Day 7 - Contextual Embeddings, Pretrained Models, and Transfer Learning

Advanced Text as Data: Natural Language Processing
Essex Summer School in Social Science Data Analysis

Burt L. Monroe (Instructor) & Sam Bestvater (TA)
Pennsylvania State University

August 4, 2021

Transfer Learning

A typical use case

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

Semi-supervised Learning Step

Model:

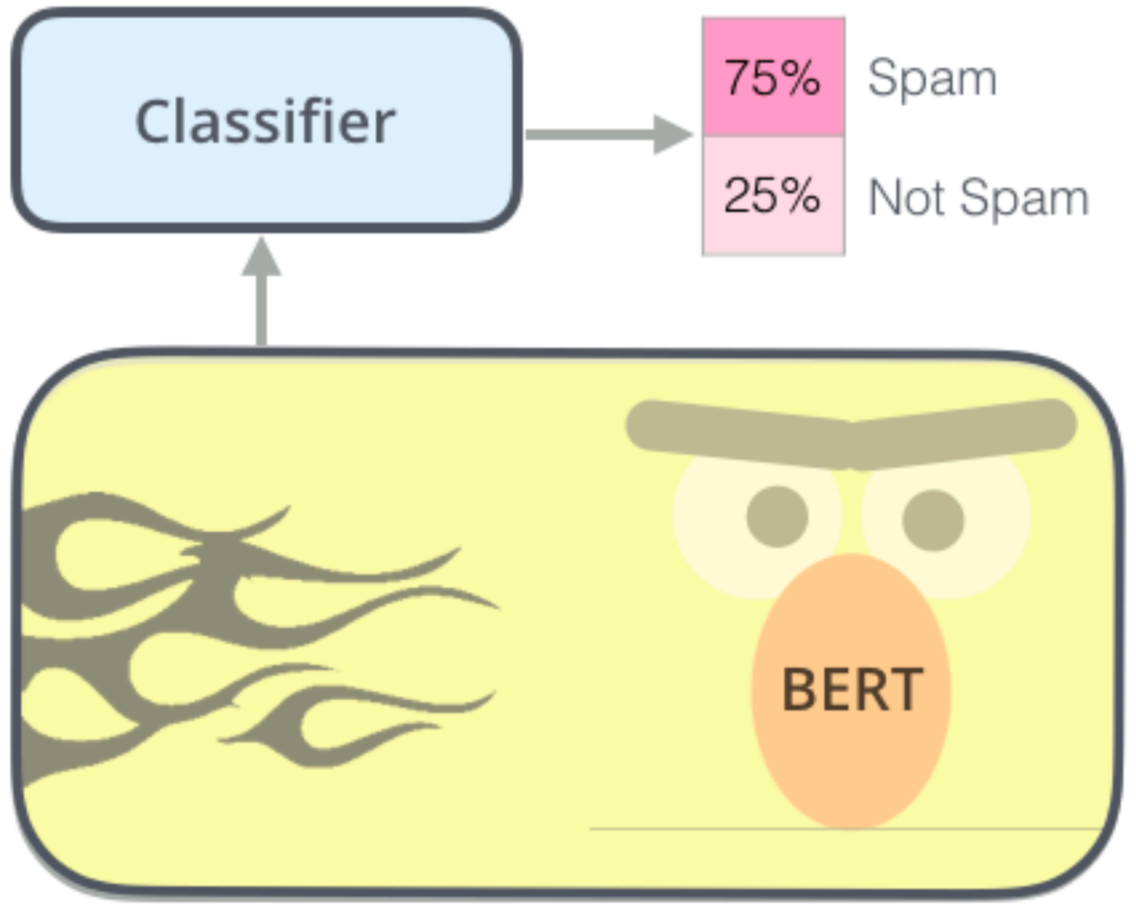
Dataset:

WIKIPEDIA
Die freie Enzyklopädie

Objective: Predict the masked word (langauge modeling)

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

Supervised Learning Step

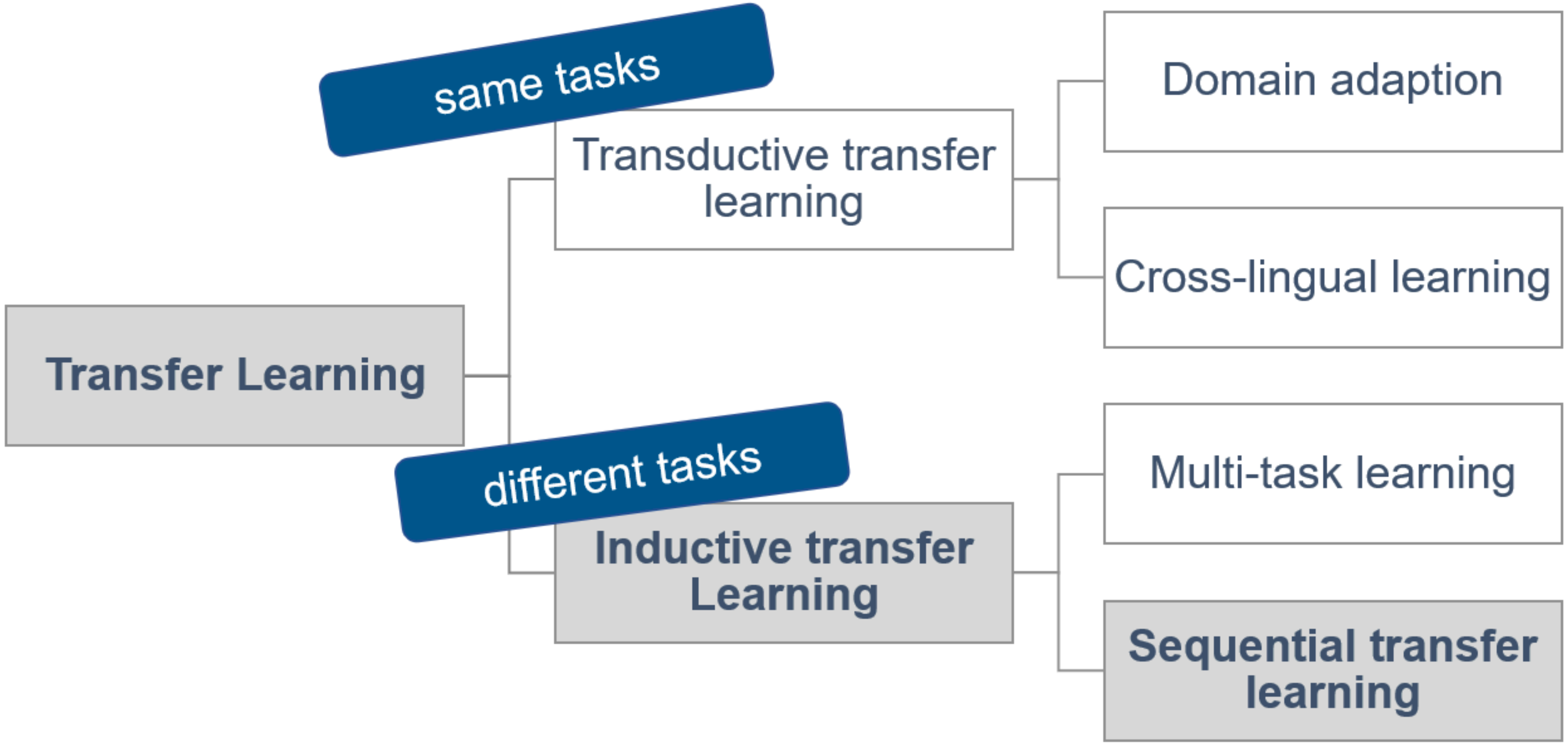


Model: (pre-trained in step #1)

Dataset:

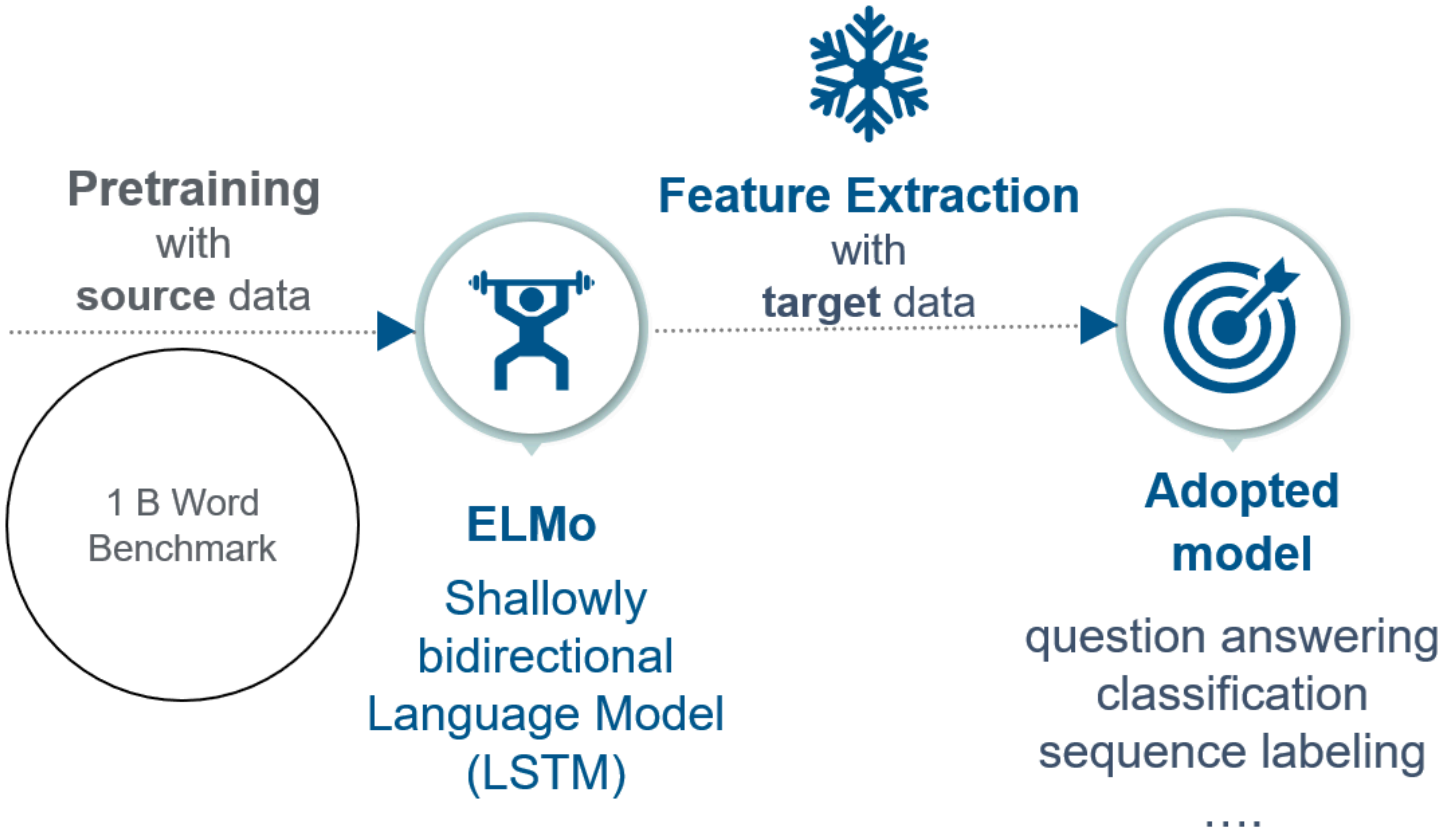
Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached…	Not Spam

Transfer Learning



Source: Becker, et al. 2020, *Modern Approaches to Natural Language Processi*

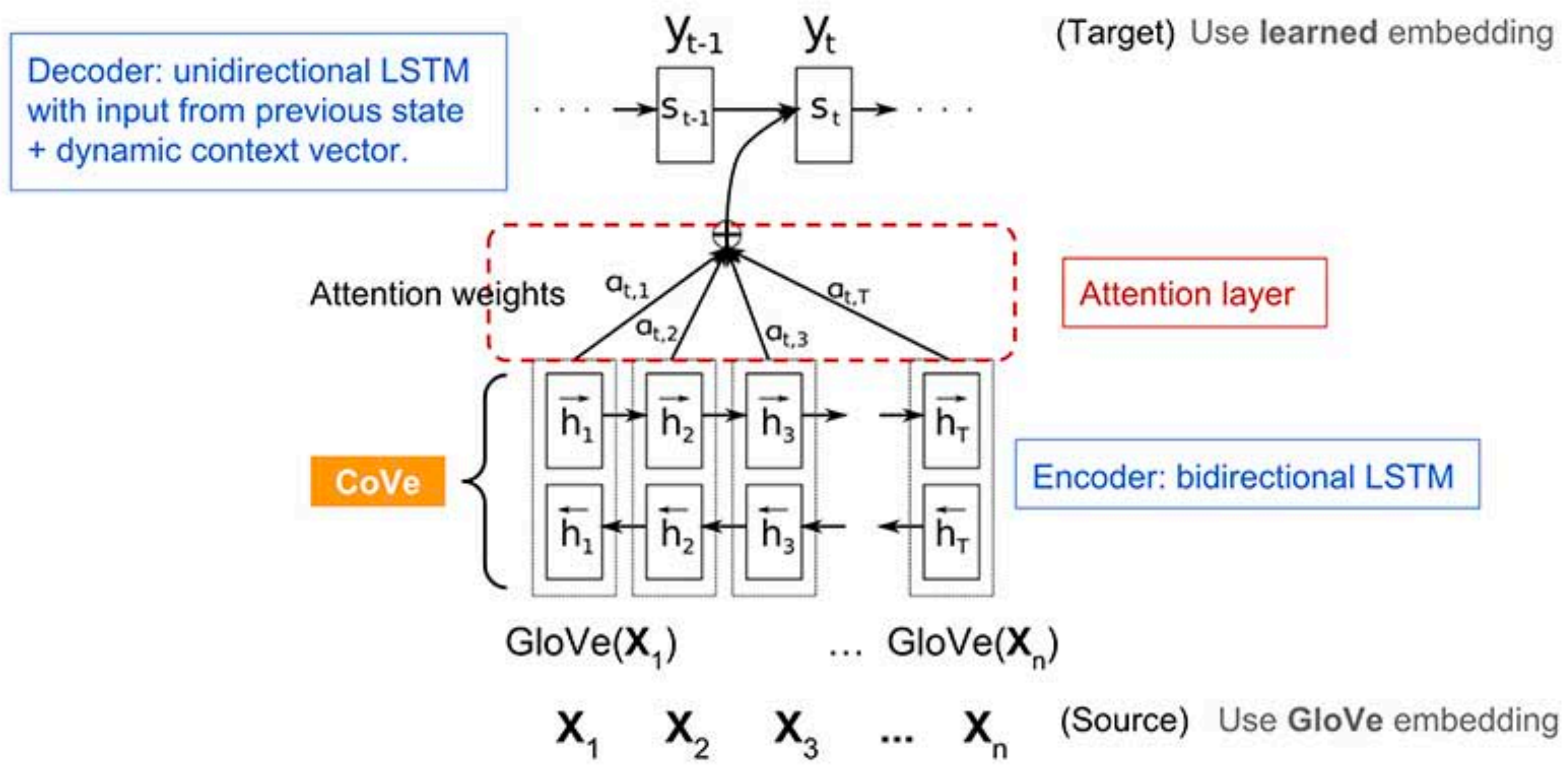
Feature Extraction / Contextualized Word Embeddings

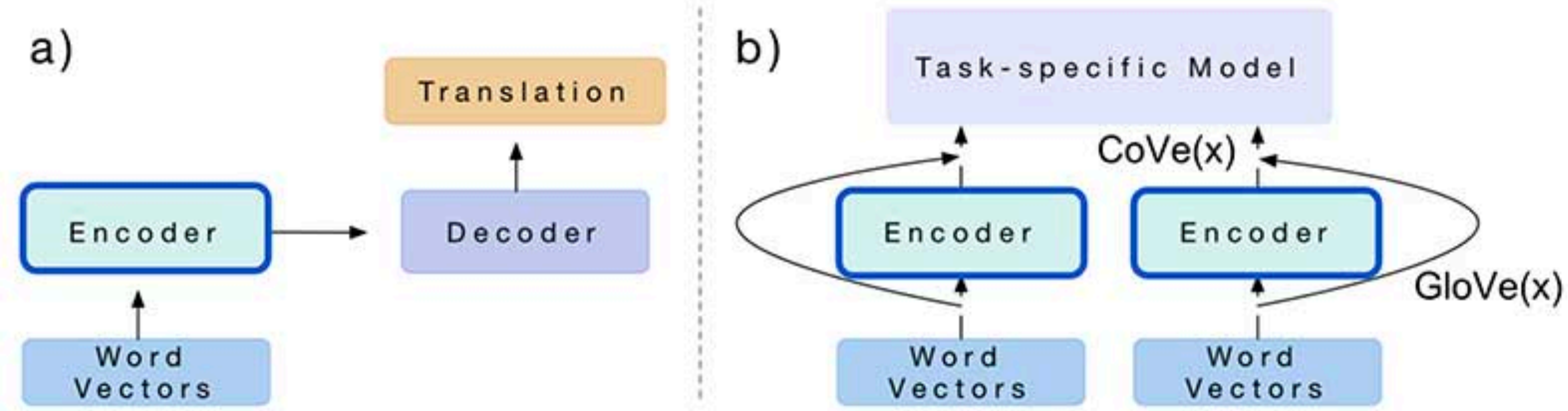


Source: Becker, et al. 2020, *Modern Approaches to Natural Language Processi*

CoVe (SalesForce, 2017)

Pretrained on machine translation.
Keep the encoder and reuse for other task.

