Prompt Tuning

昇思 MindSpore

课程目录

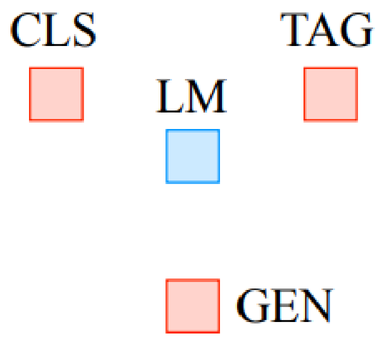
- 背景介绍：为什么需要prompt?
- prompt和prompting
- prompt tuning
  - hard prompt tuning
  - soft prompt tuning

为什么需要Prompt?

NLP发展的“第四范式”

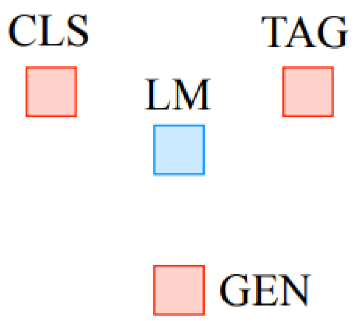
自然语言处理（NLP）发展：prompt兴起之前

监督学习（特征工程）



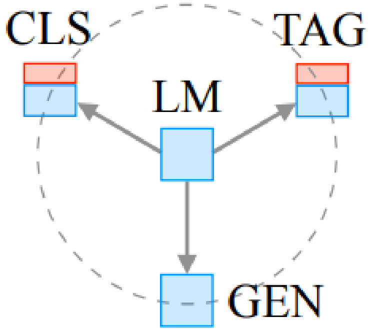
- 标注数据
- 依赖有专业领域知识的专家从数据中提取特征

监督学习（神经网络）



- 标注数据
- 神经网络从数据中提取特征并学习
- 重点在于设计网络结构

预训练，微调



- 无标注数据（预训练）
- 标注数据（微调）
- 不同任务之间的模型结构固定
- 通过微调预训练模型适配下游任务

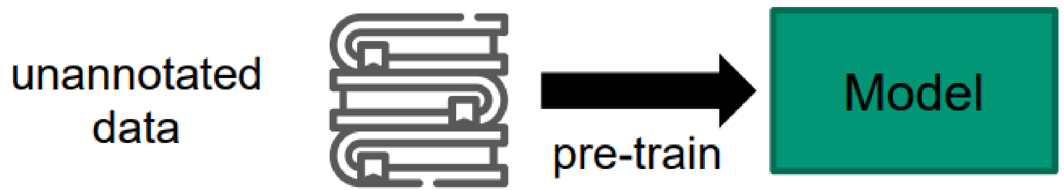
: fully unsupervised training

: fully supervised training

: fully unsupervised training

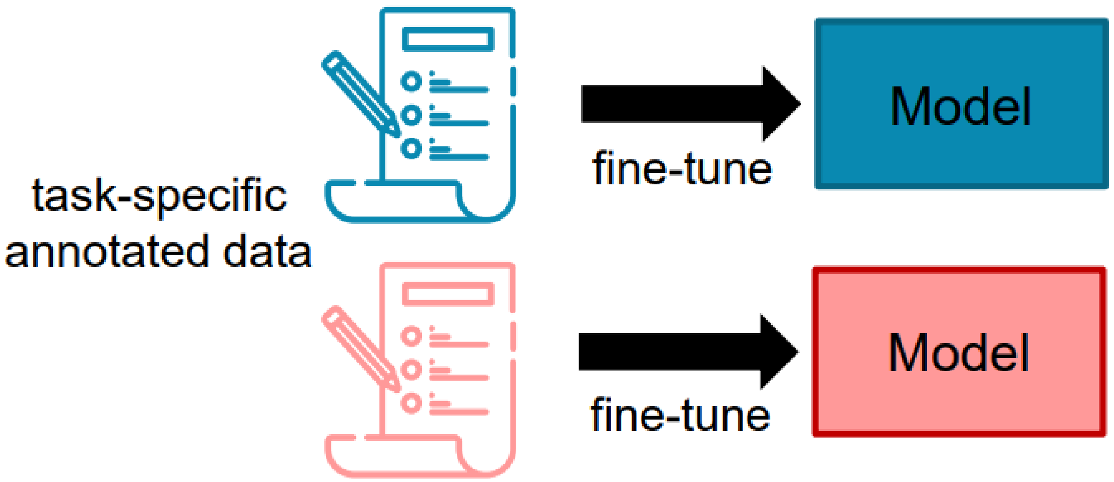
Pre-train, fine-tune

Pre-train



基于大量未标注数据进行无监督学习，得到具有一定通用能力的预训练模型

Fine-tune



针对特定下游任务，基于**标注数据**对预训练模型进行微调

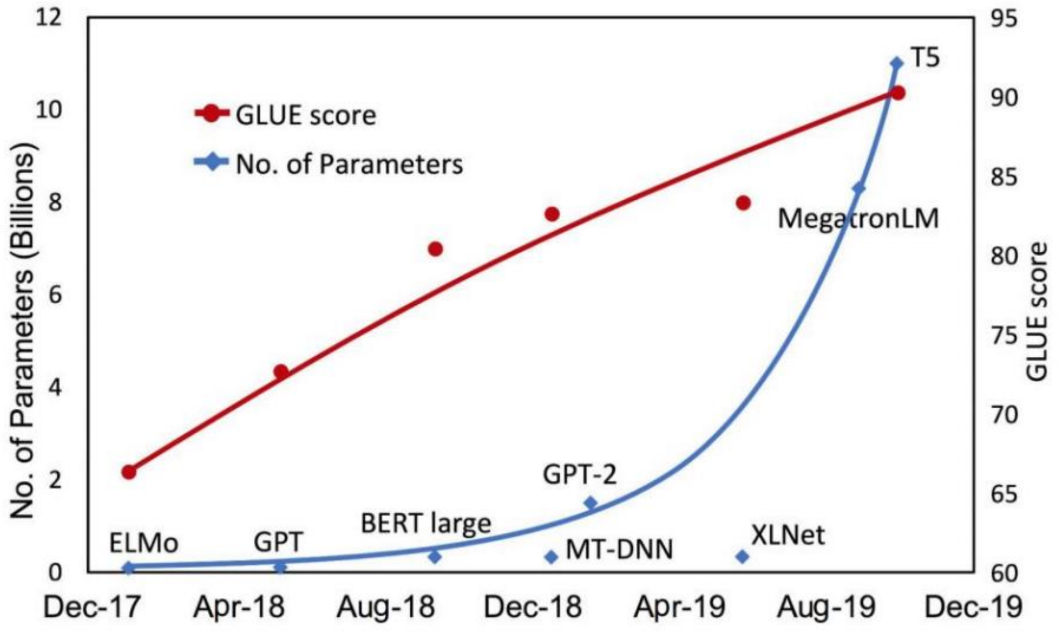
- **task-specific**：不同下游任务微调出的模型不同

Pre-train, fine-tune面临的挑战

Larger data, larger model, better performance：数据越多，参数量越大，模型效果越好

Year	Model	# of Parameters	Dataset Size
2019	BERT [39]	3.4E+08	16GB
2019	DistilBERT [113]	6.60E+07	16GB
2019	ALBERT [70]	2.23E+08	16GB
2019	XLNet (Large) [150]	3.40E+08	126GB
2020	ERNIE-GEN (Large) [145]	3.40E+08	16GB
2019	RoBERTa (Large) [74]	3.55E+08	161GB
2019	MegatronLM [122]	8.30E+09	174GB
2020	T5-11B [107]	1.10E+10	745GB
2020	T-NLG [112]	1.70E+10	174GB
2020	GPT-3 [25]	1.75E+11	570GB
2020	GShard [73]	6.00E+11	–
2021	Switch-C [43]	1.57E+12	745GB

