CSCE 689: Special Topics in Trustworthy NLP

Lecture 7: Natural Language Processing Basics (6)

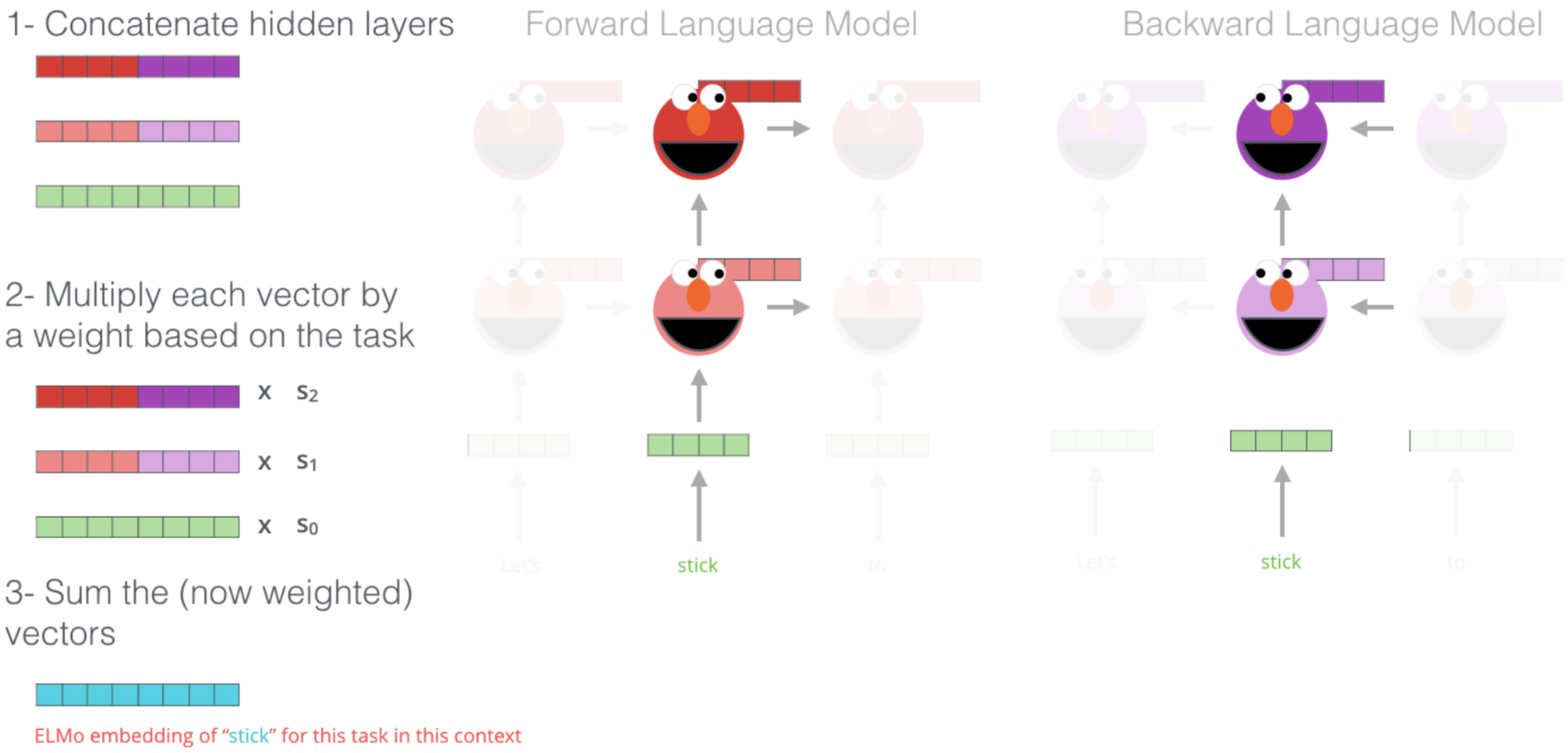
Kuan-Hao Huang
khhuang@tamu.edu

(Some slides adapted from Chris Manning, Karthik Narasimhan)

Lecture Plan

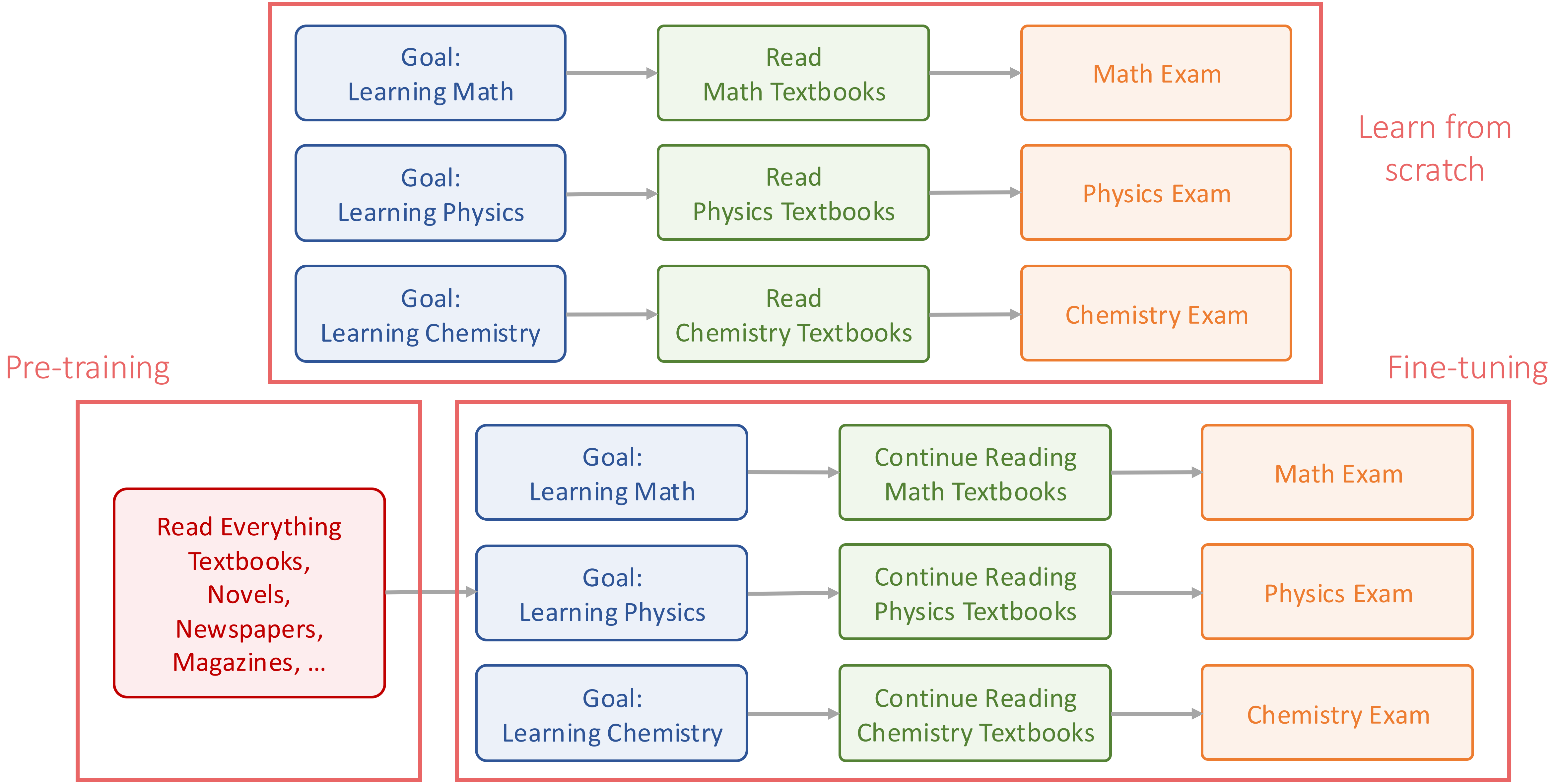
- Natural Language Processing Basics
- Pre-Training
- Language Models

