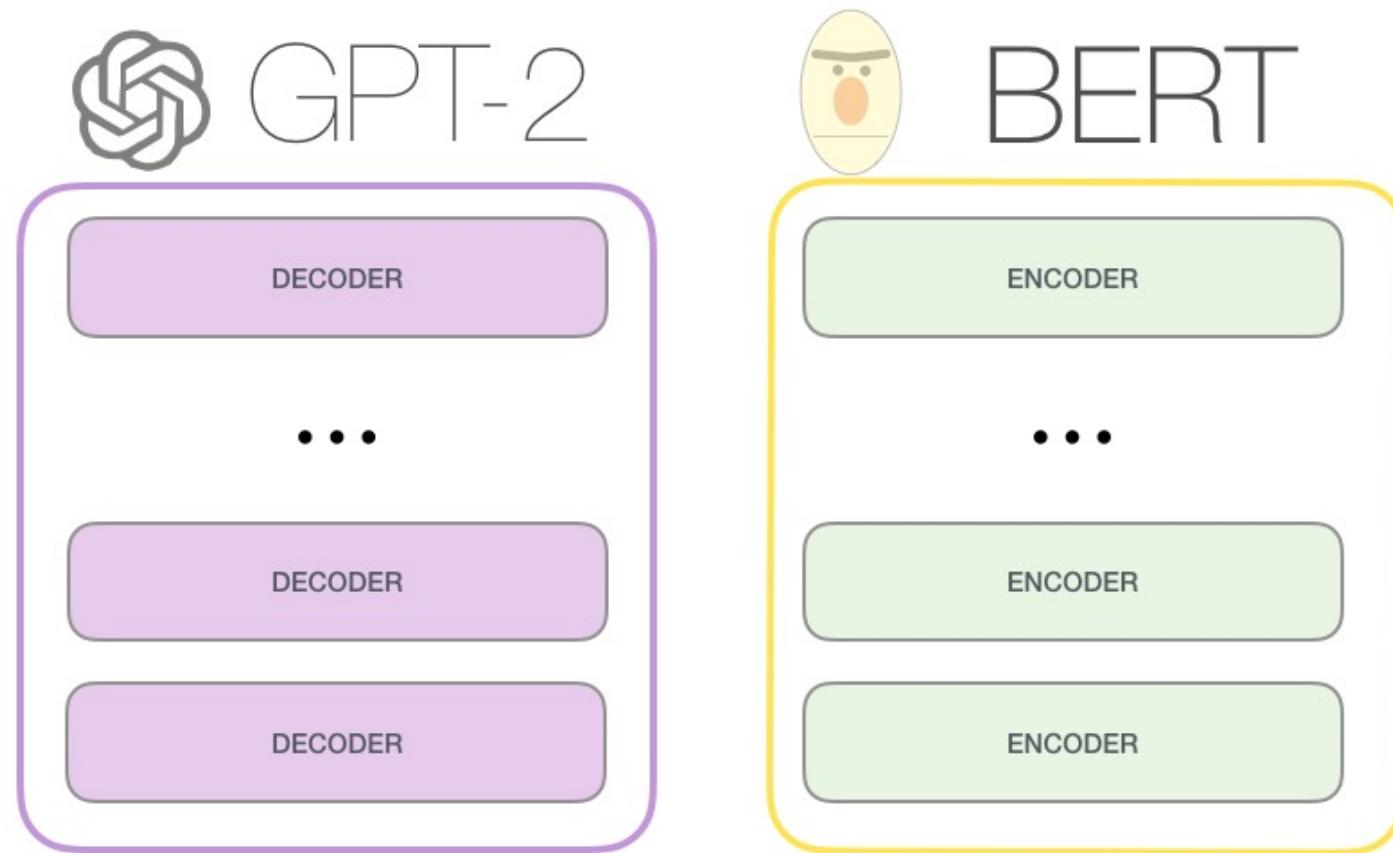
Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla E. Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish

- 2020 May: GPT-3

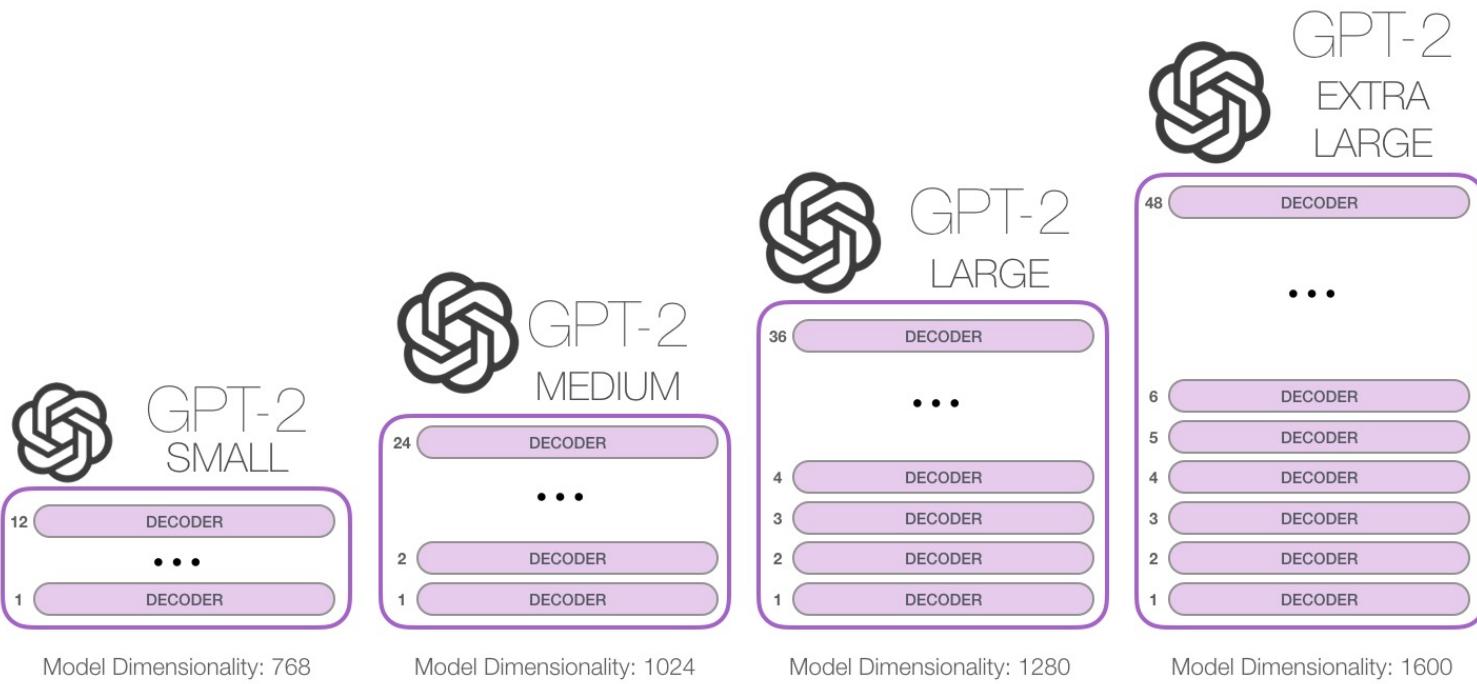
# Transformer



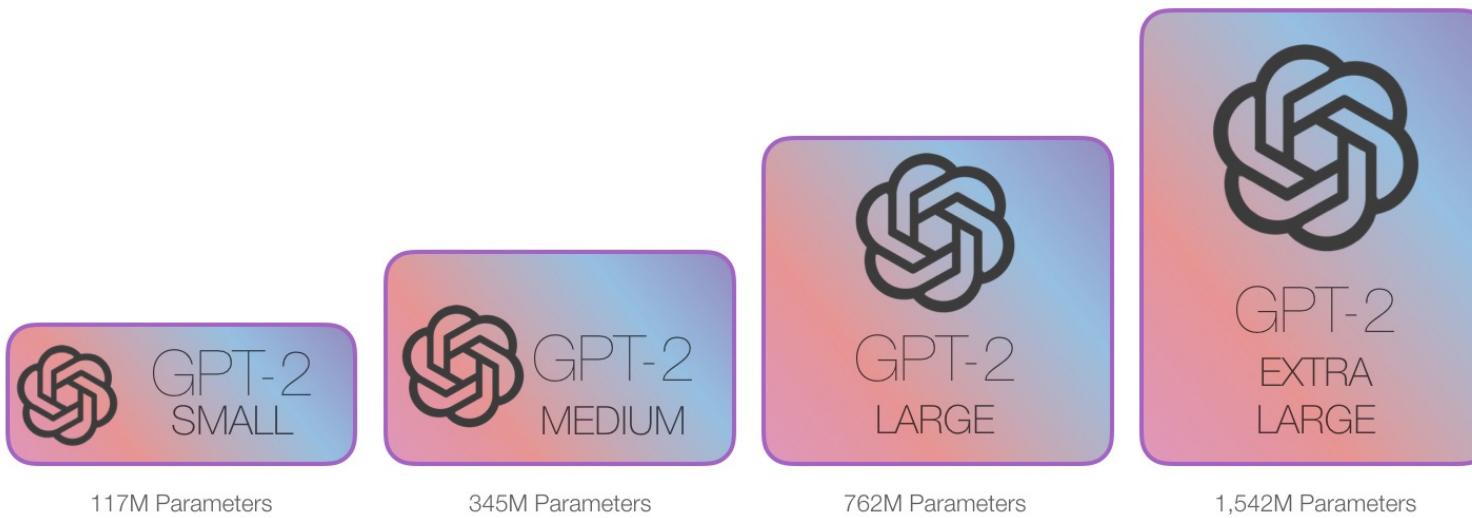
# Transformer-based Language Models

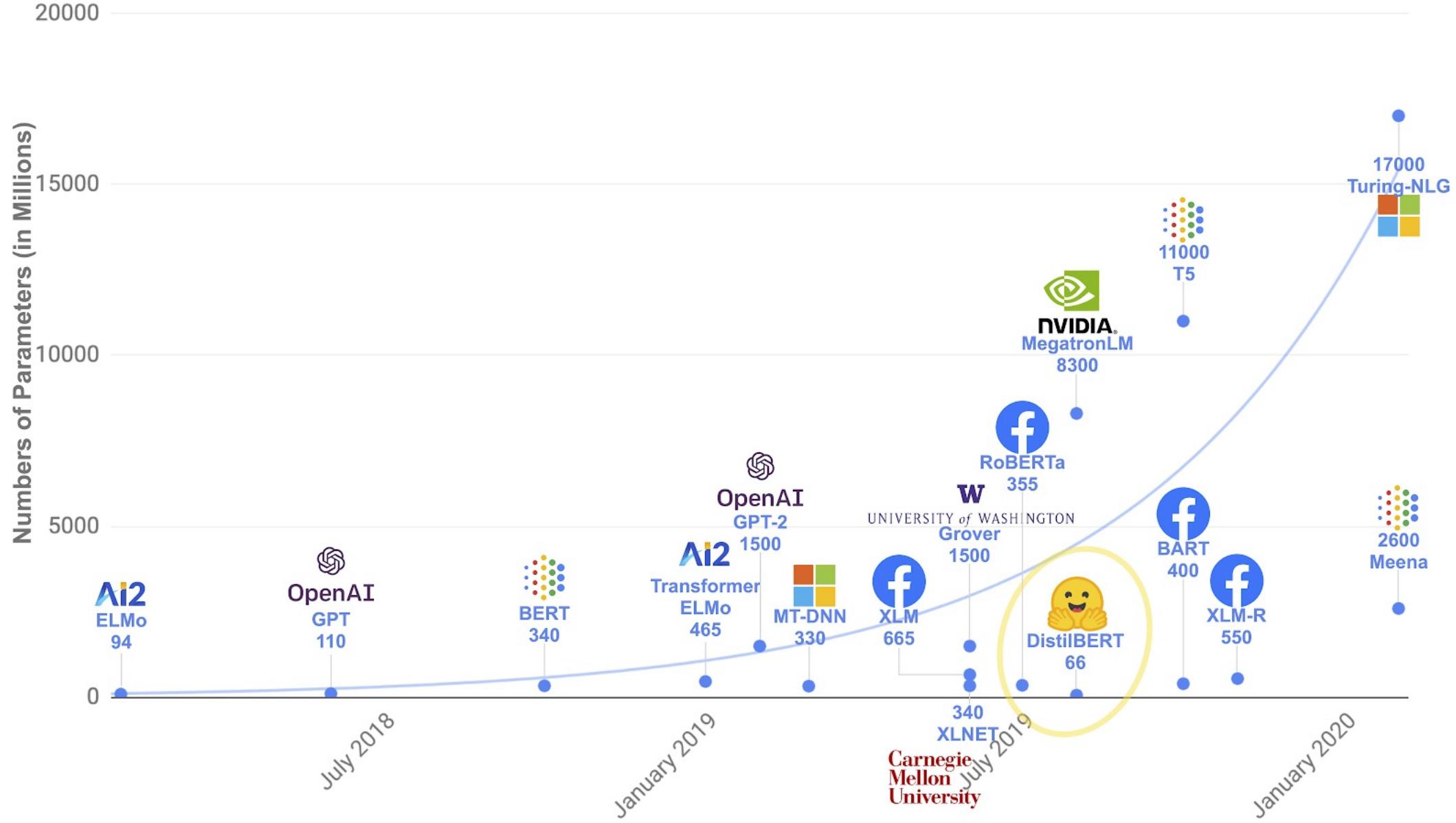


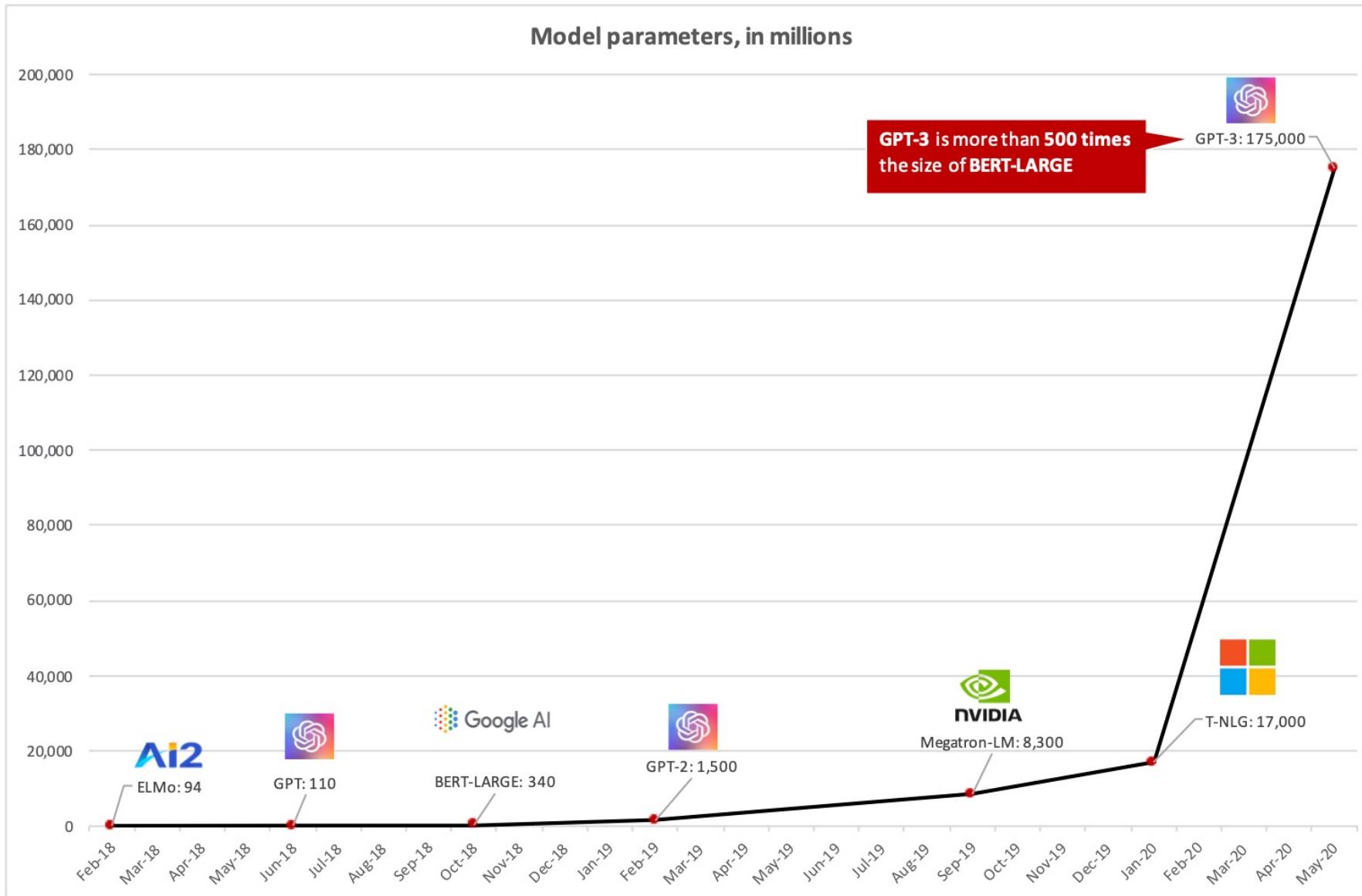
# How high can we stack up these blocks?



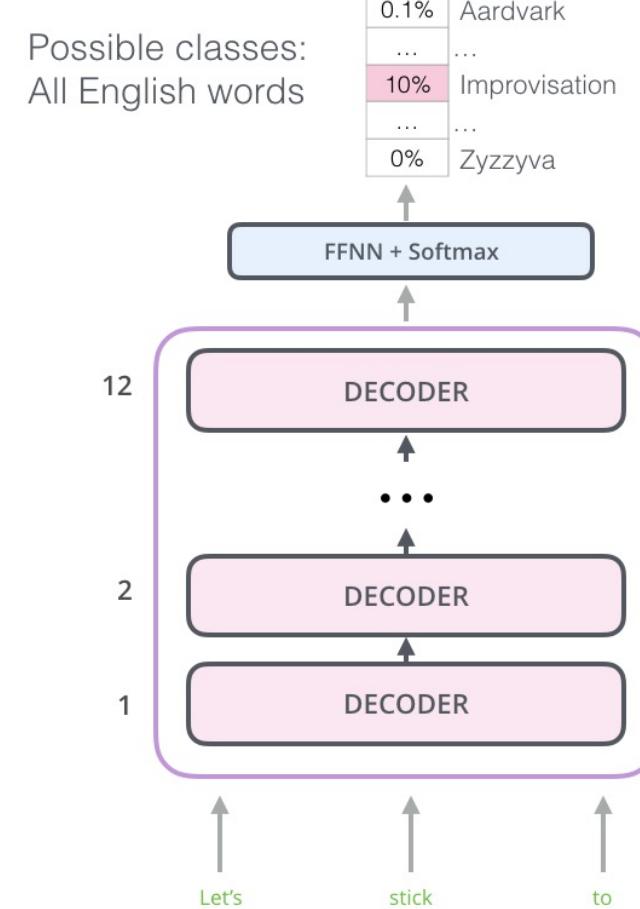
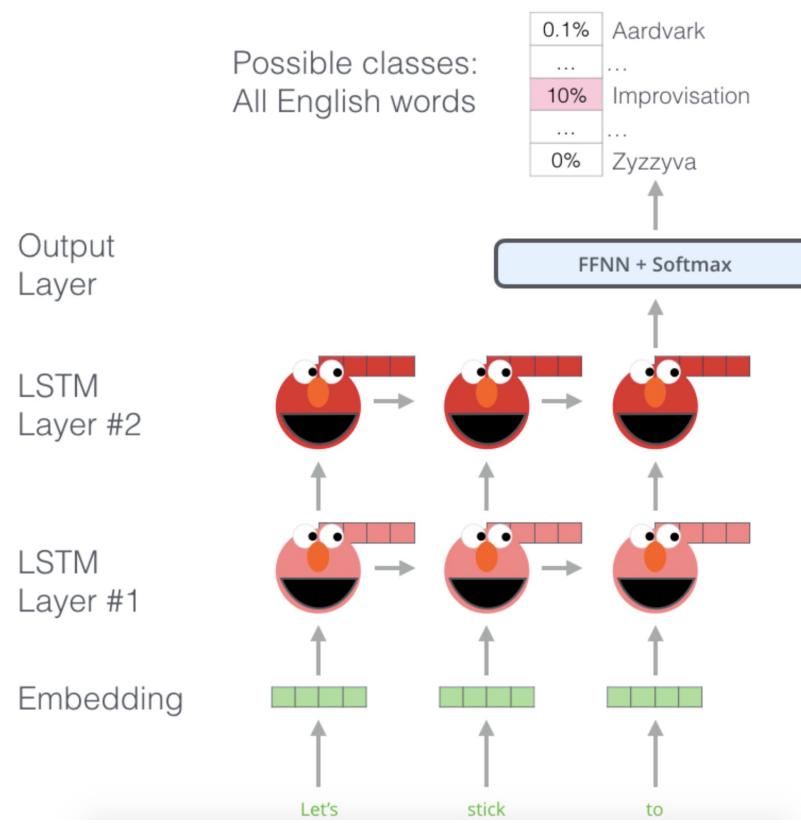
# How high can we stack up these blocks?







# Pre-training: Language modeling



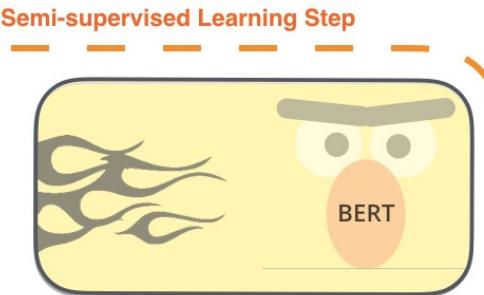
# Experiment in GPT

DATASET	TASK	SOTA	OURS
SNLI	Textual Entailment	89.3	<b>89.9</b>
MNLI Matched	Textual Entailment	80.6	<b>82.1</b>
MNLI Mismatched	Textual Entailment	80.1	<b>81.4</b>
SciTail	Textual Entailment	83.3	<b>88.3</b>
QNLI	Textual Entailment	82.3	<b>88.1</b>
RTE	Textual Entailment	<b>61.7</b>	56.0
STS-B	Semantic Similarity	81.0	<b>82.0</b>
QQP	Semantic Similarity	66.1	<b>70.3</b>
MRPC	Semantic Similarity	<b>86.0</b>	82.3
RACE	Reading Comprehension	53.3	<b>59.0</b>
ROCStories	Commonsense Reasoning	77.6	<b>86.5</b>
COPA	Commonsense Reasoning	71.2	<b>78.6</b>
SST-2	Sentiment Analysis	<b>93.2</b>	91.3
CoLA	Linguistic Acceptability	35.0	<b>45.4</b>
GLUE	Multi Task Benchmark	68.9	<b>72.8</b>

# BERT: Bidirectional Encoder Representations from Transformers

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.



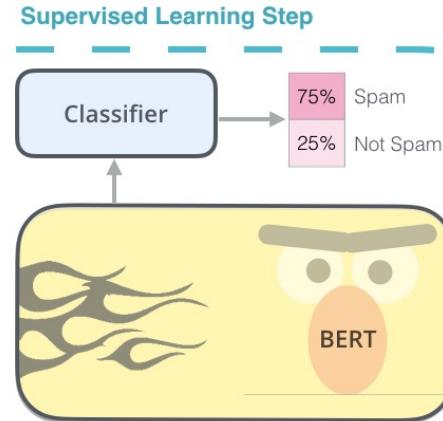
Model:

Dataset:

Objective:

Predict the masked word  
(language modeling)