Table 1: Overview of recent large language models

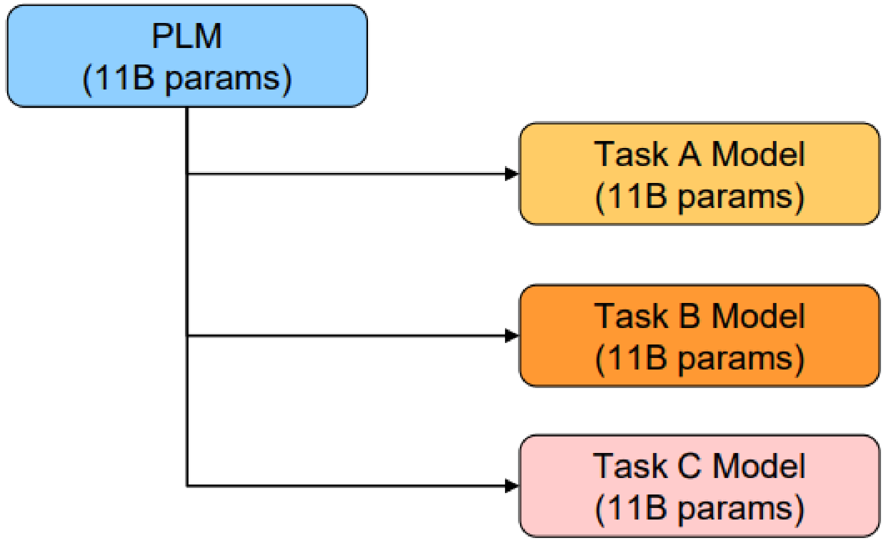


Pre-train, fine-tune面临的挑战

- **Data Scarcity**: 需要一定规模的标注数据才能使模型的微调达到较好的效果

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

- **High training cost and space requirement**: 每一个下游任务都需要重新微调出一个模型，对算力和空间的需求很大

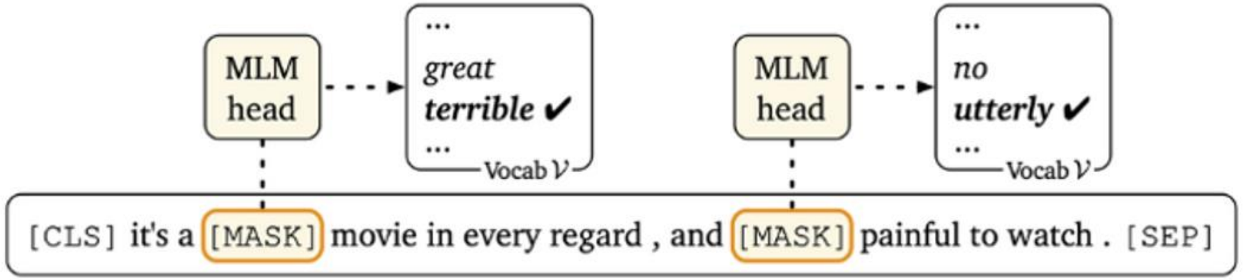


Pre-train, fine-tune面临的挑战

- **Gap between pre-train and fine-tune task**：预训练和微调的任务类型不一致，导致微调时不能很好地调用预训练模型的能力

以BERT为例：

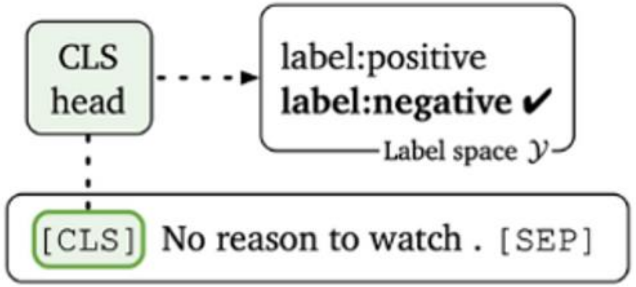
Pre-train



- **完型填空**
- 输出**自然语言文本**，形成完整通顺的句子

$\neq$

Fine-tune



- **情感分类**
- 输出**标签**，对文本情感色彩进行分类

Pre-train, fine-tune到pre-train, prompt, predict的过渡

pre-train, fine-tune范式中，预训练模型每次适配下游任务都需要基于不同的数据集微调出不同的模型

- 对数据量、算力、空间要求高
- 由于微调和预训练任务不同，导致预训练模型的能力无法完全发挥出来

能否不对模型权重进行修改，仅依赖预训练模型本身的能力，就可以使预训练模型适配到不同的下游任务？

Prompt!!!

Prompt和Prompting

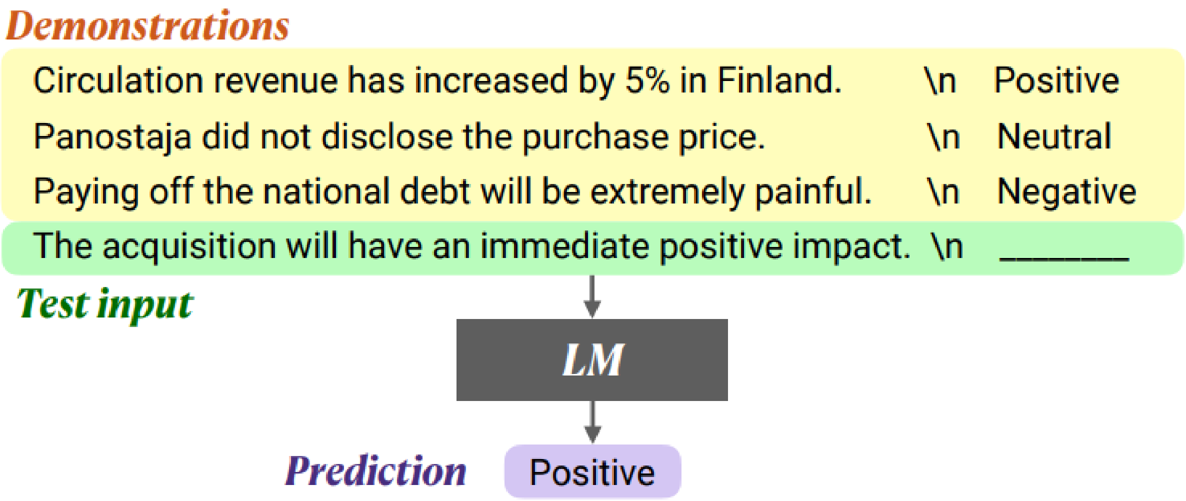
从模型迁就任务到任务迁就模型

关于预训练模型能力的测试：what makes in-context learning work?

能否不对模型权重进行修改，仅依赖预训练模型本身的能力，就可以使预训练模型适配到不同的下游任务？

- 预训练模型本身是否有能力来执行任务？

以情感分类任务为例：

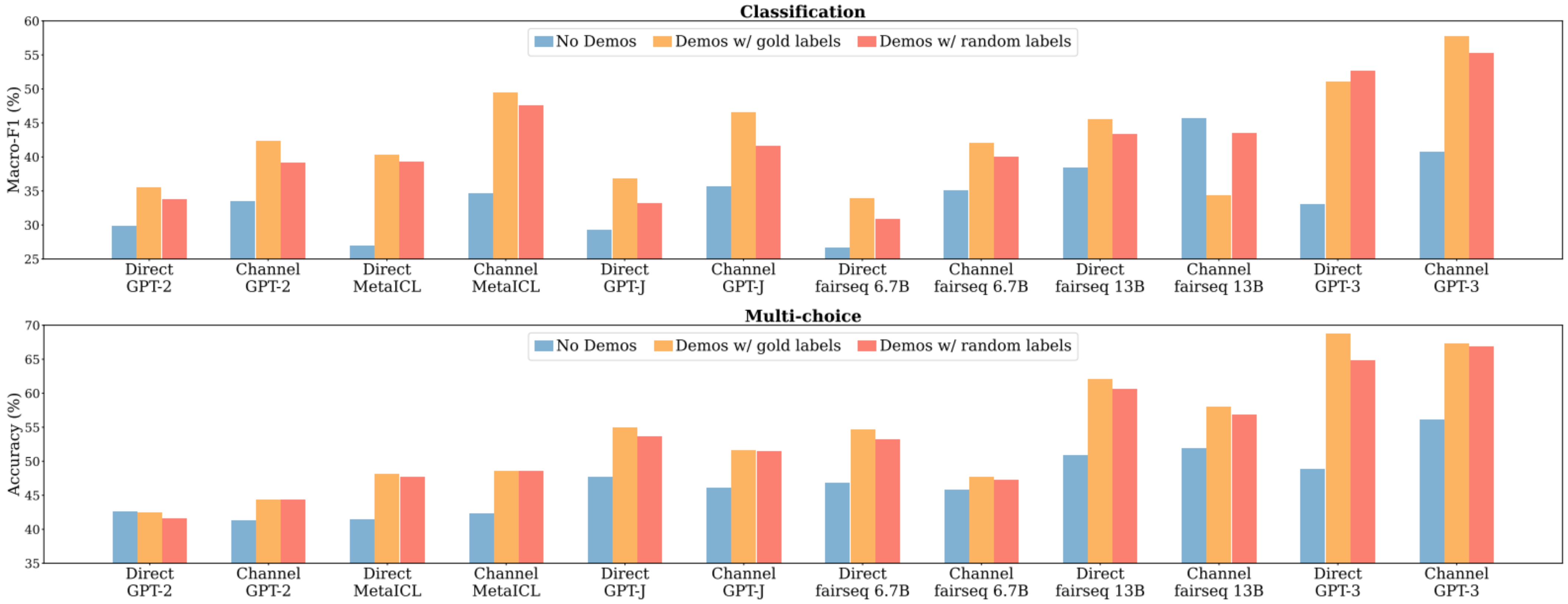


测试一 **Gold label vs random label**

Q：给定一些带有错误的标签的例子，模型预测的正确率如何？

关于预训练模型能力的测试：what makes in-context learning work?

