Applied Deep Learning

Prompt-Based Learning

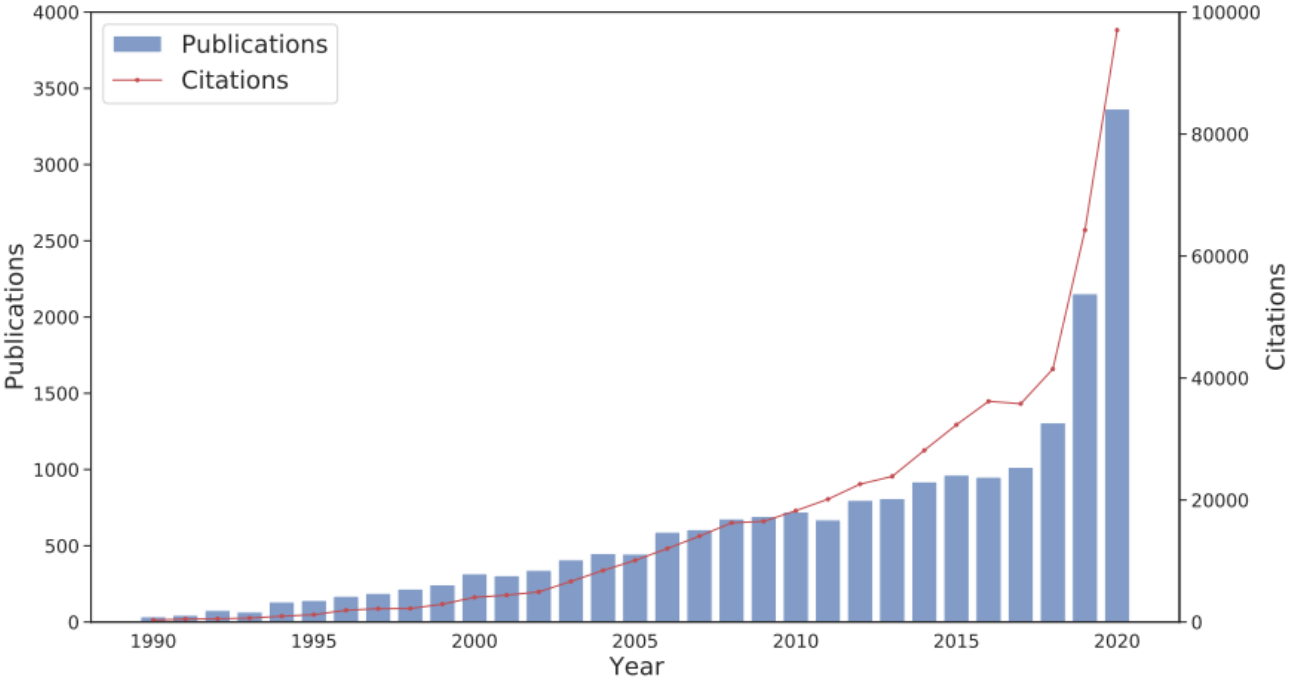
November 9th, 2023 **http://adl.miulab.tw**

National Taiwan University 國立臺灣大學

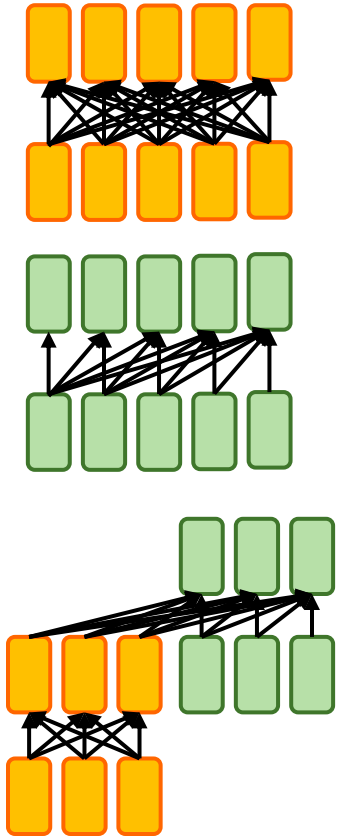
Slides credited from Mohit Lyyer (Umass) and Hung-Yi Lee

Wide Usage of PLMs (Han et al., 2021)

- Increasing usage of PLMs

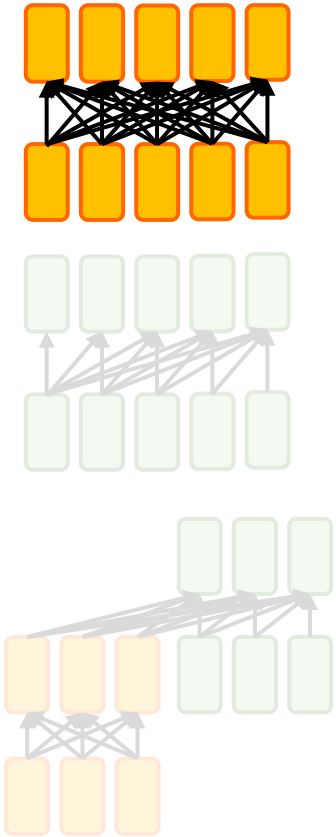


Three Types of Model Pre-Training



- Encoder
  - Bidirectional context
  - Examples: BERT and its variants
- Decoder
  - Language modeling; better for generation
  - Example: GPT, GPT-2, GPT-3
- Encoder-Decoder
  - Sequence-to-sequence model
  - Examples: Transformer, BART, T5

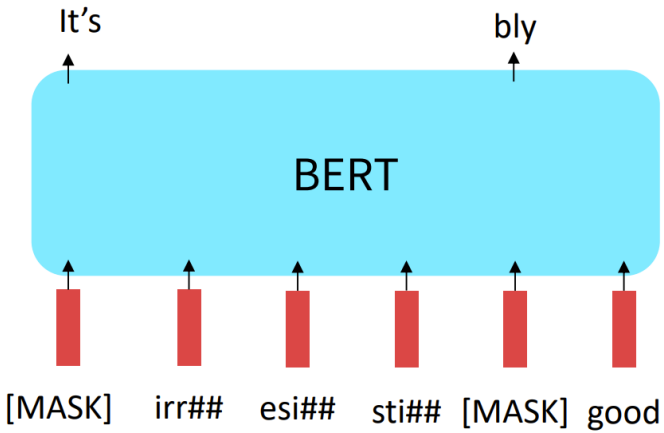
Three Types of Model Pre-Training

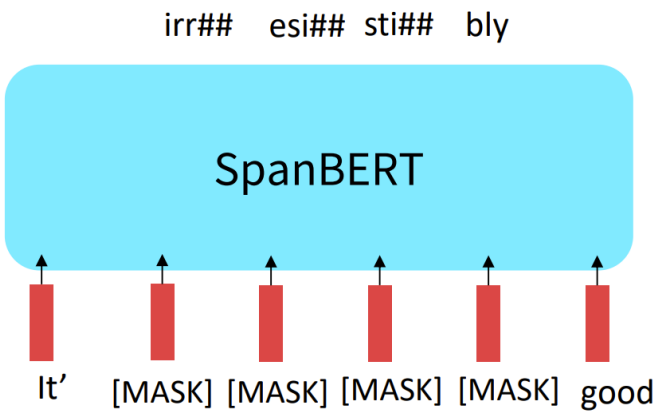


- Encoder
  - Bidirectional context
  - Examples: BERT and its variants
- Decoder
  - Language modeling; better for generation
  - Example: GPT, GPT-2, GPT-3
- Encoder-Decoder
  - Sequence-to-sequence model
  - Examples: Transformer, BART, T5

BERT Variants

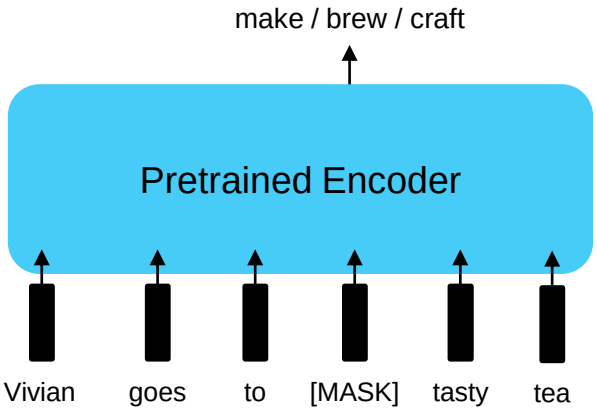
- Improvements to the BERT pretraining:
  - RoBERTa: mainly train BERT on *more data* and *longer*
  - SpanBERT: masking contiguous spans of words makes a harder, more useful pretraining task

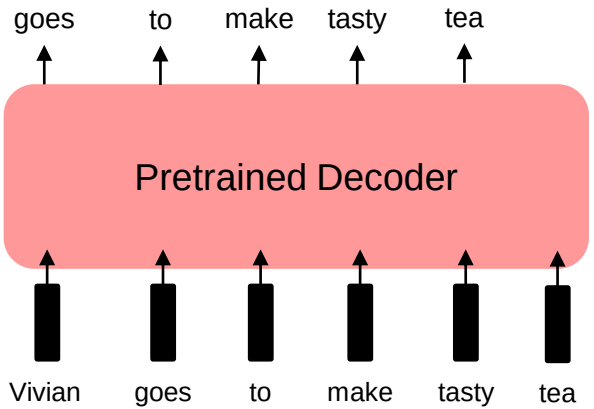




Need of Decoder

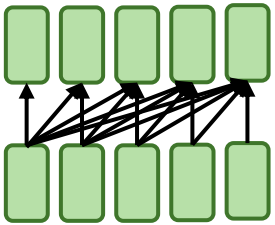
- Generation tasks
  - BERT and other pretrained encoders don't naturally lead to *autoregressive (1-word-at-a-time)* generation methods





Three Types of Model Pre-Training

- Encoder
  - Bidirectional context
  - Examples: BERT and its variants

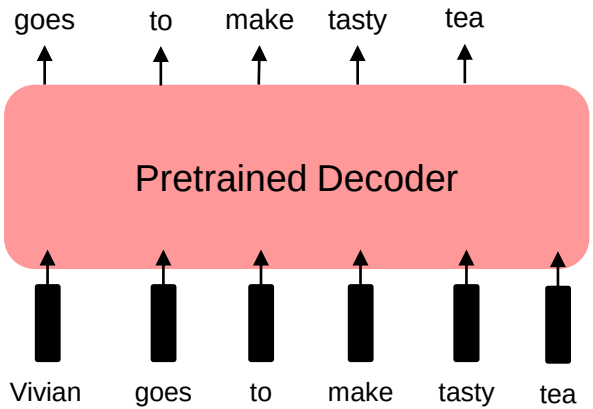


- **Decoder**
  - Language modeling; better for generation
  - Example: GPT, GPT-2, GPT-3

- Encoder-Decoder
  - Sequence-to-sequence model
  - Examples: Transformer, BART, T5

GPT: Generative Pretrained Transformer
(Radford et al., 2018)

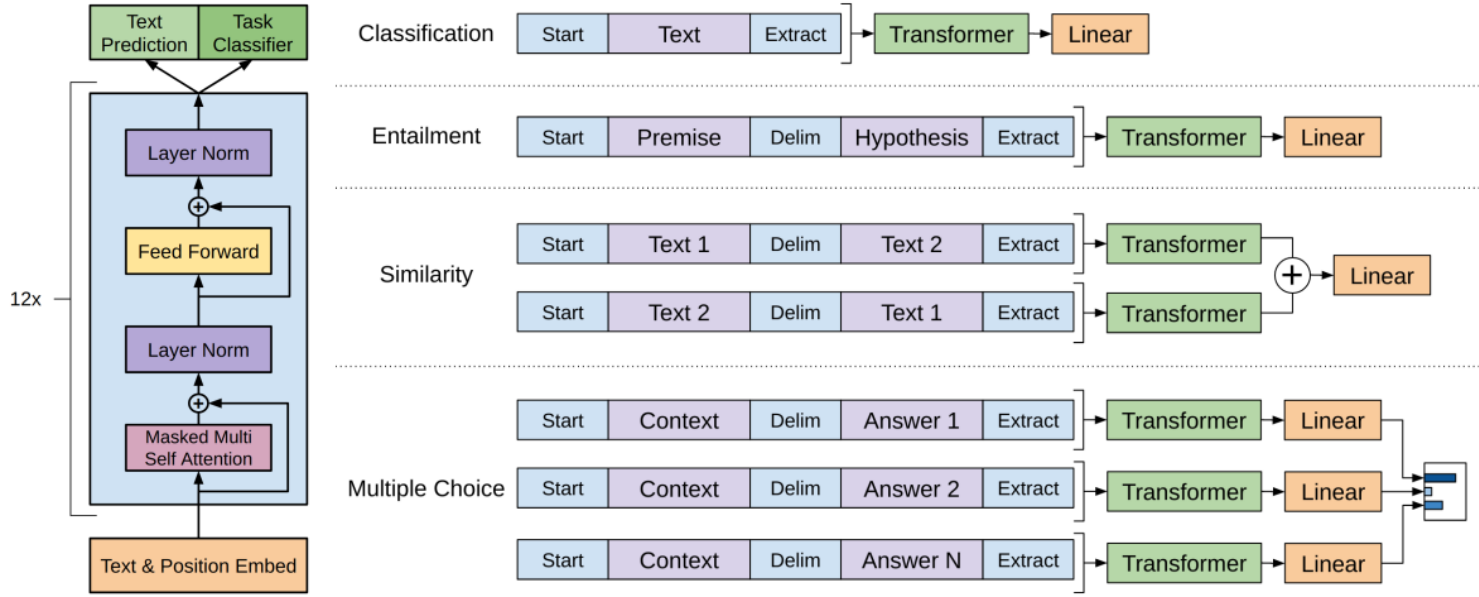
- Transformer decoder
  - Pre-trained on BooksCorpus (~7000 books; 5GB)
    - Transformer decoder with 12 layers
    - 768-dim hidden states, 3072-dim feed-forward hidden layers
    - BPE with 40,000 merges



Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. "Improving language understanding by generative pre-training." (2018).

GPT: Generative Pretrained Transformer
(Radford et al., 2018)

- Transformer decoder
  - Supervised fine-tuning for the target tasks
  - Next word prediction is kept during fine-tuning



GPT-2 (Radford et al., 2019)

- Transformer decoder
  - Pre-trained on *more data*
    - WebText from Raddit (40GB)
  - Good for NLG

Context (WebText test)

Outfit 1: Typical

This pairing was the first outfit I thought of when I bought the shoes. It's like a summer version of this Jake Grantham outfit; in fact, my shoes are close to the colors of his Nike Racers! Instead of a heavy Harris Tweed jacket and denim shirt, I'm wearing a cotton DB jacket and and a linen shirt. Both fabrics (in these colors) are an absolute must for summer, as they go with both dark and and light pants! As you can see, they pair wonderfully with the dark jeans and shoes. It's a pseudo menswear/prep outfit.

Overall, this is a very casual outfit which is why I paired my sneakers with it. I'm not about wearing a full wool suit with sneakers (as GQ shows a lot) but I'm definitely open to keeping things casual, like this cotton DB. Casual fabrics are key to pulling off your sneakers in a dressed down menswear outfit. I'd even suggest to wear these sneakers with a khaki chino suit or a white linen suit. Just be sure to ditch the tie or wear a tee or polo; wearing a tie with sneakers is a bit too much

Small model completion	**GPT-2 completion**
.	for me!
Outfit 2: High End	Outfit 2: Casual Jacket
This outfit was just in the eye of the storm. It's a hybrid of what ASICS does with streetwear (like J. Crew) and what J. Crew does with casual pieces that are a bit more traditional. Again, as I mentioned before, it's a low budget item and they can be very expensive. These will definitely be an issue if you're trying to work as a casual sneakermaker, and that's OK but don't hesitate if you want to try the lower price versions of their outfits!	I know that some people might be opposed to wearing sneakers in a turtleneck, but I wanted to be true to myself, so I went with a slim fitting turtleneck in a color more similar to my favorite color of the day. Overall, it's a classic turtleneck, with nothing too flashy. I did swap my shoes for these ASICS Tiger Killshots and I wish I hadn't did that. The shoe itself is very comfortable, however, I found that wearing them in these sneakers made my feet look a bit larger. I'd wear them with denim, chinos, whatever, but would never recommend wearing them alone. There are just too many variables involved.
This version features a very low cut-cut neckline (which is not a bad thing), slightly longer legs, and a pair of long legs to keep things nice. My preferred color of color these items are blue-gray (which is my first colors, as they are a light white), or black (which is my favorite of my designs). As always, my pick of colors from this store have nothing to do with that other than the de-sign. For a high end, I would recommend this to everyone!	Outfit 3: Dress
Outfit 3: High Light	This is the final outfit I wore the ASICS Tiger Killshots. This version of the ASICS is a much more modern and streamlined version. The shoe has the most modern and minimal design that I can think of! With a lot of black, these sneakers look very cool. They'll definitely be very comfortable for summer.
This was a "high light" look while I had a chance to get some color. I think the color is based on	To me, wearing them with a suit is the easiest way to have a good pair of sneaker shoes. That's pretty much it! If

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language models are unsupervised multitask learners." *OpenAI blog* 1, no. 8 (2019): 9.

GPT-3 (Brown et al., 2020)

- Transformer decoder
  - Pre-trained on *more data* (45TB)

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

- ***Common Crawl***: web data over 8 years (metadata & text with filtering)
- ***WebText2***: web pages from all outbound Reddit links from posts with 3+ upvotes
- ***Books1 & Books2***: internet-based books corpora
- ***Wikipedia***: English pages

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

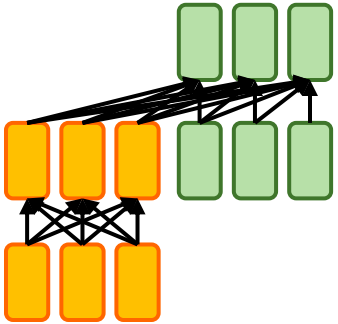
OpenAI GPT Paradigm

Model	#Parameters	Pre-Trained Data
GPT (Radford et al., 2018)	0.117 B	5GB
GPT-2 (Radford et al., 2019)	1.5 B	40GB
GPT-3 (Brown et al., 2020)	175 B	45TB
GPT-4 (OpenAI, 2023)	?	?

