别再Prompt了！谷歌提出tuning新方法，强力释放GPT-3潜力！

原创 Yimin_饭煲 夕小瑶的卖萌屋 2021-09-07 22:20

收录于话题
#卖萌屋@自然语言处理 69个 >

文 | Yimin_饭煲
编 | 炼丹学徒

如果评选NLP圈的2020年度十大关键词，那么**GPT-3（Language Models are Few shot Learners)** 一定榜上有名。

GPT-3庞大的参数量，优异的性能至今仍让圈内圈外人都津津乐道，而OpenAI发布的OpenAI API，更是为自然语言处理技术的**大规模可扩展**商业应用提供了一个极有前景的方向。不过，作为NLP研究者，我认为GPT-3对前沿研究的最大贡献是，展现了 **Prompt-tuning** 技术在通用任务（**特别是零样本和小样本场景下**）上的应用潜力。在GPT-3之前，**Prompt-tuning** 大多仅被用来探索语言模型中蕴藏的世界知识，而GPT-3之后，**Prompt-tuning** 就“登堂入室”，被用到了各种类型的NLP任务上（甚至还有多模态任务、代码分析任务），成为了近两年来发Paper的一个热点。不太熟悉 **Prompt-tuning** 的读者可以参考CMU最新发布在arxiv上的综述 （**Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing**）。

卖萌屋之前也转载过该论文的中文解析 **《Fine-tune之后的NLP新范式：Prompt越来越火，CMU华人博士后出了篇综述文章》**

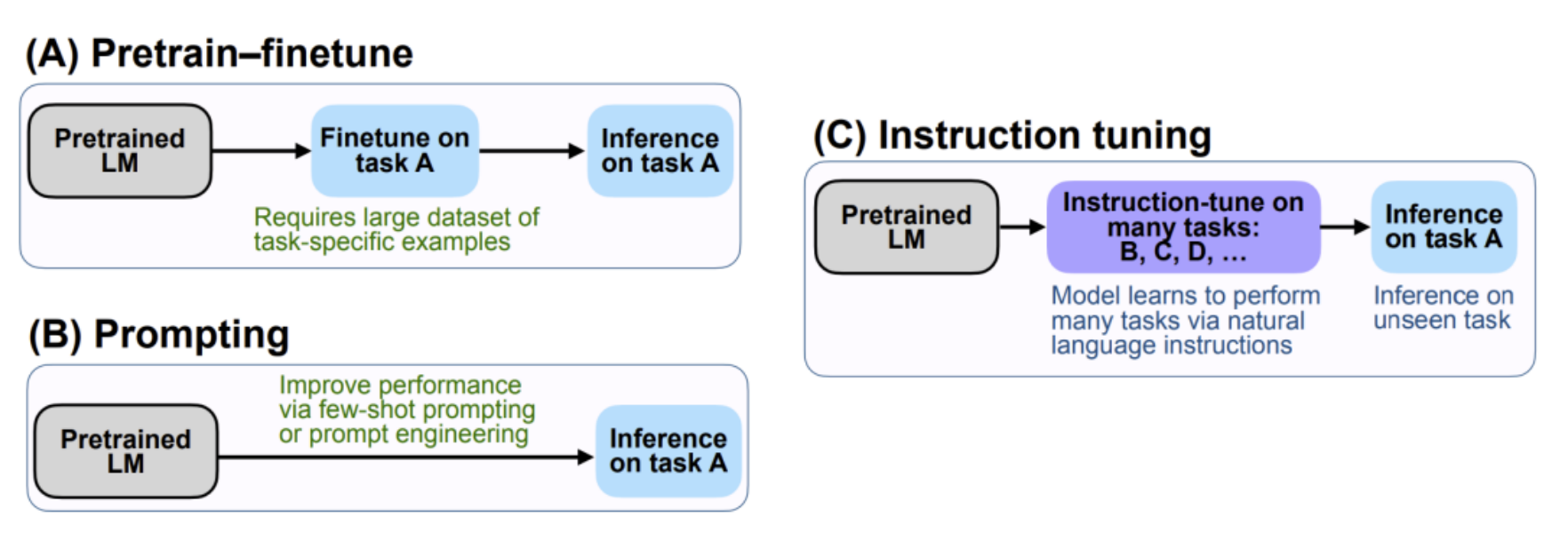
Prompt-tuning 和GPT-3互相成就，成为了NLP发展历史中不可忽略的里程碑。基于 **Prompt-tuning** 让GPT-3处理各种类型的任务，固然取得了不错的表现，但仔细想想，GPT-3这样一个创造了各种奇迹的巨大模型，是否还有 **更大的零样本和小样本学习能力** 尚未挖掘？**Prompt-tuning** 一定是利用GPT-3的最好方式吗？Google的研究员们不甘心止步于此，提出了 **instruction-tuning** ，利用比GPT-3更少的参数量，在25个任务中的19个上**显著超越GPT-3**，告诉世界：**GPT-3，还能更强！**

