Prompt & Question Answering

2022.10.19

Fine-tuning:

Paradigm: Update the entire set of model parameters for a target task.

Shortage: Memory consuming and PTM are usually large.

Prompting:

LMs also contain factual and commonsense knowledge that can be elicited with a prompt.

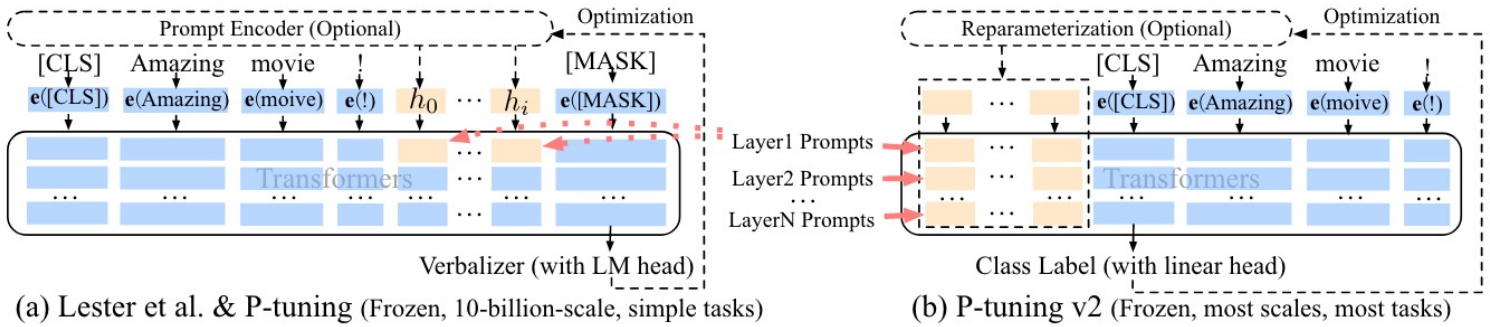
Paradigm: freezing all parameters of a PTM and uses a natural language prompt to query it.

Method: (hard) prompt design, soft-prompt, prefix-tuning (deep prompt tuning).

Example:

Mozart$_x$ was born in _____$_y$

_____$_x$ v1 v2 v3 v4 v5 _____$_y$

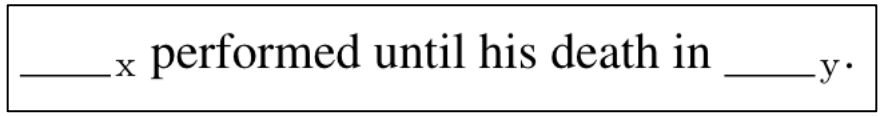


Learning How to Ask: Querying LMs with Mixtures of Soft Prompts

NAACL 2021

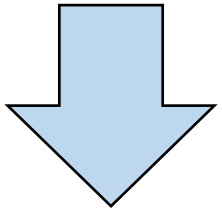
Natural-language prompts using few-shot and fill in the blank to perform NLP tasks

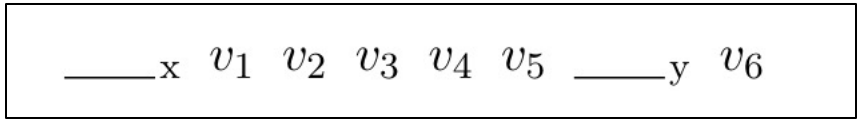
Task: extracting relational knowledge from LMs



misleading, ambiguous, or overly specific.

Hard Prompt : LAMA, LPAQA, AutoPrompt





learn a mixture of soft prompts

Qin G, Eisner J. Learning how to ask: Querying lms with mixtures of soft prompts[J]. arXiv preprint arXiv:2104.06599, 2021.

Soft Prompt:

In soft prompt, the tokens can be arbitrary vectors in embedding space.

perturbations

(1) Deeply Perturbed Prompts: additively perturbing each $v_i^{(\ell)}$ by a small $\Delta_i^{(\ell)}$ → 6 * d * (L + 1)

(2) Mixture Modeling：

$$p(y \mid x, r) = \sum_{\mathbf{t} \in \mathcal{T}_r} p(\mathbf{t} \mid r) \cdot p_{\mathrm{LM}}(y \mid \mathbf{t}, x)$$

(3) Data-Dependent Mixture Modeling：

$$p(\mathbf{t} \mid r, x)$$

Qin G, Eisner J. Learning how to ask: Querying lms with mixtures of soft prompts[J]. arXiv preprint arXiv:2104.06599, 2021.