Recap: Contextualized Word Embeddings

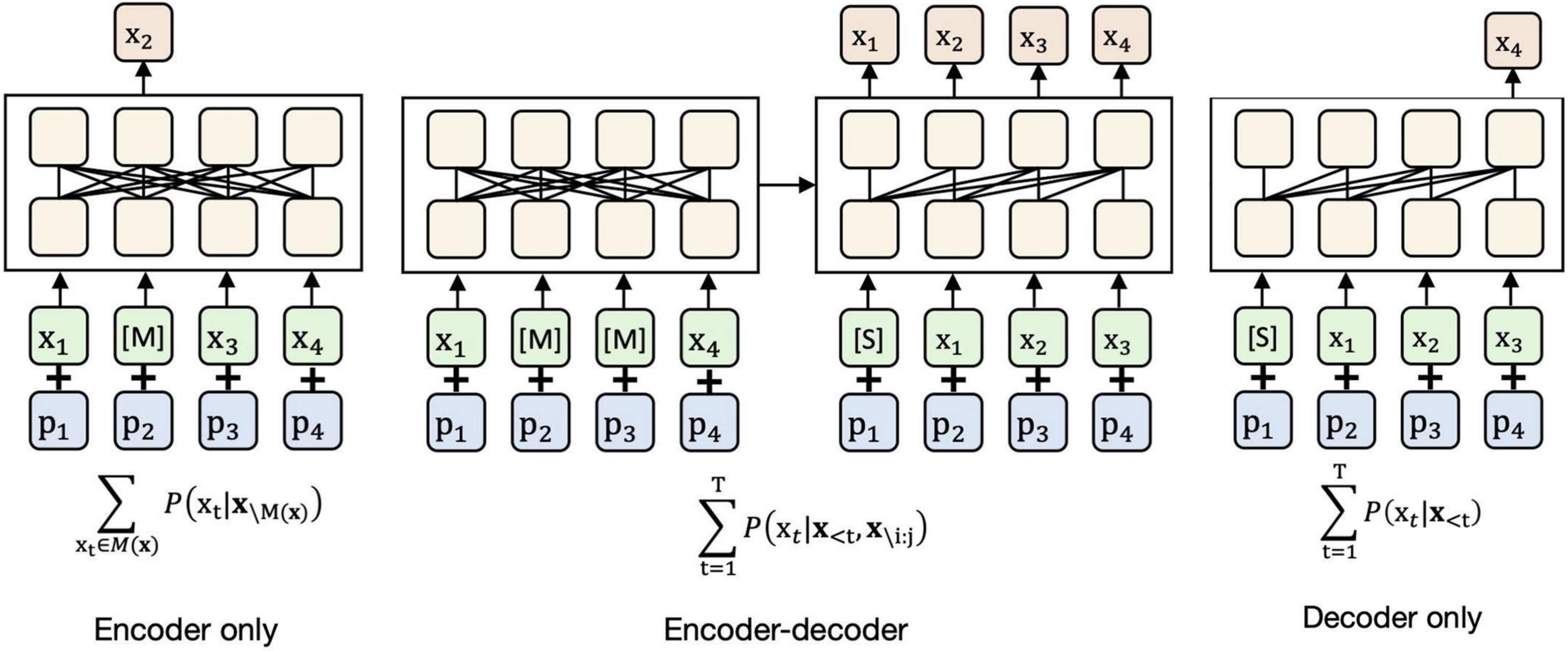


Source: http://jalammar.github.io/illustrated-bert/

Recap: Pre-Training



Three Types of Pre-Training



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

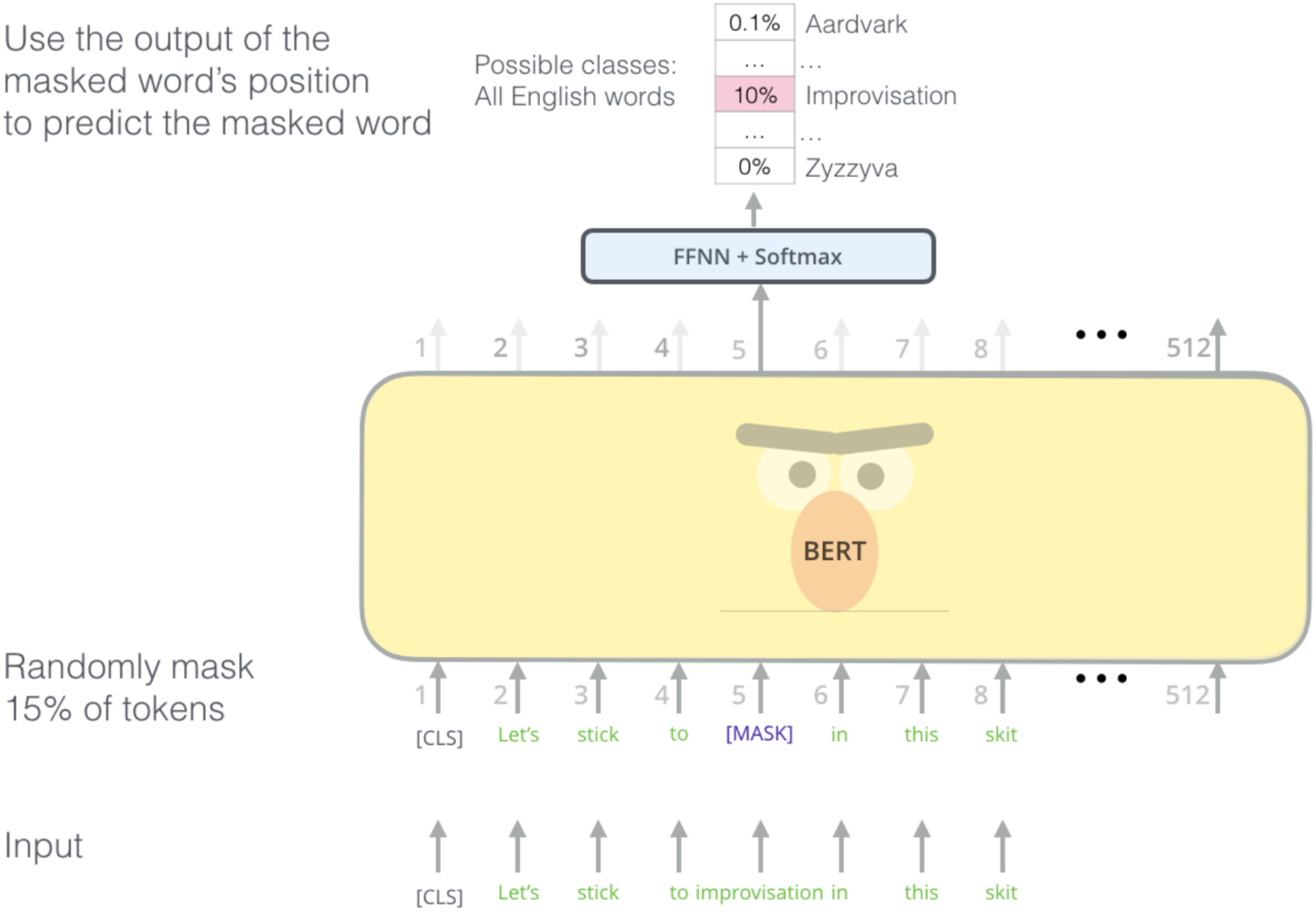
Encoder-Only: BERT

- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL 2019
  - **B**idirectional **E**ncoder **R**epresentations from **T**ransformers (BERT)
  - Learn general knowledge with a large corpus
  - Re-use model weights for fine-tuning

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

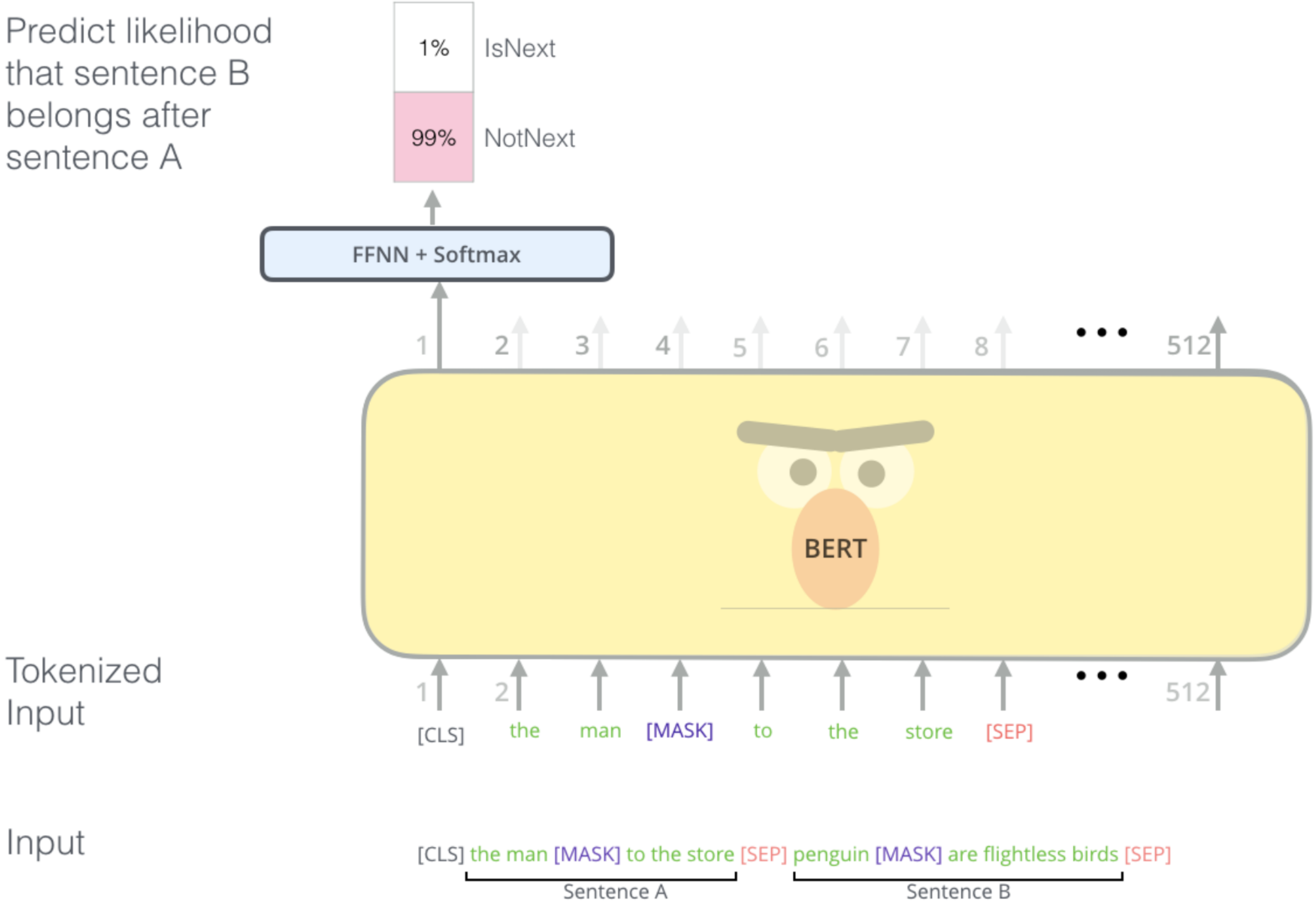
Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Pre-Training Task: Masked Language Modeling



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

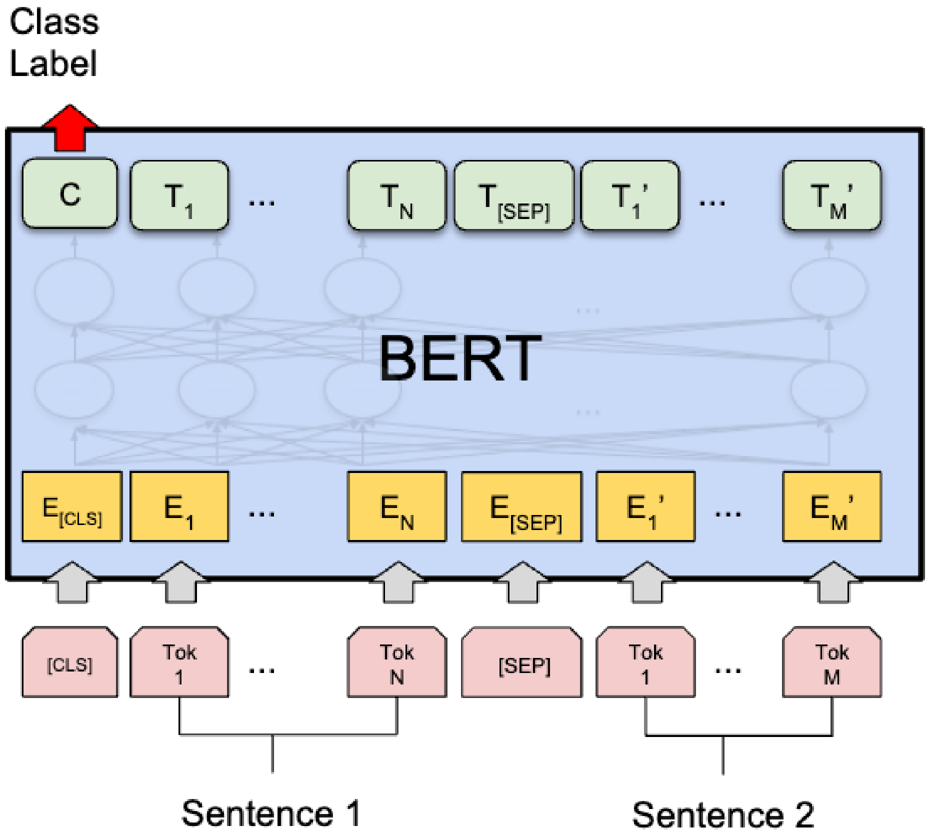
Pre-Training Task: Next Sentence Prediction