Three Types of Model Pre-Training

- Encoder
  - Bidirectional context
  - Examples: BERT and its variants

- Decoder
  - Language modeling; better for generation
  - Example: GPT, GPT-2, GPT-3

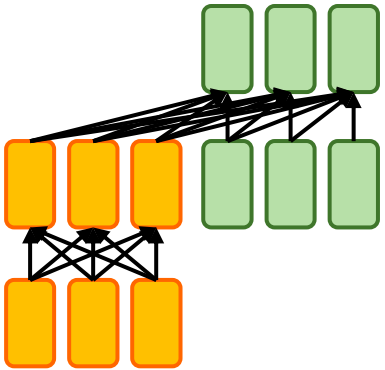


- Encoder-Decoder
  - Sequence-to-sequence model
  - Examples: Transformer, BART, T5

Encoder-Decoder Pre-Training

- The encoder portion benefits from bidirectional context; the decoder portion is used to train the whole model through language modeling.
- Pre-training objective: span corruption (denoising)
  - implemented in preprocessing
  - similar to language modeling at the decoder side

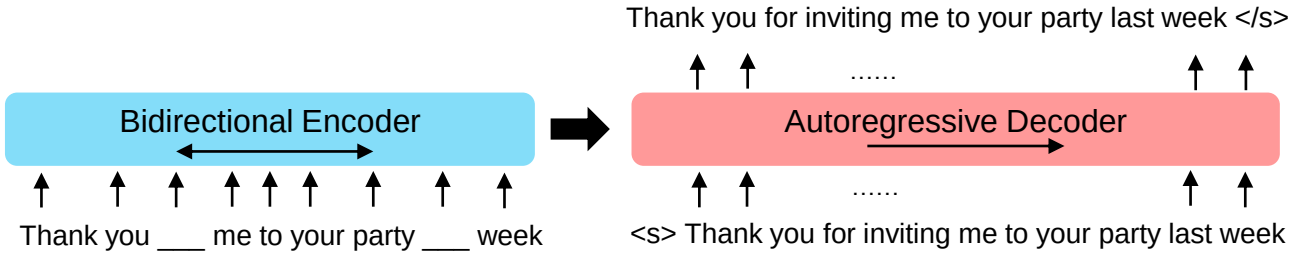
Thank you ~~for inviting~~ me to your party ~~last~~ week



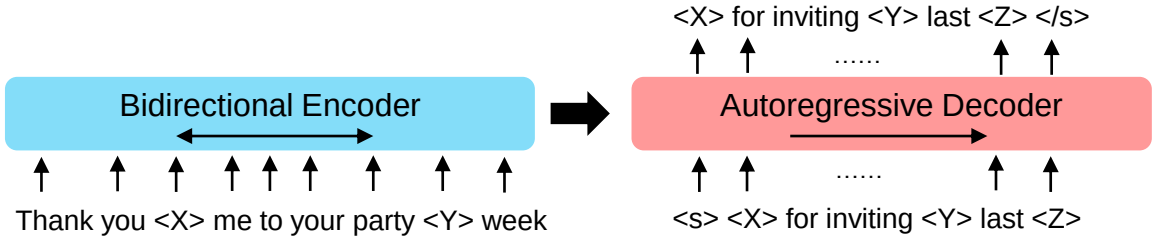
Denoising for Pre-Training

Thank you ~~for inviting~~ me to your party ~~last~~ week

- BART: output the whole sentence (Lewis et al., 2019)

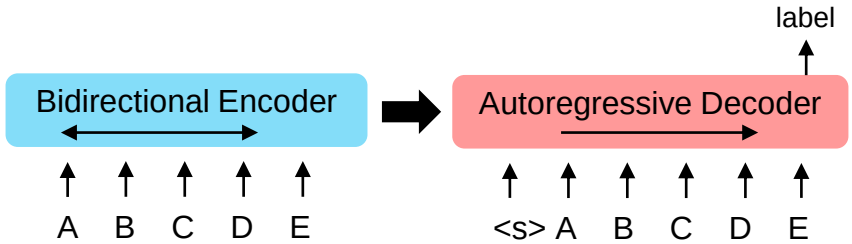


- T5: output the missing parts (Raffel et al., 2020)

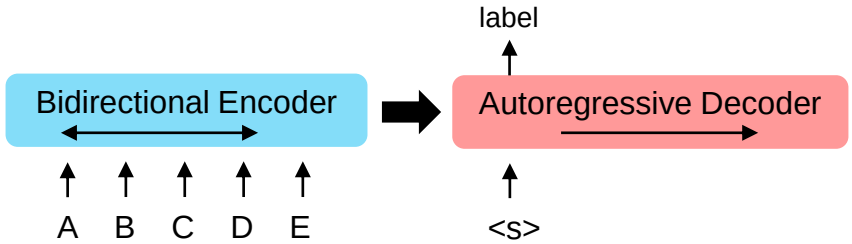


Fine-Tuning for Classification

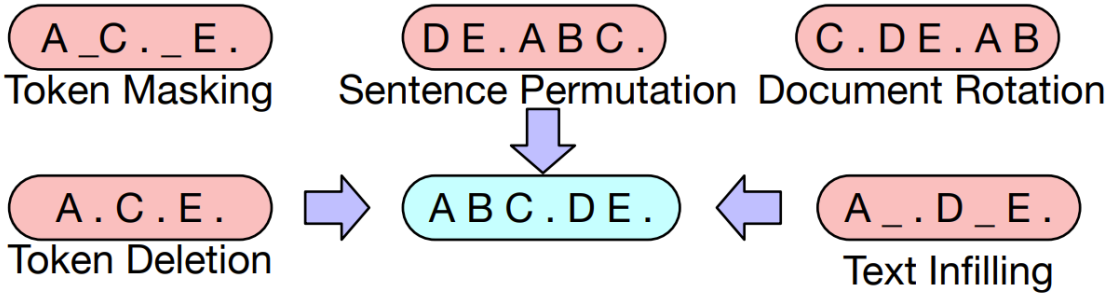
- BART: repeat input in decoder (Lewis et al., 2019)



- T5: treat it as a seq2seq task (Raffel et al., 2020)

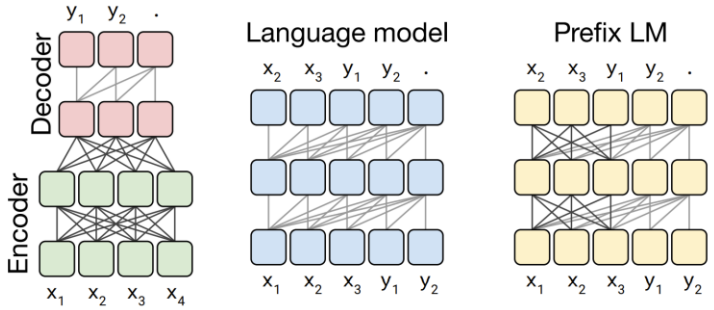


Diverse Noises in BART



Model	SQuAD 1.1 F1	MNLI Acc	ELI5 PPL	XSum PPL	ConvAI2 PPL	CNN/DM PPL
BERT Base (Devlin et al., 2019)	88.5	**84.3**	-	-	-	-
BART Base						
w/ Token Masking	90.4	84.1	25.05	7.08	11.73	6.10
w/ Token Deletion	90.4	84.1	24.61	6.90	11.46	5.87
w/ Text Infilling	**90.8**	84.0	24.26	**6.61**	**11.05**	5.83
w/ Document Rotation	77.2	75.3	53.69	17.14	19.87	10.59
w/ Sentence Shuffling	85.4	81.5	41.87	10.93	16.67	7.89
w/ Text Infilling + Sentence Shuffling	**90.8**	83.8	24.17	6.62	11.12	**5.41**

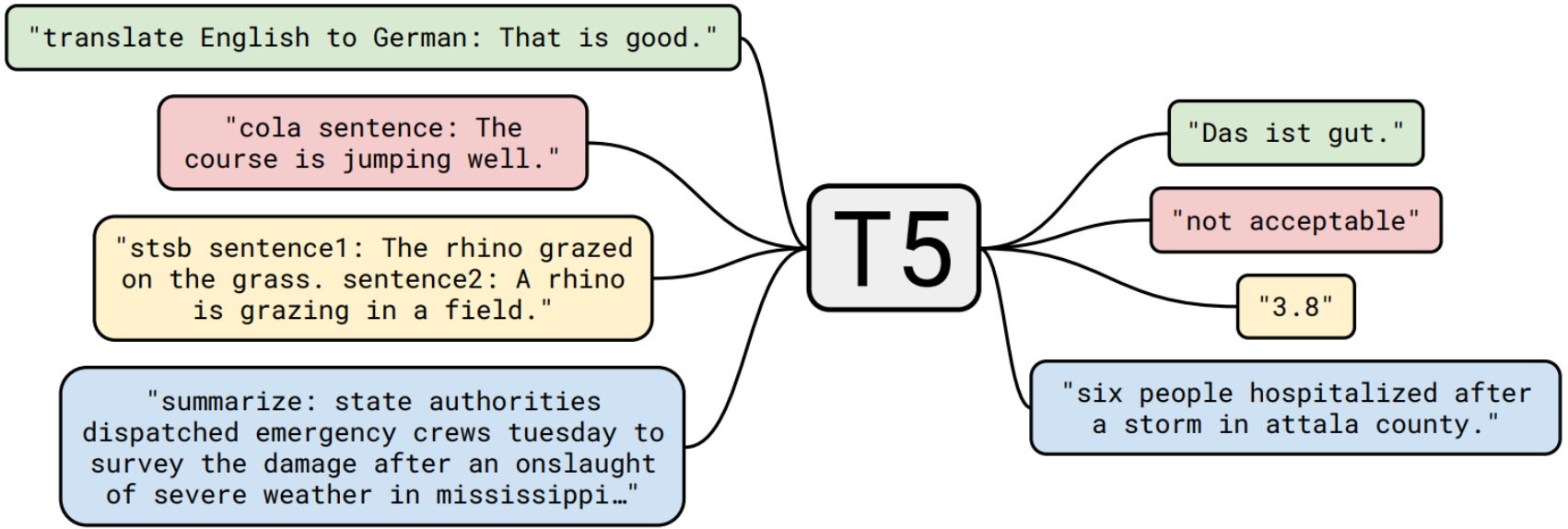
Effectiveness of Denoising in T5



Architecture	Objective	Params	Cost	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Encoder-decoder	Denoising	$2P$	M	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Enc-dec, shared	Denoising	P	M	82.81	18.78	**80.63**	**70.73**	26.72	39.03	**27.46**
Enc-dec, 6 layers	Denoising	P	$M/2$	80.88	18.97	77.59	68.42	26.38	38.40	26.95
Language model	Denoising	P	M	74.70	17.93	61.14	55.02	25.09	35.28	25.86
Prefix LM	Denoising	P	M	81.82	18.61	78.94	68.11	26.43	37.98	27.39
Encoder-decoder	LM	$2P$	M	79.56	18.59	76.02	64.29	26.27	39.17	26.86
Enc-dec, shared	LM	P	M	79.60	18.13	76.35	63.50	26.62	39.17	27.05
Enc-dec, 6 layers	LM	P	$M/2$	78.67	18.26	75.32	64.06	26.13	38.42	26.89
Language model	LM	P	M	73.78	17.54	53.81	56.51	25.23	34.31	25.38
Prefix LM	LM	P	M	79.68	17.84	76.87	64.86	26.28	37.51	26.76

T5: Text-to-Text Transfer Transformer

- **Multi-task pre-training**: learning multiple tasks via seq2seq



BART vs. T5

- Differences
  - Training data size: BART > T5 (about 2x)
  - Model size:
    - BART-large: 12 encoder, 12 decoder, 1024 hidden
    - T5-base: 12encoder, 12decoder, 768 hidden, 220M parameters (2x BERT-base)
    - T5-large: 24encoder, 24decoder, 1024hidden, 770M parameters
  - Position encoding: learnable absolute position (BART) & relative position (T5)
- Understanding performance

	SQuAD	MNLI	SST	QQP	QNLI	STS-B	RTE	MRPC	CoLA
BART	88.8 / 94.6	89.9 / 90.1	96.6	92.5	94.9	91.2	87.2	90.4	62.8
T5	86.7 / 93.8	89.9 / 89.6	96.3	89.9	94.8	89.9	87.0	89.9	61.2

- Generation performance (summarization)

CNN/DailyMail	ROUGE-1	ROUGE-2	ROUGE-3
BART	45.14	21.28	37.25
T5	42.50	20.68	39.75

Fine-Tuning on Pretrained LMs

- (Standard) fine-tuning: use the pre-trained LMs for initialization and tuning the parameters for a **downstream** task

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
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**

Issue 1: Data Scarcity

- Downstream annotated data may not be large

Task	MNLI	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE
Size	391K	363K	108K	67K	8.5K	5.7K	3.5K	2.5K

→ More practical cases are few-shot, one-shot or even zero-shot settings

Fine-Tuning vs. In-Context Learning

Pre-Training & Fine-Tuning

unannotated data

pre-train

Model

task-specific annotated data

fine-tune

Model

fine-tune

Model

Pre-Training & In-Context Learning

pre-train

Model

no learning

Model

no learning

Model

GPT-3 “In-Context” Learning

題組一：詞彙與結構
本部分共15題，每題含一個空格。請就試題中 A、B、C、D 四個選項中選出最適合題意的字或詞。

題型說明

例：
It’s eight o’clock now. Sue _____ in her bedroom.
A. study
B. studies
C. studied
D. is studying

正確答案為D。

少數範例

GPT-3 “In-Context” Learning

- Zero-Shot

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

- One-Shot

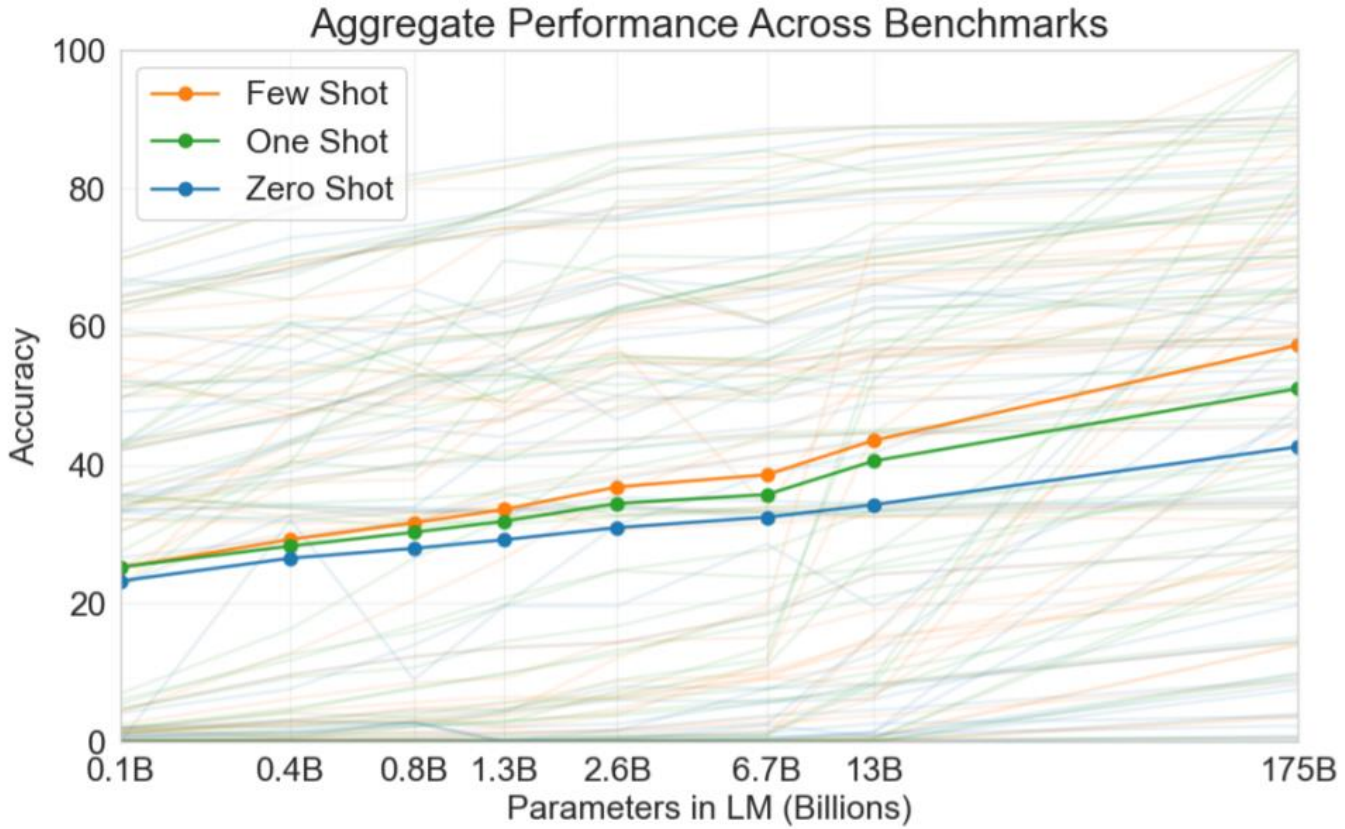
```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

- Few-Shot

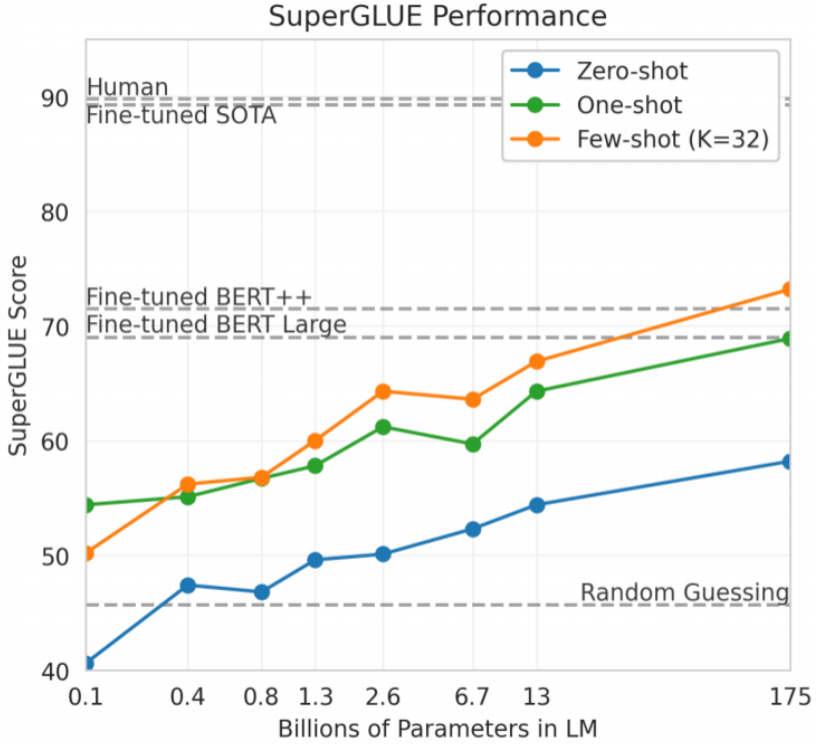
```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```

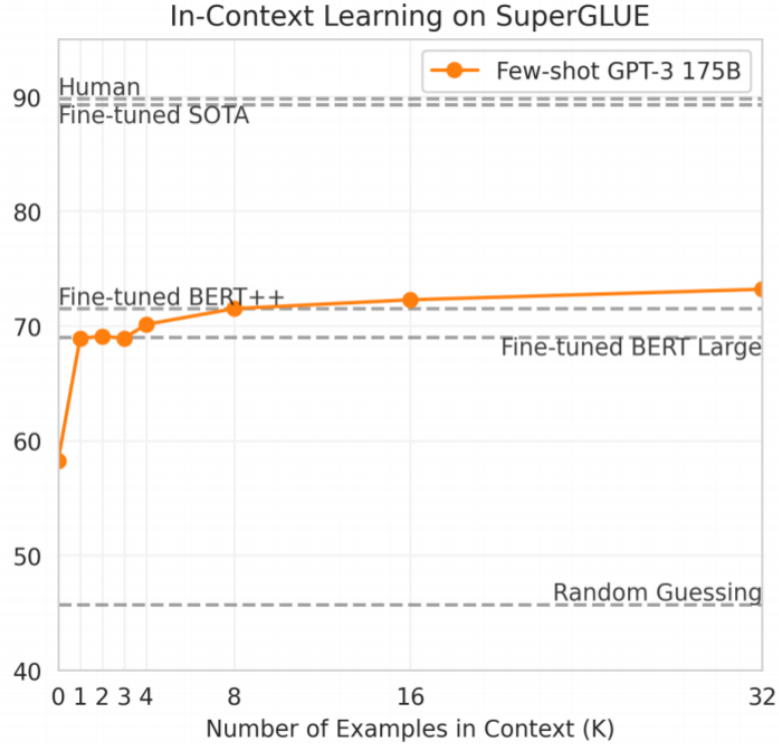
- Traditional Fine-Tuning

Benchmark 42 NLU Tasks



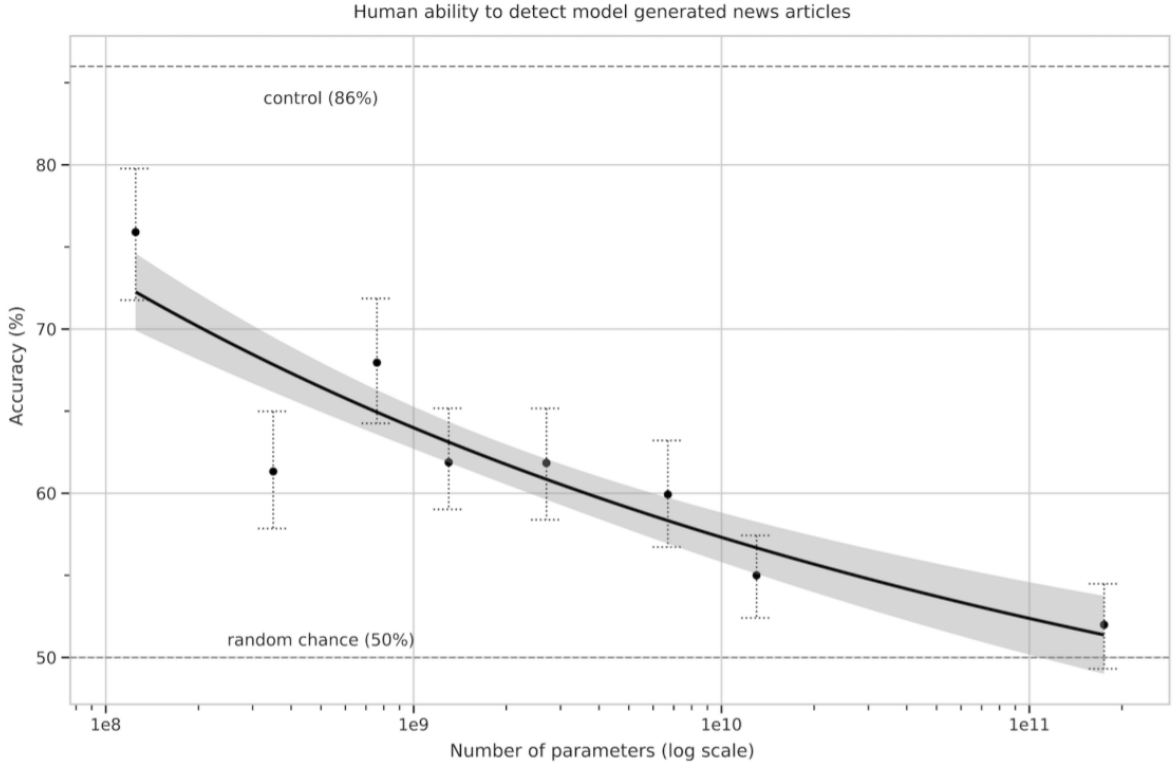
NLU Performance in SuperGLUE





NLG Performance

- Human identify if the article is generated



NLG Performance

- Using a new word in a sentence (few-shot)

A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is:
We were traveling in Africa and we saw these very cute whatpus.

