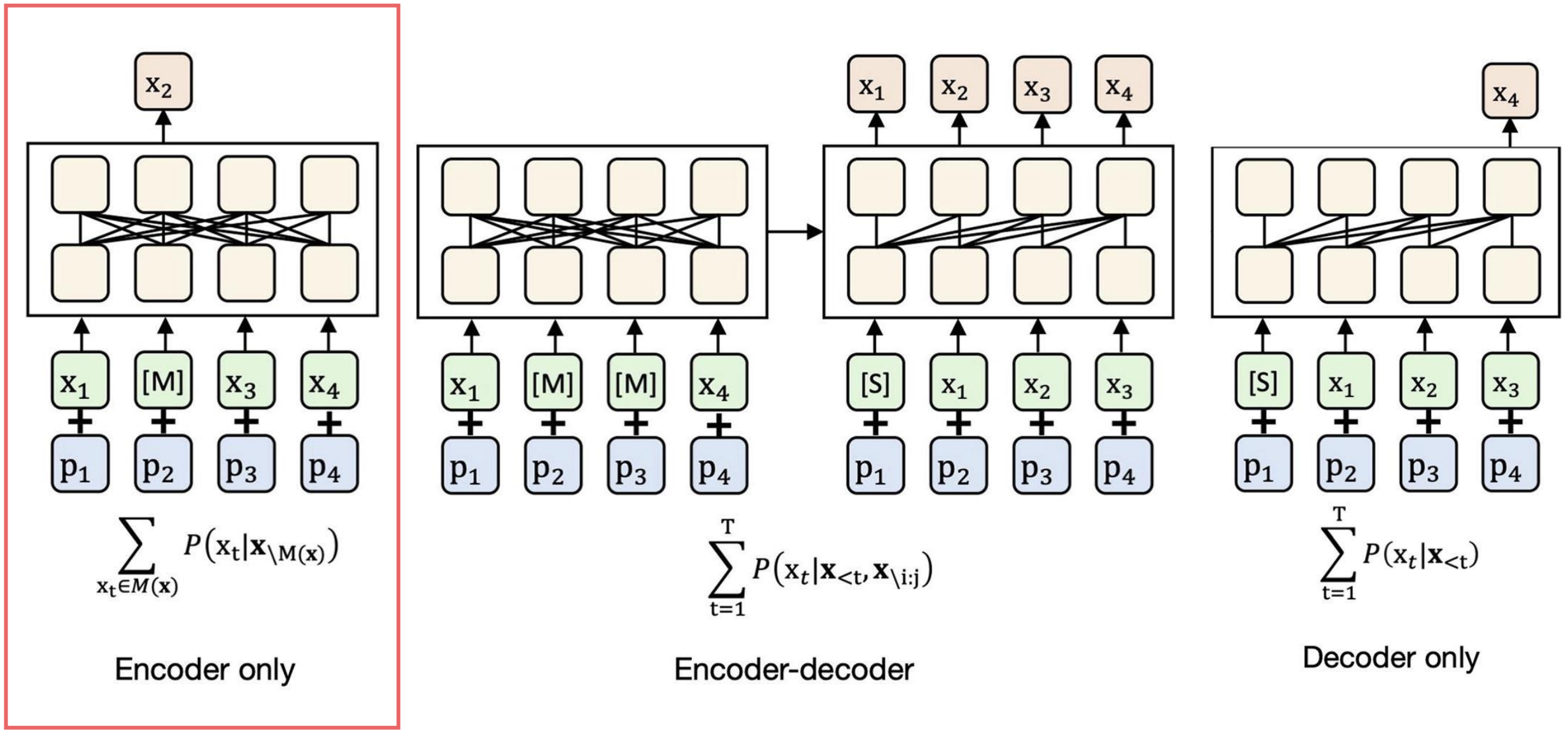
CSCE 689: Special Topics in Advanced NLP

Lecture 6: Large Language Models, Text Similarity

Kuan-Hao Huang

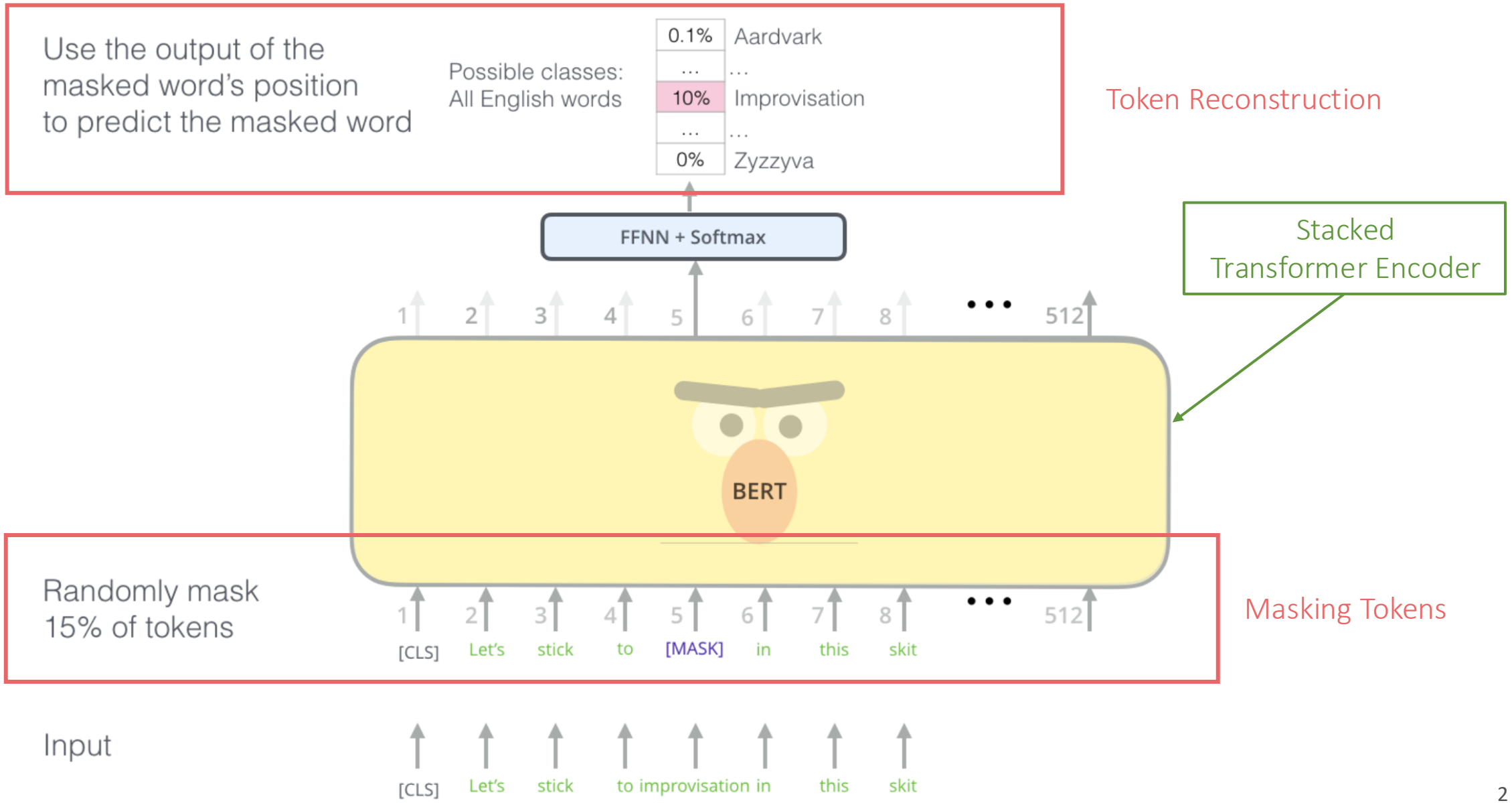
Fall 2026

Types of Pre-Training



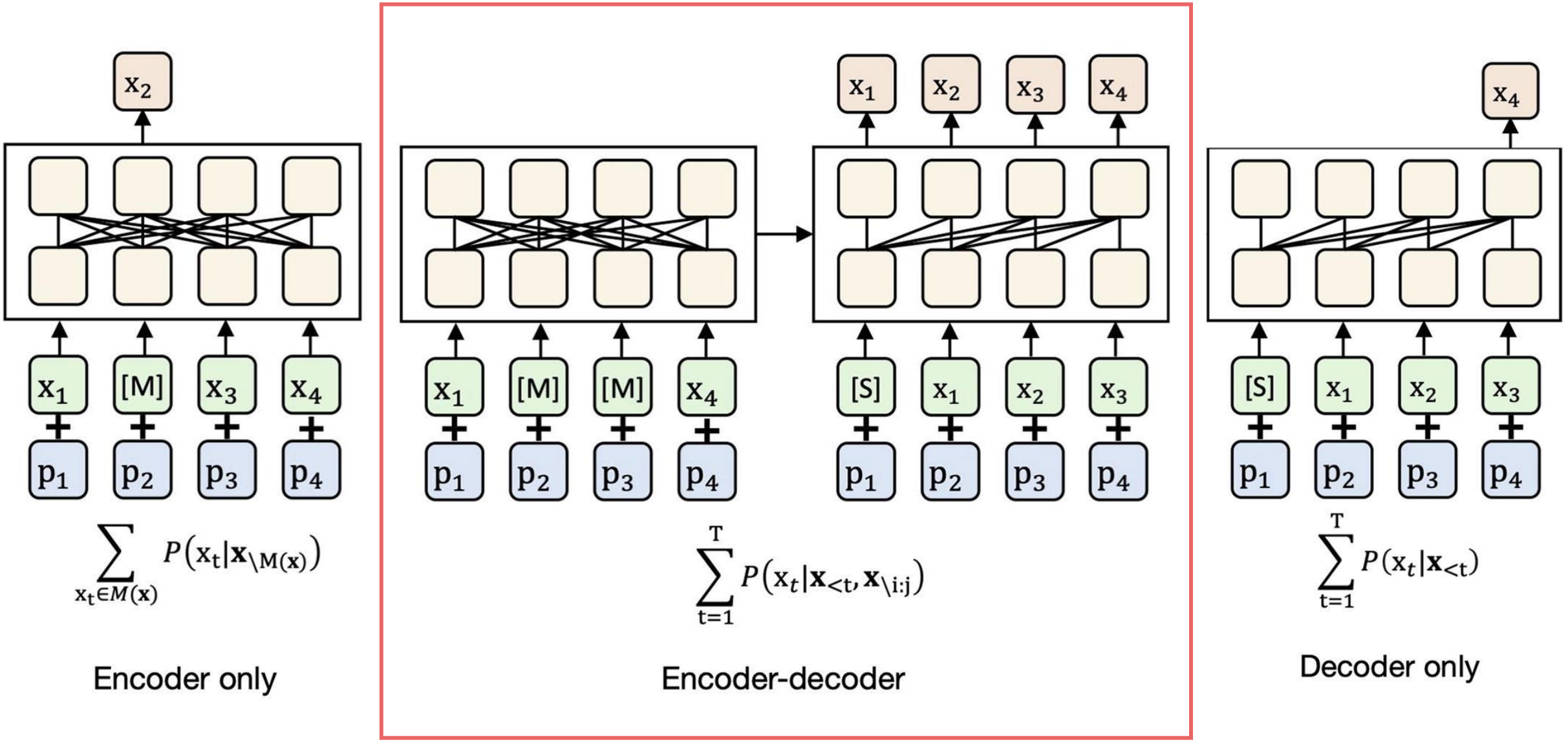
https://www.sciencedirect.com/science/article/pii/S2095809922006324

Pre-Training Task: Masked Language Modeling



http://jalammar.github.io/illustrated-bert/

Types of Pre-Training



https://www.sciencedirect.com/science/article/pii/S2095809922006324

Encoder-Decoder: BART

- **B**idirectional and **A**uto-**R**egressive **T**ransformers (BART)

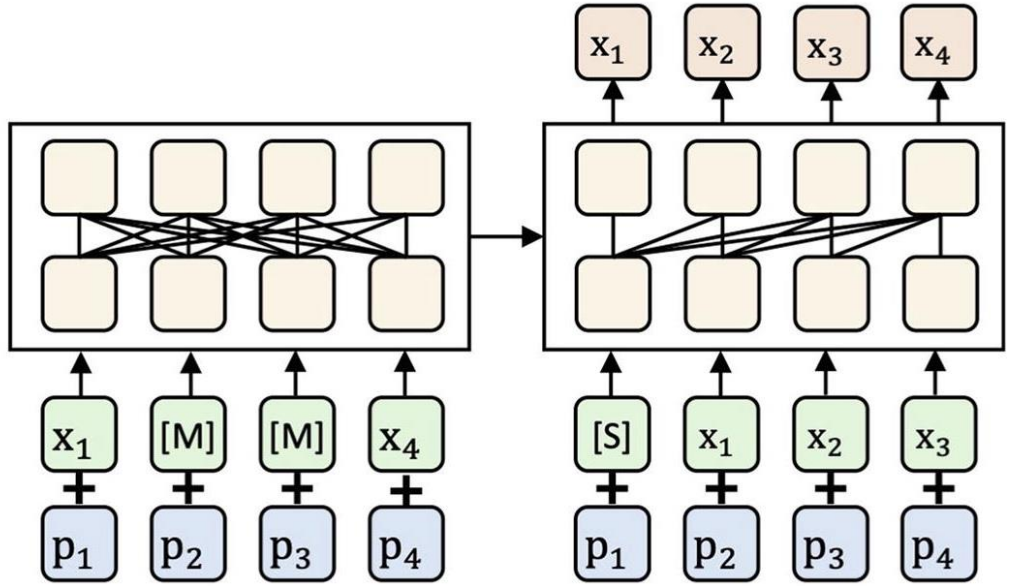
BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

Mike Lewis*, Yinhan Liu*, Naman Goyal*, Marjan Ghazvininejad,
Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer
Facebook AI
{mikelewis,yinhanliu,naman}@fb.com

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, 2020

Encoder-Decoder: BART

- Transformer Encoder-Decoder
- Pre-training for generation tasks but can be also used for representations

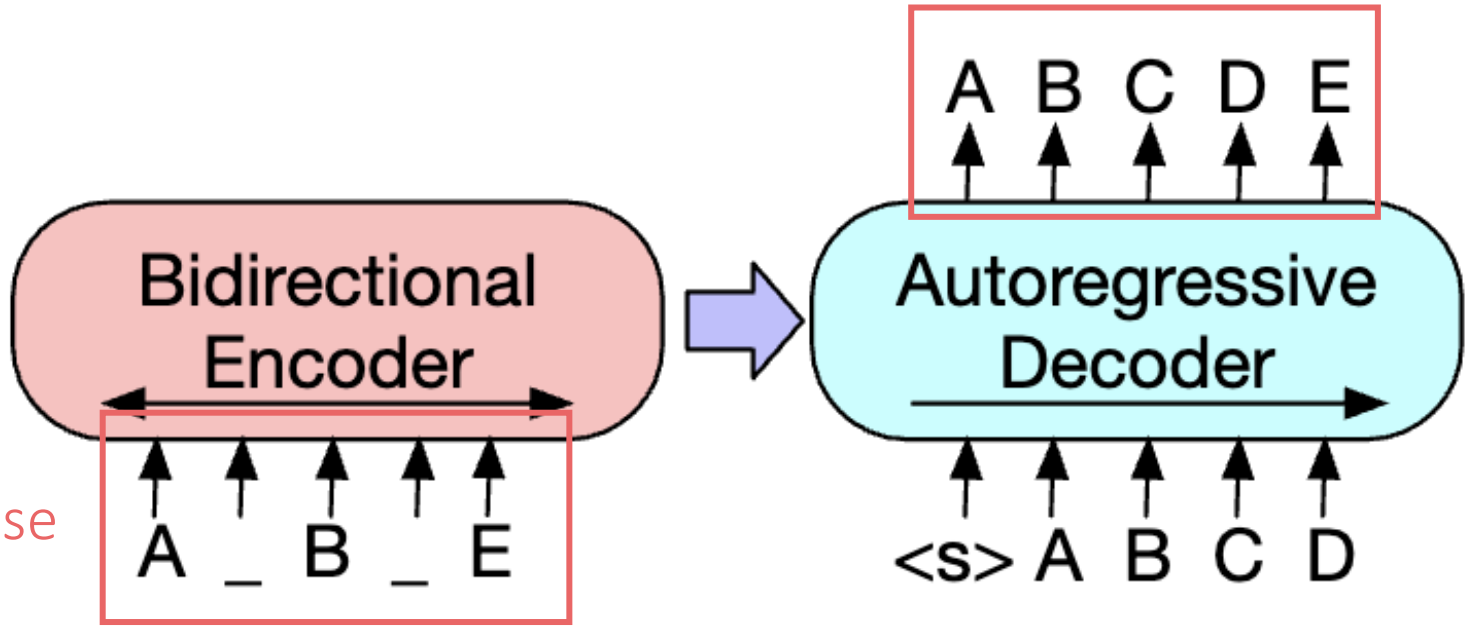


$$\sum_{t=1}^{T} P\left(\mathrm{x}_t | \mathbf{x}_{<t}, \mathbf{x}_{\backslash \mathrm{i:j}}\right)$$

Encoder-decoder

Denoising Autoencoder

Generate original input



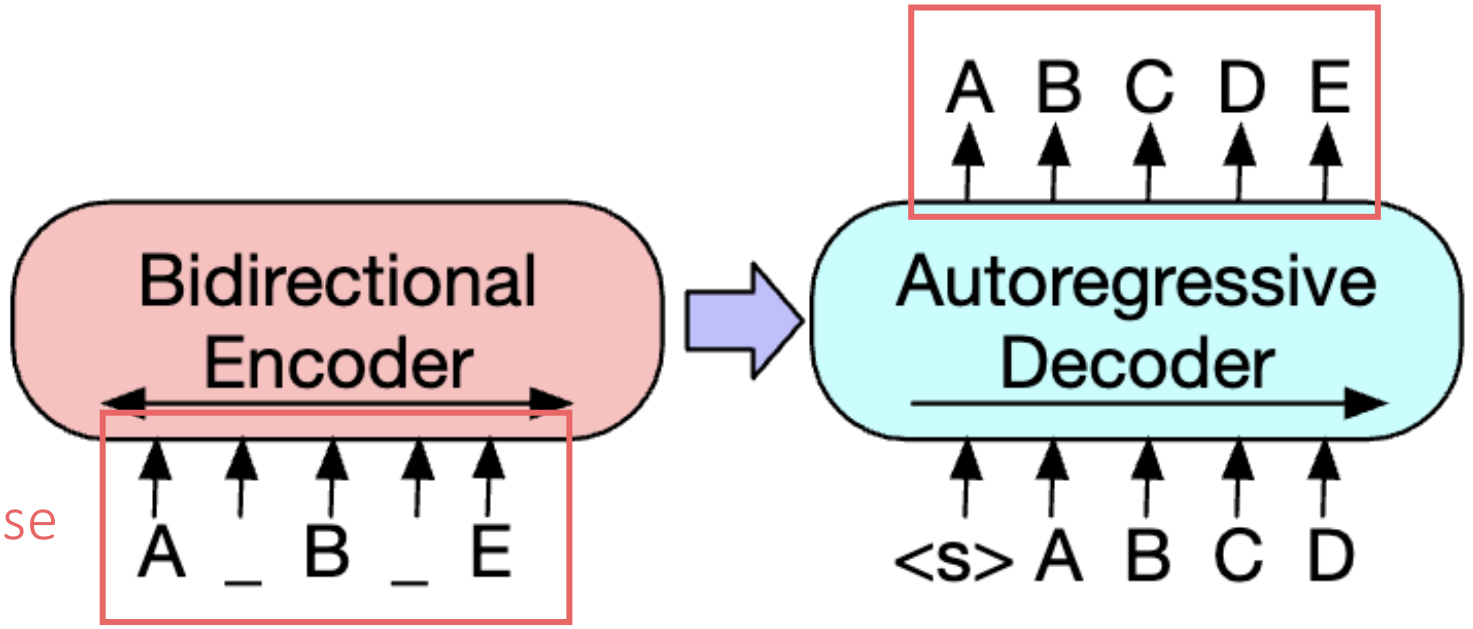
Adding noise

Denoising Objective

- Token Masking
  - A<mask>CD<mask>F. → ABCDEF.
- Token Deletion
  - ACDF. → ABCDEF.
- Text Infilling
  - A<mask>D<mask>F. → ABCDEF.
- Sentence Permutation
  - FG. ABC. DE. → ABC. DE. FG.
- Document Rotation
  - E. FG. ABC. D → ABC. DE. FG.