Prompts Initialization:

(1) single: LAMA
(2) paraphrases: LPAQA
(3) Mining: LPAQA
(4) random

Dataset:

T-Rex original, T-Rex extended, Google-RE, ConceptNet

Model	P@1	P@10	MRR
LAMA	9.7†	27.0†	15.6†
LPAQA	10.6†	23.7†	15.3†
Soft (sin.)	**11.2** (+1.5)	**33.5** (+ 6.5)	**18.9** (+3.3)
Soft (min.)	**12.9** (+2.3)	**34.7** (+11.0)	**20.3** (+5.0)
Soft (par.)	**11.5** (+0.9)	**31.4** (+ 7.7)	**18.3** (+3.0)

Table 2: Results on Google-RE dataset obtained by querying the BERT-large-cased model.

Model	T-REx orig.	T-REx ext.
LAMA (BEb)	31.1	26.4
LPAQA(BEb)	34.1	31.2
AutoPrompt	43.3	45.6
Soft (sin., BEb)	47.7 (+**16.6**?)	49.6 (+**23.2**?)
Soft (min., BEb)	**50.7**? (+**16.6**?)	**50.5**? (+**19.3**?)
Soft (par., BEb)	48.4 (+**12.8**?)	49.7 (+**18.5**?)
Soft (ran., BEb)	48.1 (+**47.4**)	50.6 (+**49.8**)
LAMA (BEl)	28.9†	24.0†
LPAQA(BEl)	39.4†	37.8†
Soft (sin., BEl)	51.1 (+**22.2**)	51.4 (+**27.4**)
Soft (min., BEl)	**51.6** (+**12.2**)	**52.5** (+**14.7**)
Soft (par., BEl)	51.1 (+**11.7**)	51.7 (+**13.9**)
Soft (ran., BEl)	51.9 (+**47.1**)	51.9 (+**50.5**)
AutoPrompt	40.0	-
Soft (min., Rob)	**40.6**? (+**39.4**)	-

Qin G, Eisner J. Learning how to ask: Querying lms with mixtures of soft prompts[J]. arXiv preprint arXiv:2104.06599, 2021.