论文题目:
Finetuned Language Models Are Zero-Shot Learners

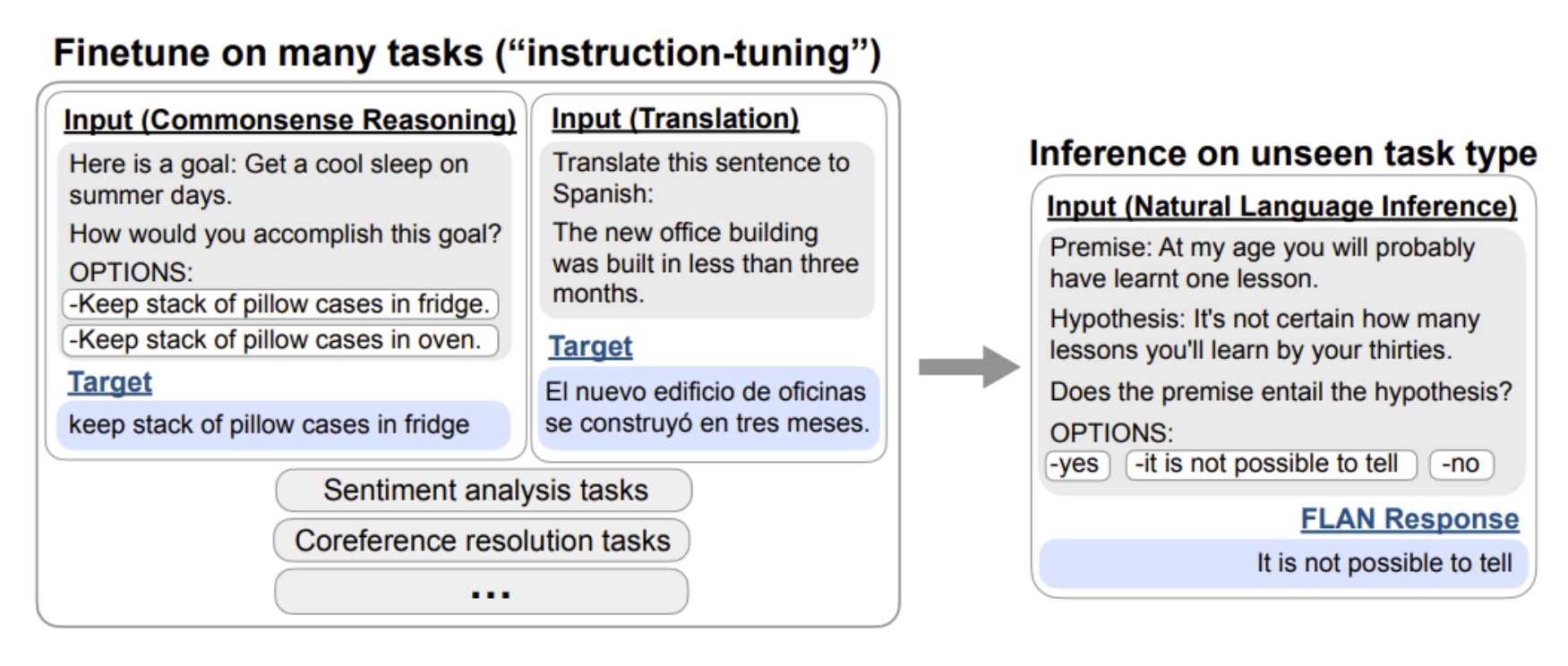
论文链接:
https://arxiv.org/abs/2109.01652

方法：FLAN

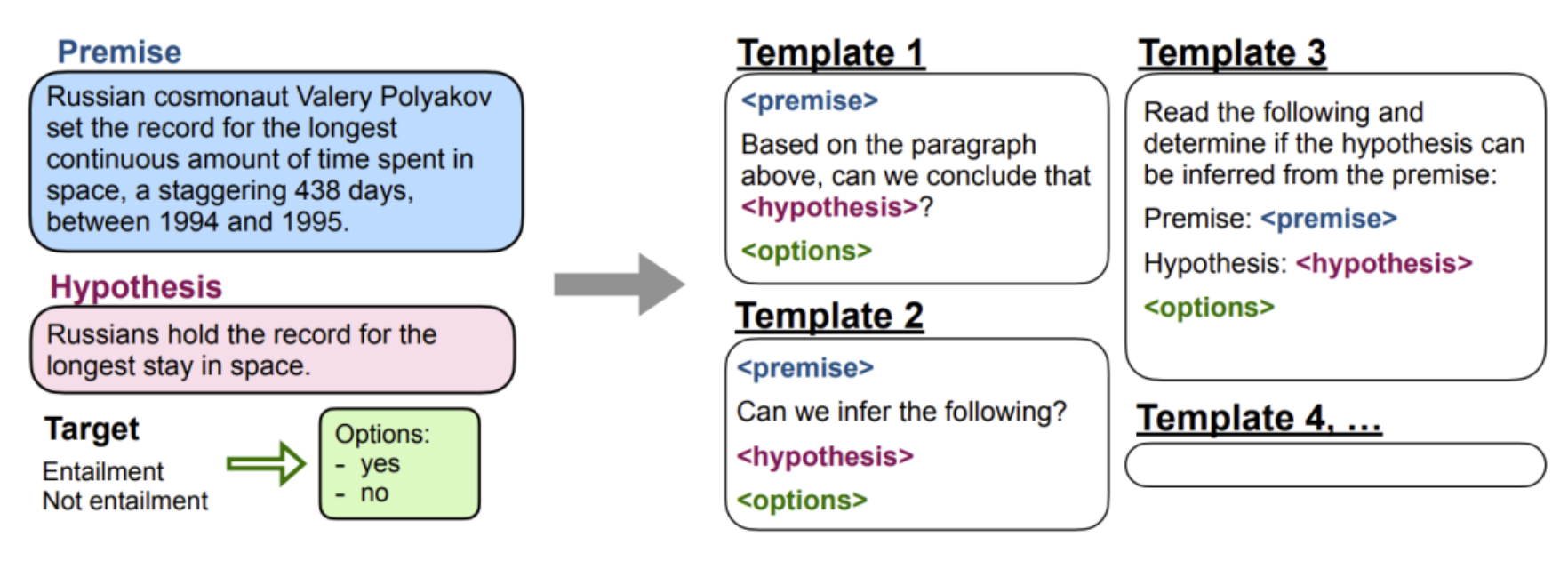
谷歌的研究员们将自己的方法取名为**FLAN** (**F**inetuned **LAN**guage Models are zero-shot Learners)，相比于GPT-3（**LAN**guage Models are zero-shot Learners），区别在于**F**inetune. **FLAN**的核心思想是，当面对给定的任务A时，首先将模型在大量的其他**不同类型**的任务上进行微调，微调的方式是将**任务的指令**与数据进行拼接（可以理解为一种Prompt)，随后**给出任务A的指令，直接进行推断**。具体示例可见下图。



例如，我们的最终目标是**推理任务**，FLAN首先讲语言模型在**其他任务上**进行微调，包括给定任务指令的**翻译**、**常识推理**、**情感分类**等。在面对翻译任务时可以给出指令“**请把这句话翻译成西班牙语**”。在面对常识推理任务时可以给出指令“**请预测下面可能发生的情况**”，而当模型根据这些“指令”完成了微调阶段的各种任务后（将指令拼接在微调数据的前面），在面对从未见过的最终需要的**自然语言推理任务的指令“这段话能从假设中推导出来吗？”** 时，就能更好地调动出已有的知识回答问题。