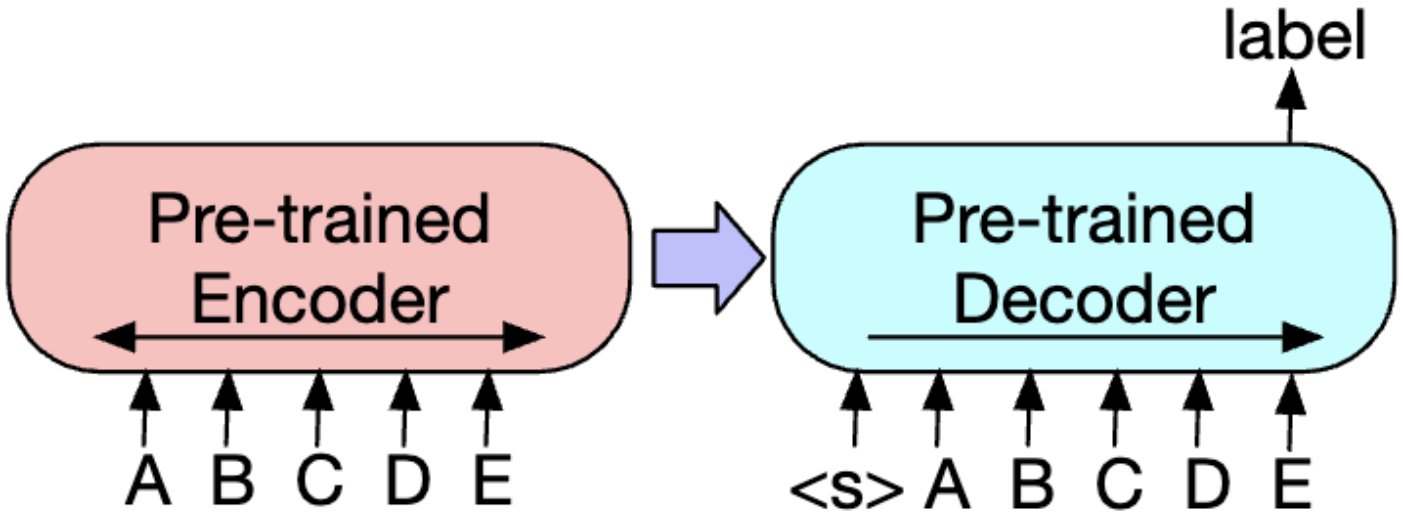
Denoising Autoencoder

Generate original input

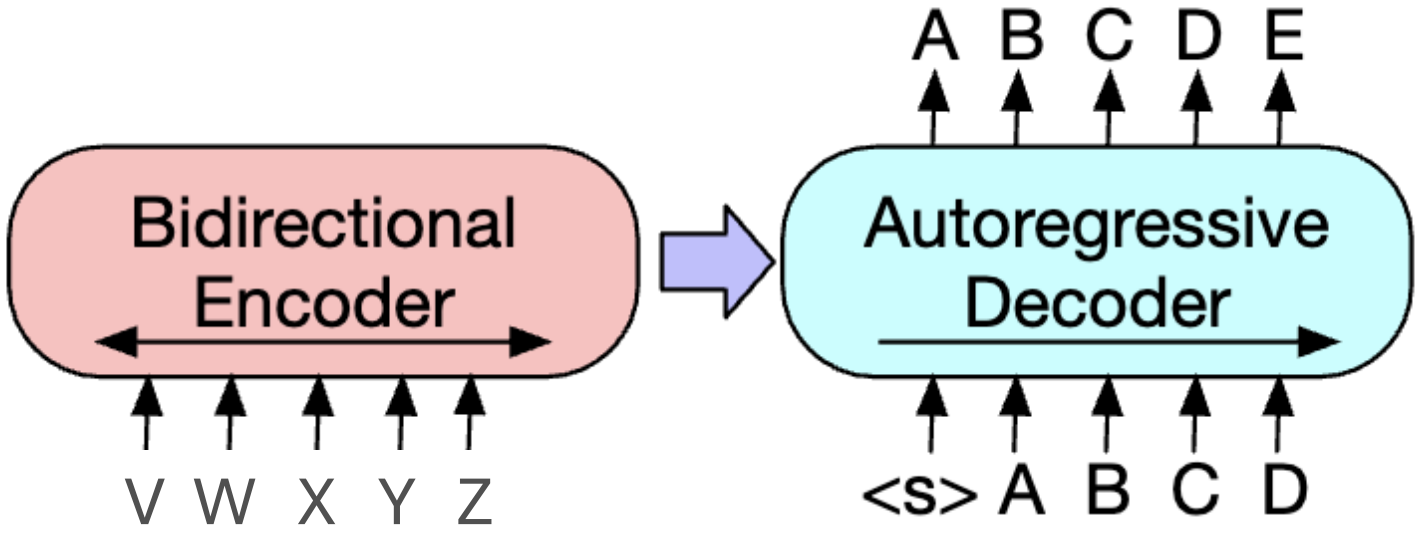


Adding noise

Fine-Tuning: Sentence-Level Tasks



Fine-Tuning: Sequence-to-Sequence



Comparable Performance on Classification Tasks

	SQuAD 1.1 EM/F1	**SQuAD 2.0** EM/F1	**MNLI** m/mm	**SST** Acc	**QQP** Acc	**QNLI** Acc	**STS-B** Acc	**RTE** Acc	**MRPC** Acc	**CoLA** Mcc
BERT	84.1/90.9	79.0/81.8	86.6/-	93.2	91.3	92.3	90.0	70.4	88.0	60.6
RoBERTa	88.9/**94.6**	**86.5/89.4**	**90.2/90.2**	96.4	92.2	94.7	**92.4**	86.6	**90.9**	**68.0**
BART	88.8/**94.6**	86.1/89.2	89.9/90.1	**96.6**	**92.5**	**94.9**	91.2	**87.0**	90.4	62.8

Better Performance on Generation Tasks

Summarization

	CNN/DailyMail			XSum		
	R1	R2	RL	R1	R2	RL
Lead-3	40.42	17.62	36.67	16.30	1.60	11.95
PTGEN (See et al., 2017)	36.44	15.66	33.42	29.70	9.21	23.24
PTGEN+COV (See et al., 2017)	39.53	17.28	36.38	28.10	8.02	21.72
UniLM	43.33	20.21	40.51	-	-	-
BERTSUMABS (Liu & Lapata, 2019)	41.72	19.39	38.76	38.76	16.33	31.15
BERTSUMEXTABS (Liu & Lapata, 2019)	42.13	19.60	39.18	38.81	16.50	31.27
BART	**44.16**	**21.28**	**40.90**	**45.14**	**22.27**	**37.25**

Question Answering

	ELI5		
	R1	R2	RL
Best Extractive	23.5	3.1	17.5
Language Model	27.8	4.7	23.1
Seq2Seq	28.3	5.1	22.8
Seq2Seq Multitask	28.9	5.4	23.1
BART	**30.6**	**6.2**	**24.3**

Translation

	RO-EN
Baseline	36.80
Fixed BART	36.29
Tuned BART	**37.96**

Use BART

Hugging Face

- BART-base
  - 6 layers for both encoder and decoder, hidden size = 768, 12 attention heads
  - # parameters ≈ 139M
- BART-large
  - 12 layers for both encoder and decoder, hidden size = 1024, 16 attention heads
  - # parameters ≈ 406M

Encoder-Decoder: T5

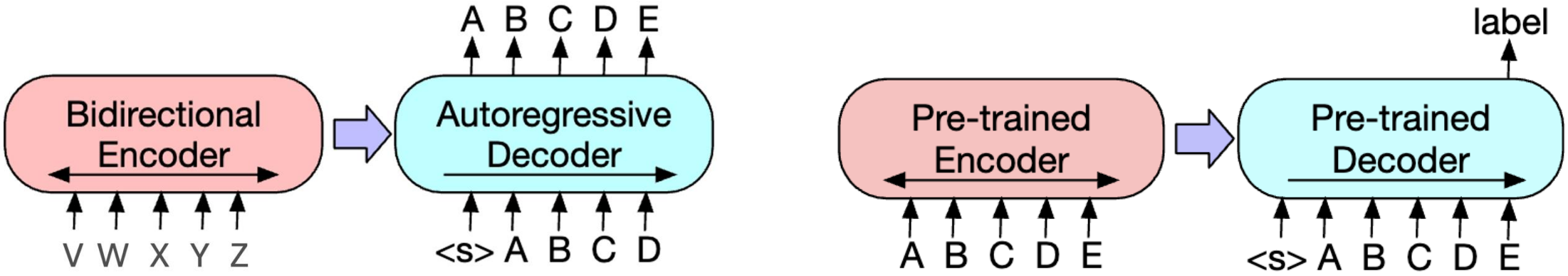
- Text-to-Text Transfer Transformer (T5)

Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

Colin Raffel*	CRAFFEL@GMAIL.COM
Noam Shazeer*	NOAM@GOOGLE.COM
Adam Roberts*	ADAROB@GOOGLE.COM
Katherine Lee*	KATHERINELEE@GOOGLE.COM
Sharan Narang	SHARANNARANG@GOOGLE.COM
Michael Matena	MMATENA@GOOGLE.COM
Yanqi Zhou	YANQIZ@GOOGLE.COM
Wei Li	MWEILI@GOOGLE.COM
Peter J. Liu	PETERJLIU@GOOGLE.COM

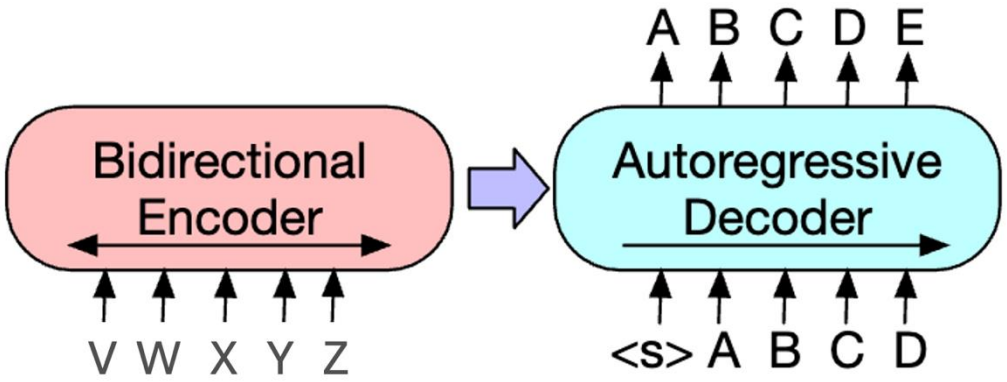
Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, 2020

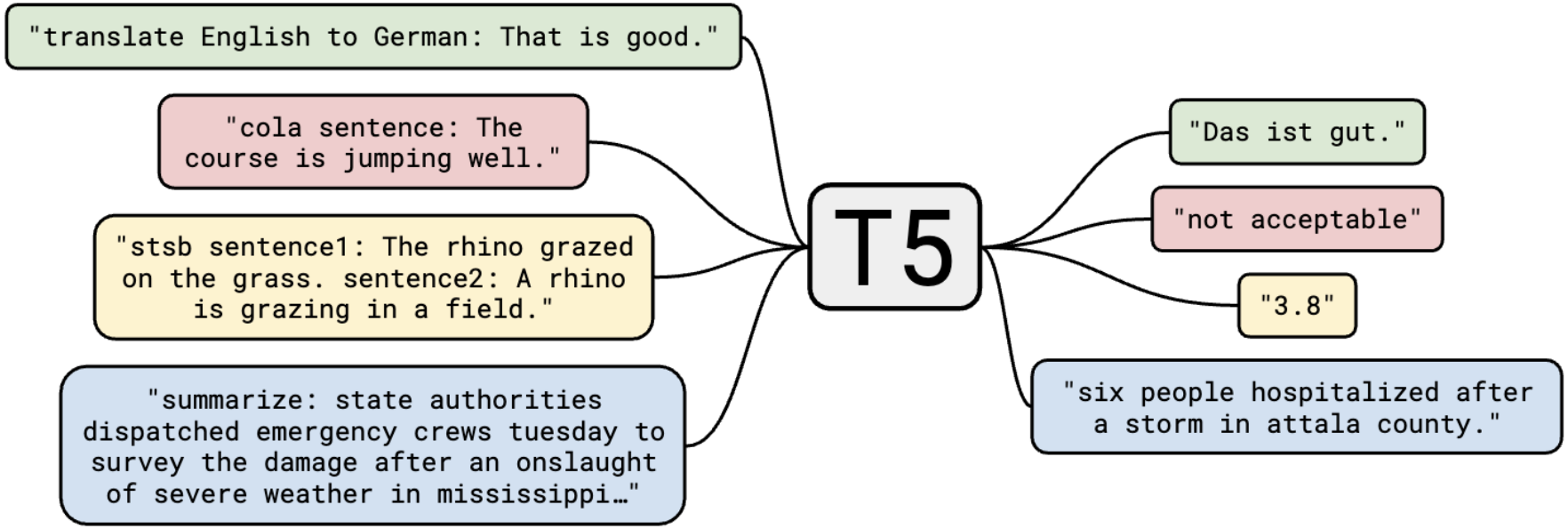
Motivation: BART



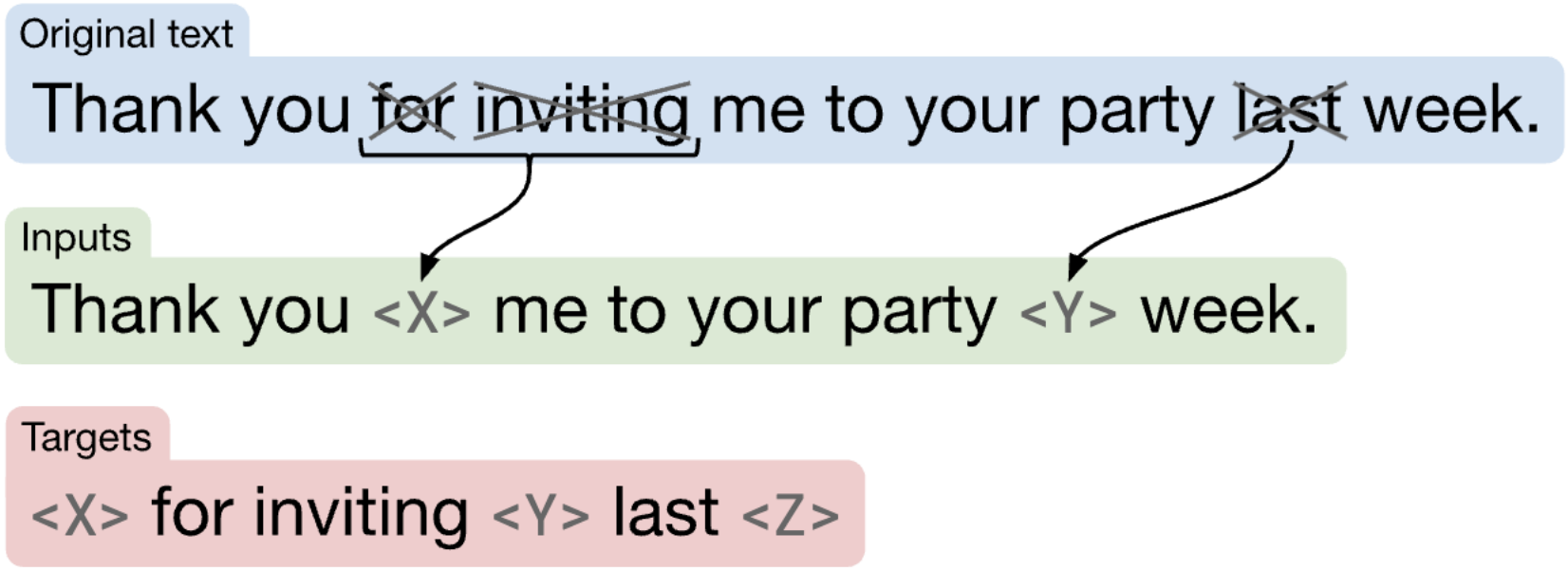
Different ways when considering classification and seq2seq generation

Convert Everything to Text-to-Text Tasks





Masked Span Reconstruction (Seq2Seq Version)



Multi-Task Learning

- Convert everything to text-to-text tasks
- Jointly fine-tune them together

Multi-Task Learning

D.7 SST2

Original input:

Sentence: it confirms fincher 's status as a film maker who artfully bends technical know-how to the service of psychological insight .

Processed input: sst2 sentence: it confirms fincher 's status as a film maker who artfully bends technical know-how to the service of psychological insight .

Original target: 1

Processed target: positive

Multi-Task Learning

D.4 MRPC

Original input:

Sentence 1: We acted because we saw the existing evidence in a new light , through the prism of our experience on 11 September , " Rumsfeld said .

Sentence 2: Rather , the US acted because the administration saw " existing evidence in a new light , through the prism of our experience on September 11 " .

Processed input: mrpc sentence1: We acted because we saw the existing evidence in a new light , through the prism of our experience on 11 September , " Rumsfeld said . sentence2: Rather , the US acted because the administration saw " existing evidence in a new light , through the prism of our experience on September 11 " .

Original target: 1

Processed target: equivalent

Multi-Task Learning

D.16 WMT English to German

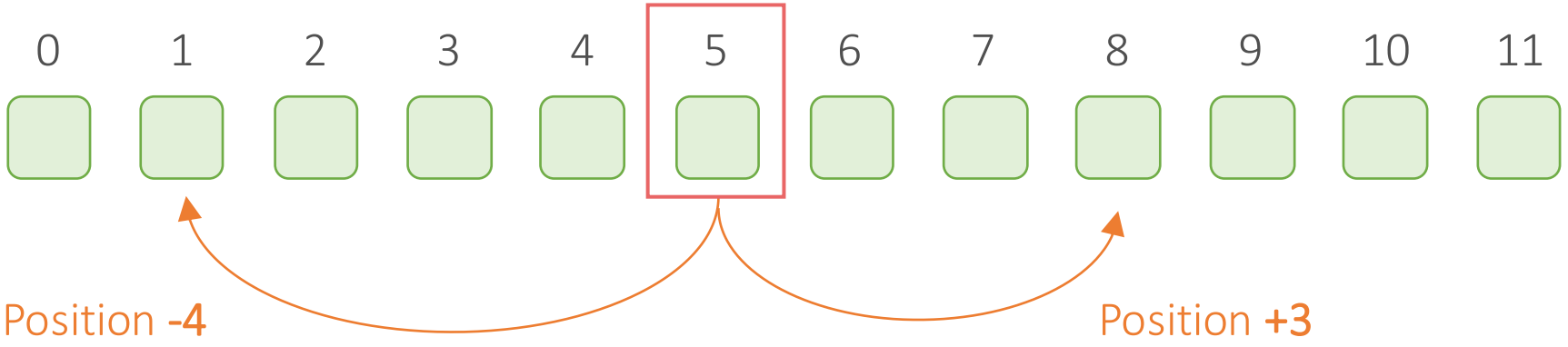
Original input: "Luigi often said to me that he never wanted the brothers to end up in court," she wrote.

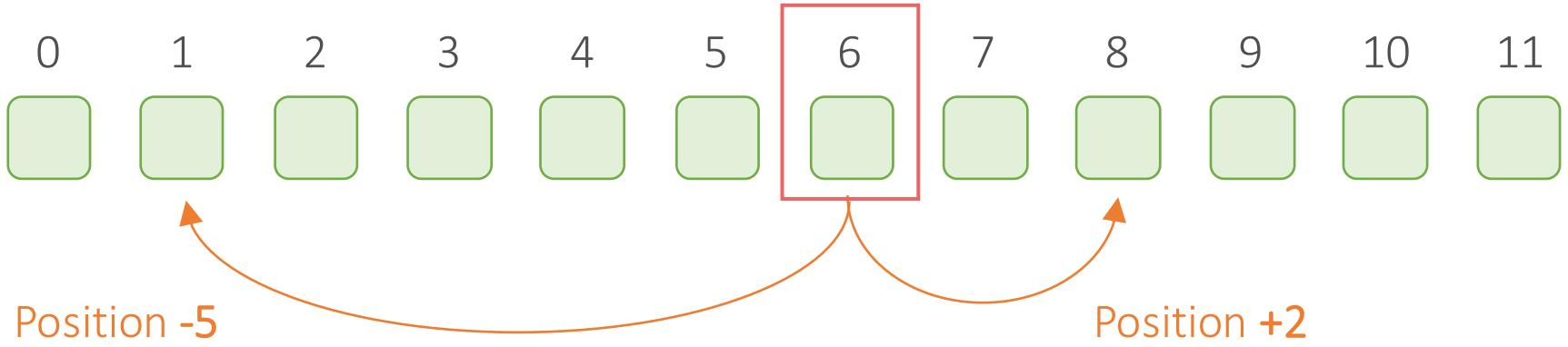
Processed input: translate English to German: "Luigi often said to me that he never wanted the brothers to end up in court," she wrote.

Original target: "Luigi sagte oft zu mir, dass er nie wollte, dass die Brüder vor Gericht landen", schrieb sie.

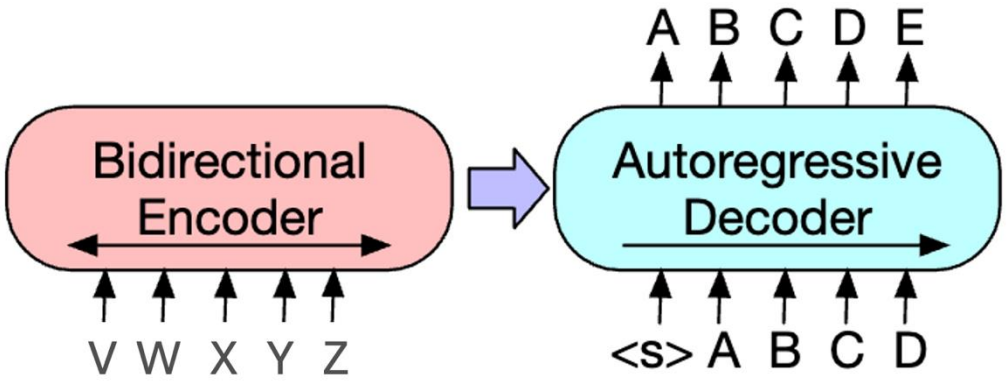
Processed target: "Luigi sagte oft zu mir, dass er nie wollte, dass die Brüder vor Gericht landen", schrieb sie.

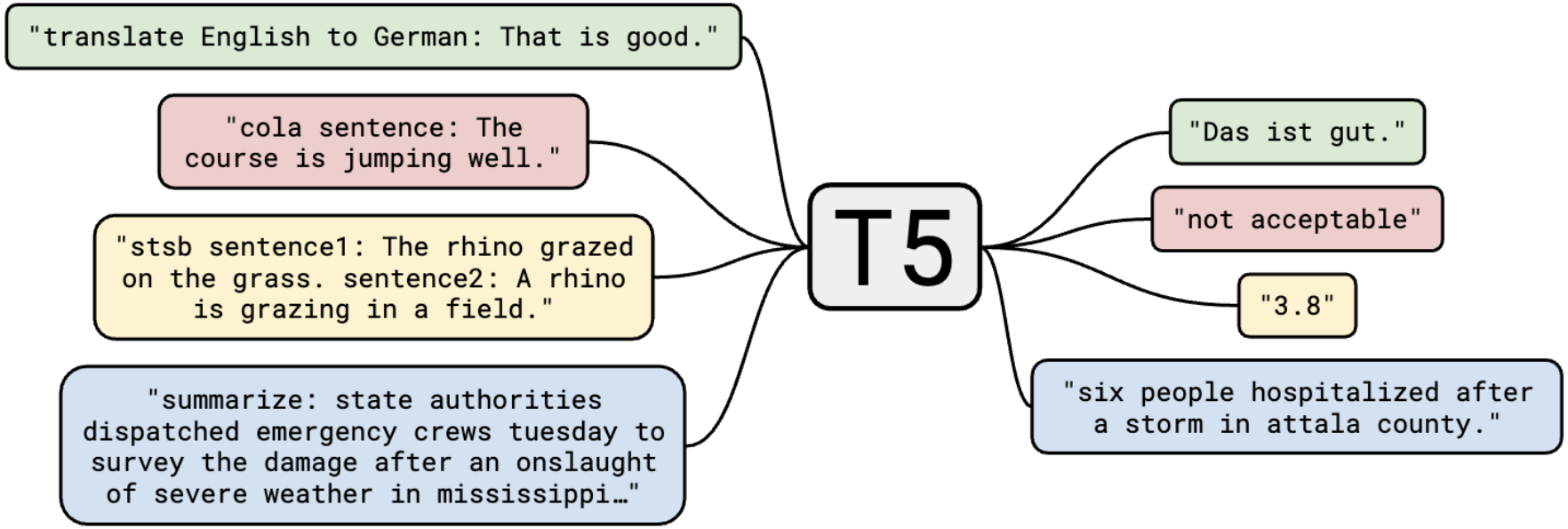
Relative Position





Fine-Tuning: Text-to-Text For Everything





Promising Results

Model	QQP F1	QQP Accuracy	MNLI-m Accuracy	MNLI-mm Accuracy	QNLI Accuracy	RTE Accuracy	WNLI Accuracy
Previous best	74.8[c]	**90.7**[b]	91.3[a]	91.0[a]	**99.2**[a]	89.2[a]	91.8[a]
T5-Small	70.0	88.0	82.4	82.3	90.3	69.9	69.2
T5-Base	72.6	89.4	87.1	86.2	93.7	80.1	78.8
T5-Large	73.9	89.9	89.9	89.6	94.8	87.2	85.6
T5-3B	74.4	89.7	91.4	91.2	96.3	91.1	89.7
T5-11B	**75.1**	90.6	**92.2**	**91.9**	96.9	**92.8**	**94.5**

Model	SQuAD EM	SQuAD F1	SuperGLUE Average	BoolQ Accuracy	CB F1	CB Accuracy	COPA Accuracy
Previous best	90.1[a]	95.5[a]	84.6[d]	87.1[d]	90.5[d]	95.2[d]	90.6[d]
T5-Small	79.10	87.24	63.3	76.4	56.9	81.6	46.0
T5-Base	85.44	92.08	76.2	81.4	86.2	94.0	71.2
T5-Large	86.66	93.79	82.3	85.4	91.6	94.8	83.4
T5-3B	88.53	94.95	86.4	89.9	90.3	94.4	92.0
T5-11B	**91.26**	**96.22**	**88.9**	**91.2**	**93.9**	**96.8**	**94.8**