测试一 Gold label vs random label：给定一些带有错误的标签的例子，模型预测的正确率如何？

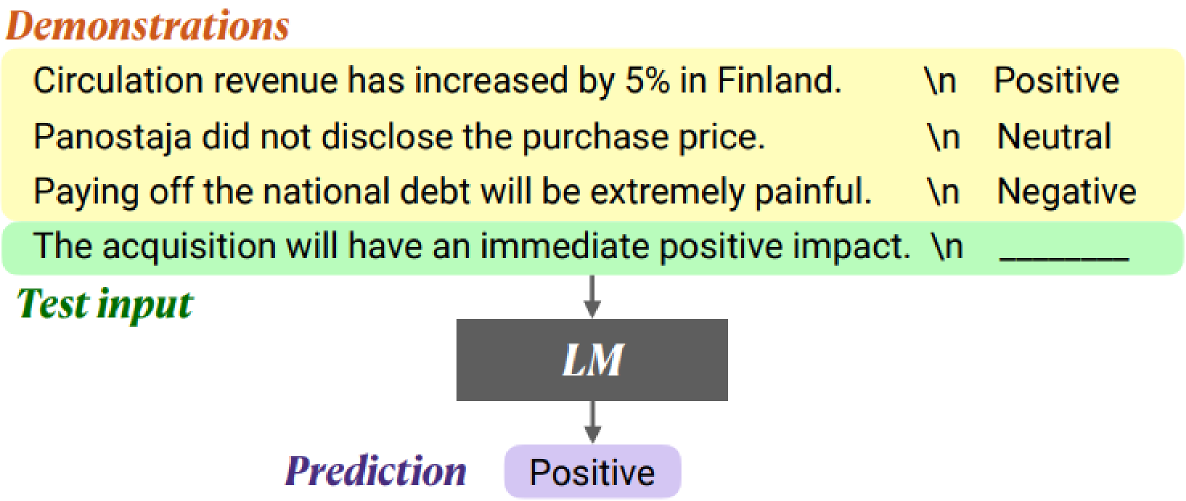


关于预训练模型能力的测试：what makes in-context learning work?

能否不对模型权重进行修改，仅依赖预训练模型本身的能力，就可以使预训练模型适配到不同的下游任务？

- 预训练模型本身是否有能力来执行任务？

以情感分类任务为例：



测试一 **Gold label vs random label**

Q：给定一些带有错误的标签的例子，模型预测的正确率如何？

A：模型的预测准确率并没有因为标签错误而收到多少的影响

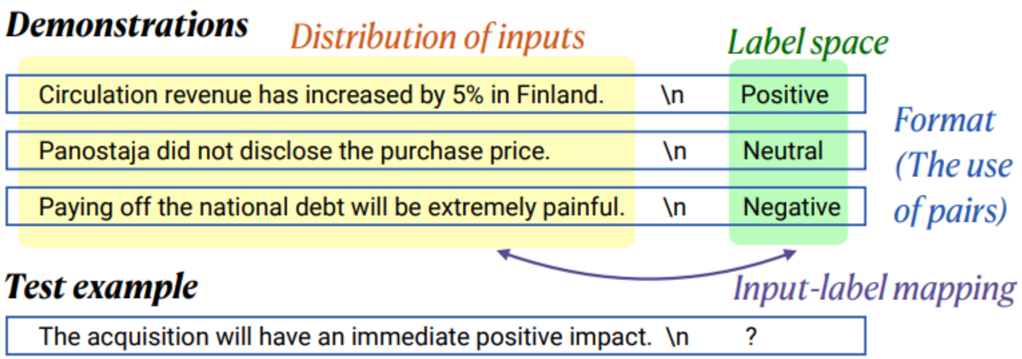
测试二 **OOD demonstrations**

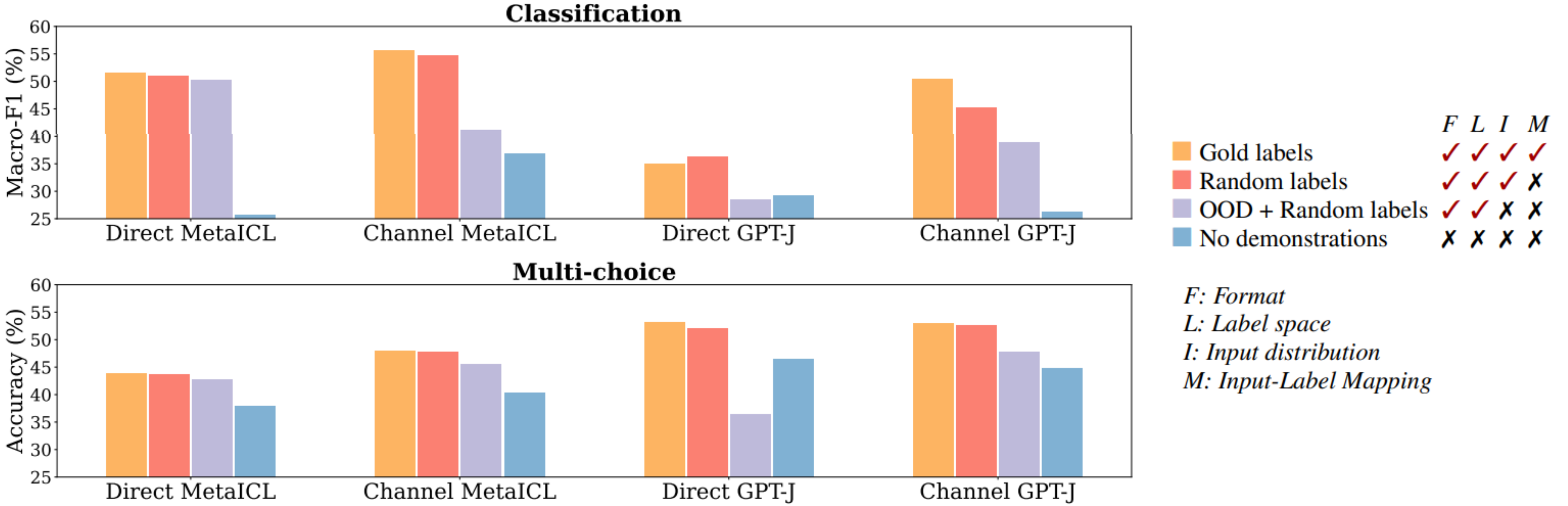
Q：给定一些和任务无关的输入，模型预测的正确率如何？

关于预训练模型能力的测试：what makes in-context learning work?

测试二 OOD demonstrations：

给定一些和任务无关的输入，模型预测的正确率如何?



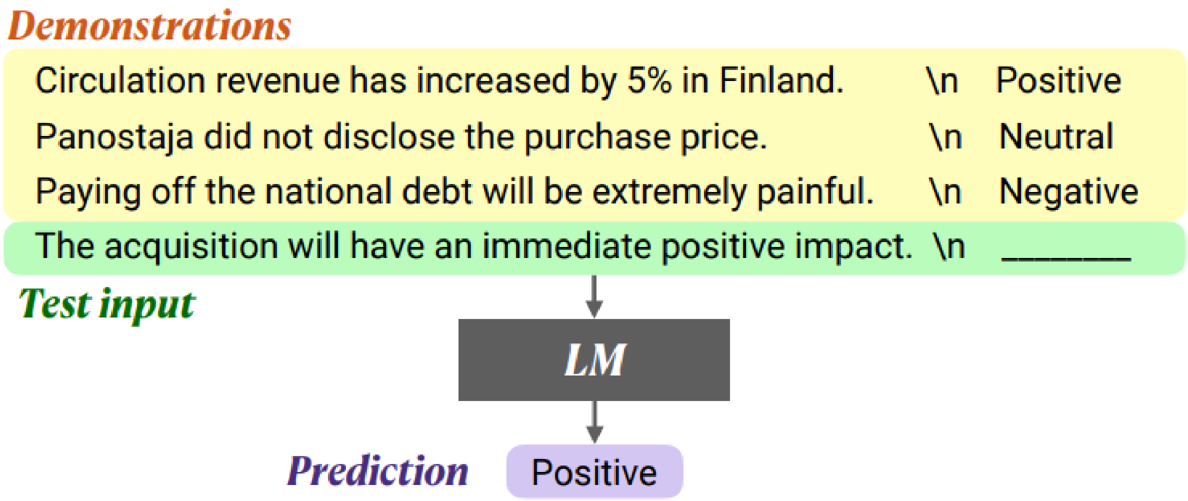


关于预训练模型能力的测试：what makes in-context learning work?