LM	Method	Precision@1	Precision@10	MRR
		init → soft → deep	init → soft → deep	init → soft → deep
BEb	LAMA	31.1	59.5	40.3
	LPAQA	34.1	62.0	43.6
	Soft (sin.)	31.1 $\xrightarrow{+14.6^?}$ 45.7 $\xrightarrow{+2.0}$ 47.7	59.5 $\xrightarrow{+16.3^?}$ 75.8 $\xrightarrow{+3.2}$ 79.0	40.3 $\xrightarrow{+15.9^?}$ 56.2 $\xrightarrow{+2.2}$ 58.4
	Soft (min.)	34.1 $\xrightarrow{+14.7^?}$ 48.8 $\xrightarrow{+1.9}$ **50.7**$^?$	62.0 $\xrightarrow{+15.6^?}$ 79.6 $\xrightarrow{+1.1}$ **80.7**$^?$	43.6 $\xrightarrow{+15.8^?}$ 59.4 $\xrightarrow{+1.7}$ **61.1**$^?$
	Soft (par.)	34.1 $\xrightarrow{+12.8^?}$ 46.9 $\xrightarrow{+1.5}$ 48.4	62.0 $\xrightarrow{+16.8^?}$ 78.8 $\xrightarrow{+0.8}$ 79.6	43.6 $\xrightarrow{+14.2^?}$ 57.8 $\xrightarrow{+1.3}$ 59.1
	Soft (ran.)	0.7 $\xrightarrow{+46.6}$ 47.3 $\xrightarrow{+0.8}$ 48.1	4.6 $\xrightarrow{+74.0}$ 79.1 $\xrightarrow{+0.0}$ 79.1	2.3 $\xrightarrow{+56.1}$ 58.4 $\xrightarrow{+0.5}$ 58.9
BEl	LAMA	28.9†	57.7†	38.7†
	LPAQA	39.4†	67.4†	49.1†
	Soft (sin.)	28.9 $\xrightarrow{+16.9}$ 45.8 $\xrightarrow{+5.3}$ 51.1	57.7 $\xrightarrow{+19.0}$ 76.7 $\xrightarrow{+4.4}$ 81.1	38.7 $\xrightarrow{+17.8}$ 56.5 $\xrightarrow{+5.0}$ 61.5
	Soft (min.)	39.4 $\xrightarrow{+11.6}$ 51.0 $\xrightarrow{+0.6}$ **51.6**	67.4 $\xrightarrow{+14.0}$ 81.4 $\xrightarrow{+0.5}$ **81.9**	49.1 $\xrightarrow{+12.5}$ 61.6 $\xrightarrow{+0.5}$ **62.1**
	Soft (par.)	39.4 $\xrightarrow{+9.2}$ 48.6 $\xrightarrow{+2.5}$ 51.1	67.4 $\xrightarrow{+12.6}$ 80.0 $\xrightarrow{+1.7}$ 81.7	49.1 $\xrightarrow{+10.5}$ 59.6 $\xrightarrow{+2.1}$ 61.7
	Soft (ran.)	2.3 $\xrightarrow{+47.1}$ 49.4 $\xrightarrow{+1.9}$ 51.3	8.0 $\xrightarrow{+73.0}$ 81.0 $\xrightarrow{+0.7}$ 81.7	4.5 $\xrightarrow{+55.9}$ 60.4 $\xrightarrow{+1.5}$ 61.9
Rob	LPAQA	1.2†	9.1†	4.2†
	AutoPrompt	40.0	68.3	49.9
	Soft (min.)	1.2 $\xrightarrow{+39.4}$ **40.6** $\xrightarrow{-7.3}$ 33.2	9.1 $\xrightarrow{+66.3}$ **75.4** $\xrightarrow{-22.3}$ 53.0	4.2 $\xrightarrow{+48.8}$ **53.0** $\xrightarrow{-12.1}$ 40.8
BAb	LPAQA	0.8†	5.7†	2.9†
	Soft (min.)	0.8 $\xrightarrow{+39.1}$ **39.9**	5.7 $\xrightarrow{+69.7}$ **75.4**	2.9 $\xrightarrow{+49.2}$ **52.1**
BAl	LPAQA	3.5†	5.6†	4.8†
	Soft (min.)	3.5 $\xrightarrow{+22.3}$ **25.8**	5.6 $\xrightarrow{+62.4}$ **68.0**	4.8 $\xrightarrow{+36.2}$ **41.0**

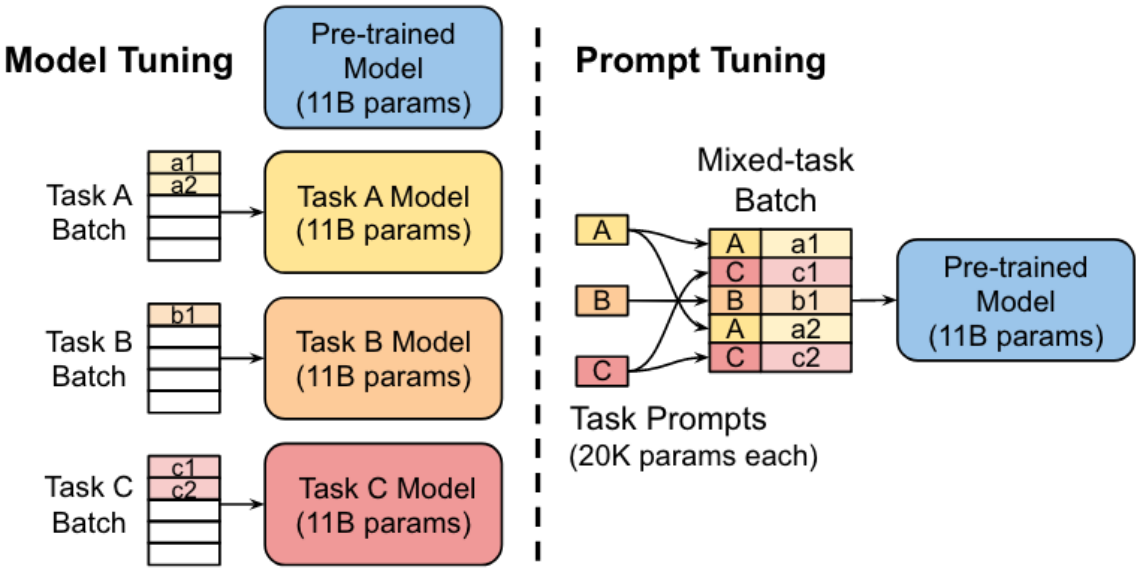
Qin G, Eisner J. Learning how to ask: Querying lms with mixtures of soft prompts[J]. arXiv preprint arXiv:2104.06599, 2021.