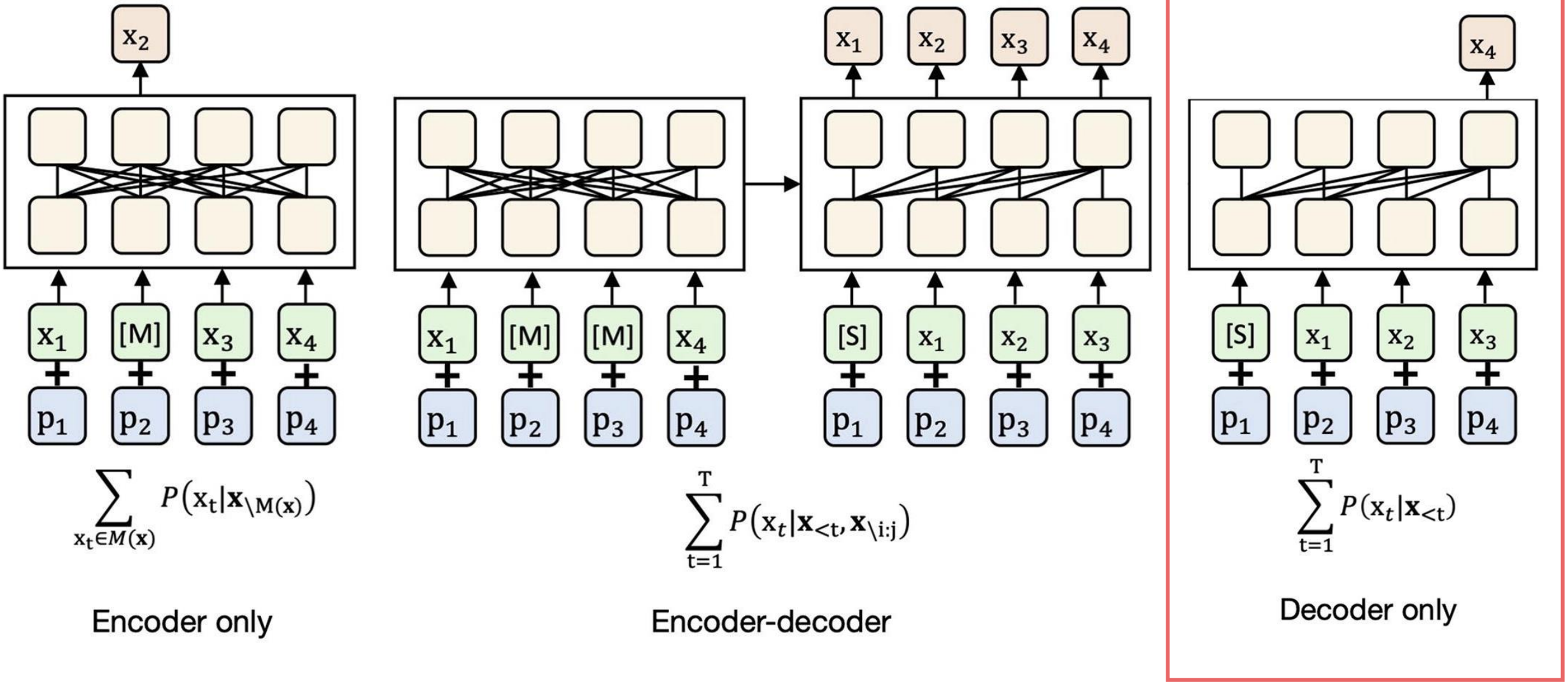
Model	MultiRC F1a	MultiRC EM	ReCoRD F1	ReCoRD Accuracy	RTE Accuracy	WiC Accuracy	WSC Accuracy
Previous best	84.4[d]	52.5[d]	90.6[d]	90.0[d]	88.2[d]	69.9[d]	89.0[d]
T5-Small	69.3	26.3	56.3	55.4	73.3	66.9	70.5
T5-Base	79.7	43.1	75.0	74.2	81.5	68.3	80.8
T5-Large	83.3	50.7	86.8	85.9	87.8	69.3	86.3
T5-3B	86.8	58.3	91.2	90.4	90.7	72.1	90.4
T5-11B	**88.1**	**63.3**	**94.1**	**93.4**	**92.5**	**76.9**	**93.8**

Use T5

Hugging Face

- T5-small:
  - # parameters ≈ 60M
- T5-base:
  - # parameters ≈ 220M
- T5-large:
  - # parameters ≈ 770M
- T5-3B: #
  - parameters ≈ 3B
- T5-11B:
  - # parameters ≈ 11B

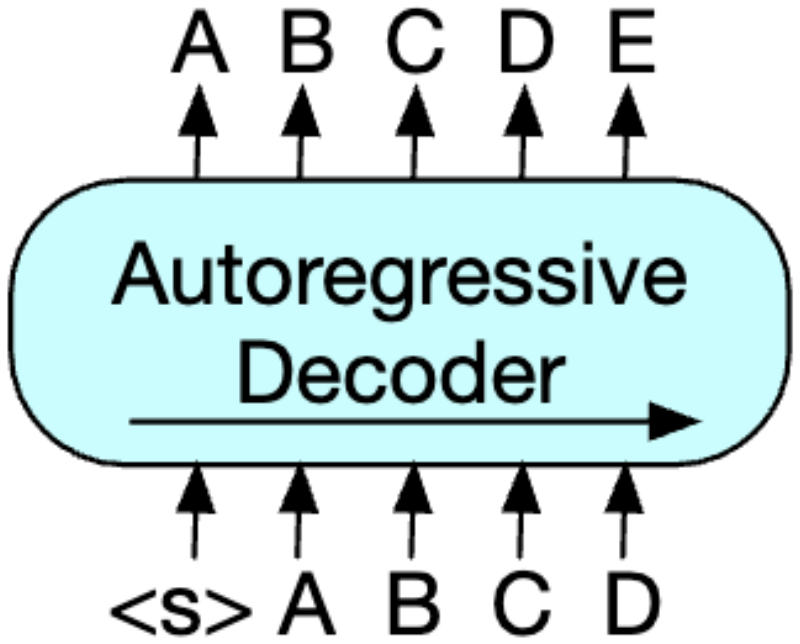
Types of Pre-Training



https://www.sciencedirect.com/science/article/pii/S2095809922006324

Language Modeling

- Next word prediction
- Trained with large corpus

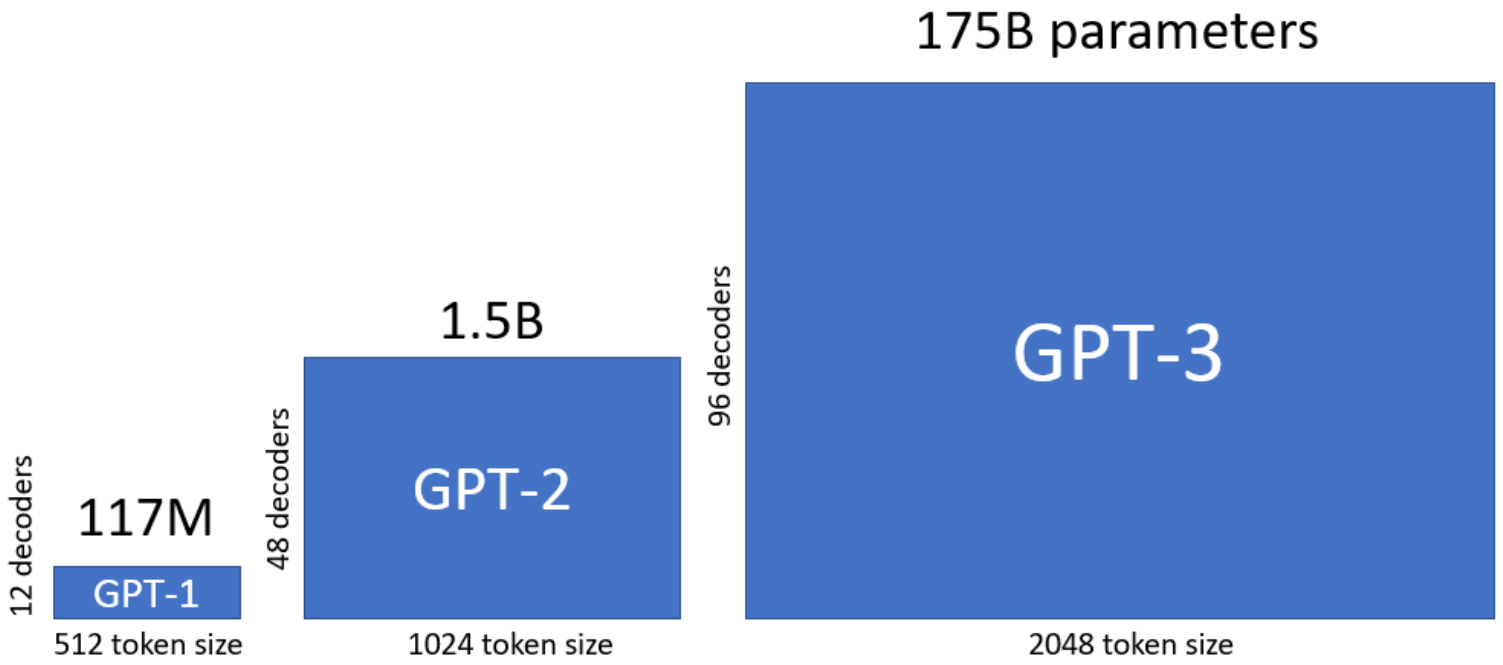


Decoder-Only: Generative Pre-trained Transformer (GPT)

- Improving Language Understanding by Generative Pre-Training, OpenAI 2018
  - **G**enerative **P**re-trained **T**ransformer (GPT)
- Language Models are Unsupervised Multitask Learners, OpenAI 2019
  - GPT-2
- Language Models are Few-Shot Learners, OpenAI 2020
  - GPT-3

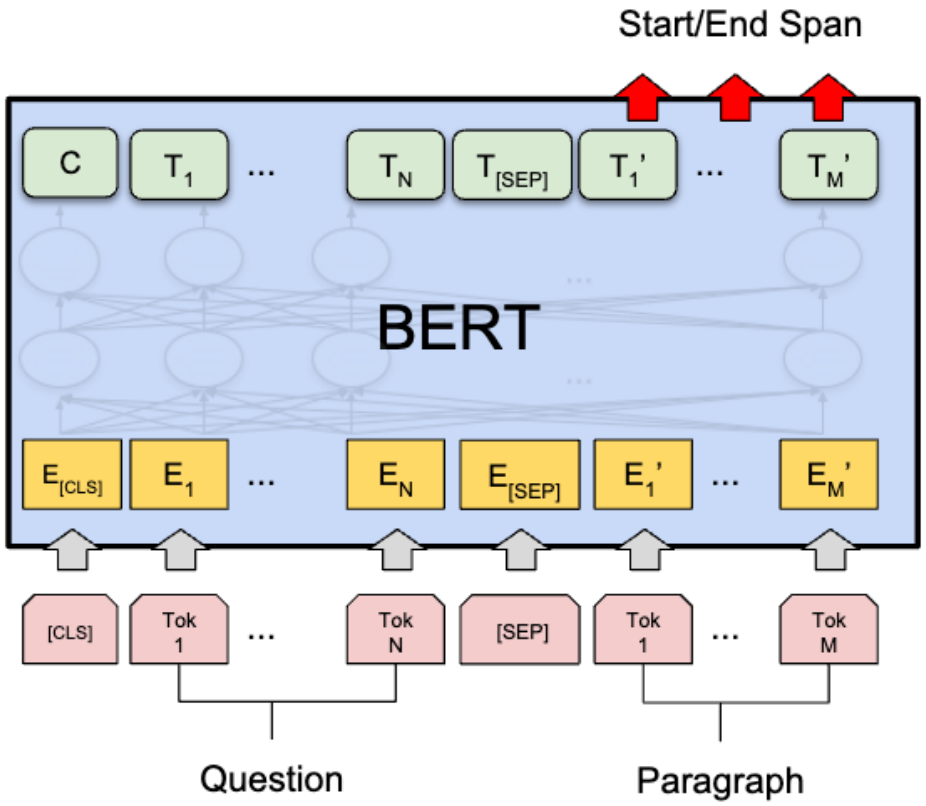
GPT-3

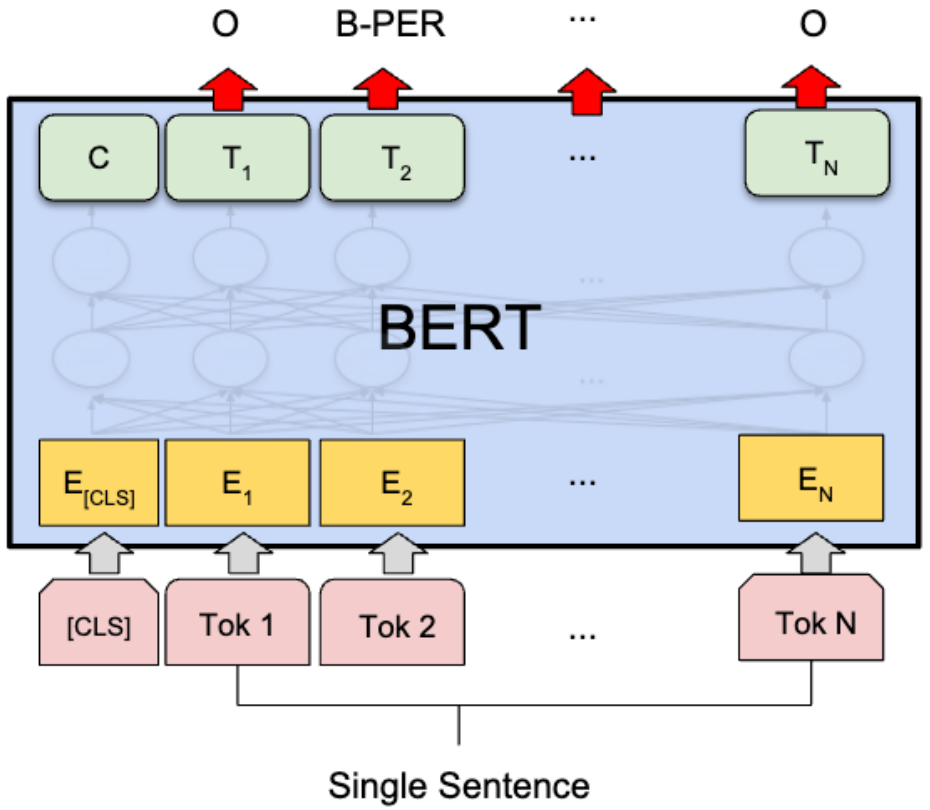
- Even larger training data, even larger model size



From Text Representations to Text Generation

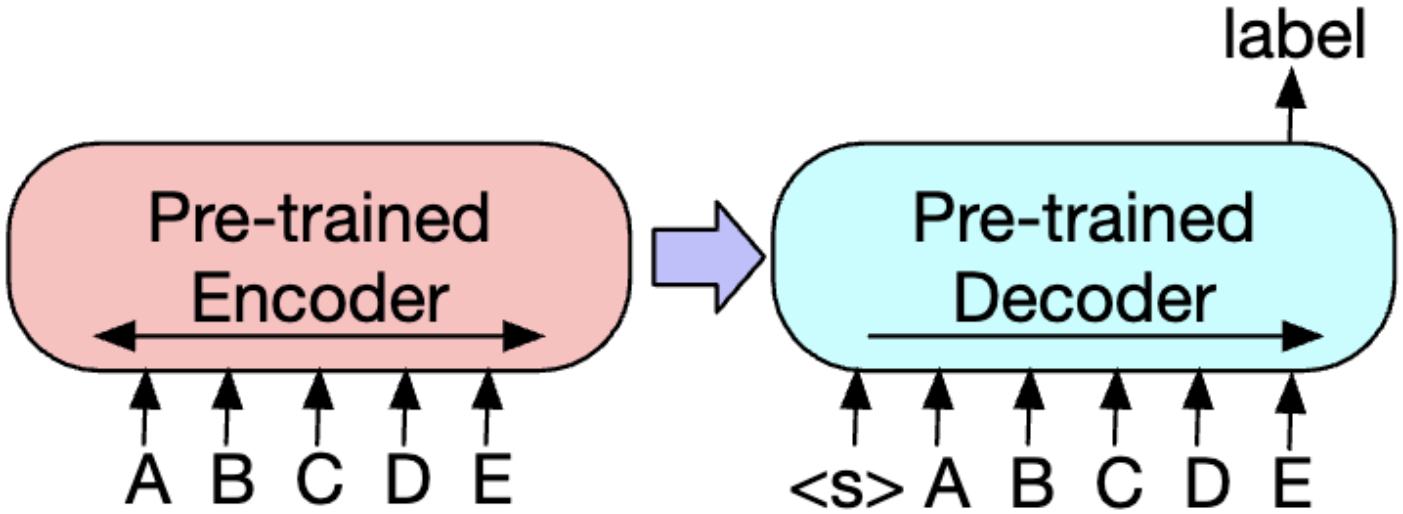
- BERT provides a good weight initialization (text representations) for fine-tuning classification tasks





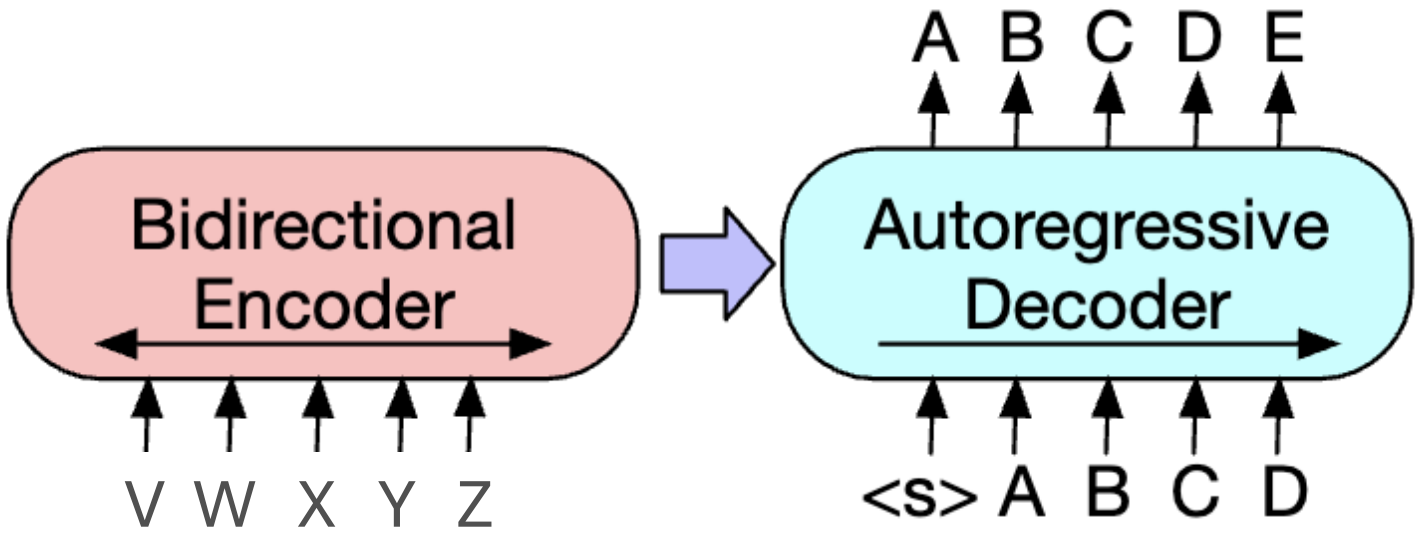
From Text Representations to Text Generation

- BART provides a good weight initialization (text representations) for fine-tuning classification tasks



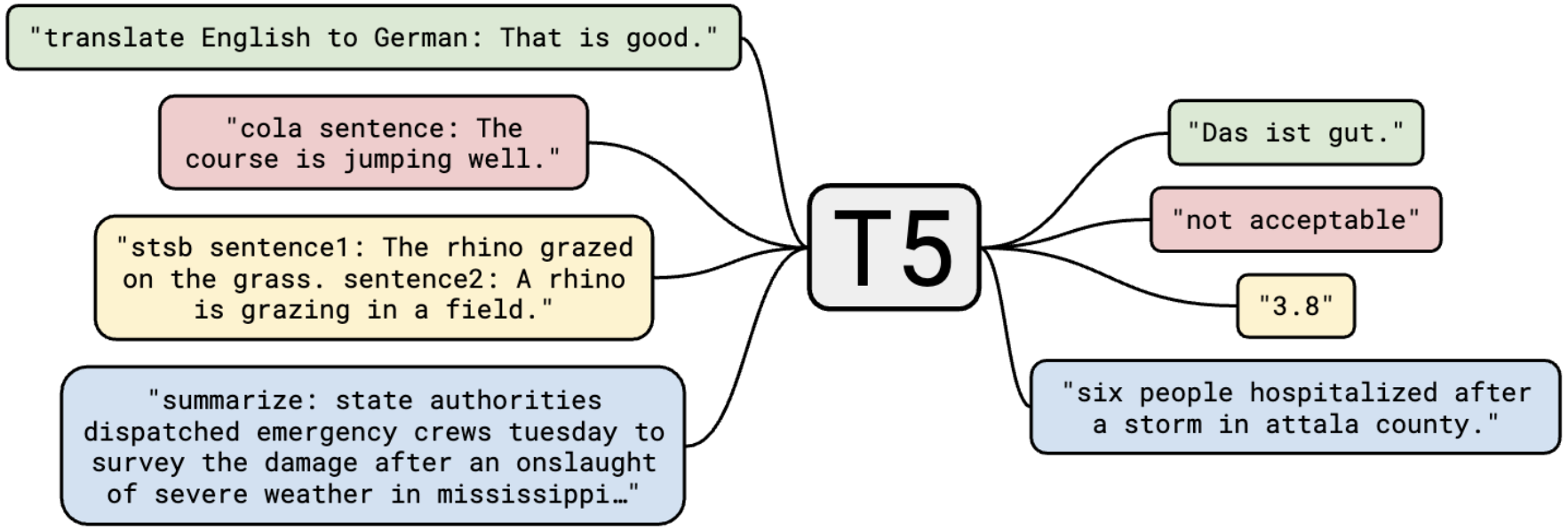
From Text Representations to Text Generation

- BART also provides a good weight initialization for fine-tuning generation tasks



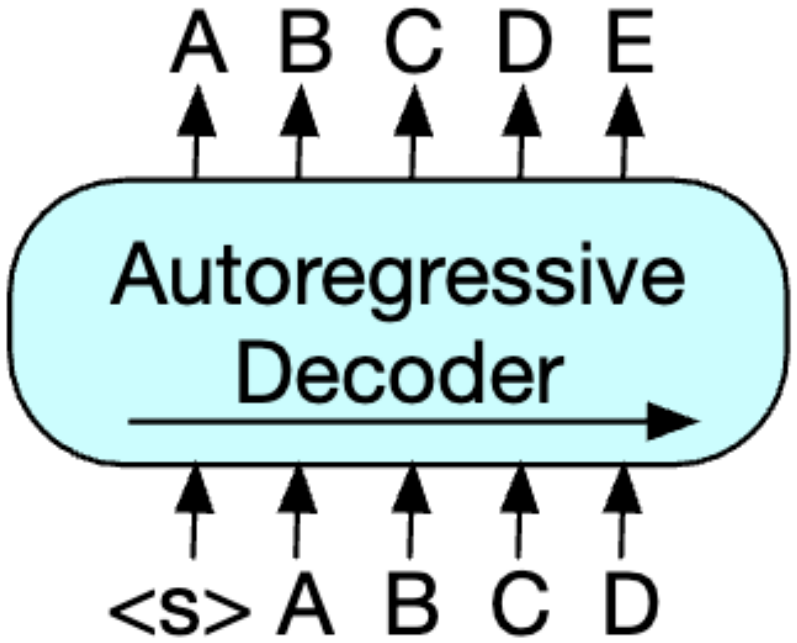
From Text Representations to Text Generation