To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:
One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.
A "yalubalu" is a type of vegetable that looks like a big pumpkin. An example of a sentence that uses the word yalubalu is:
I was on a trip to Africa and I tried this yalubalu vegetable that was grown in a garden there. It was delicious.

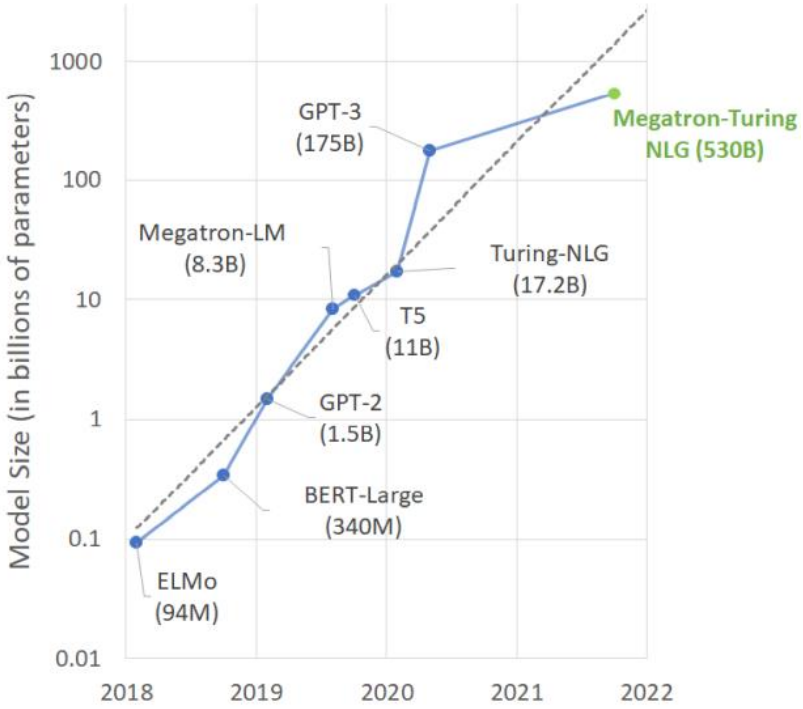
A "Burringo" is a car with very fast acceleration. An example of a sentence that uses the word Burringo is:
In our garage we have a Burringo that my father drives to work every day.

A "Gigamuru" is a type of Japanese musical instrument. An example of a sentence that uses the word Gigamuru is:
I have a Gigamuru that my uncle gave me as a gift. I love to play it at home.

Issue 2: Large-Scale PLMs

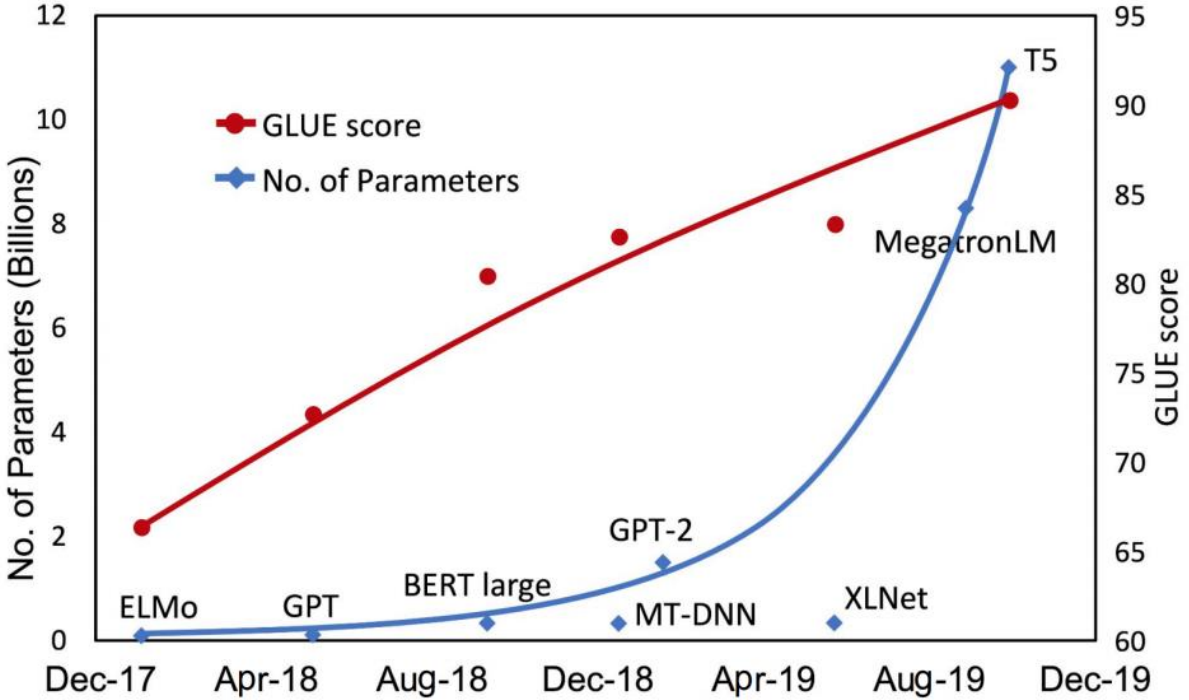
- PLMs are larger and larger

Model	#Params	#Layers
ELMo	93M	2 (BiLSTM)
BERT Base	110M	12
BERT Large	340M	24
GPT-3 Small	125M	12
GPT-3 Medium	350M	24
GPT-3 Large	760M	24
GPT-3 XL	1.3B	24
GPT-3 2.7B	2.7B	32
GPT-3 6.7B	6.7B	32
GPT-3 13B	13B	40
GPT-3 175B (“GPT-3”)	175.0B	96



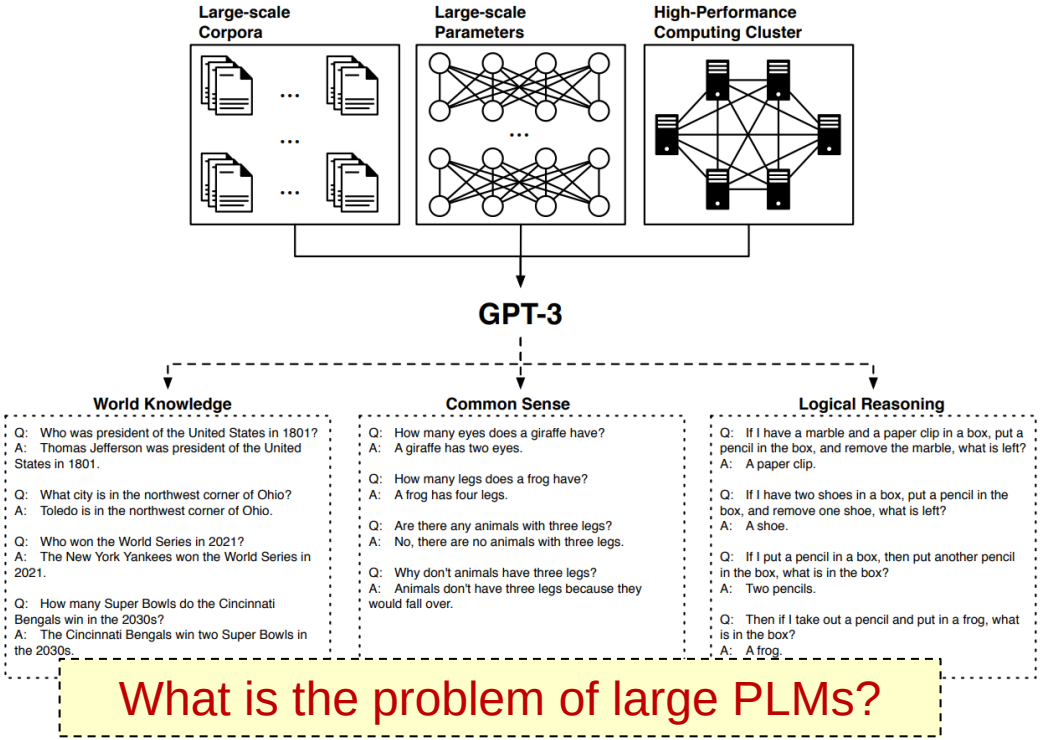
Better Performance from Larger Models

- Language understanding performance (Ahmet & Abdullah, 2021)

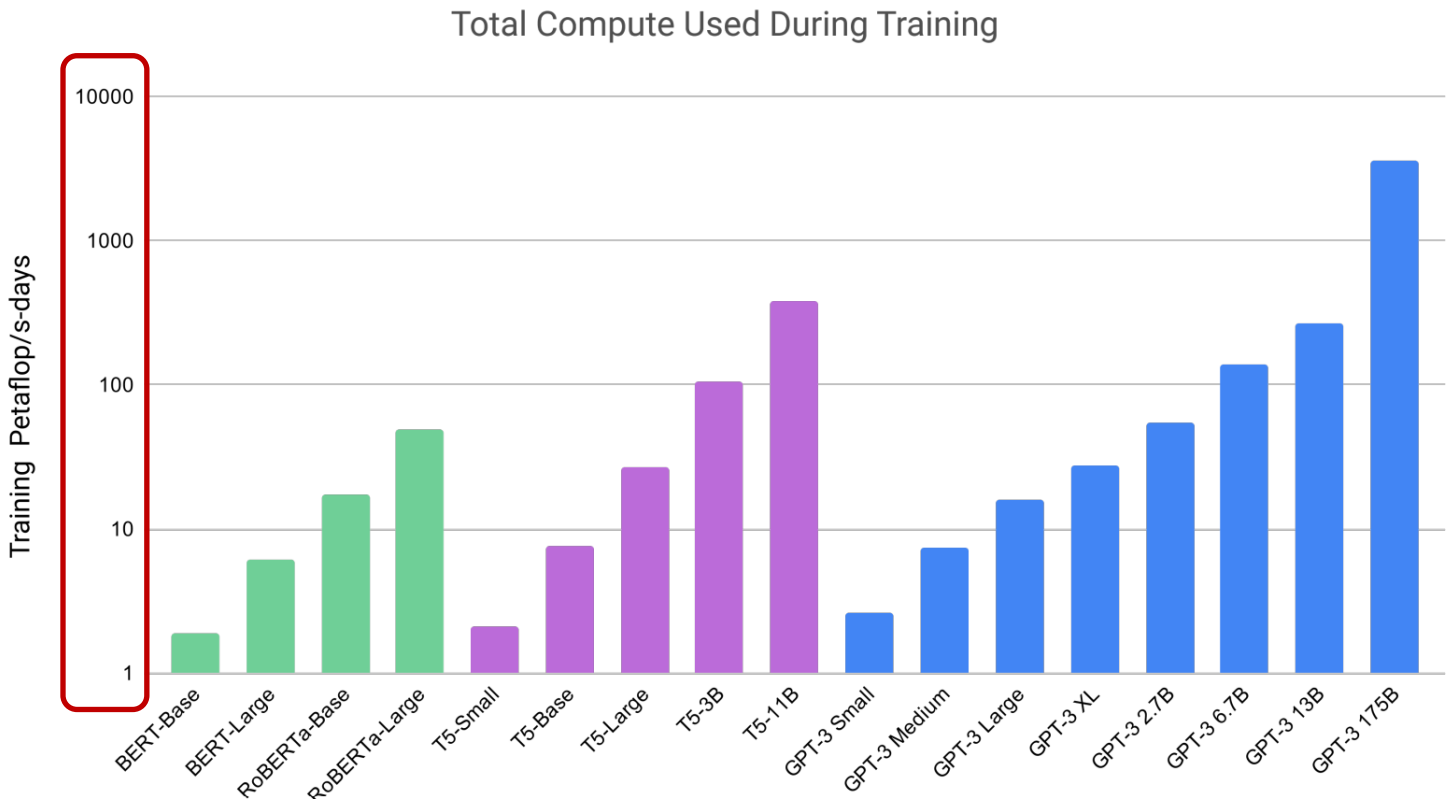


Better Performance from Large Models

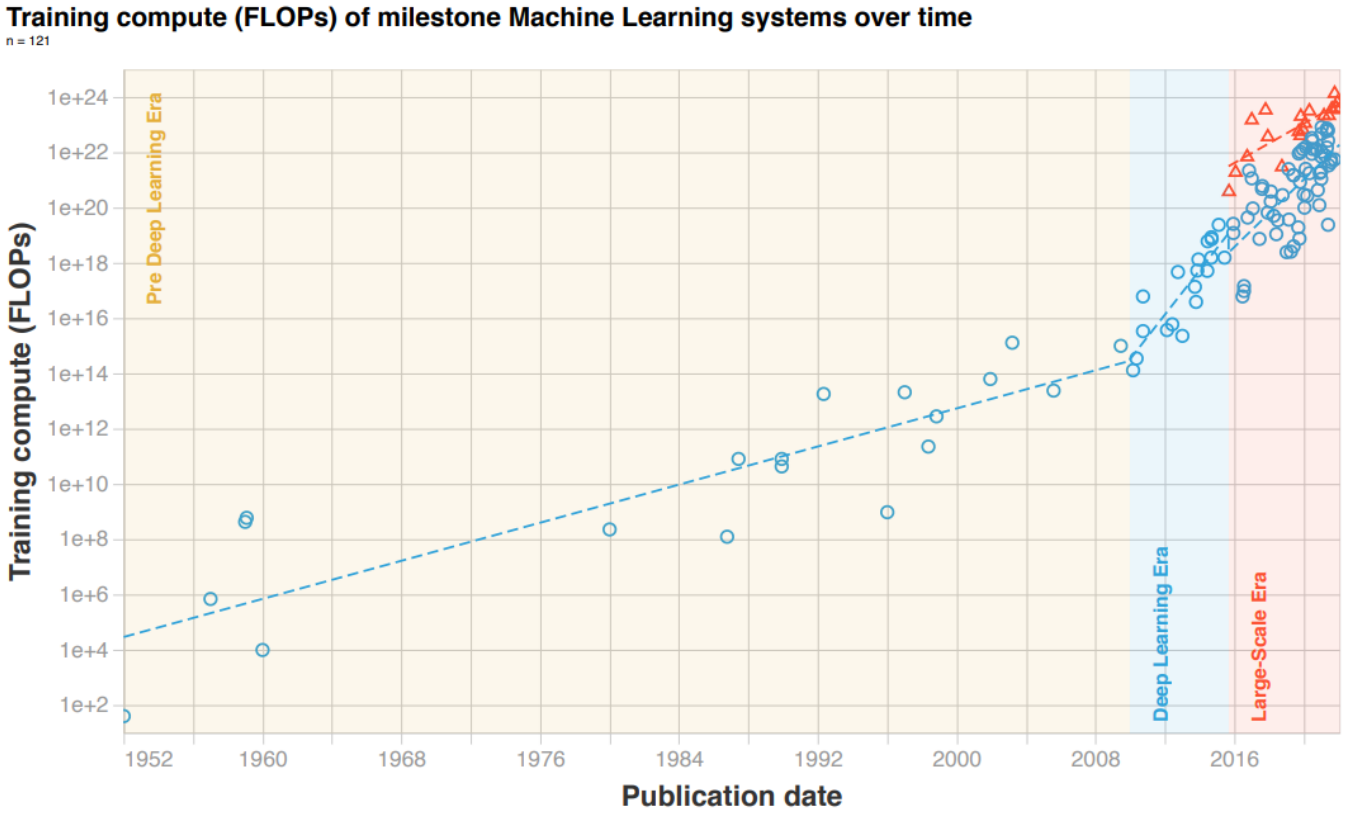
- More types of data for pre-training → diverse capability



Training Cost of Large PLMs

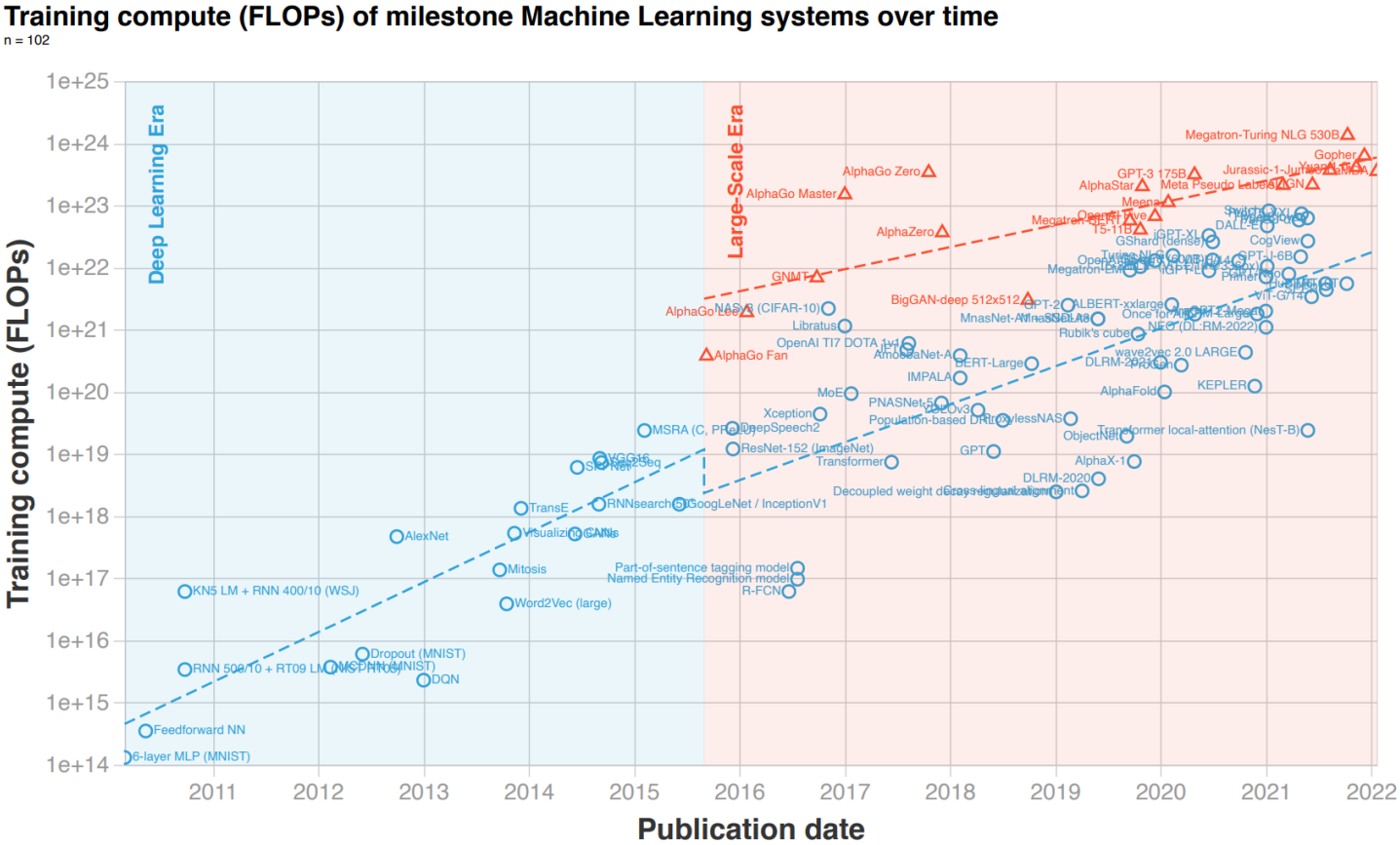


Training Cost of Large PLMs



Sevilla et al., “Compute Trends Across Three Eras of Machine Learning,” in arXiv:2202.05924, 2022.