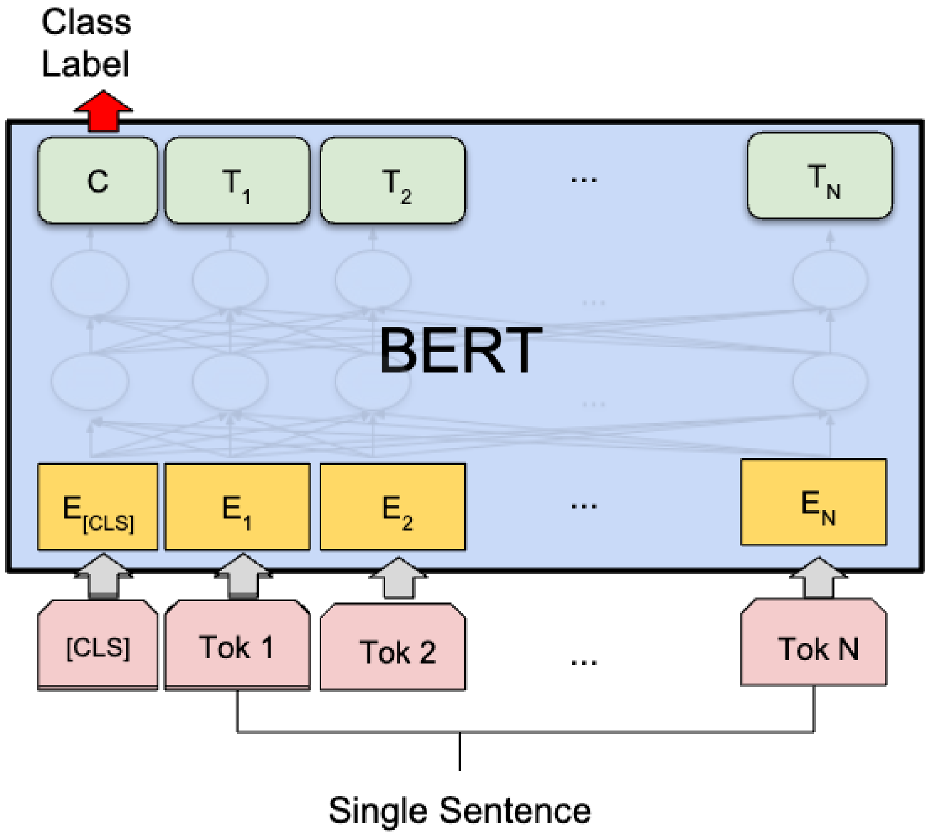


Source: http://jalammar.github.io/illustrated-bert/

Fine-Tuning: Sentence-Level Tasks

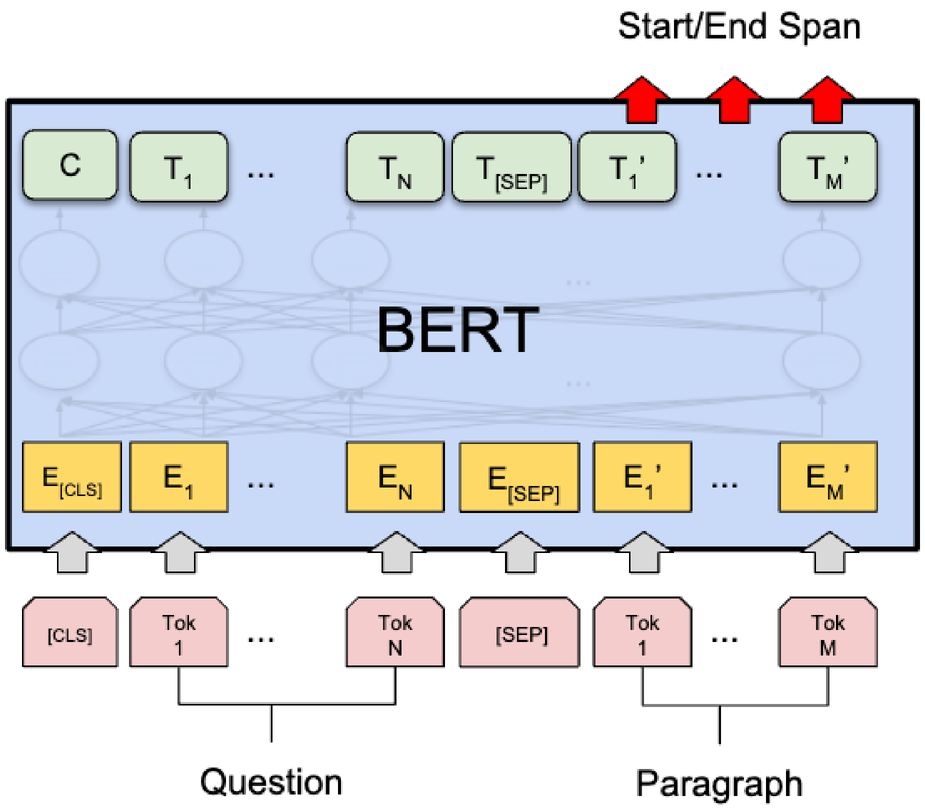
- Pre-training provides a good weight initialization

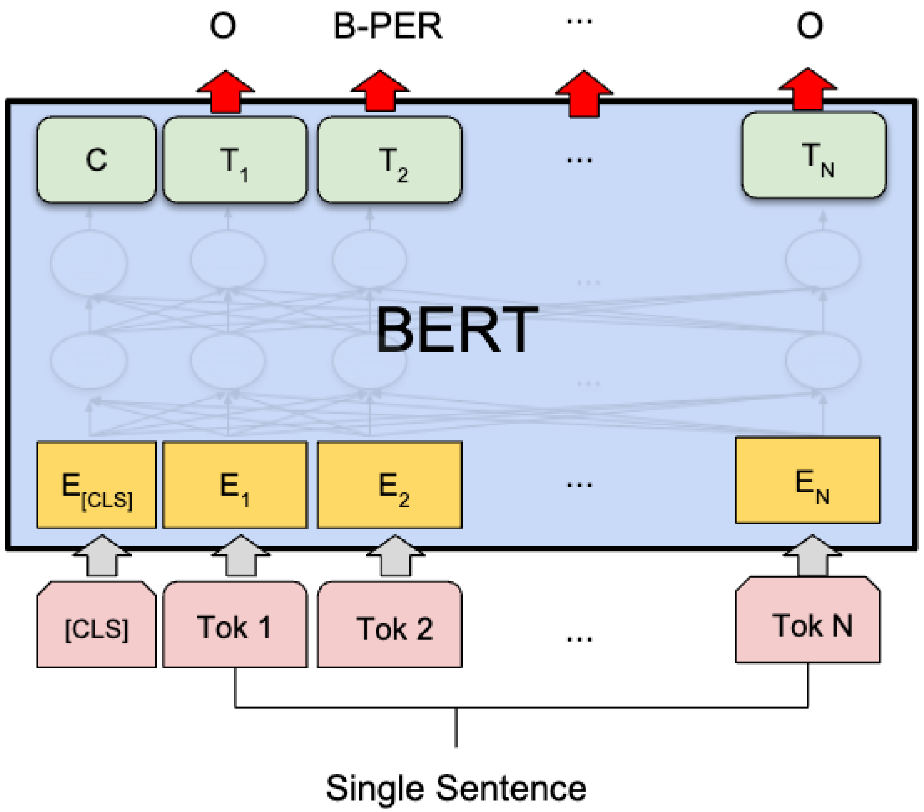




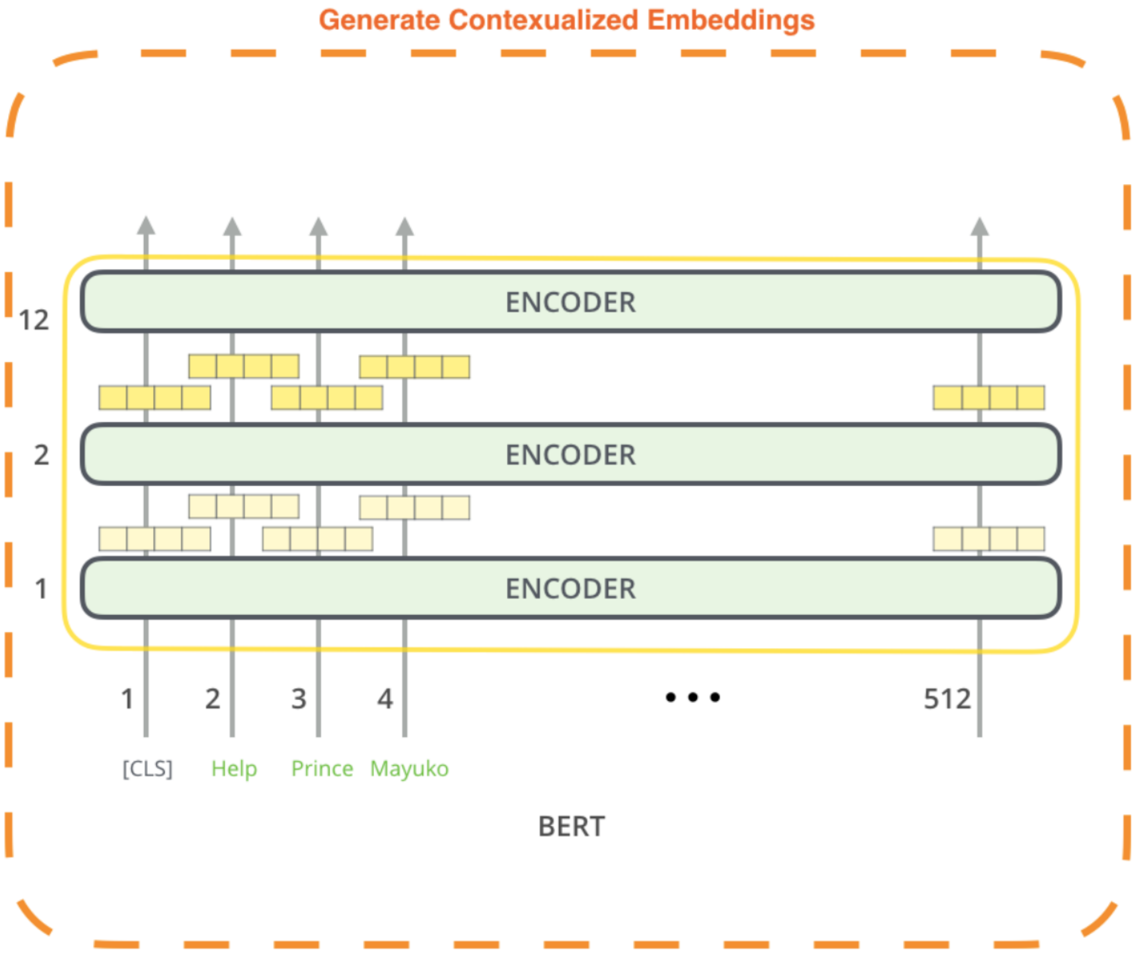
Fine-Tuning: Token-Level Tasks

- Pre-training provides a good weight initialization

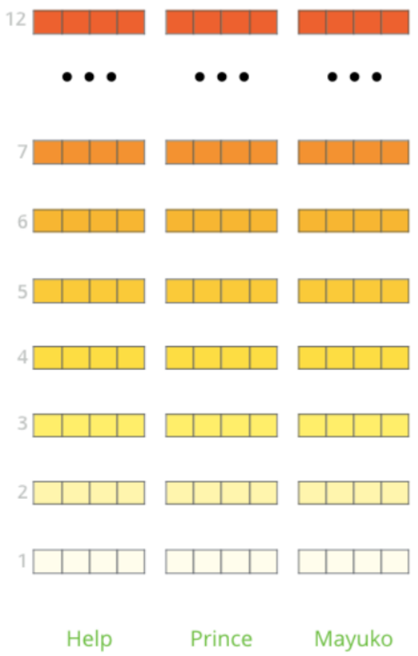




BERT as General Contextualized Representations



The output of each encoder layer along each token's path can be used as a feature representing that token.



But which one should we use?