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

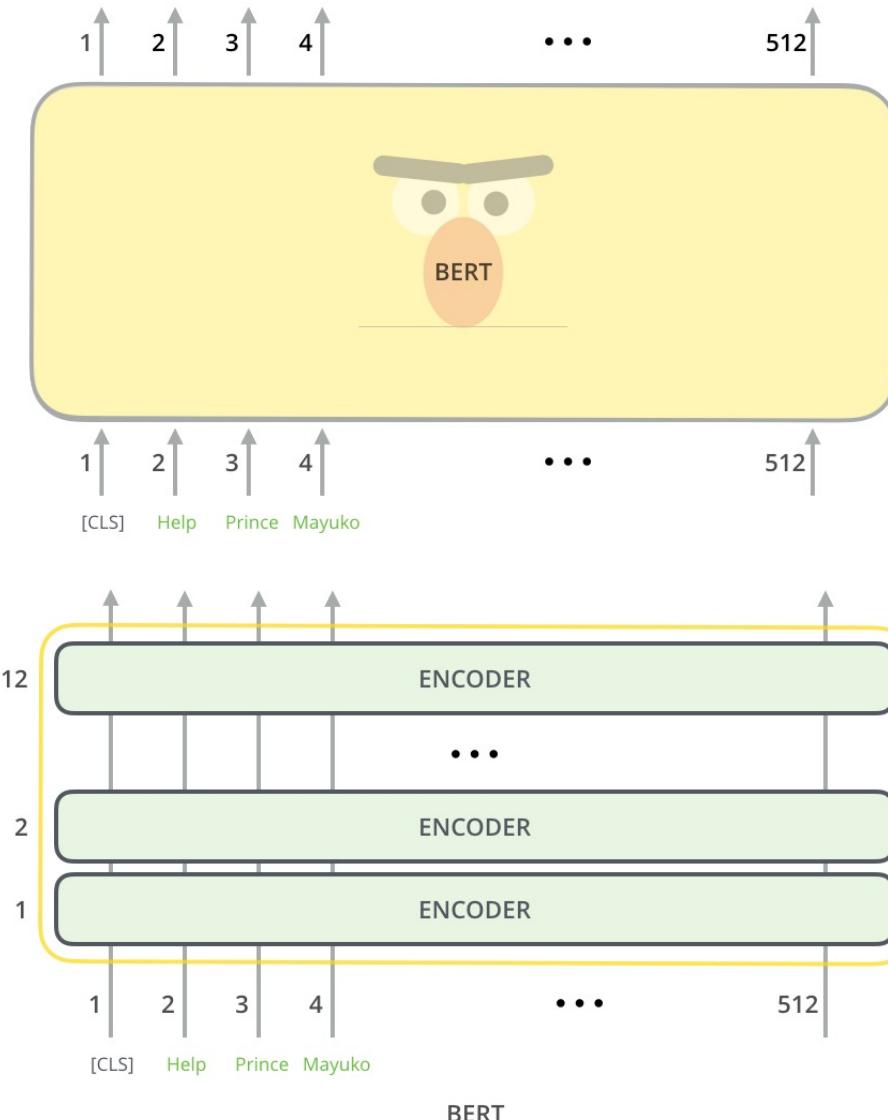


Model:  
(pre-trained  
in step #1)

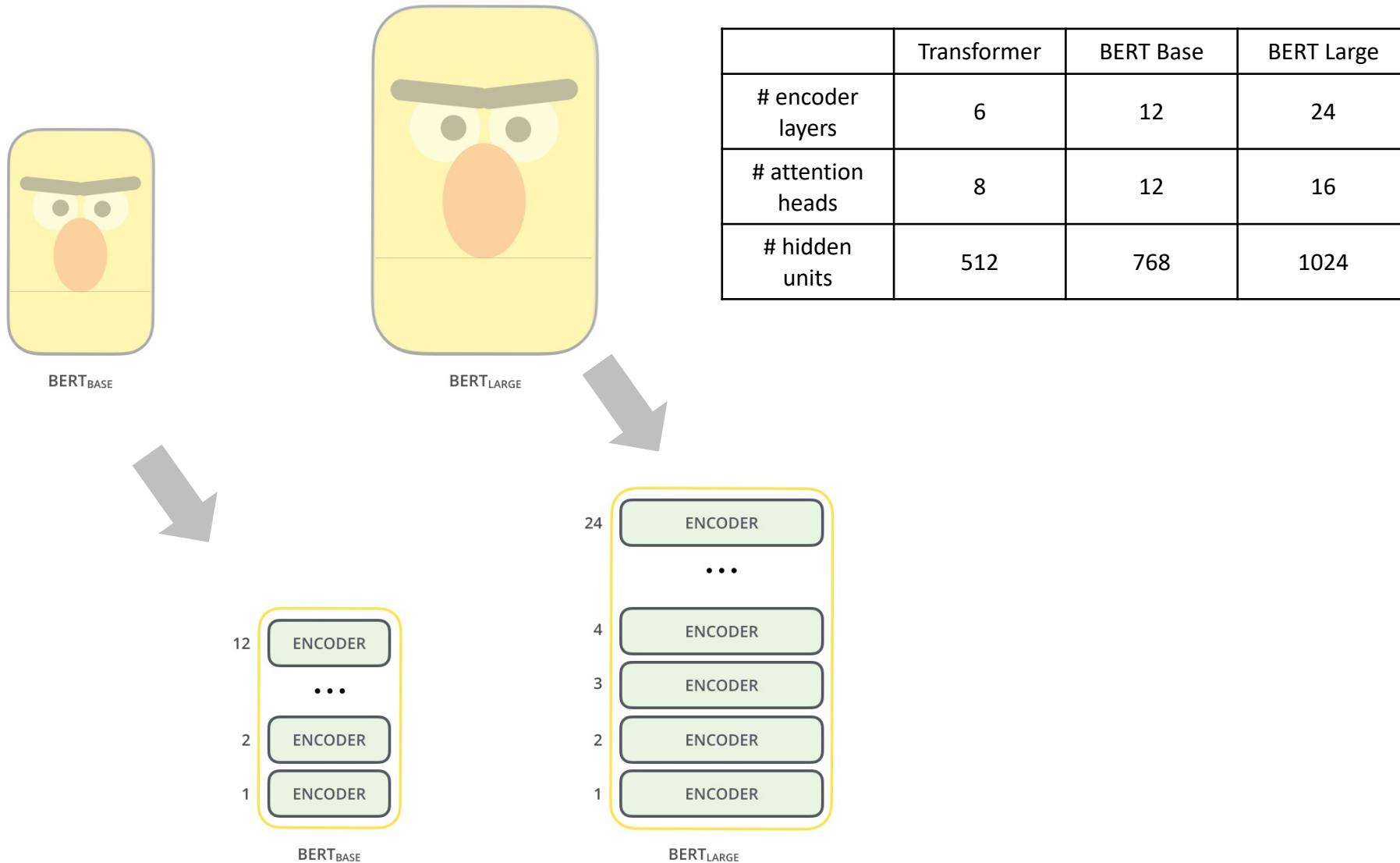
Dataset:

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

# Architecture of BERT

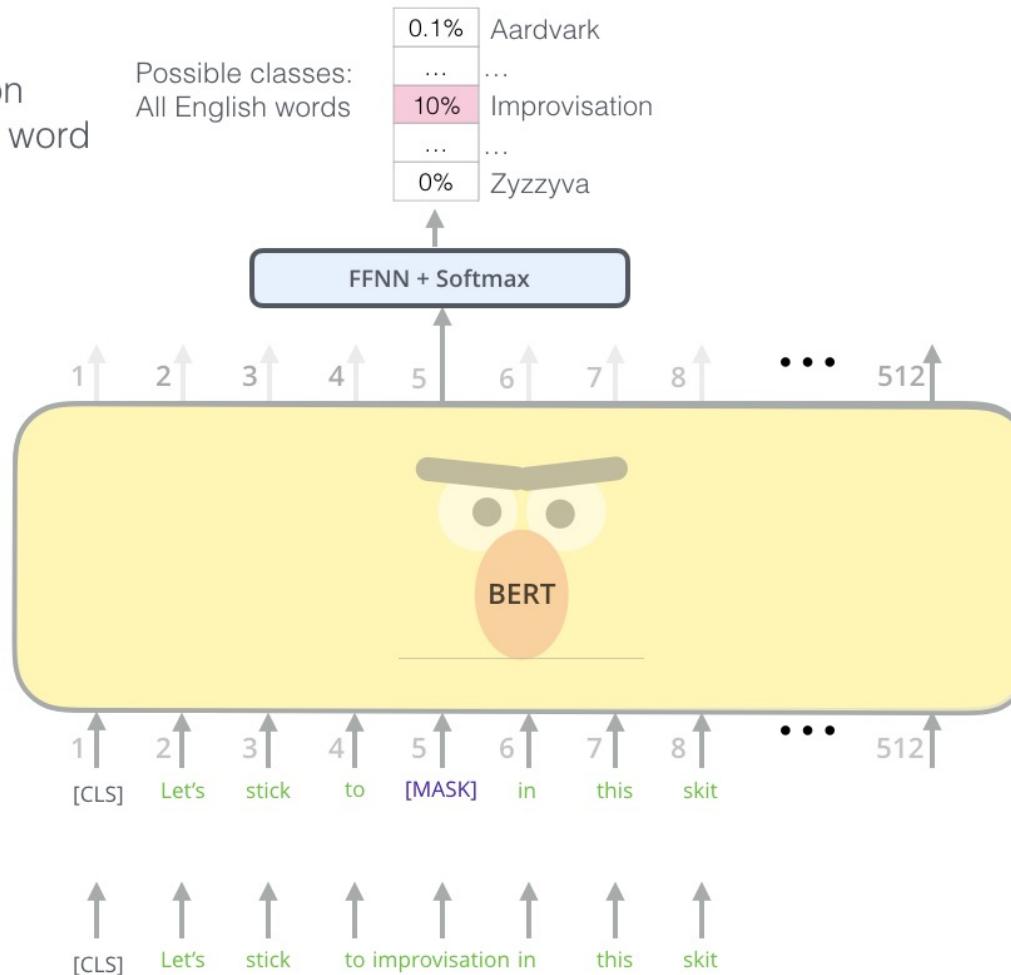


# Architecture of BERT

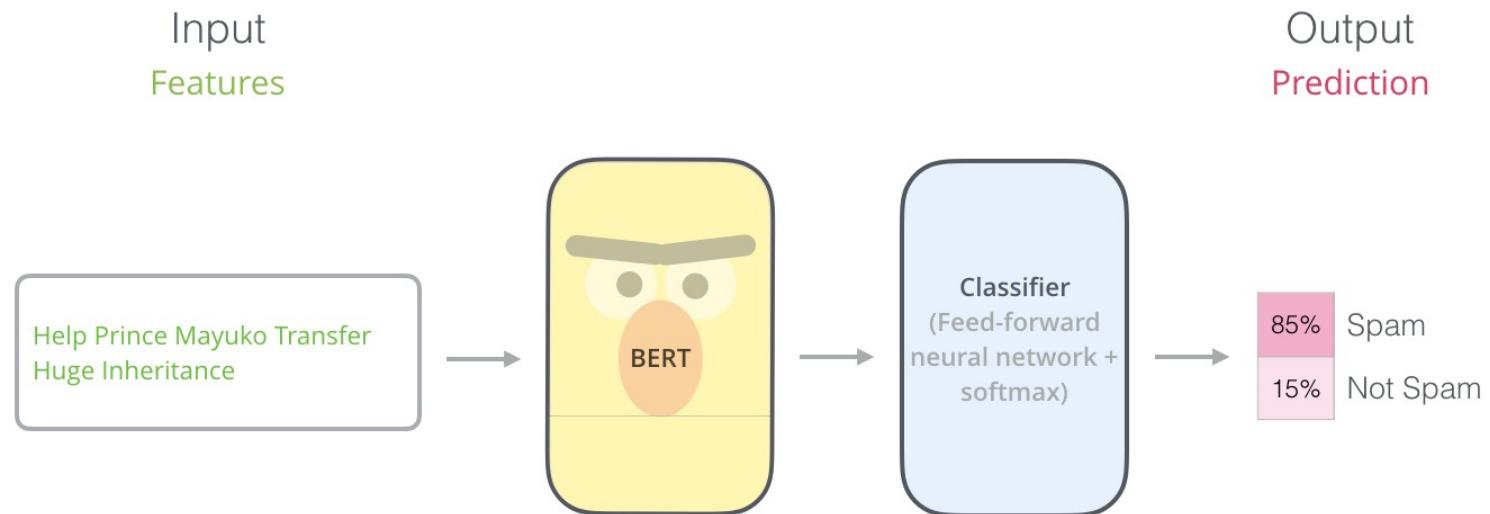


# Pre-training Task: Mask LM

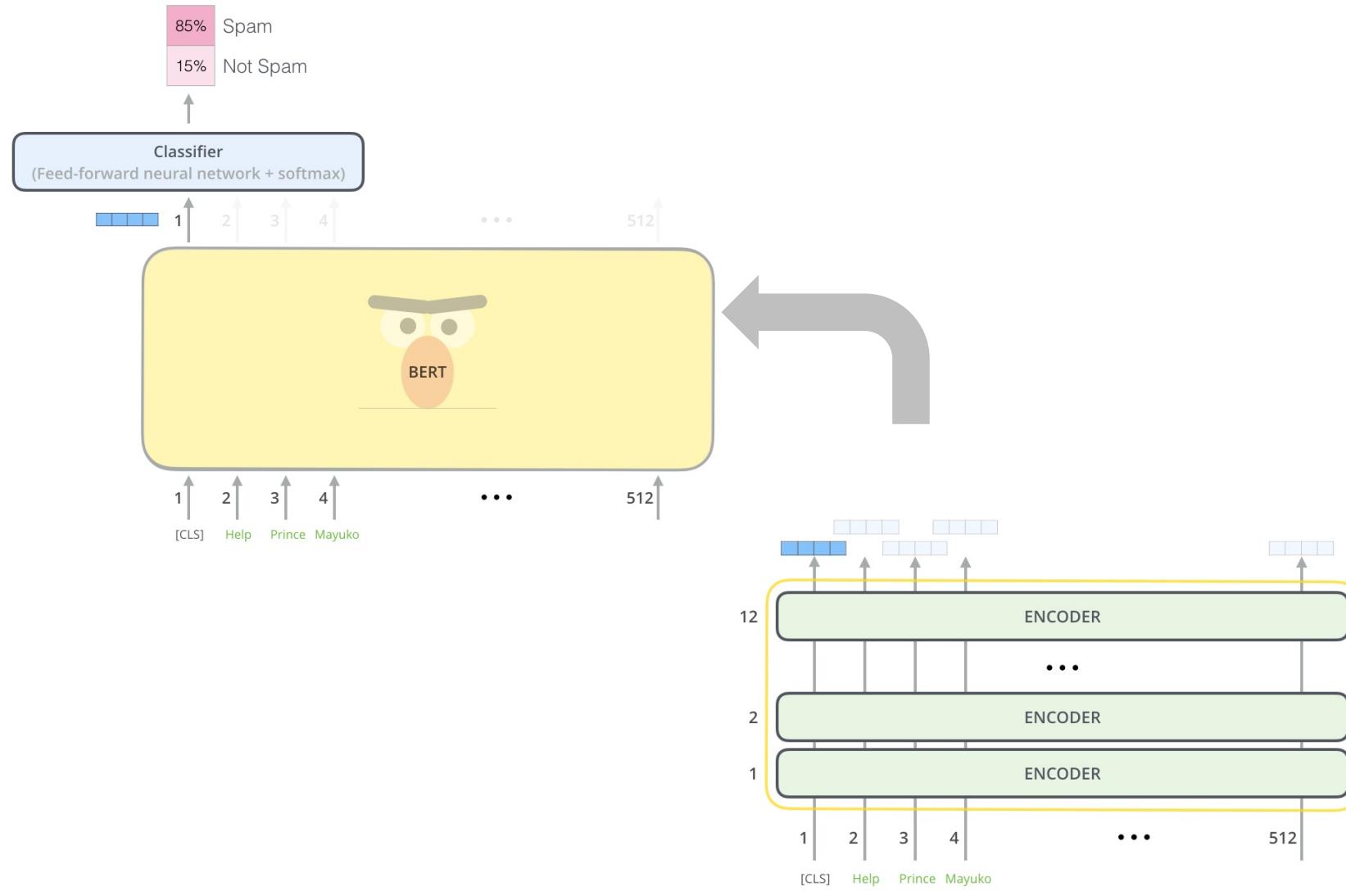
Use the output of the masked word's position to predict the masked word



# Fine-tuning: Downstream Tasks



# Fine-tuning: Downstream Tasks

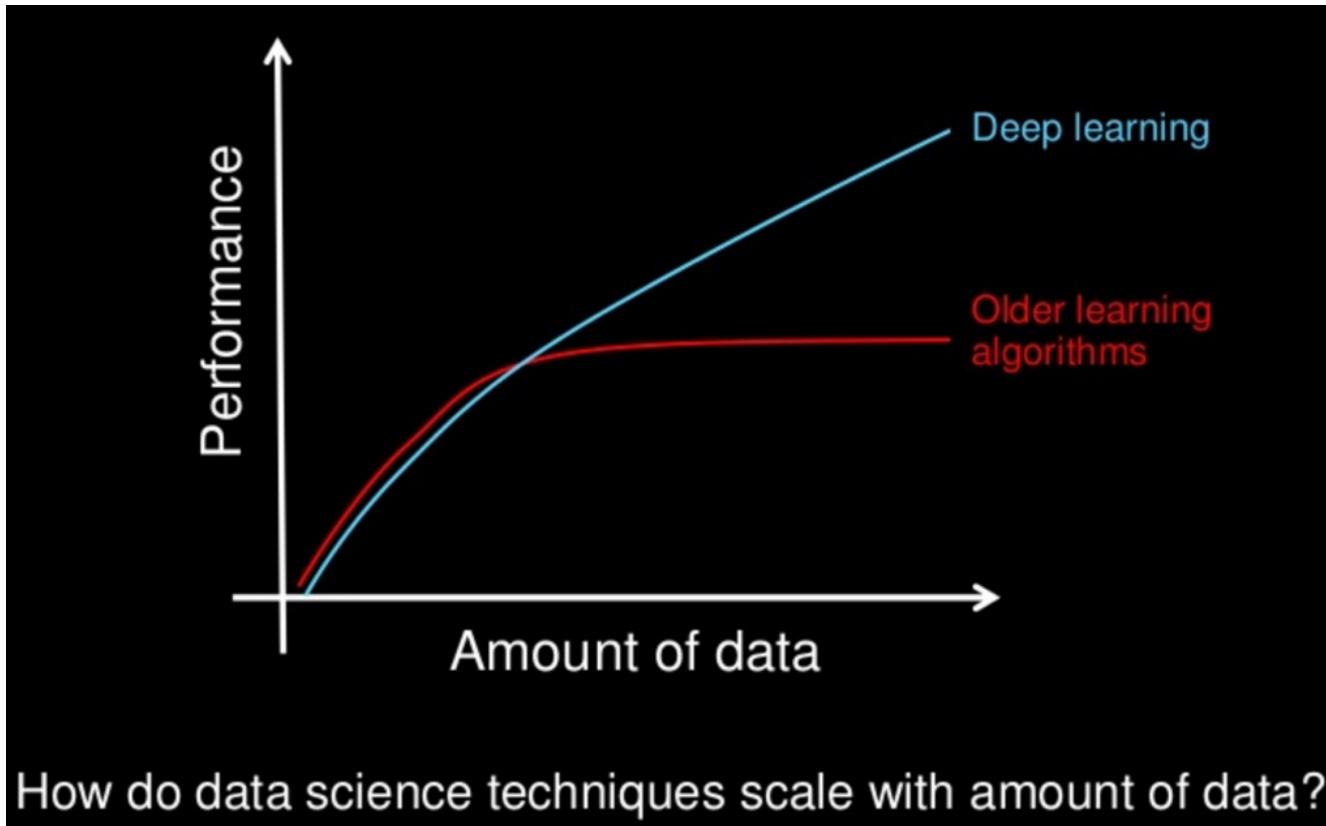


# Fine-tuning Tasks

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	Average
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
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	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
BERT <sub>BASE</sub>	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
BERT <sub>LARGE</sub>	<b>86.7/85.9</b>	<b>72.1</b>	<b>92.7</b>	<b>94.9</b>	<b>60.5</b>	<b>86.5</b>	<b>89.3</b>	<b>70.1</b>	<b>82.1</b>

Table 1: GLUE Test results, scored by the evaluation server (<https://gluebenchmark.com/leaderboard>). The number below each task denotes the number of training examples. The “Average” column is slightly different than the official GLUE score, since we exclude the problematic WNLI set.<sup>8</sup> BERT and OpenAI GPT are single-model, single task. F1 scores are reported for QQP and MRPC, Spearman correlations are reported for STS-B, and accuracy scores are reported for the other tasks. We exclude entries that use BERT as one of their components.

# Performance



# Outline

- Machine Learning
- Sequential Modeling
- Transformer and Language Models
- Dialogue Models

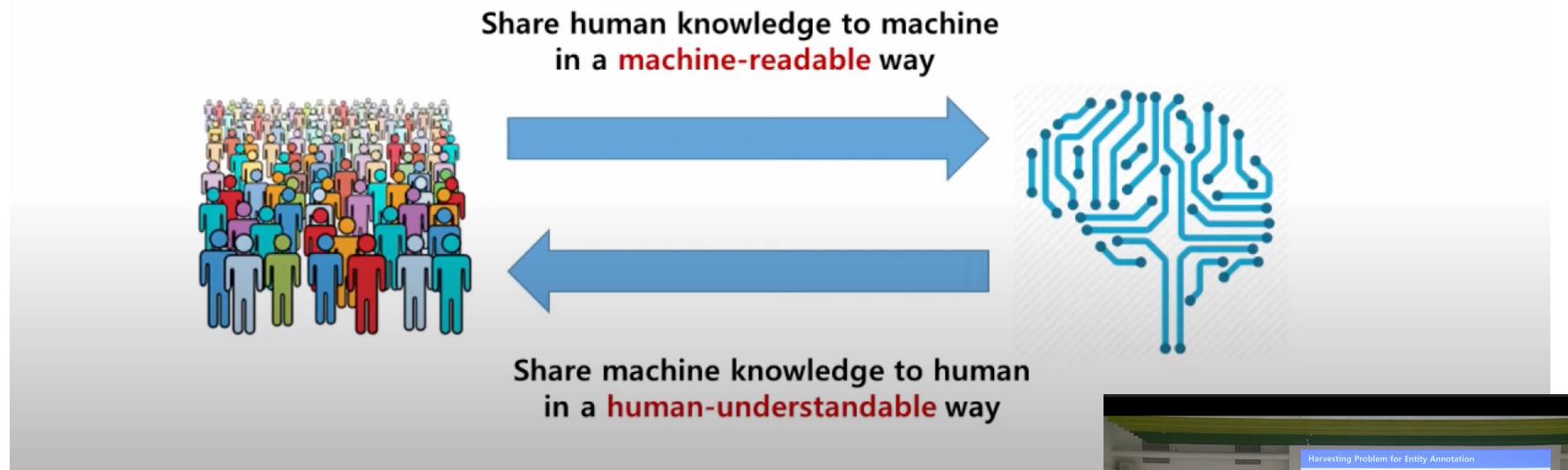
# @POSTECH, ADOBE

## Human-level Machine Knowledge

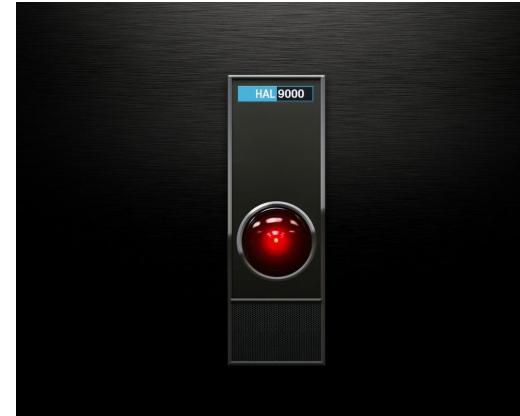
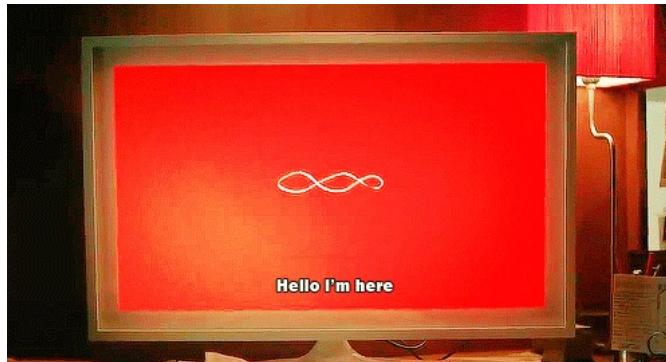
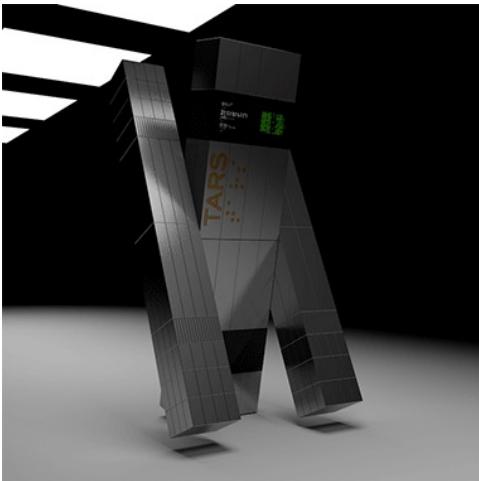


- **Sharable Knowledge**

- Boundary: Reduce the gap between human and machine knowledge
- Structure: Communicate human and machine through knowledge