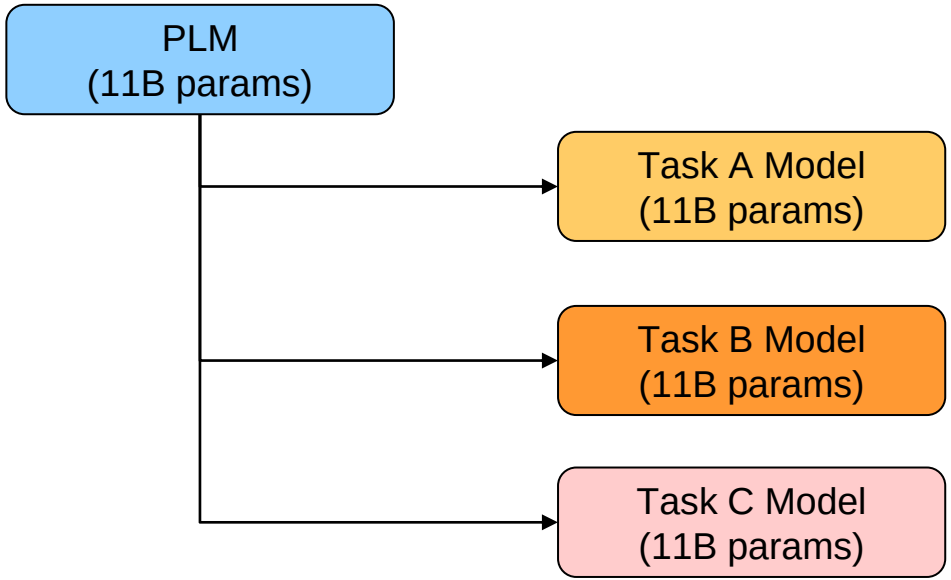
Training Cost of Large PLMs



Sevilla et al., “Compute Trends Across Three Eras of Machine Learning,” in arXiv:2202.05924, 2022.

Large Space Requirement

- Each task requires a copy of a large model



Practical Issues of PLMs

1) Data scarcity
2) Large PLMs
   - Higher training cost
   - Larger space requirement

→ Solution: Prompt-Based Learning

38

Prompt-Based Learning

Leveraging big pre-trained models

GPT-3 “In-Context” Learning

- Zero-Shot

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

- One-Shot

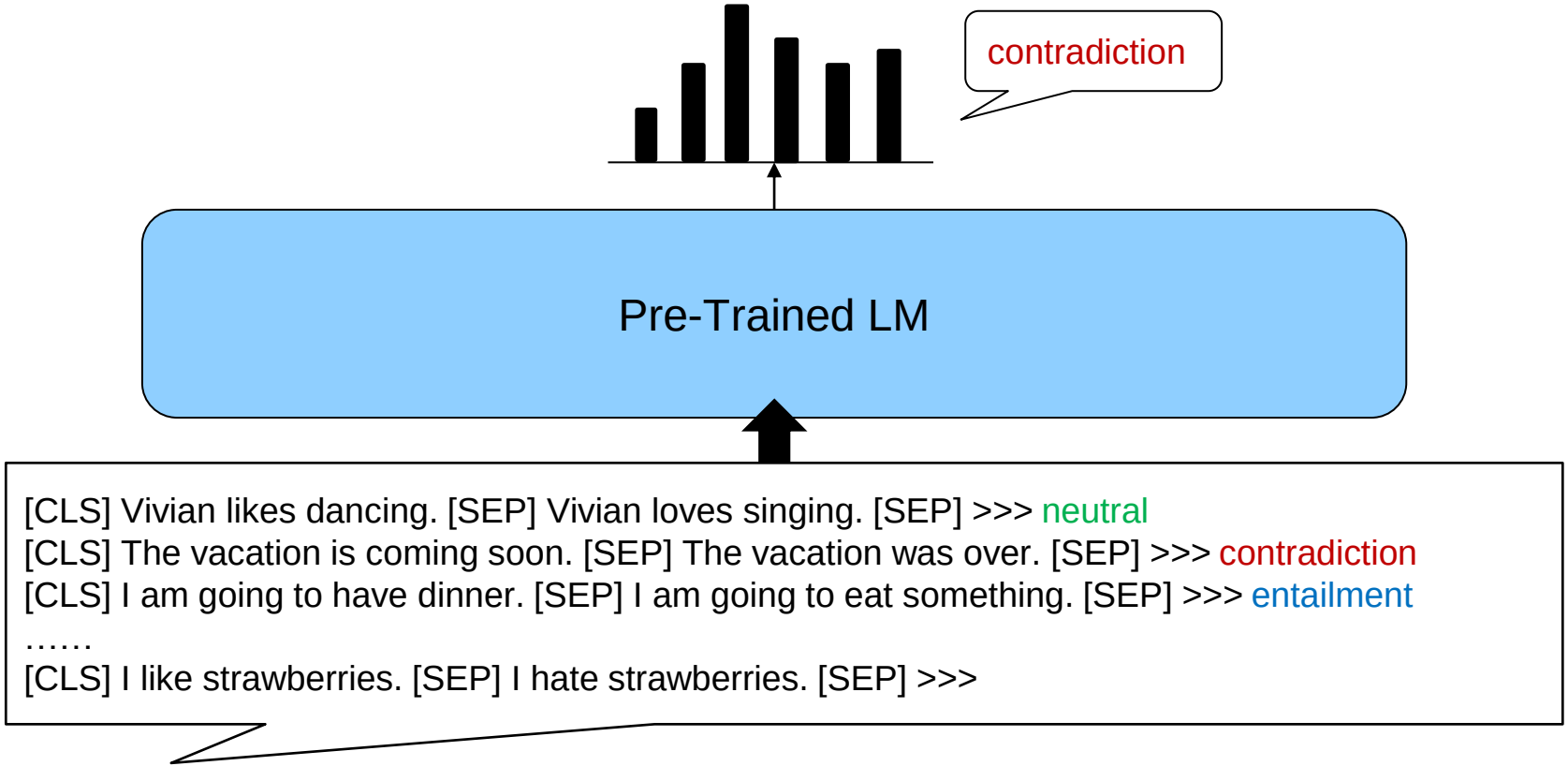
```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

- Few-Shot

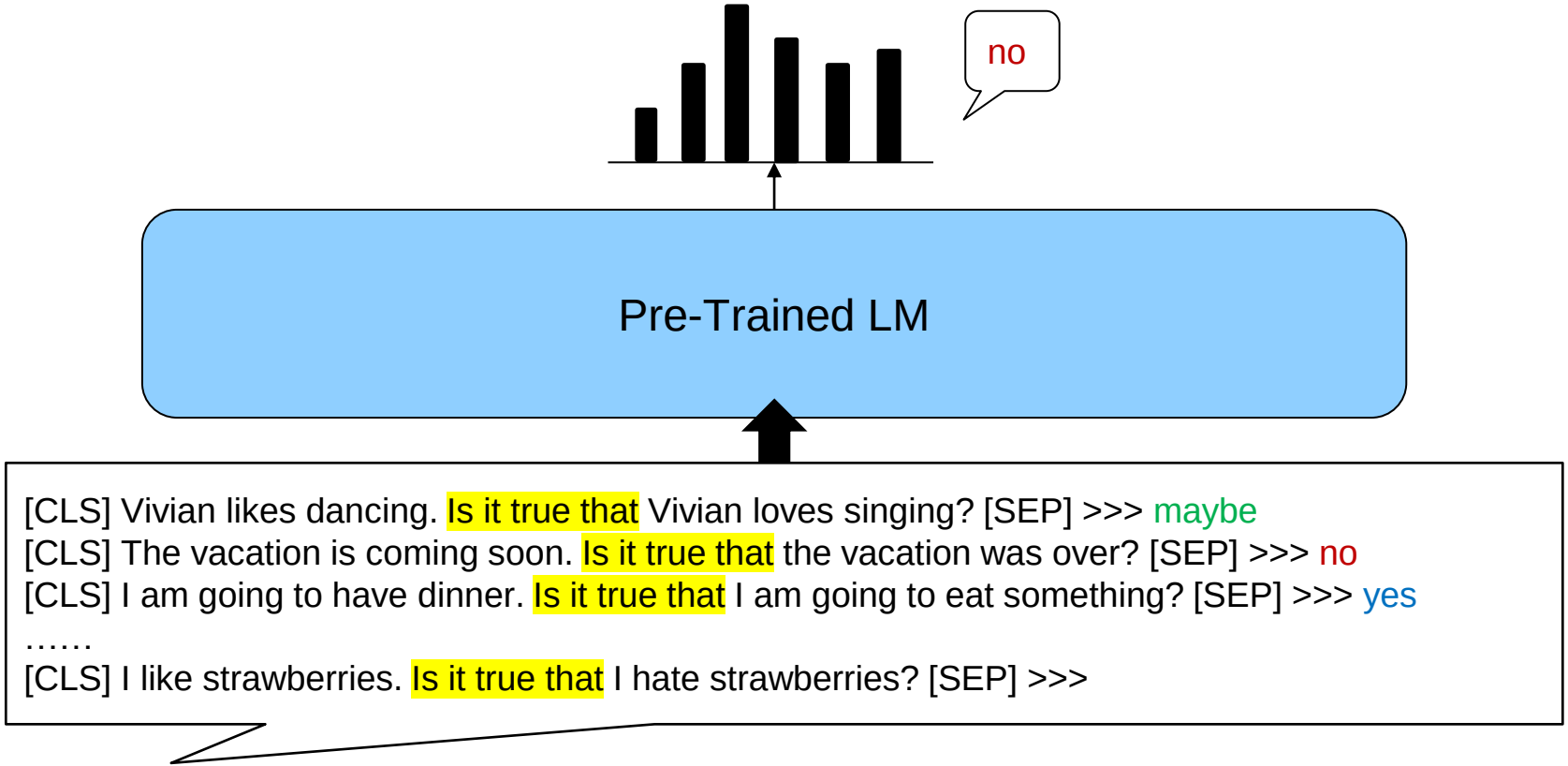
```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```

natural language instruction and/or a few task demonstrations

Prompt-Tuning



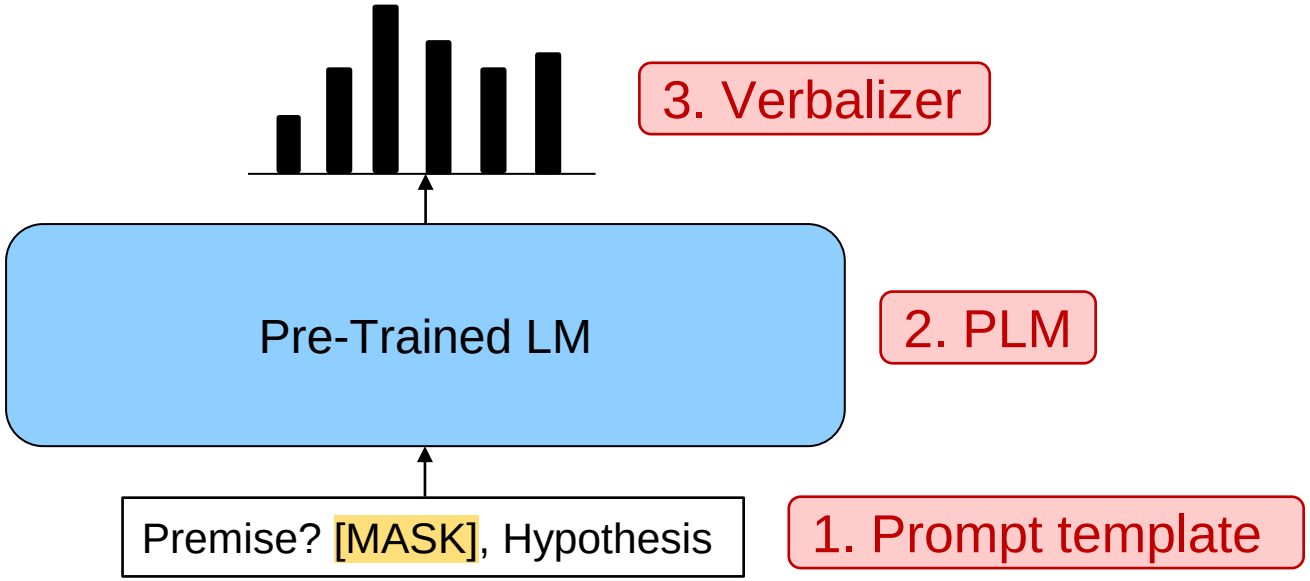
Prompt-Tuning



Prompt-Tuning

- Idea: convert data into natural language prompts

  → better for few-shot, one-shot, or zero-shot cases



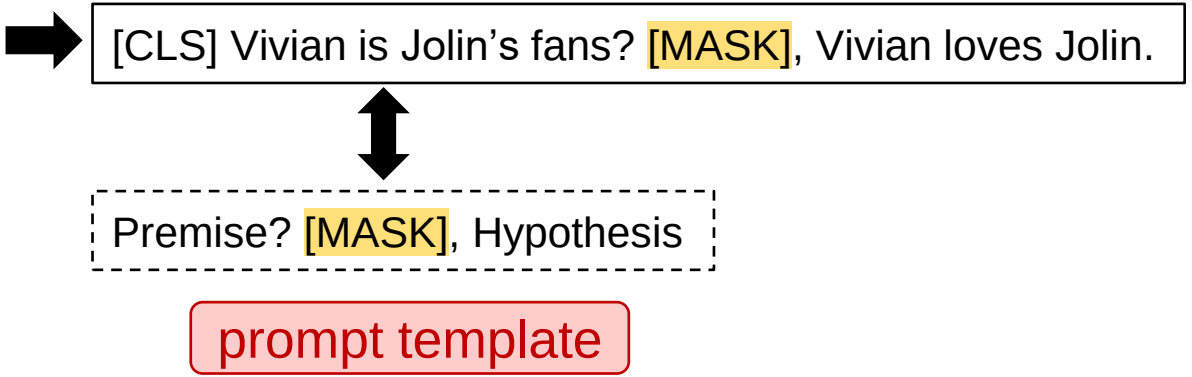
Prompt-Tuning

1. Prompt template: manually designed natural language input for a task

NLI sample datapoint

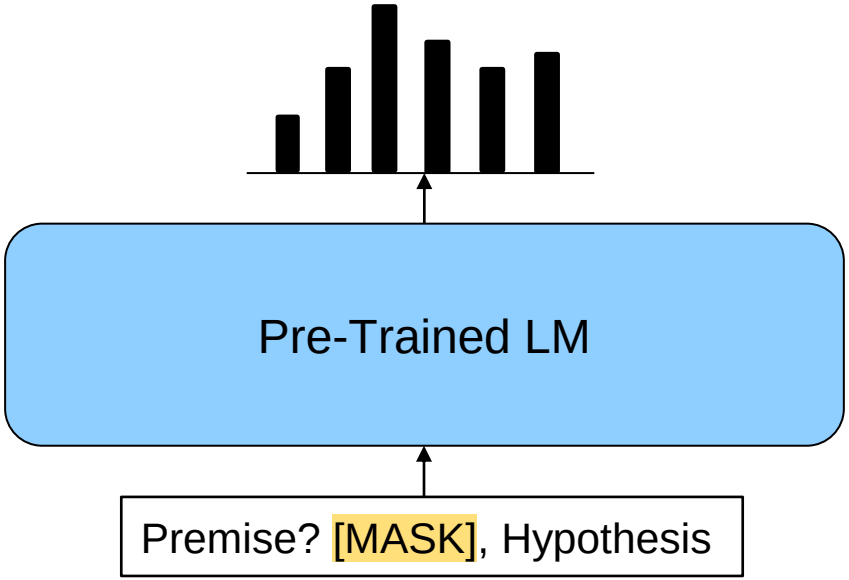
Premise	Vivian is Jolin's fans
Hypothesis	Vivian loves Jolin.
Label	0

0: "entailment"
1: "neutral"
2: "contradiction"



Prompt-Tuning

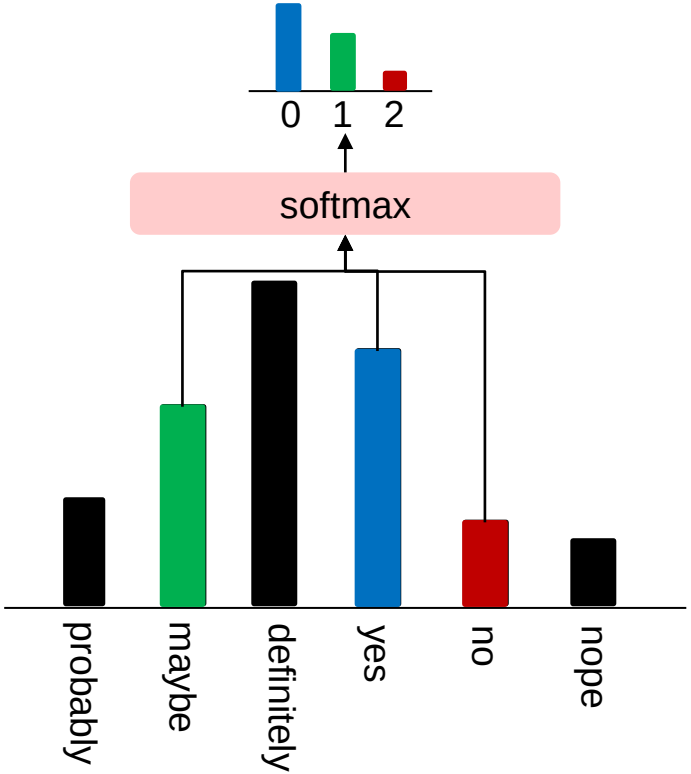
2. PLM: perform language modeling (masked LM or auto-regressive LM)