能否不对模型权重进行修改，仅依赖预训练模型本身的能力，就可以使预训练模型适配到不同的下游任务？

- 预训练模型本身是否有能力来执行任务？

以情感分类任务为例：



预训练模型本身具有执行任务的能力，
但需要“提示”模型现在正在处理什么样的任务

测试一 **Gold label vs random label**

Q：给定一些带有错误的标签的例子，模型预测的正确率如何？

A：模型的预测准确率并没有因为标签错误而收到多少的影响

测试二 **OOD demonstrations**

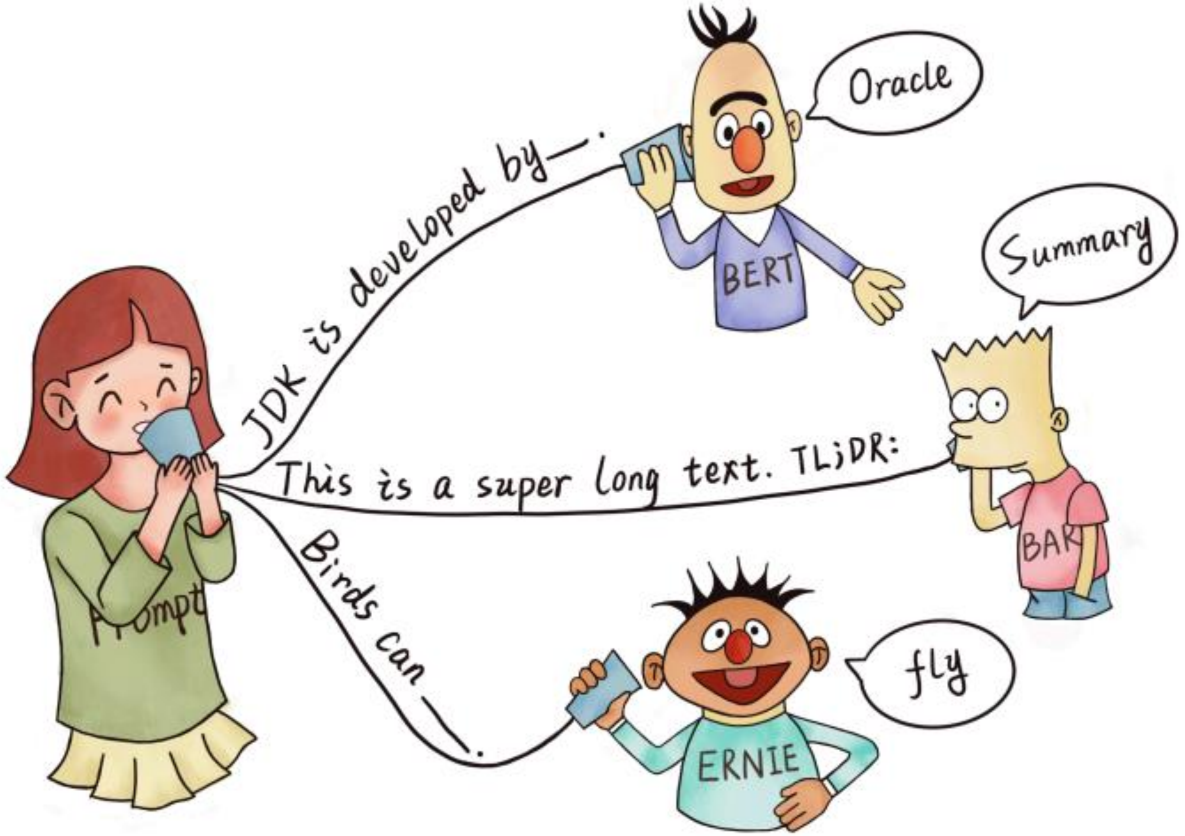
Q：给定一些和任务无关的输入，模型预测的正确率如何？

A：当输入和任务无关时，相比之前的任务相关输入，模型表现有一定下降

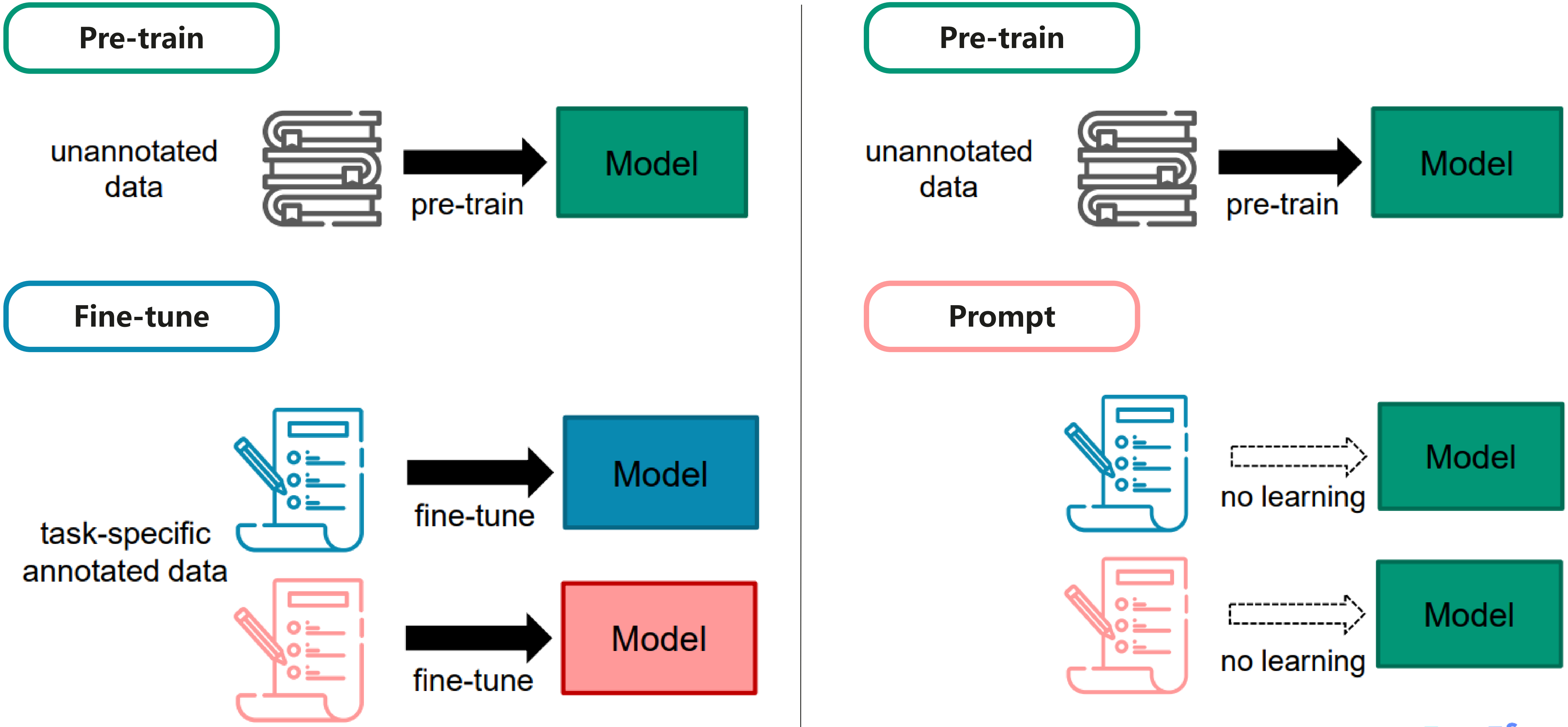
什么是prompt?

通过提供明确任务要求的“提示”（prompt），来引导预训练模型对特定任务进行输出。

- 可以理解为将所有任务都表述为“完型填空”问题
- 模型不做任何任何改动，直接根据提示进行预测

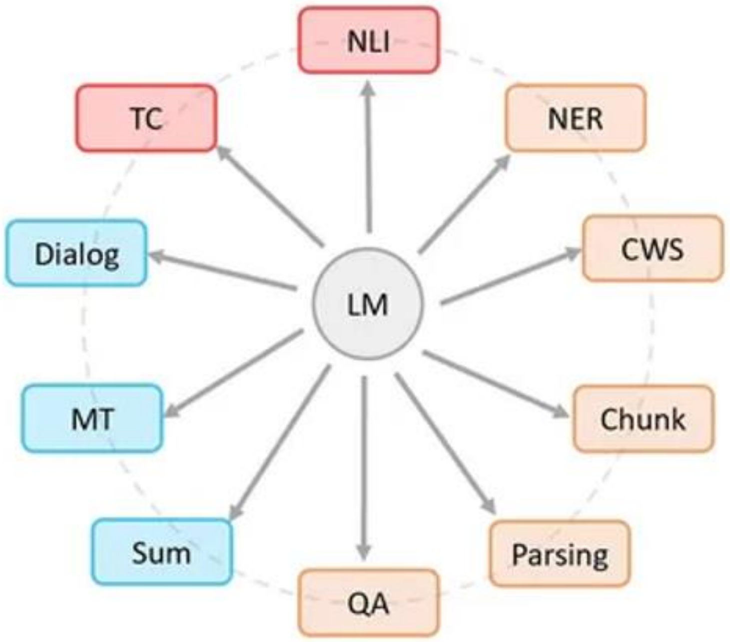


Pre-train, fine-tune vs pre-train, prompt, predict



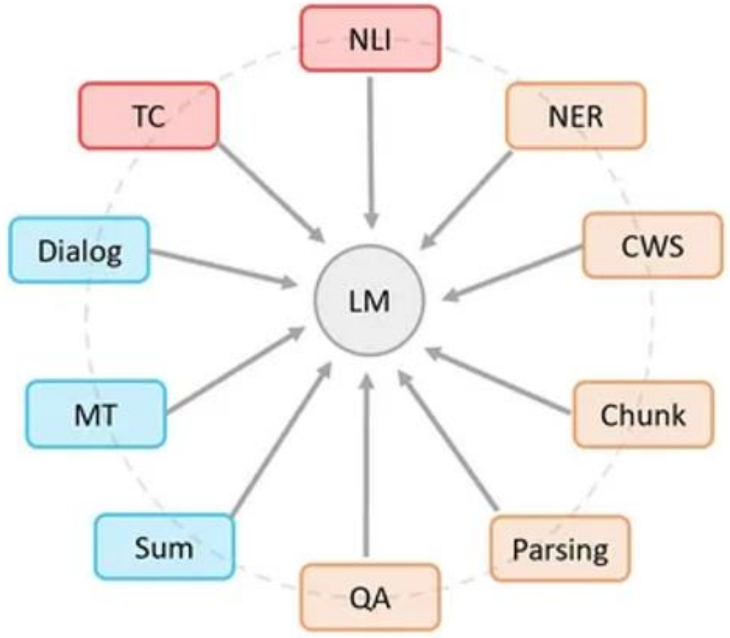
Pre-train, fine-tune vs pre-train, prompt, predict