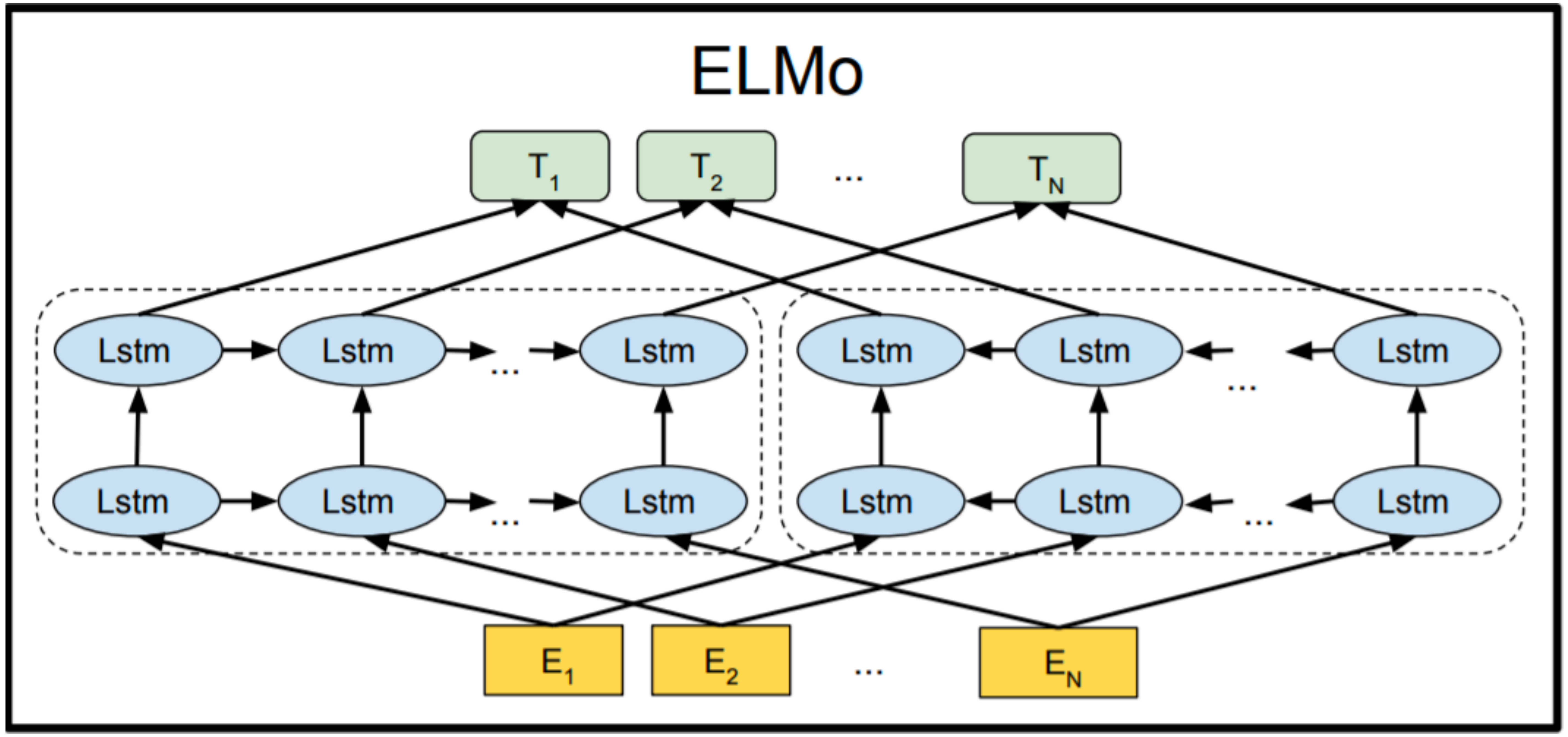




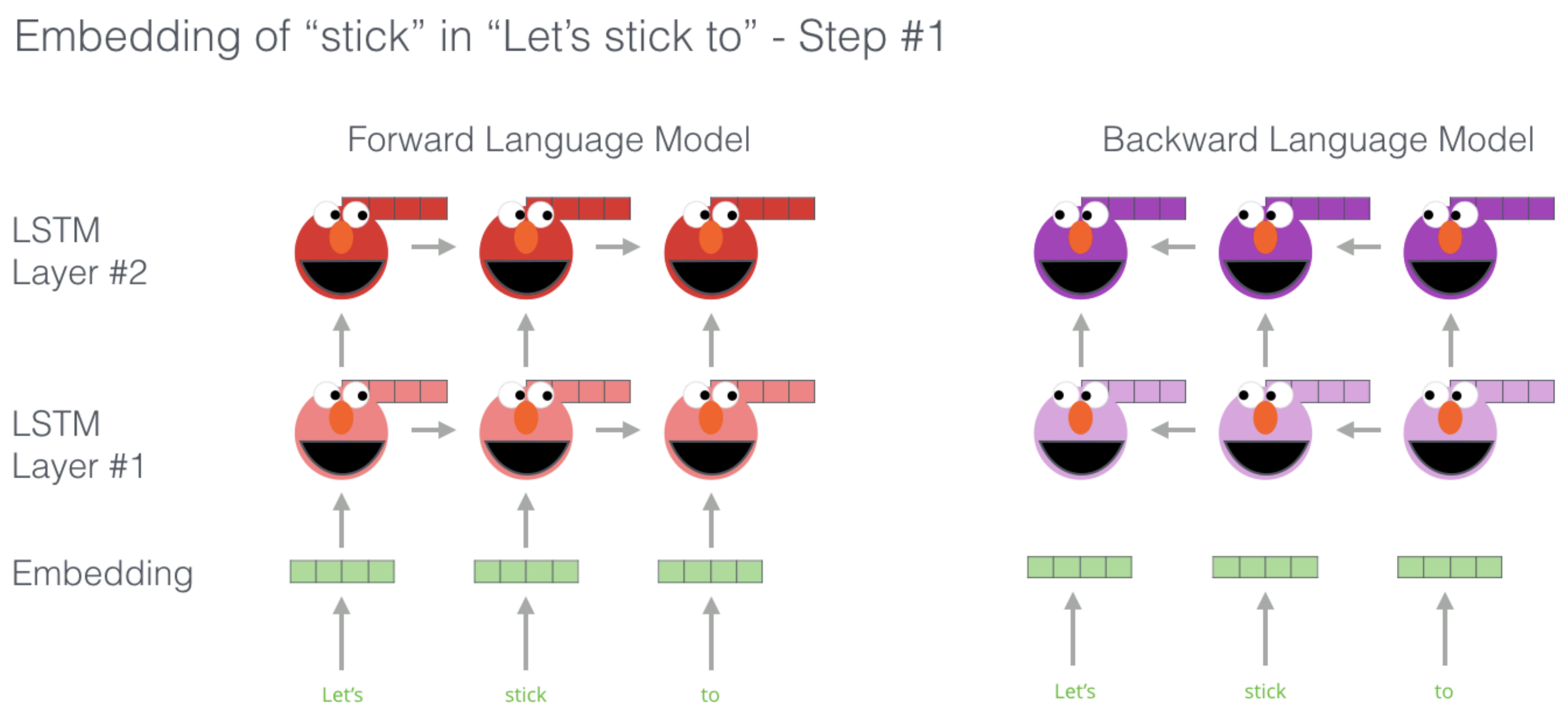
Supervised Pre-training

Downstream tasks

(ELMo, AllenNLP 2018)



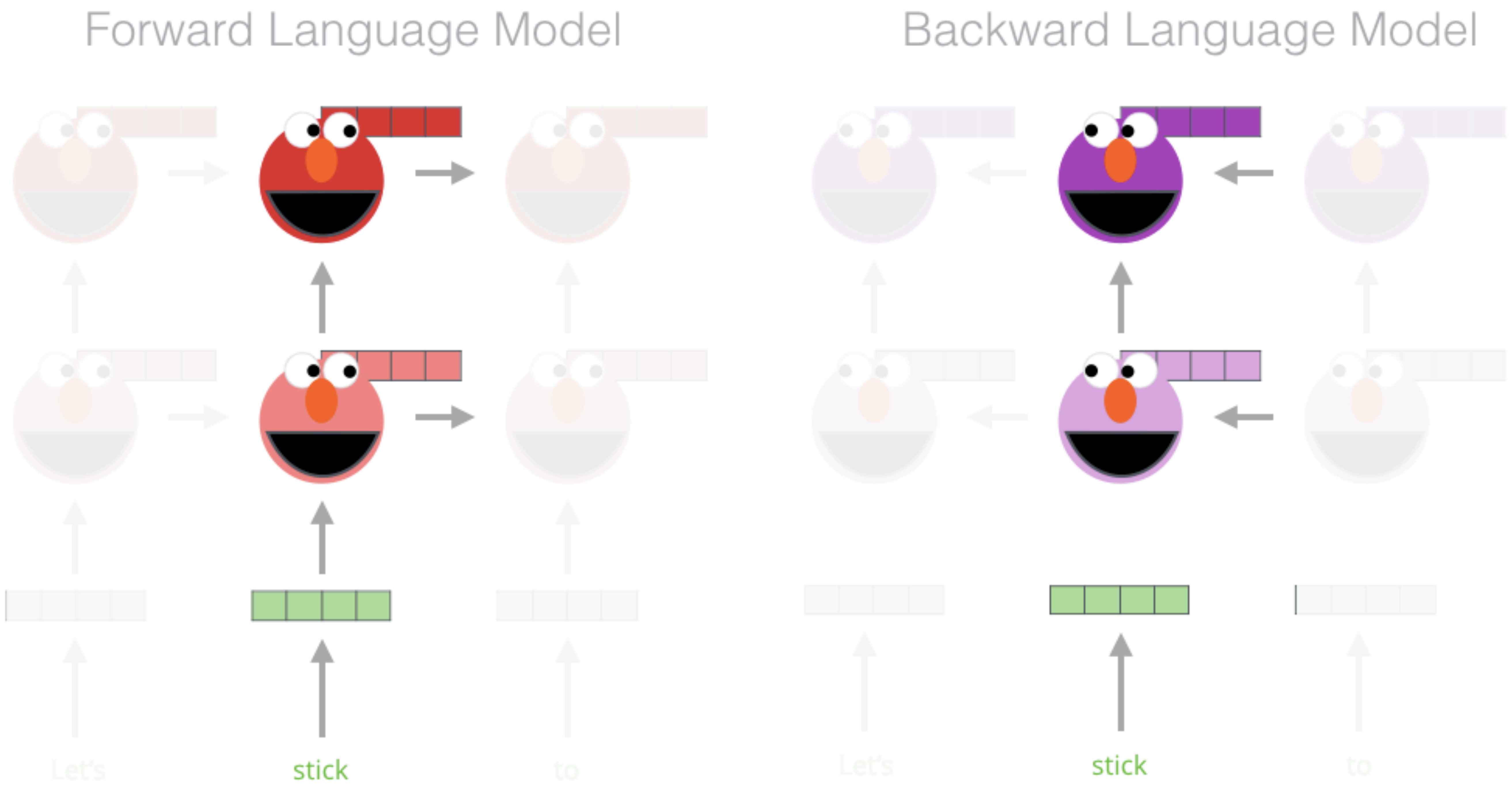
Language Model: Predict the next word



ELMo combines a forward and backward language model

Embedding of “stick” in “Let’s stick to” - Step #2

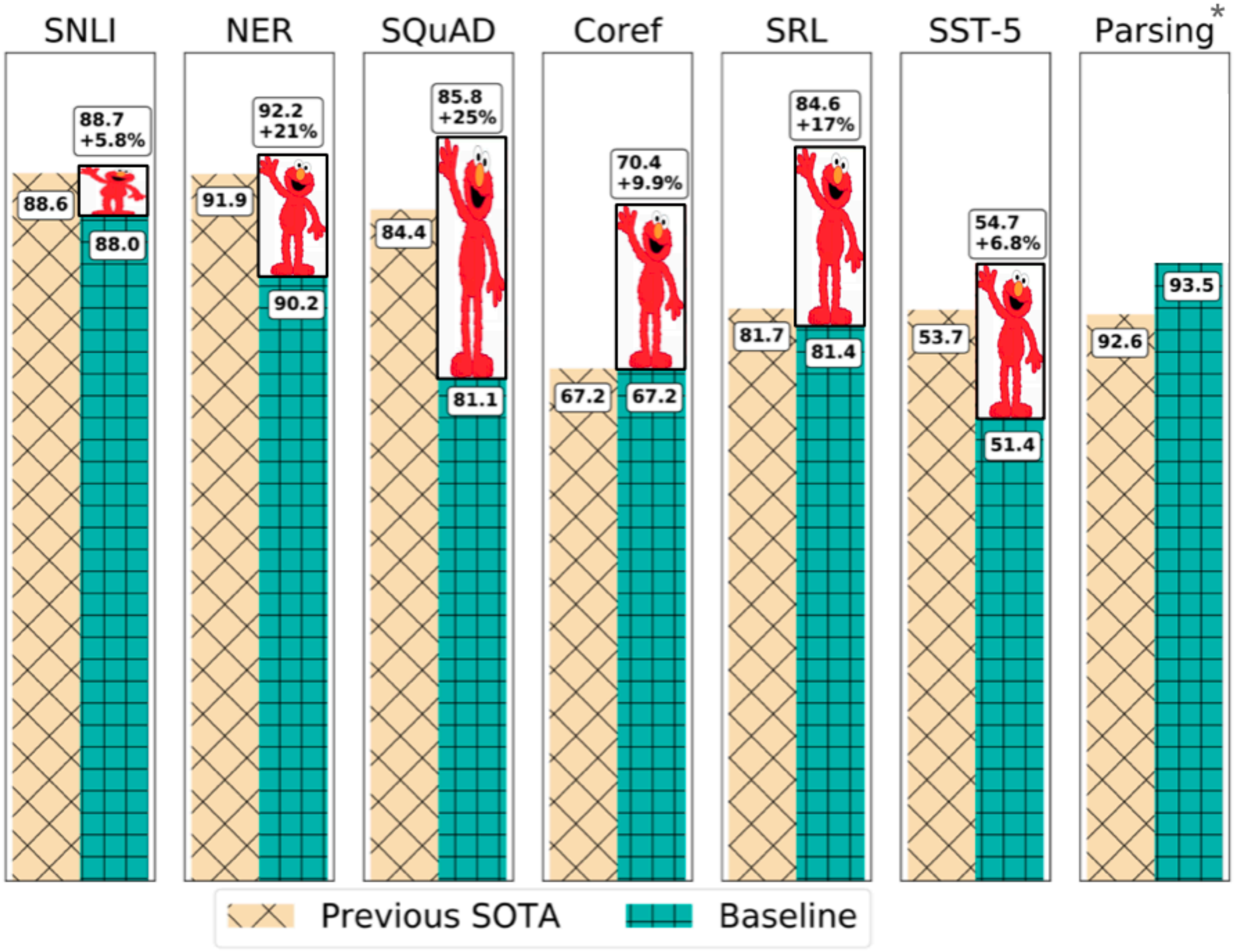
1- Concatenate hidden layers



2- Multiply each vector by a weight based on the task

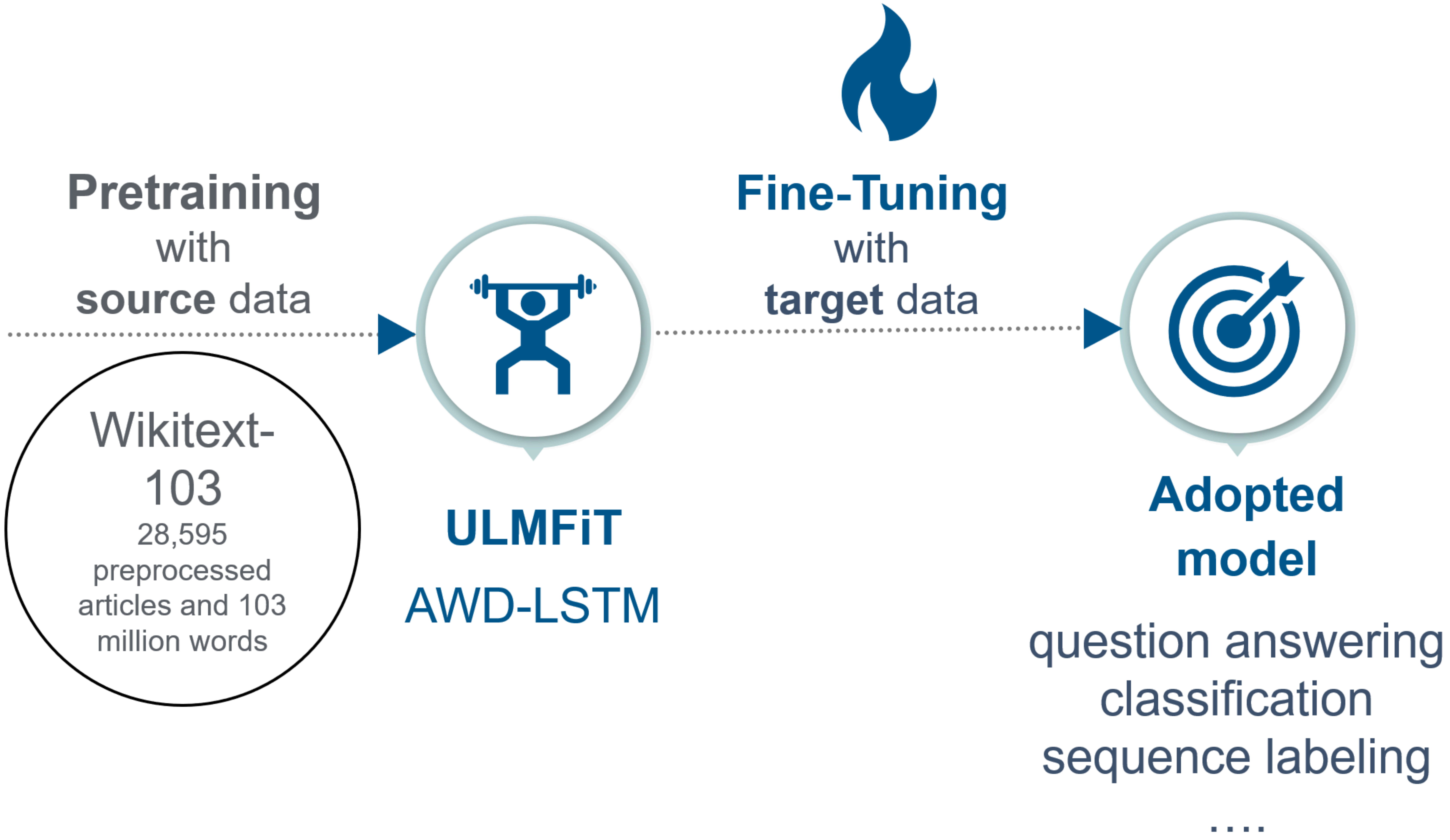
3- Sum the (now weighted) vectors

ELMo embedding of “stick” for this task in this context



*Kitaev and Klein, ACL 2018 (see also Joshi et al., ACL 2018)

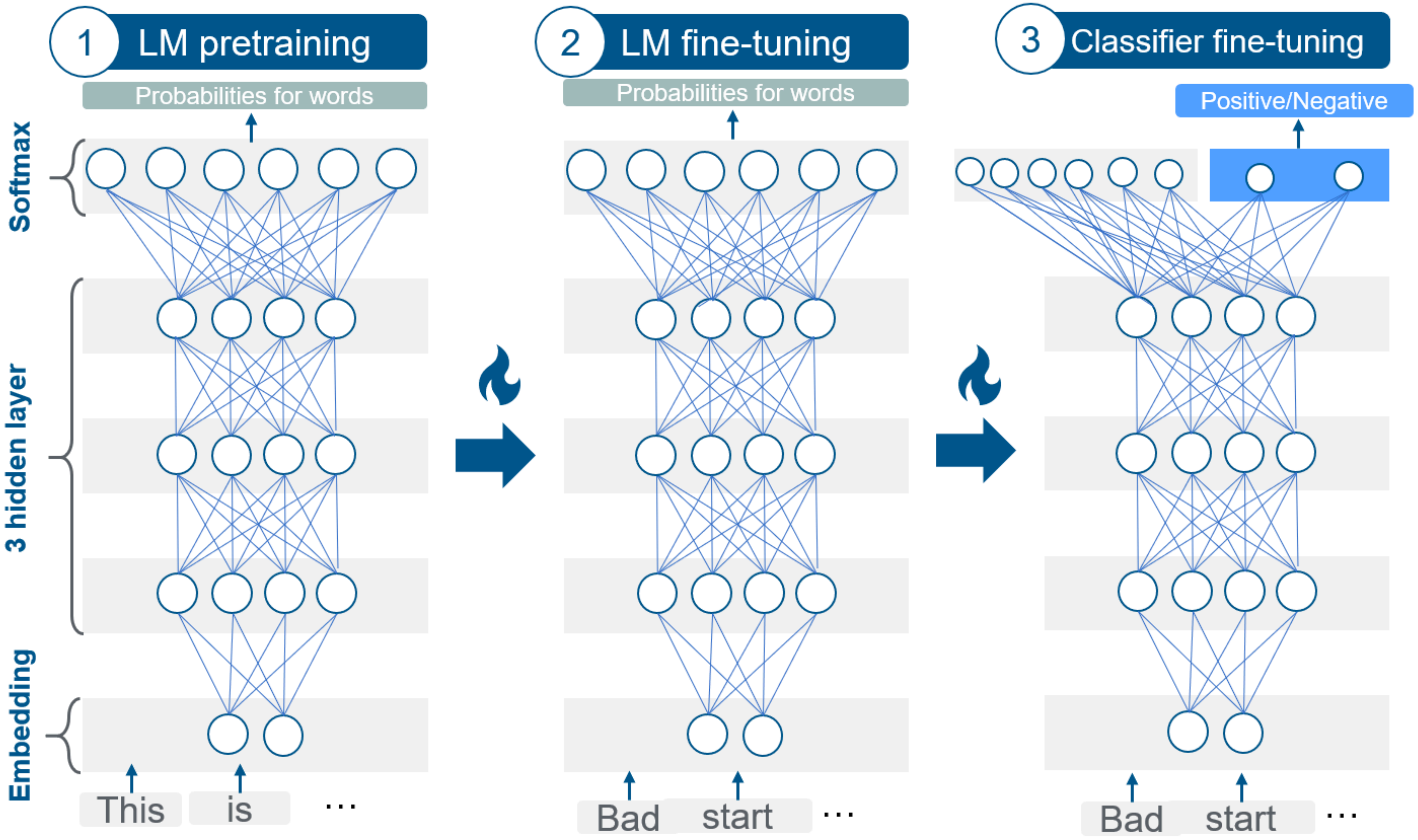
Fine-Tuning (ULMFiT - Howard & Ruder 2018)



Source: Becker, et al. 2020, *Modern Approaches to Natural Language Processi*

Fine-Tuning (ULMFiT - Howard & Ruder 2018)

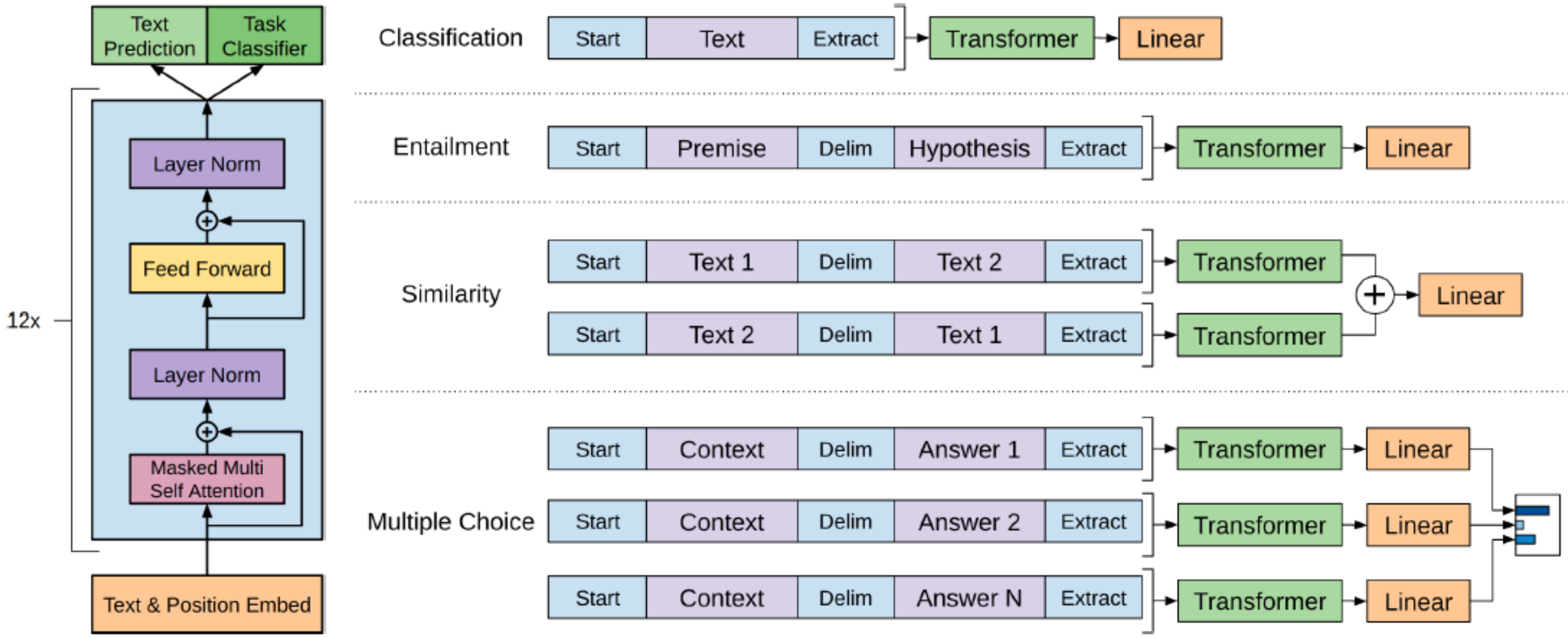
Techniques like "gradual unfreezing" help to prevent "catastrophic forgetting."