Amazing Performance

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
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

Encoder-Only: RoBERTa

- RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv 2019
  - **R**obustly **o**ptimized **BERT a**pproach (RoBERTa)
  - BERT is still under-trained
  - Improve the robustness of training BERT

RoBERTa: A Robustly Optimized BERT Pretraining Approach

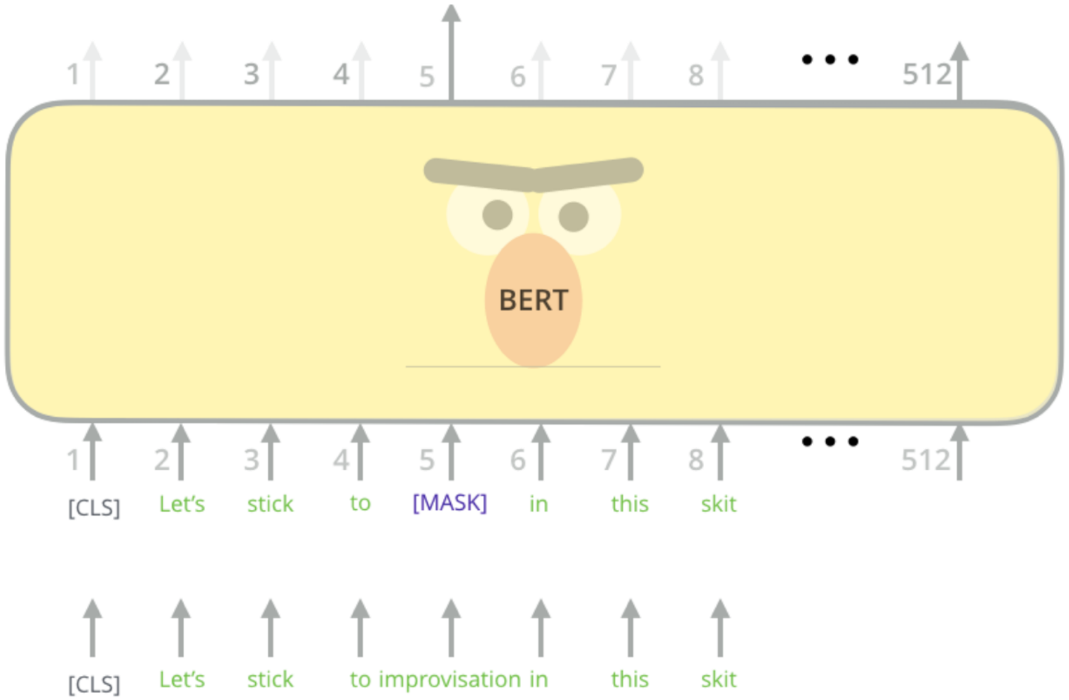
Yinhan Liu[*§] **Myle Ott**[*§] **Naman Goyal**[*§] **Jingfei Du**[*§] **Mandar Joshi**[†]
Danqi Chen[§] **Omer Levy**[§] **Mike Lewis**[§] **Luke Zettlemoyer**[†§] **Veselin Stoyanov**[§]

[†] Paul G. Allen School of Computer Science & Engineering,
University of Washington, Seattle, WA
{mandar90,lsz}@cs.washington.edu

[§] Facebook AI
{yinhanliu,myleott,naman,jingfeidu,
danqi,omerlevy,mikelewis,lsz,ves}@fb.com

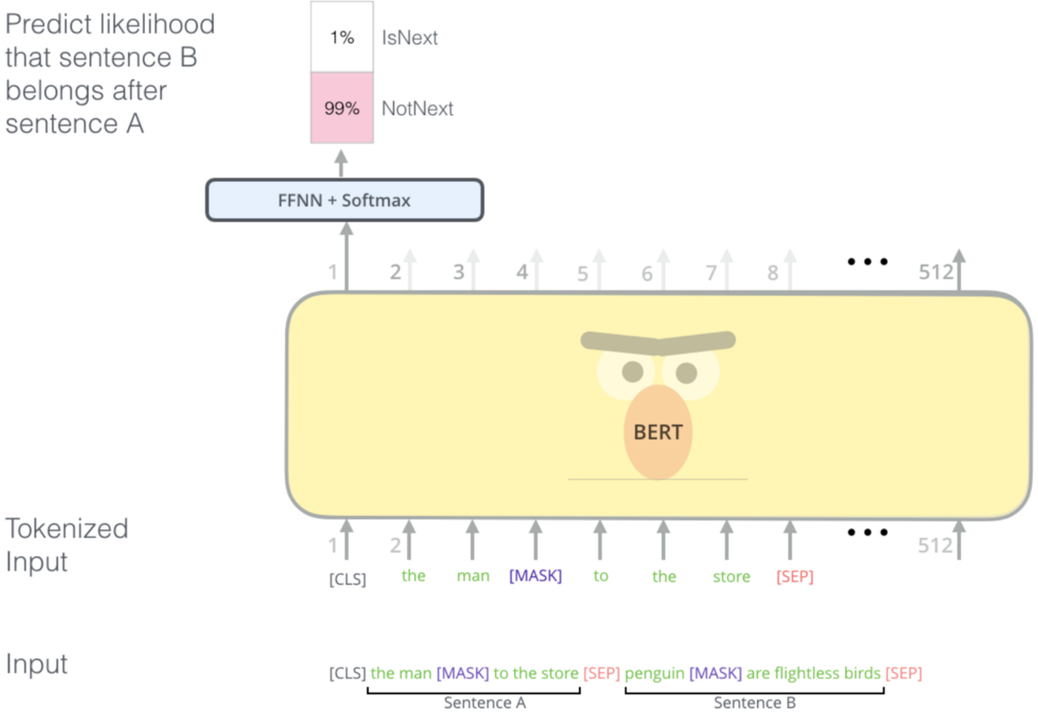
Static Masking vs. Dynamic Masking

- Static masking: decide masked words during data pre-processing
- Dynamic masking: decide masked words right before feeding into models



Masking	SQuAD 2.0	MNLI-m	SST-2
static	78.3	84.3	92.5
dynamic	78.7	84.0	92.9

Removing Next Sentence Prediction Task



Model	SQuAD 1.1/2.0	MNLI-m	SST-2	RACE
Our reimplementation (with NSP loss):				
SEGMENT-PAIR	90.4/78.7	84.0	92.9	64.2
SENTENCE-PAIR	88.7/76.2	82.9	92.1	63.0
Our reimplementation (without NSP loss):				
FULL-SENTENCES	90.4/79.1	84.7	92.5	64.8
DOC-SENTENCES	90.6/79.7	84.7	92.7	65.6

True Byte-Pair Encoding (BPE)

- BERT: BPE with unicode characters
  - Vocabulary size: 30K
- RoBERTa: BPE with bytes
  - Vocabulary size: 50K

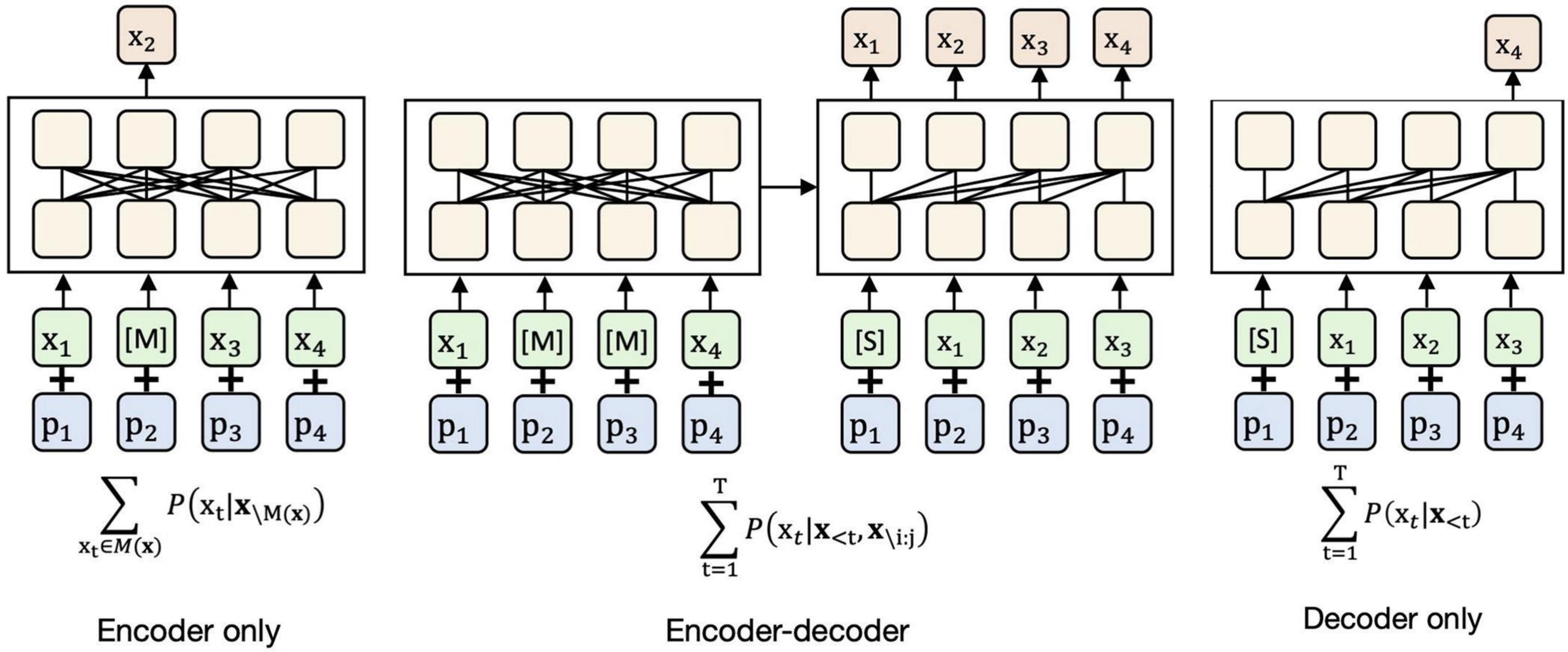
Training Details

- Trained longer
- 10x data
- Bigger batch sizes

Much Better Performance Than BERT

Model	data	bsz	steps	SQuAD (v1.1/2.0)	MNLI-m	SST-2
RoBERTa						
with BOOKS + WIKI	16GB	8K	100K	93.6/87.3	89.0	95.3
+ additional data (§3.2)	160GB	8K	100K	94.0/87.7	89.3	95.6
+ pretrain longer	160GB	8K	300K	94.4/88.7	90.0	96.1
+ pretrain even longer	160GB	8K	500K	**94.6/89.4**	**90.2**	**96.4**
$\text{BERT}_{\text{LARGE}}$						
with BOOKS + WIKI	13GB	256	1M	90.9/81.8	86.6	93.7