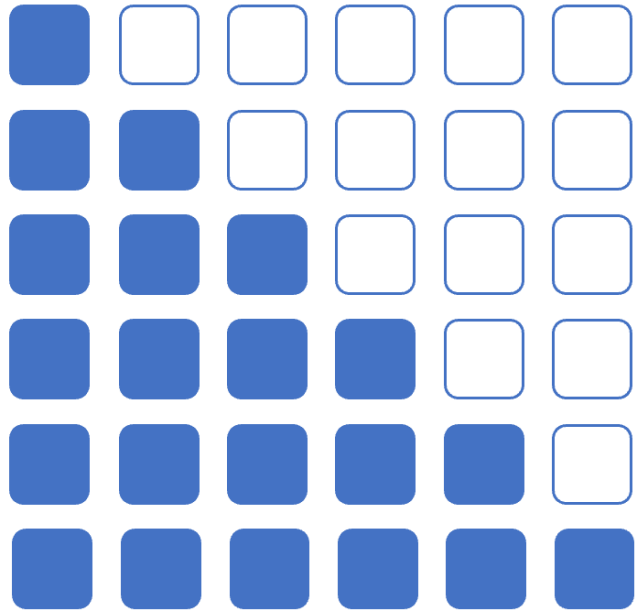
- T5 provides a good weight initialization for fine-tuning both classification and generation tasks via generation framework



From Text Representations to Text Generation

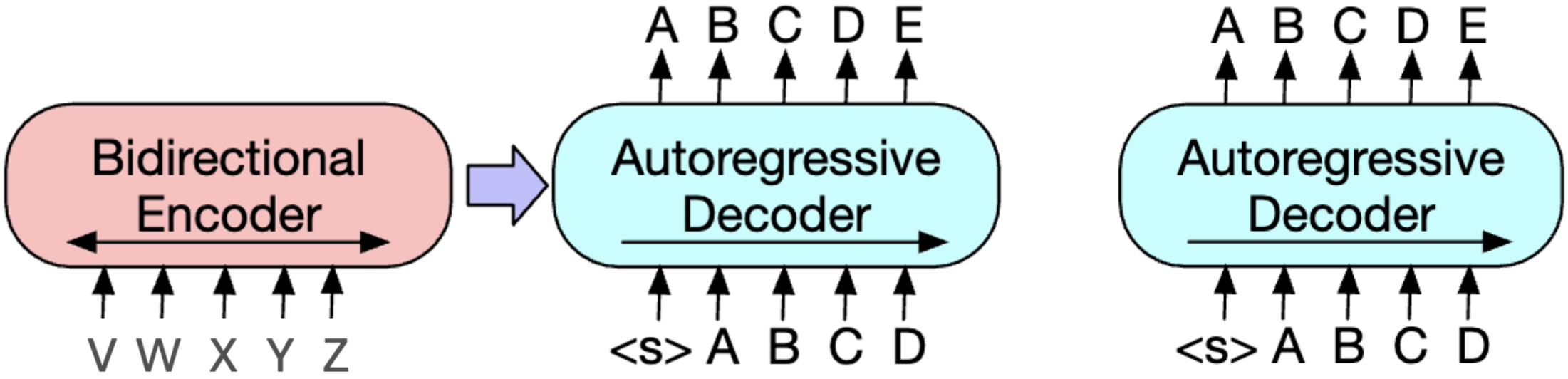
- GPT provides a good weight initialization for fine-tuning both classification and generation tasks via generation framework



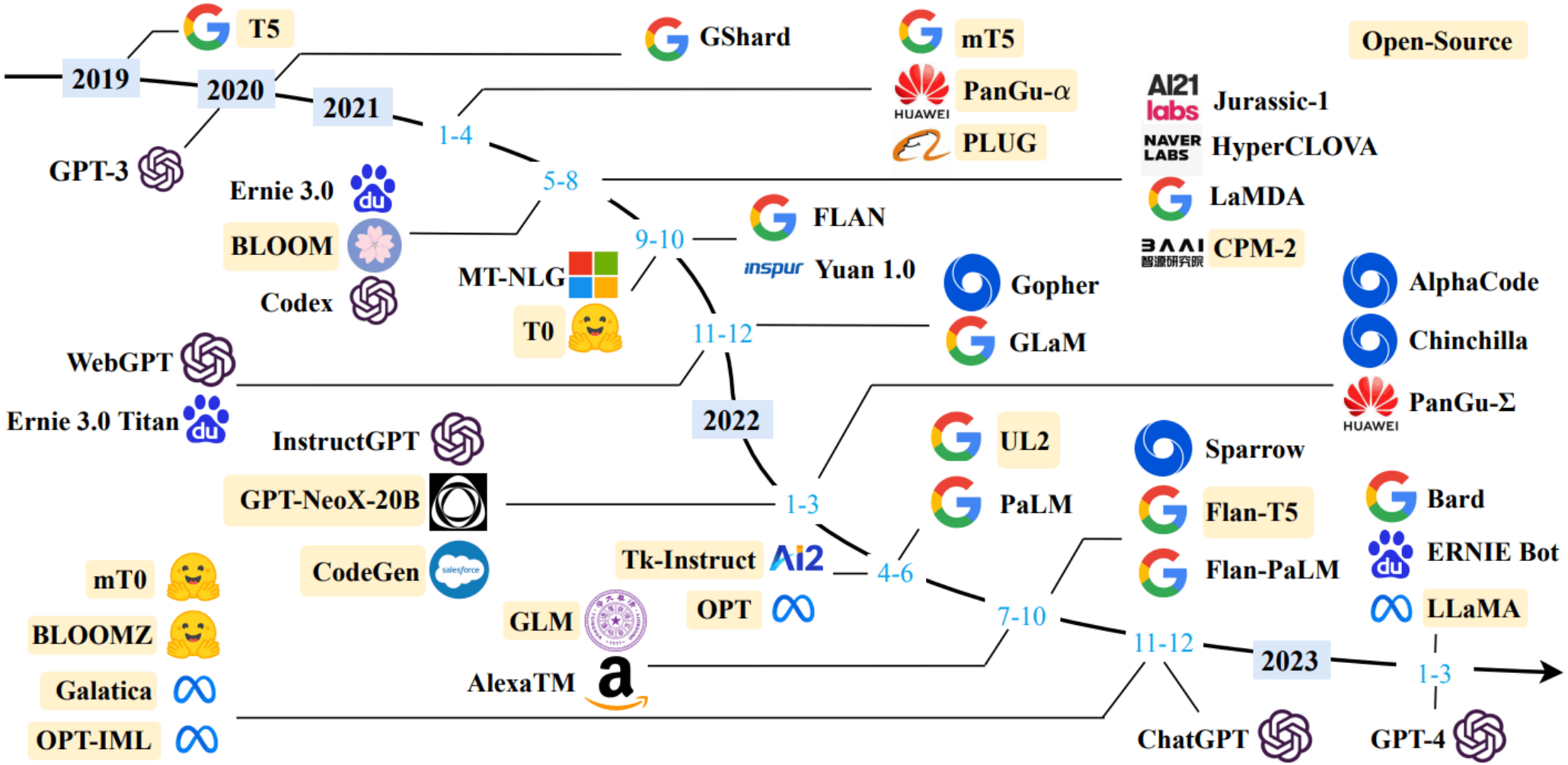


Causal Masking

Encoder-Decoder vs. Decoder Only

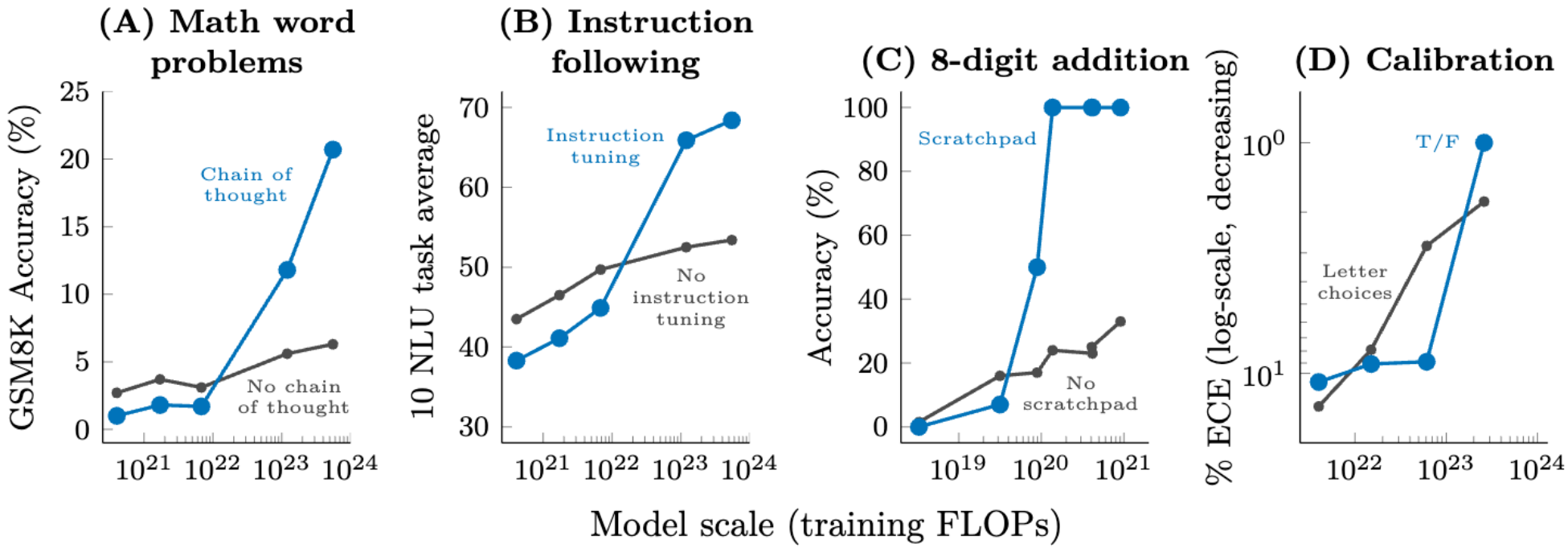


Large Language Models



A Survey of Large Language Models, 2023

Scaling Is The Key



Emergent Abilities of Large Language Models, 2022

Zero-Shot Prompting

- Prompt → Completion
  - Continue writing

Prompt

This place is incredible! The lobster is the best I've ever had. The sentiment of the above sentence is

positive.

Completion

Zero-Shot Prompting

- Prompt → Completion
  - Continue writing

Prompt

Stephen Curry's clutch barrage seals another Olympic gold for USA. The topic of the above sentence is

sport.

Completion

A New Way to Use NLP Models

- Task-specific features + task-specific model
- General embeddings + task-specific model
- General embeddings + general model + task-specific fine-tuning
- General embeddings + general model + task-specific prompting

Zero-Shot Prompting

Prompt

This place is incredible! The lobster is the best I've ever had. The sentiment of the above sentence is

positive.

Completion

Prompt

Stephen Curry's clutch barrage seals another Olympic gold for USA. The topic of the above sentence is

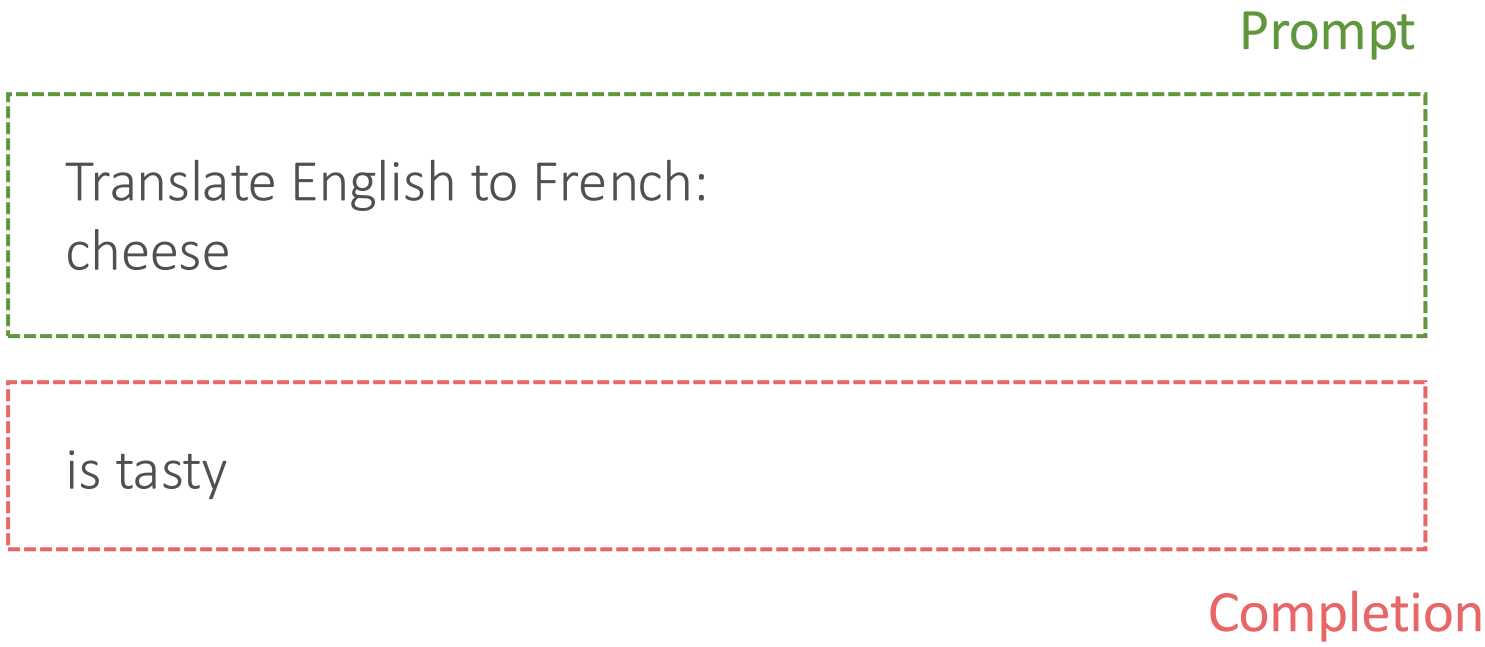
sport.

Completion

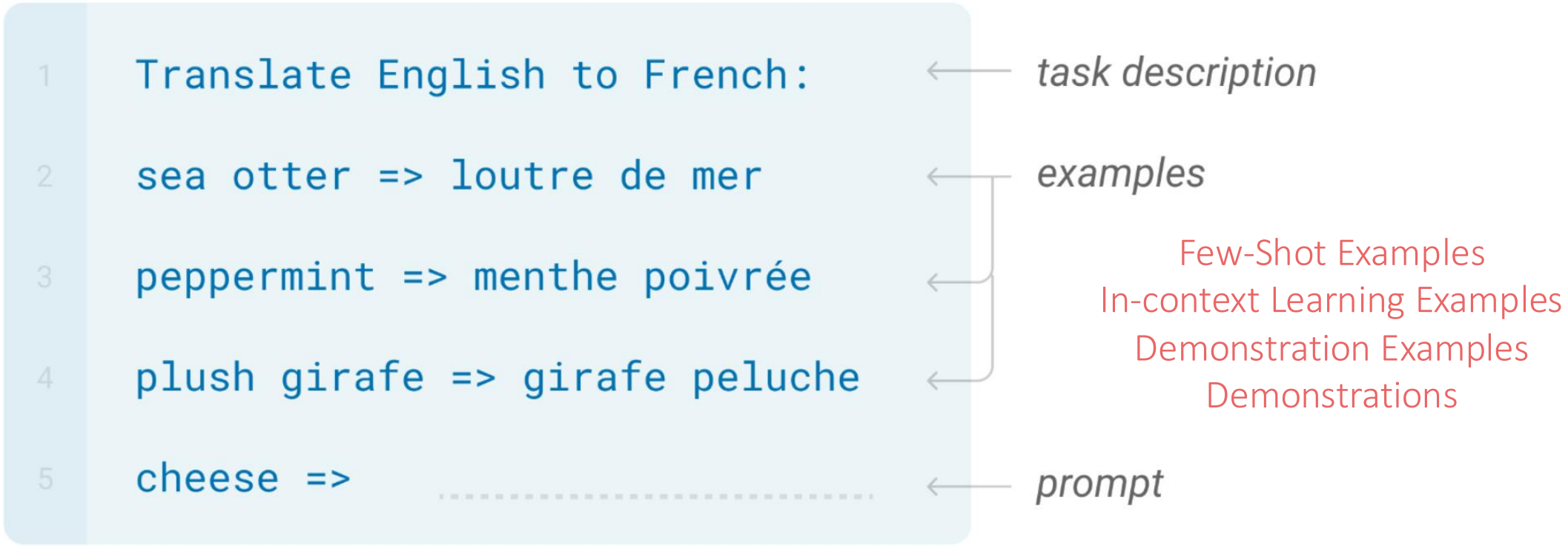
Any Issues?

Zero-Shot Prompting

- Prompt → Completion
  - Continue writing



Few-Shot Prompting / In-Context Learning



Language Models are Few-Shot Learners, 2020

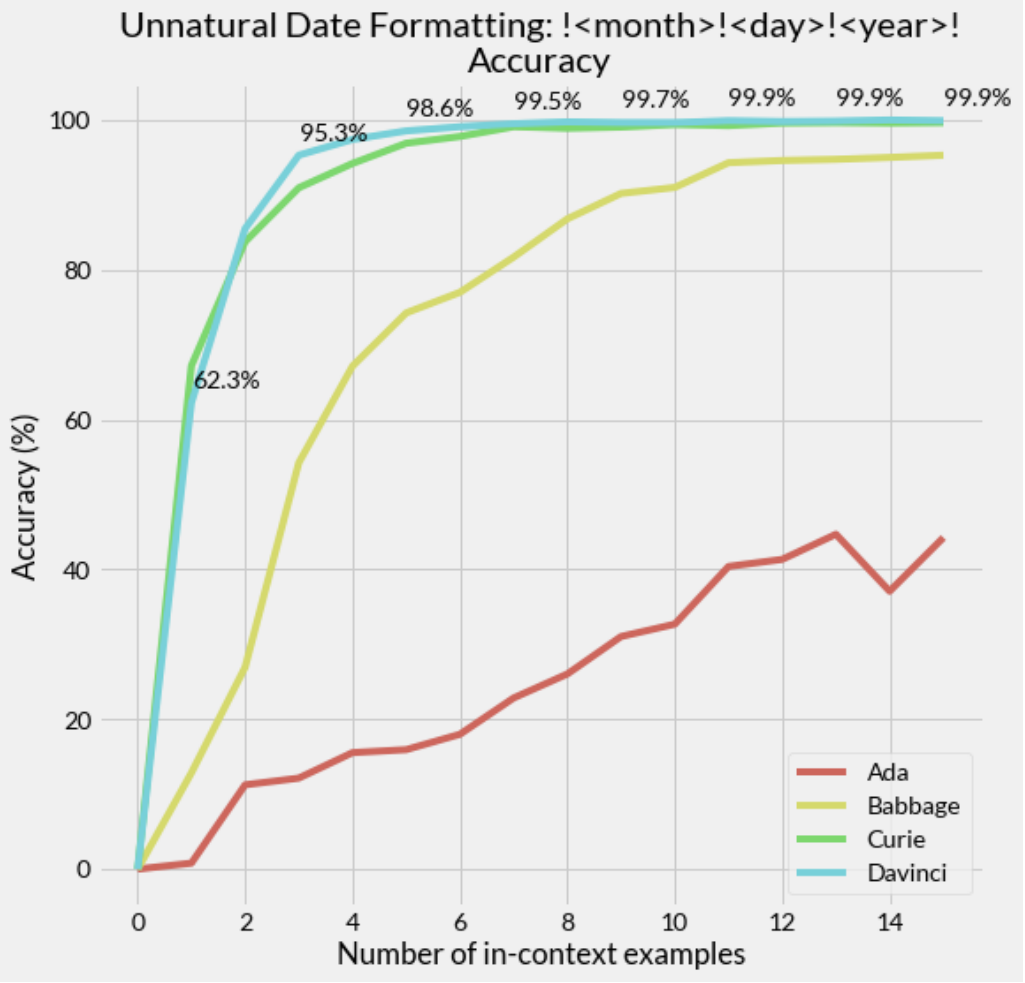
Few-Shot Prompting / In-Context Learning

```
Input: 2014-06-01
Output: !06!01!2014!
Input: 2007-12-13
Output: !12!13!2007!
Input: 2010-09-23
Output: !09!23!2010!
Input: 2005-07-23
Output: !07!23!2005!
```