Source: Becker, et al. 2020, *Modern Approaches to Natural Language Processi*

Use a transformer! GPT (Open-AI / Radford, et al., 2018)

Transformer architecture dramatically sped up training, allowing for deeper models (12 layers in the original GPT), and bigger training data (Books Corpus in the



GPT used just the decoder, on an LM task.

Image Source: Radford, et al. 2018

BERT (Google, 2018)

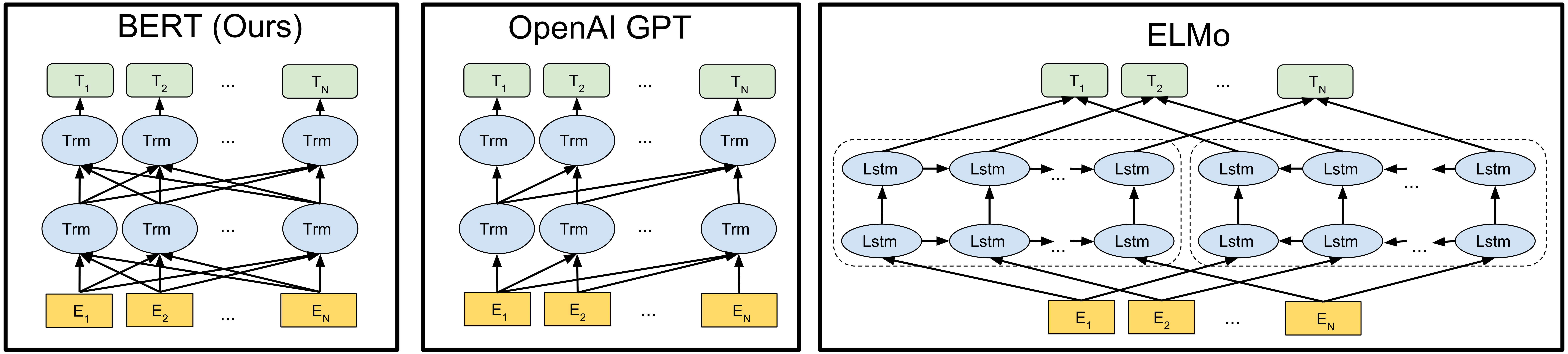
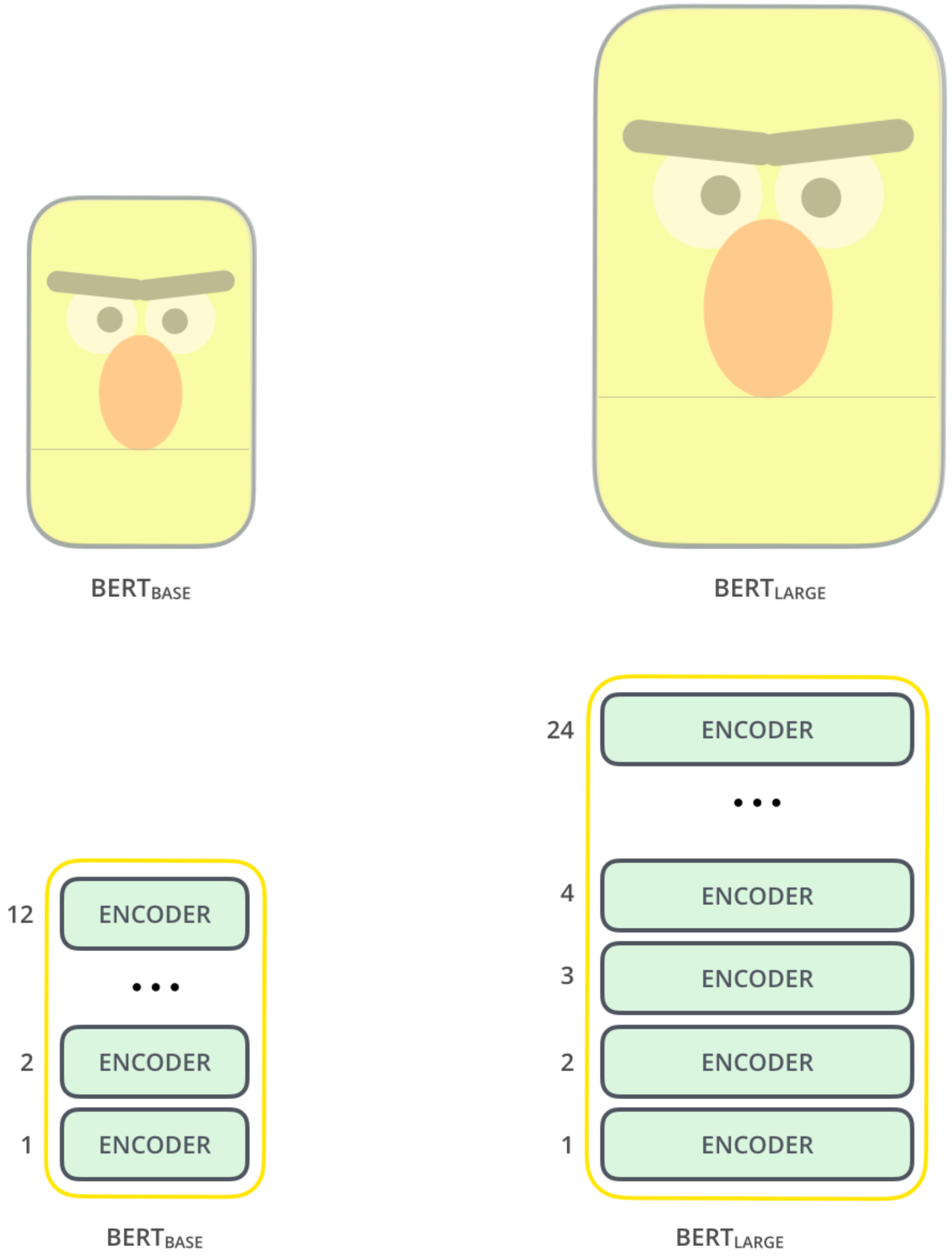
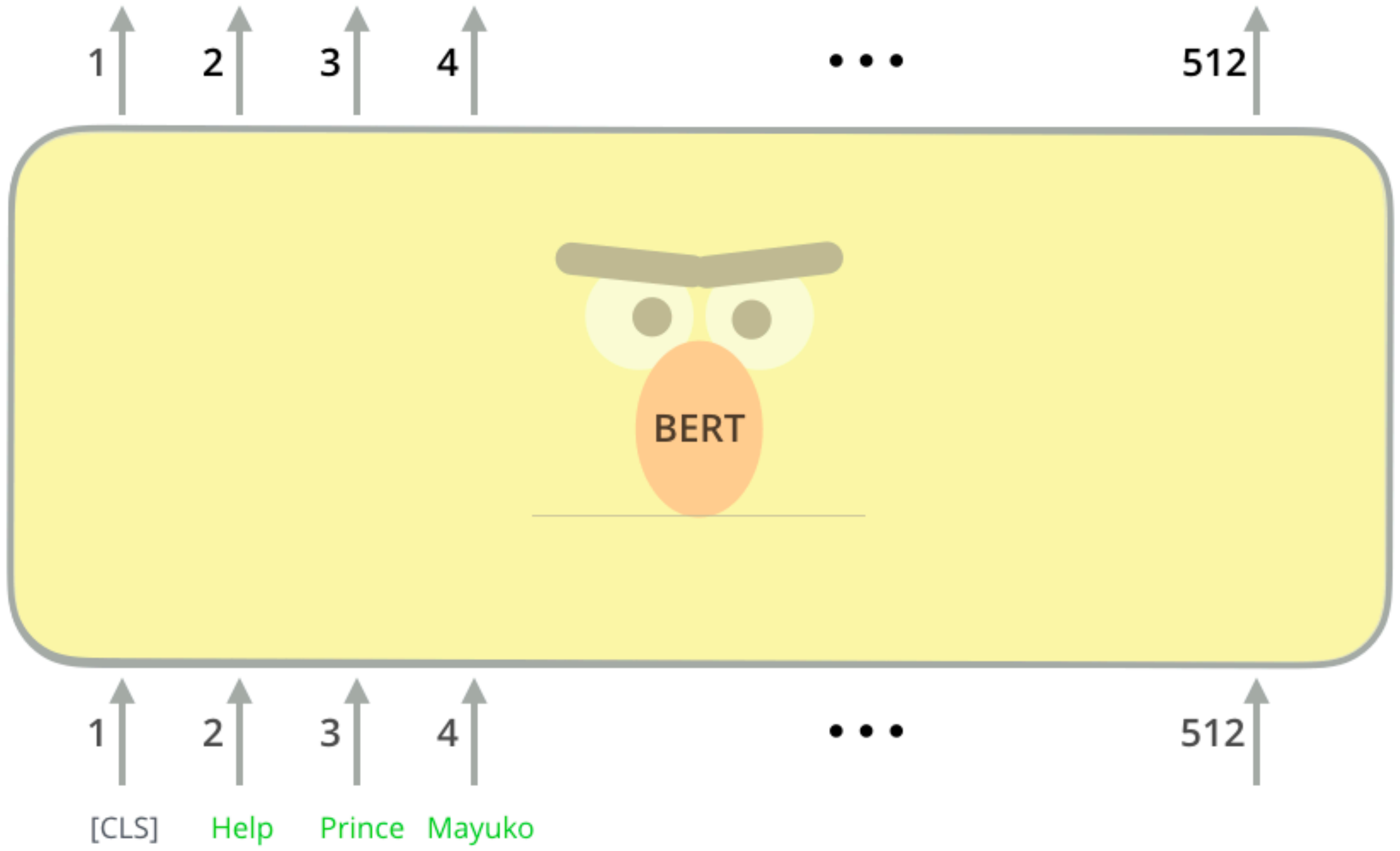
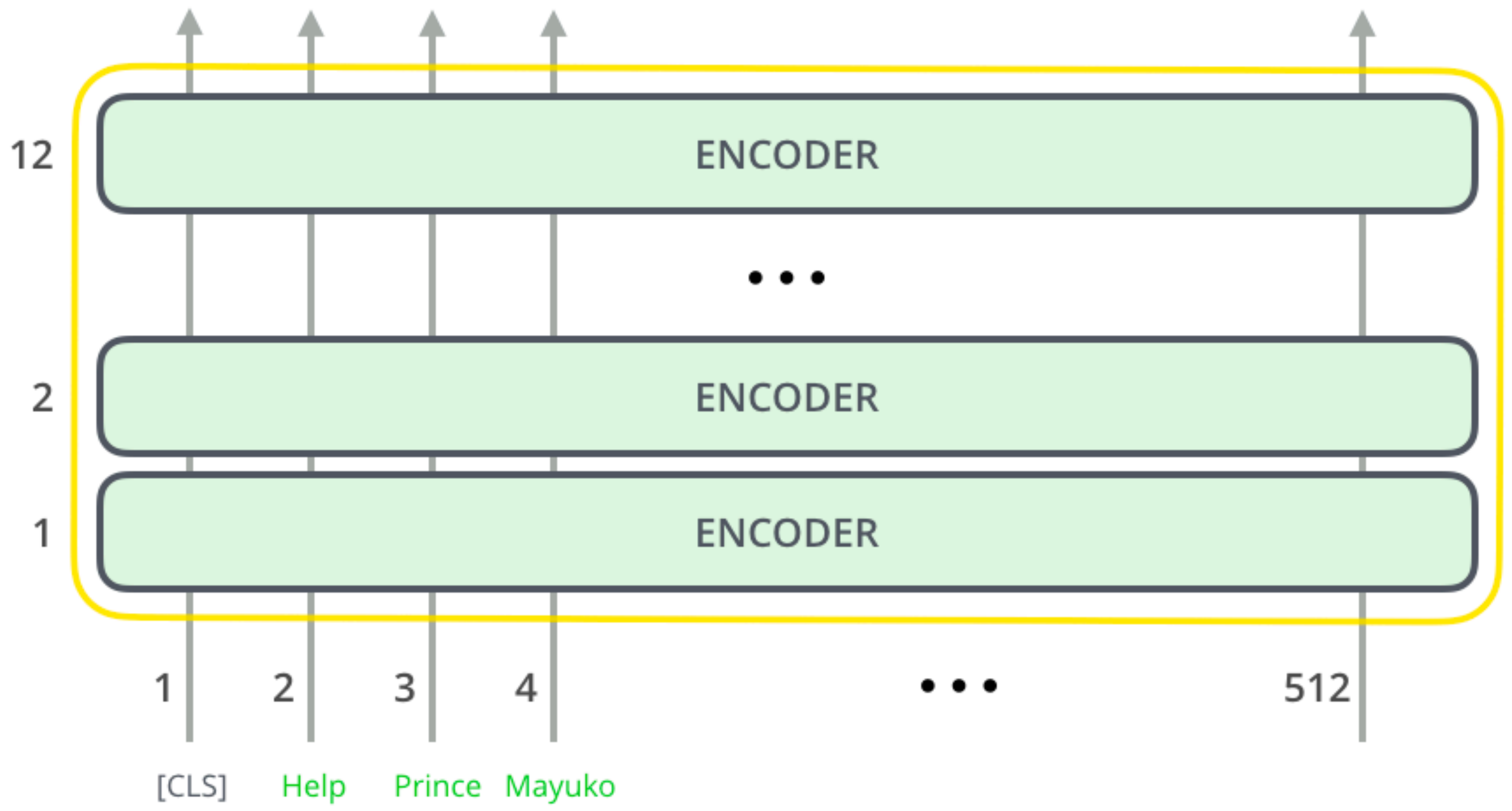


Figure 3: Differences in pre-training model architectures. BERT uses a bidirectional Transformer. OpenAI GPT uses a left-to-right Transformer. ELMo uses the concatenation of independently trained left-to-right and right-to-left LSTMs to generate features for downstream tasks. Among the three, only BERT representations are jointly conditioned on both left and right context in all layers. In addition to the architecture differences, BERT and OpenAI GPT are fine-tuning approaches, while ELMo is a feature-based approach.


BERT BASE
BERT LARGE
24
ENCODER
• • •
4
ENCODER
3
ENCODER
2
ENCODER
1
ENCODER
12
ENCODER
• • •
2
ENCODER
1
ENCODER
BERT BASE
BERT LARGE

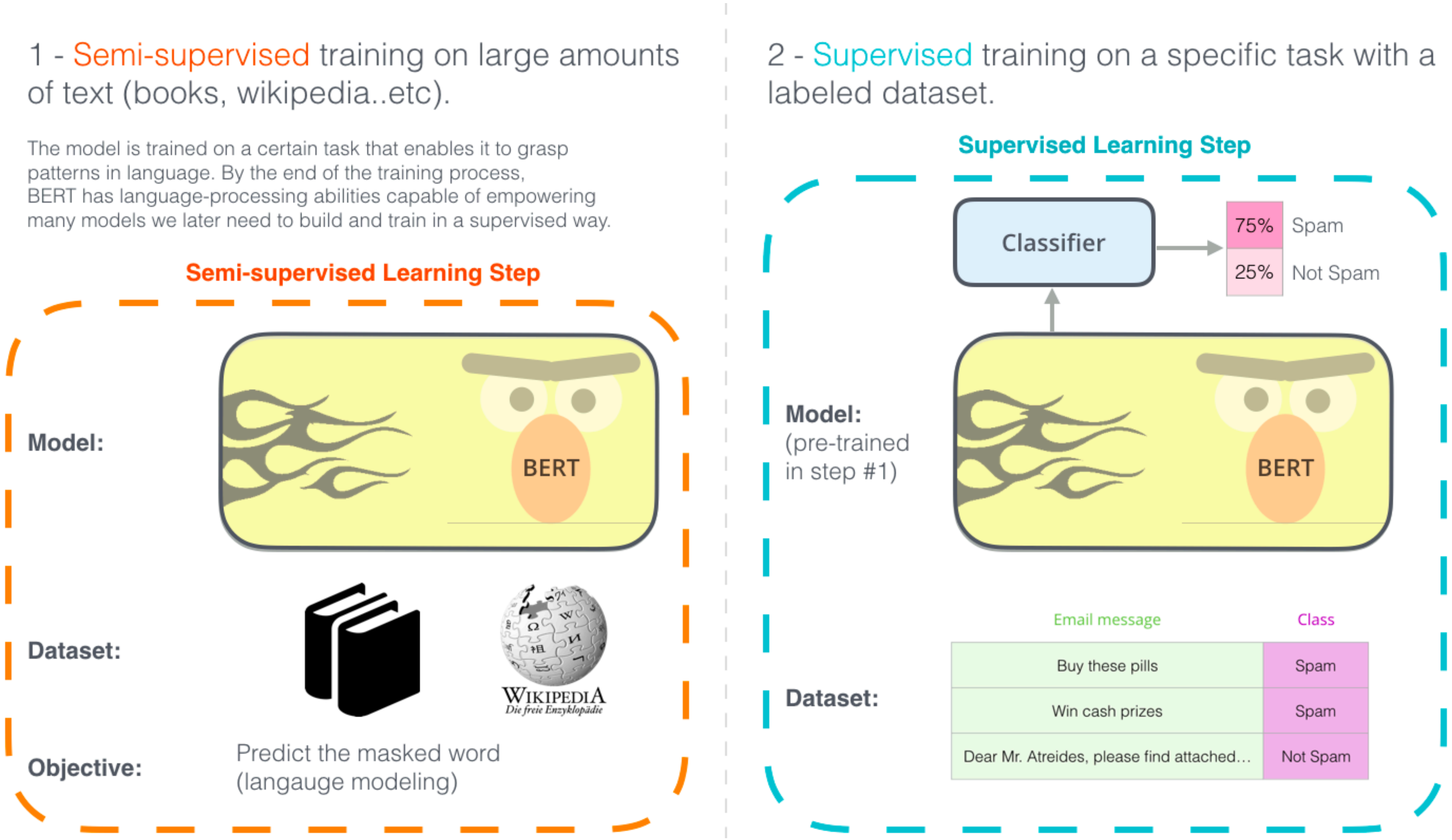

1
2
3
4
512
BERT
1
2
3
4
512
[CLS]
Help
Prince
Mayuko


12
ENCODER
ENCODER
2
ENCODER
1
1
2
3
4
512
[CLS]
Help
Prince
Mayuko

BERT input representation

BERT - so what?