Prompt-Tuning

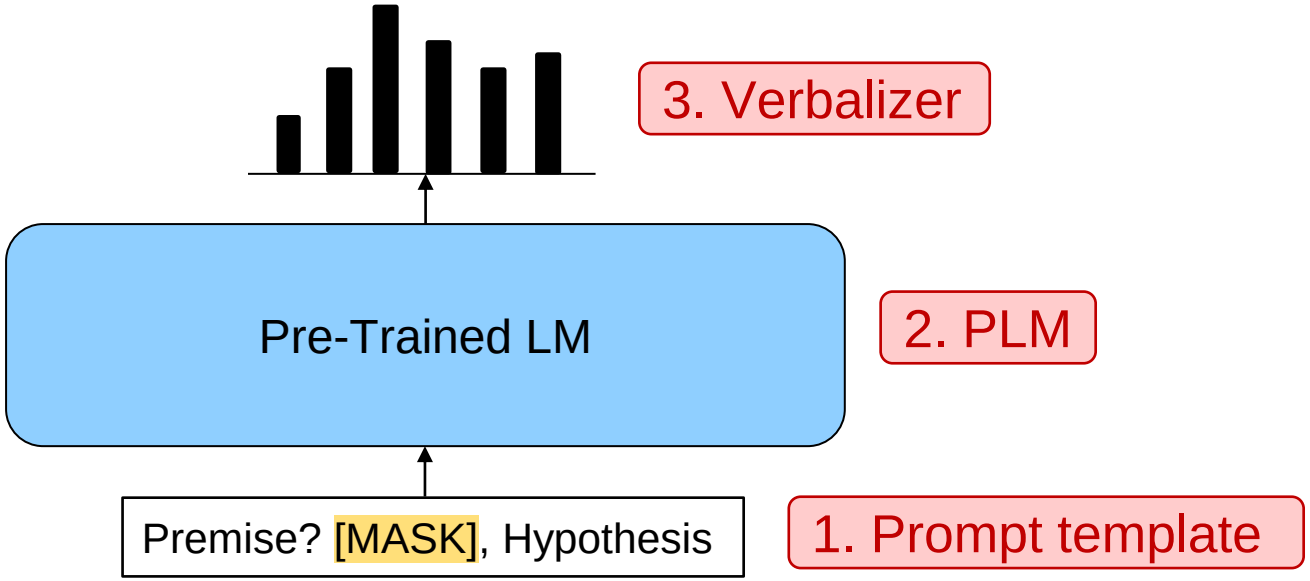
3. Verbalizer: mapping from the vocabulary to labels

0: “entailment” → yes
1: “neutral” → maybe
2: “contradiction” → no



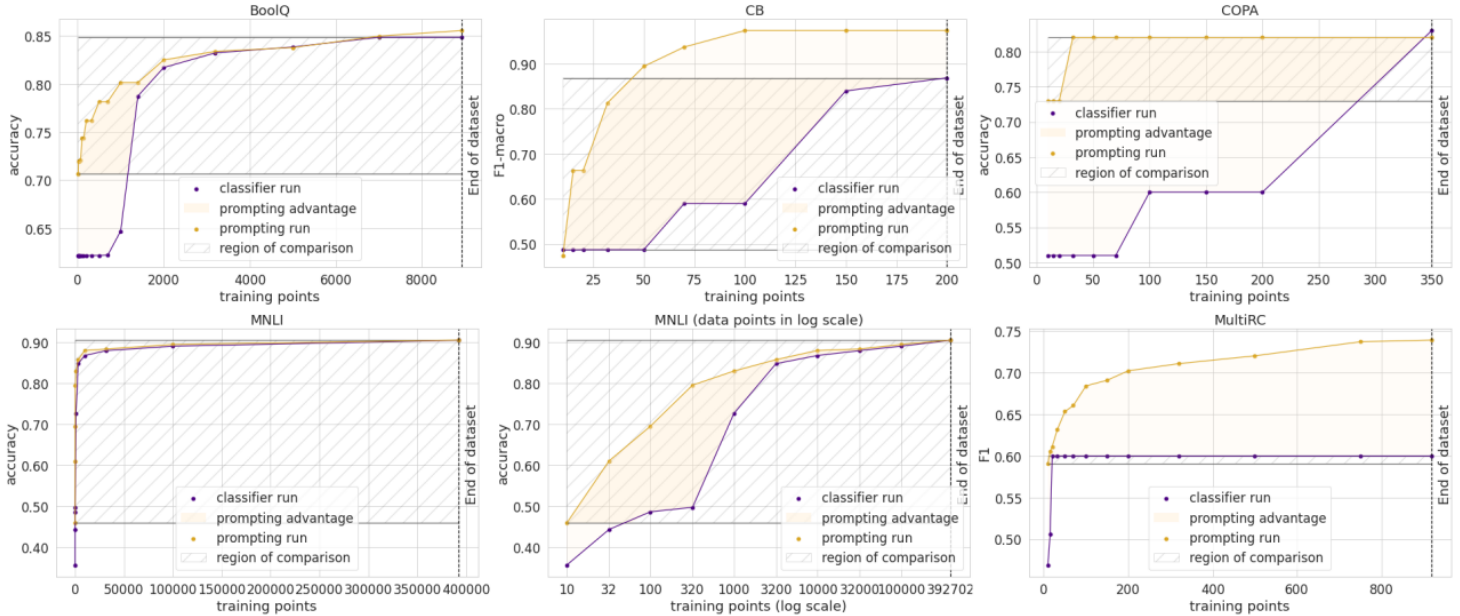
Prompt-Tuning

- Fine-tuning PLMs based on few annotated data samples
  - No parameter tuning when zero-shot settings



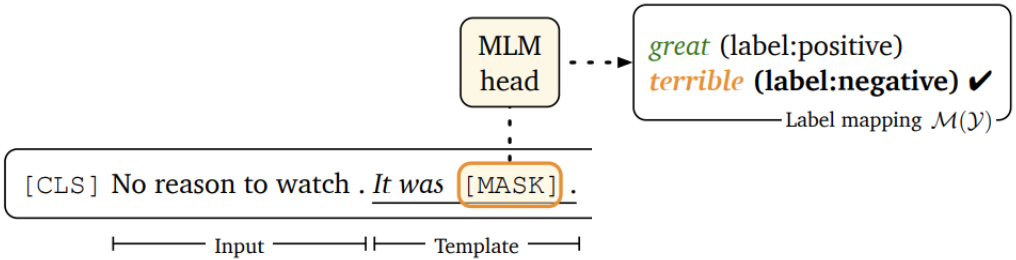
Prompt-Tuning

- Prompt-tuning is better under data scarcity (Le and Rush, 2021) due to
  - It better leverages pre-trained knowledge
  - Pre-trained knowledge can be kept

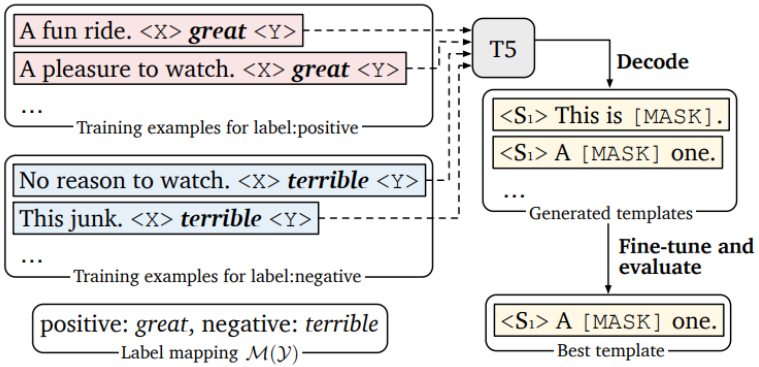


LM-BFF: Better Few-shot Fine-tuning of Language Models (Gao et al., 2021)

- Idea: prompt + demonstration for few-shot learning



- template generation



LM-BFF: Better Few-shot Fine-tuning of Language Models (Gao et al., 2021)

- Performance with RoBERTa-Large

	SST-2 (acc)	SST-5 (acc)	MR (acc)	CR (acc)	MPQA (acc)	Subj (acc)	TREC (acc)	CoLA (Matt.)
Majority†	*50.9*	*23.1*	*50.0*	*50.0*	*50.0*	*50.0*	*18.8*	*0.0*
Prompt-based zero-shot‡	83.6	35.0	80.8	79.5	67.6	51.4	32.0	2.0
"GPT-3" in-context learning	84.8 (1.3)	30.6 (0.9)	80.5 (1.7)	87.4 (0.8)	63.8 (2.1)	53.6 (1.0)	26.2 (2.4)	-1.5 (2.4)
Fine-tuning	81.4 (3.8)	43.9 (2.0)	76.9 (5.9)	75.8 (3.2)	72.0 (3.8)	90.8 (1.8)	88.8 (2.1)	**33.9** (14.3)
Prompt-based FT (man)	92.7 (0.9)	47.4 (2.5)	87.0 (1.2)	90.3 (1.0)	84.7 (2.2)	91.2 (1.1)	84.8 (5.1)	9.3 (7.3)
+ demonstrations	92.6 (0.5)	**50.6** (1.4)	86.6 (2.2)	90.2 (1.2)	**87.0** (1.1)	**92.3** (0.8)	87.5 (3.2)	18.7 (8.8)
Prompt-based FT (auto)	92.3 (1.0)	49.2 (1.6)	85.5 (2.8)	89.0 (1.4)	85.8 (1.9)	91.2 (1.1)	88.2 (2.0)	14.0 (14.1)
+ demonstrations	**93.0** (0.6)	49.5 (1.7)	**87.7** (1.4)	**91.0** (0.9)	86.5 (2.6)	91.4 (1.8)	**89.4** (1.7)	21.8 (15.9)
Fine-tuning (full)†	*95.0*	*58.7*	*90.8*	*89.4*	*87.8*	*97.0*	*97.4*	*62.6*

	MNLI (acc)	MNLI-mm (acc)	SNLI (acc)	QNLI (acc)	RTE (acc)	MRPC (F1)	QQP (F1)	STS-B (Pear.)
Majority†	*32.7*	*33.0*	*33.8*	*49.5*	*52.7*	*81.2*	*0.0*	-
Prompt-based zero-shot‡	50.8	51.7	49.5	50.8	51.3	61.9	49.7	-3.2
"GPT-3" in-context learning	52.0 (0.7)	53.4 (0.6)	47.1 (0.6)	53.8 (0.4)	60.4 (1.4)	45.7 (6.0)	36.1 (5.2)	14.3 (2.8)
Fine-tuning	45.8 (6.4)	47.8 (6.8)	48.4 (4.8)	60.2 (6.5)	54.4 (3.9)	76.6 (2.5)	60.7 (4.3)	53.5 (8.5)
Prompt-based FT (man)	68.3 (2.3)	70.5 (1.9)	77.2 (3.7)	64.5 (4.2)	69.1 (3.6)	74.5 (5.3)	65.5 (5.3)	71.0 (7.0)
+ demonstrations	**70.7** (1.3)	**72.0** (1.2)	**79.7** (1.5)	**69.2** (1.9)	68.7 (2.3)	77.8 (2.0)	**69.8** (1.8)	73.5 (5.1)
Prompt-based FT (auto)	68.3 (2.5)	70.1 (2.6)	77.1 (2.1)	68.3 (7.4)	**73.9** (2.2)	76.2 (2.3)	67.0 (3.0)	75.0 (3.3)
+ demonstrations	70.0 (3.6)	**72.0** (3.1)	77.5 (3.5)	68.5 (5.4)	71.1 (5.3)	**78.1** (3.4)	67.7 (5.8)	**76.4** (6.2)
Fine-tuning (full)†	*89.8*	*89.5*	*92.6*	*93.3*	*80.9*	*91.4*	*81.7*	*91.9*

Issues of Discrete/Hard Prompts

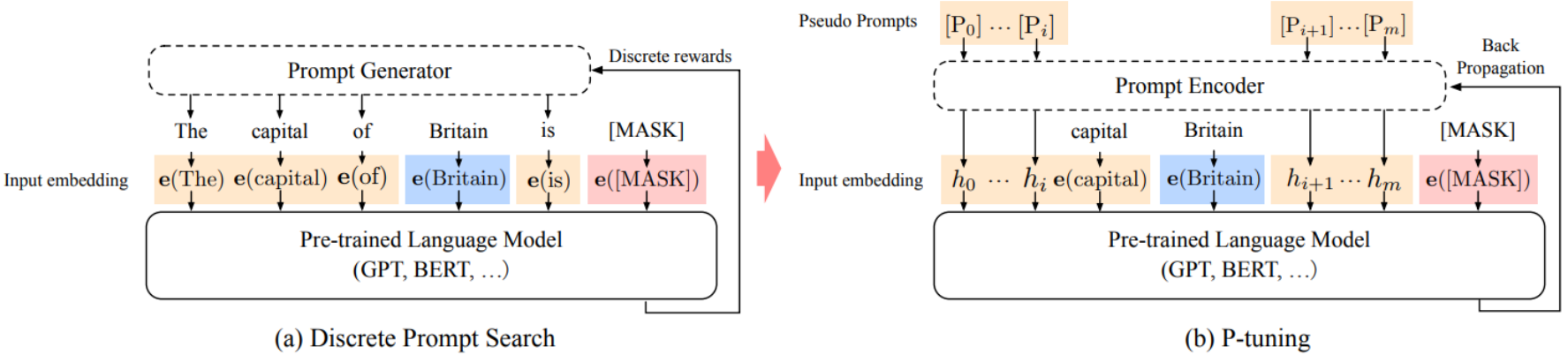
- Difficulty of manually designing prompts
  - Prompts that humans consider reasonable is not necessarily effective for LMs (Liu et al., 2021)
  - Pre-trained LMs are sensitive to the choice of prompts (Zhao et al., 2021)

Prompt	P@1
[X] is located in [Y]. *(original)*	31.29
[X] is located in which country or state? [Y].	19.78
[X] is located in which country? [Y].	31.40
[X] is located in which country? In [Y].	51.08

P-Tuning (Liu et al., 2021)

- Idea: direct optimize the embeddings instead of prompt tokens

prompt search for “The capital of Britain is [MASK]”.

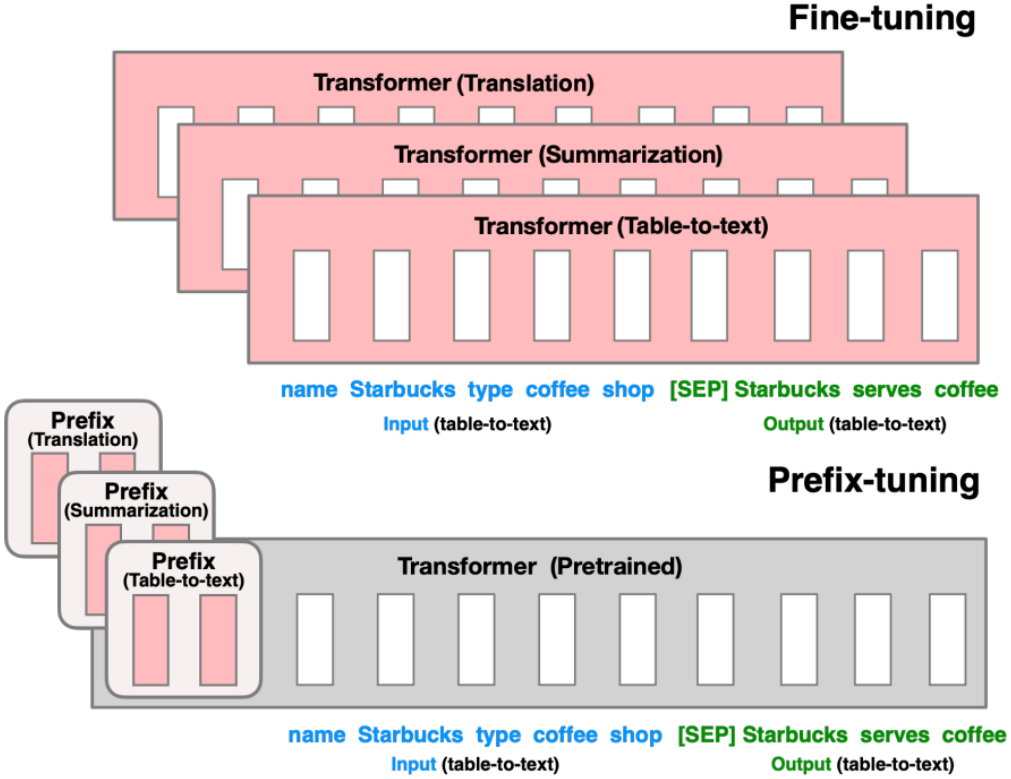


(a) Discrete Prompt Search (b) P-tuning

Prompt	$\mathcal{D}_{dev}$ Acc.	$\mathcal{D}_{dev32}$ Acc.
Does [PRE] agree with [HYP]? [MASK].	57.16	53.12
Does [HYP] agree with [PRE]? [MASK].	51.38	50.00
Premise: [PRE] Hypothesis: [HYP] Answer: [MASK].	68.59	55.20
[PRE] question: [HYP]. true or false? answer: [MASK].	70.15	53.12
P-tuning	76.45	56.25

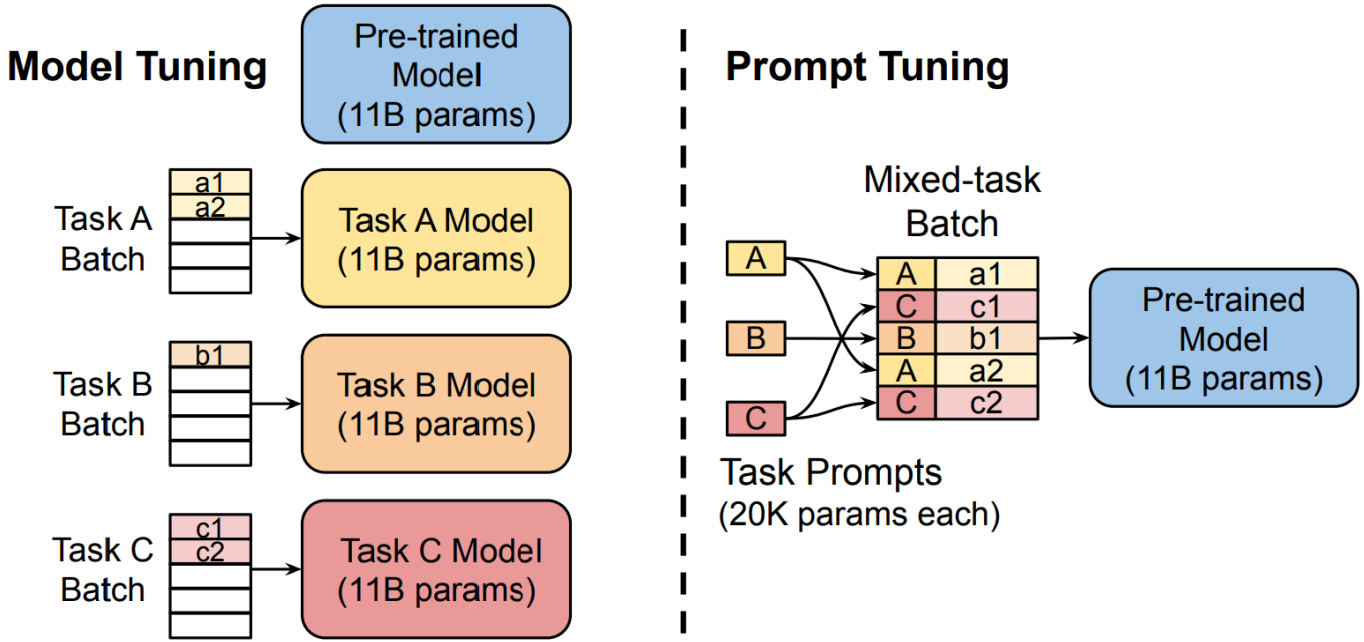
Prefix-Tuning (Li and Liang, 2021)