in-context examples

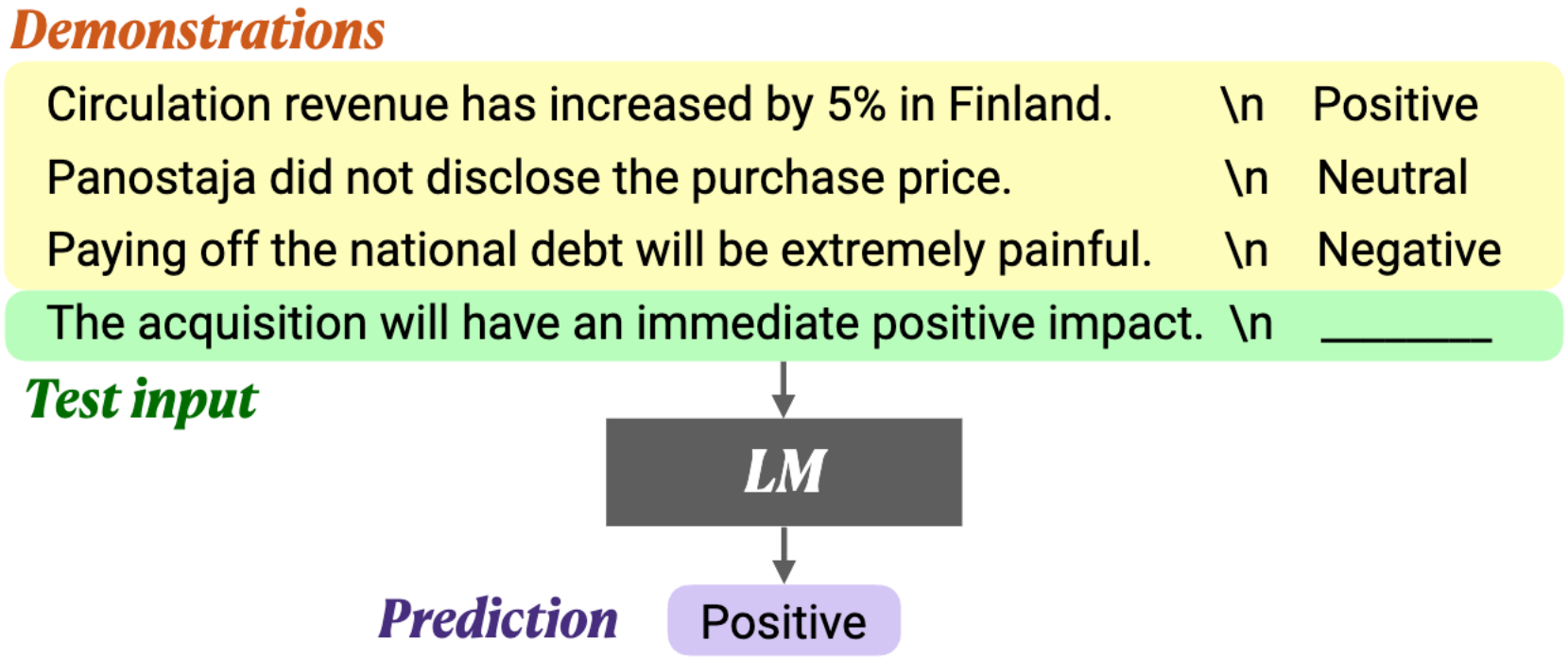
test example

model completion



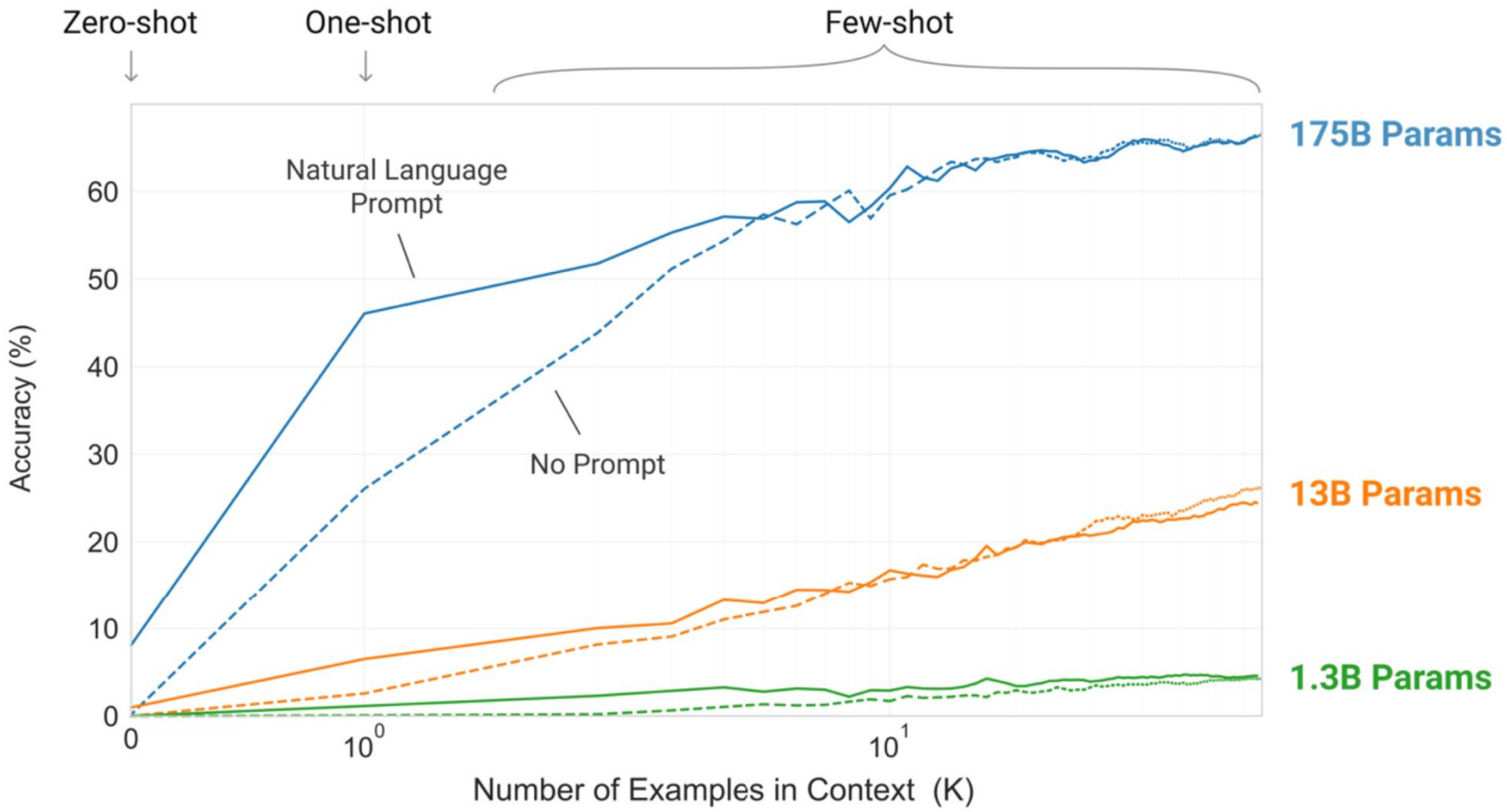
http://ai.stanford.edu/blog/in-context-learning/

Few-Shot Prompting / In-Context Learning



Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?, 2022

Few-Shot Prompting / In-Context Learning



Language Models are Few-Shot Learners, 2020

Many-Shot In-Context Learning

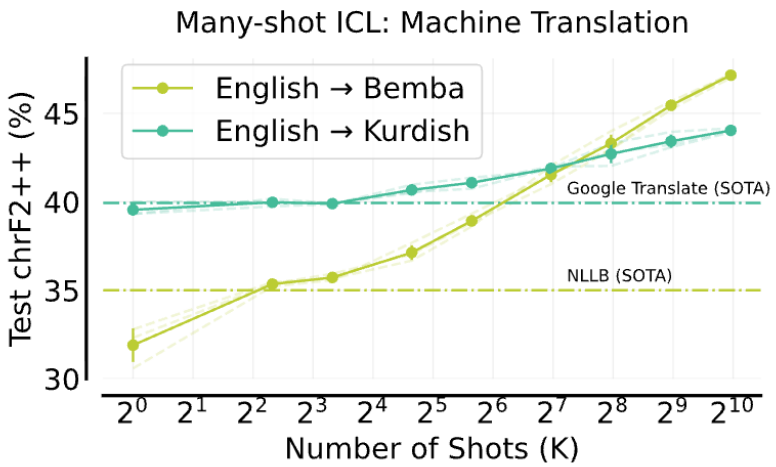
Google DeepMind

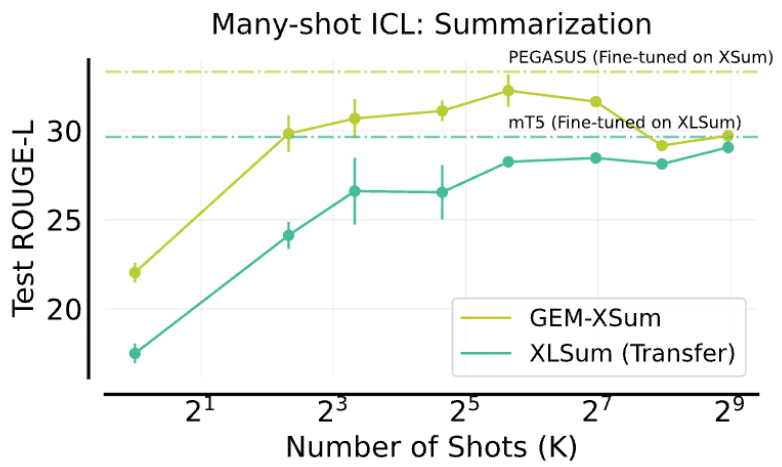
2024-10-18

Many-Shot In-Context Learning

Rishabh Agarwal*, Avi Singh*, Lei M. Zhang†, Bernd Bohnet†, Luis Rosias†, Stephanie C.Y. Chan†, Biao Zhang†, Ankesh Anand , Zaheer Abbas , Azade Nova , John D. Co-Reyes , Eric Chu , Feryal Behbahani , Aleksandra Faust and Hugo Larochelle

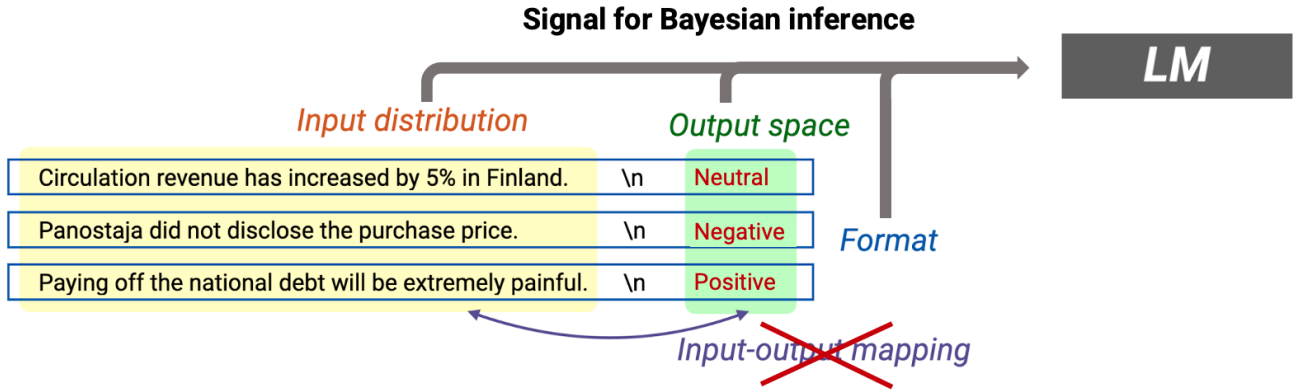
*Contributed equally, †Key contribution

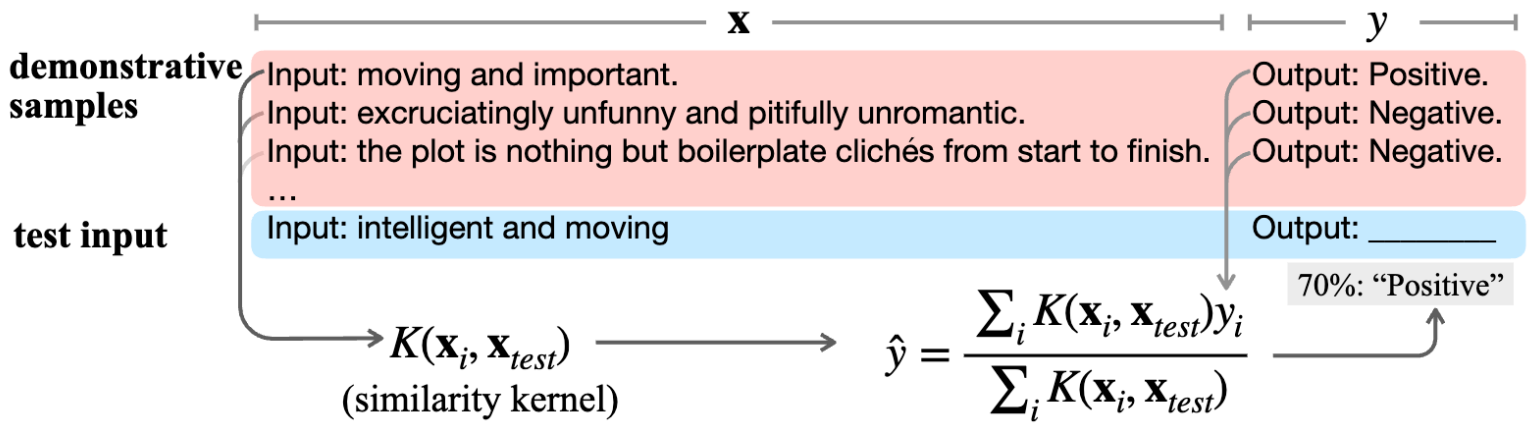




What Makes In-Context Learning Work?

- Still an open research problem





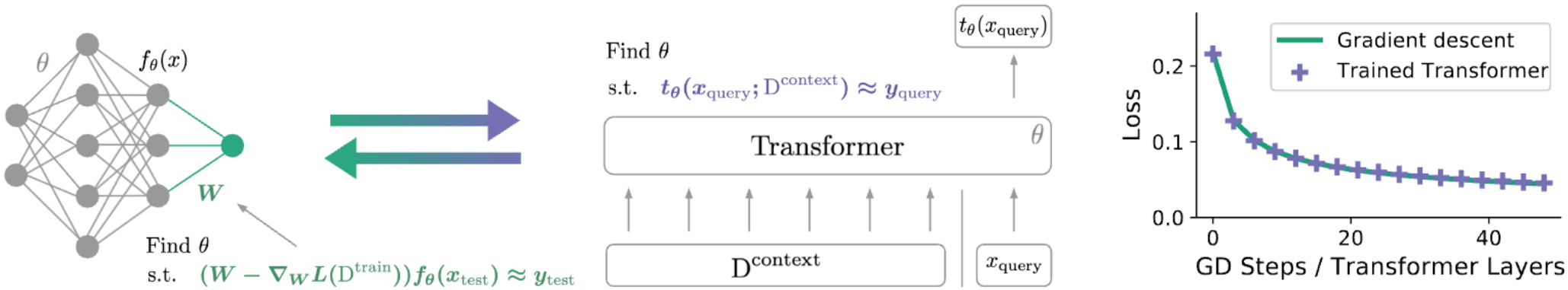
https://ai.stanford.edu/blog/understanding-incontext/ Understanding Emergent In-Context Learning from a Kernel Regression Perspective,2025

What Makes In-Context Learning Work?

- Still an open research problem

Transformers Learn In-Context by Gradient Descent

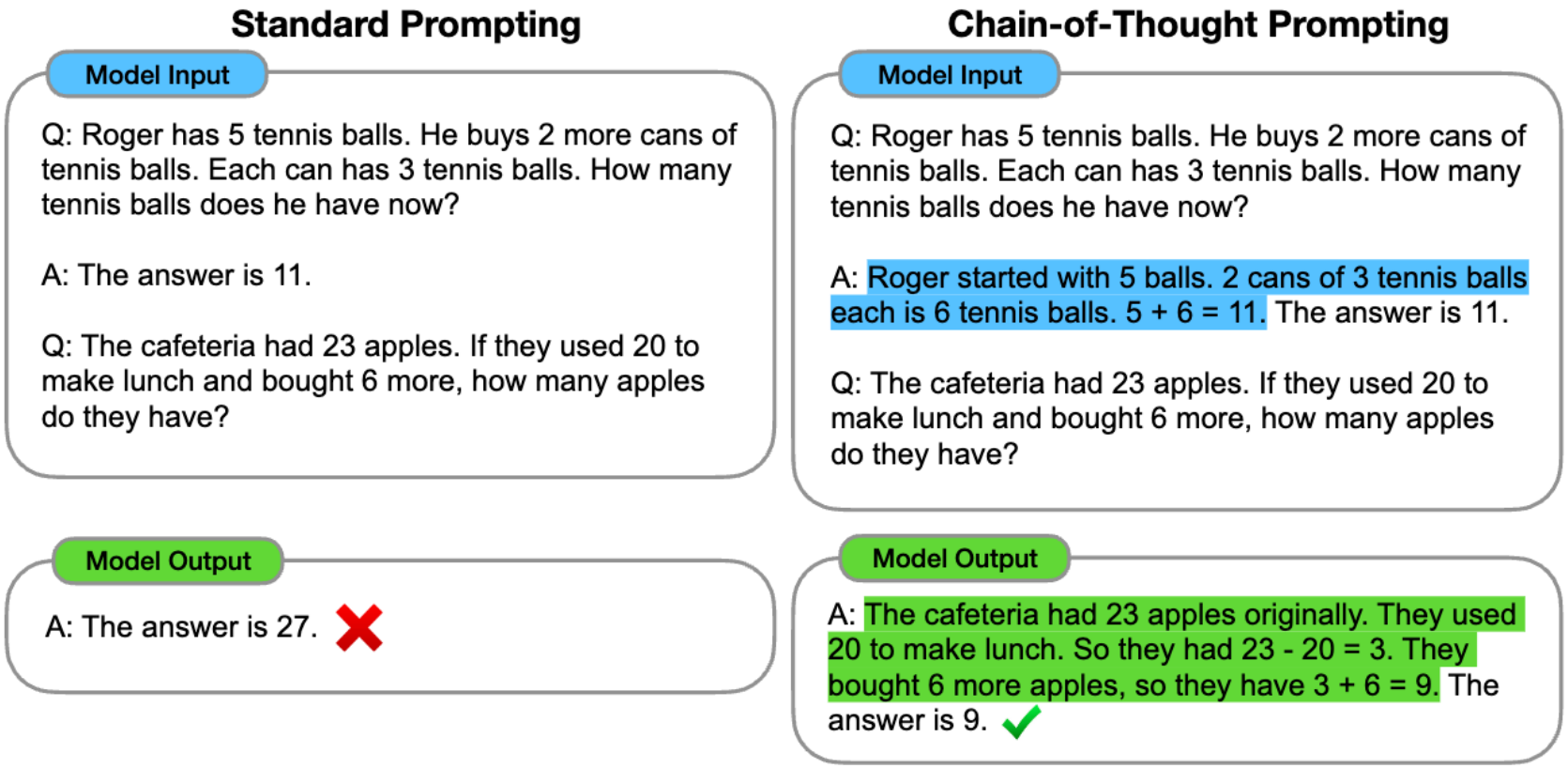
Johannes von Oswald [1 2] **Eyvind Niklasson** [2] **Ettore Randazzo** [2] **João Sacramento** [1]
Alexander Mordvintsev [2] **Andrey Zhmoginov** [2] **Max Vladymyrov** [2]



Transformers learn in-context by gradient descent, 2022

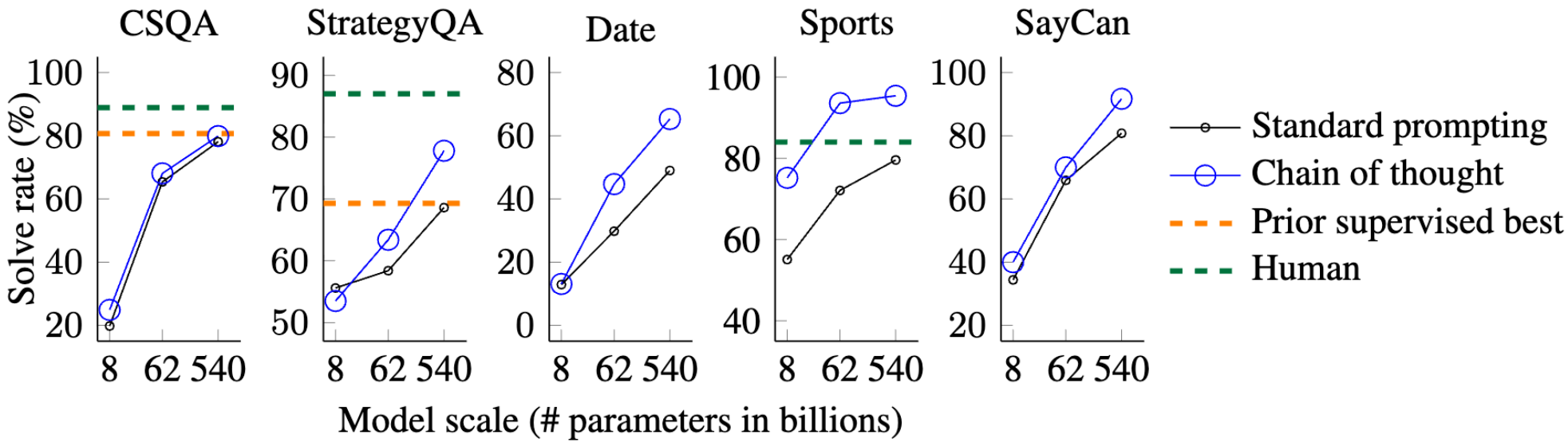
Chain-of-Thought (CoT) Prompting

- Provide reasoning chain improves performance



Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, 2022

Chain-of-Thought (CoT) Prompting



Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, 2022

What Makes Chain-of-Thought Work?

- Explicit reasoning steps
  - Models can think
- Knowledge expansion
  - Model can retrieve and use internal knowledge
- Possibility to refine answers
  - Model can do self-correction

Zero-Shot Chain-of-Thought Prompting

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The answer is 8. ✗

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are 16 / 2 = 8 golf balls. Half of the golf balls are blue. So there are 8 / 2 = 4 blue golf balls. The answer is 4. ✓

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: The answer (arabic numerals) is

(Output) 8 ✗

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: ***Let's think step by step.***

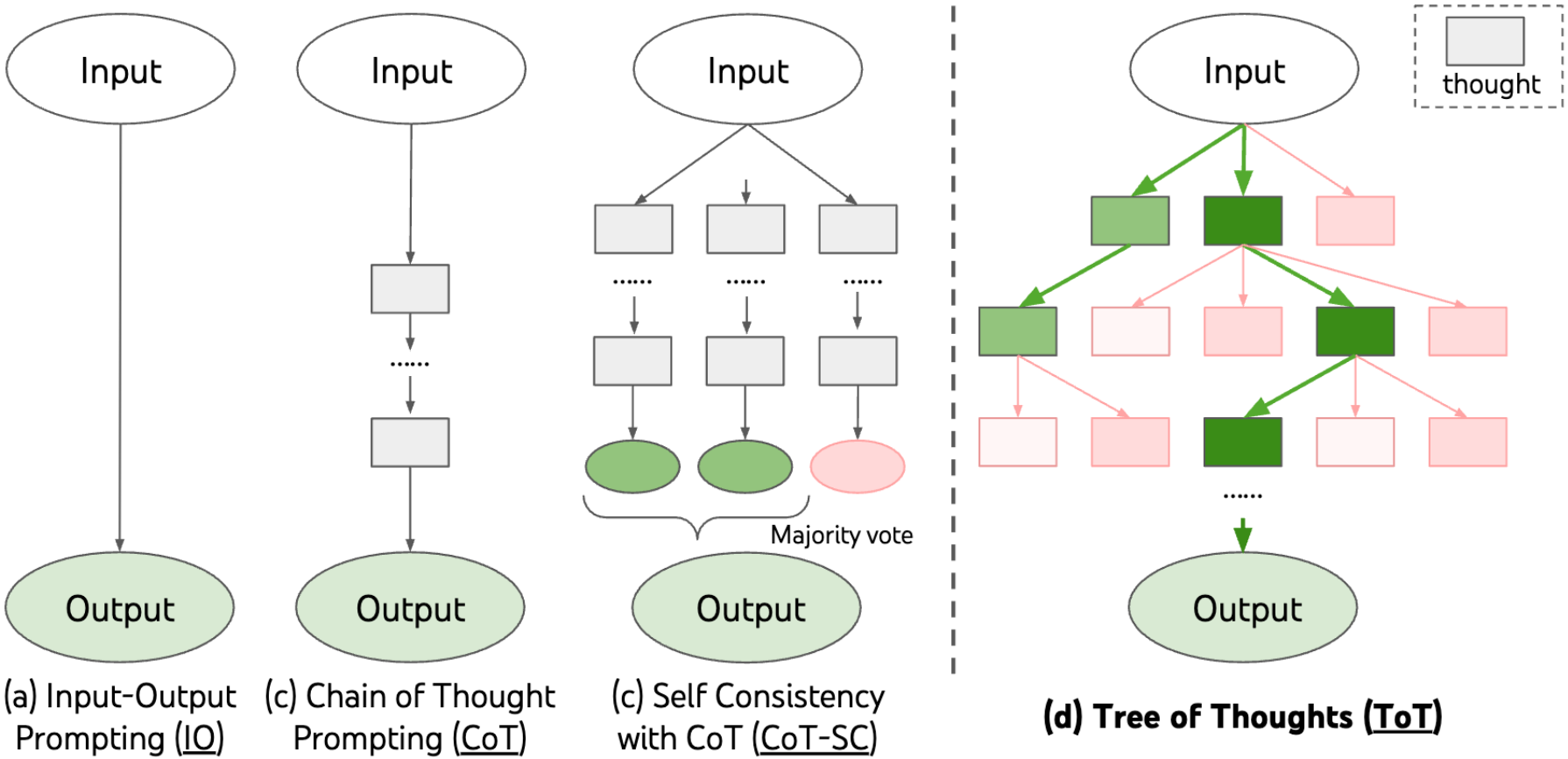
(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