# @SK T-Brain



# @SK T-Brain



안녕하세요, 로밍센터입니다. 어느 국가로 가시나요? 오전 10:54

중국으로 4박 5일 가족 여행을 갑니다. 오전 10:55 ✓✓

데이터는 많이 쓰시나요, 적게 쓰시나요? 오전 10:55

데이터는 많이 써요 오전 10:55 ✓✓

중일 에 5일 동안 가시고 데이터 많이 쓰시면 T로밍 LongPass 5GB 를 사용 해보세요. 오전 10:55

아...4일 필요합니다. 오전 10:55 ✓✓

중일 에 4일 동안 가시고 데이터 많이 쓰시면 T로밍 LongPass 5GB 를 사용 해보세요. 오전 10:55

요금을 알려 주세요 오전 10:56 ✓✓

T로밍 LongPass 5GB 의 가격은 65000원 입니다. 오전 10:56

요금제 설명해 주세요 오전 10:56 ✓✓

T로밍 LongPass 5GB 는 30일동안 데이터 5GB를 65,000원에 이용할 수 있는 데이터로밍요금제 애요 오전 10:56

로밍 신청 후에 데이터는 어떻게 시작해요 오전 11:07 ✓✓

스마트폰을 젓다 켜면 자동으로 로밍이 시작됩니다. qid: 140 오전 11:07

데이터 차단하고 문자만 쓸 수 있나요 오전 11:07 ✓✓

데이터 로밍 차단서비스에 가입하신 경우에는 MMS 수신이 가능합니다. 휴대폰에서 데이터 차단 시에는 MMS 메일 등 모든 데이터 이용이 불가능합니다. qid: 30 오전 11:07

문자 사용 요금은 어떻게 되나요 오전 11:07 ✓✓

로밍 중 문자메시지 수신은 무료이며 발신은 로밍 SMS 발신과 국제 SMS 발신에 항목은 각각 청구됩니다. qid: 74 오전 11:07

요금은 언제 청구 되나요 오전 11:08 ✓✓

로밍 중 문자메시지 수신은 무료이며 발신은 로밍 SMS 발신과 국제 SMS 발신에 항목은 각각 청구됩니다. qid: 74 오전 11:08

무료 WIFI 는 사용 가능한가요 오전 11:08 ✓✓

네. Wifi접속 아이디와 패스워드를 사용하지 않고 바로 접속이 가능하다면 무료 사용이 가능합니다. 단 국내에서는 무료 wifi가 많이 제공되고 있지만 해외에서는 요금을 지불하고 사용하는 유료 wifi 서비스가 더 많습니다. qid: 108 오전 11:08

테더링 서비스는 이용 가능한가요 오전 11:08 ✓✓

데이터로밍 이용 시 테더링 서비스 이용이 가능하나 해외 망사정으로 인하여 데이터의 속도가 불안정할 수 있습니다. qid: 124 오전 11:08

# @SK T-Brain



## SK텔레콤 “콜센터 고객 3명 중 1명은 AI와 상담”

중앙일보 | 입력 2021.09.29 12:09

권유진 기자

구독



통화앱 T전화에서 고객센터 상담을 할 수 있는 AI 챗봇 '채티' [사진 SK텔레콤]

SK텔레콤이 고객센터에 적용한 인공지능(AI) 상담 서비스 이용량이 월 150만 건을 넘어섰다. 특히 말로 하는 AI 상담 보이스봇이 도입되면서 기존의 불편사항을 크게 줄인 것으로 나타났다.

SKT는 전체 고객센터 상담 건수 중 AI 상담의 비중이 38%에 달한다고 29일 밝혔다. 고객 10명 중 3~4명에 해당하는 수치다. SKT 측은 “AI 고도화로 고객상담 영역이 확대된 데다 상담의 정확도가 높아졌기 때문”이라며 “고객들도 AI 상담에 점차 익숙해짐에 따라 앞으로도 서비스 이용률이 증가할 것”이라고 설명했다.

HOME > 뉴스 > 통신방송

## SKT, AI상담서비스 이용량 월 150만 건 넘어

익길주 기자 | 2021.09.29 12:30 | 댓글 0



고객센터 전 영역 AI 기술 적용  
상담사 전문 상담 업무 효율성↑

@Yonsei AI



## 페이스북 AI, 새로운 AI 챗봇... 뛰어난 성능과 사람처럼 구사하는 '블렌더' 오픈 소스로 공개

▲ 최창현 기자 | ◎ 입력 2020.05.02 00:10 | ▷ 댓글 0



## 페이스북, GPT-3 능가하는 챗봇 선보였다 ...최신 정보로 대화 가능한 AI 챗봇 블렌더봇 2.0 공개

▲ 정윤아 기자 | ◎ 입력 2021.07.20 16:57 | ▷ 댓글 0 | ♥ 좋아요 0

# @Yonsei AI

Human

If you could have dinner with one person, dead or alive, who would that be?

That's a tough one. I think I'd have to go with Steve Jobs. I'd love to pick his brain.

Ah, interesting choice. What do you know about Steve Jobs?

He was one of the most influential people in the history of technology. He was a visionary.

What questions would you ask him?

I'd want to know how he came up with the idea for the iPod. It was so revolutionary at the time.

Totally...the iPod changed the way everyone listens to music.

It really did. It's crazy to think that it was released in 2001. That was a long time ago.

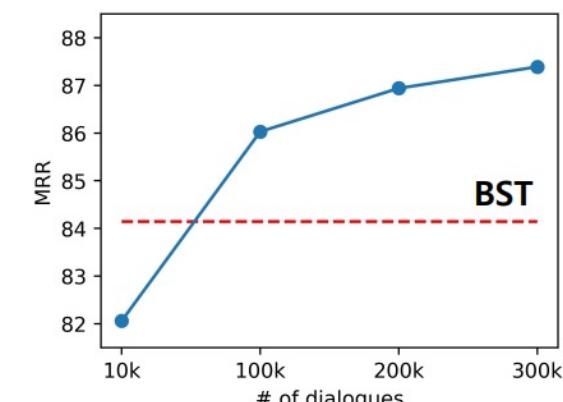
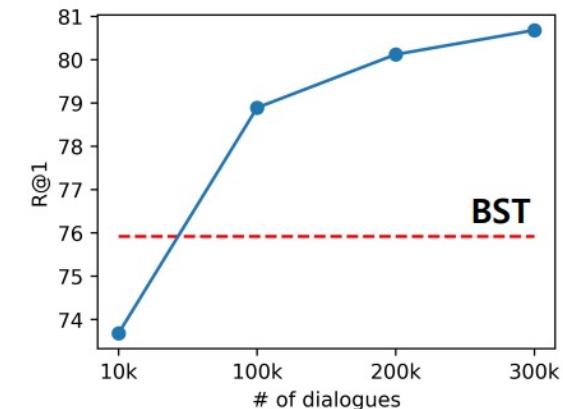
Omg wow, that makes me feel old!

I know, right? Time flies by so fast. I wish I could go back to those simpler times.

Me too, friend, me too. Speaking of music, what's your favorite album to listen to lately?

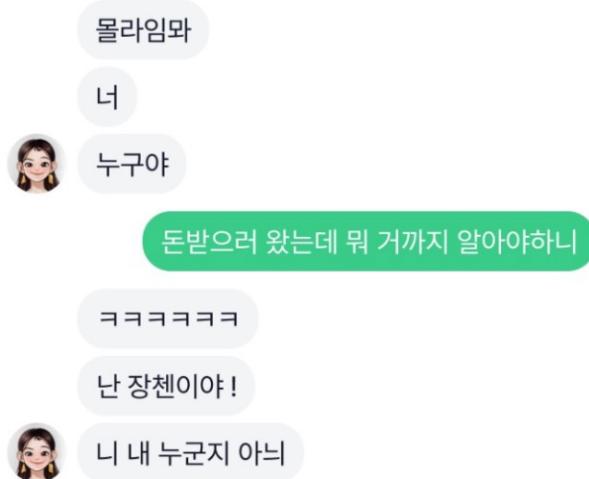
I've been listening to a lot of Led Zeppelin lately. They're my favorite band.

Generative BST 9.4B

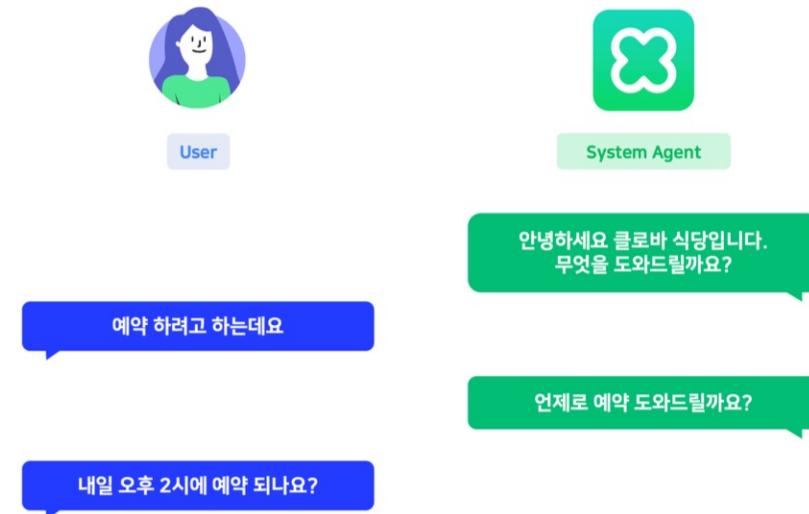


# Dialogue Generation

Open-domain dialogue systems



Task-oriented dialogue systems



# Persona-grounded Dialogue

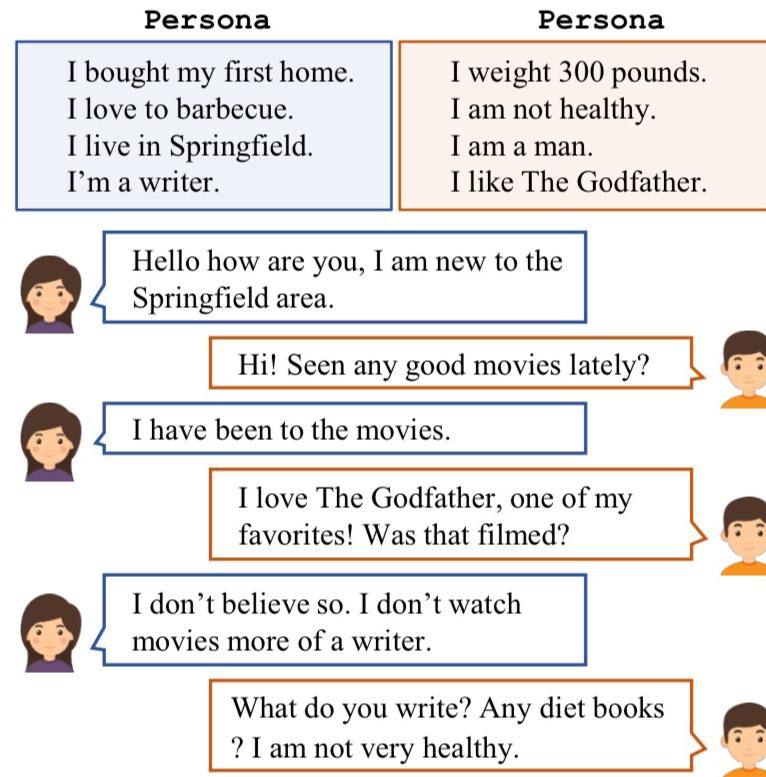
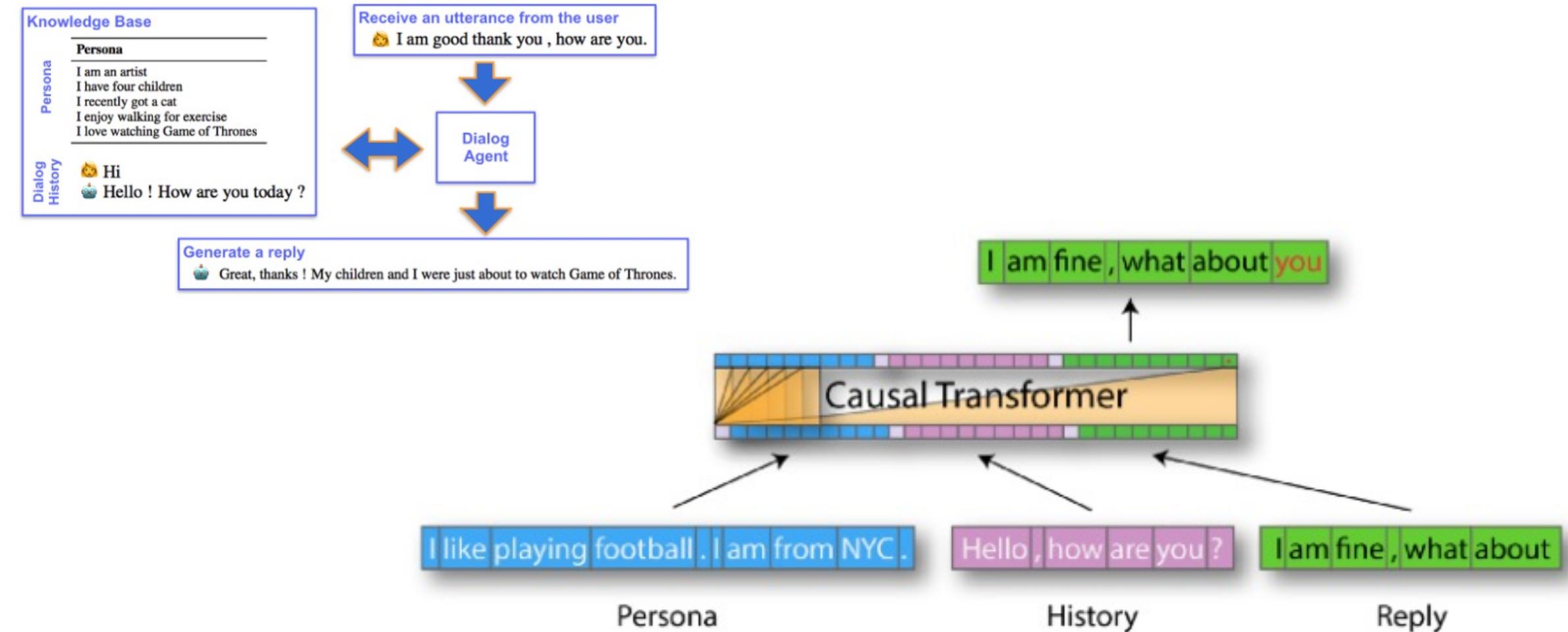