Finetuning



模型迁就任务：通过对模型权重进行更新，使模型适配下游任务

Prompting



任务迁就模型：将不同下游任务描述统一为自然语言表述格式，模型权重不变

Prompt如何应用到不同任务

自然语言生成（NLG）

Label

Prediction: Je

PLM: Pretrained LM

Prompt: English: I am a student. French: [MASK]

自然语言理解（NLU）

Label: positive

positive	great, **fantastic**...
negative	bad, terrible...

Prediction: fantastic

PLM: Pretrained LM

Prompt: I love this movie. The movie is [MASK]

Prompting全流程

如何设计prompt很重要

Input: x="I love this movie"

⇩

Template: [x] Overall, this movie is [z]

⇩

Prompting: x'="I love this movie. This movie is [z]."

⇩

Predicting: x'="I love this movie. This movie is fantastic."

⇩

Mapping (answer ->label): fantastic => positive

Prompt Template

对于输入x，我们通过以下两步将其转换为prompt:

- **定义prompt模板**，模板中需包含两处带填充的空缺，一处是输入位[x]，一处是答案位[z]
- 在输入位填充输入x

模型预测

预训练模型基于prompt，预测答案位[z]的内容

Verbalizer

将模型的预测结果（自然语言表示），映射至对应标签

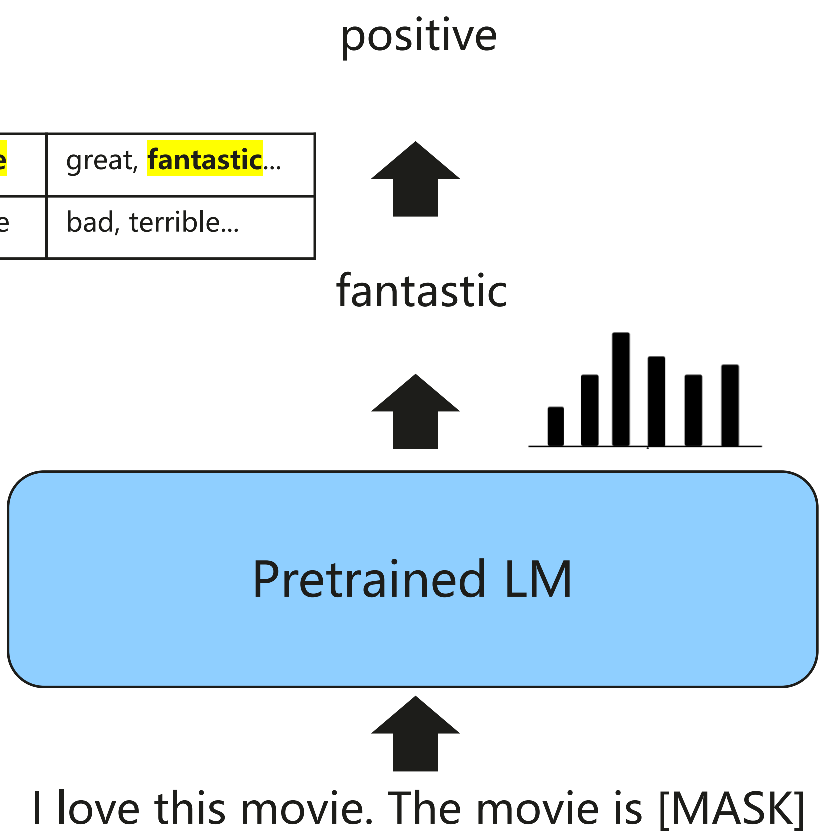
Hard prompt tuning

自然语言提示模型输出

Hard prompt tuning

将任务表示为自然语言提示，通过调整input prompt，使得模型输出更好的预测结果。

positive	great, fantastic...
negative	bad, terrible...



Verbalizer 如何将模型预测结果和标签相对应

Pretrained LM Masked LM or Autoregressive LM

Prompt Template 怎样设计贴合任务、预训练模型的模板

Prompt based on task

Type	Task	Input([X])	Template	Answer([Z])
Text CLS	Sentiment	I love this movie.	[X] The movie is [Z].	great, fantastic, ...
	Topics	He prompted the LM.	[X] The text is about [Z].	sports, science, ...
	Intention	What is taxi fare to Denver?	[X] The question is about [Z].	quantity, city, ...
Text-span CLS	Aspect Sentiment	Poor service but good food.	[X] What about service?[Z].	Bad, Terrible, ...
Text-pair CLS	NLI	[X1]:An old man with ... [X2]:A man walks ..	[X1]?[Z],[X2]	Yes, No, ...
Tagging	NER	[X1]:Mike went to Paris. [X2]:Paris	[X1][X2] is a [Z] entity.	organization location ...
Text Generation	Summarization	Las Vegas police ..	[X] TL;DR:[Z]	The victim . A woman ...
	Translation	Je vous aime.	French:[X] English:[Z]	I love you. I fancy you. ...

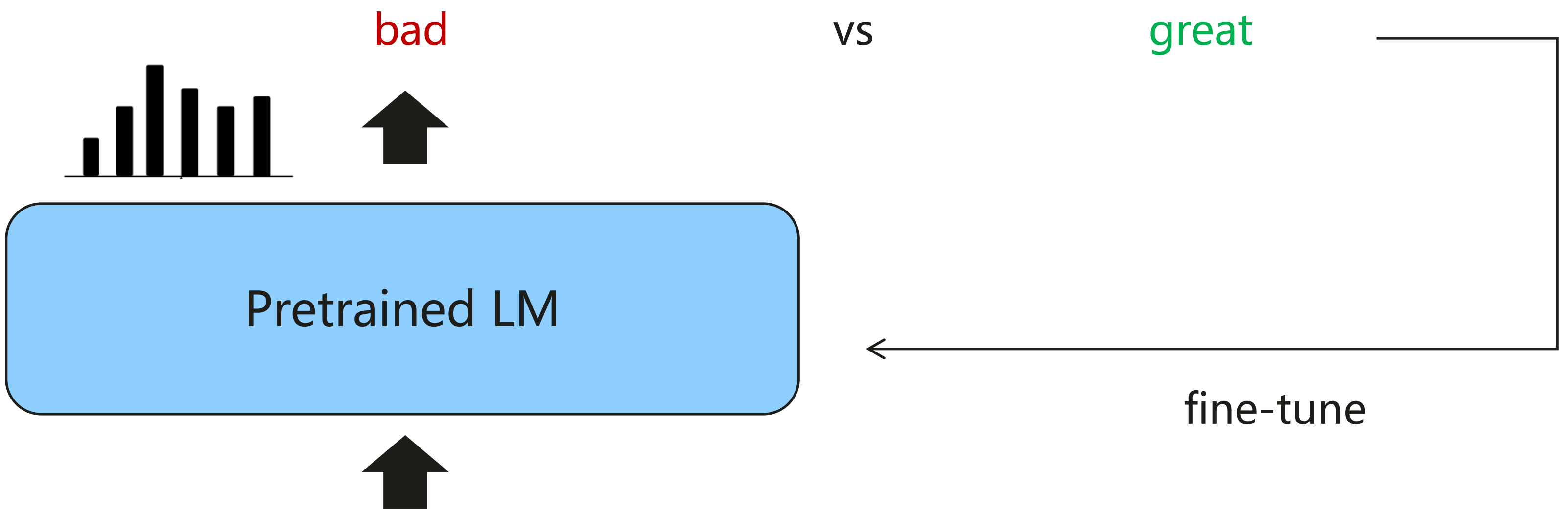
Prompt based on pretrained LM

Model	Template shape	Template	Answer([Z])
Masked LM（BERT...)	cloze prompt	[X] Overall it is a [Z] movie.	great, fantastic, ...
Autoregressive LM（GPT...)	prefix prompt	[X] The movie is [Z].	great, fantastic, ...