在GPT-3等研究***Prompt-tuning*的工作中，研究者们发现*Prompt的质量*对模型的性能有着较大的影响，因此产生了**Prompt Engineering**这一技术，通过选取更好的**Prompt**让模型取得更好的表现。谷歌的研究者们同样为每一类任务手动设计了10个 **不同的指令** 。以自然语言推断任务为例，如下图所示。



实验结果

实验设置

作者们选取了12类共计62个常见的自然语言处理和生成任务开展实验：

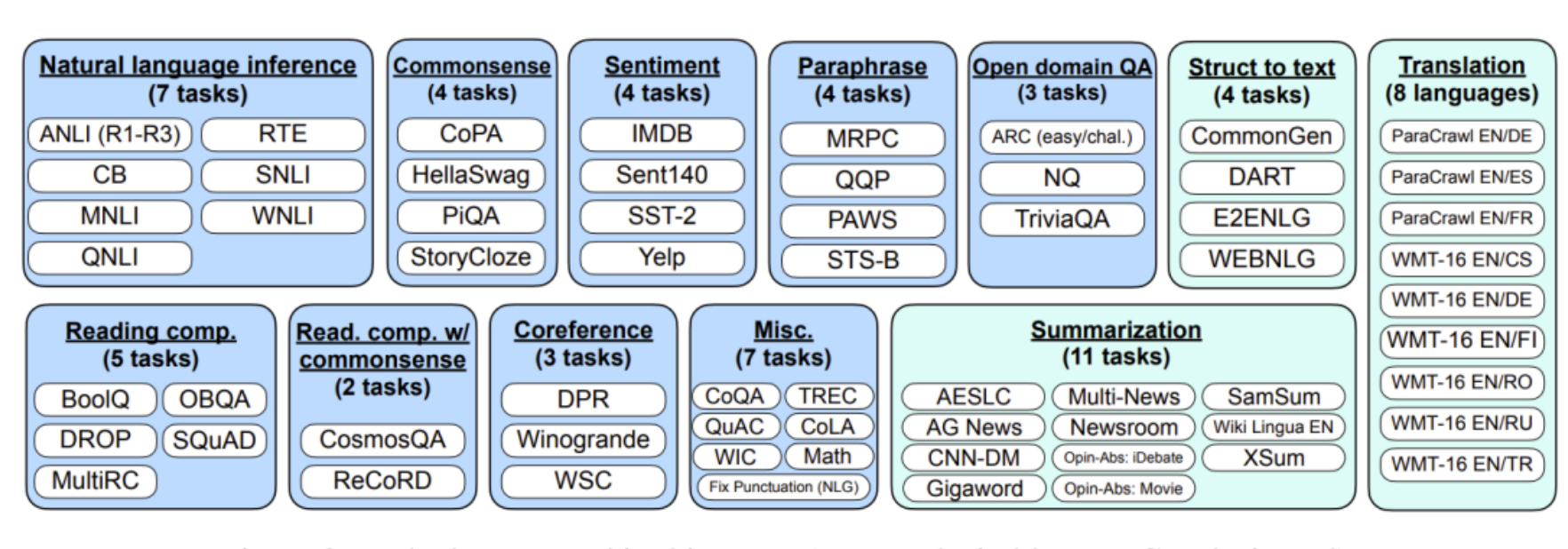


Figure 3: Task clusters used in this paper (NLU tasks in blue; NLG tasks in teal).

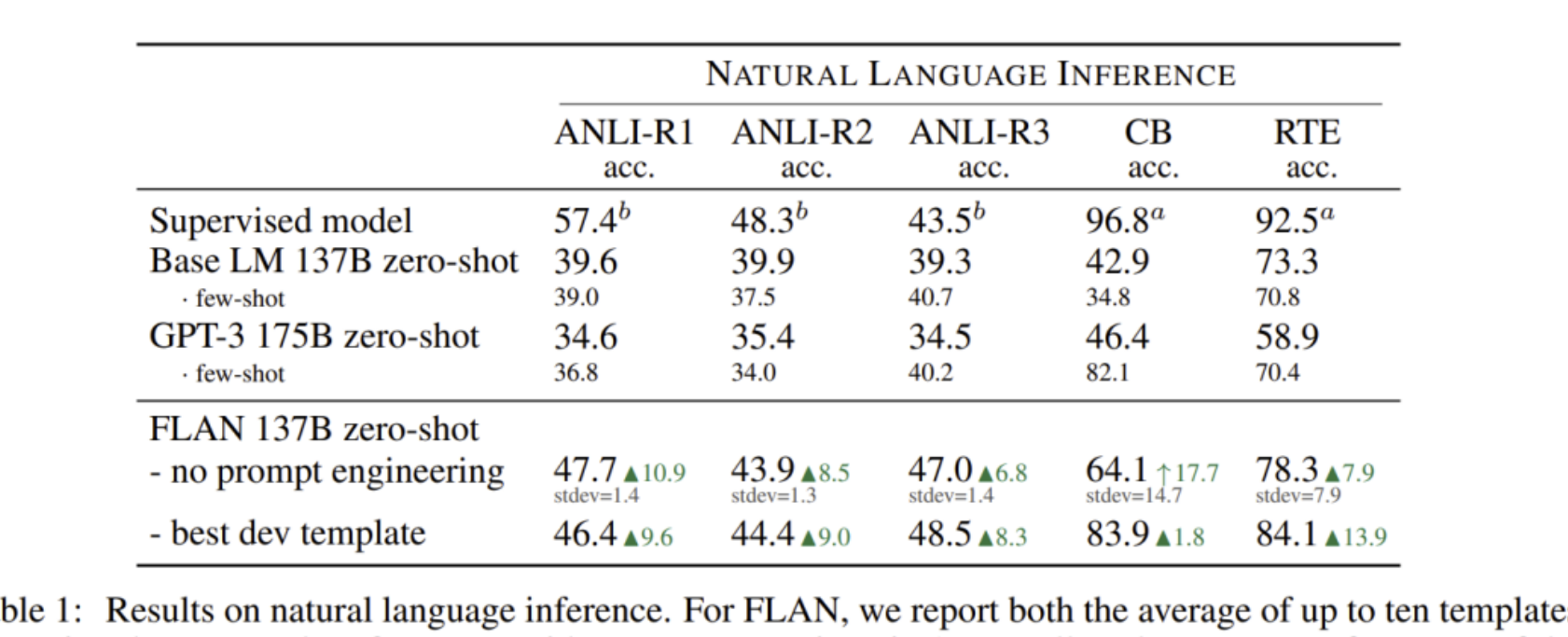
作者们训练了一个大小为**137B**的自回归语言模型（**Base LM**）作为基准语言模型。通过简单的重采样方法，作者们让不同大小的数据集在总的微调语料中占比相对平衡。在**微调**时，对每一个输入样本都随机选取该类任务十条指令中的一条以增加模型的多样性。为了保证实验对比的公平性，当评测在某一个任务（例如RTE）上的表现时，所有的同类别数据集（例如：SNLI，WNLI，MNLI...)和一些强相关类别的数据集（例如Paraphrase)将不会在微调阶段出现。