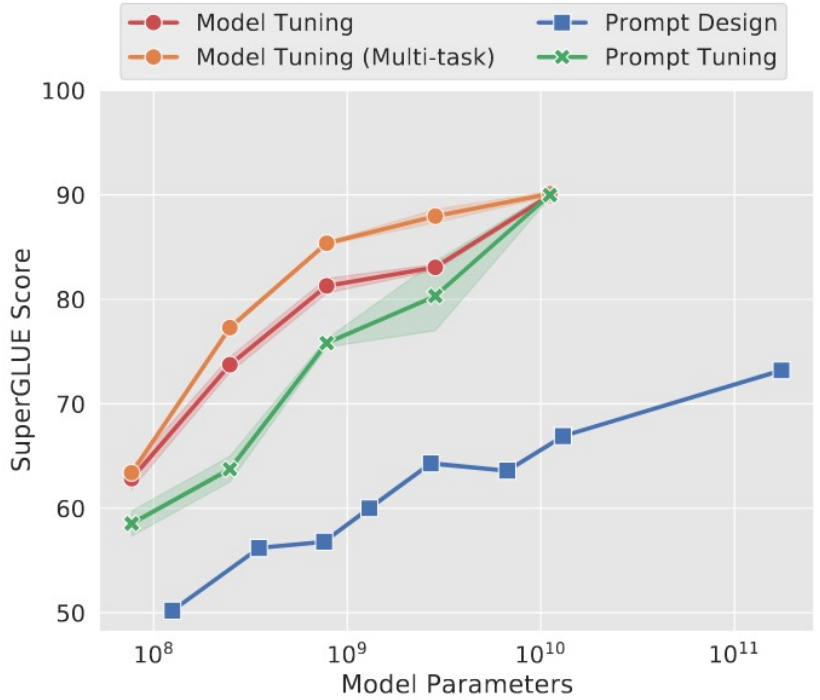
The Power of Scale for Parameter-Efficient Prompt Tuning

EMNLP 2021

GPT3 prompt design VS prompt-tuning:

SuperGLUE: GPT-3 175B fewshot performance is 17.5 points below fine-tuned T5-XXL (71.8 vs. 89.3) despite using 16 times more parameters.