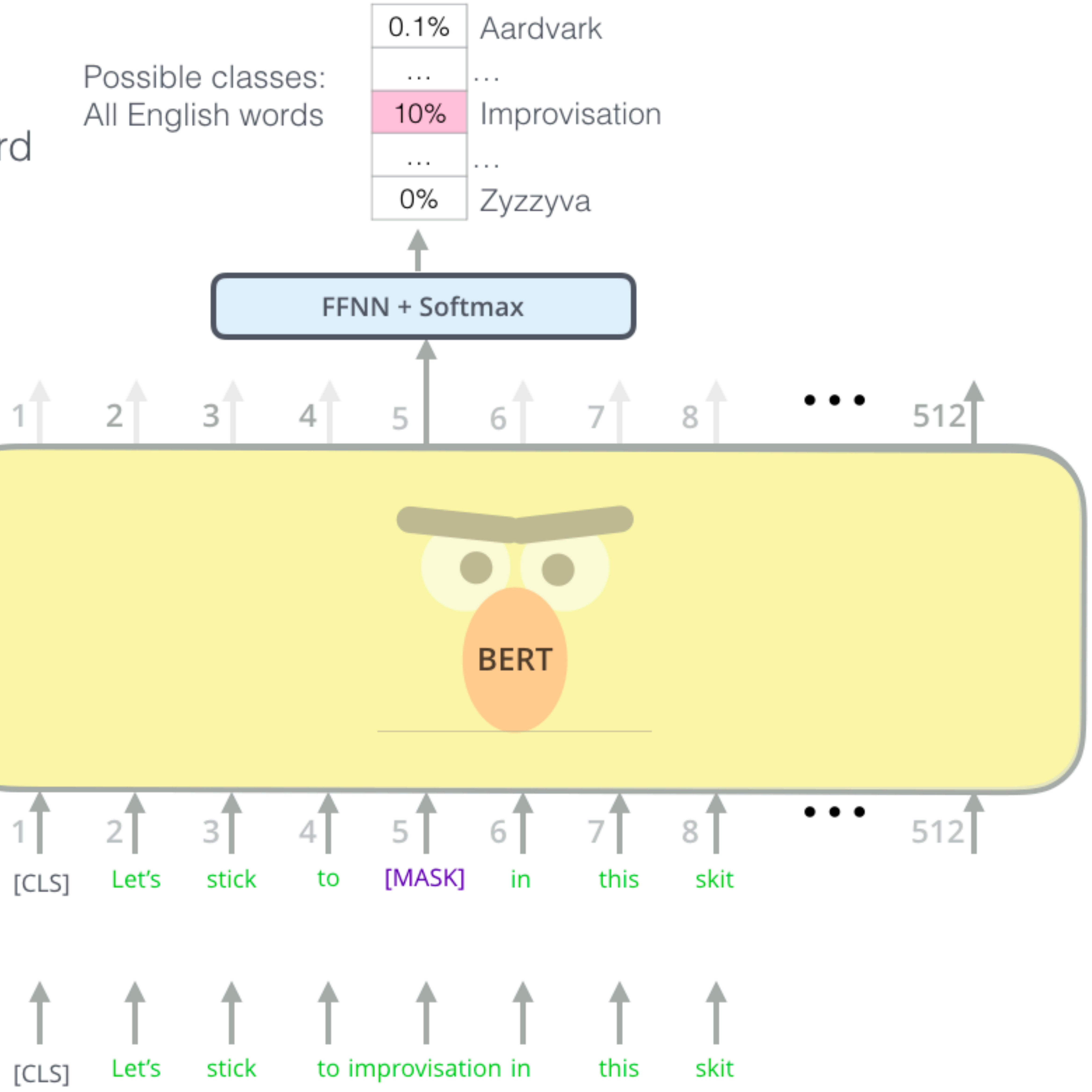
- BERT is a pretrained language model that can be fine-tuned to a specific task.



Source: Alammar (2018)

Pretraining #1: Masked Language Modeling

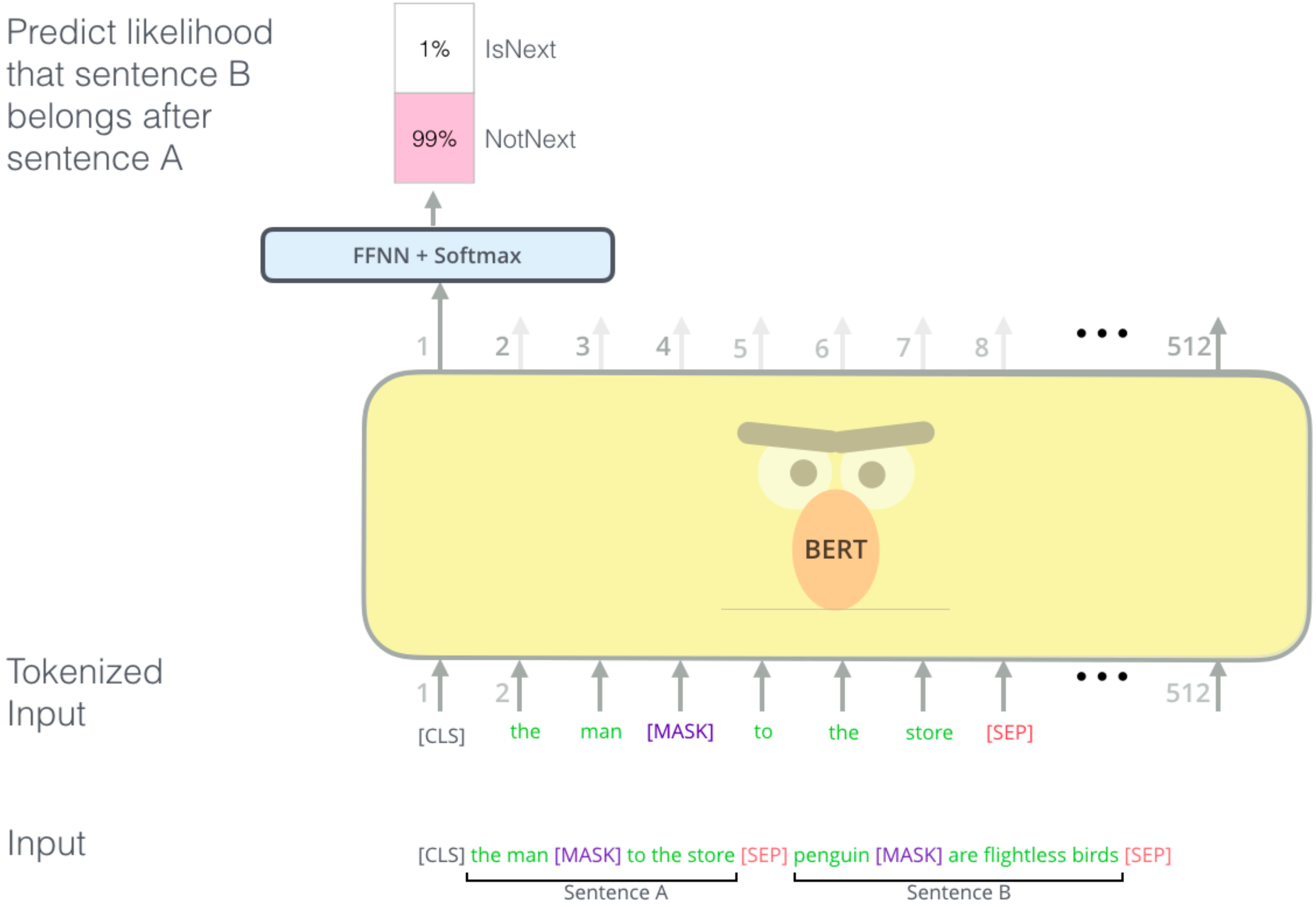
Use the output of the masked word's position to predict the masked word

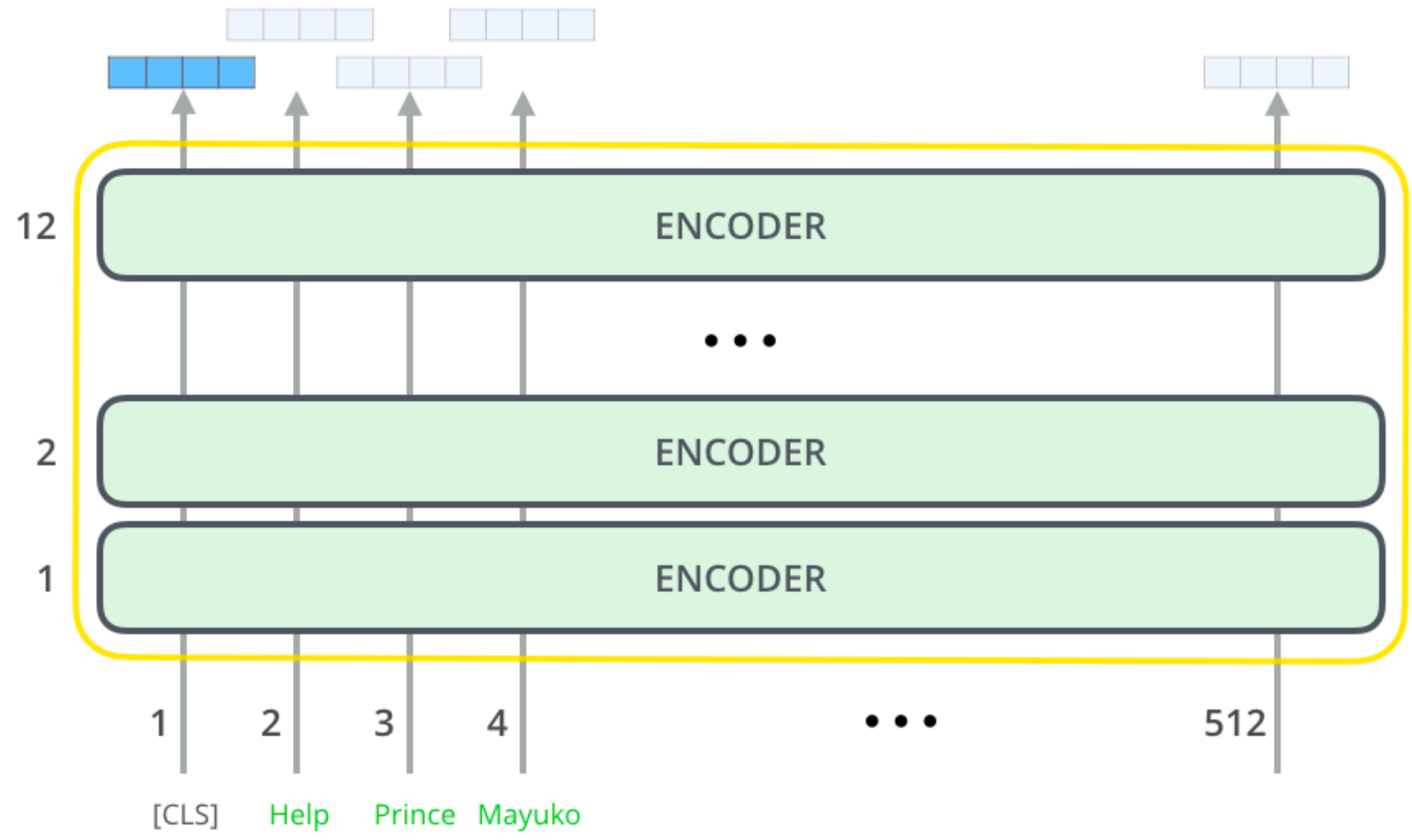


Randomly mask 15% of tokens

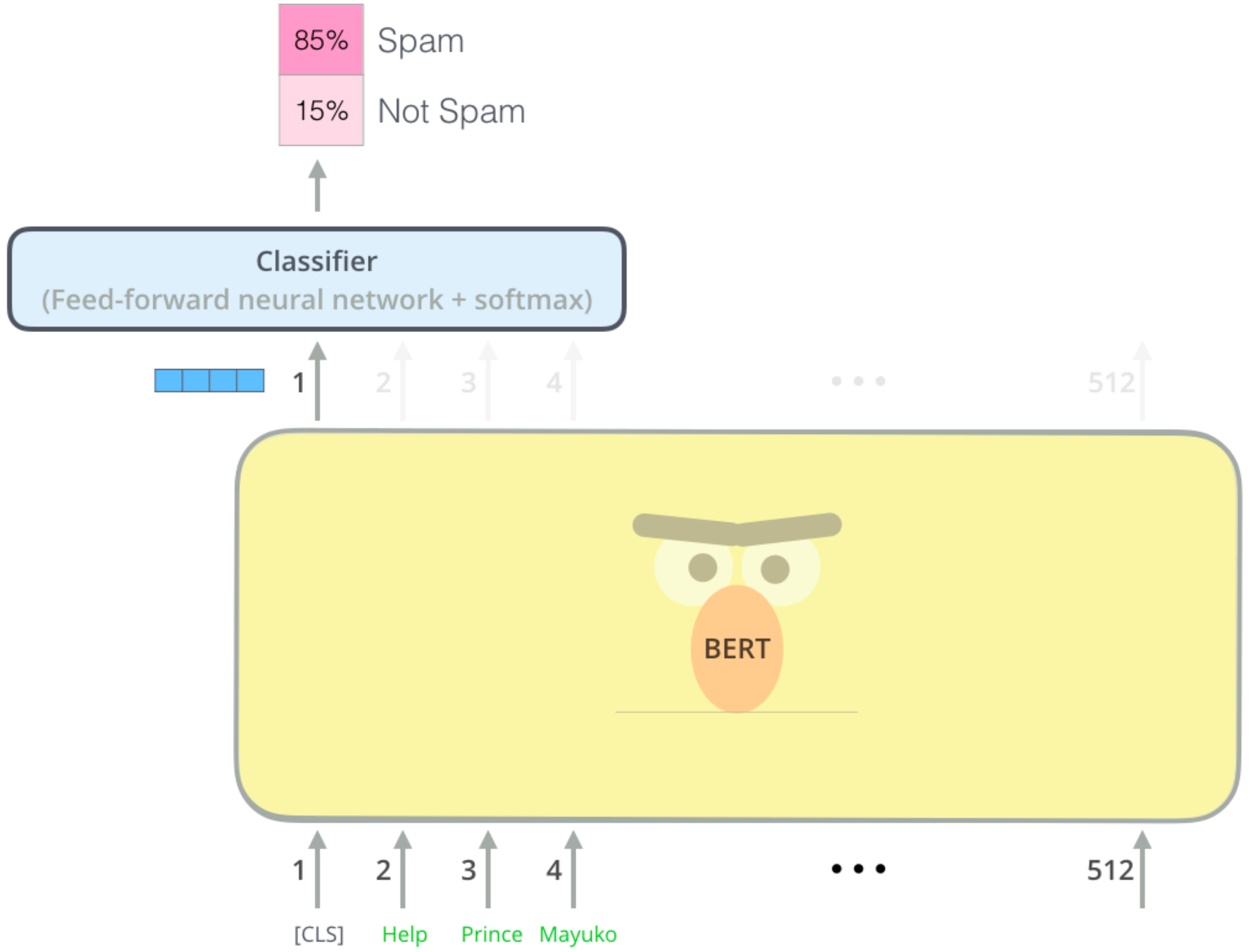
Input

Pretraining #2: Next Sentence Prediction




12
ENCODER
2
ENCODER
1
ENCODER
1
2
3
4
512
[CLS]
Help
Prince
Mayuko

BERT


85%
Spam
15%
Not Spam
Classifier
(Feed-forward neural network + softmax)
1
BERT
1
2
3
4
512
[CLS]
Help
Prince
Mayuko

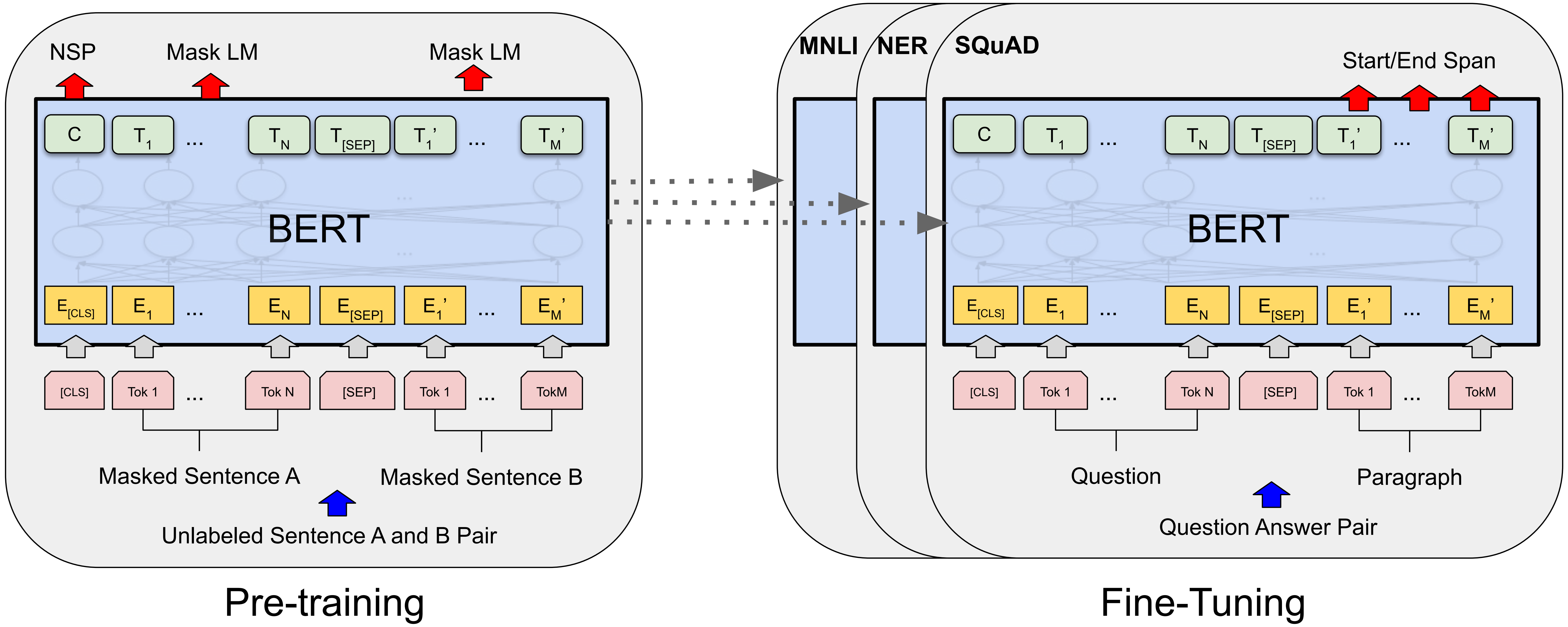
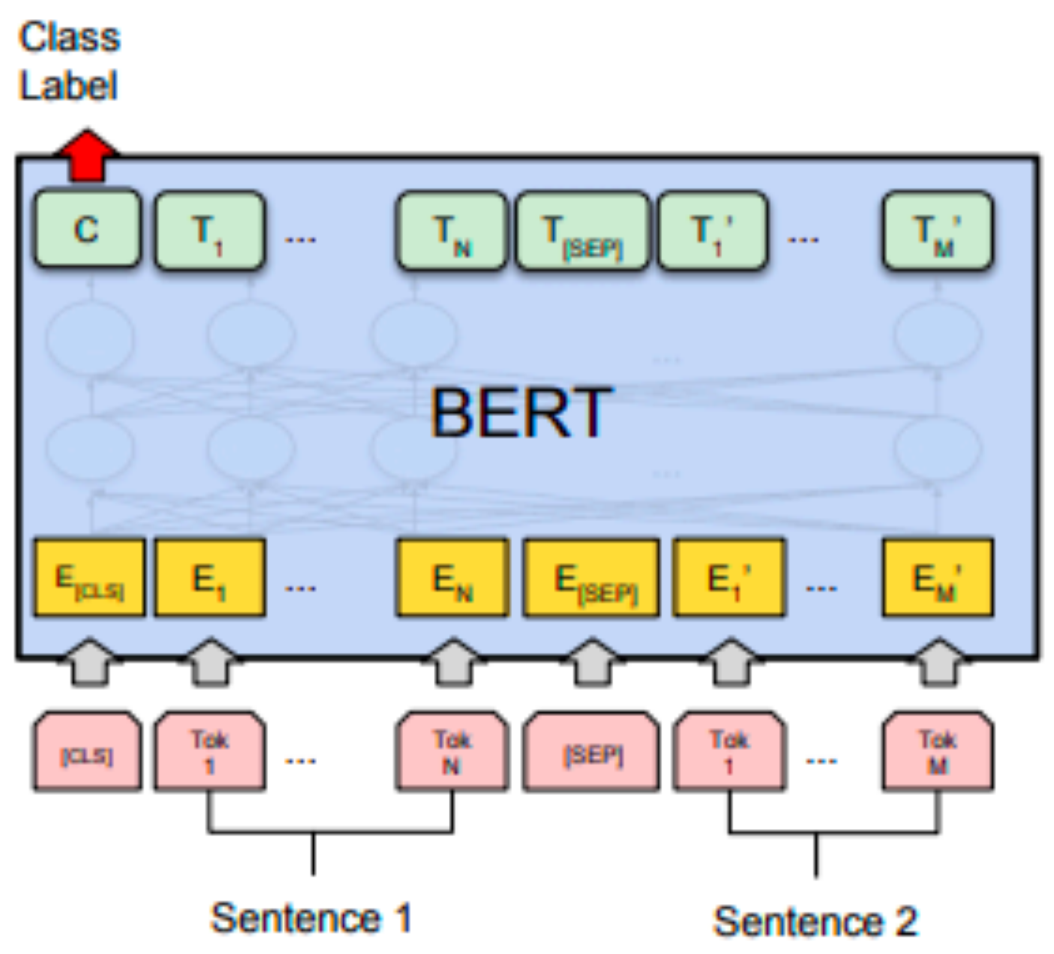
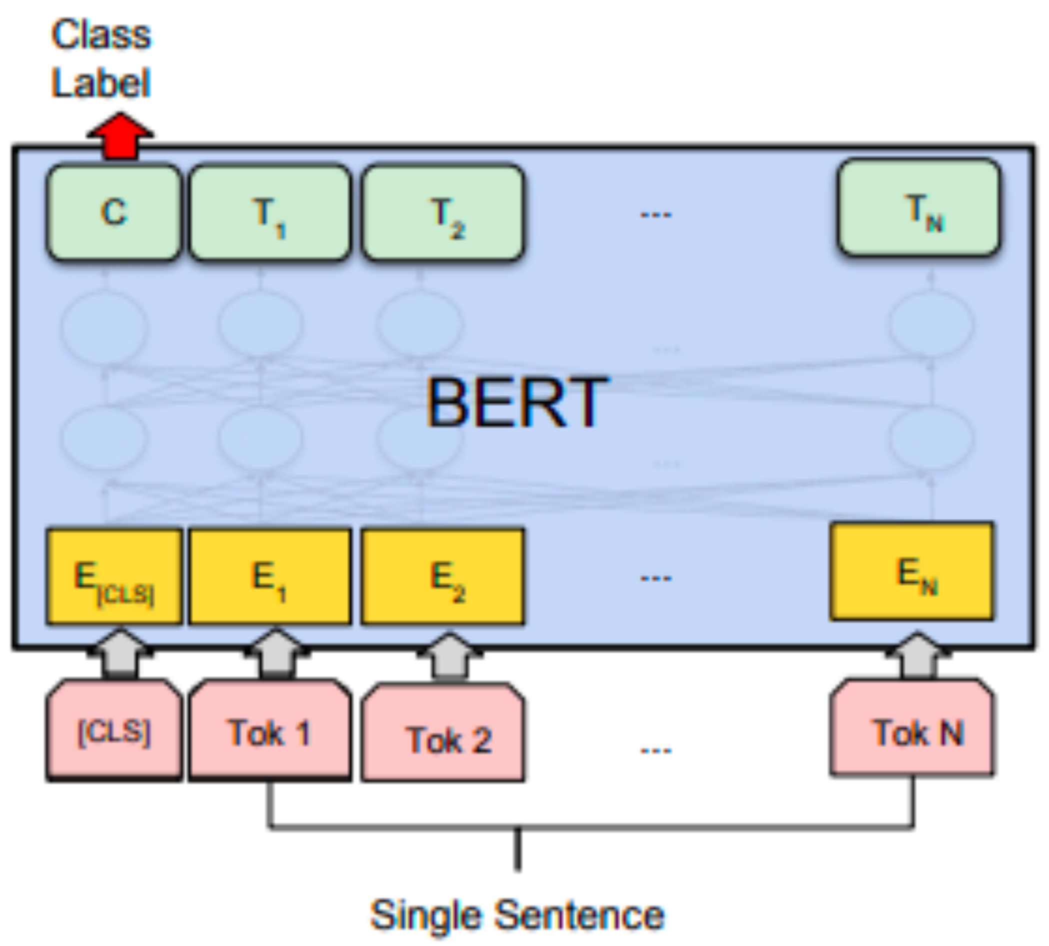