让预训练模型“学习”提示信息（few shot场景可用）

如果不存在标注数据，直接prompt提示模型推理（zero shot），模型的权重并不会变动。

如果有少量的标注数据，也可以通过这样prompt tuning的方法对预训练模型进行微调（few shot）。



prompt I love this movie. The movie is [MASK]

examples

No reason to watch. The movie is terrible.

This movie is a nightmare. The movie is awful.

I would recommend this movie to my friends. The movie is amazing.

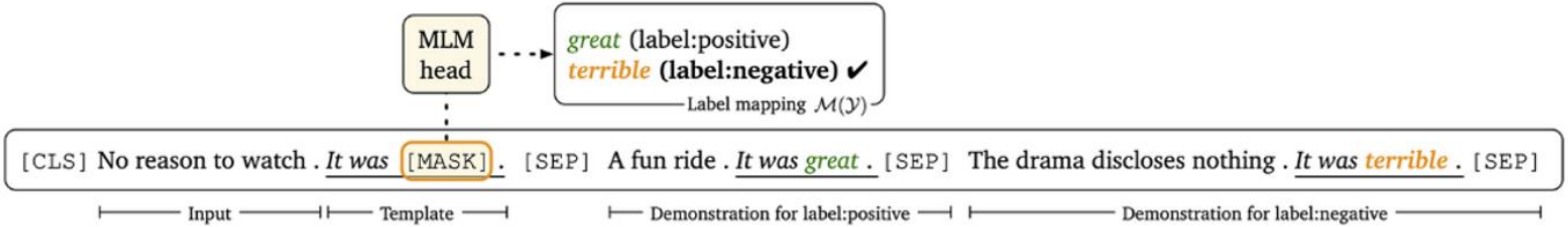
Hard Prompt的问题

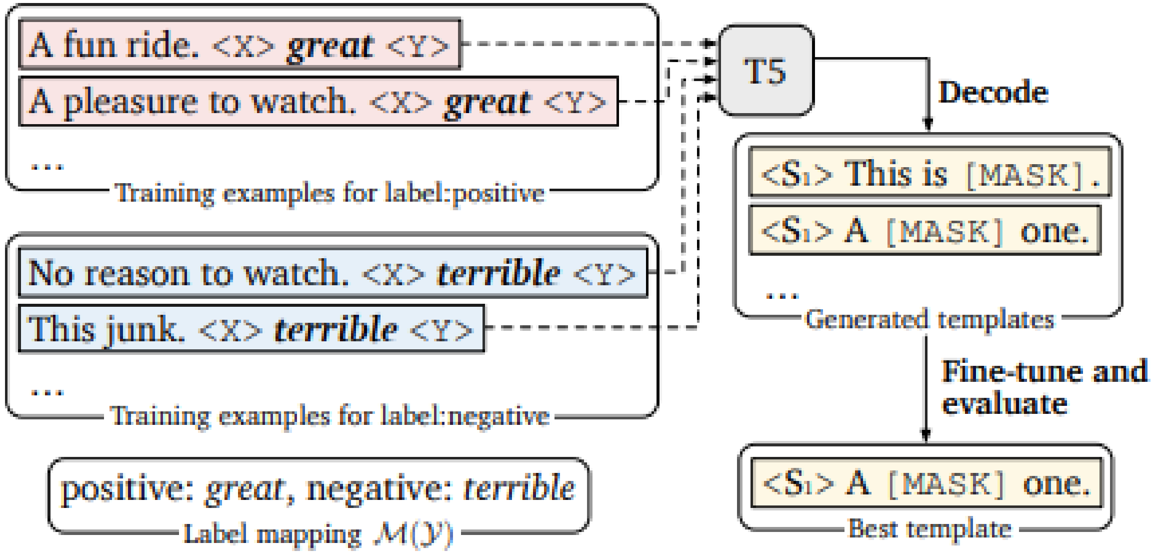
人类认为最优的prompt并不一定是最适合模型的prompt。

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

能否让模型选择最能激发自己能力的prompt?

LM-BFF：Better Few-Shot Fine-tuning of Language Models





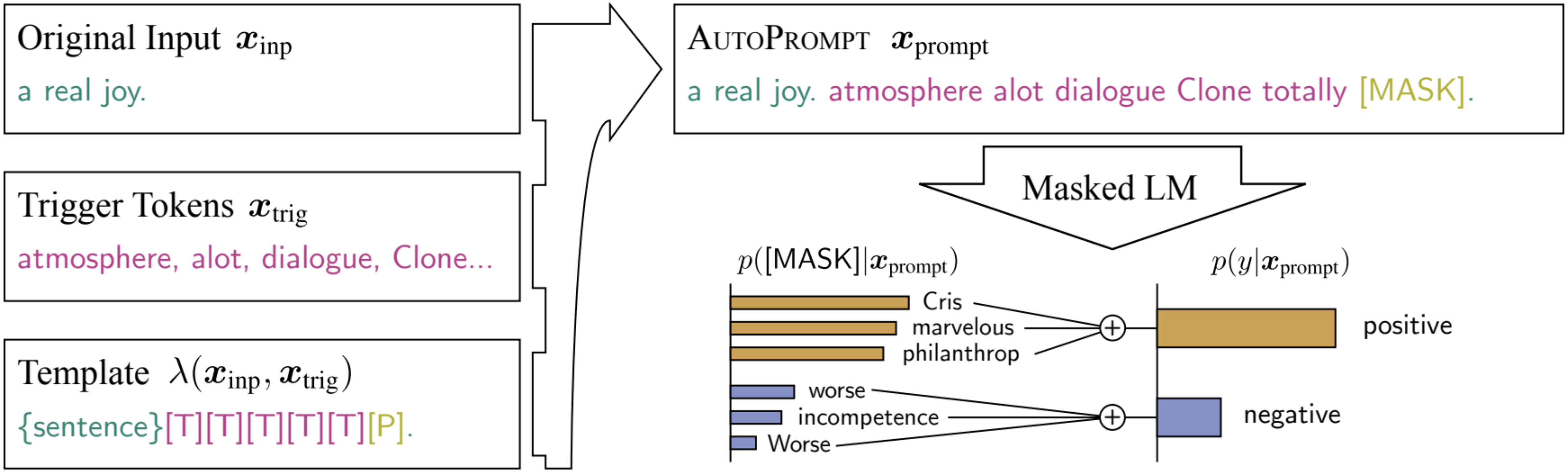
template generation：

- T5模型生成prompt模板
- 模板代入到标注数据中，比较哪个template的预测效果最好

AutoPrompt: 自动检索生成prompt

Gradient-based search of prompts based on existing tokens

- 定义trigger tokens集合，每个trigger token [T]通过[MASK]初始化，然后不断迭代更新
- 不用在意prompt是否可读，对模型效果好就行



Hard Prompt的问题

- 人工设计prompt的成本高
- 人类认为最优的prompt并不一定是最适合模型的prompt
- prompt的效果受制于模型输入的大小
- 预训练模型对hard prompt的选择较为敏感

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

Hard prompt特点：（人工设计）、离散、可读，但无法参与学习优化

真的有必要设计可读（human readable）的自然语言文本类型的prompt吗?

可不可以一切交给模型，人不理解没关系，模型能理解，甚至能进行优化就好。

Soft prompt tuning

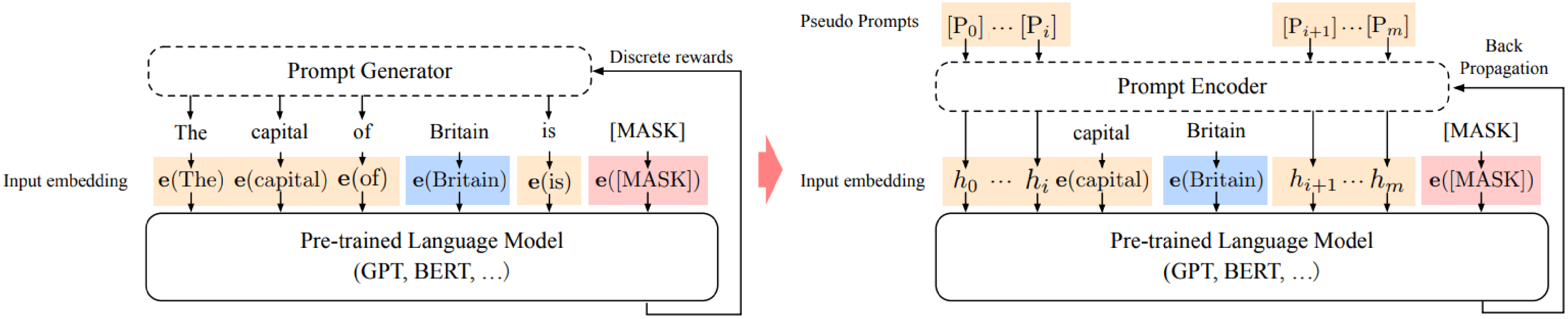
提示交给模型搞定

P-Tuning：从discrete prompt到continuous prompt的转变

direct optimize embedding：直接优化prompt template的embedding

- input embedding不变，与任务有关的archor token embedding不变
- 预训练模型权重不变
- prompt template的embedding进行优化