Zero-Shot Chain-of-Thought Prompting

	MultiArith	GSM8K
Zero-Shot	**17.7**	**10.4**
Few-Shot (2 samples)	33.7	15.6
Few-Shot (8 samples)	33.8	15.6
Zero-Shot-CoT	**78.7**	**40.7**
Few-Shot-CoT (2 samples)	84.8	41.3
Few-Shot-CoT (4 samples : First) (*1)	89.2	-
Few-Shot-CoT (4 samples : Second) (*1)	90.5	-
Few-Shot-CoT (8 samples)	93.0	48.7
Zero-Plus-Few-Shot-CoT (8 samples) (*2)	**92.8**	**51.5**
Finetuned GPT-3 175B [Wei et al., 2022]	-	33
Finetuned GPT-3 175B + verifier [Wei et al., 2022]	-	55
PaLM 540B: Zero-Shot	**25.5**	**12.5**
PaLM 540B: Zero-Shot-CoT	**66.1**	**43.0**
PaLM 540B: Zero-Shot-CoT + self consistency	**89.0**	**70.1**
PaLM 540B: Few-Shot [Wei et al., 2022]	-	17.9
PaLM 540B: Few-Shot-CoT [Wei et al., 2022]	-	56.9
PaLM 540B: Few-Shot-CoT + self consistency [Wang et al., 2022]	-	74.4

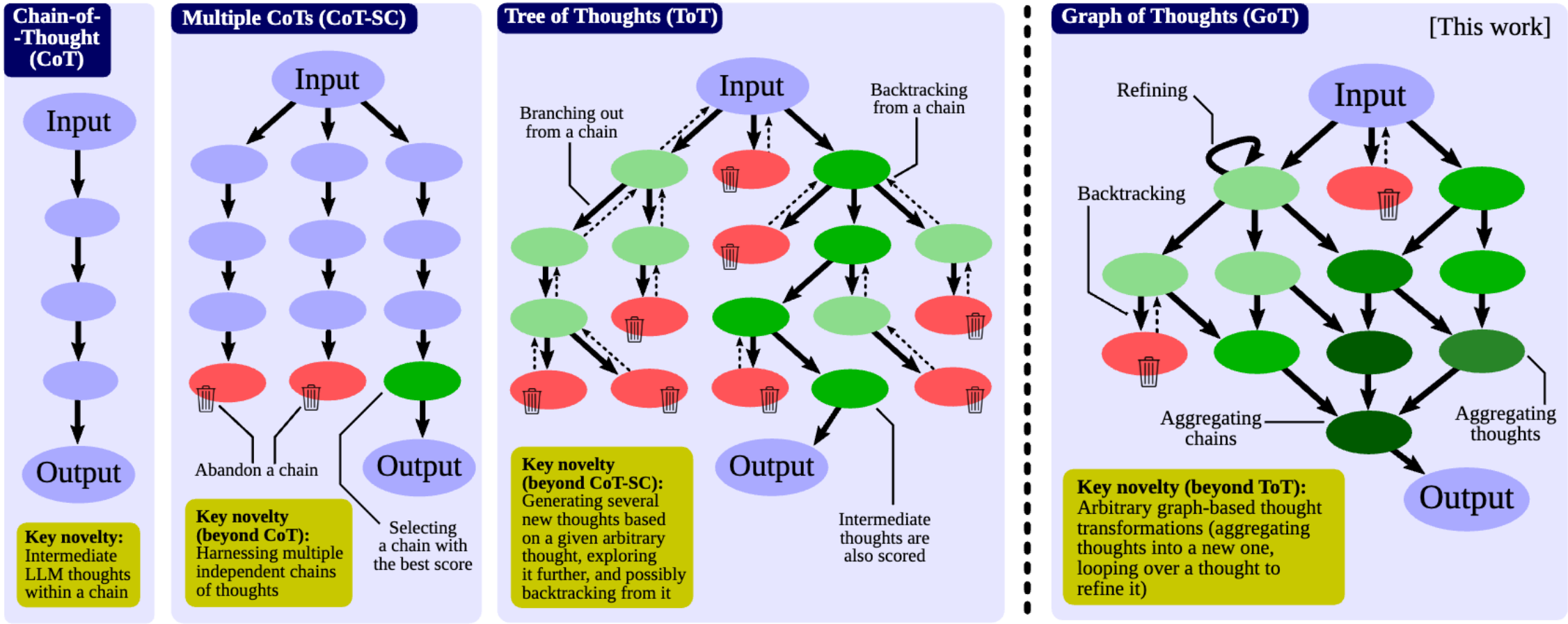
Large Language Models are Zero-Shot Reasoners, 2022

Tree-of-Thoughts



Tree of Thoughts: Deliberate Problem Solving with Large Language Models, 2023

Graph-of-Thoughts



Graph of Thoughts: Solving Elaborate Problems with Large Language Models, 2023

Model Reasoning is The Key

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: ***Let's think step by step.***

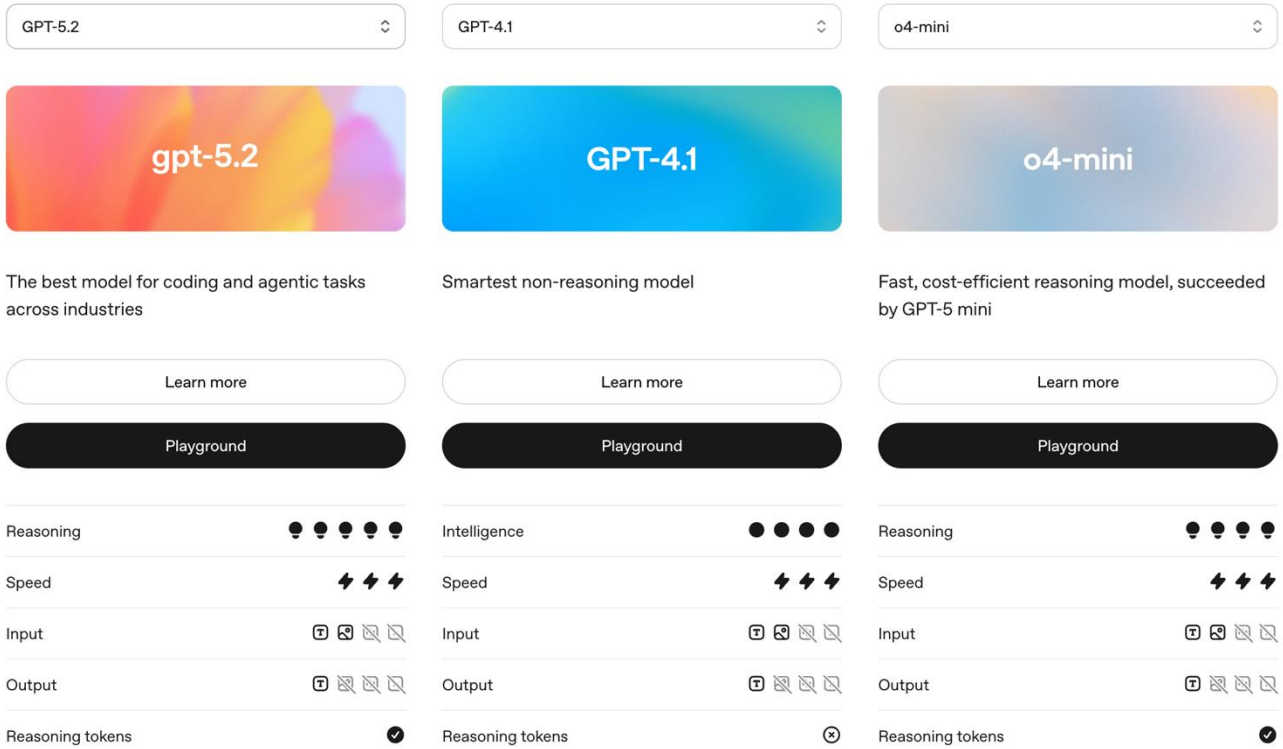
(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

How to encourage model to generate better reasoning chain?

Reasoning Models

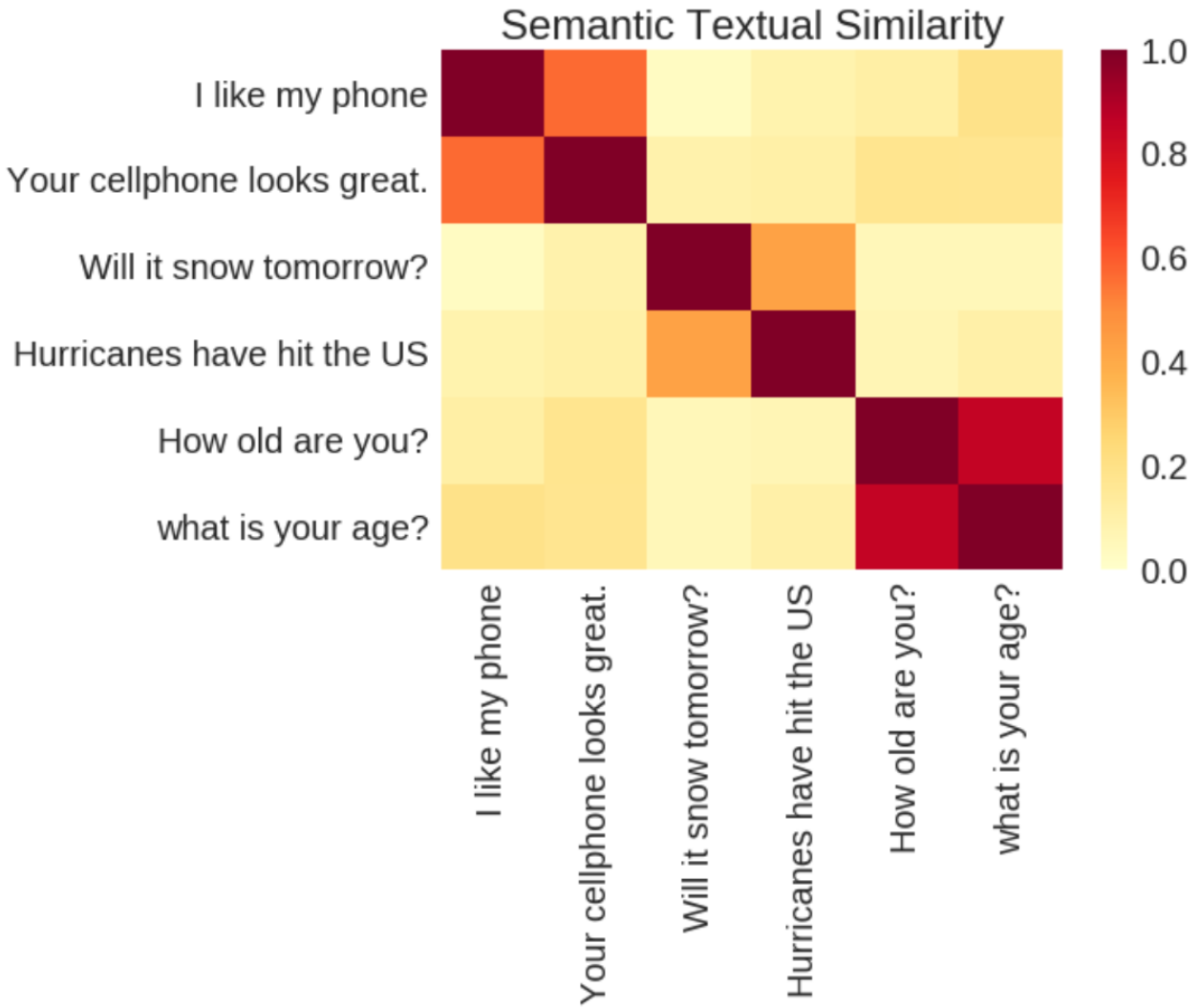
Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a+x}} = x$ is equal to

Response: <think>
To solve the equation $\sqrt{a - \sqrt{a+x}} = x$, let's start by squaring both $\cdots$
$\left(\sqrt{a - \sqrt{a+x}}\right)^2 = x^2 \implies a - \sqrt{a+x} = x^2.$
Rearrange to isolate the inner square root term:
$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$
$\cdots$
Wait, wait. Wait. That's an aha moment I can flag here.
Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$
We started with the equation:
$\sqrt{a - \sqrt{a+x}} = x$
First, let's square both sides:
$a - \sqrt{a+x} = x^2 \implies \sqrt{a+x} = a - x^2$
Next, I could square both sides again, treating the equation: $\cdots$
$\cdots$

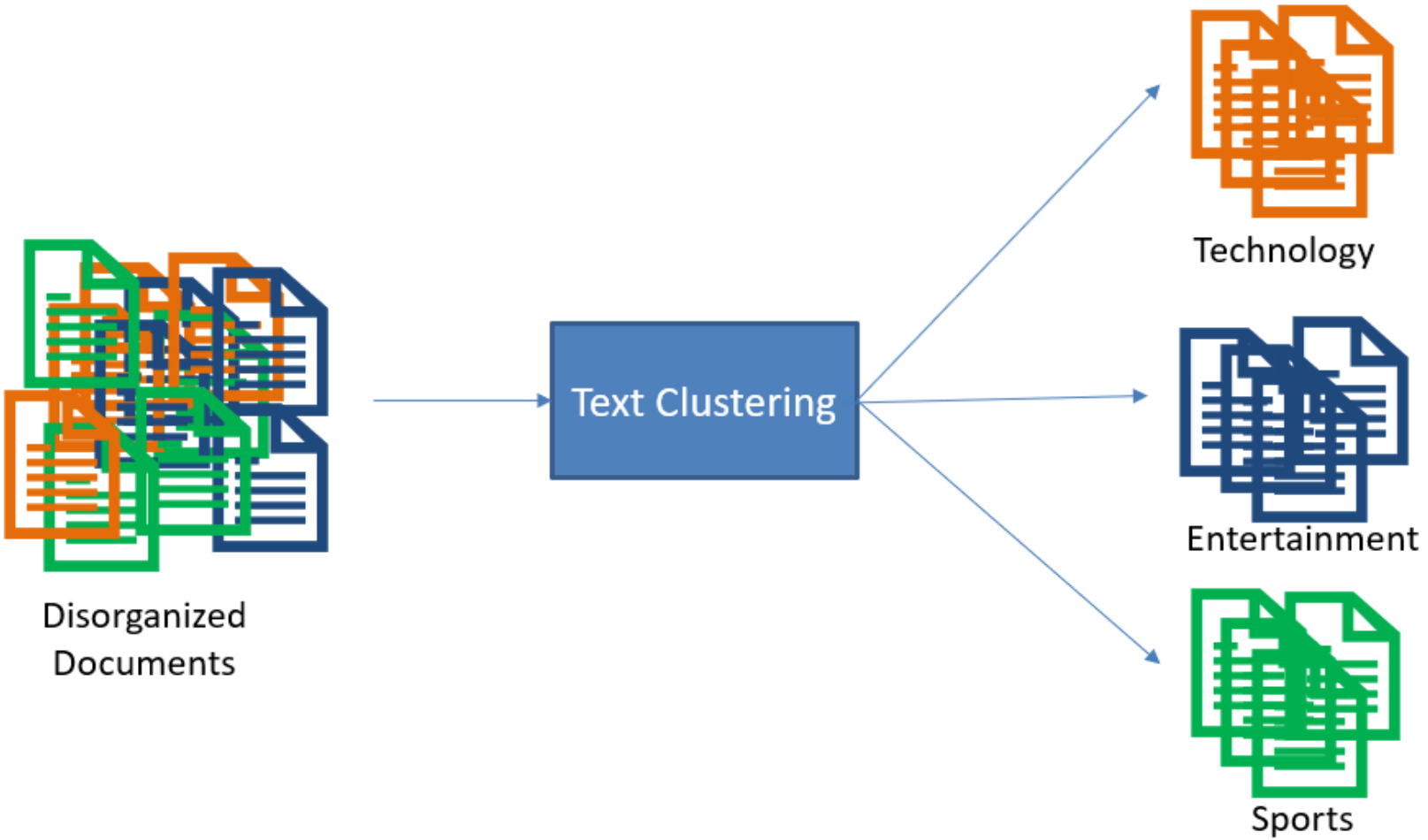


https://developers.openai.com/api/docs/models/compare
DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, 2025

Text Similarity

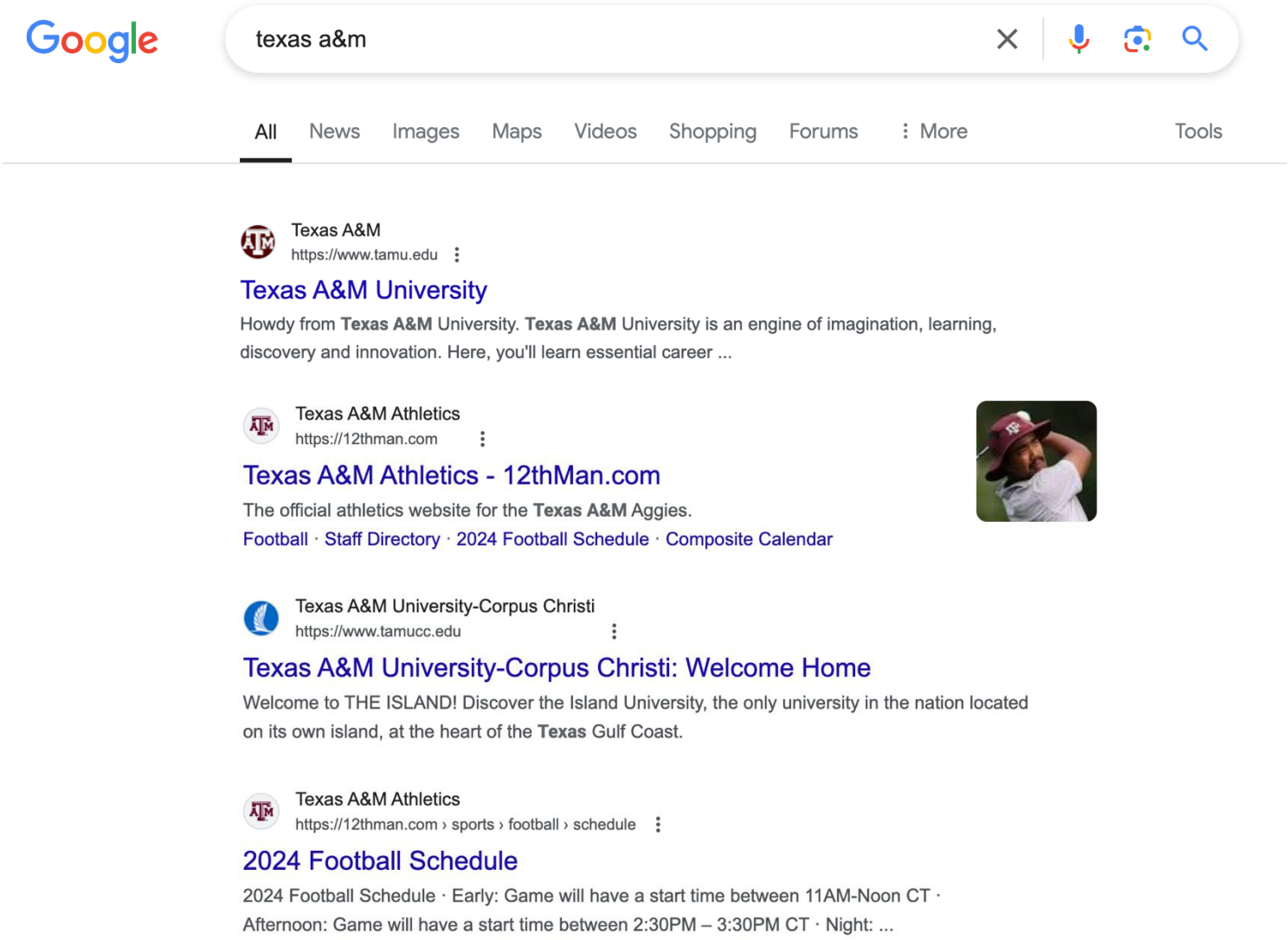


Document Clustering



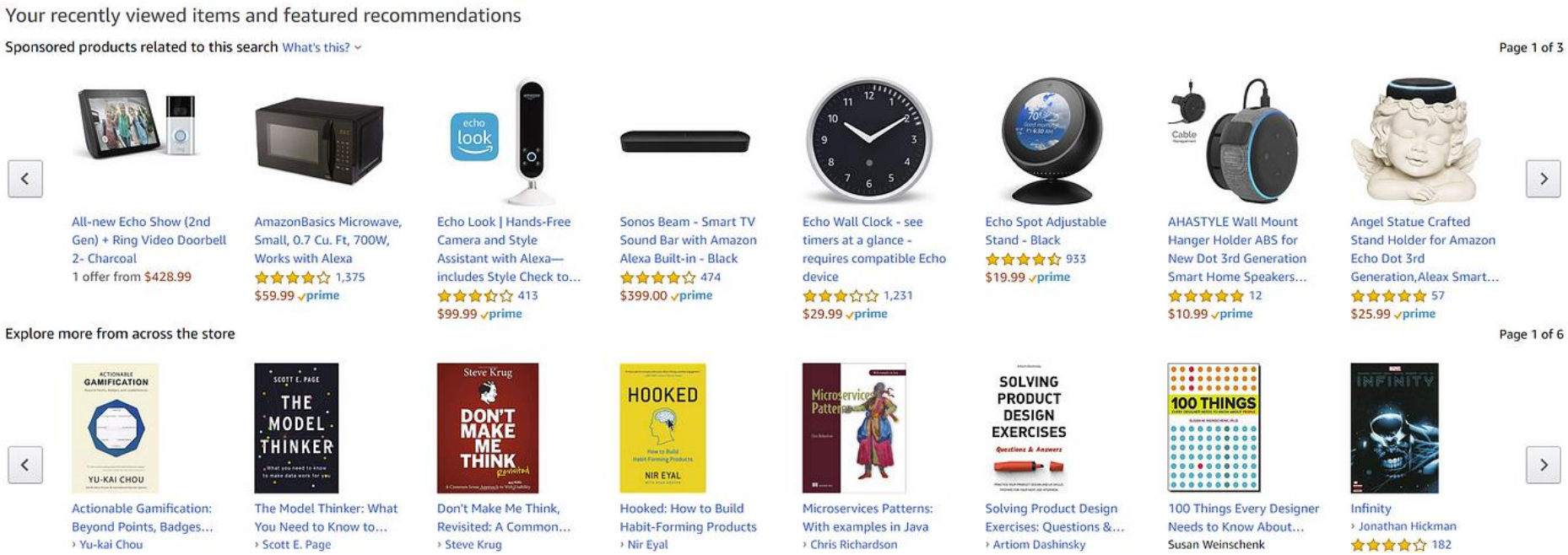
https://medium.com/@danielafrimi/text-clustering-using-nlp-techniques-c2e6b08b6e95

Information Retrieval



https://www.google.com/

Recommendation Systems



https://www.amazon.com/

Semantic Quality Control

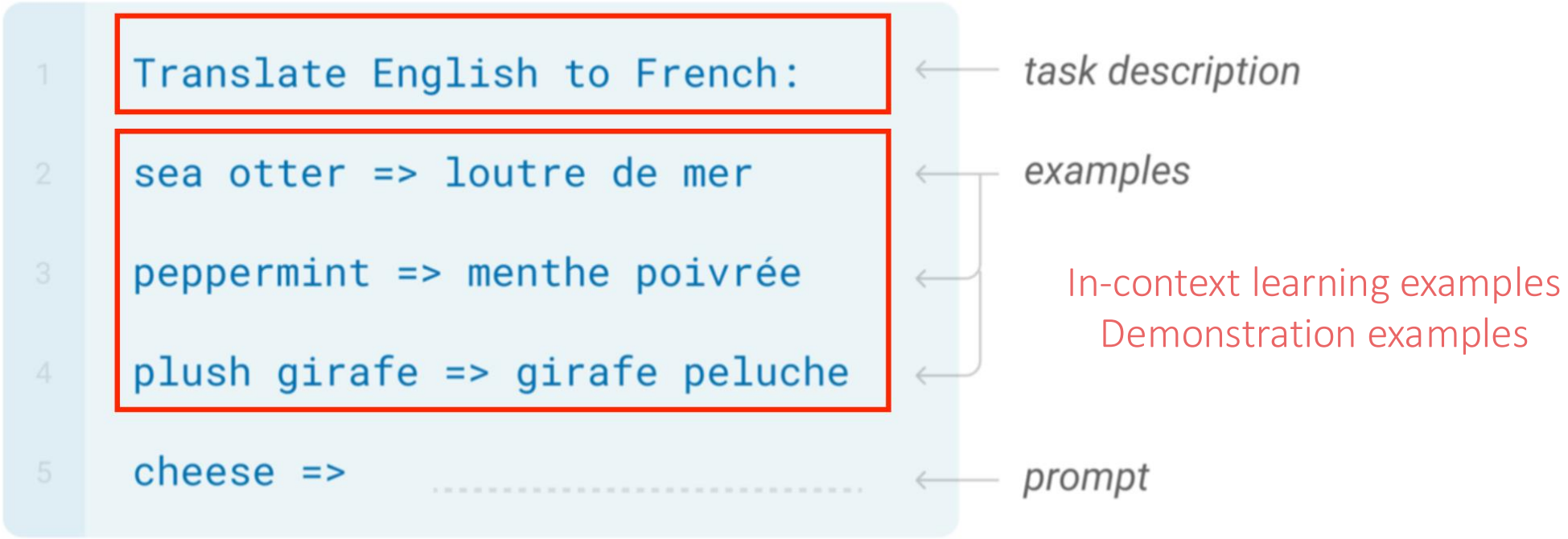
- Paraphrase generation

We will go hiking if tomorrow is a sunny day.