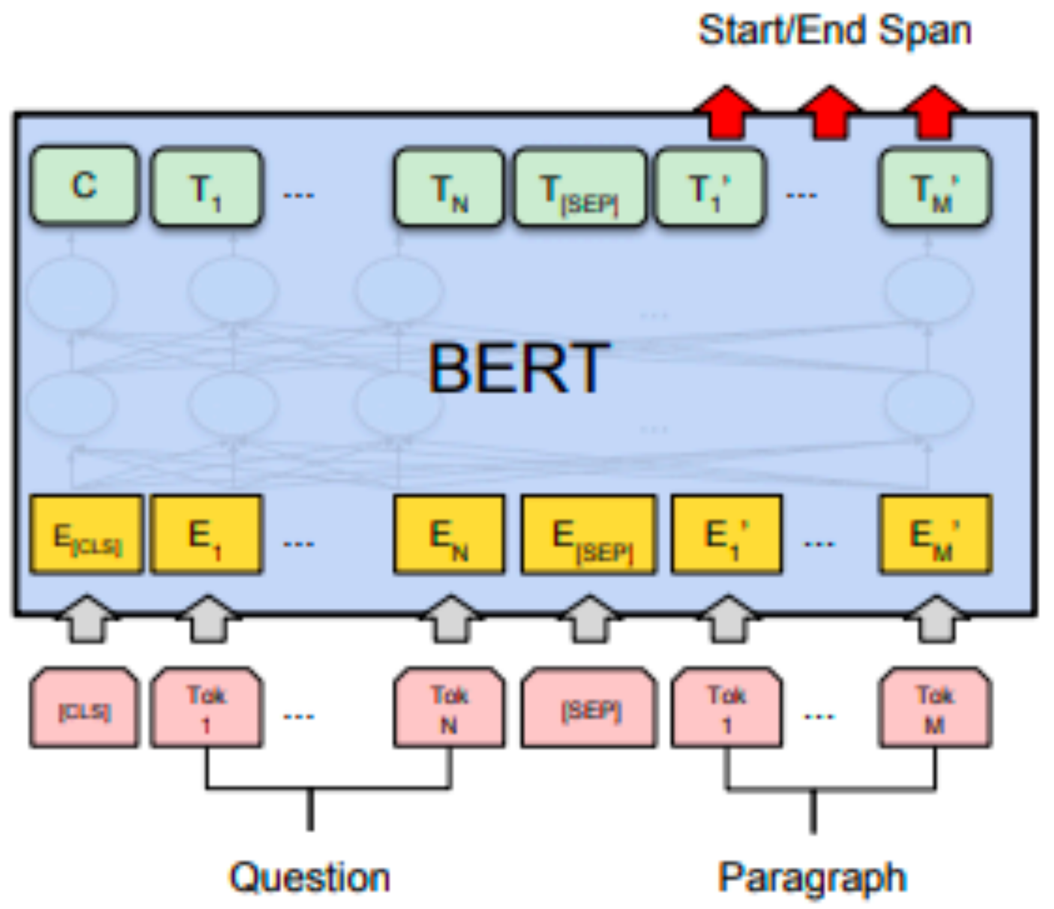
Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers).



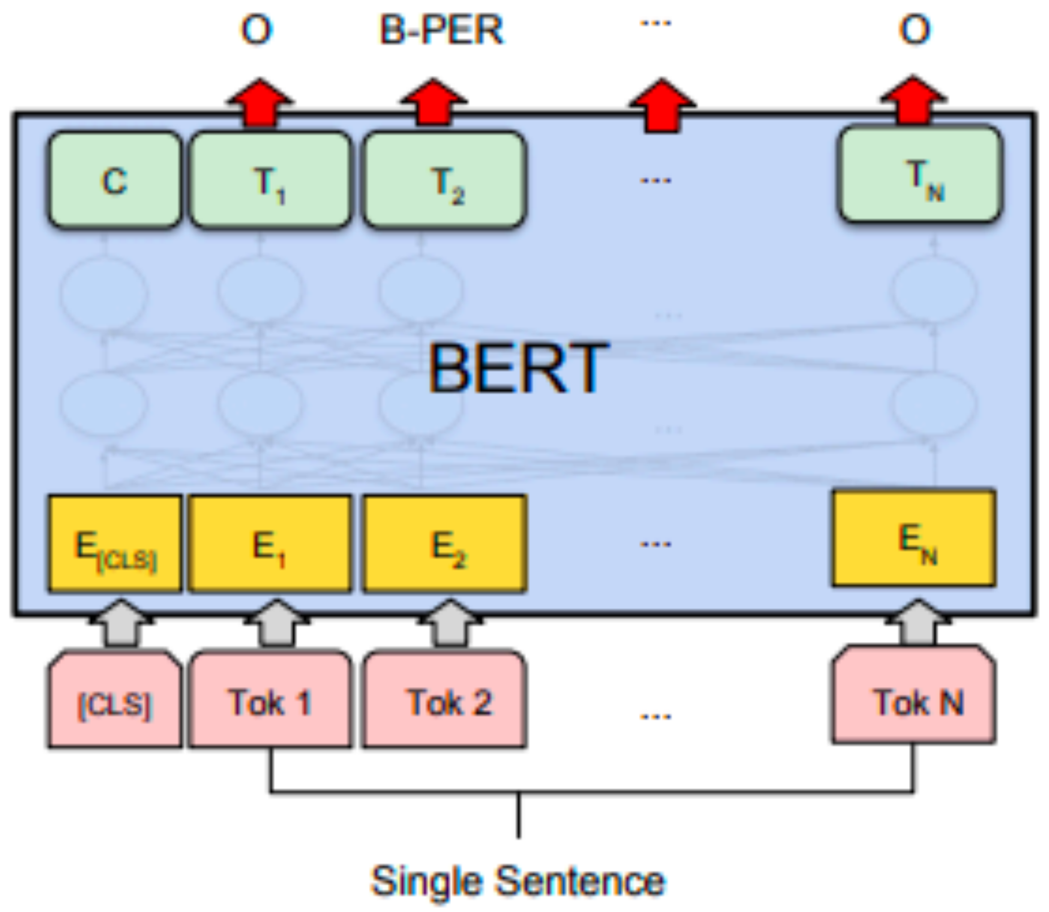
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



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

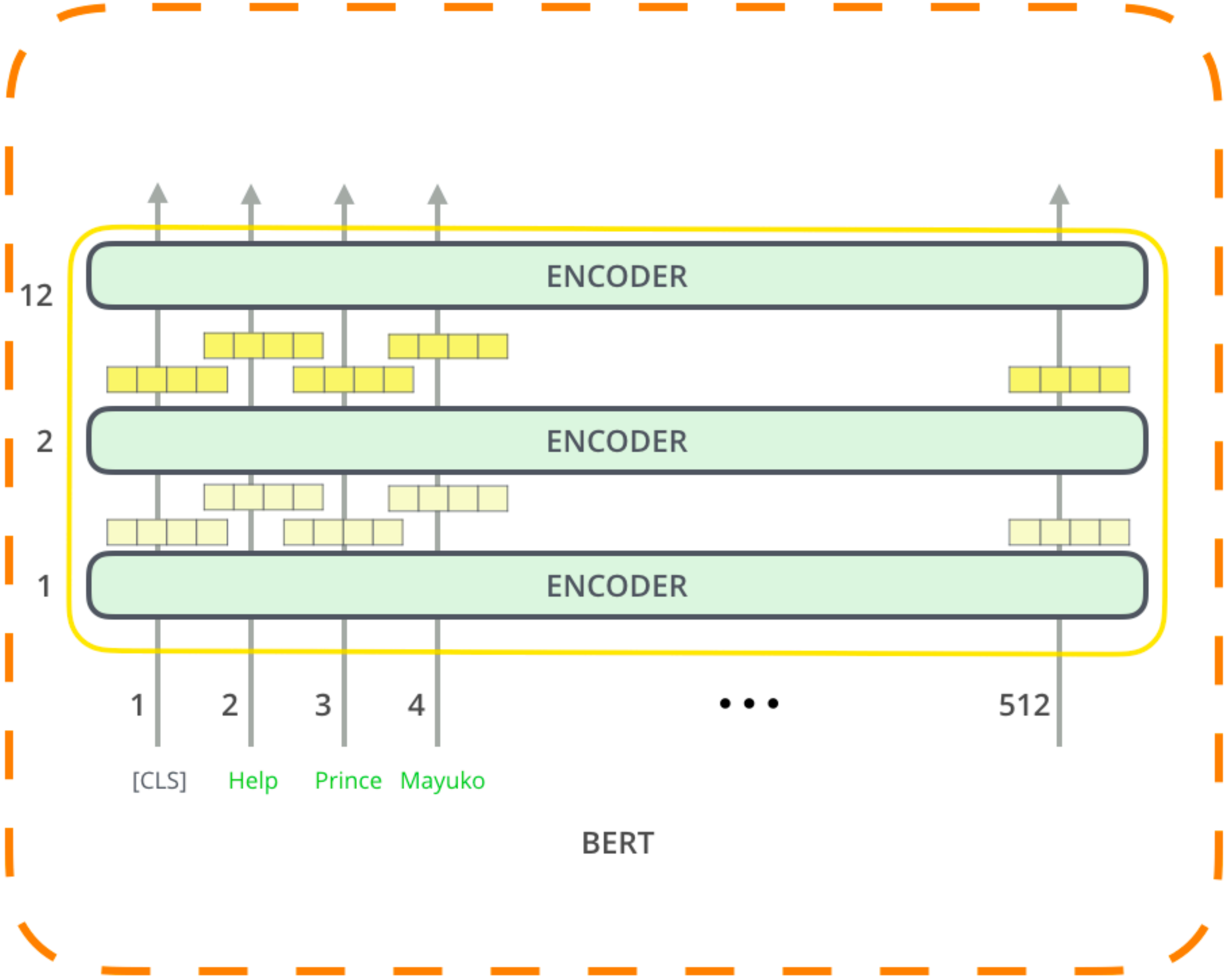


(c) Question Answering Tasks: SQuAD v1.1

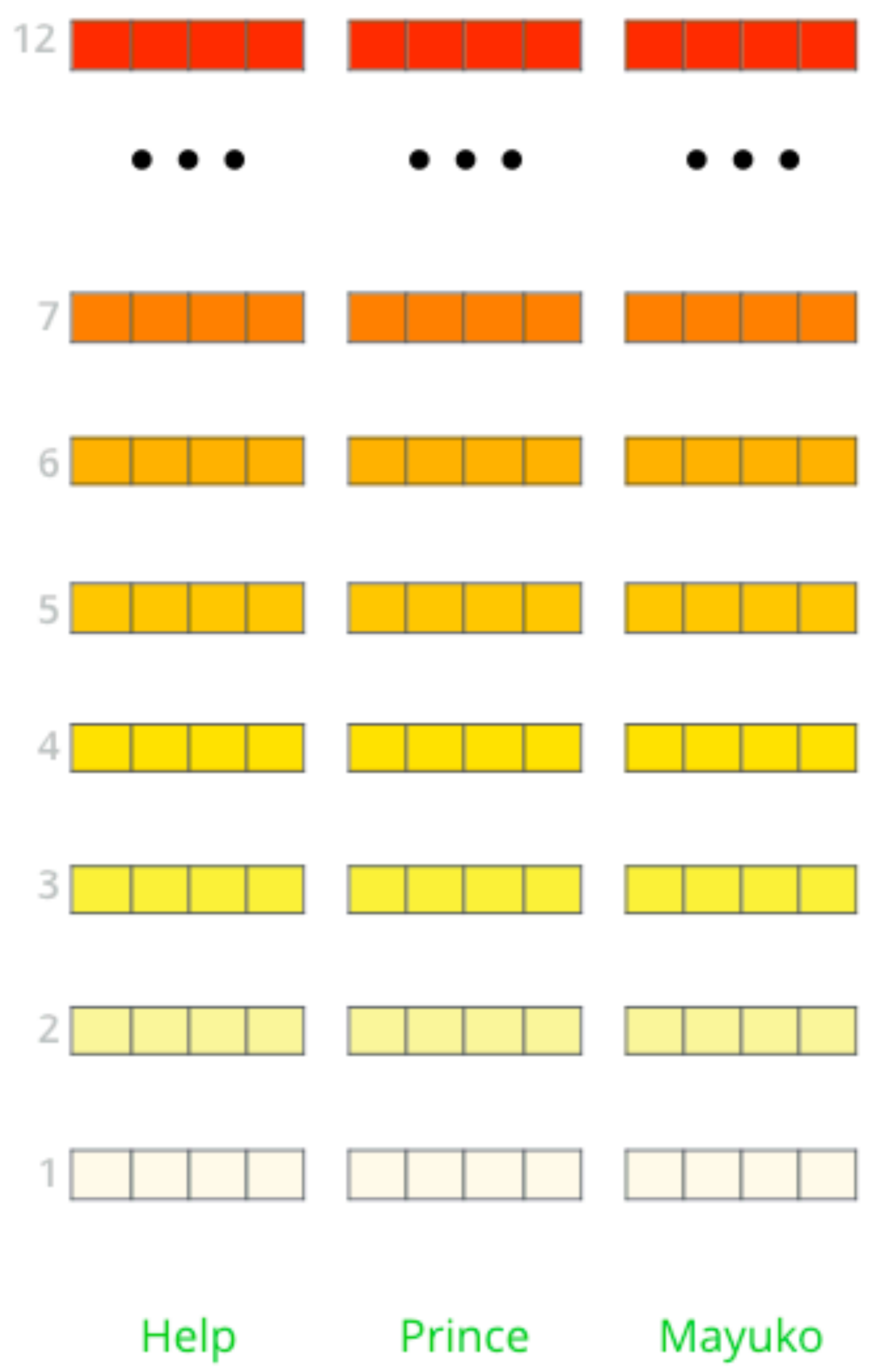


(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Generate Contexualized Embeddings



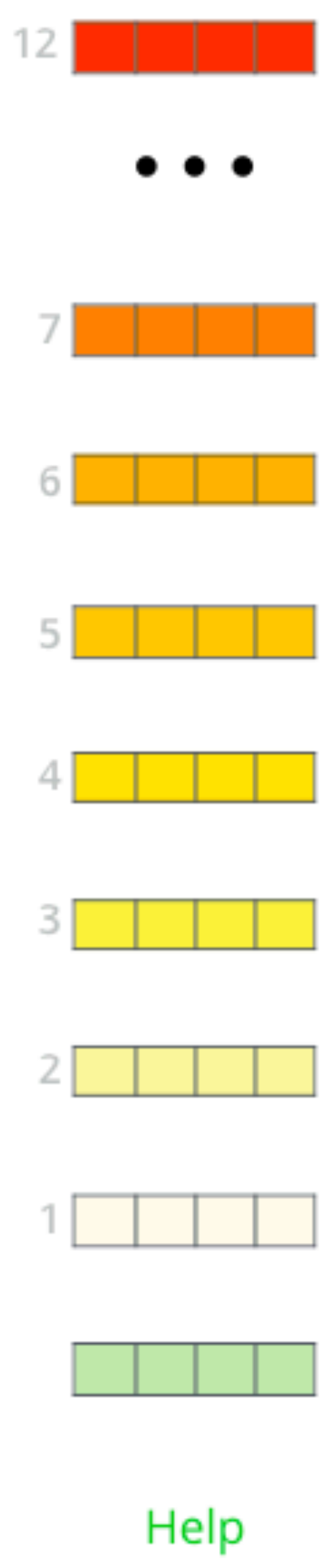
The output of each encoder layer along each token's path can be used as a feature representing that token.



But which one should we use?

What is the best contextualized embedding for “Help” in that context?