Three Types of Pre-Training



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

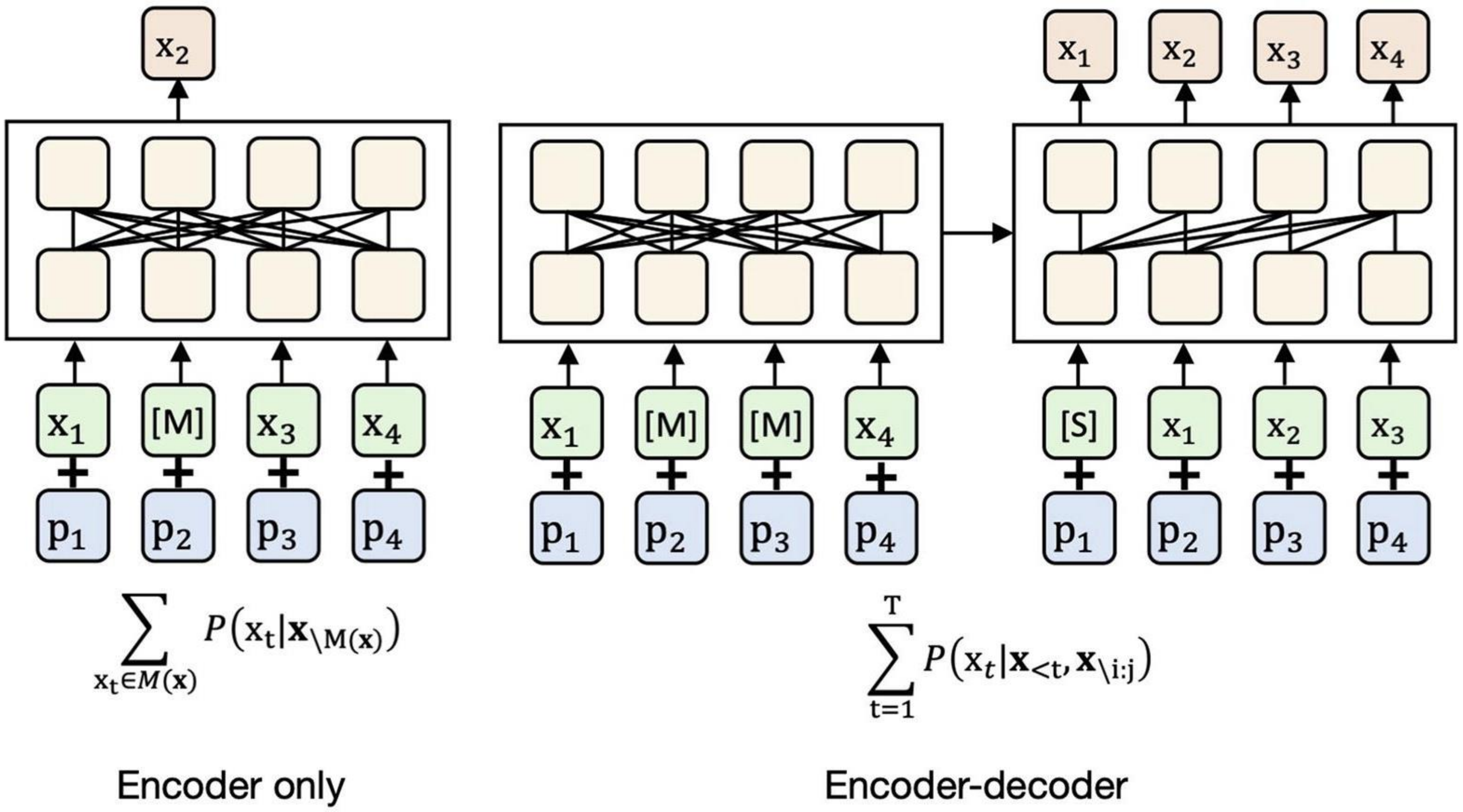
Encoder-Decoder: BART

- BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, ACL 2020
  - **B**idirectional and **A**uto-**R**egressive **T**ransformers (BART)
  - Pre-training for generation tasks

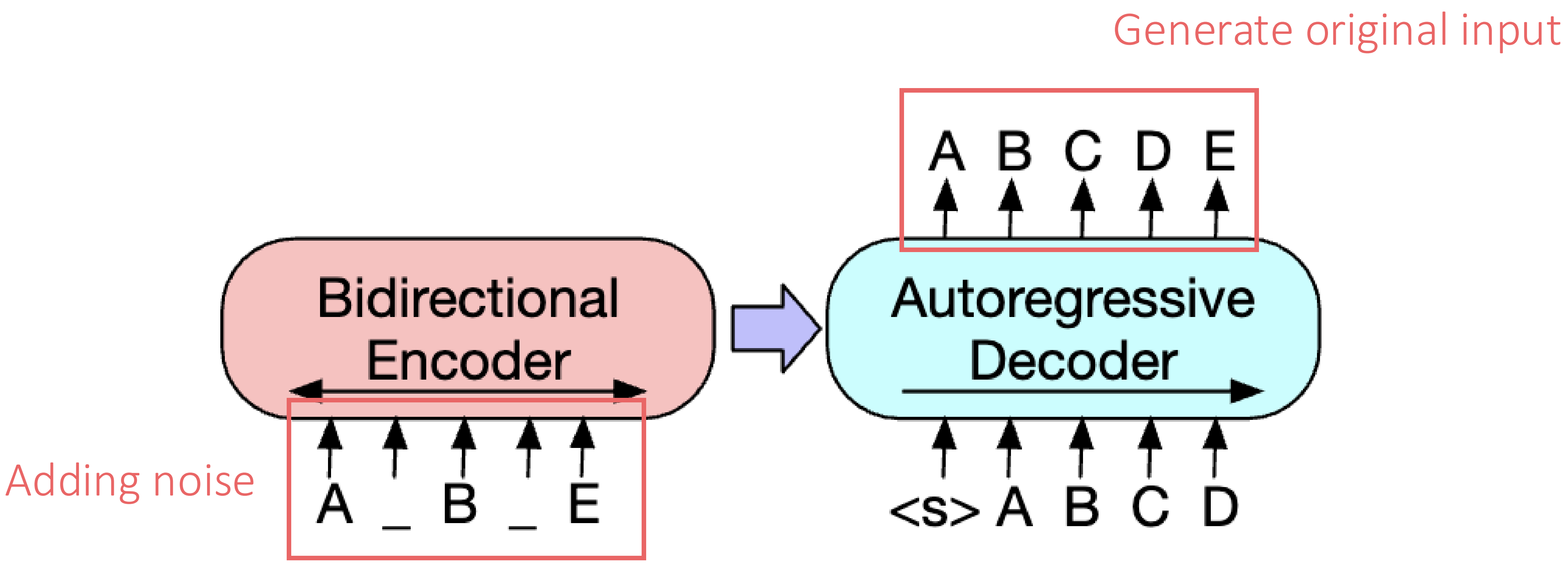
BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

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

Encoder-Only vs. Encoder-Decoder



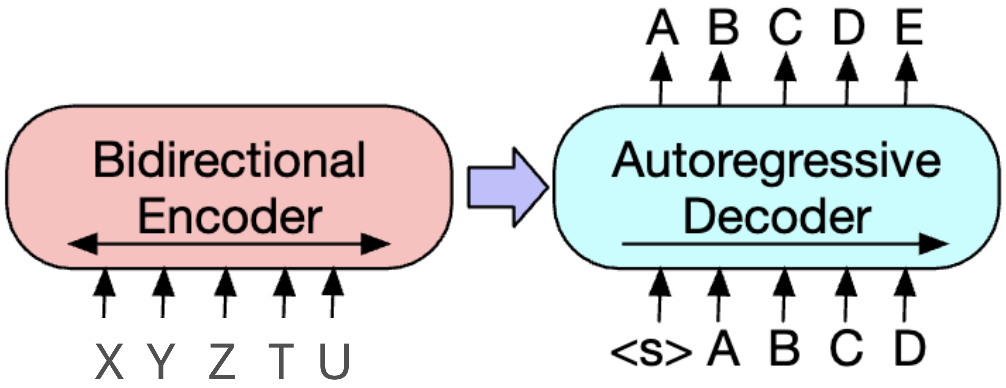
Denoising Autoencoder



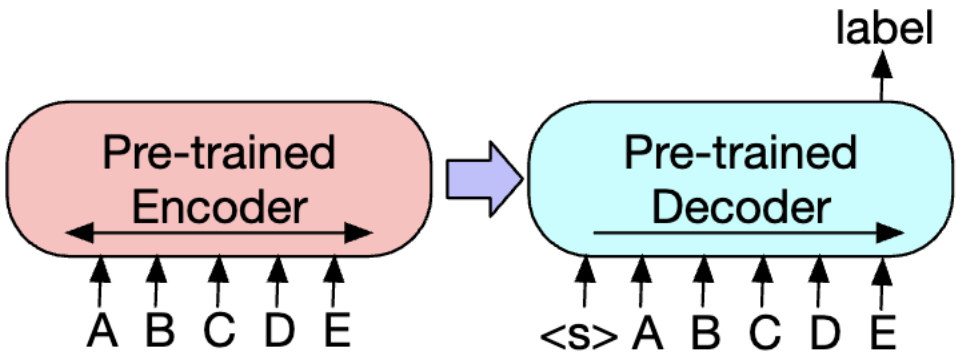
Denoising Objective

- 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.

Fine-Tuning



Sequence-to-Sequence



Classification

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

	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**

	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**

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

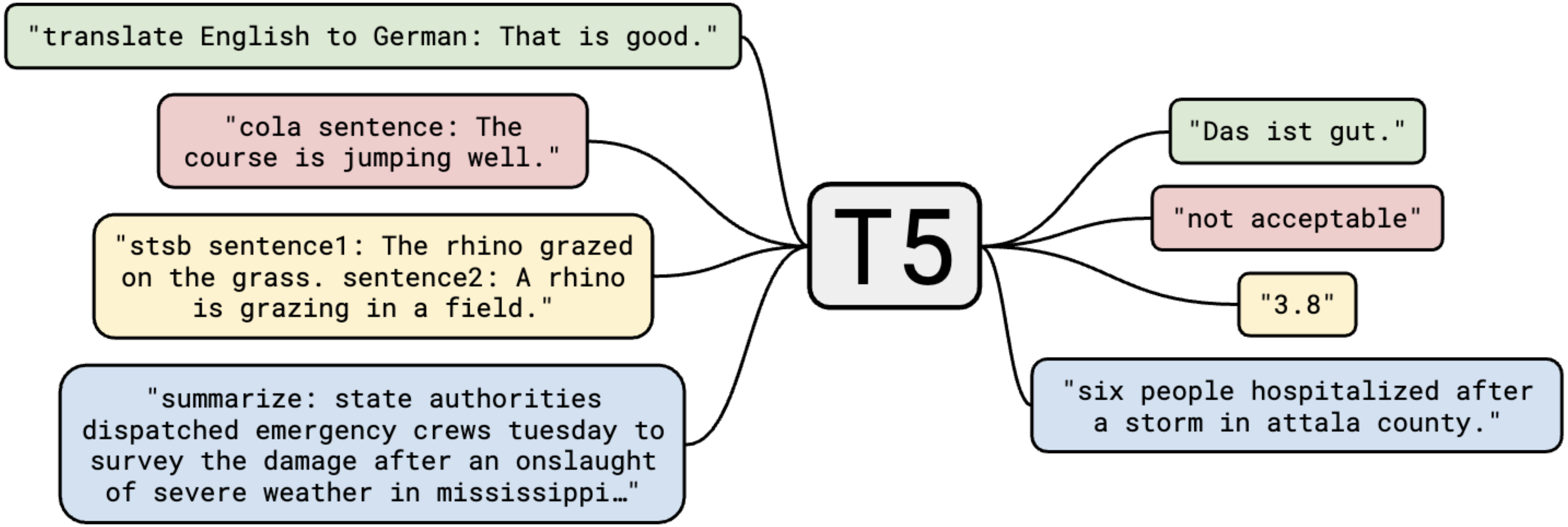
Encoder-Decoder: T5

- Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, JMLR 2020
  - Text-to-Text Transfer Transformer (T5)