If it is sunny tomorrow, we will go hiking.

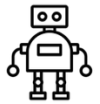
- Style transfer
- Plagiarism detection
- …

In-Context Example Selection



Language Models are Few-Shot Learners, 2020

Semantic Textual Similarity Benchmark

			
A soccer player is kicking the soccer ball into the goal from a long way down the field.	*A soccer player kicks the ball into the goal.*	3.25	3.94
Earlier this month, RIM had said it expected to report second-quarter earnings of between 7 cents and 11 cents a share.	*Excluding legal fees and other charges it expected a loss of between 1 and 4 cents a share.*	1.2	0.5
…	…	…	…
David Beckham Announces Retirement From Soccer.	*David Beckham retires from football.*	4.4	3.8

Pearson's Correlation Coefficient

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

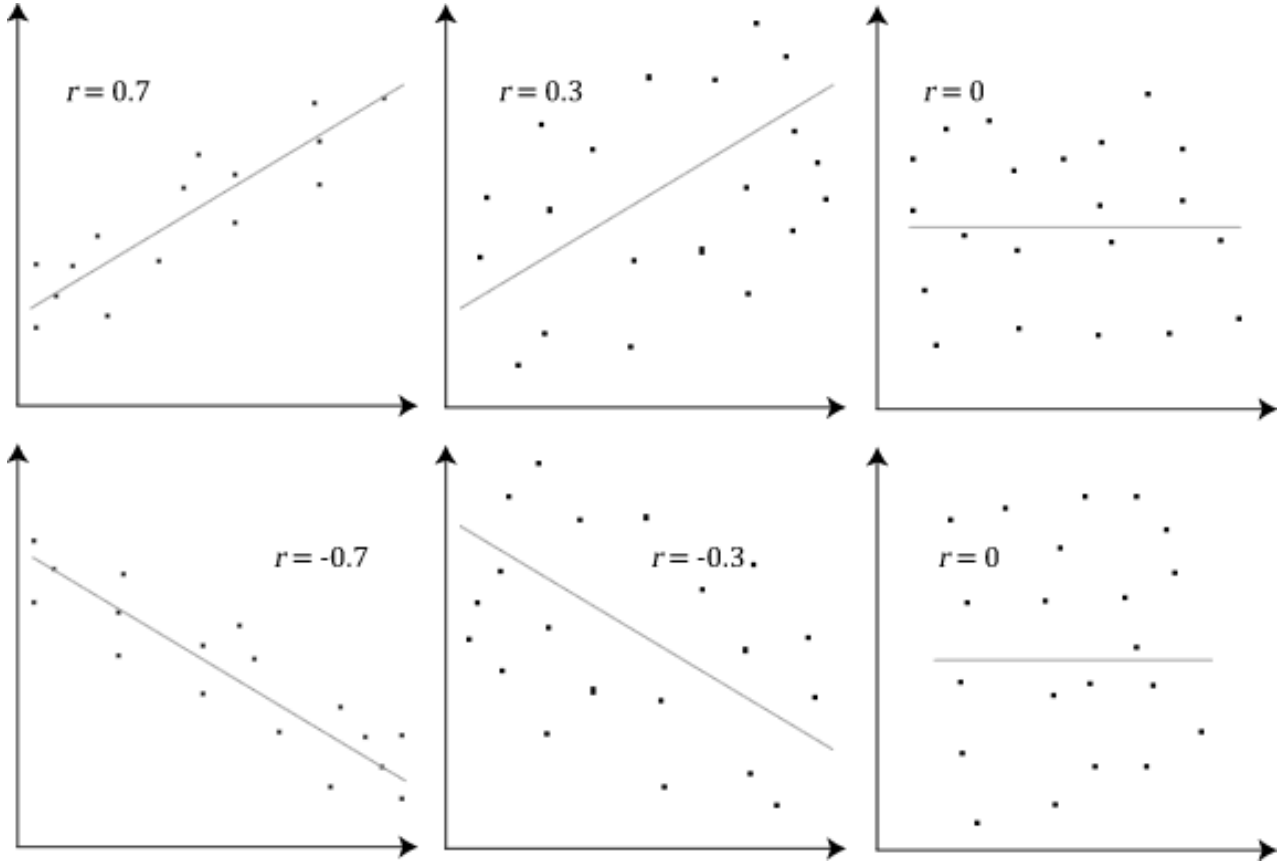
r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

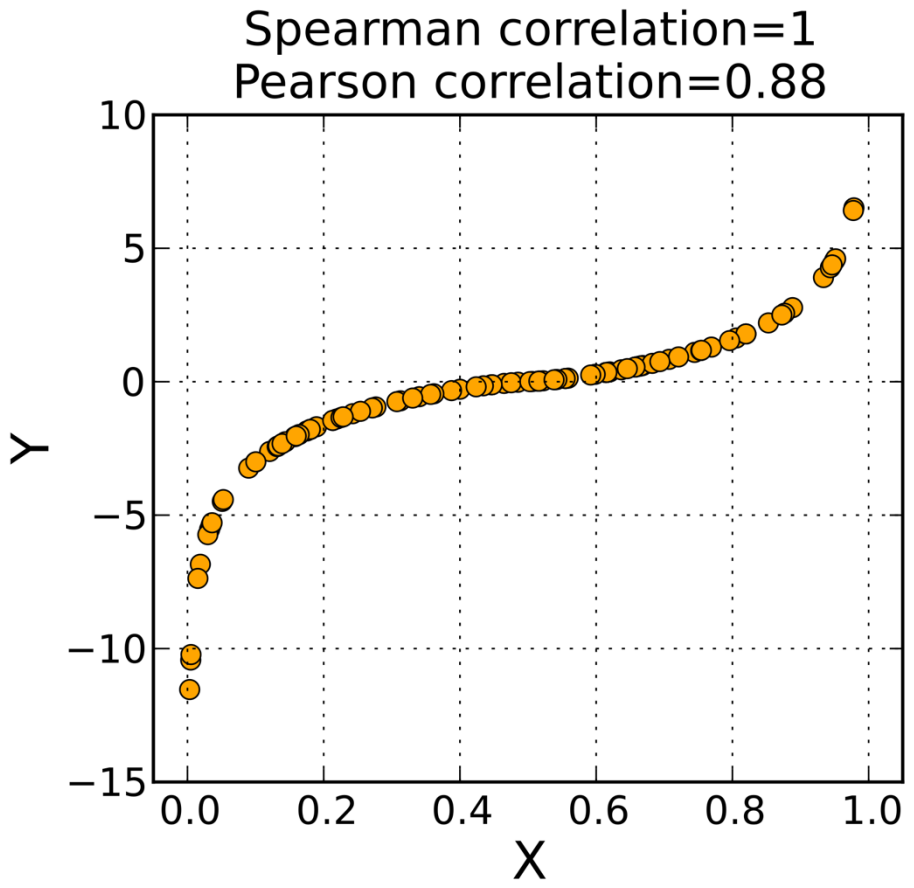
$\bar{y}$ = mean of the values of the y-variable



https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

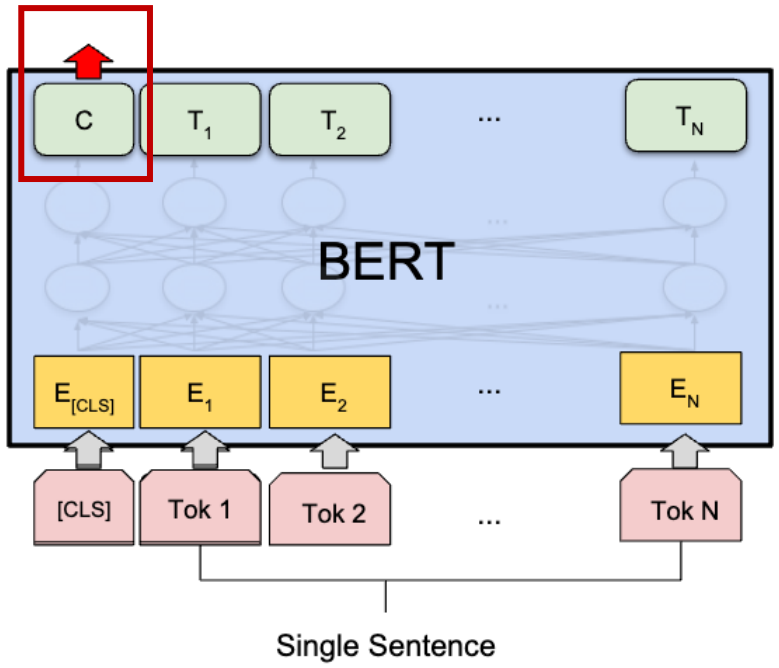
Spearman's Correlation Coefficient

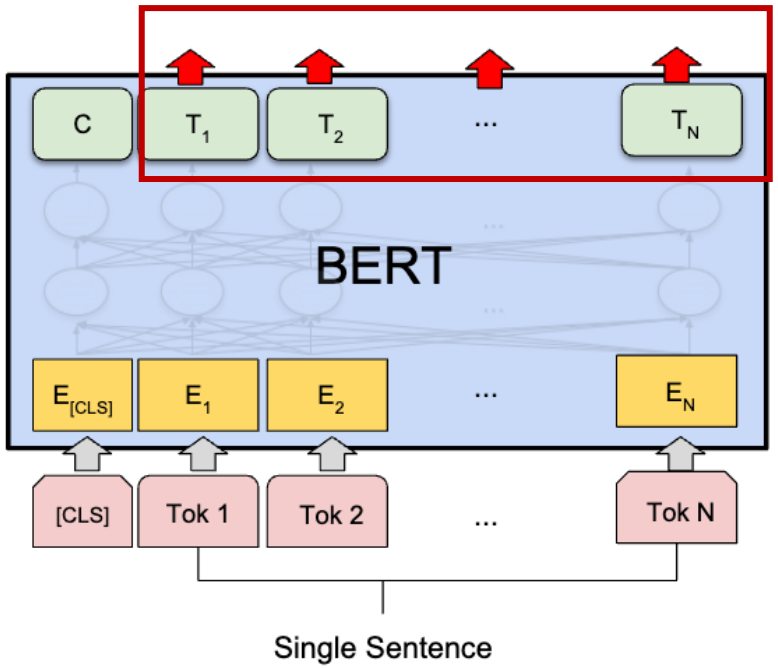
- Pearson's correlation coefficient on rank
- Score
  - Human: [1.2, 3.4, 2.5, 0.7, 4.0]
  - Machine: [0.5, 3.3, 1.0, 1.2, 3.4]
- Rank
  - Human: [4, 2, 3, 5, 1]
  - Machine: [5, 2, 4, 3, 1]
- Assesses monotonic relationships
  - whether linear or not



https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

A Simple Approach: Text Encoder + Cosine Similarity





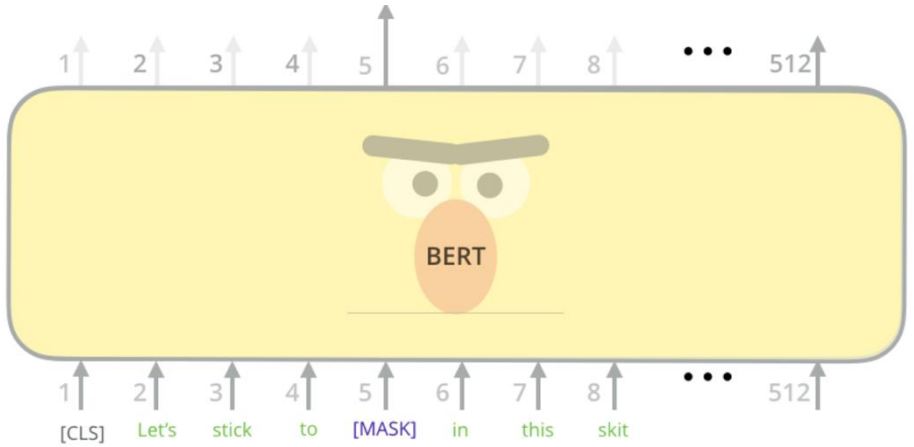
$$E_1 = \text{Encoder}(S_1)$$

$$E_2 = \text{Encoder}(S_2)$$

$$\text{Similarity}(S_1, S_2) = \frac{E_1 \cdot E_2}{\|E_1\|\|E_2\|}$$

Unfortunately, the performance is bad (why?)

A Simple Approach: Text Encoder + Cosine Similarity



Pre-trained BERT embeddings are more about lexical information

If it is sunny tomorrow, we will go hiking.

We will go hiking if tomorrow is a sunny day.

Good classification performance ≠ Good similarity

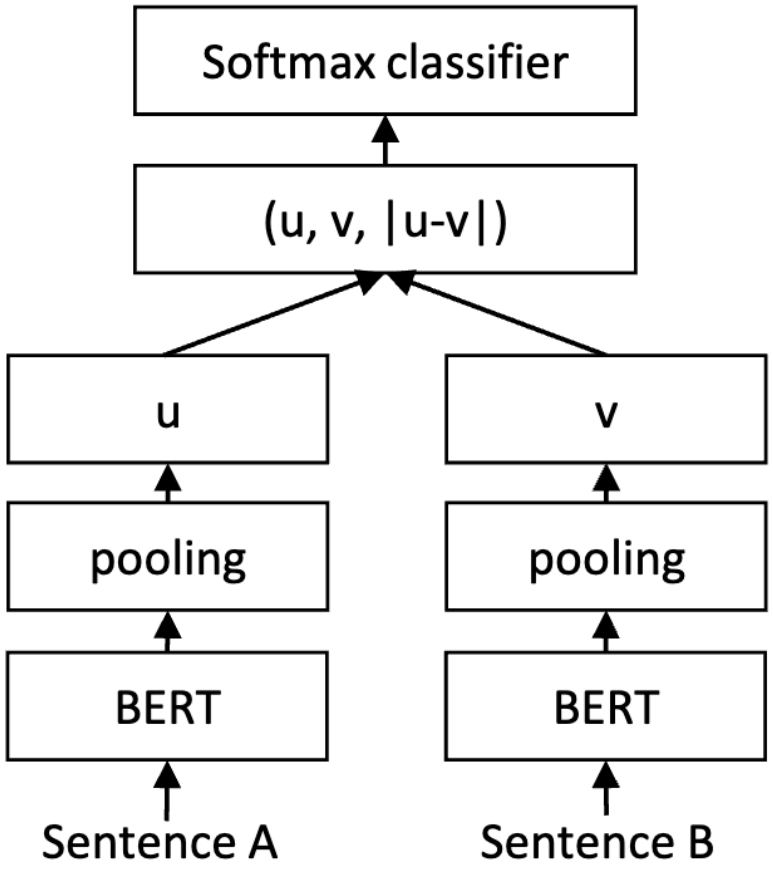
Sentence-BERT

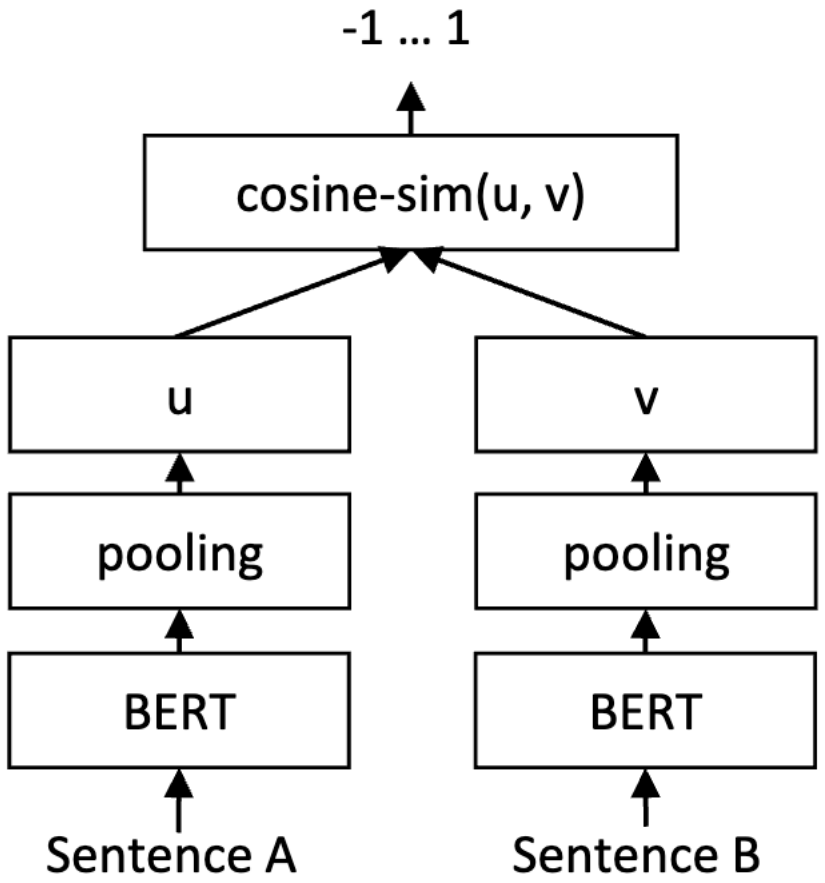
- Consider SNLI dataset
  - Stanford Natural Language Inference

A boy is jumping on skateboard in the middle of a red bridge.	*The boy skates down the sidewalk.*	Contradiction
A boy is jumping on skateboard in the middle of a red bridge.	*The boy is wearing safety equipment.*	Neutral
A boy is jumping on skateboard in the middle of a red bridge.	*The boy does a skateboarding trick.*	Entailment

Sentence-BERT

Contradiction Neutral Entailment

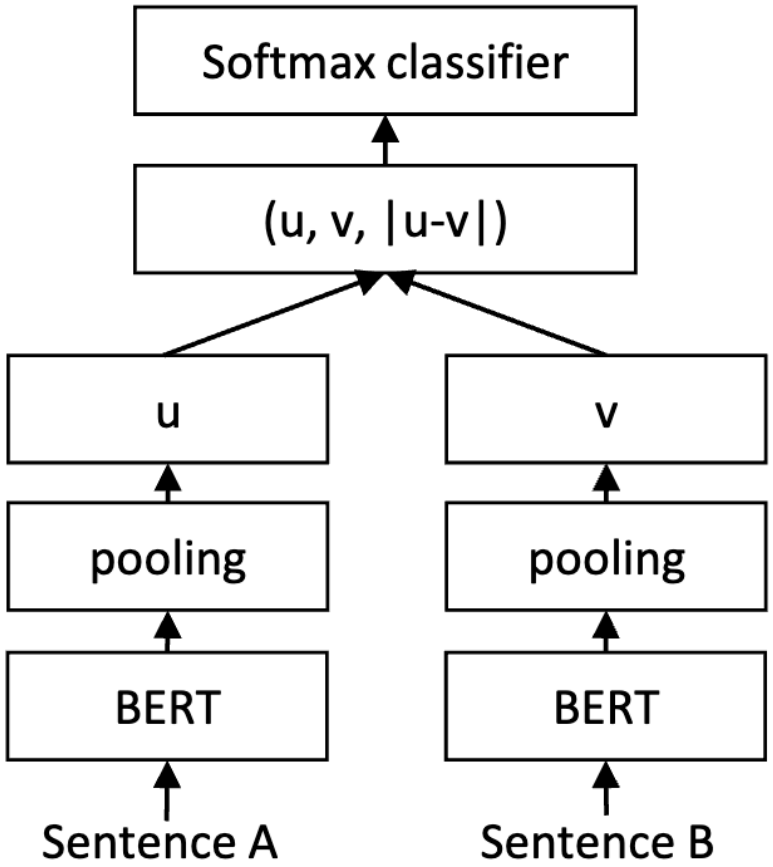




Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks

Sentence-BERT

Contradiction Neutral Entailment



Cross Entropy Loss

$$o = \text{softmax}(W_t(u, v, |u - v|))$$

Triplet Loss

$$max(||s_a - s_p|| - ||s_a - s_n|| + \epsilon, 0)$$

Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks

Sentence-BERT

If it is sunny tomorrow, we will go hiking.

We will go hiking if tomorrow is a sunny day.