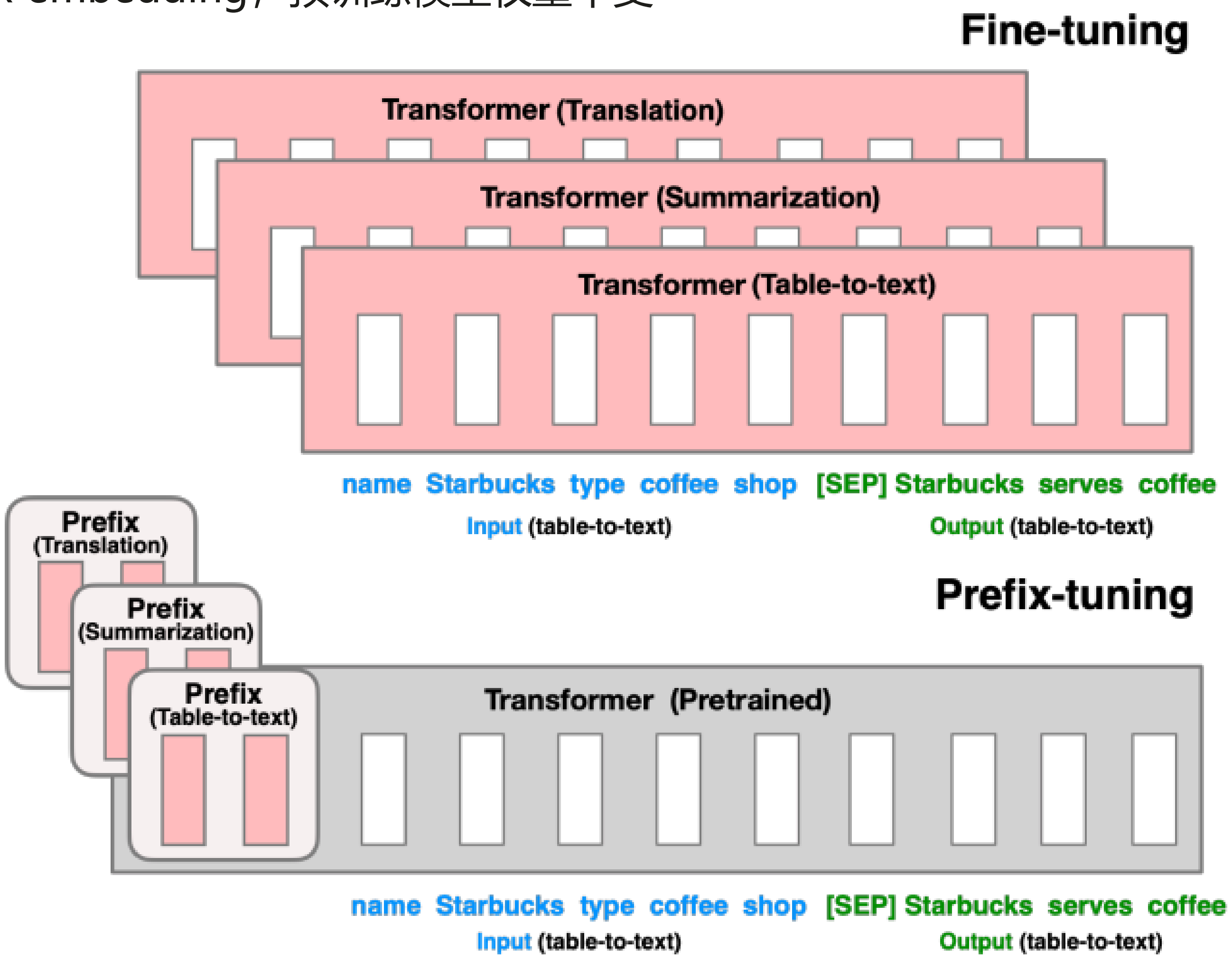
例: The capital of Britain is [MASK]