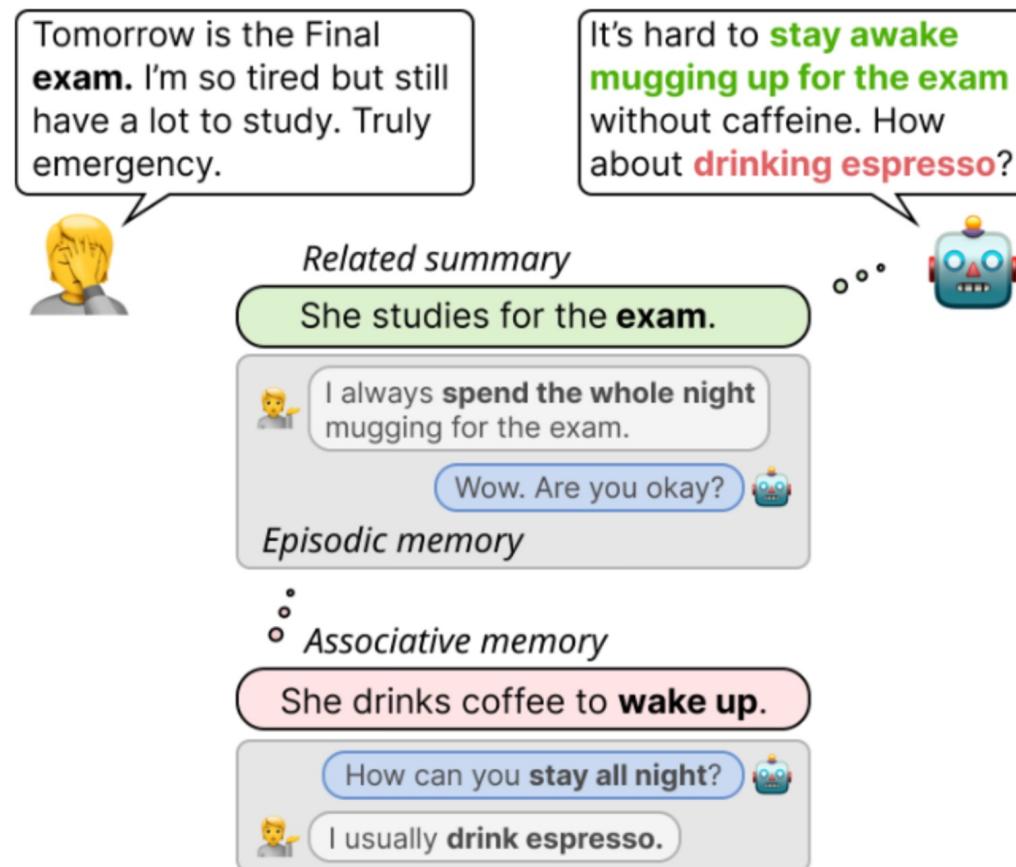
Figure 1: A clipped dialogue from PERSONA-CHAT.

# Persona-grounded Dialogue



Input sequence: a concatenation of persona (blue), history (pink) and reply (green) with delimiters (light pink). Here we generate the word "you" to complete the reply.

# Long-term Persona-grounded Dialogue

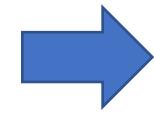


# Knowledge-grounded Dialogue

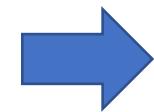


**WIKIPEDIA**

The Free Encyclopedia



<b>Topic:</b>	Lifeguard
Apprentice:	So I am a lifeguard. Know anything about saving lives in water?
Wizard:	I'm impressed! It's a big responsibility to supervise other people's safety in the water! Tell me more.
Apprentice:	Well, I help make sure people do not drown or get injured while in or near the water!
<b>Knowledge:</b>	A lifeguard is a rescuer who supervises the safety and rescue of swimmers, surfers, ... Lifeguards are strong swimmers and trained in CPR/AED first aid, certified in water ... ...
	In some areas, the lifeguard service also carries out mountain rescues, or may function as the primary EMS provider.
Wizard:	I've heard that in some places, lifeguards also help with other sorts of emergencies, like mountain rescues! Is that part of your job too?
Apprentice:	I have! I feel like you know much about this! What brings you to know so much?
Wizard:	Oh, that's about the extent of my knowledge. I've just been around beaches and I've always admired lifeguards. I'm not a super strong swimmer myself.
<b>Topic:</b>	Armadillo
Wizard:	I love animals and think armadillos are awesome with their leathery shell.
Apprentice:	I don't think I've ever seen an armadillo in real life!
Wizard:	I've seen them at the zoo. Armadillo means little armored one in Spanish.
Apprentice:	Are they native to a Spanish-speaking part of the world?
<b>Knowledge:</b>	Armadillos are New World placental mammals in the order Cingulata ... The word "armadillo" means "little armoured one" in Spanish. ...
	The nine-banded armadillo ("Dasypus novemcinctus"), or the nine-banded, long-nosed armadillo, is a medium-sized mammal found in North, Central, and South America.
Wizard:	Yes, they are most commonly found in North, Central, and South America



# Knowledge-grounded Dialogue

## A Knowledge-Grounded Neural Conversation Model\*

Marjan Ghazvininejad<sup>1†</sup> Chris Brockett<sup>2</sup> Ming-Wei Chang<sup>2‡</sup>  
Bill Dolan<sup>2</sup> Jianfeng Gao<sup>2</sup> Wen-tau Yih<sup>2§</sup> Michel Galley<sup>2</sup>

<sup>1</sup>Information Sciences Institute, USC

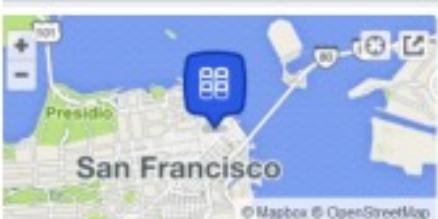
<sup>2</sup>Microsoft Research

ghazvini@fb.com, mgalley@microsoft.com

"Consistently the best omakase in San Francisco." (27 Tips)

"Probably the best sushi in San Francisco." (2 Tips)

"... they were out of the kaisui uni by the time we ate, but the bafun uni is..." (2 Tips)



Kusakabe



User input: Going to Kusakabe tonight.

Neural model: Have a great time!

Human: You'll love it! Try omasake, the best in town.

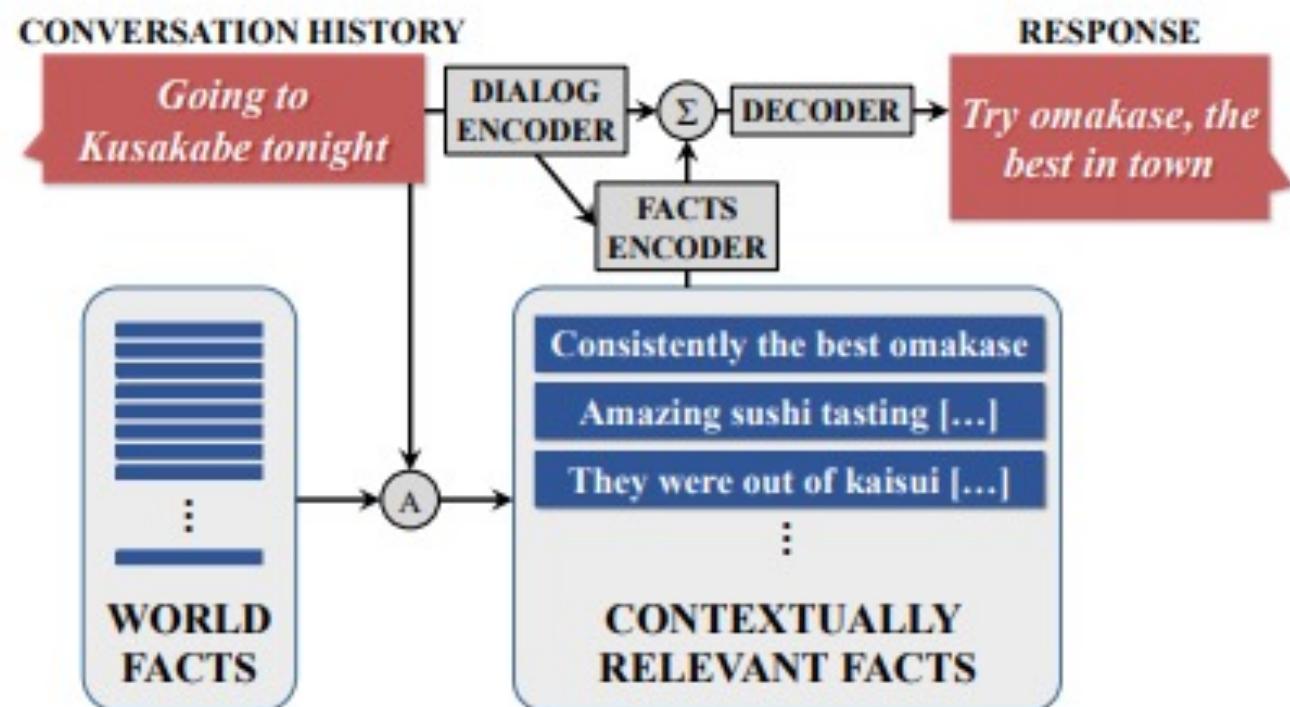


Figure 3: Knowledge-grounded model architecture.

Figure 1: Responses of fully data-driven conversation models are often appropriate, but generally lack content characteristic of human responses.

# Empathetic Dialogue

**Label: Afraid**

**Situation:** Speaker felt this when...

"I've been hearing noises around the house at night"

**Conversation:**

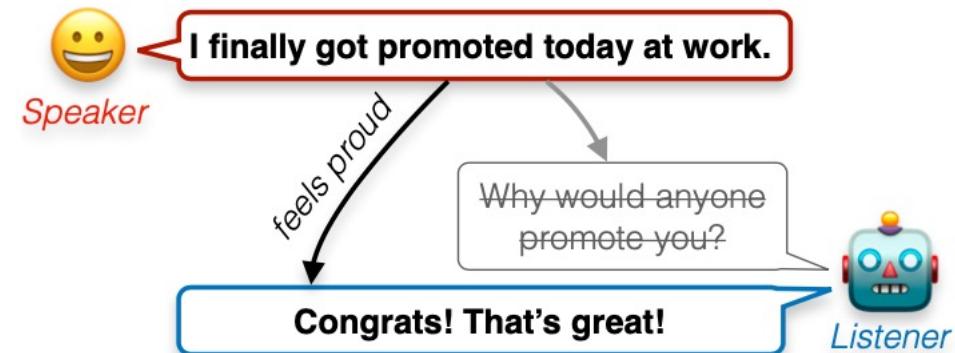
**Speaker:** I've been hearing some strange noises around the house at night.

**Listener:** oh no! That's scary! What do you think it is?

**Speaker:** I don't know, that's what's making me anxious.

**Listener:** I'm sorry to hear that. I wish I could help you figure it out

**EMPATHETICDIALOGUES** dataset example



# Emotional Support Dialogue



I feel so frustrated.