作者们使用了T5-11B和GPT-3作为基线模型。对于FLAN方法，作者同时给出了在目标任务上选择**随机指令(no prompt engineering)** 和在目标任务验证集上**最优指令(best dev template)**

结果

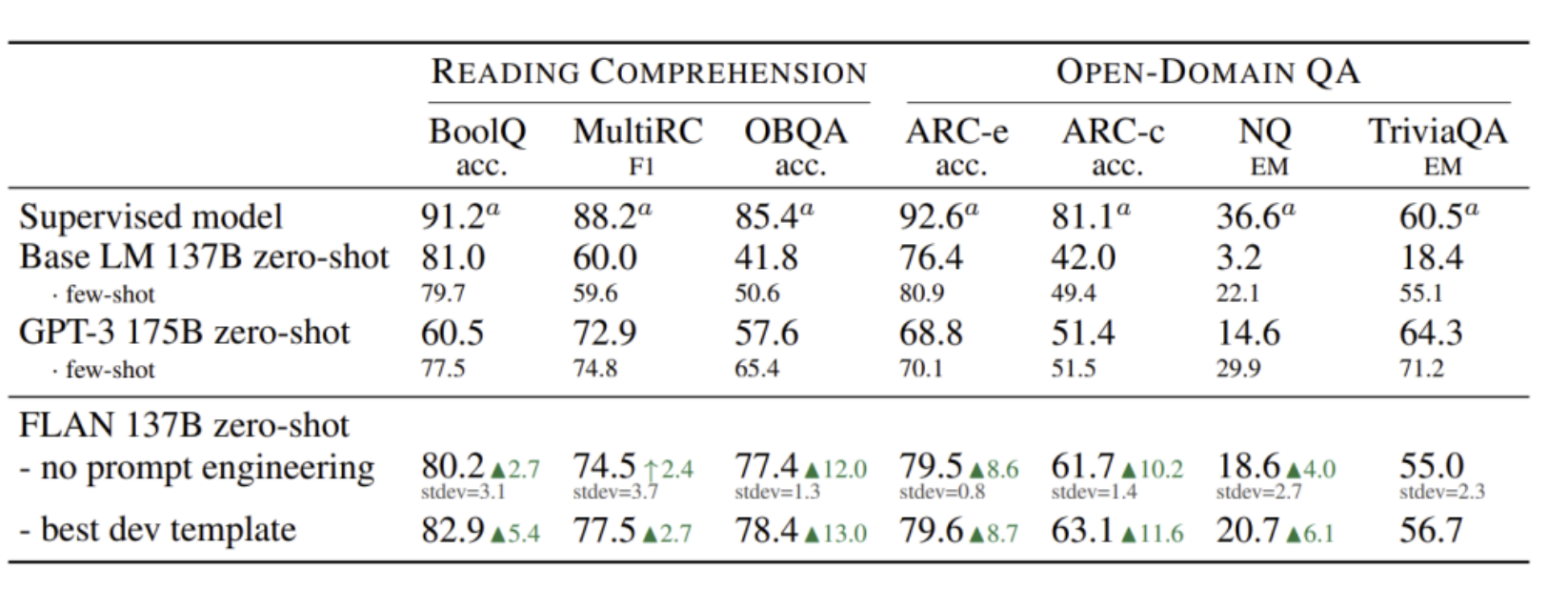
作者们发现，FLAN这一方法在与**指令**更相关的任务上表现更好（例如自然语言推断，问答），而在与常识更相关的任务上表现较为普通。

在自然语言推断任务和问答任务上，**FLAN在零样本场景下就已经超过了小样本GPT-3的效果，在许多任务上甚至与有监督模型达到了相当的表现。**



	NATURAL LANGUAGE INFERENCE				
	ANLI-R1 acc.	ANLI-R2 acc.	ANLI-R3 acc.	CB acc.	RTE acc.
Supervised model	57.4^b	48.3^b	43.5^b	96.8^a	92.5^a
Base LM 137B zero-shot	39.6	39.9	39.3	42.9	73.3
· few-shot	39.0	37.5	40.7	34.8	70.8
GPT-3 175B zero-shot	34.6	35.4	34.5	46.4	58.9
· few-shot	36.8	34.0	40.2	82.1	70.4
FLAN 137B zero-shot					
- no prompt engineering	47.7 ▲10.9 stdev=1.4	43.9 ▲8.5 stdev=1.3	47.0 ▲6.8 stdev=1.4	64.1 ↑17.7 stdev=14.7	78.3 ▲7.9 stdev=7.9
- best dev template	46.4 ▲9.6	44.4 ▲9.0	48.5 ▲8.3	83.9 ▲1.8	84.1 ▲13.9