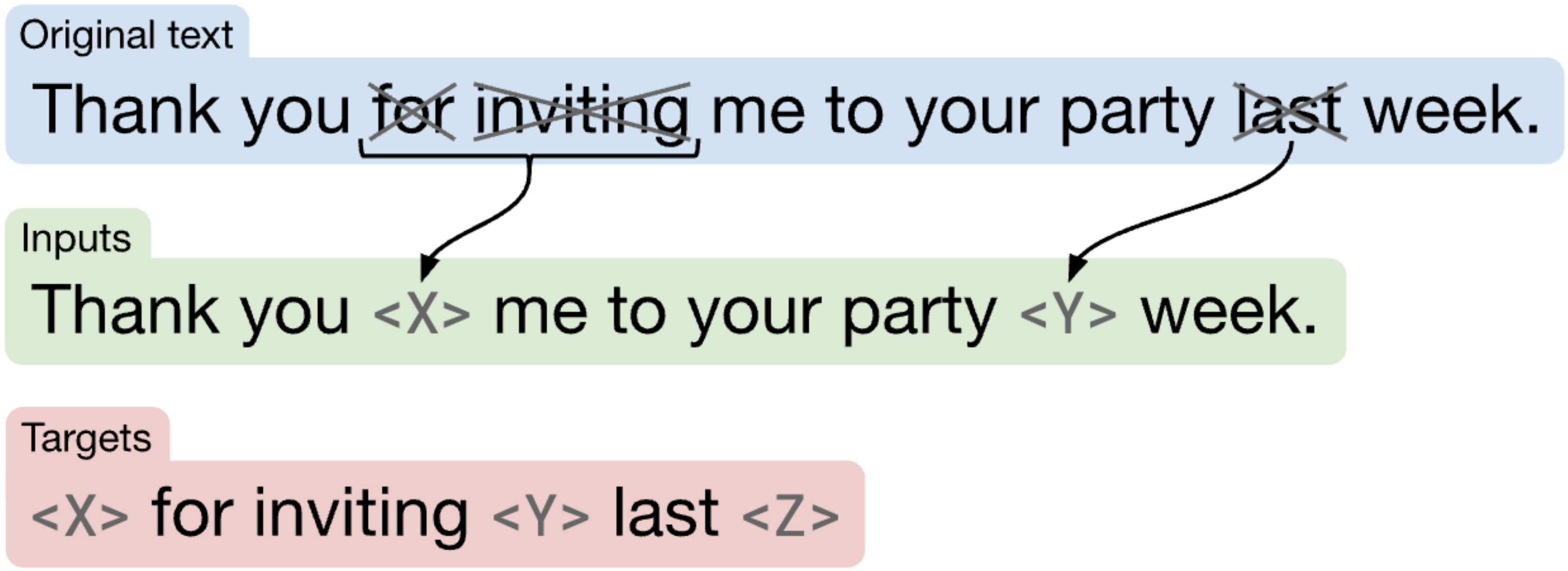
Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

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

Convert Everything to Text-to-Text Tasks



Unsupervised Objective



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.`

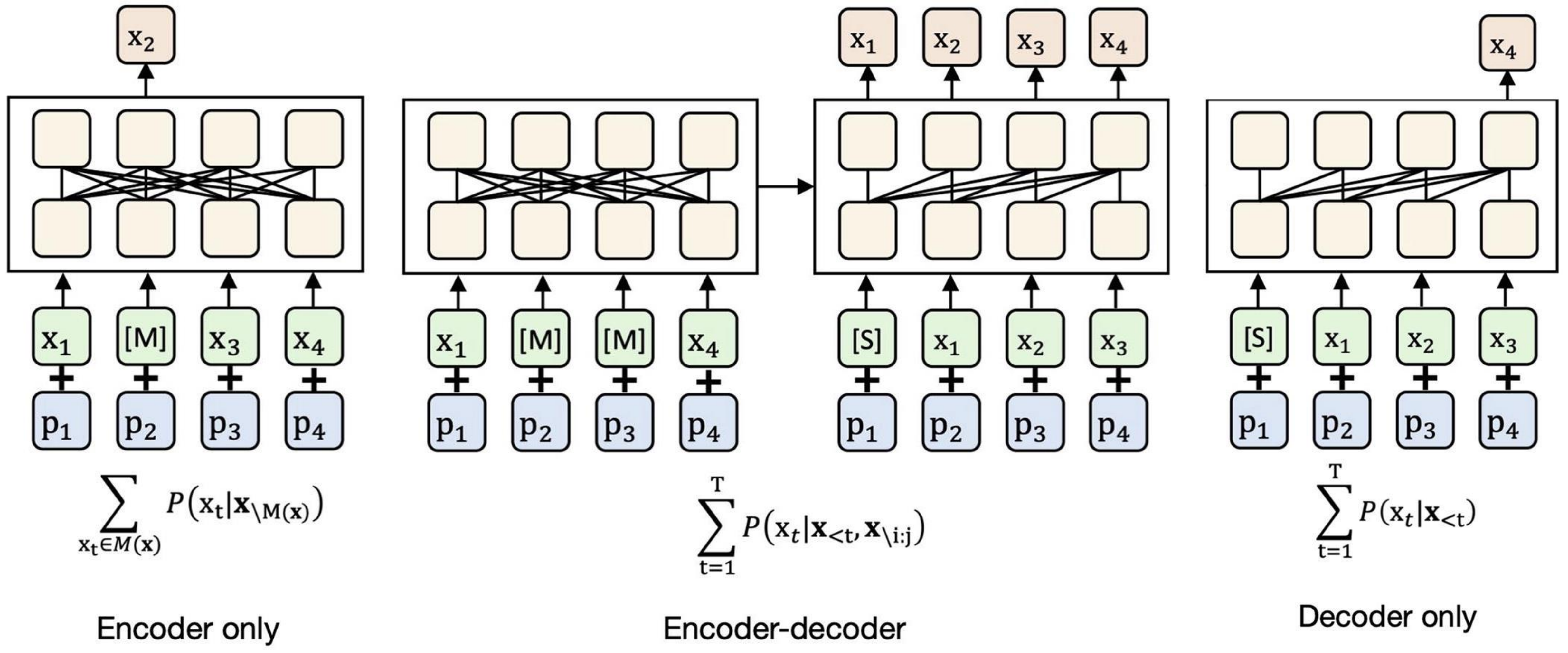
Promising Results

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**

Three Types of Pre-Training



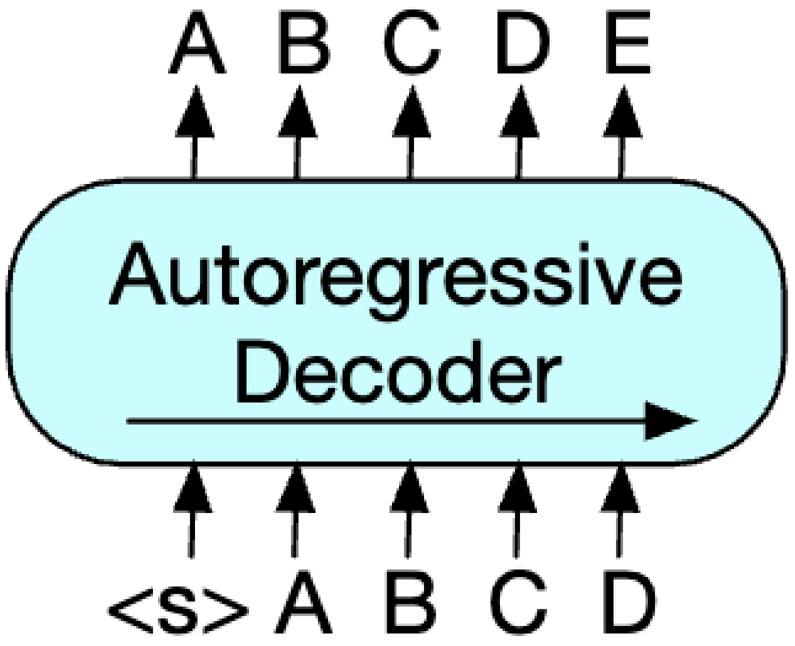
Source: https://www.sciencedirect.com/science/article/pii/S2095809922006324