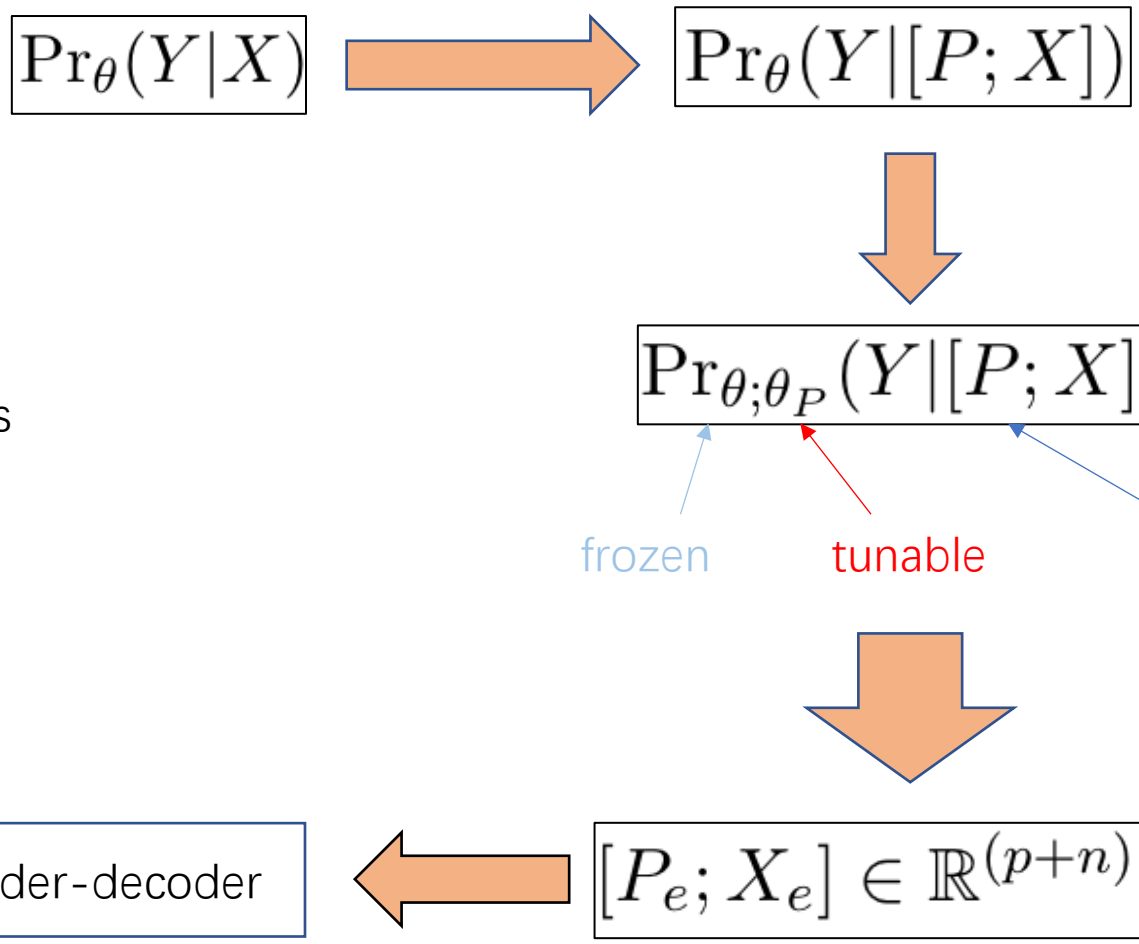




Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning[J]. arXiv preprint arXiv:2104.08691, 2021.

Prompt Tuning:

For "text-to-text" approach of T5:



Design Decisions:

(1) initialize the prompt representations

(2) the length of the prompt

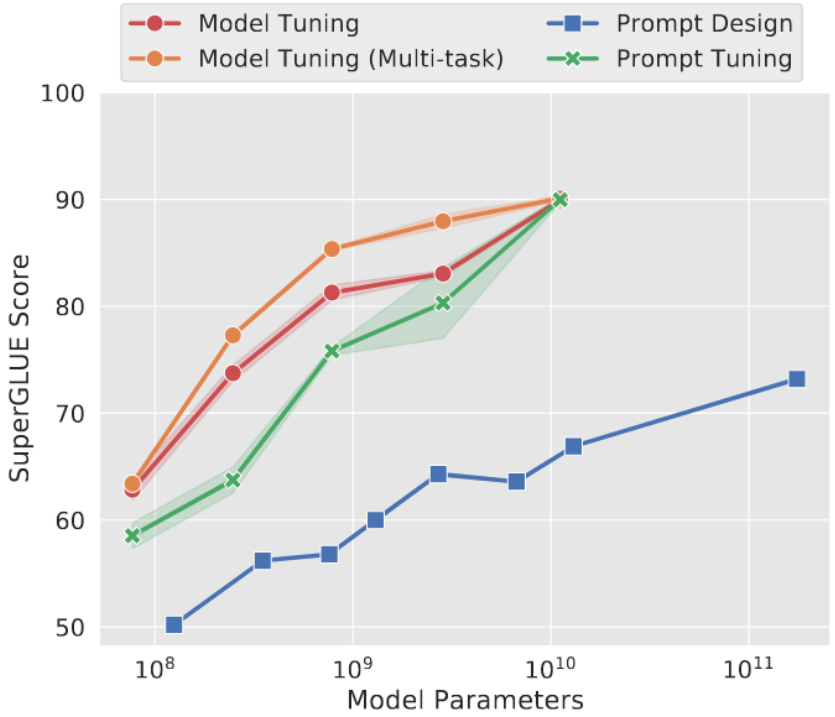
Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning[J]. arXiv preprint arXiv:2104.08691, 2021.

Unlearning Span Corruption(T5):

"LM Adaptation" （100k）: transform T5 into a model more similar to GPT-3.