- Idea: only optimize the prefix embeddings (all layers) for efficiency



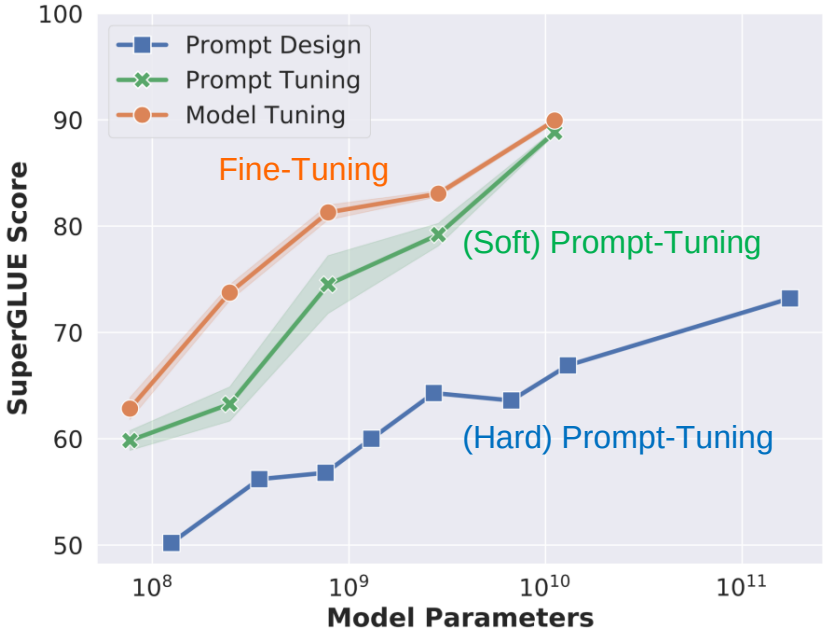
(Soft) Prompt-Tuning (Lester et al., 2021)

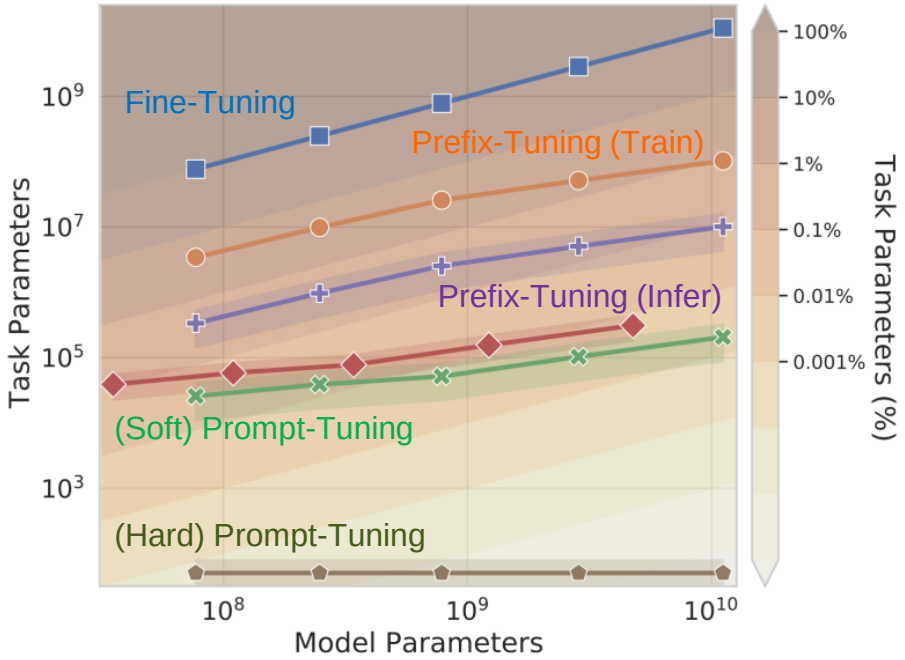
- Idea: only require storing a small task-specific prompt (one layer) for each task and enables mixed-task inference using the original PLMs



(Soft) Prompt-Tuning (Lester et al., 2021)

- Competitive performance and better space efficiency





Instruction Tuning (Wei et al., 2022)

- Idea: improve model's capability of understanding the task description

LM for sentence completion

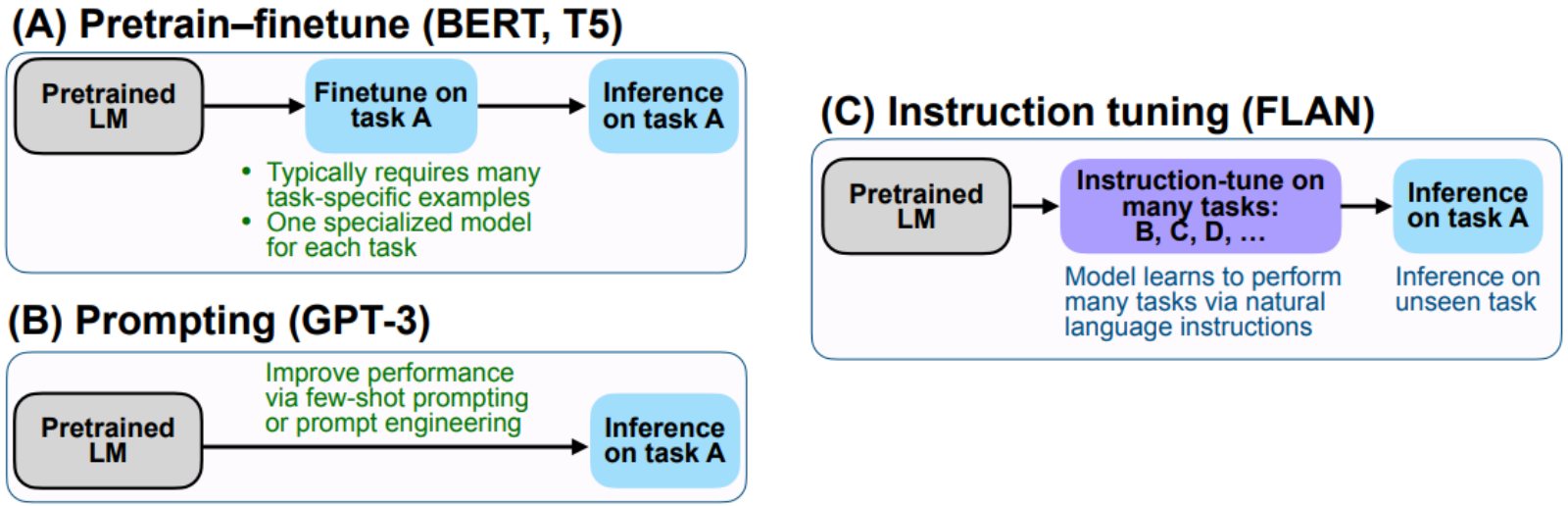
I went to Jolin's concert last night. I really loved her songs and dancing. It was _____

Detailed task instruction for LM generation

Decide the sentiment of the following sentences:
I went to Jolin's concert last night. I really loved her songs and dancing.
OPTIONS: - positive – negative - neutral

FLAN: Finetuned LANguage Models (Wei et al., 2022)

- Idea: fine-tune LM to better understand task descriptions via other tasks



Prompt v.s. Instruction Tuning (Wei et al., 2022)

Prompt

Inference

Input (Translation)

Translate this sentence to Spanish: The new office building was built in less than three months.

Target

El nuevo edificio de oficinas se construyó en tres meses.

Instruction tuning

Training

Input (Commonsense Reasoning)

Here is a goal: Get a cool sleep on summer days.
How would you accomplish this goal?
OPTIONS:
-Keep stack of pillow cases in fridge.
-Keep stack of pillow cases in oven.

Target

keep stack of pillow cases in fridge

LM Fine-tuning

Inference

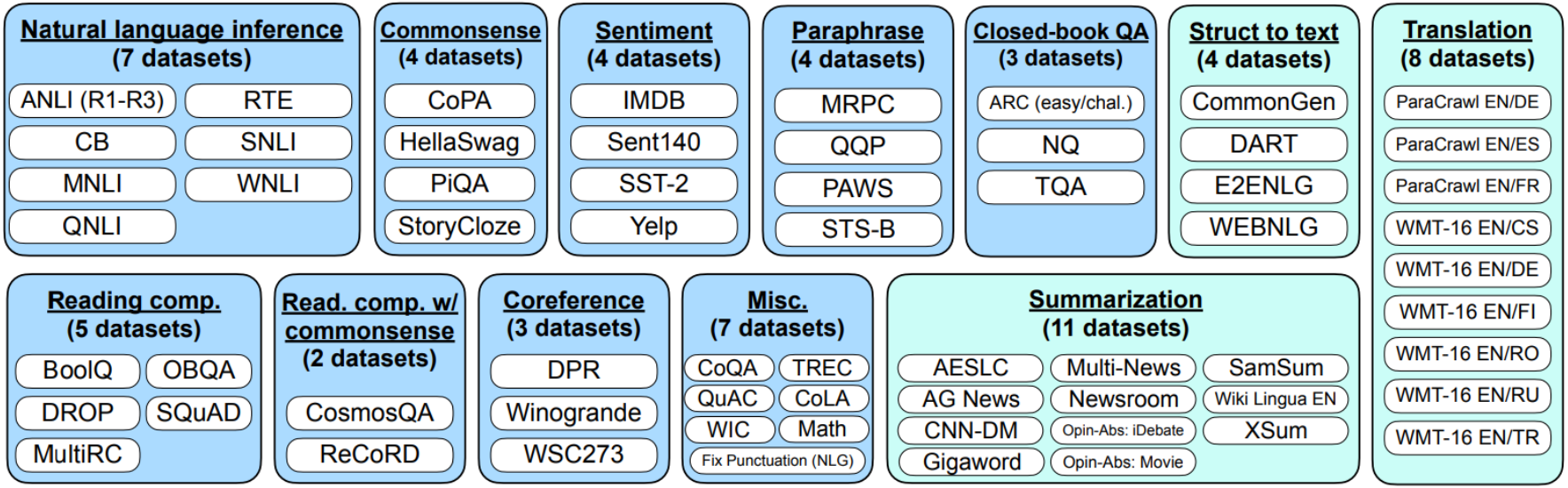
Input (Translation)

Translate this sentence to Spanish: The new office building was built in less than three months.

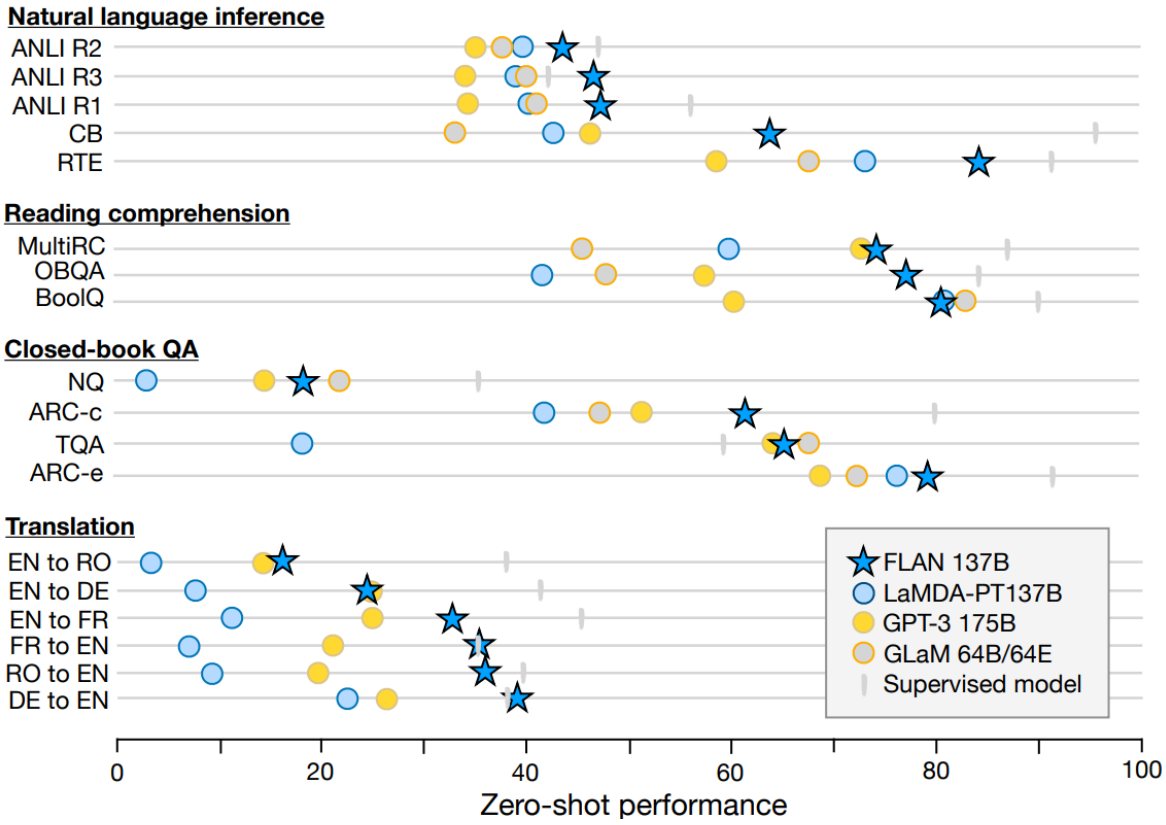
Target

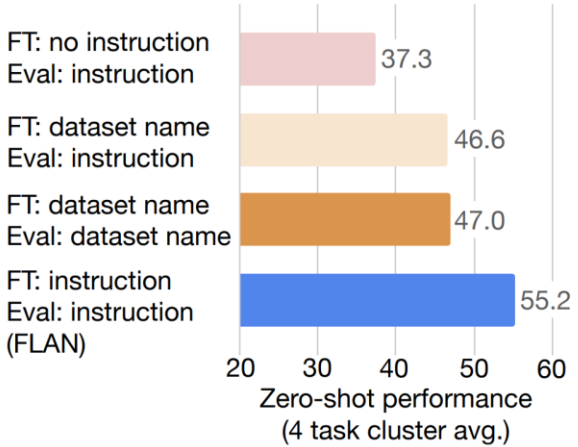
El nuevo edificio de oficinas se construyó en tres meses.

Task Clusters (Wei et al., 2022)