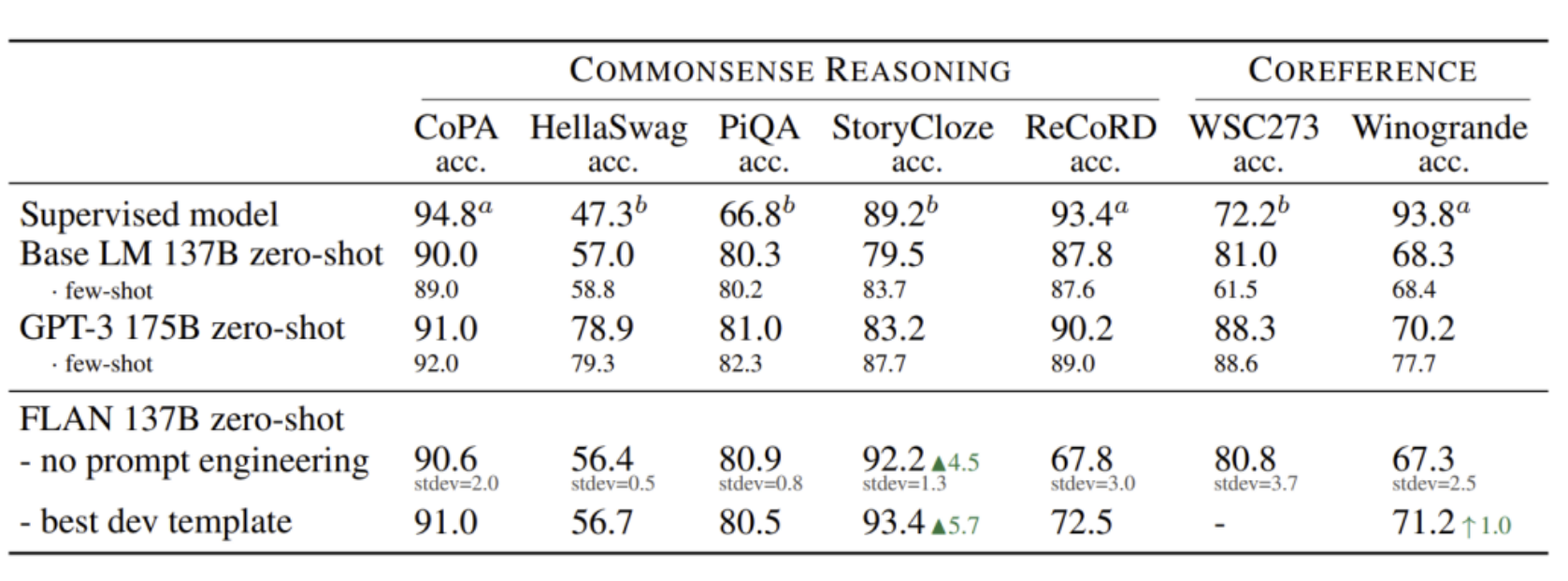
Table 1: Results on natural language inference. For FLAN, we report both the average of up to ten templates (proxying the expected performance without prompt engineering), as well as the test set performance of the template that had the highest performance on the dev set. The triangle ▲ indicates improvement over few-shot GPT-3. The up-arrow ↑ indicates improvement only over zero-shot GPT-3. ^{a}T5-11B, ^{b}BERT-large.



	READING COMPREHENSION			OPEN-DOMAIN QA			
	BoolQ acc.	MultiRC F1	OBQA acc.	ARC-e acc.	ARC-c acc.	NQ EM	TriviaQA EM
Supervised model	91.2^a	88.2^a	85.4^a	92.6^a	81.1^a	36.6^a	60.5^a
Base LM 137B zero-shot	81.0	60.0	41.8	76.4	42.0	3.2	18.4
· few-shot	79.7	59.6	50.6	80.9	49.4	22.1	55.1
GPT-3 175B zero-shot	60.5	72.9	57.6	68.8	51.4	14.6	64.3
· few-shot	77.5	74.8	65.4	70.1	51.5	29.9	71.2
FLAN 137B zero-shot							
- no prompt engineering	80.2 ▲2.7 stdev=3.1	74.5 ↑2.4 stdev=3.7	77.4 ▲12.0 stdev=1.3	79.5 ▲8.6 stdev=0.8	61.7 ▲10.2 stdev=1.4	18.6 ▲4.0 stdev=2.7	55.0 stdev=2.3
- best dev template	82.9 ▲5.4	77.5 ▲2.7	78.4 ▲13.0	79.6 ▲8.7	63.1 ▲11.6	20.7 ▲6.1	56.7

Table 2: Results on reading comprehension and open-domain question answering. For FLAN, we report both the average of up to ten templates (proxying the expected performance without prompt engineering), as well as the test set performance of the template that had the highest performance on the dev set. The triangle ▲ indicates improvement over few-shot GPT-3. The up-arrow ↑ indicates improvement only over zero-shot GPT-3. ^{a}T5-11B.

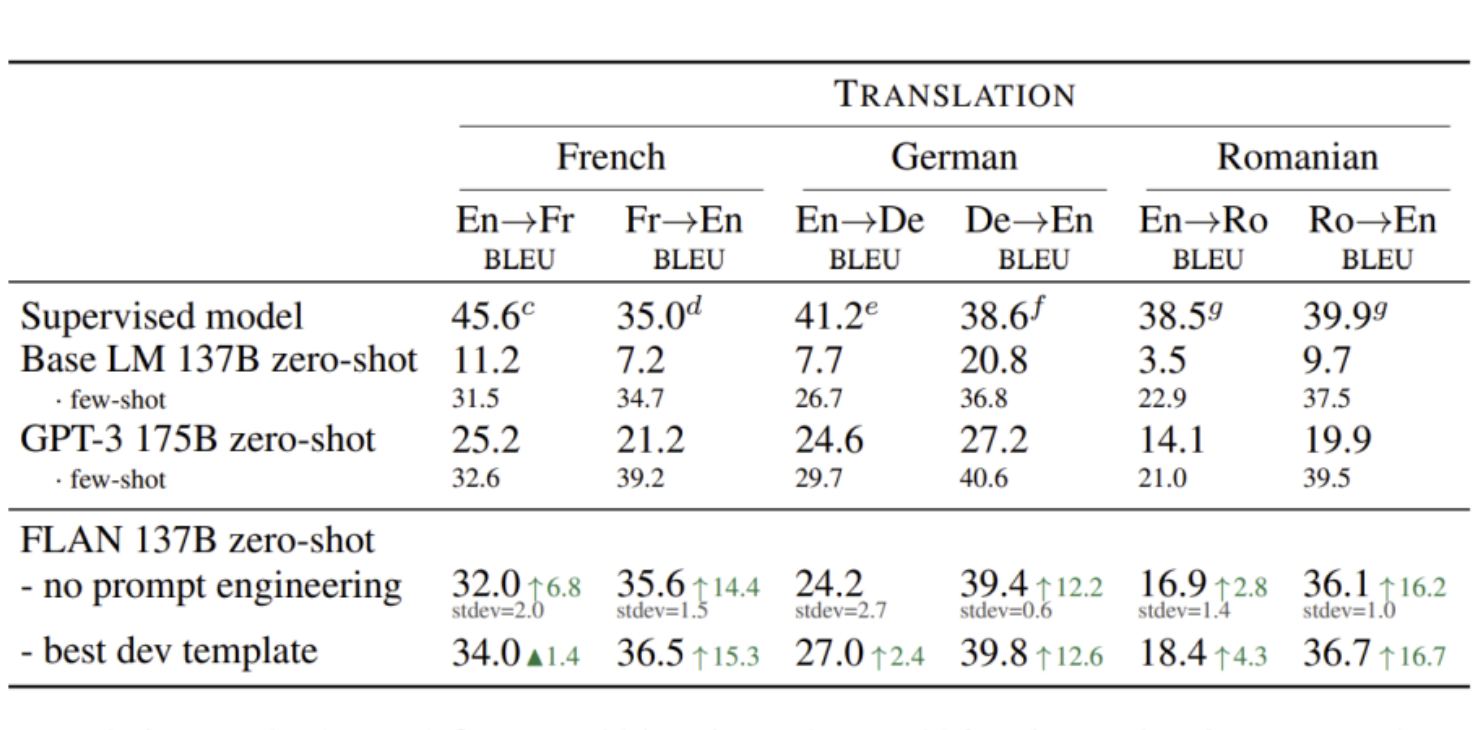
在大多数常识推理任务上，FLAN相比于GPT-3的表现基本没有提升。作者们认为主要原因是**常识推理任务对指令的依赖较小**。大多数常识推理任务可以直接转化为语言模型的形式，因此仅需要基于简单的Prompt就已经可以达到很好的表现。在作者们给出的原因之外，笔者认为另一个可能的原因是，其他任务很难提供常识推理所需要的通用世界知识，进行微调反而可能影响原有语言模型中涉及到常识的参数，损害泛化性能。