For named-entity recognition task CoNLL-2003 NER



	Dev F1 Score
First Layer (Embedding)	91.0
Last Hidden Layer (12)	94.9
Sum All 12 Layers (12 + … + 2 + 1 =)	95.5
Second-to-Last Hidden Layer (11)	95.6
Sum Last Four Hidden (12 + 11 + 10 + 9 =)	95.9
Concat Last Four Hidden (9, 10, 11, 12)	96.1

Highlights in pretrained models

- 2018 - ELMo (AllenNLP) - Contextualized word embeddings
- 2018 - ULMFit (fast.ai) - Fine-tuning a pretrained model
- 2018 - GPT - Use a transformer (decoder) - autoregressive
- 2018 - BERT (Google) - Bidirectional transformer encoder / auto-encoder, MLM/NSP
- Feb 2019 - GPT-2 (OpenAI) - 1.5 billion parameters
- Apr 2019 - ERNIE (Baidu) - masked phrases/entities in pretraining
- Jun 2019 - XLNet (CMU/Google) - permuted language modeling
- Jul 2019 - RoBERTa (Facebook) - pretraining differences, more training, more data
- Sep 2019 - ALBERT (Google) - Parameter-reduction techniques

Highlights in pretrained models

- Aug 2019 - StructBERT (Alibaba) - BERT masked LM + unshuffling scrambled word and sentence order.
- Sep 2019 - MegatronLM (Nvidia) - parallelism in pretraining, 8.3b parameter ~GPT-2, 3.9b ~BERT
- Oct 2019 - T5 (Google) - unifying text-to-text framework
- Jan 2020 - Reformer (Google) - locality-sensitive hashing allows context windows of 1m words
- Feb 2020 - Meena (Google) - 2.6B parameter chatbot
- Feb 2020 - Turing NLG (Microsoft) - 17 billion parameters
- May 2020 - GPT-3 (OpenAI) - 175 billion parameters
- May 2020 - ELECTRA (Stanford) - Efficiency from different pretraining
- Jul 2020 - DeBERTa (Microsoft) - “disentangled attention”
- May 2021 - OmniNET (Google) - “omnidirectional attention”

ERNIE (Baidu, 2019)

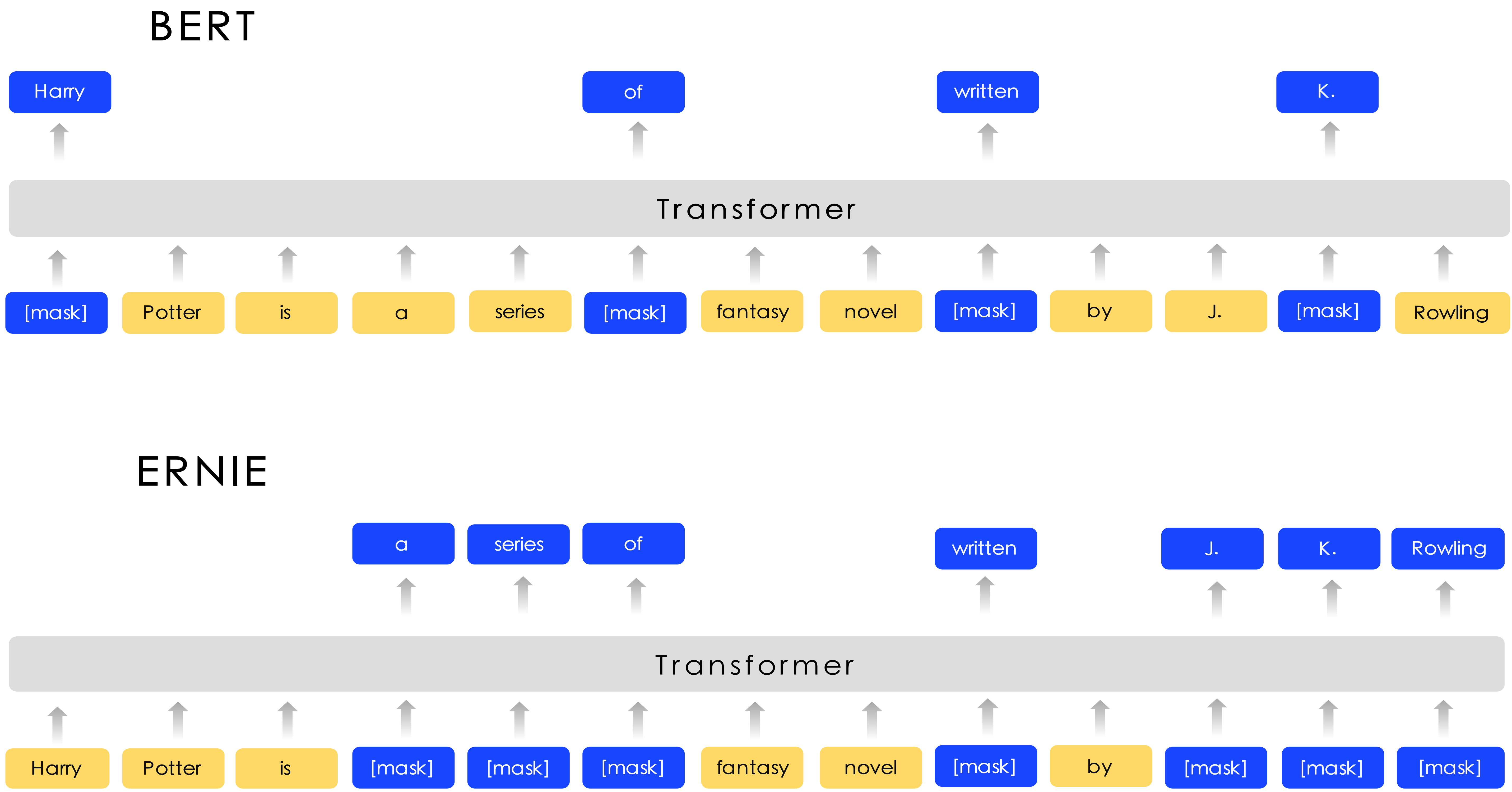


Figure 1: The different masking strategy between BERT and ERNIE

Training objective. BERT Masked LM + masked phrases and entities

StructBERT (Alibaba, 2019)

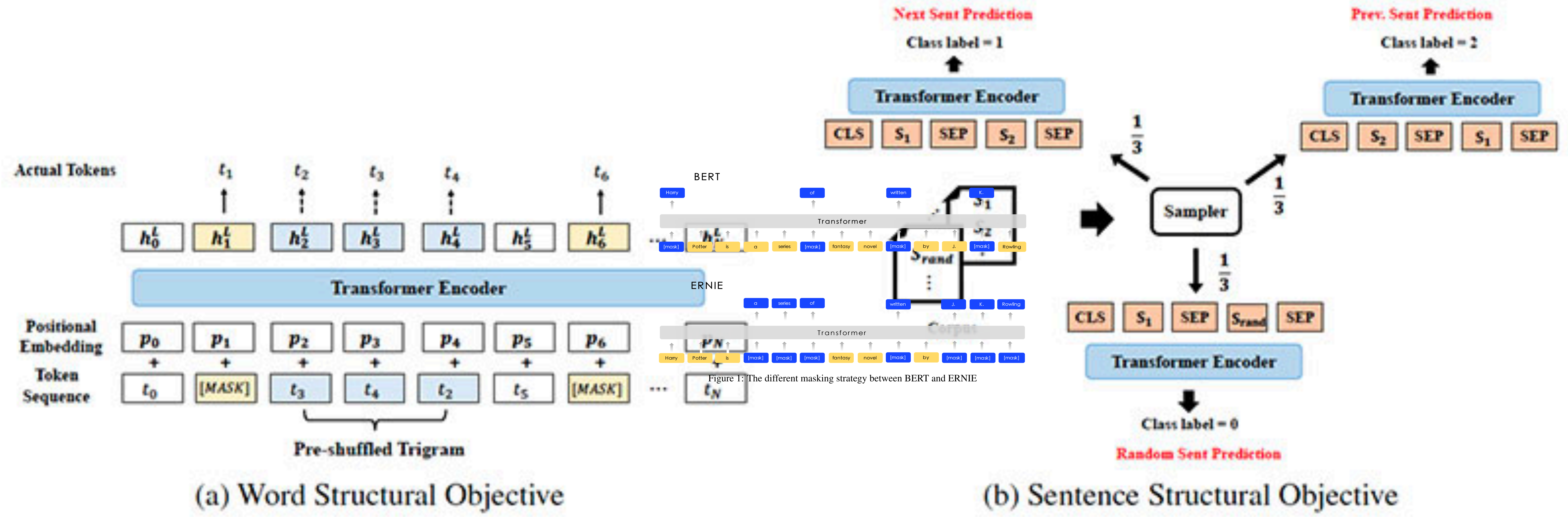
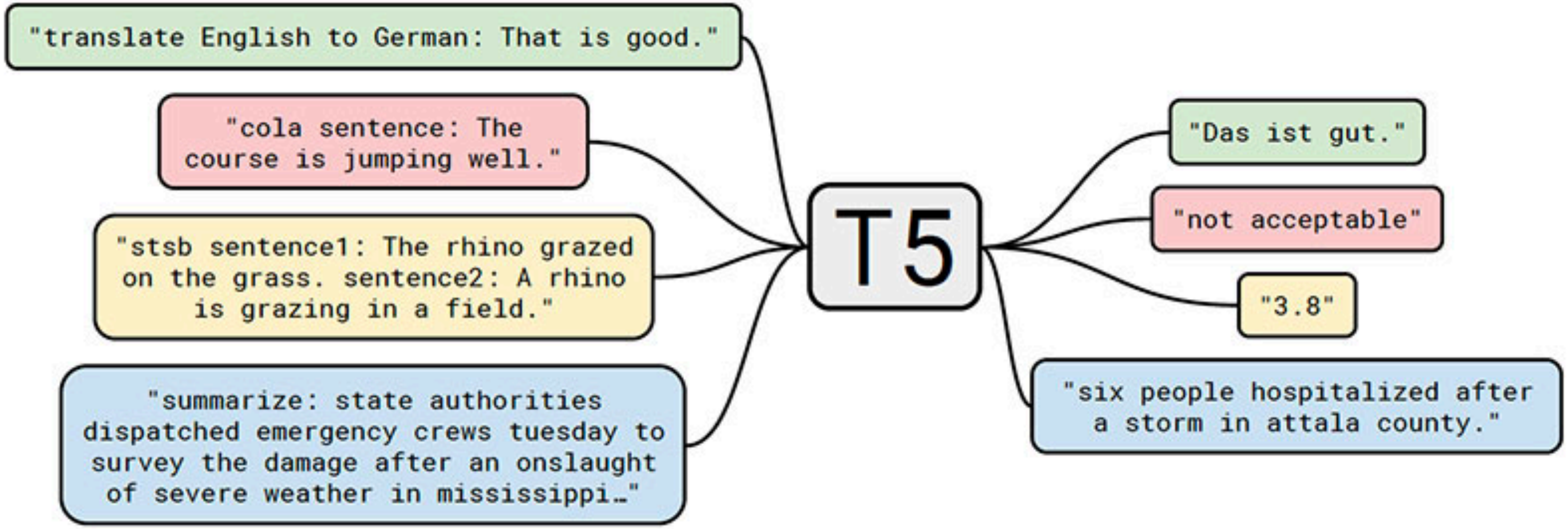


Figure 1: Illustrations of the two new pre-training objectives

Training objective. BERT Masked LM + unshuffling scrambled words and sentences.

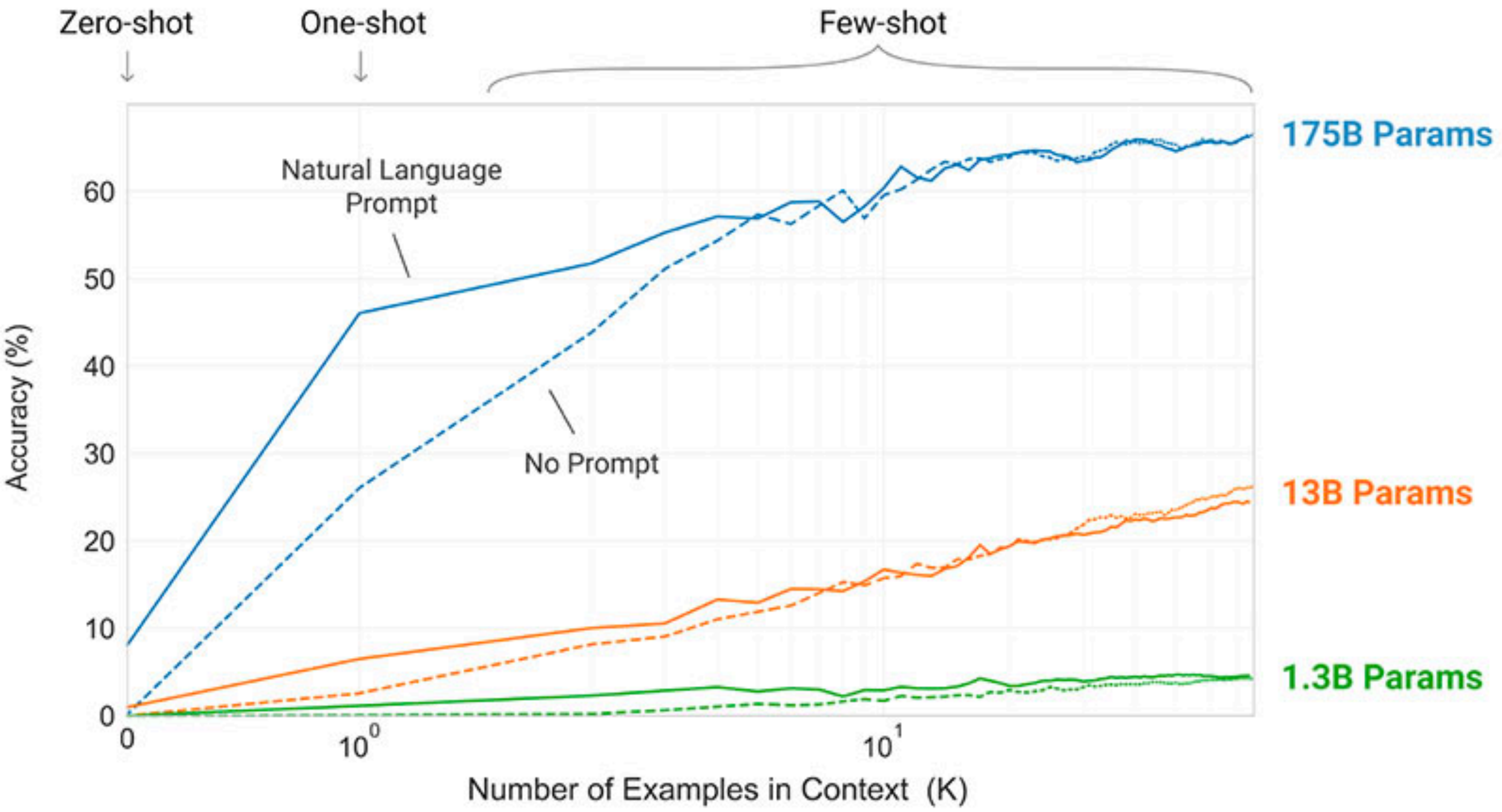
T5 (Google, 2019)

Unifying tasks as all text-to-text



Trained on C4, new “Collossal Clean Crawled Corpus.” - 11b param

GPT-3 (Google, 2019)



Autoregressive

Zero-, one-, and few-shot learning - 175 billion parameters

ALBERT (Google, 2019)

What's the core idea of this paper?

- It is not reasonable to further improve language models by making them larger because of memory limitations of available hardware, longer training times, and unexpected degradation of model performance with the increased number of parameters.
- To address this problem, the researchers introduce the **ALBERT** architecture that incorporates two parameter-reduction techniques:
  - **factorized embedding parameterization**, where the size of the hidden layers is separated from the size of vocabulary embeddings by decomposing the large vocabulary-embedding matrix into two small matrices;
  - **cross-layer parameter sharing** to prevent the number of parameters from growing with the depth of the network.
- The performance of ALBERT is further improved by introducing the self-supervised loss for **sentence-order prediction** to address BERT's limitations with regard to inter-sentence coherence.

What's the key achievement?

- With the introduced parameter-reduction techniques, the ALBERT configuration with 18× fewer parameters and 1.7× faster training compared to the original BERT-large model achieves only slightly worse performance.
- The much larger ALBERT configuration, which still has fewer parameters than BERT-large, outperforms all of the current state-of-the-art language modes by getting:
  - 89.4% accuracy on the RACE benchmark;
  - 89.4 score on the GLUE benchmark; and
  - An F1 score of 92.2 on the SQuAD 2.0 benchmark.

XLNet (CMU/Google 2019)

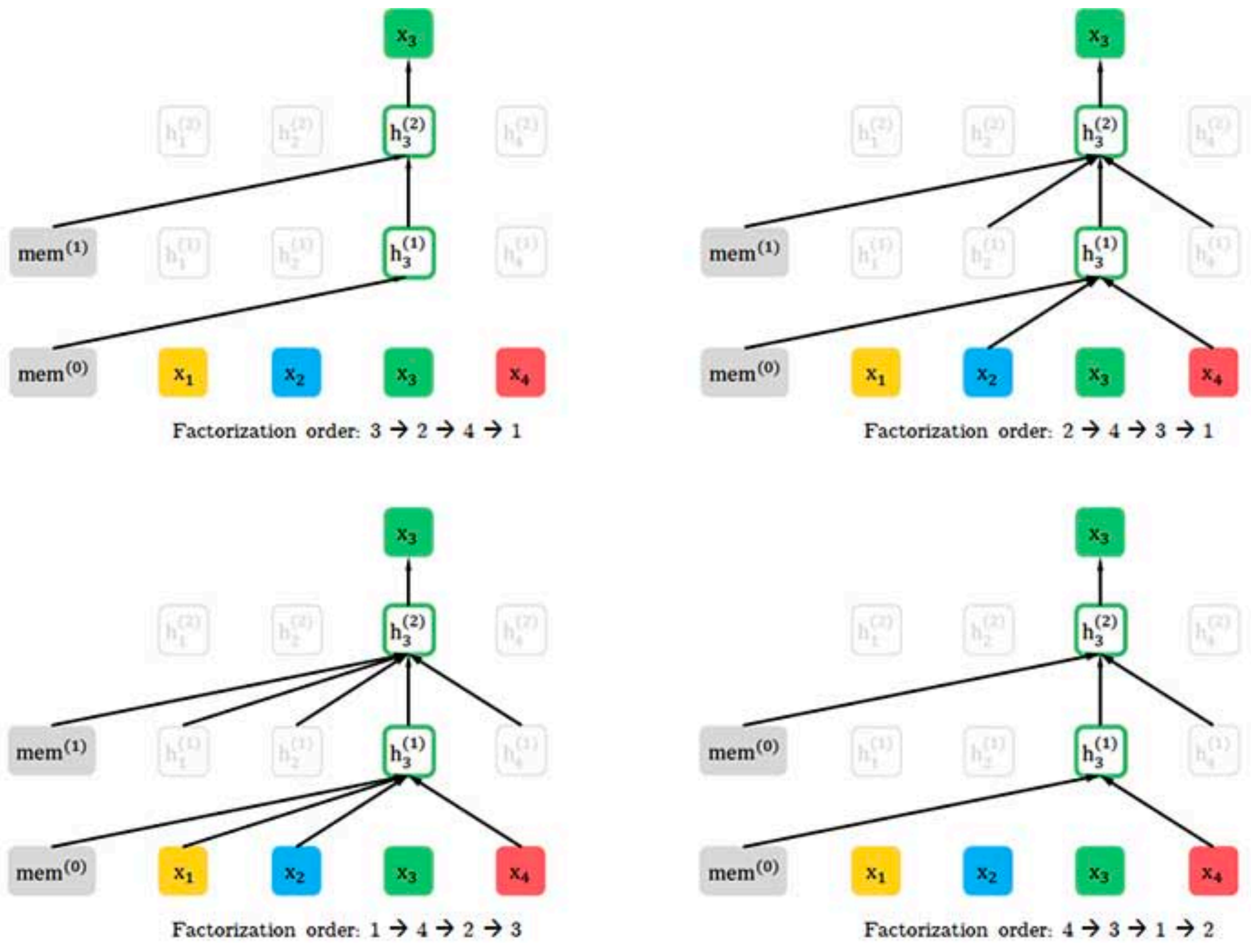


Figure 1: Illustration of the permutation language modeling objective for predicting x_3 given the same input sequence $\mathbf{x}$ but with different factorization orders.

Pretraining: Permuted Language Modeling

RoBERTa (Facebook, 2019)

What's the core idea of this paper?

- The Facebook AI research team found that BERT was significantly undertrained and suggested an improved recipe for its training, called RoBERTa:
    - More data: 160GB of text instead of the 16GB dataset originally used to train BERT.
    - Longer training: increasing the number of iterations from 100K to 300K and then further to 500K.
    - Larger batches: 8K instead of 256 in the original BERT base model.
    - Larger byte-level BPE vocabulary with 50K subword units instead of character-level BPE vocabulary of size 30K.
    - Removing the next sequence prediction objective from the training procedure.
    - Dynamically changing the masking pattern applied to the training data.

What's the key achievement?

- RoBERTa outperforms BERT in all individual tasks on the General Language Understanding Evaluation (GLUE) benchmark.
- The new model matches the recently introduced XLNet model on the GLUE benchmark and sets a new state of the art in four out of nine individual tasks.