Decoder-Only: 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

Language Modeling

- Next word prediction
- Trained with large corpus

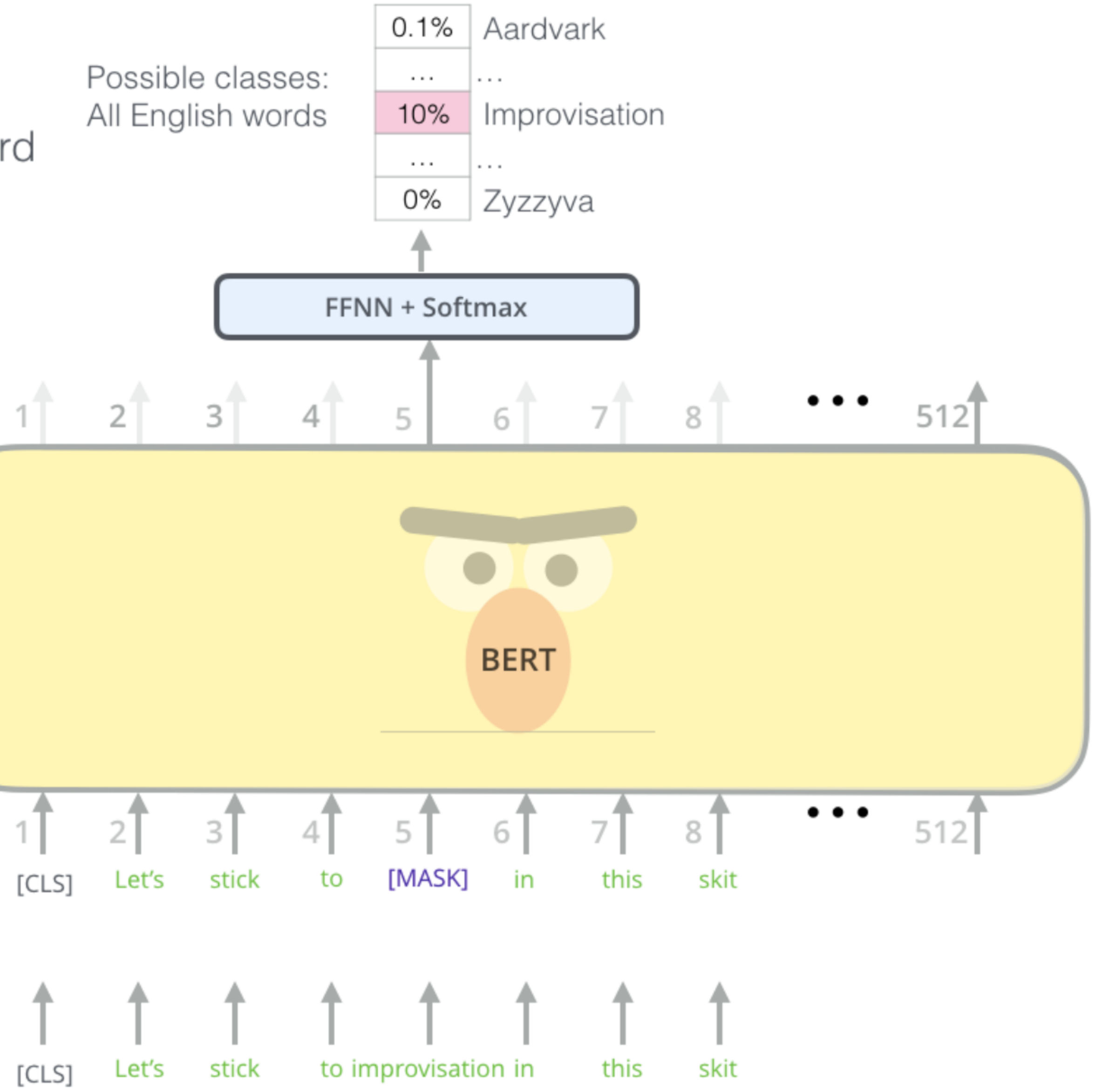


Comparison: Masked Language Models

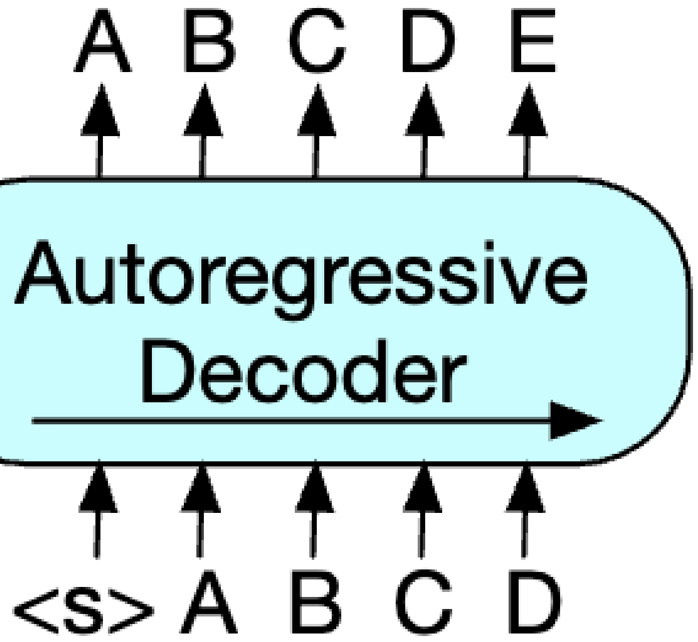
Use the output of the
masked word's position
to predict the masked word

Randomly mask
15% of tokens

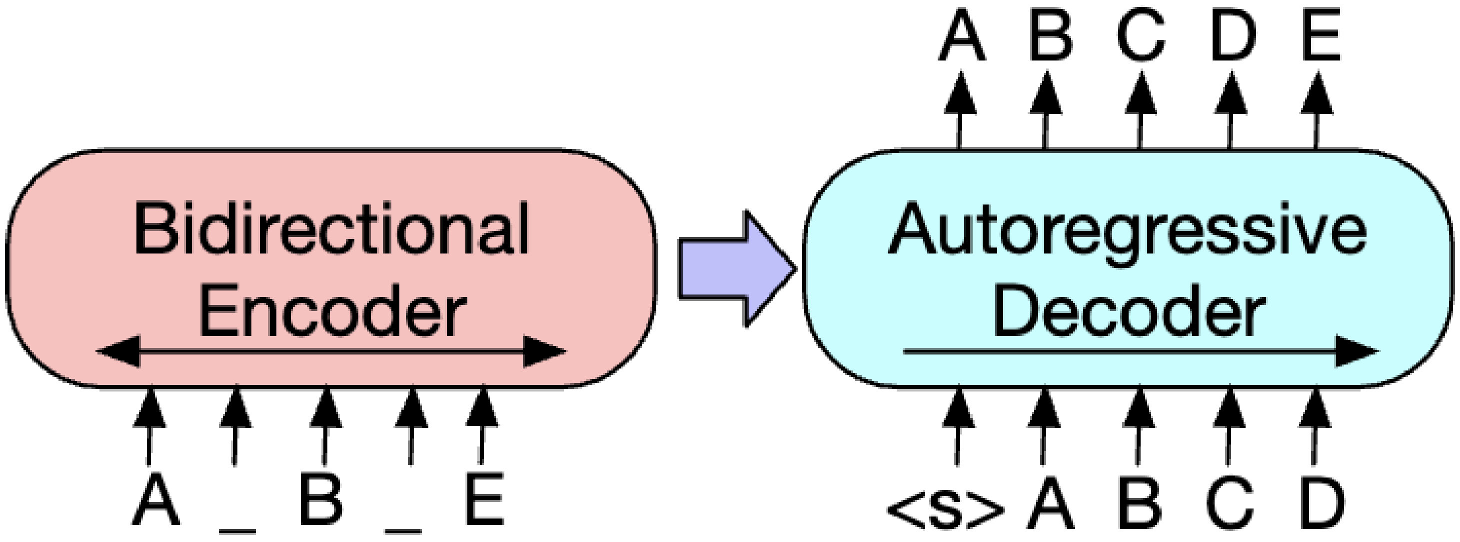
Input



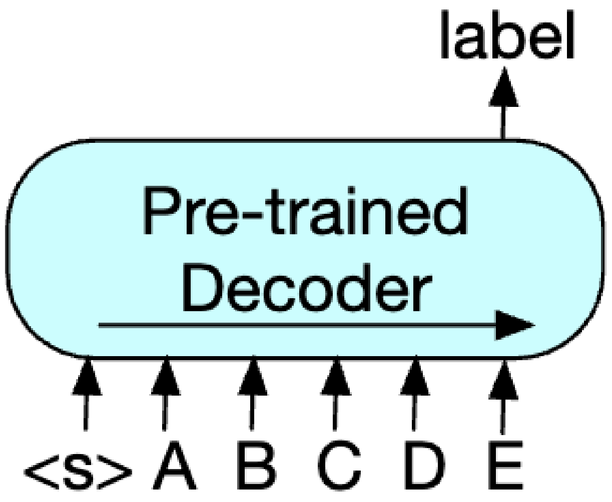
Comparison: Causal Language Models

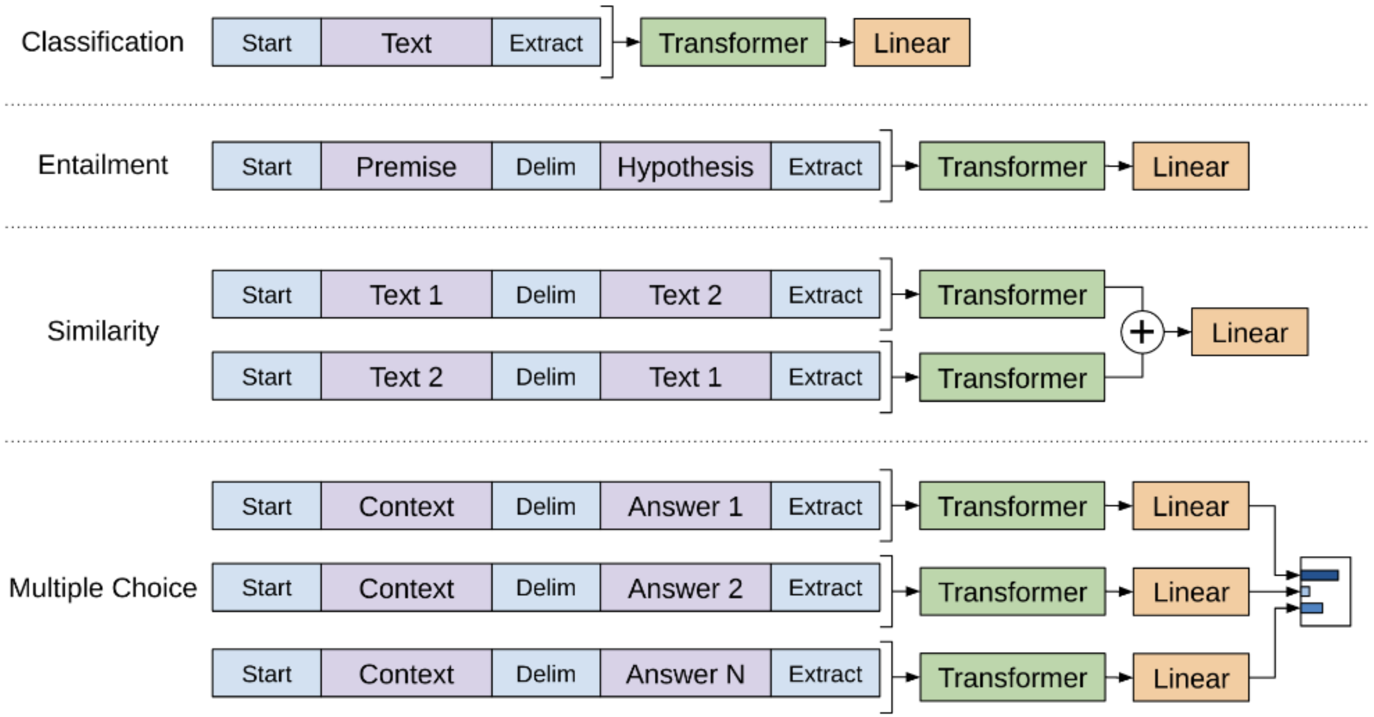


Comparison: Seq2Seq Models



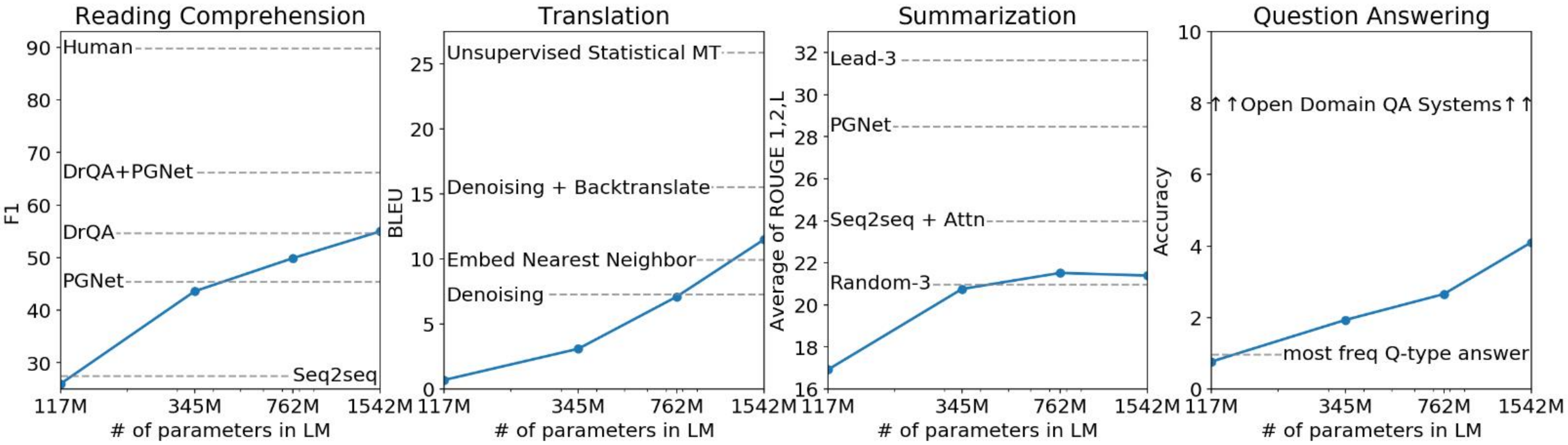
GPT-1: Good Contextualized Representations