	COMMONSENSE REASONING					COREFERENCE	
	CoPA acc.	HellaSwag acc.	PiQA acc.	StoryCloze acc.	ReCoRD acc.	WSC273 acc.	Winogrande acc.
Supervised model	94.8^a	47.3^b	66.8^b	89.2^b	93.4^a	72.2^b	93.8^a
Base LM 137B zero-shot	90.0	57.0	80.3	79.5	87.8	81.0	68.3
· few-shot	89.0	58.8	80.2	83.7	87.6	61.5	68.4
GPT-3 175B zero-shot	91.0	78.9	81.0	83.2	90.2	88.3	70.2
· few-shot	92.0	79.3	82.3	87.7	89.0	88.6	77.7
FLAN 137B zero-shot							
- no prompt engineering	90.6 stdev=2.0	56.4 stdev=0.5	80.9 stdev=0.8	92.2 ▲4.5 stdev=1.3	67.8 stdev=3.0	80.8 stdev=3.7	67.3 stdev=2.5
- best dev template	91.0	56.7	80.5	93.4 ▲5.7	72.5	-	71.2 ↑1.0

Table 3: Results (accuracy in %) for commonsense reasoning and coreference resolution. For FLAN, we report both the average of up to ten templates (proxying the expected performance without prompt engineering), as well as the test set performance of the template that had the highest performance on the dev set. ^{a}T5-11B, ^{b}BERT-large. The triangle ▲ indicates improvement over few-shot GPT-3. The up-arrow ↑ indicates improvement only over zero-shot GPT-3.

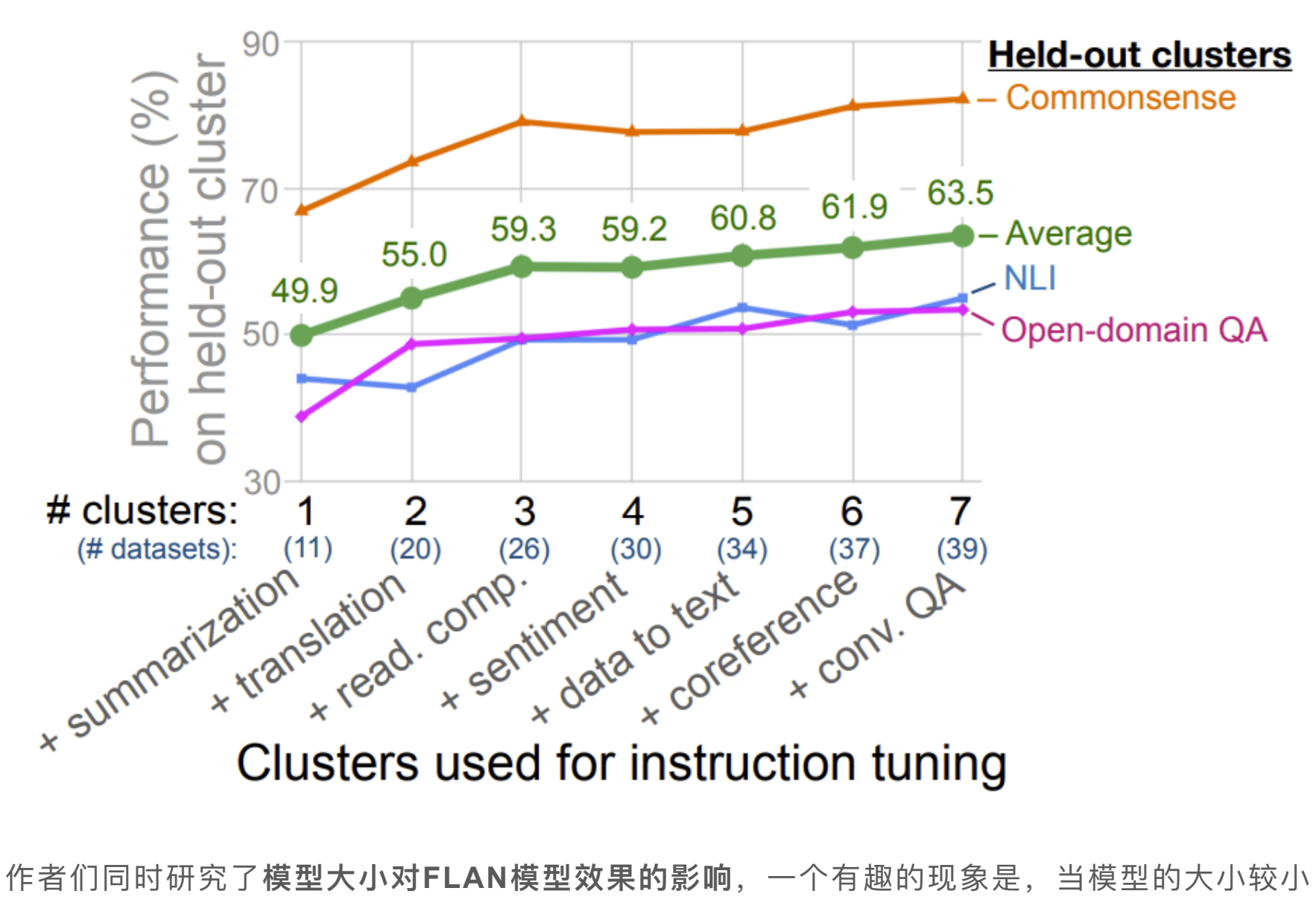
在翻译任务上，零样本场景下的FLAN明显优于GPT-3，但相比于小样本GPT-3的表现仍然有差距。