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

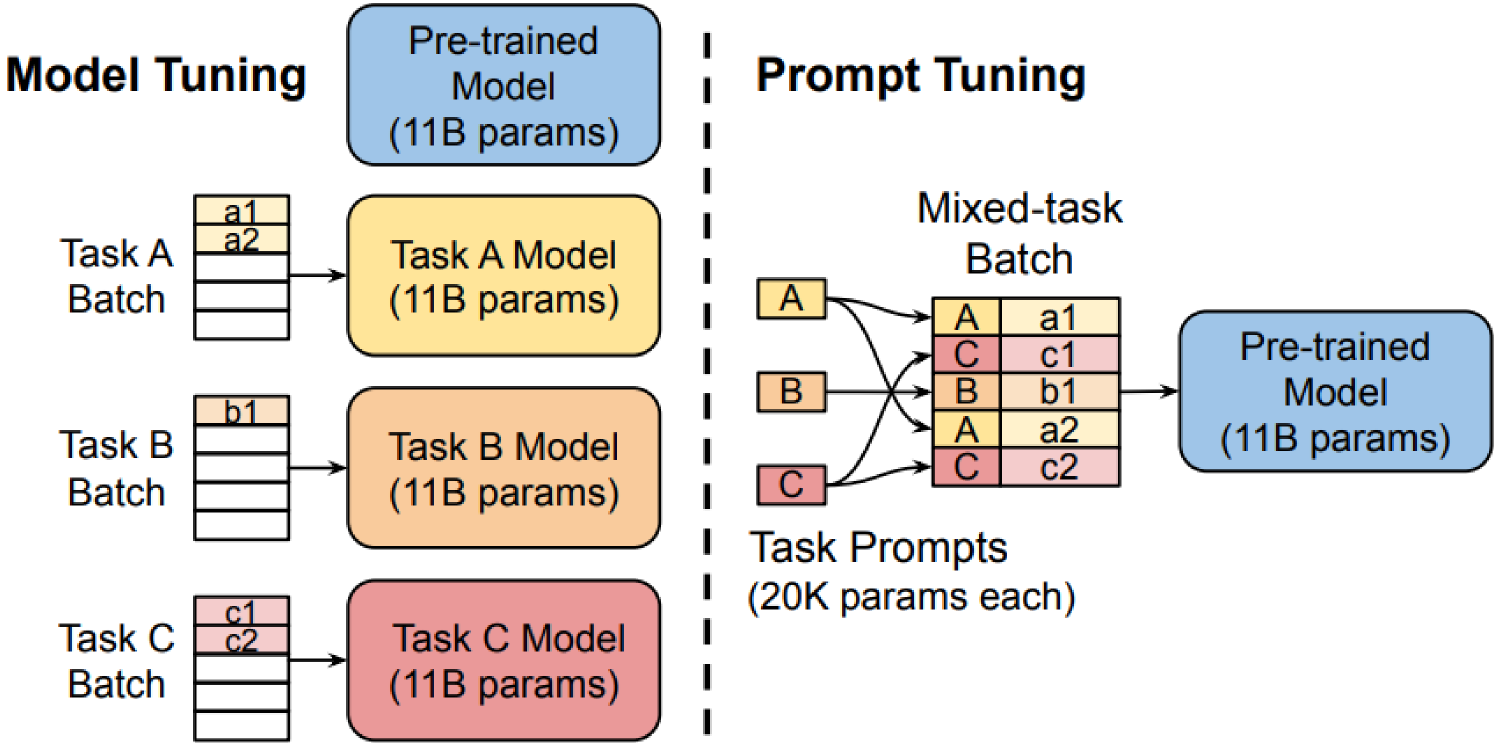
Prefix-Tuning：在预训练模型层中添加embedding

- 每一层前面添加prefix embedding
- 微调时只微调prefix embedding，预训练模型权重不变

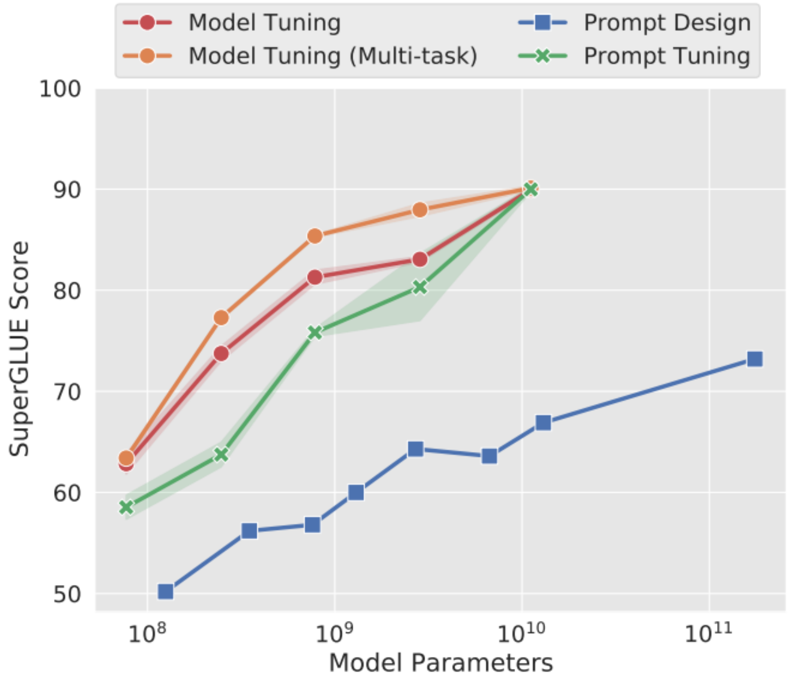


(Soft) Prompt-Tuning：仅在输入添加prompt embedding

- 仅在输入添加可学习的embedding作为prompt，且没有对预训练模型结构进行修改
- mixed-task inference

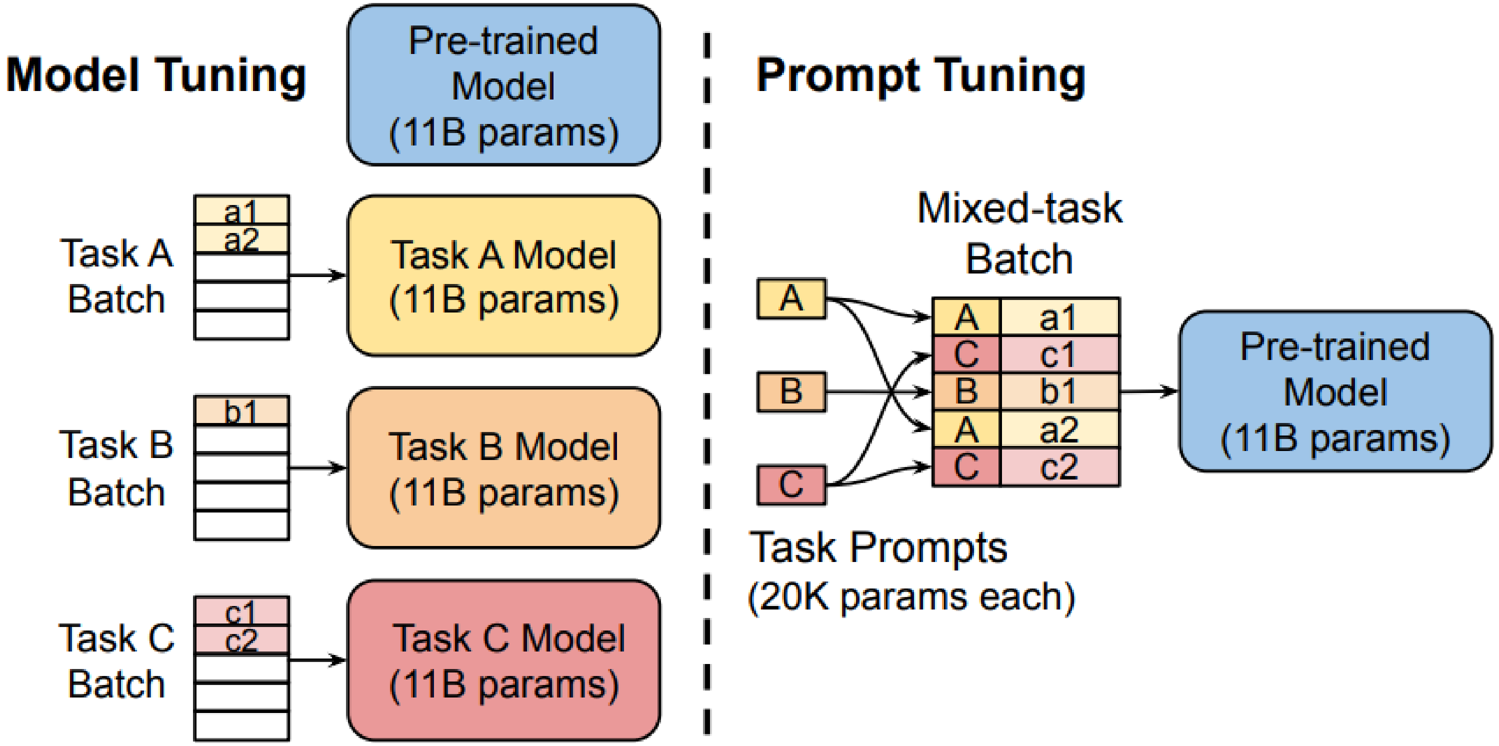


模型效果OK

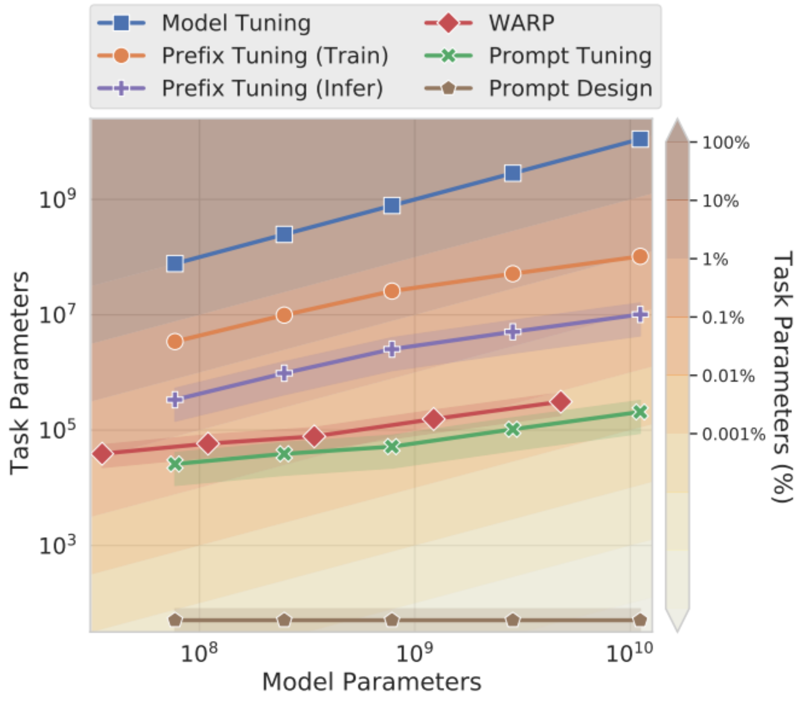


(Soft) Prompt-Tuning：仅在输入添加prompt embedding

- 仅在输入添加可学习的embedding作为prompt，且没有对预训练模型结构进行修改
- mixed-task inference



参数量少



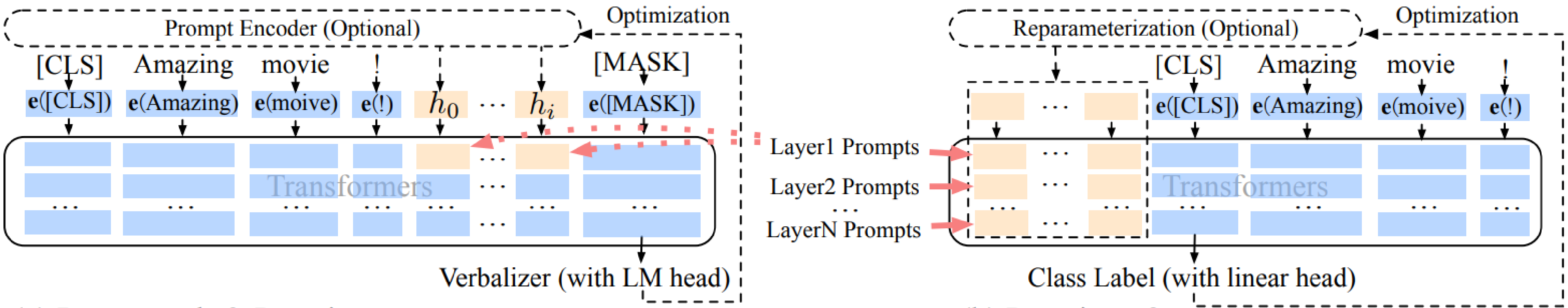
P-Tuning v2：P-Tuning的改进，纯continuous prompt

Limitation of P-Tuning

- 因为输入文本长度的限制，约束可学习优化的参数量
- 模型层数加深时，仅第一层的prompt对后续层影响难以预估，影响模型tuning的稳定性

P-Tuning v2

- 和prefix tuning类似，每层前添加可学习的prompt embedding
- 多任务学习



(a) Lester et al. & P-tuning (Frozen, 10-billion-scale, simple tasks)

(b) P-tuning v2 (Frozen, most scales, most tasks)

实操演练 —— Roberta模型prompt tuning

在线实验：

https://pangu.huaweicloud.com/gallery/asset-detail.html?id=016991f8-0e0d-44c8-96f7-8b2cad54c592

（在线实验运行指南请参考《昇思MindSpore在线实验指导手册——AI Gallery》）

Thanks