GPT-2: Unsupervised Pre-Training Helps Supervised Tasks

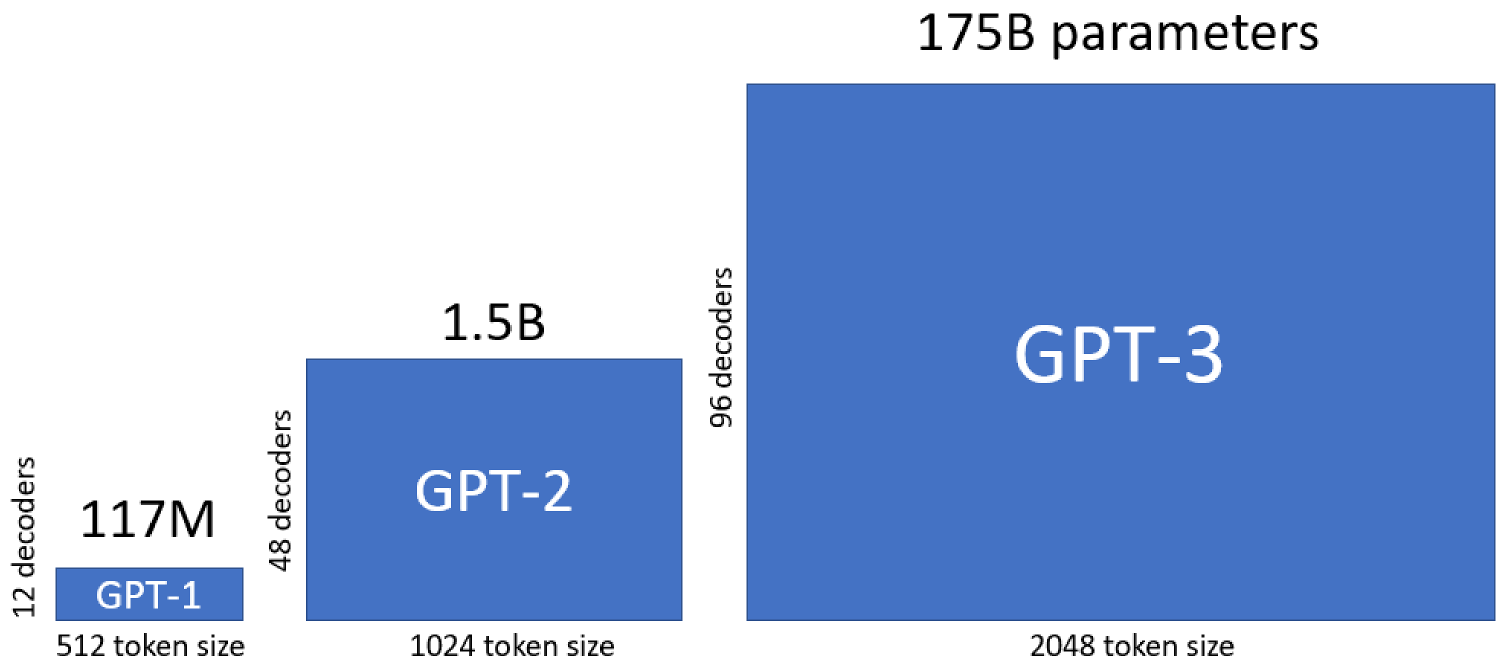
- Larger training data, larger model size



Demonstrate zero-shot ability on certain tasks

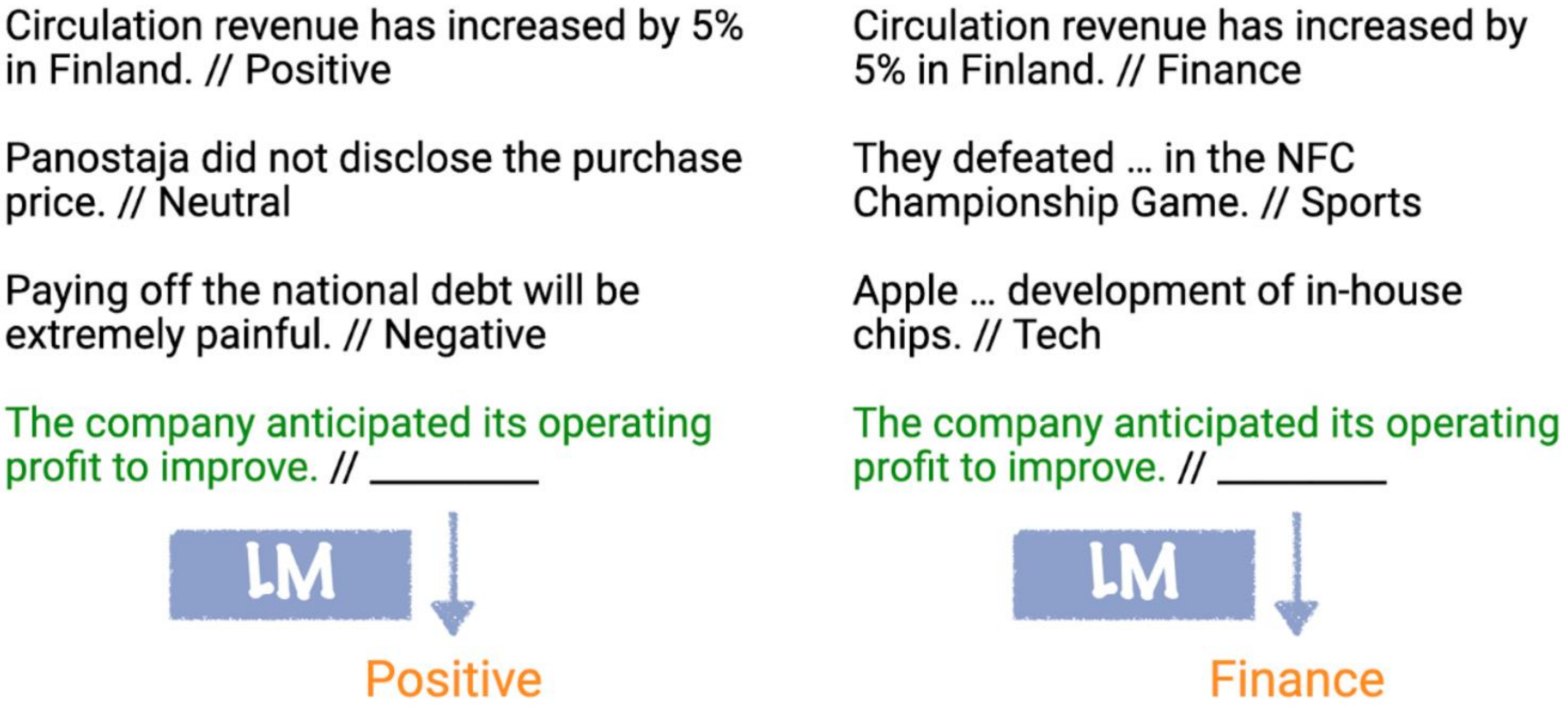
GPT-3: From Fine-Tuning to Few-Shot Learning

- Even larger training data, even larger model size

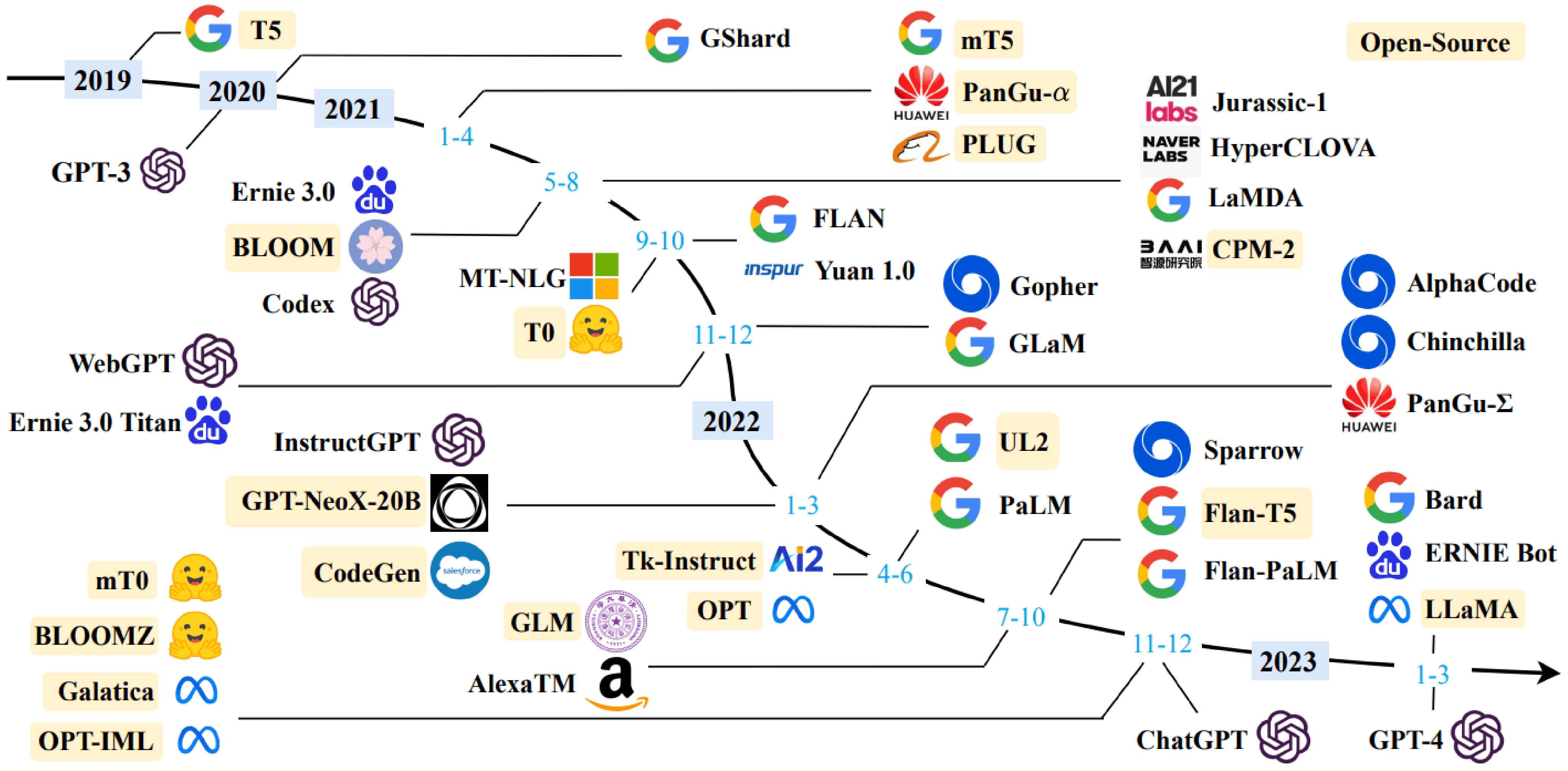


GPT-3: From Fine-Tuning to Few-Shot Learning

- Solve entirely new tasks by few-shot learning (in-context learning)



Large Language Models



Next Lecture

- Natural Language Processing Basics
- Large Language Models
- Prompting
- In-Context Learning