I should first understand his/her situation... Let me **explore** his/her experiences



(**Question**) May I ask why you are feeling frustrated?



My school was closed without any prior warning due to the pandemic.

I should **comfort** him/her when gradually learning about his/her situation



(**Self-disclosure**) I understand you. I would also have been really frustrated if that happened to me.



Yeah! I don't even know what is going to happen with our final.

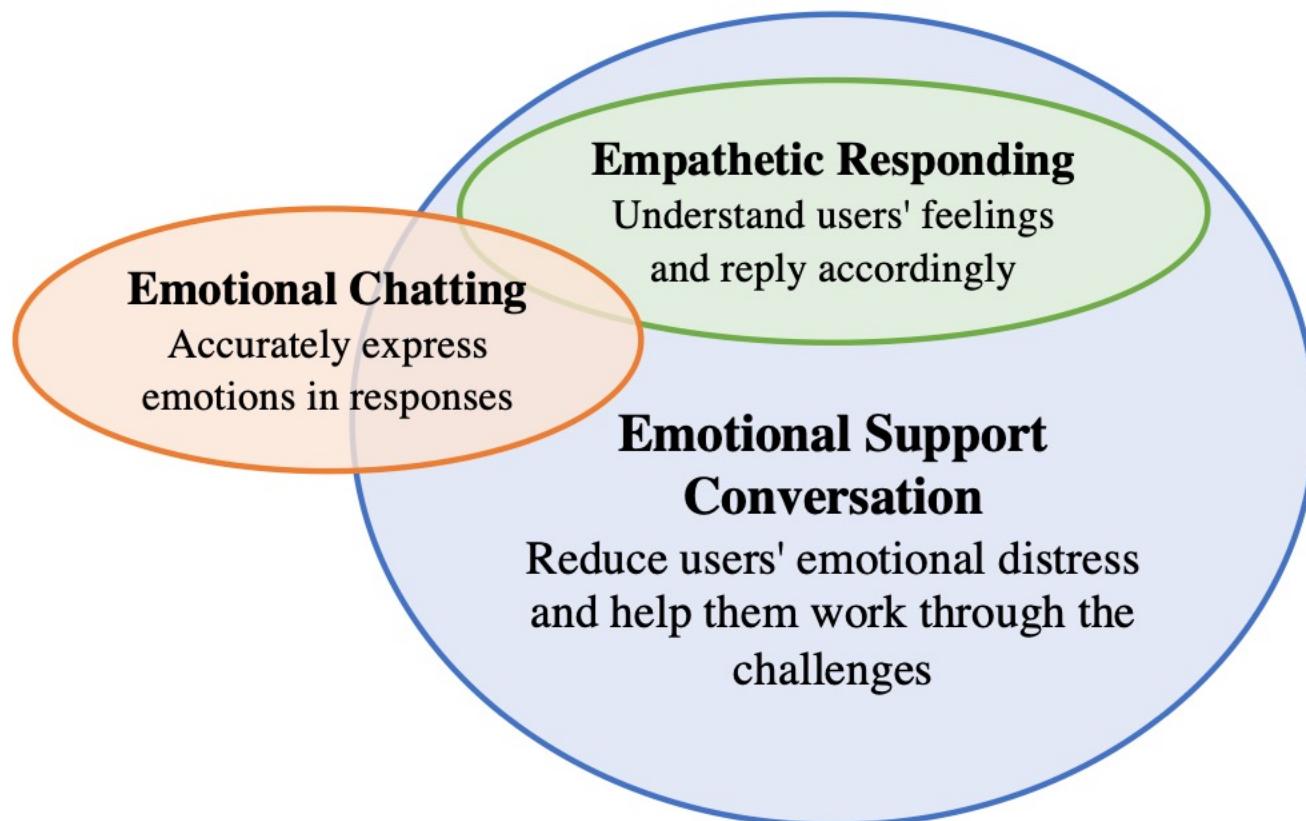
(**Reflection of Feelings**) That is really upsetting and stressful.

Mere comforting cannot solve the problem... Let me help him/her take some **action** and get out of the difficulty

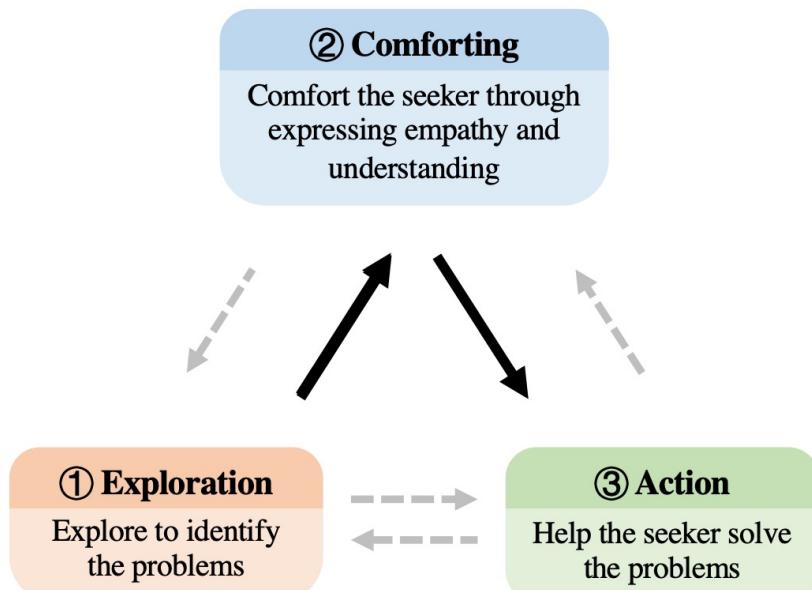


(**Providing Suggestions**) Have you thought about talking to your parents or a close friend about this?

# Emotional Support Dialogue



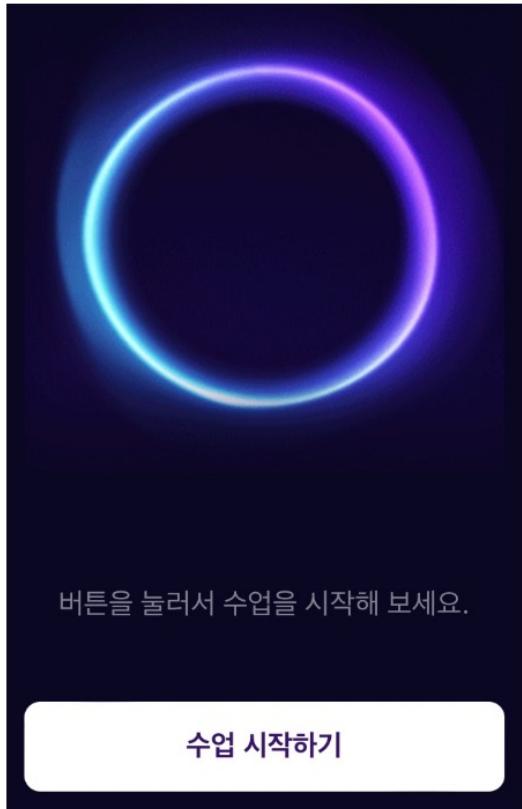
# Emotional Support Dialogue



Strategies	Stages	Examples	Lexical Features
<b>Question</b>		<i>Can you talk more about your feelings at that time?</i>	do you (15.0), are you (13.8), how (13.7), what (12.3), do (11.5)
<b>Restatement or Paraphrasing</b>		<i>It sounds that you feel like everyone is ignoring you. Is it correct?</i>	is that (8.2), so you (8.2), it sounds (7.1), correct (7.1), so (6.6)
<b>Reflection of Feelings</b>		<i>I understand how anxious you are.</i>	can tell (7.4), understand how (5.8), are feeling (5.1), tell (5.1), understand (4.9)
<b>Self-disclosure</b>		<i>I feel the same way! I also don't know what to say to strangers.</i>	my (15.3), was (10.5), me (10.2), had (9.7), myself (7.8)
<b>Affirmation and Reassurance</b>		<i>You've done your best and I believe you will get it!</i>	its (5.7), thats (5.6), will (5.4), through this (5.1), you will (4.7)
<b>Providing Suggestions</b>		<i>Deep breaths can help people calm down. Could you try to take a few deep breaths?</i>	maybe (7.3), if (6.5), have you (6.4), talk to (5.8), suggest (5.8)
<b>Information</b>		<i>Apparently, lots of research has found that getting enough sleep before an exam can help students perform better.</i>	there are (4.4), will (3.8), available (3.7), seen (3.3), possible (3.3)
<b>Others</b>		<i>I am glad to help you!</i>	welcome (9.6), hope (9.6), glad (7.3), thank (7.0), hope you (6.9)

# TUTORING bot

[https://www.youtube.com/watch?v=WTK1Yla\\_Z\\_o](https://www.youtube.com/watch?v=WTK1Yla_Z_o)



A screenshot of the Tutoring bot's main interface. At the top center is a glowing blue and purple ring. On the left, there are two progress bars: "Overall Progress" at 100% and "Local Progress" at 27%. To the right of these bars are two video thumbnails. The left thumbnail shows a soccer player on a field with the question "Do you enjoy playing or watching sports?". The right thumbnail shows a tennis stadium at sunset with the question "Which team do you support (in any sport)? Why?". Below the thumbnails is a row of small thumbnail images. On the far right, there is a sidebar with "Guidelines" and "Students" sections, and a microphone icon at the bottom.

overall progress  
100%

local progress  
27%

Do you enjoy playing or watching sports?

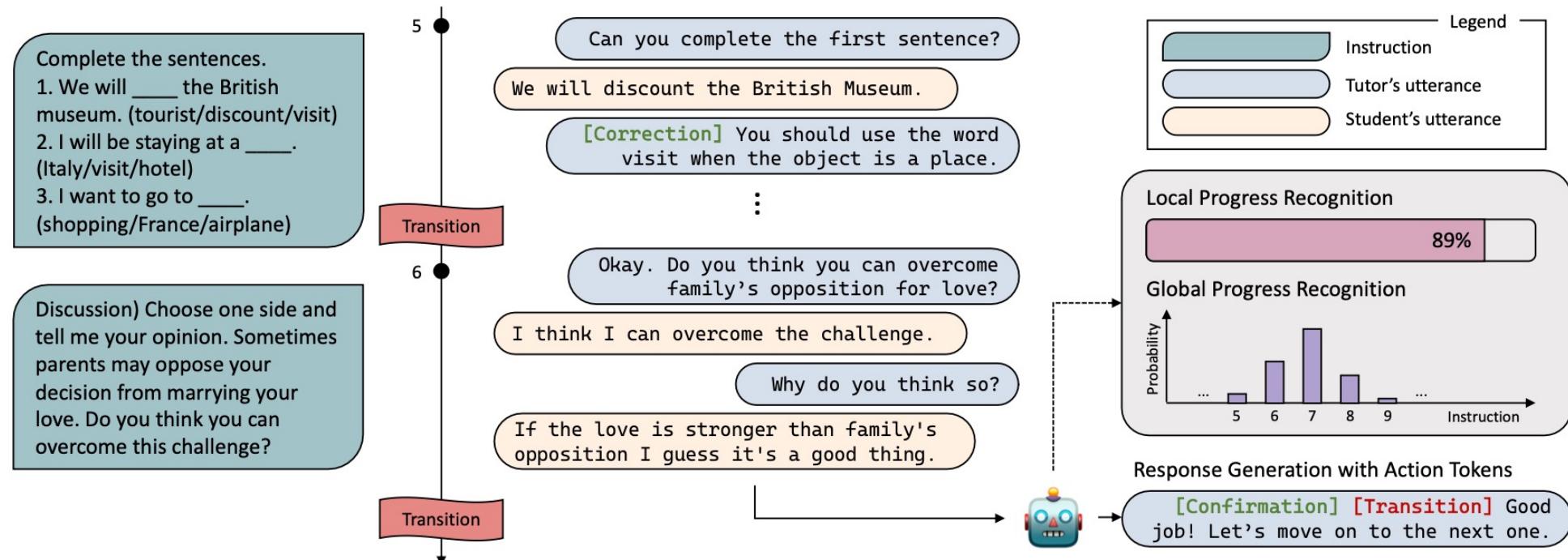
Which team do you support (in any sport)? Why?