Sentence-BERT: Performance

Model	STS12	STS13	STS14	STS15	STS16	STSb	SICK-R	Avg.
Avg. GloVe embeddings	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
Avg. BERT embeddings	38.78	57.98	57.98	63.15	61.06	46.35	58.40	54.81
BERT CLS-vector	20.16	30.01	20.09	36.88	38.08	16.50	42.63	29.19
InferSent - Glove	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
Universal Sentence Encoder	64.49	67.80	64.61	76.83	73.18	74.92	**76.69**	71.22
SBERT-NLI-base	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
SBERT-NLI-large	72.27	**78.46**	**74.90**	80.99	76.25	**79.23**	73.75	76.55
SRoBERTa-NLI-base	71.54	72.49	70.80	78.74	73.69	77.77	74.46	74.21
SRoBERTa-NLI-large	**74.53**	77.00	73.18	**81.85**	**76.82**	79.10	74.29	**76.68**

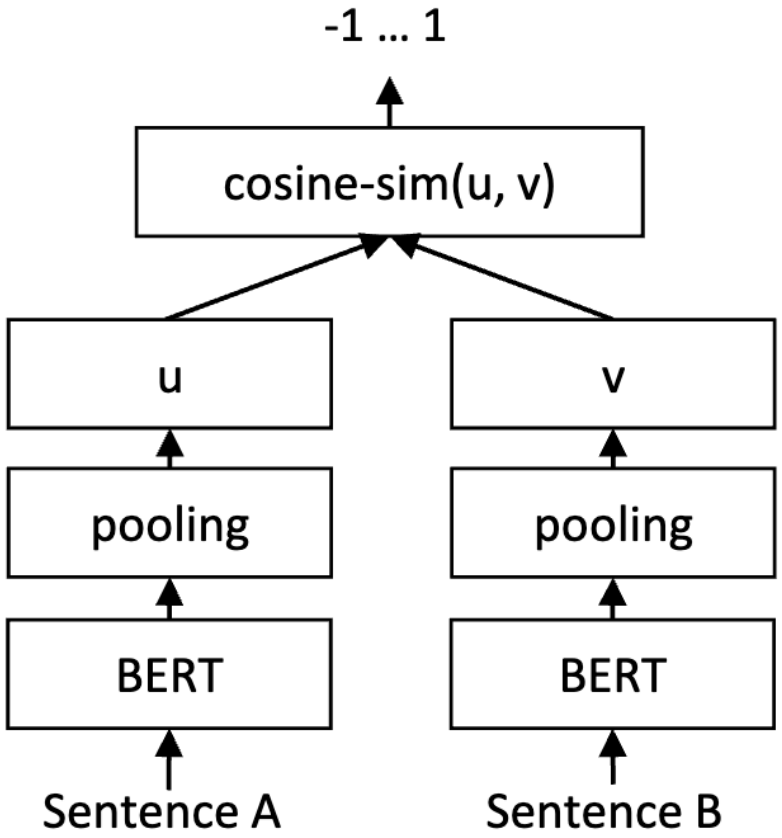
SentenceTransformers

- https://sbert.net/

```python
from sentence_transformers import SentenceTransformer

# 1. Load a pretrained Sentence Transformer model
model = SentenceTransformer("all-MiniLM-L6-v2")

# The sentences to encode
sentences = [
    "The weather is lovely today.",
    "It's so sunny outside!",
    "He drove to the stadium.",
]

# 2. Calculate embeddings by calling model.encode()
embeddings = model.encode(sentences)
print(embeddings.shape)
# [3, 384]

# 3. Calculate the embedding similarities
similarities = model.similarity(embeddings, embeddings)
print(similarities)
# tensor([[1.0000, 0.6660, 0.1046],
#         [0.6660, 1.0000, 0.1411],
#         [0.1046, 0.1411, 1.0000]])
```

SimCSE

- Simple Contrastive Learning of Sentence Embeddings



Contrastive Loss

$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_j^+)/\tau}}$$

SimCSE: Supervised Contrastive Learning

Sentence 1A	*Sentence 1B*	Positive
Sentence 2A	*Sentence 2B*	Negative
Sentence 3A	*Sentence 3B*	Negative
Sentence 4A	*Sentence 4B*	Negative
Sentence 5A	*Sentence 5B*	Negative

Contrastive Loss

$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_j^+)/\tau}}$$

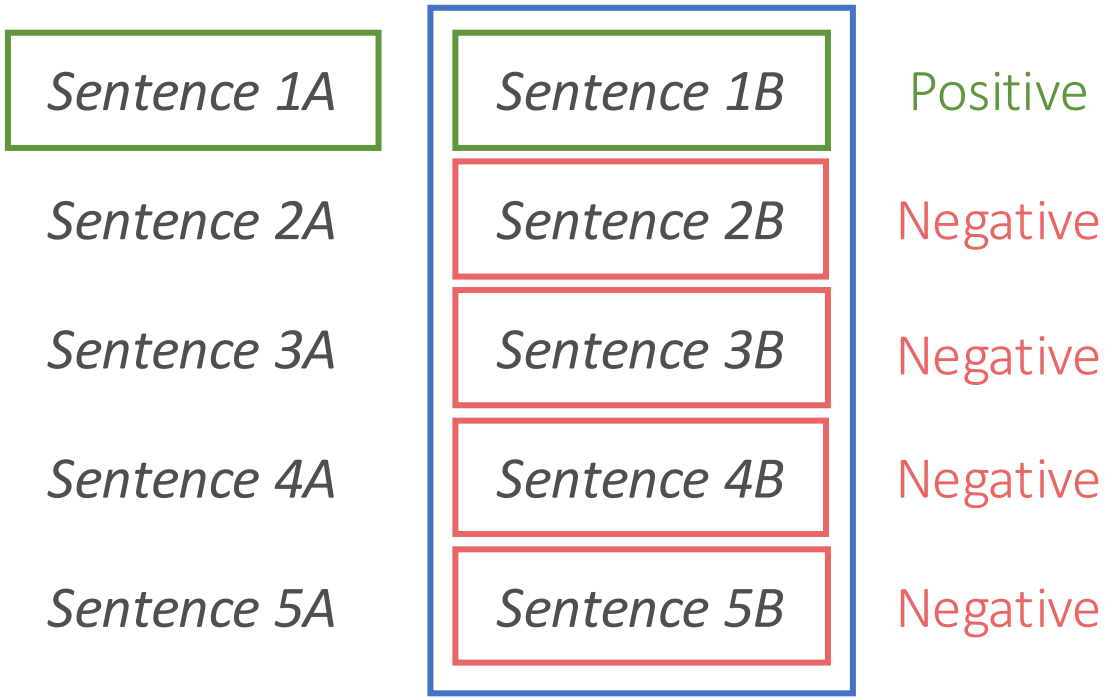
SimCSE: Supervised Contrastive Learning

Sentence 1A	*Sentence 1B*	Positive
Sentence 2A	*Sentence 2B*	
Sentence 3A	*Sentence 3B*	
Sentence 4A	*Sentence 4B*	
Sentence 5A	*Sentence 5B*	

Contrastive Loss

$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_j^+)/\tau}}$$

SimCSE: Supervised Contrastive Learning



Contrastive Loss

$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_j^+)/\tau}}$$

SimCSE: Supervised Contrastive Learning

Sentence 1A	*Sentence 1B*	Negative
Sentence 2A	*Sentence 2B*	Positive
Sentence 3A	*Sentence 3B*	Negative
Sentence 4A	*Sentence 4B*	Negative
Sentence 5A	*Sentence 5B*	Negative

Contrastive Loss

$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_j^+)/\tau}}$$

SimCSE: Supervised Contrastive Learning

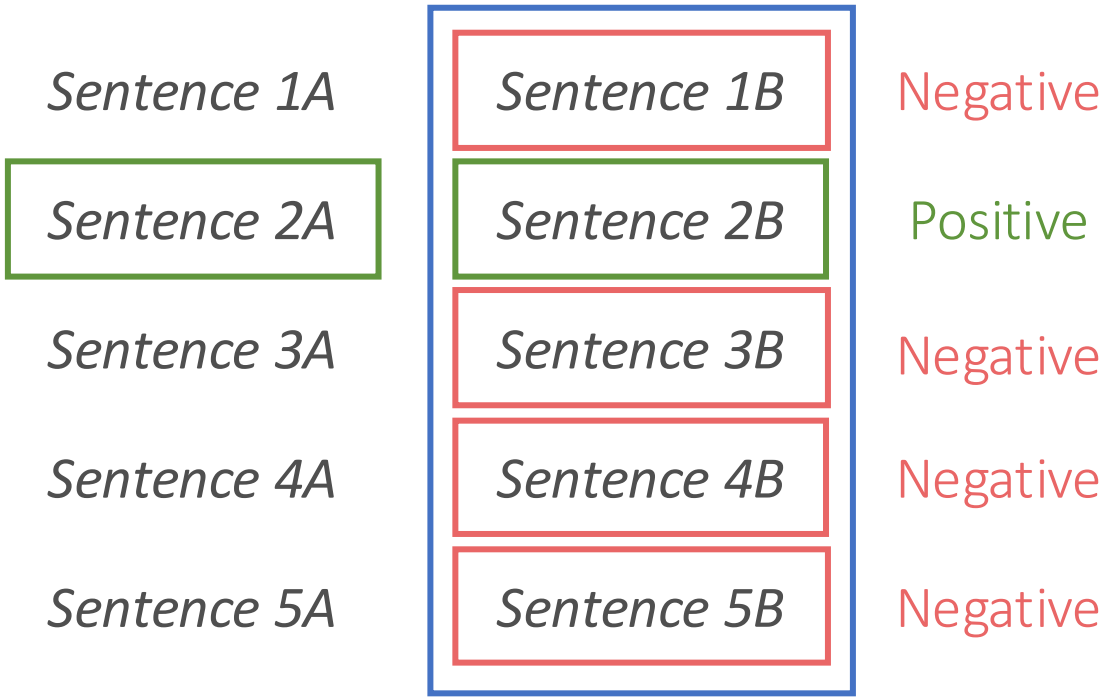
Sentence 1A	*Sentence 1B*
Sentence 2A	*Sentence 2B*
Sentence 3A	*Sentence 3B*
Sentence 4A	*Sentence 4B*
Sentence 5A	*Sentence 5B*

Positive

Contrastive Loss

$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_j^+)/\tau}}$$

SimCSE: Supervised Contrastive Learning

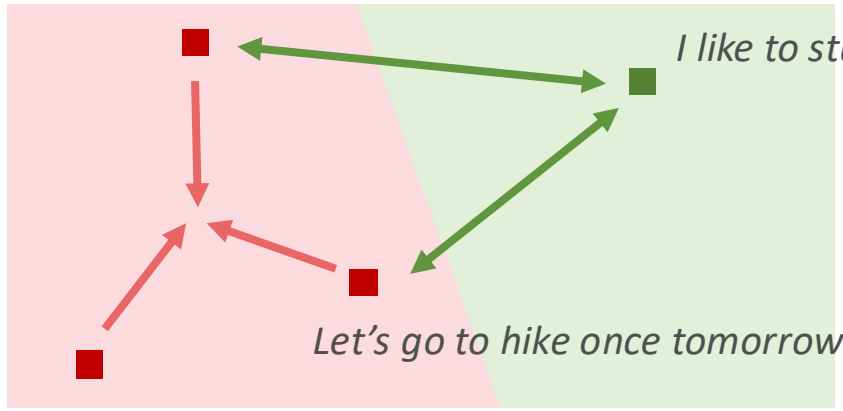


Contrastive Loss

$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_j^+)/\tau}}$$

SimCSE: Supervised Contrastive Learning

If it is sunny tomorrow, we will go hiking.



I like to study large language models.

Let's go to hike once tomorrow is sunny.

We will go hiking if tomorrow is a sunny day.

SimCSE: Performance

Supervised models								
InferSent-GloVe♣	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
Universal Sentence Encoder♣	64.49	67.80	64.61	76.83	73.18	74.92	76.69	71.22
$\text{SBERT}_{\text{base}}$♣	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
$\text{SBERT}_{\text{base}}$-flow	69.78	77.27	74.35	82.01	77.46	79.12	76.21	76.60
$\text{SBERT}_{\text{base}}$-whitening	69.65	77.57	74.66	82.27	78.39	79.52	76.91	77.00
$\text{CT-SBERT}_{\text{base}}$	74.84	83.20	78.07	83.84	77.93	81.46	76.42	79.39
∗ $\text{SimCSE-BERT}_{\text{base}}$	**75.30**	**84.67**	**80.19**	**85.40**	**80.82**	**84.25**	**80.39**	**81.57**
$\text{SRoBERTa}_{\text{base}}$♣	71.54	72.49	70.80	78.74	73.69	77.77	74.46	74.21
$\text{SRoBERTa}_{\text{base}}$-whitening	70.46	77.07	74.46	81.64	76.43	79.49	76.65	76.60
∗ $\text{SimCSE-RoBERTa}_{\text{base}}$	**76.53**	**85.21**	**80.95**	**86.03**	**82.57**	**85.83**	**80.50**	**82.52**
∗ $\text{SimCSE-RoBERTa}_{\text{large}}$	**77.46**	**87.27**	**82.36**	**86.66**	**83.93**	**86.70**	**81.95**	**83.76**

Supervised Data

- Consider SNLI dataset
  - Stanford Natural Language Inference

A boy is jumping on skateboard in the middle of a red bridge.	*The boy skates down the sidewalk.*	Contradiction
A boy is jumping on skateboard in the middle of a red bridge.	*The boy is wearing safety equipment.*	Neutral
A boy is jumping on skateboard in the middle of a red bridge.	*The boy does a skateboarding trick.*	Entailment

What if we don't have supervised data?

SimCSE: Unsupervised Contrastive Learning

Sentence 1	*Sentence 1'*	Positive
Sentence 2		Negative
Sentence 3		Negative
Sentence 4		Negative
Sentence 5		Negative

Contrastive Loss

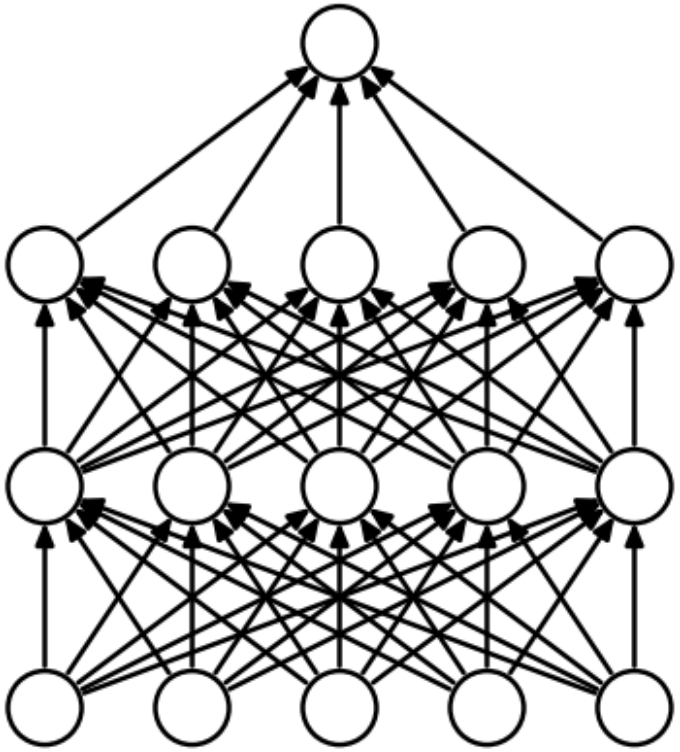
$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_j^+)/\tau}}$$

Generate positive example with masking

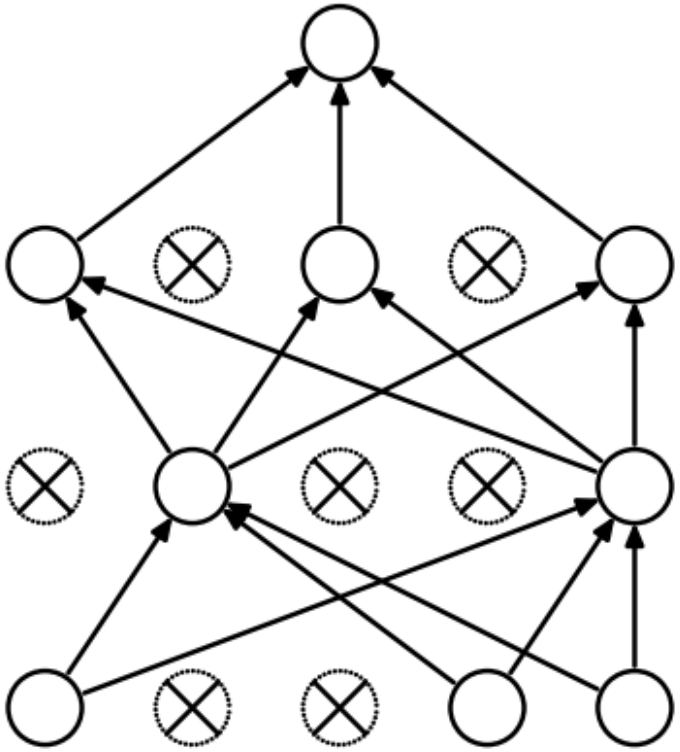
If it is sunny tomorrow, we will go hiking.

If [mask] is sunny tomorrow, we [mask] go hiking.

Dropout



(a) Standard Neural Net



(b) After applying dropout.

Generate positive example with neuron masking

Dropout: A Simple Way to Prevent Neural Networks from Overfitting

SimCSE: Unsupervised Contrastive Learning

Sentence 1 *Sentence 1'* Positive

Sentence 2

Sentence 3

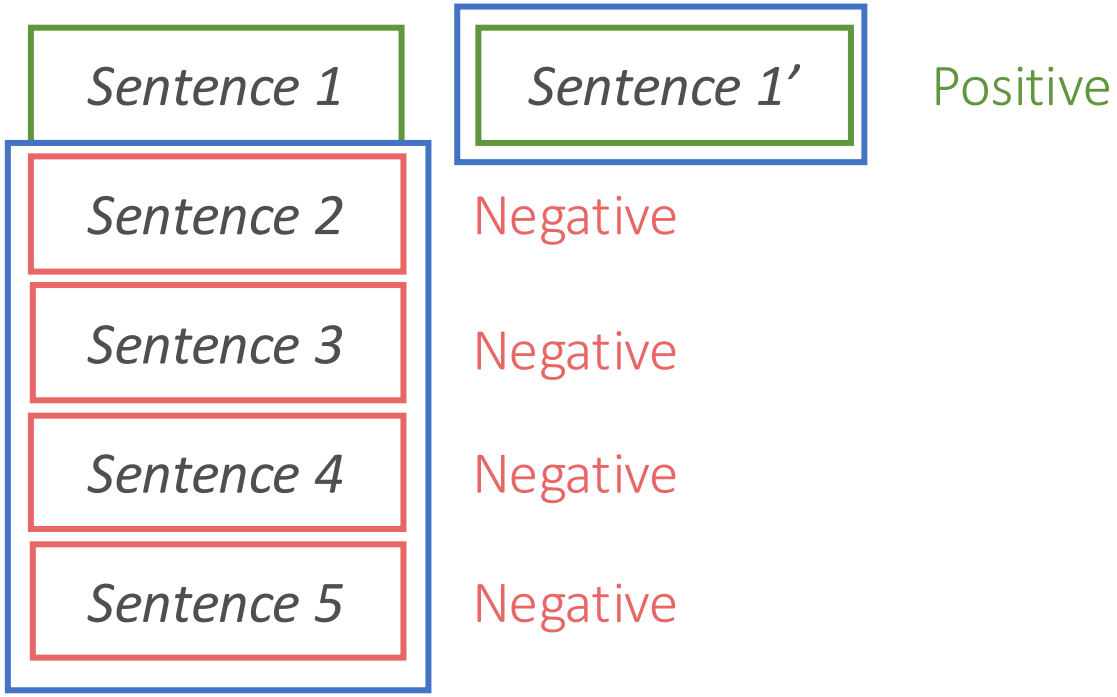
Sentence 4

Sentence 5

Contrastive Loss

$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_j^+)/\tau}}$$

SimCSE: Unsupervised Contrastive Learning



Contrastive Loss

$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_j^+)/\tau}}$$

SimCSE: Unsupervised Contrastive Learning

Sentence 1

Sentence 2

Sentence 2'

Positive

Sentence 3

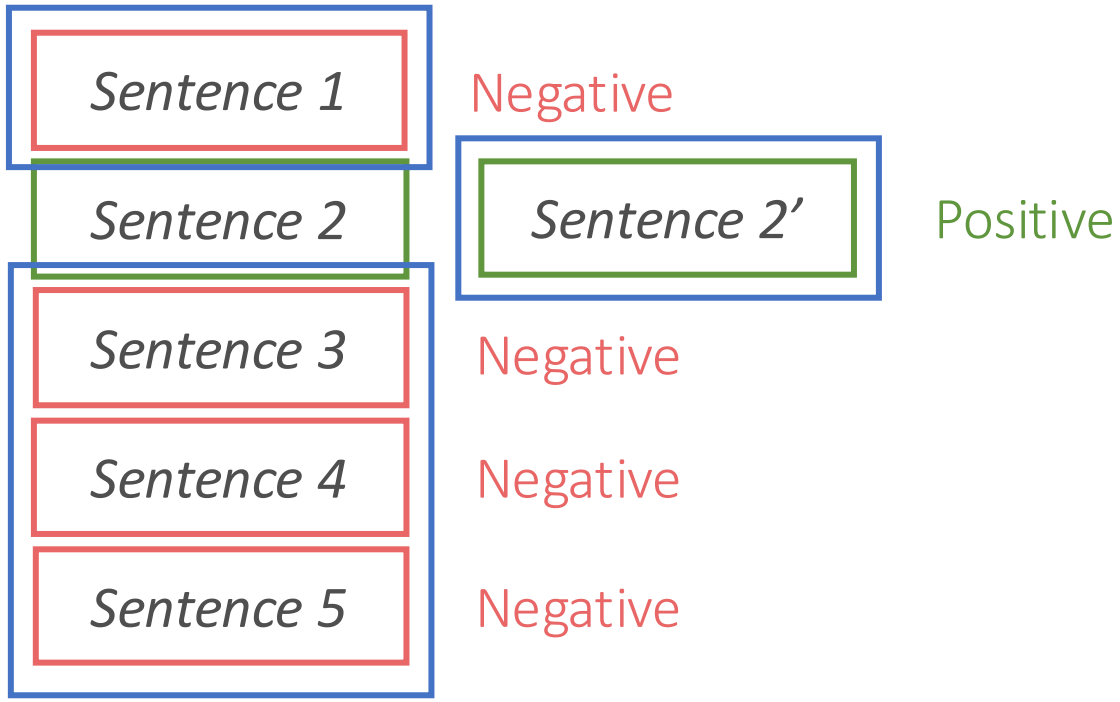
Sentence 4

Sentence 5

Contrastive Loss

$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_j^+)/\tau}}$$

SimCSE: Unsupervised Contrastive Learning



Contrastive Loss

$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_j^+)/\tau}}$$

SimCSE: Performance

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
Unsupervised models								
GloVe embeddings (avg.)♣	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
$\text{BERT}_{\text{base}}$ (first-last avg.)	39.70	59.38	49.67	66.03	66.19	53.87	62.06	56.70
$\text{BERT}_{\text{base}}$-flow	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
$\text{BERT}_{\text{base}}$-whitening	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
IS-$\text{BERT}_{\text{base}}$♡	56.77	69.24	61.21	75.23	70.16	69.21	64.25	66.58
CT-$\text{BERT}_{\text{base}}$	61.63	76.80	68.47	77.50	76.48	74.31	69.19	72.05
∗ SimCSE-$\text{BERT}_{\text{base}}$	**68.40**	**82.41**	**74.38**	**80.91**	**78.56**	**76.85**	**72.23**	**76.25**
$\text{RoBERTa}_{\text{base}}$ (first-last avg.)	40.88	58.74	49.07	65.63	61.48	58.55	61.63	56.57
$\text{RoBERTa}_{\text{base}}$-whitening	46.99	63.24	57.23	71.36	68.99	61.36	62.91	61.73
DeCLUTR-$\text{RoBERTa}_{\text{base}}$	52.41	75.19	65.52	77.12	78.63	72.41	**68.62**	69.99
∗ SimCSE-$\text{RoBERTa}_{\text{base}}$	**70.16**	**81.77**	**73.24**	**81.36**	**80.65**	**80.22**	68.56	**76.57**
∗ SimCSE-$\text{RoBERTa}_{\text{large}}$	**72.86**	**83.99**	**75.62**	**84.77**	**81.80**	**81.98**	**71.26**	**78.90**
Supervised models								
InferSent-GloVe♣	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
Universal Sentence Encoder♣	64.49	67.80	64.61	76.83	73.18	74.92	76.69	71.22
$\text{SBERT}_{\text{base}}$♣	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
$\text{SBERT}_{\text{base}}$-flow	69.78	77.27	74.35	82.01	77.46	79.12	76.21	76.60
$\text{SBERT}_{\text{base}}$-whitening	69.65	77.57	74.66	82.27	78.39	79.52	76.91	77.00
CT-$\text{SBERT}_{\text{base}}$	74.84	83.20	78.07	83.84	77.93	81.46	76.42	79.39
∗ SimCSE-$\text{BERT}_{\text{base}}$	**75.30**	**84.67**	**80.19**	**85.40**	**80.82**	**84.25**	**80.39**	**81.57**
$\text{SRoBERTa}_{\text{base}}$♣	71.54	72.49	70.80	78.74	73.69	77.77	74.46	74.21
$\text{SRoBERTa}_{\text{base}}$-whitening	70.46	77.07	74.46	81.64	76.43	79.49	76.65	76.60
∗ SimCSE-$\text{RoBERTa}_{\text{base}}$	**76.53**	**85.21**	**80.95**	**86.03**	**82.57**	**85.83**	**80.50**	**82.52**
∗ SimCSE-$\text{RoBERTa}_{\text{large}}$	**77.46**	**87.27**	**82.36**	**86.66**	**83.93**	**86.70**	**81.95**	**83.76**

Questions?