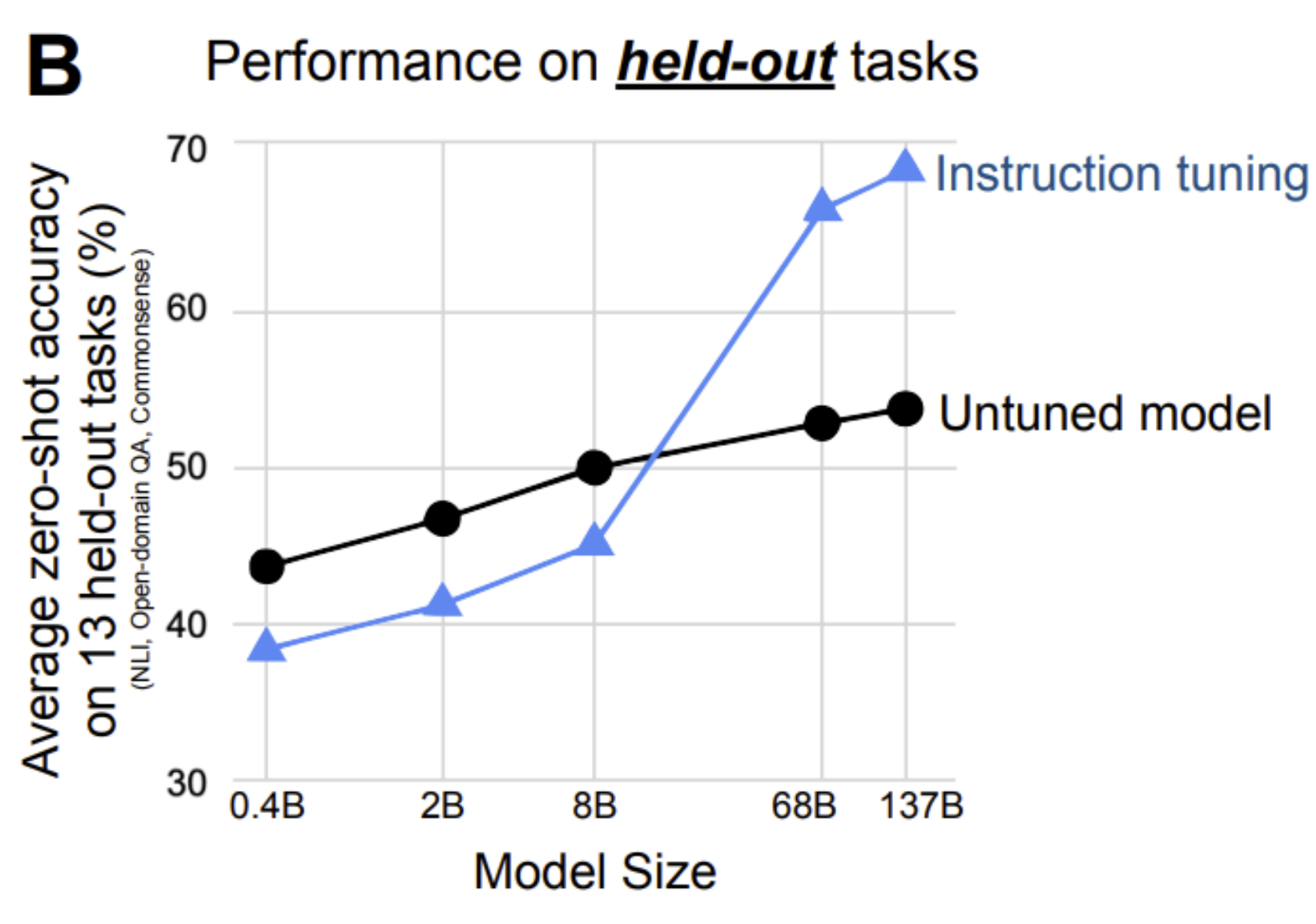
	TRANSLATION					
	French		German		Romanian	
	En→Fr BLEU	Fr→En BLEU	En→De BLEU	De→En BLEU	En→Ro BLEU	Ro→En BLEU
Supervised model	45.6^c	35.0^d	41.2^e	38.6^f	38.5^g	39.9^g
Base LM 137B zero-shot	11.2	7.2	7.7	20.8	3.5	9.7
· few-shot	31.5	34.7	26.7	36.8	22.9	37.5
GPT-3 175B zero-shot	25.2	21.2	24.6	27.2	14.1	19.9
· few-shot	32.6	39.2	29.7	40.6	21.0	39.5
FLAN 137B zero-shot						
- no prompt engineering	32.0 ↑6.8 stdev=2.0	35.6 ↑14.4 stdev=1.5	24.2 stdev=2.7	39.4 ↑12.2 stdev=0.6	16.9 ↑2.8 stdev=1.4	36.1 ↑16.2 stdev=1.0
- best dev template	34.0 ▲1.4	36.5 ↑15.3	27.0 ↑2.4	39.8 ↑12.6	18.4 ↑4.3	36.7 ↑16.7

Table 4: Translation results (BLEU) for WMT'14 En/Fr and WMT'16 En/De and En/Ro. For FLAN, we report both the average of up to ten templates (proxying the expected performance without prompt engineering), as well as the test set performance of the template that had the highest performance on the dev set. ^{c}Edunov et al. (2018), ^{d}Durrani et al. (2014), ^{e}Wang et al. (2019b), ^{f}Sennrich et al. (2016), ^{g}Liu et al. (2020). The triangle ▲ indicates improvement over few-shot GPT-3. The up-arrow ↑ indicates improvement only over zero-shot GPT-3.