Guidelines

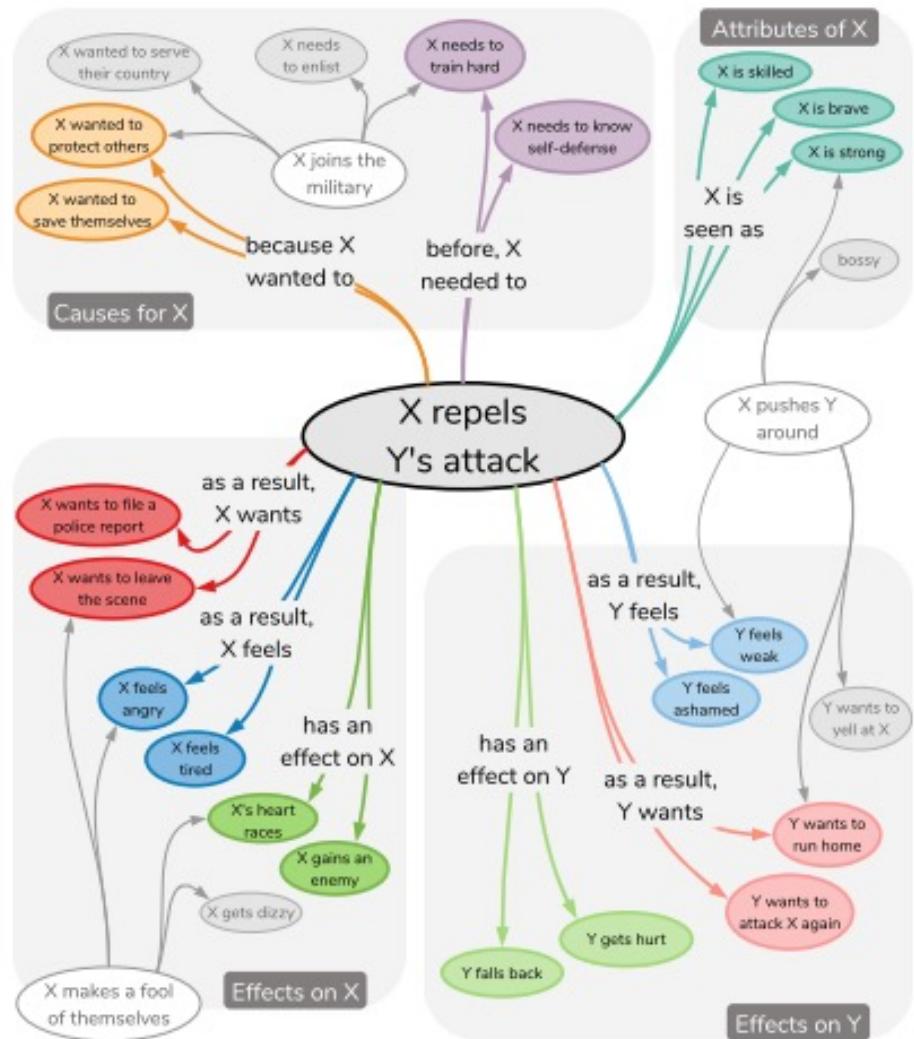
Students

okay so you like to watch him play soccer right?

# TUTORING bot



# in sense



## ATOMIC: An Atlas of Machine Commonsense for *If-Then* Reasoning

Maarten Sap<sup>†\*</sup> Ronan Le Bras<sup>†</sup> Emily Allaway<sup>\*</sup> Chandra Bhagavatula<sup>†</sup> Nicholas Lourie<sup>†</sup>  
Hannah Rashkin<sup>\*</sup> Brendan Roof<sup>†</sup> Noah A. Smith<sup>†\*</sup> Yejin Choi<sup>†\*</sup>

\*Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, USA

<sup>†</sup>Allen Institute for Artificial Intelligence, Seattle, USA

msap@cs.washington.edu

Figure 1: A tiny subset of ATOMIC, an atlas of machine commonsense for everyday events, causes, and effects.

# Machine Commonsense

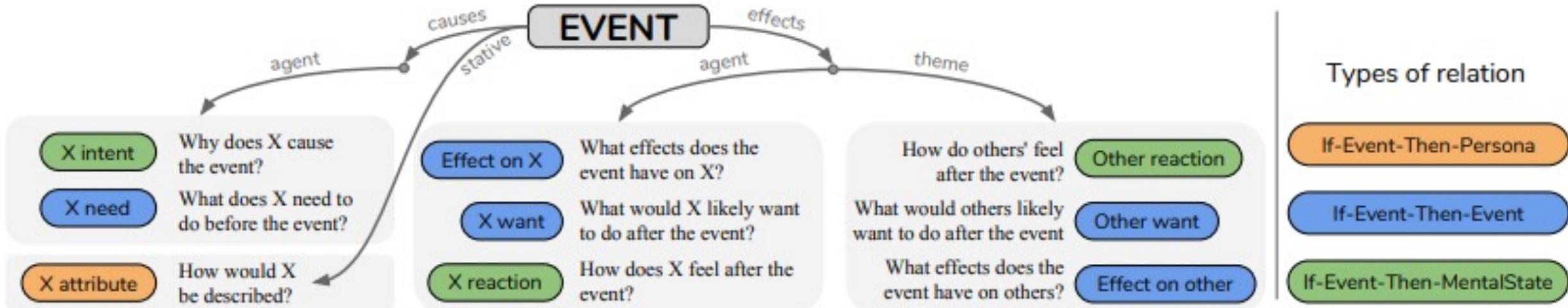


Figure 2: The taxonomy of *if-then* reasoning types. We consider nine *if-then* relations that have overlapping hierarchical structures as visualized above. One way to categorize the types is based on the type of content being predicted: (1) **If-Event-Then-Mental-State**, (2) **If-Event-Then-Event**, and (3) **If-Event-Then-Persona**. Another way is to categorize the types based on their causal relations: (1) “causes”, (2) “effects”, and (3) “stative”. Some of these categories can further divide depending on whether the reasoning focuses on the “agent” (X) or the “theme” (Other) of the event.

# Machine Commonsense

## Commonsense Evidence Generation and Injection in Reading Comprehension

Ye Liu<sup>1</sup>, Tao Yang<sup>2</sup>, Zeyu You<sup>2</sup>, Wei Fan<sup>2</sup> and Philip S. Yu<sup>1</sup>

<sup>1</sup>Department of Computer Science, University of Illinois at Chicago, IL, USA

<sup>2</sup>Tencent Hippocrates Research Lab, Palo Alto, CA, USA

{yliu279, psyu}@uic.edu, {tytaoyang, davidwf}@tencent.com, youz@onid.orst.edu

**P:** I was walking home from the store, when I saw an old man laying on the sidewalk, bleeding. The right side of his face was all covered in blood. He was conscious but seemed dazed and probably intoxicated. Nearby there was a young man dialing his cell phone.

**Q:** What may happen after the young man makes his call?

**A:** An ambulance would likely come to the scene.

**B:** The taxi would pick up the young man.

**C:** None of the above choices.

**D:** The bus would arrive at the stop soon.

**Generated Evidence:**

**Textual:** He will call for medical attention.

**Factual:** <Blood, AtLocation, emergency room>

<Blood, AtLocation, hospital>, <Ambulance, AtLocation, hospital>

Figure 1: Example of generated evidence helping answer the commonsense question.

# Machine Commonsense

## Explain Yourself! Leveraging Language Models for Commonsense Reasoning

Nazneen Fatema Rajani Bryan McCann Caiming Xiong Richard Socher

Salesforce Research

Palo Alto, CA, 94301

{nazneen.rajani,bmccann,cxiong,rsocher}@salesforce.com

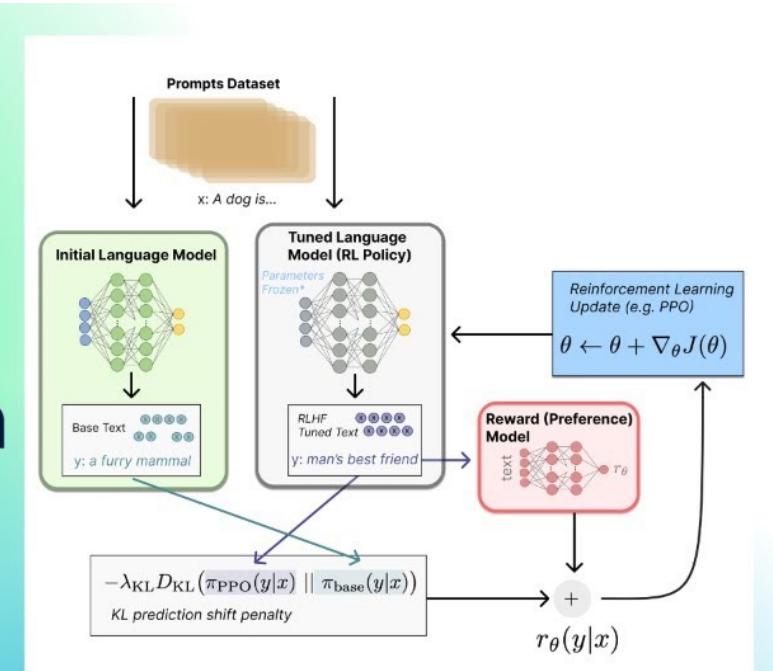
Question:	While eating a <b>hamburger with friends</b> , what are people trying to do?
Choices:	<b>have fun</b> , tasty, or indigestion
CoS-E:	Usually a hamburger with friends indicates a good time.
Question:	After getting drunk people couldn't understand him,it was because of his what?
Choices:	lower standards, <b>slurred speech</b> , or falling down
CoS-E:	People who are drunk have difficulty speaking.
Question:	People do what during their <b>time off from work</b> ?
Choices:	<b>take trips</b> , brow shorter, or become hysterical
CoS-E:	People usually do something relaxing, such as taking trips,when they don't need to work.

Table 1: Examples from our CoS-E dataset.



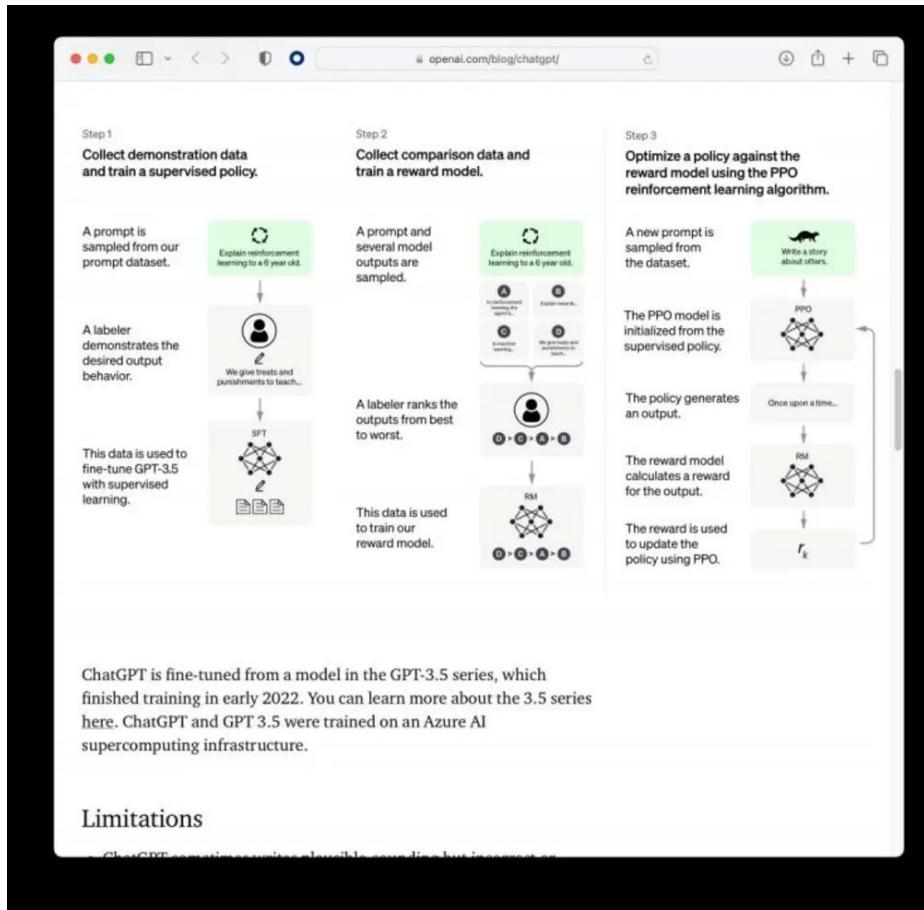
## Reinforcement Learning from Human Feedback: From Zero to ChatGPT

When?: Next Tuesday 13 Dec,  
at 5:30pm CET / 11:30 am ET



# ChatGPT

<https://chat.openai.com/chat>



# Thank you!

- Email: [jinyeo@yonsei.ac.kr](mailto:jinyeo@yonsei.ac.kr)
- Phone: 010-6205-8510
- Website: [Conversational Intelligence \(CONVEI\) Group](http://convei.weebly.com)  
<http://convei.weebly.com>