RoBERTa (Facebook, 2019)

	BERT	RoBERTa	DistilBERT	XLNet
Size (millions)	**Base**: 110 **Large**: 340	**Base**: 110 **Large**: 340	**Base**: 66	**Base**: ~110 **Large**: ~340
Training Time	**Base**: 8 x V100 x 12 days* **Large:** 64 TPU Chips x 4 days (or 280 x V100 x 1 days*)	**Large:** 1024 x V100 x 1 day; 4-5 times more than BERT.	**Base**: 8 x V100 x 3.5 days; 4 times less than BERT.	**Large:** 512 TPU Chips x 2.5 days; 5 times more than BERT.
Performance	Outperforms state-of-the-art in Oct 2018	2-20% improvement over BERT	3% degradation from BERT	2-15% improvement over BERT
Data	16 GB BERT data (Books Corpus + Wikipedia). 3.3 Billion words.	160 GB (16 GB BERT data + 144 GB additional)	16 GB BERT data. 3.3 Billion words.	**Base**: 16 GB BERT data **Large**: 113 GB (16 GB BERT data + 97 GB additional). 33 Billion words.
Method	BERT (Bidirectional Transformer with MLM and NSP)	BERT without NSP**	BERT Distillation	Bidirectional Transformer with Permutation based modeling

DeBERTa (Microsoft 2020)

Disentangled Attention

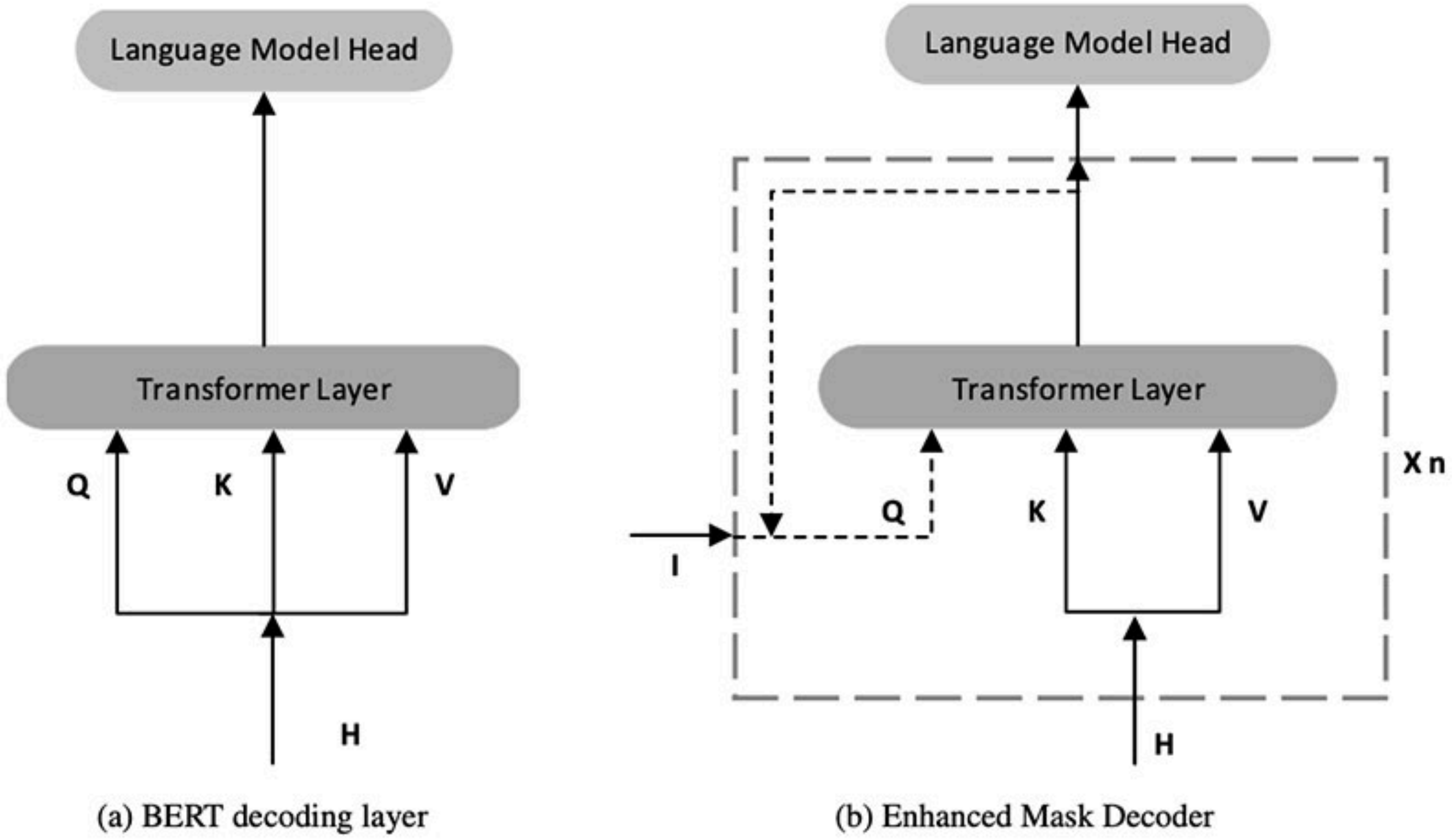


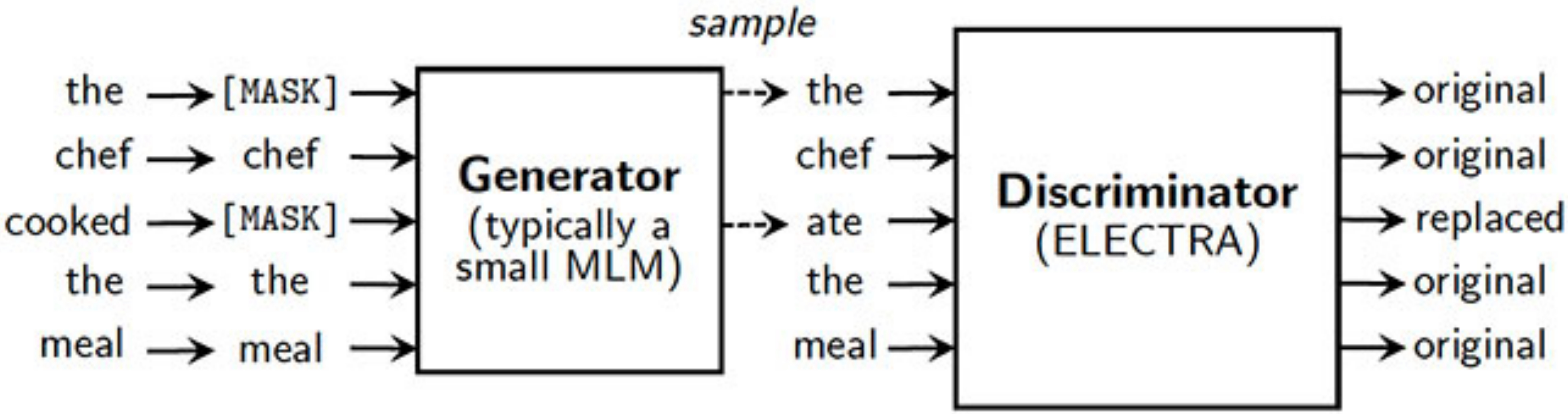
Figure 2: Comparison of the decoding layer.

What's the key achievement?

- Compared to the current state-of-the-art method RoBERTa-Large, the DeBERTA model trained on half the training data achieves:
    - an improvement of +0.9% in accuracy on MNLI (91.1% vs. 90.2%),
    - an improvement of +2.3% in accuracy on SQuAD v2.0 (90.7% vs. 88.4%),
    - an improvement of +3.6% in accuracy on RACE (86.8% vs. 83.2%)
- A single scaled-up variant of DeBERTa surpasses the human baseline on the SuperGLUE benchmark for the first time (89.9 vs. 89.8). The ensemble DeBERTa is the top-performing method on SuperGLUE at the time of this publication, outperforming the human baseline by a decent margin (90.3 versus 89.8).

ELECTRA (Stanford, 2020)

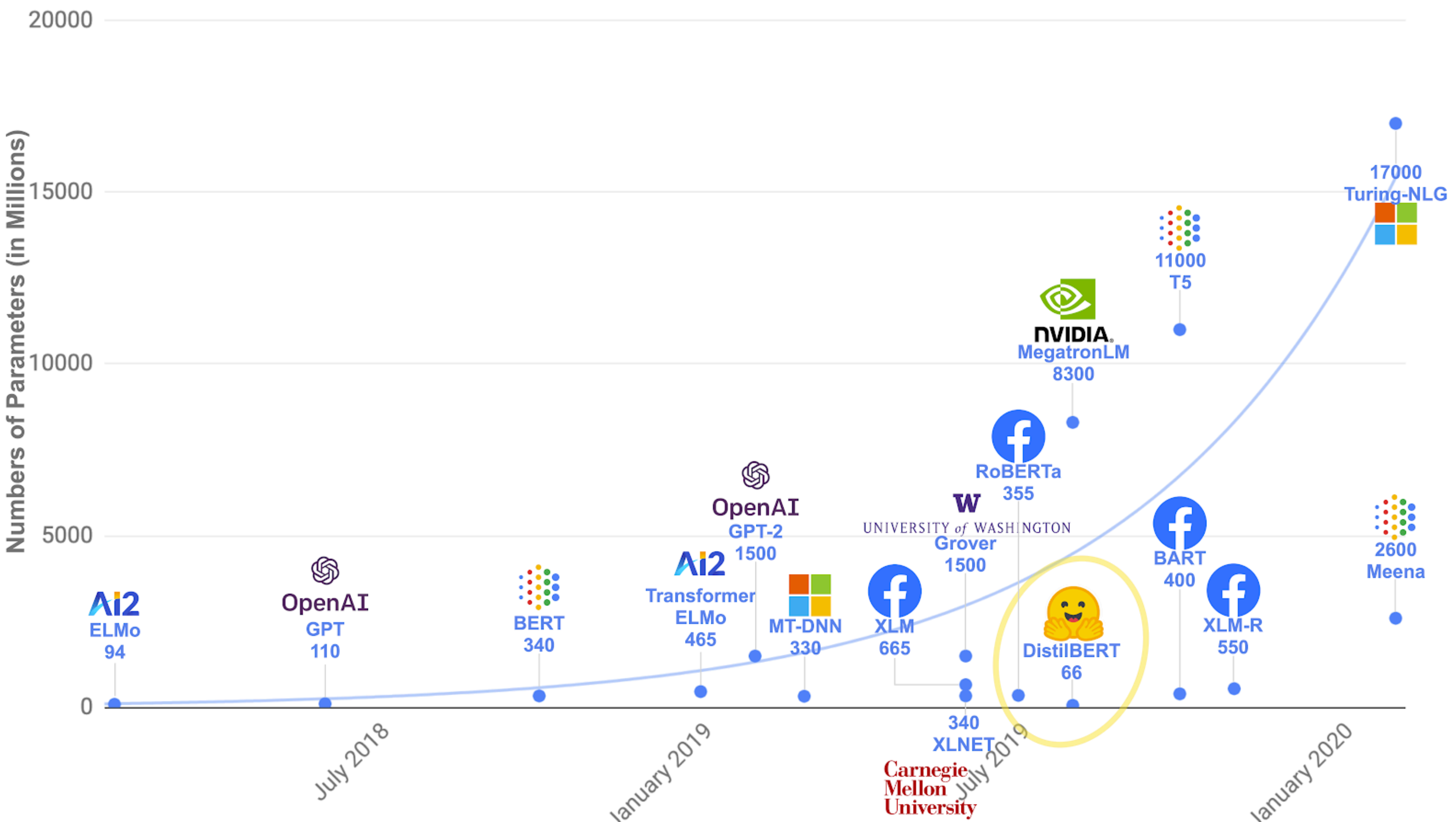
Training objective. Replaced token detection.



Outperforms BERT w/ similar parameters,
matches RoBERTa and XLNet w/ 25% compute.

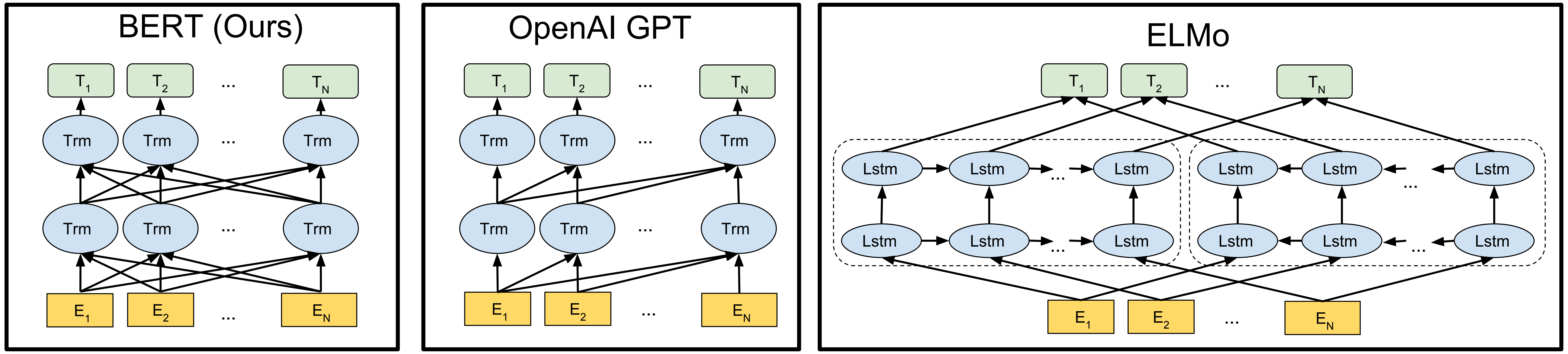
Arms Race

GPT-3 - 175000!



Overview / Taxonomy

Pretraining Model Architectures

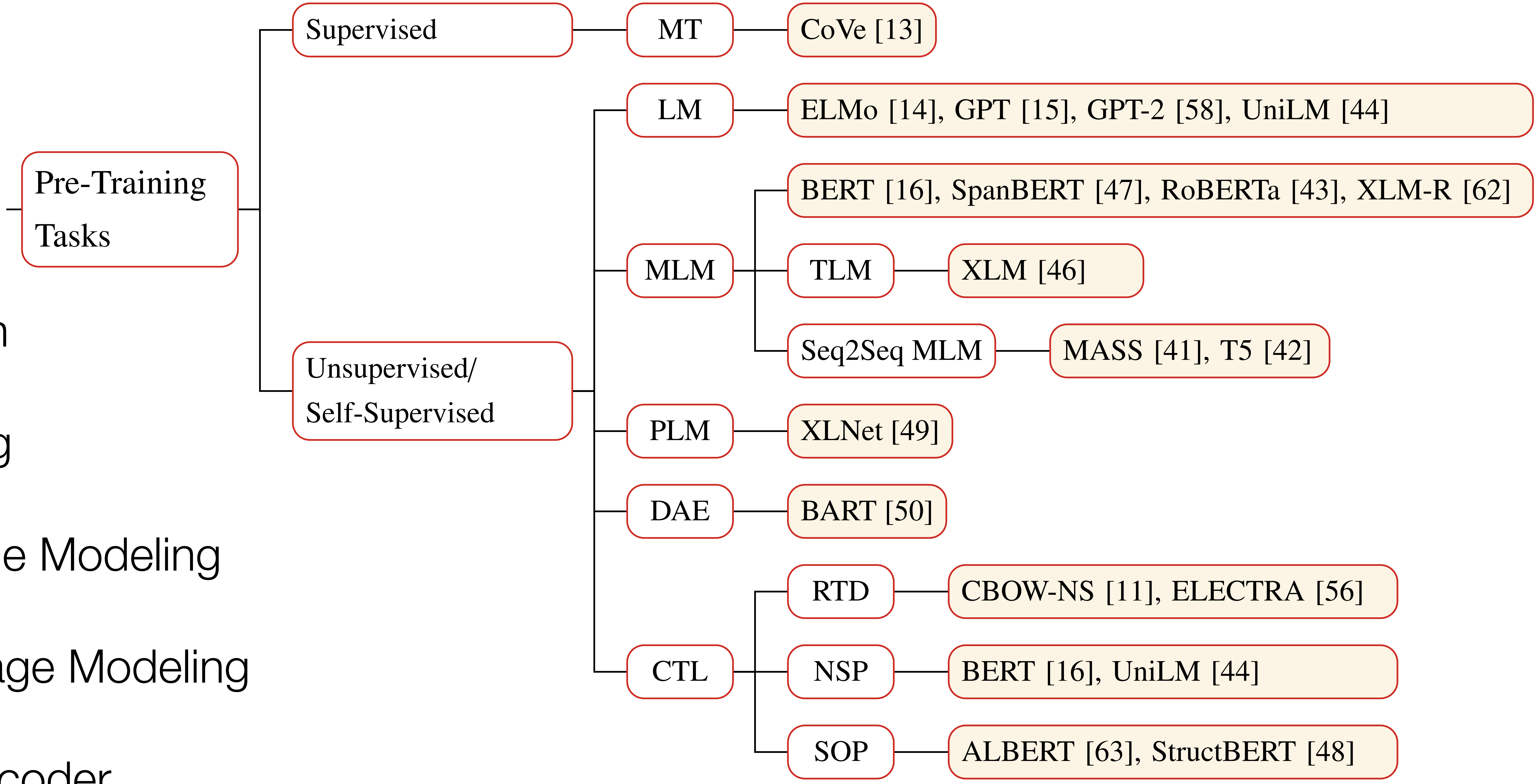


- Architectures
  - LSTM — LM-LSTM [35], Shared LSTM[5], ELMo [14], CoVe [13]
  - Transformer Enc. — BERT [16], SpanBERT [47], XLNet [49], RoBERTa [43]
  - Transformer Dec. — GPT [15], GPT-2 [58],GPT-3 [59]
  - Transformer — MASS [41], BART [50],T5 [42], XNLG [60], mBART [61]

Pre-trained Models for Natural Language Processing: A Survey

Xipeng Qiu*, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai & Xuanjing Huang

Pretraining Tasks



MT: Machine Translation

LM: Language Modeling

MLM: Masked Language Modeling

PLM: Permuted Language Modeling

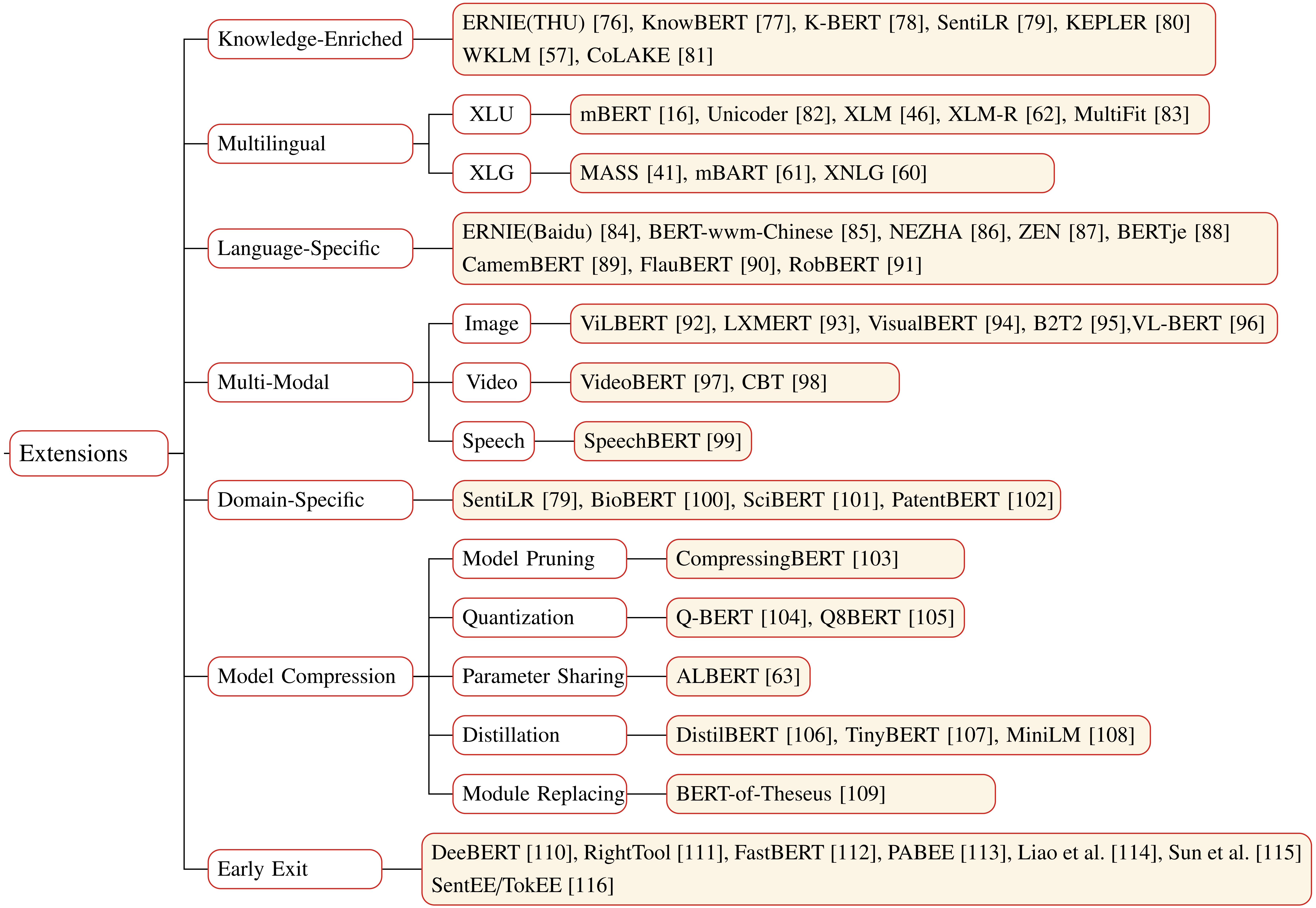
DAE: Denoising Autoencoder

CTL: Contrastive Learning
- RTD: Replaced Token Detection
- NSP: Next Sentence Prediction
- SOP: Sentence Order Prediction