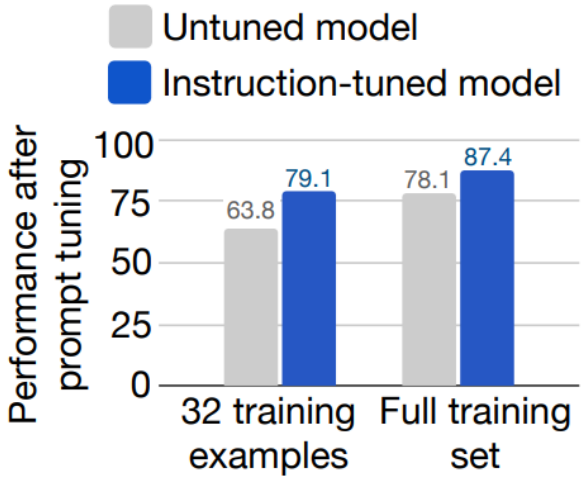
Zero-Shot Performance of FLAN



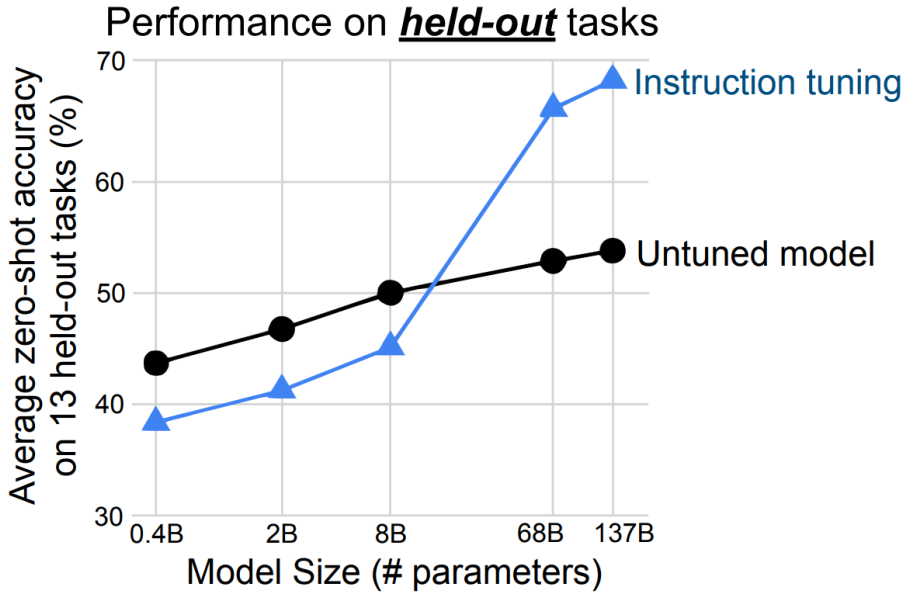


Zero-Shot Performance of FLAN

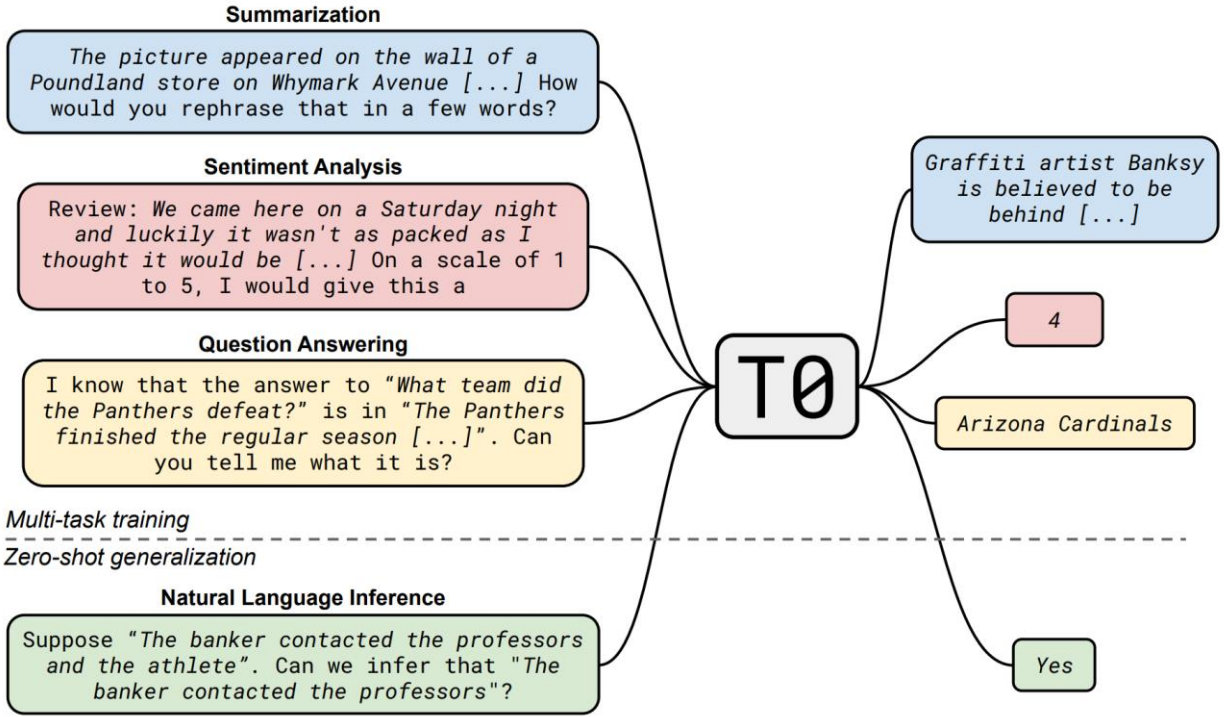
- Combine with prompt-tuning



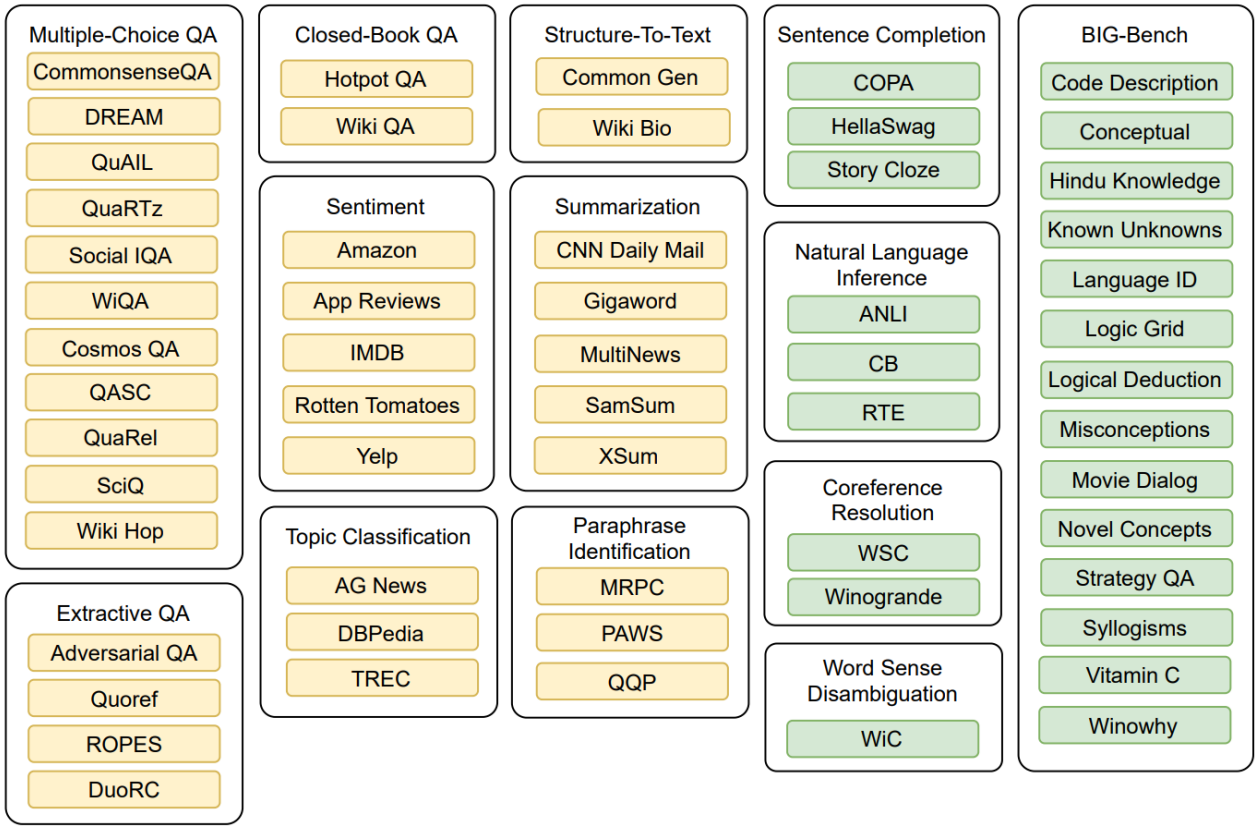
- Model size requirement



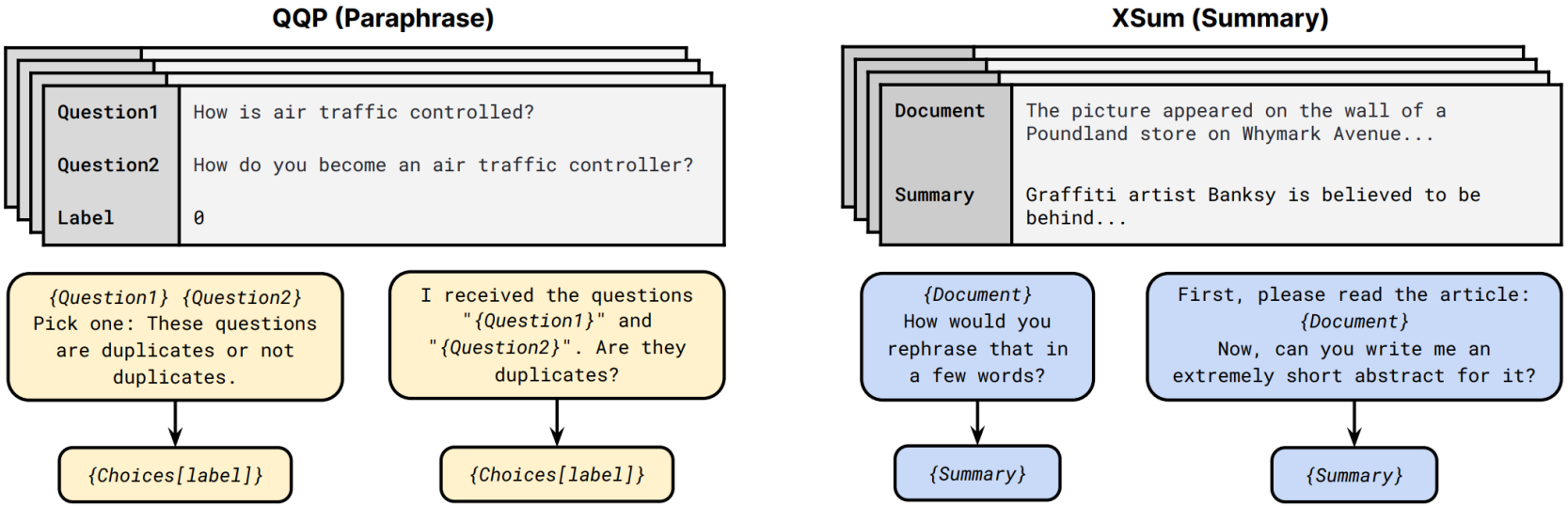
T0: Multitask Prompted Training (Sanh et al., 2022)



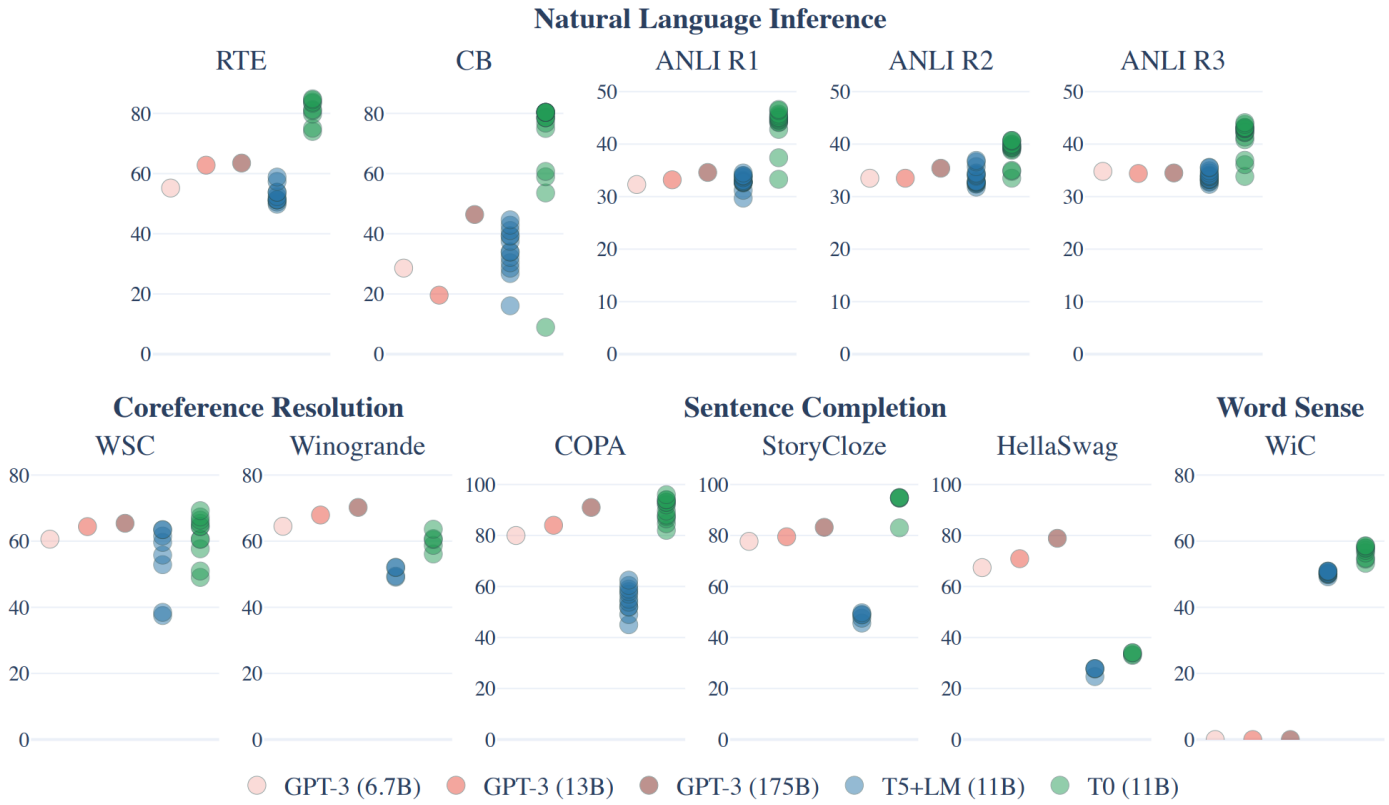
Task Clusters (Sanh et al., 2022)



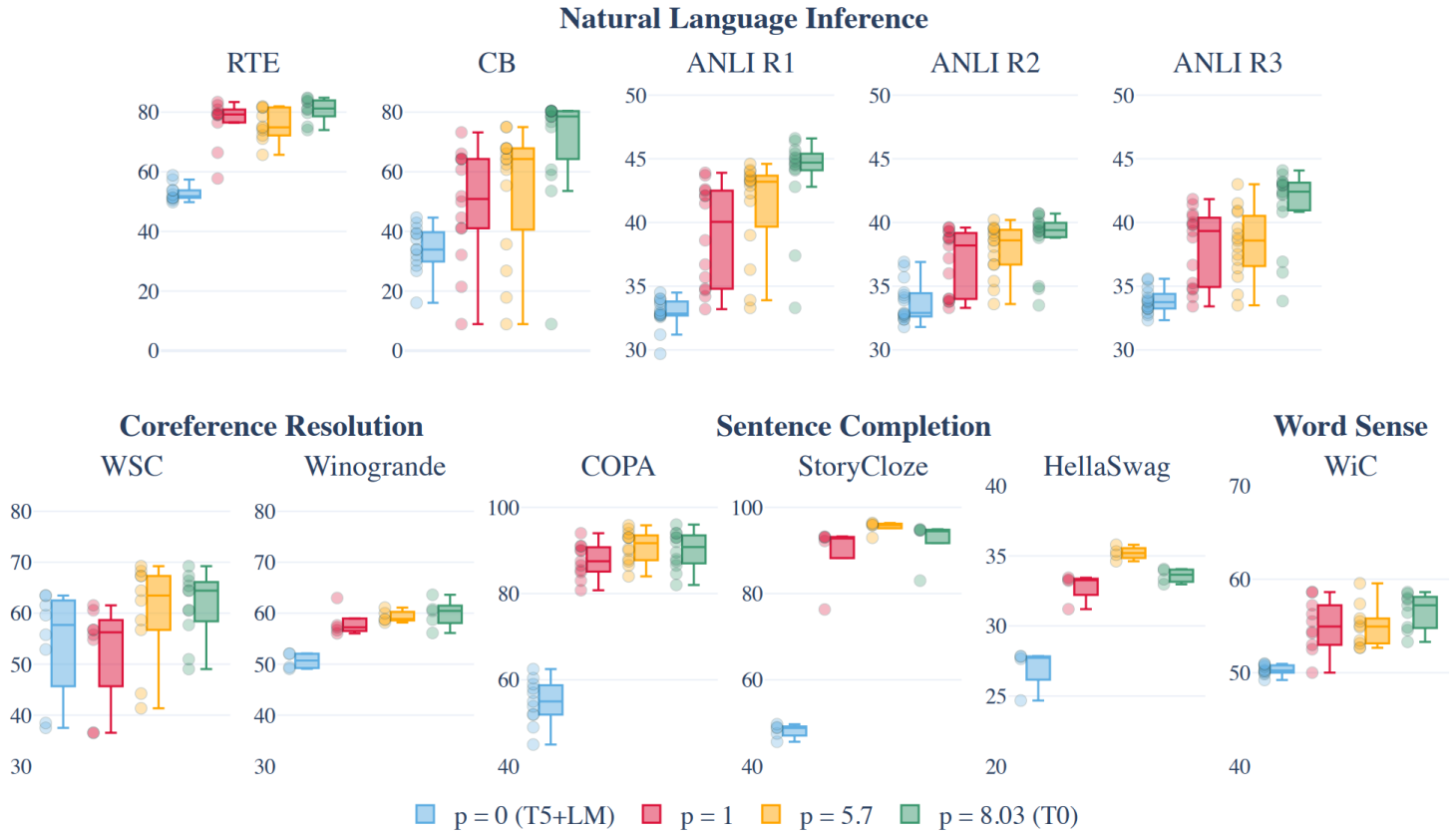
Prompt Templates (Sanh et al., 2022)



Performance of T0



Effect of #Prompts



Chain-of-Thought (CoT) (Wei et al., 2022)

Standard Prompting

Model Input

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: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ✗

Chain-of-Thought Prompting

Model Input

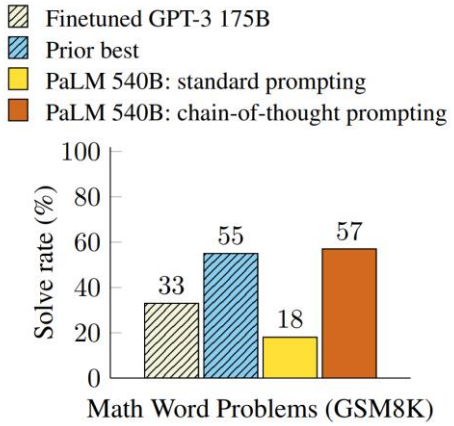
Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

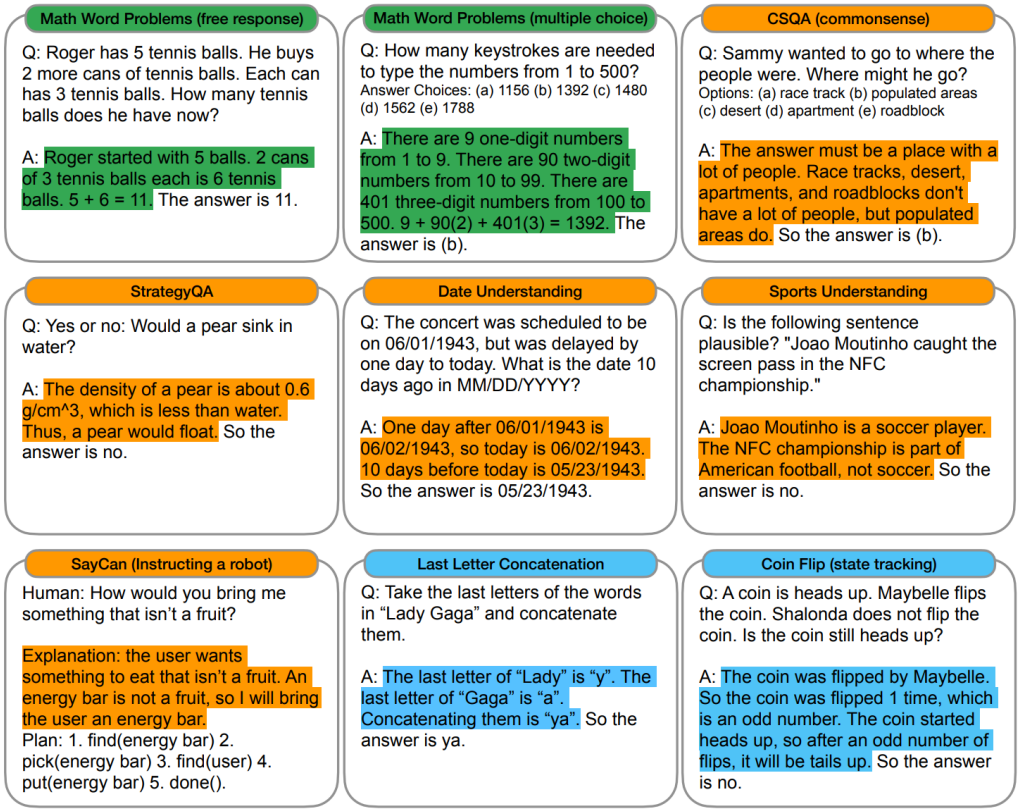
Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

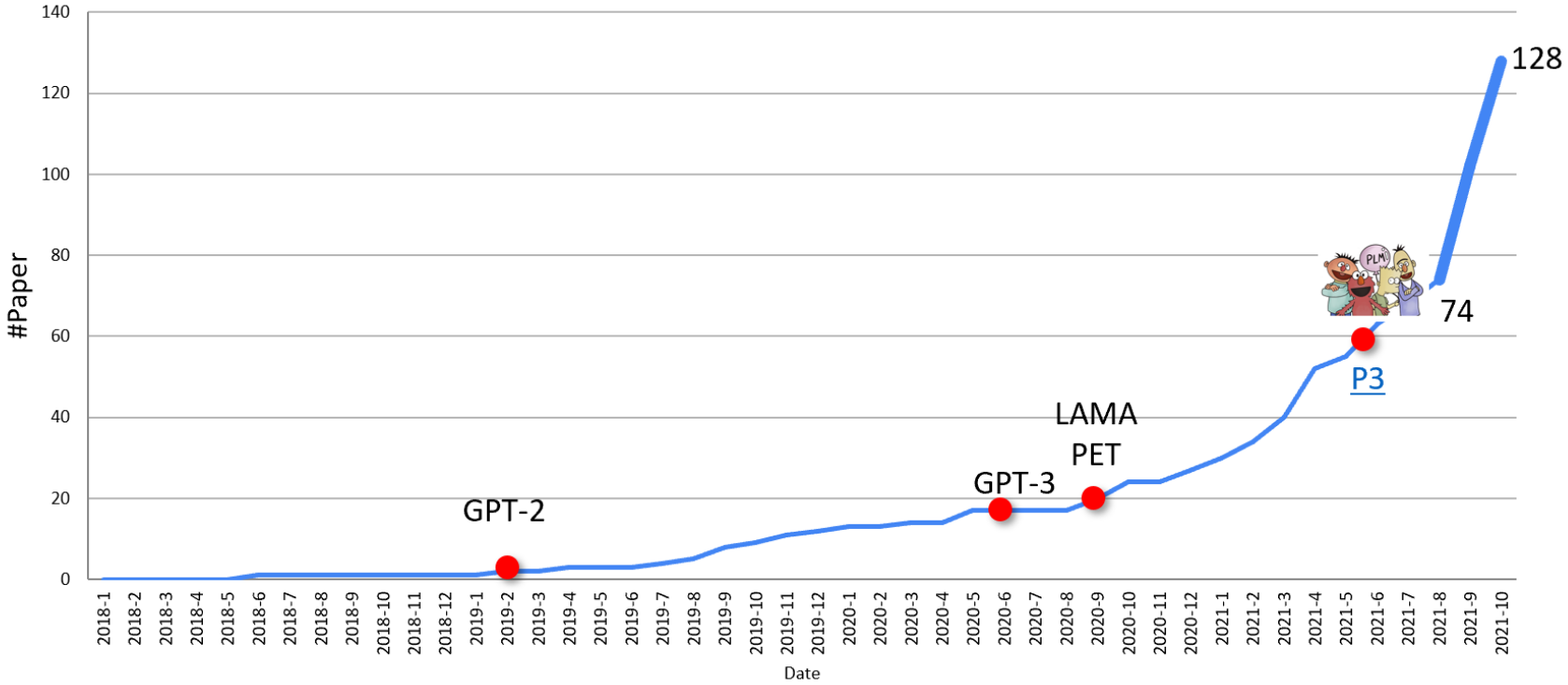
A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓



Chain-of-Thought (CoT) (Wei et al., 2022)

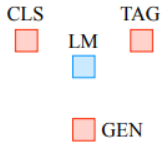
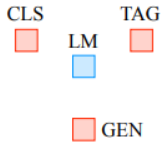
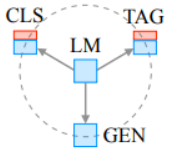
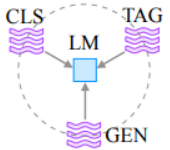


Trend of Prompt-Based Research



http://pretrain.nlpedia.ai/

Prompting Paradigm (Liu et al., 2021)

Paradigm	Engineering	Task Relation
a. Fully Supervised Learning (Non-Neural Network)	Features (e.g. word identity, part-of-speech, sentence length)	
b. Fully Supervised Learning (Neural Network)	Architecture (e.g. convolutional, recurrent, self-attentional)	
c. Pre-train, Fine-tune	Objective (e.g. masked language modeling, next sentence prediction)	
d. Pre-train, Prompt, Predict	Prompt (e.g. cloze, prefix)	

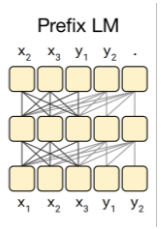
: unsupervised

: supervised

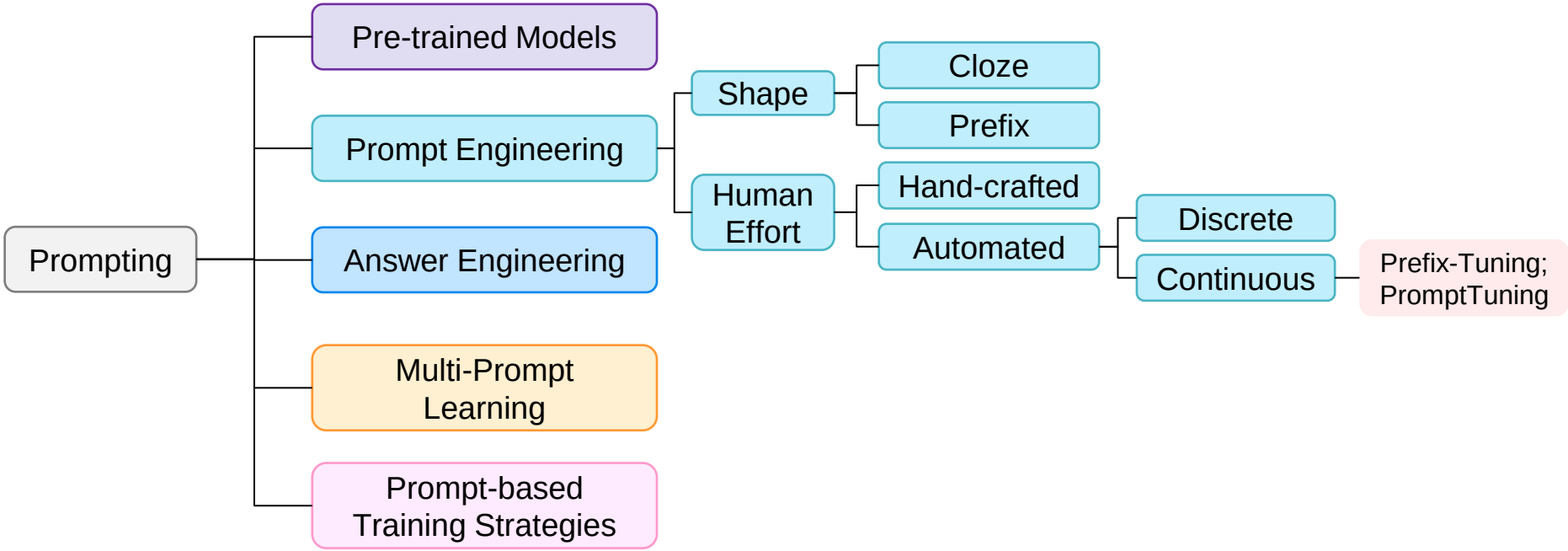
: textual prompt

Prompting Typology (Liu et al., 2021)

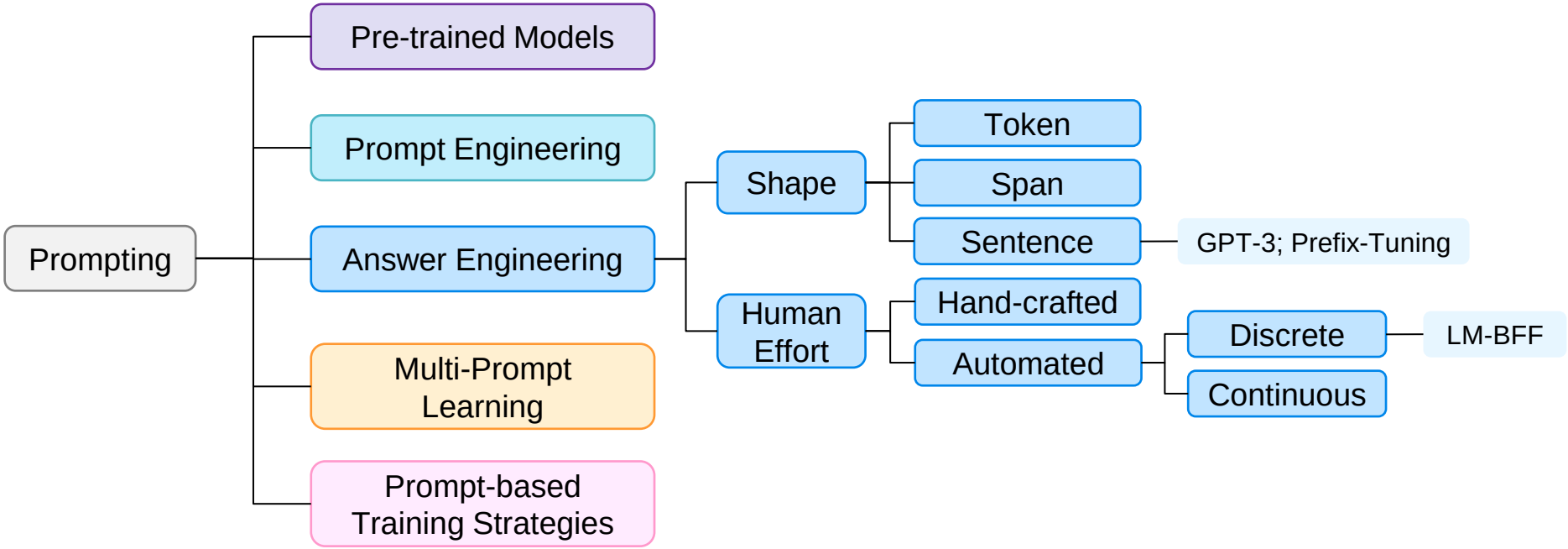
- Prompting
  - Pre-trained Models
    - Left-to-Right LM — GPT; GPT-2; GPT-3
    - Masked LM — BERT; RoBERTa
    - Prefix LM — UniLM
    - Encoder-Decoder — T5; BART
  - Prompt Engineering
  - Answer Engineering
  - Multi-Prompt Learning
  - Prompt-based Training Strategies



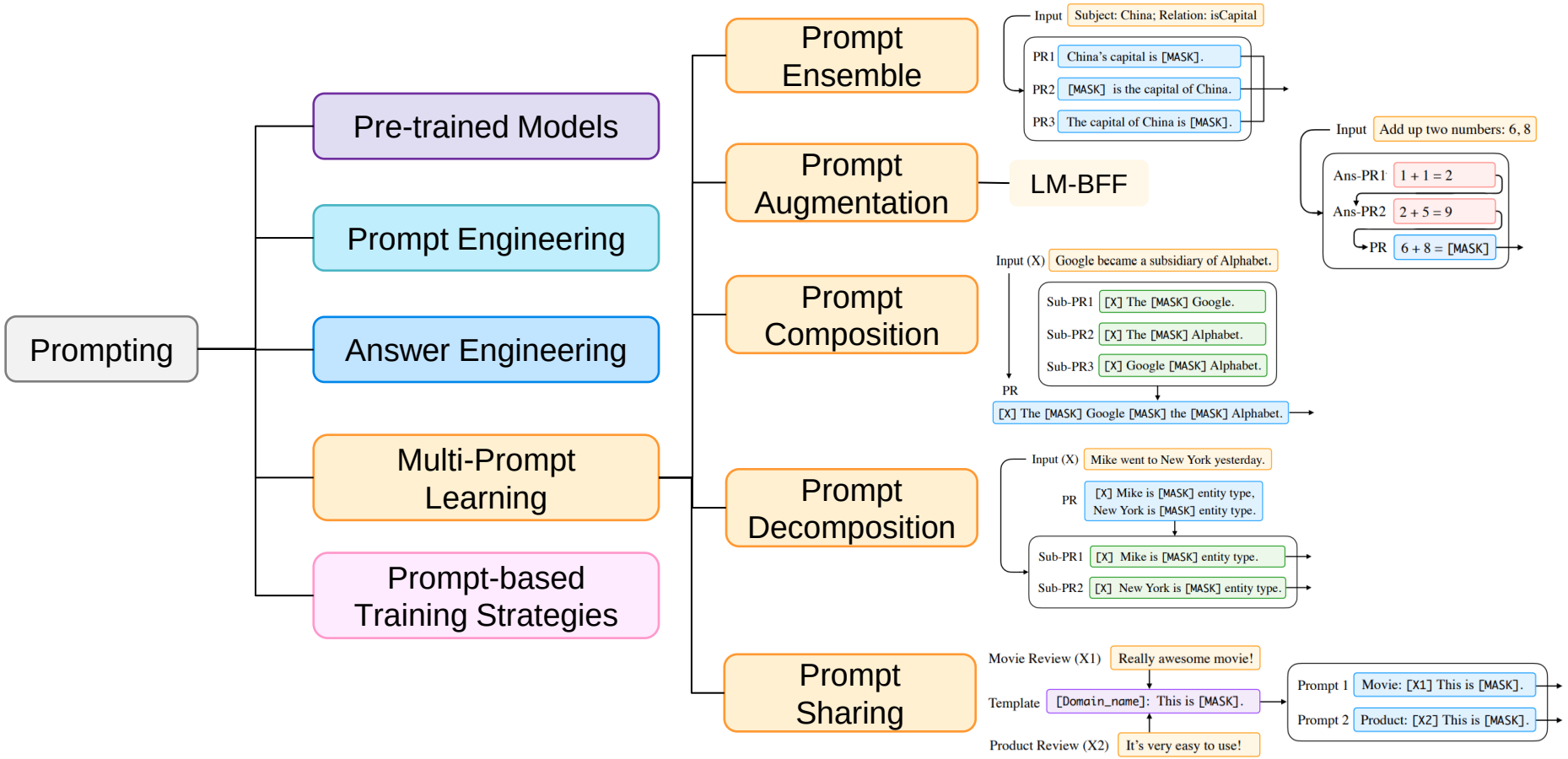
Prompting Typology (Liu et al., 2021)



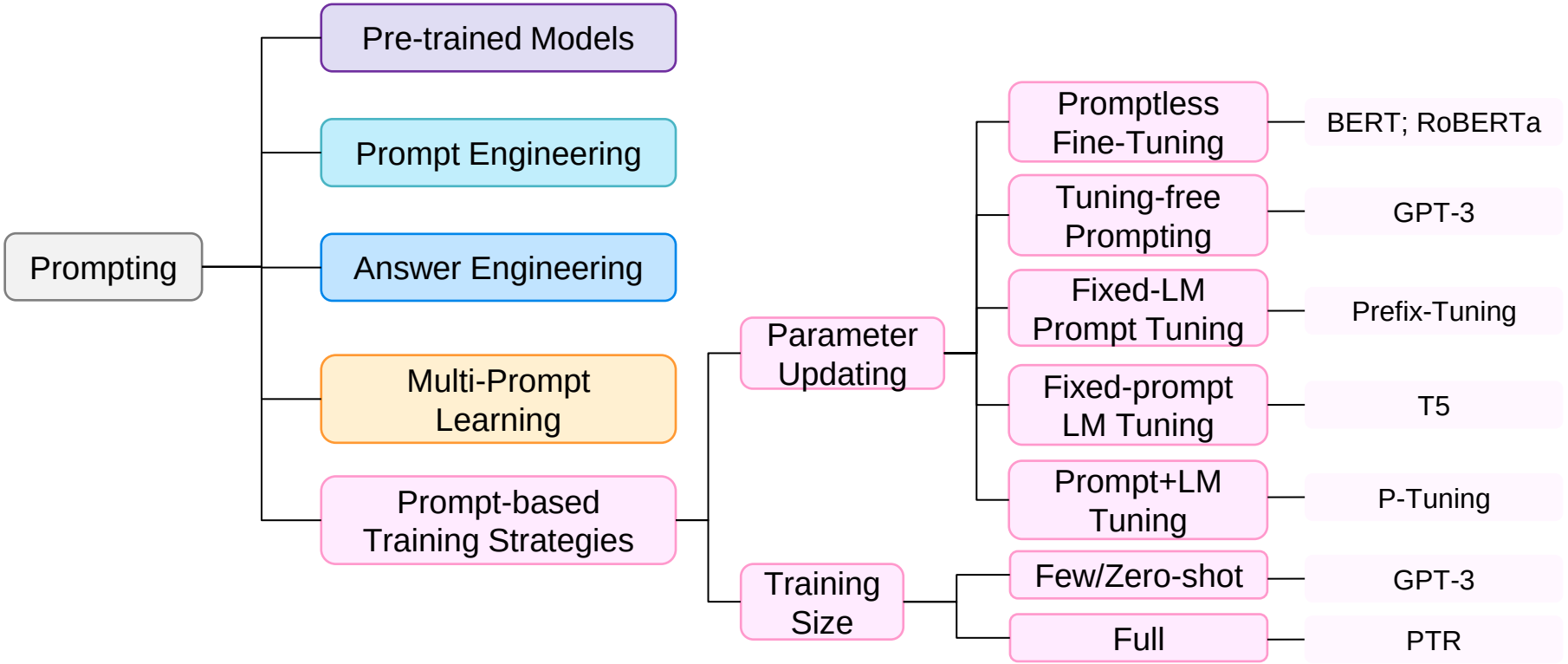
Prompting Typology (Liu et al., 2021)



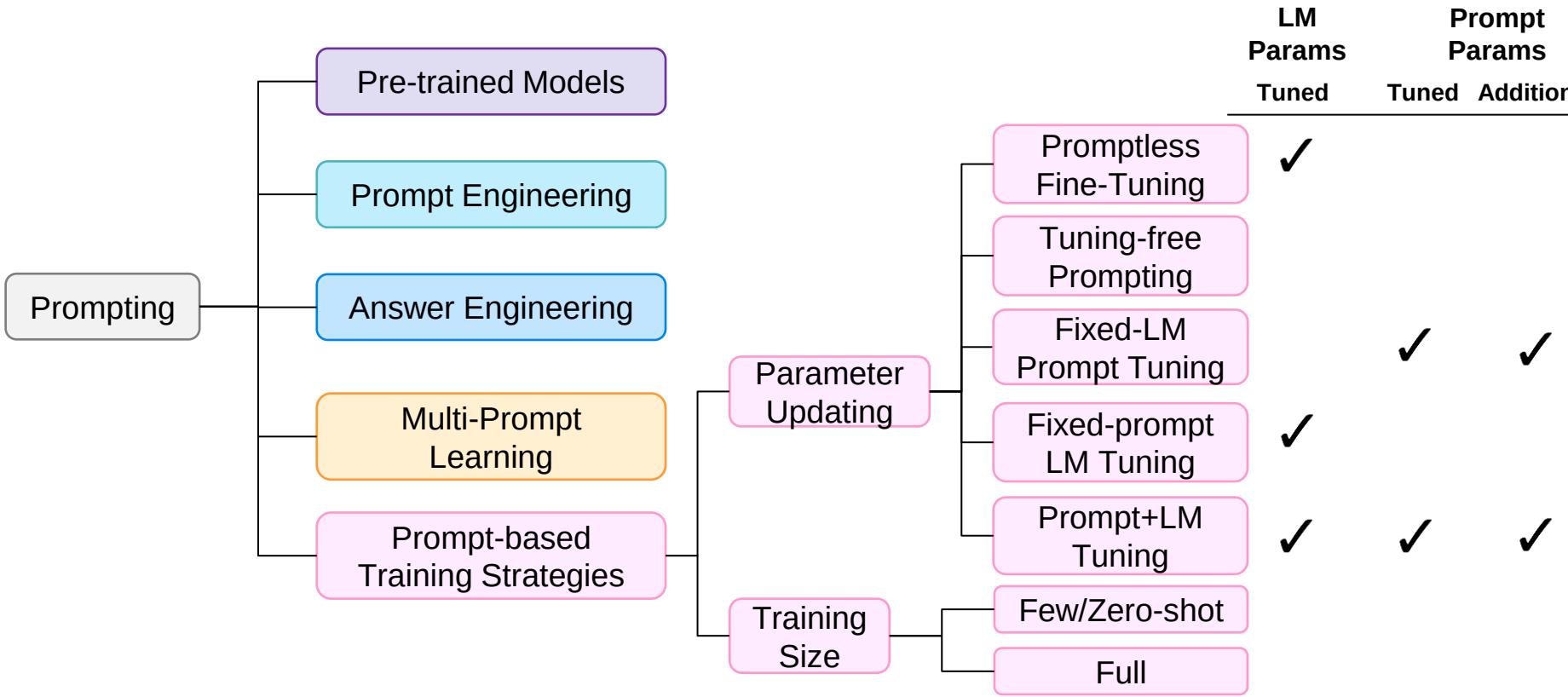
Prompting Typology (Liu et al., 2021)



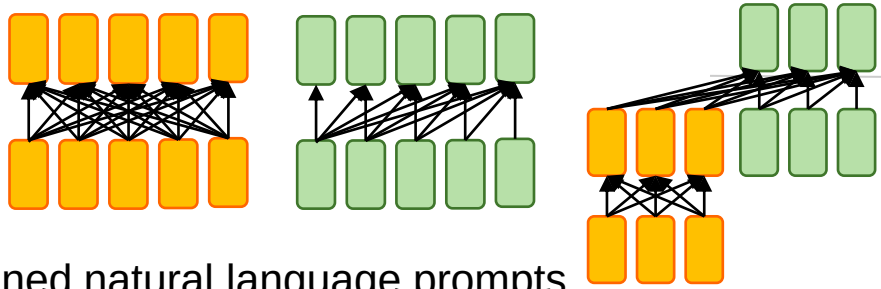
Prompting Typology (Liu et al., 2021)



Prompting Typology (Liu et al., 2021)



Concluding Remarks



- **(Hard) Prompt-Tuning**
  - **(Hard) Prompt-Tuning**: manually designed natural language prompts
    - Human-understandable prompts
    - Sensitive to choices of prompts
  - **LM-BFF**: prompt-tuning + demonstration + template generation
    - Better performance
- **(Soft) Prompt-Tuning**
  - **P-Tuning**: tuning the input (prompt) embeddings
    - Better performance via soft prompts
  - **Prefix-Tuning**: only optimize the prefix embeddings (all layers)
    - Better training time/space efficiency
- **Instruction Tuning**: tuning LMs for understanding task instructions
  - Better zero-shot performance