作者们还研究了**增加指令微调阶段任务的数目**对FLAN模型效果的影响。结果表明，随着指令微调任务数目的增加，模型在各种任务上都能够取得更好的表现。



作者们同时研究了**模型大小对FLAN模型效果的影响**，一个有趣的现象是，当模型的大小较小时，**指令微调**反而会让模型的表现变差。作者认为的原因时，当模型较小时，在大量的任务上做**指令微调**会“填满”模型的容量，损害模型的泛化能力，使得模型在新任务上表现较差。



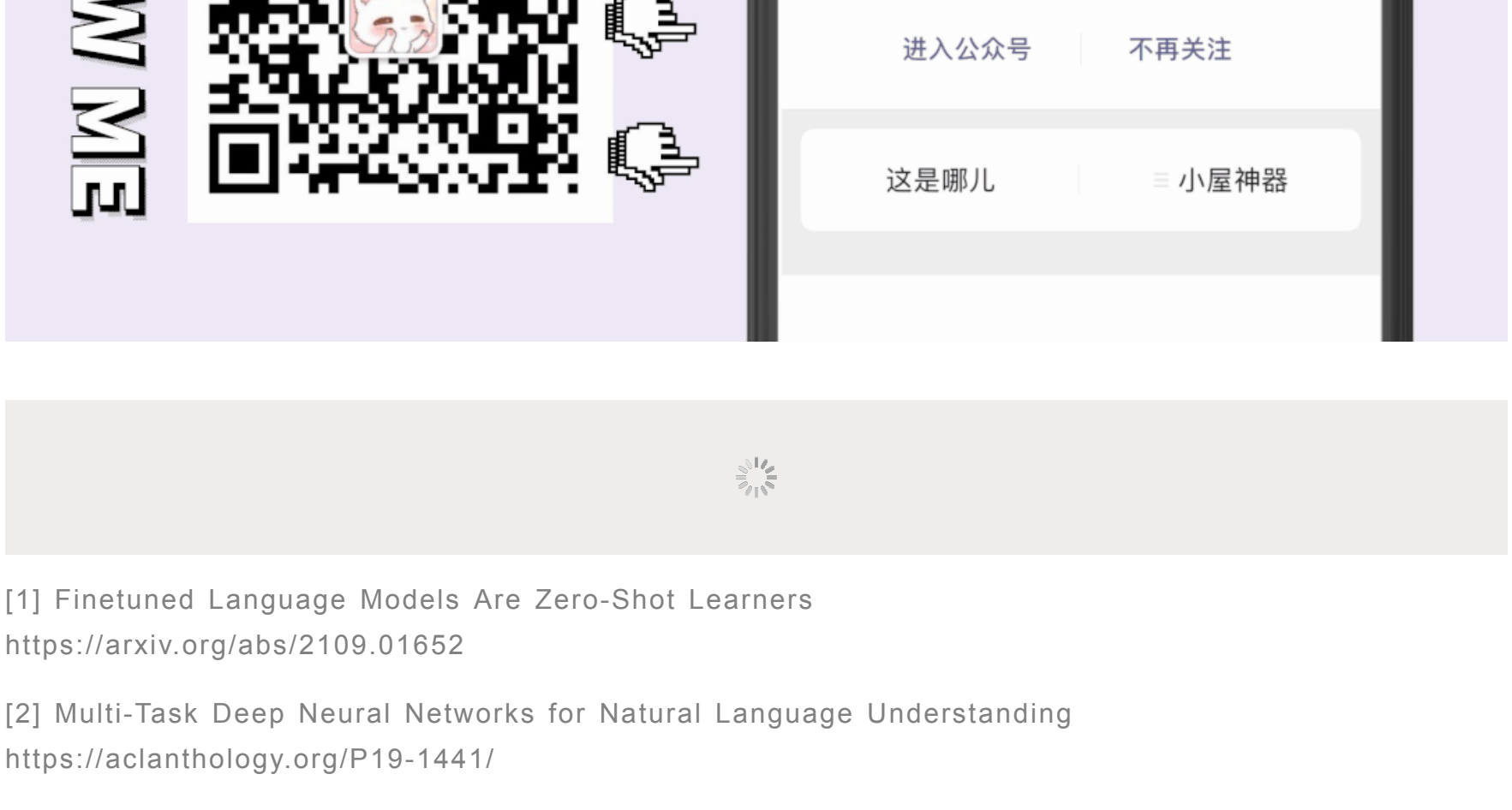
结语

熟悉NLP相关领域的同行们，也许会认为这篇文章又是一篇“A+B”的工作（A= Prompt-tuning, B= Multi-task Learning).基于Prompt的工作正值大热的时期（你敢相信九月的第一周就有四篇和Prompt有关的NLP论文挂Arxiv吗~），而通过在不同种类的微调任务上多任务学习提升性能也并不新颖，例如早期Microsoft的工作MT-DNN，Facebook的工作MUPPET。

不过，笔者认为，这样的A+B，或许是未来**通用自然语言处理模型**的一个可能的解决方案。

首先通过大量的无标记语料训练千亿参数级别的大规模自回归预训练模型，第二步，通过设计指令（Instruction Tuning）的方式让这样的模型能够对理解和生成任务进行微调。在微调的过程中可以采用类似于课程学习的方式，先学习底层的任务（如命名实体识别，序列语义标注），再学习上层的任务（如逻辑推理，问答）；先学习资源丰富的任务（如英语/大数据任务），再学习资源较少的任务（如小语种、少数据任务），并利用适配器(Adapter)保留模型中任务专用的部分。最后，给出指令让模型面对新数据、新任务进行推理。

这样通用性更强的工作应该不会太远，也许资源丰富的大厂们已经在搞了呢~



[1] Finetuned Language Models Are Zero-Shot Learners
https://arxiv.org/abs/2109.01652

[2] Multi-Task Deep Neural Networks for Natural Language Understanding
https://aclanthology.org/P19-1441/

[3] Muppet: Massive Multi-task Representations with Pre-Finetuning
https://arxiv.org/abs/2101.11038