Pre-trained Models for Natural Language Processing: A Survey

Xipeng Qiu*, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai & Xuanjing Huang

Variants



Pre-trained Models for Natural Language Processing: A Survey

Xipeng Qiu*, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai & Xuanjing Huang

Benchmarks / leaderboards

GLUE Benchmarks - 9 Language Understanding Tasks (NYU)

Dataset	Description	Data example	Metric
CoLA	Is the sentence grammatical or ungrammatical?	"This building is than that one." = **Ungrammatical**	Matthews
SST-2	Is the movie review positive, negative, or neutral?	"The movie is funny , smart , visually inventive , and most of all , alive ." = **.93056 (Very Positive)**	Accuracy
MRPC	Is the sentence B a paraphrase of sentence A?	A) "Yesterday , Taiwan reported 35 new infections , bringing the total number of cases to 418 ." B) "The island reported another 35 probable cases yesterday , taking its total to 418 ." = **A Paraphrase**	Accuracy / F1
STS-B	How similar are sentences A and B?	A) "Elephants are walking down a trail." B) "A herd of elephants are walking along a trail." = **4.6 (Very Similar)**	Pearson / Spearman
QQP	Are the two questions similar?	A) "How can I increase the speed of my internet connection while using a VPN?" B) "How can Internet speed be increased by hacking through DNS?" = **Not Similar**	Accuracy / F1
MNLI-mm	Does sentence A entail or contradict sentence B?	A) "Tourist Information offices can be very helpful." B) "Tourist Information offices are never of any help." = **Contradiction**	Accuracy
QNLI	Does sentence B contain the answer to the question in sentence A?	A) "What is essential for the mating of the elements that create radio waves?" B) "Antennas are required by any radio receiver or transmitter to couple its electrical connection to the electromagnetic field." = **Answerable**	Accuracy
RTE	Does sentence A entail sentence B?	A) "In 2003, Yunus brought the microcredit revolution to the streets of Bangladesh to support more than 50,000 beggars, whom the Grameen Bank respectfully calls Struggling Members." B) "Yunus supported more than 50,000 Struggling Members." = **Entailed**	Accuracy
WNLI	Sentence B replaces sentence A's ambiguous pronoun with one of the nouns - is this the correct noun?	A) "Lily spoke to Donna, breaking her concentration." B) "Lily spoke to Donna, breaking Lily's concentration." = **Incorrect Referent**	Accuracy

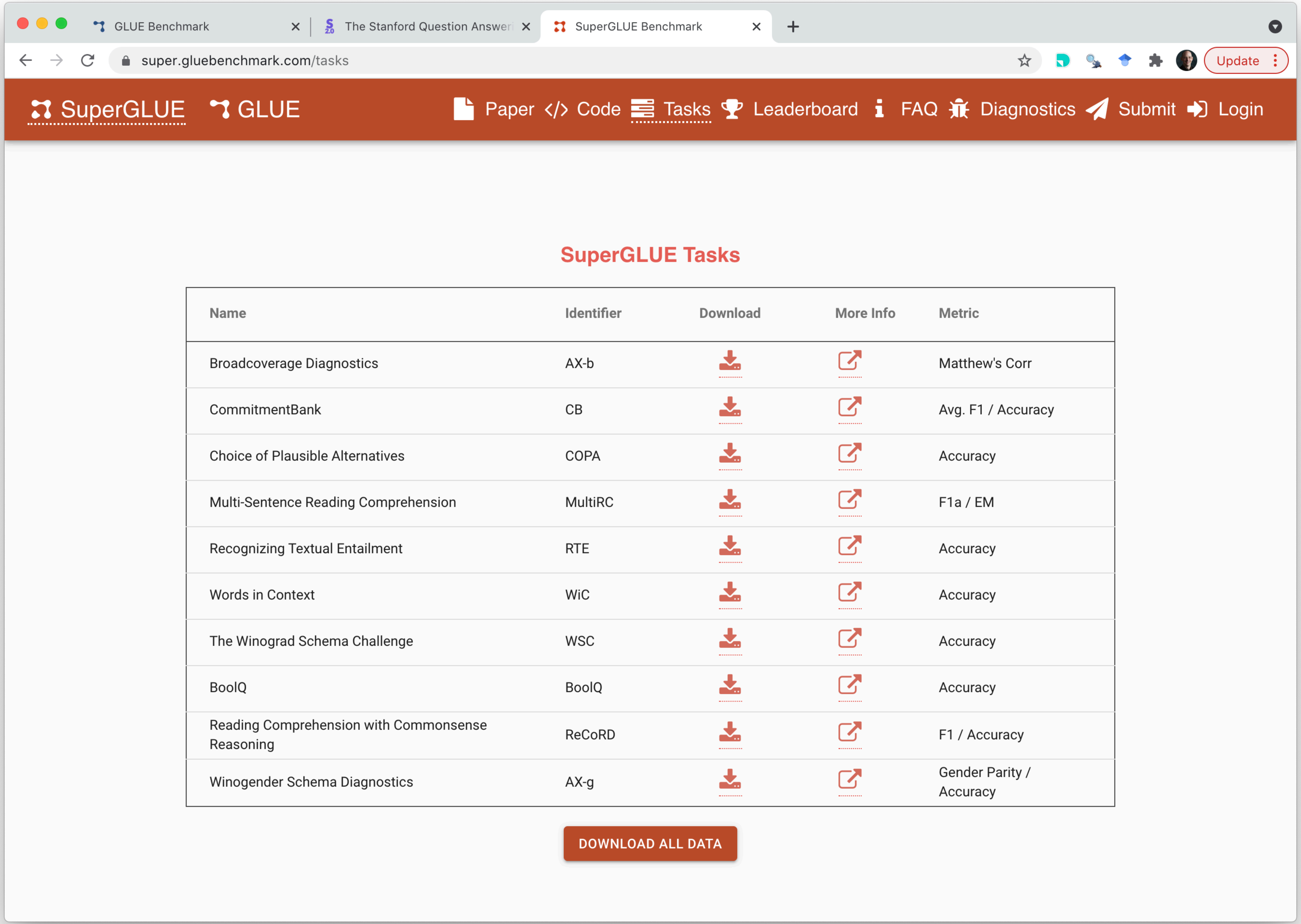
GLUE Benchmark

gluebenchmark.com/leaderboard

Update

Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RT
1	AliceMind & DIRL	StructBERT + CLEVER		91.0	75.3	97.7	93.9/91.9	93.5/93.1	75.6/90.8	91.7	91.5	97.4	92.
2	ERNIE Team - Baidu	ERNIE		90.9	74.4	97.8	93.9/91.8	93.0/92.6	75.2/90.9	91.9	91.4	97.3	92.
3	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.8	71.5	97.5	94.0/92.0	92.9/92.6	76.2/90.8	91.9	91.6	99.2	93.
4	HFL iFLYTEK	MacALBERT + DKM		90.7	74.8	97.0	94.5/92.6	92.8/92.6	74.7/90.6	91.3	91.1	97.8	92.
5	PING-AN Omni-Sinitic	ALBERT + DAAF + NAS		90.6	73.5	97.2	94.0/92.0	93.0/92.4	76.1/91.0	91.6	91.3	97.5	91.
6	liangzhu ge	Deberta + adv (ensemble)		90.4	72.7	97.3	92.7/90.3	93.2/92.9	75.6/90.8	91.7	91.5	96.4	92.
7	T5 Team - Google	T5		90.3	71.6	97.5	92.8/90.4	93.1/92.8	75.1/90.6	92.2	91.9	96.9	92.
8	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2	91.0	90.8	99.2	89.
9	Huawei Noah's Ark Lab	NEZHA-Large		89.8	71.7	97.3	93.3/91.0	92.4/91.9	75.2/90.7	91.5	91.3	96.2	90.
10	Zihang Dai	Funnel-Transformer (Ensemble B10-10-10H1024)		89.7	70.5	97.5	93.4/91.2	92.6/92.3	75.4/90.7	91.4	91.1	95.8	90.
11	ELECTRA Team	ELECTRA-Large + Standard Tricks		89.4	71.7	97.1	93.1/90.7	92.9/92.5	75.6/90.8	91.3	90.8	95.8	89.
12	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.4	68.0	96.8	93.1/90.8	92.3/92.1	74.8/90.3	91.1	90.7	95.6	88.
13	Junjie Yang	HIRE-RoBERTa		88.3	68.6	97.1	93.0/90.7	92.4/92.0	74.3/90.2	90.7	90.4	95.5	87.
14	Facebook AI	RoBERTa		88.1	67.8	96.7	92.3/89.8	92.2/91.9	74.3/90.2	90.8	90.2	95.4	88.
15	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6	68.4	96.5	92.7/90.3	91.1/90.7	73.7/89.9	87.9	87.4	96.0	86.
16	GLUE Human Baselines	GLUE Human Baselines		87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.
17	Adrian de Wynter	Bort (Alexa AI)		[illegible]	63.9	96.2	94.1/92.3	89.2/88.3	66.0/85.9	88.1	87.8	92.3	82.

Click on a submission to see more information



SuperGLUE Tasks

Name	Identifier	Download	More Info	Metric
Broadcoverage Diagnostics	AX-b			Matthew's Corr
CommitmentBank	CB			Avg. F1 / Accuracy
Choice of Plausible Alternatives	COPA			Accuracy
Multi-Sentence Reading Comprehension	MultiRC			F1a / EM
Recognizing Textual Entailment	RTE			Accuracy
Words in Context	WiC			Accuracy
The Winograd Schema Challenge	WSC			Accuracy
BoolQ	BoolQ			Accuracy
Reading Comprehension with Commonsense Reasoning	ReCoRD			F1 / Accuracy
Winogender Schema Diagnostics	AX-g			Gender Parity / Accuracy

DOWNLOAD ALL DATA

GLUE Benchmark | The Stanford Question Answ… | SuperGLUE Benchmark | SuperGLUE Benchmark | RACE (Reading Comprehens…

super.gluebenchmark.com/leaderboard

Update

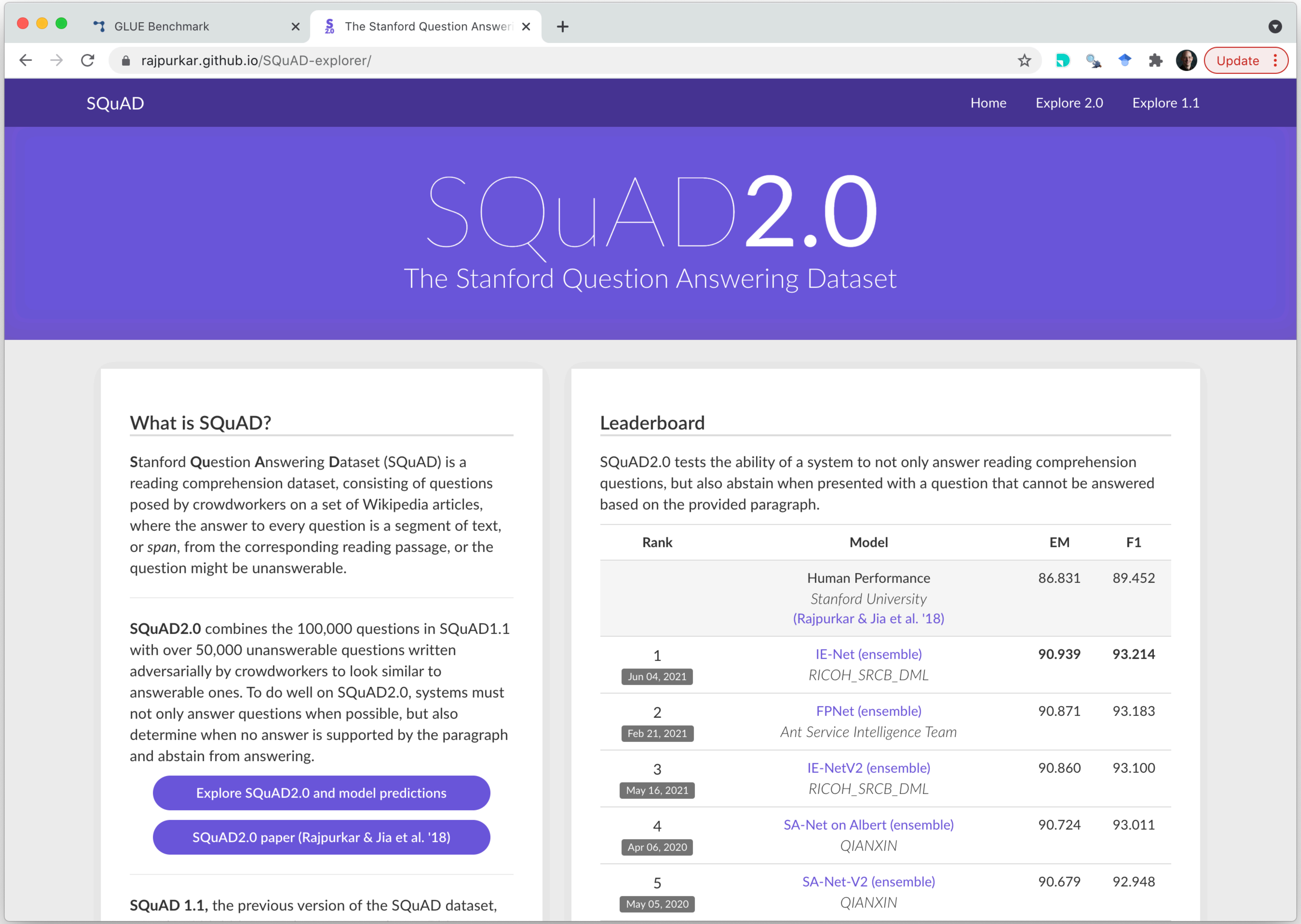
SuperGLUE GLUE Paper Code Tasks Leaderboard FAQ Diagnostics Submit Login

Leaderboard Version: **2.0**

	Rank	Name	Model	URL	Score	BoolQ	CB	COPA	MultiRC	ReCoRD	RTE	WiC	WSC	AX-b	AX-g
	1	ERNIE Team - Baidu	ERNIE 3.0		90.6	91.0	98.6/99.2	97.4	88.6/63.2	94.7/94.2	92.6	77.4	97.3	68.6	92.7/94.7
+	2	Zirui Wang	T5 + Meena, Single Model (Meena Team - Google Brain)		90.4	91.4	95.8/97.6	98.0	88.3/63.0	94.2/93.5	93.0	77.9	96.6	69.1	92.7/91.9
+	3	DeBERTa Team - Microsoft	DeBERTa / TuringNLRv4		90.3	90.4	95.7/97.6	98.4	88.2/63.7	94.5/94.1	93.2	77.5	95.9	66.7	93.3/93.8
	4	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8	89.0	95.8/98.9	100.0	81.8/51.9	91.7/91.3	93.6	80.0	100.0	76.6	99.3/99.7
+	5	T5 Team - Google	T5		89.3	91.2	93.9/96.8	94.8	88.1/63.3	94.1/93.4	92.5	76.9	93.8	65.6	92.7/91.9
+	6	Huawei Noah's Ark Lab	NEZHA-Plus		86.7	87.8	94.4/96.0	93.6	84.6/55.1	90.1/89.6	89.1	74.6	93.2	58.0	87.1/74.4
+	7	Alibaba PAI&ICBU	PAI Albert		86.1	88.1	92.4/96.4	91.8	84.6/54.7	89.0/88.3	88.8	74.1	93.2	75.6	98.3/99.2
+	8	Infosys : DAWN : AI Research	RoBERTa-iCETS		86.0	88.5	93.2/95.2	91.2	86.4/58.2	89.9/89.3	89.9	72.9	89.0	61.8	88.8/81.5
+	9	Tencent Jarvis Lab	RoBERTa (ensemble)		85.9	88.2	92.5/95.6	90.8	84.4/53.4	91.5/91.0	87.9	74.1	91.8	57.6	89.3/75.6
	10	Zhuiyi Technology	RoBERTa-mtl-adv		85.7	87.1	92.4/95.6	91.2	85.1/54.3	91.7/91.3	88.1	72.1	91.8	58.5	91.0/78.1
	11	Facebook AI	RoBERTa		84.6	87.1	90.5/95.2	90.6	84.4/52.5	90.6/90.0	88.2	69.9	89.0	57.9	91.0/78.1
+	12	Anuar Sharafudinov	AILabs Team, Transformers		82.6	88.1	91.6/94.8	86.8	85.1/54.7	82.8/79.8	88.9	74.1	78.8	100.0	100.0/100.0
	13	Rakesh Radhakrishnan Menon	ADAPET (ALBERT) - few-shot		[illegible]	[illegible]	[illegible]	85.4	76.2/35.7	86.1/85.5	75.0	53.5	85.6	-0.4	100.0/50.0

Click on a submission to see more information

SQuAD benchmark



GLUE Benchmark | The Stanford Question Answer

rajpurkar.github.io/SQuAD-explorer/

Update

SQuAD | Home | Explore 2.0 | Explore 1.1

SQuAD2.0

The Stanford Question Answering Dataset

What is SQuAD?

Stanford **Qu**estion **A**nswering **D**ataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage, or the question might be unanswerable.

SQuAD2.0 combines the 100,000 questions in SQuAD1.1 with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering.

Explore SQuAD2.0 and model predictions

SQuAD2.0 paper (Rajpurkar & Jia et al. '18)

SQuAD 1.1, the previous version of the SQuAD dataset,

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jun 04, 2021	IE-Net (ensemble) *RICOH_SRCB_DML*	**90.939**	**93.214**
2 Feb 21, 2021	FPNet (ensemble) *Ant Service Intelligence Team*	90.871	93.183
3 May 16, 2021	IE-NetV2 (ensemble) *RICOH_SRCB_DML*	90.860	93.100
4 Apr 06, 2020	SA-Net on Albert (ensemble) *QIANXIN*	90.724	93.011
5 May 05, 2020	SA-Net-V2 (ensemble) *QIANXIN*	90.679	92.948

RACE Benchmark

GLUE Benchmark | The Stanford Question Answeri | SuperGLUE Benchmark | RACE (Reading Comprehension

Not Secure | qizhexie.com/data/RACE_leaderboard.html

Update

RACE Reading Comprehension Dataset

The RACE dataset is a large-scale **R**e**A**ding **C**omprehension dataset collected from English **E**xaminations that are created for middle school and high school students.

Report your results: If you have new results, please send Qizhe (qizhex@cs.cmu.edu) or Guokun (guokun@cs.cmu.edu) an email with the link to your paper!

Leaderboard

Model	Report Time	Institute	RACE	RACE-M	RACE-H
Human Ceiling Performance	Apr 15, 2017	CMU	*94.5*	*95.4*	*94.2*
Amazon Mechanical Turker	Apr 15, 2017	CMU	73.3	85.1	69.4
ALBERT-SingleChoice + transfer learning (ensemble)	Nov 06, 2020	Tencent Cloud Xiaowei & Tencent Cloud TI-ONE	**91.4**	**93.6**	**90.5**
Megatron-BERT (ensemble)	Mar 13, 2020	NVIDIA Research	90.9	93.1	90.0
ALBERT-SingleChoice + transfer learning	Nov 06, 2020	Tencent Cloud Xiaowei & Tencent Cloud TI-ONE	90.7	92.8	89.8
ALBERT + DUMA (ensemble)	Mar 18, 2020	SJTU & Huawei Noah's Ark Lab	89.8	92.6	88.7
Megatron-BERT	Mar 13, 2020	NVIDIA Research	89.5	91.8	88.6
ALBERT (ensemble)	Sep 26, 2019	Google Research & TTIC	89.4	91.2	88.6
UnifiedQA	May 02, 2020	AI2 & UW	89.4	-	-
ALBERT + DUMA	Feb 08, 2020	SJTU & Huawei Noah's Ark Lab	88.0	90.9	86.7
T5*	May 02, 2020	Google	87.1	-	-
ALBERT	Sep 26, 2019	Google Research & TTIC	86.5	89.0	85.5
RoBERTa + MMM	Oct 01, 2019	MIT & Amazon Alexa AI	85.0	89.1	83.3
[illegible]	Aug 30,	[illegible]	[illegible]	[illegible]	[illegible]

Sam - Text Classification with BERT