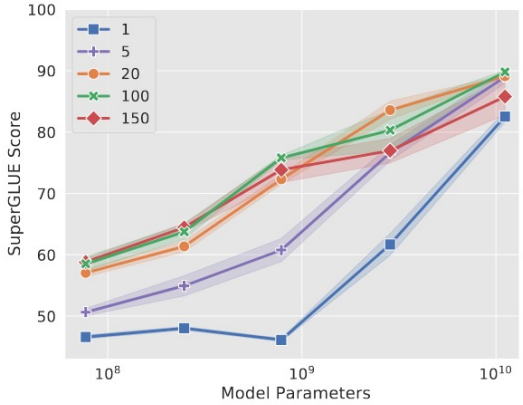
Datasets: SuperGLUE (translate into a text-to-text formate)

Results:

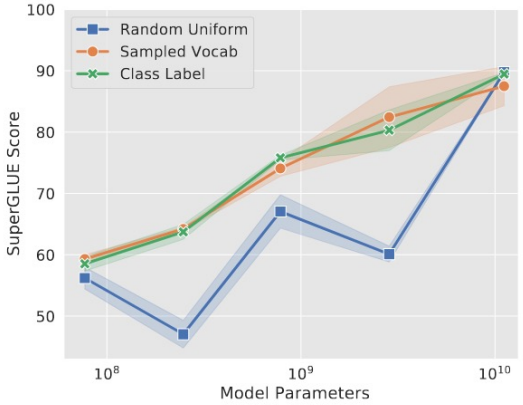


Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning[J]. arXiv preprint arXiv:2104.08691, 2021.

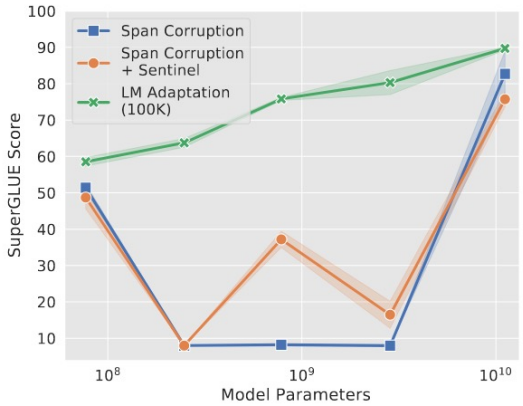
Ablation Study:



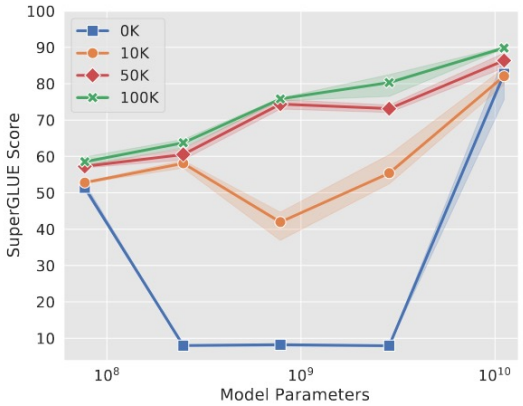
(a) Prompt length



(b) Prompt initialization



(c) Pre-training method



(d) LM adaptation steps

Comparison to Similar Approaches：soft prompt，Adapter

Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning[J]. arXiv preprint arXiv:2104.08691, 2021.

Resilience to Domain Shift:

Dataset	Domain	Model	Prompt	Δ
SQuAD	Wiki	94.9 ±0.2	94.8 ±0.1	−0.1
TextbookQA	Book	54.3 ±3.7	**66.8** ±2.9	+12.5
BioASQ	Bio	77.9 ±0.4	**79.1** ±0.3	+1.2
RACE	Exam	59.8 ±0.6	**60.7** ±0.5	+0.9
RE	Wiki	88.4 ±0.1	**88.8** ±0.2	+0.4
DuoRC	Movie	**68.9** ±0.7	67.7 ±1.1	−1.2
DROP	Wiki	**68.9** ±1.7	67.1 ±1.9	−1.8

Table 1: F1 mean and stddev for models trained on SQuAD and evaluated on out-of-domain datasets from the MRQA 2019 shared task. Prompt tuning tends to give stronger zero-shot performance than model tuning, especially on datasets with large domain shifts like TextbookQA.

Train	Eval	Tuning	Accuracy	F1
QQP	MRPC	Model	73.1 ±0.9	81.2 ±2.1
		Prompt	**76.3** ±0.1	**84.3** ±0.3
MRPC	QQP	Model	74.9 ±1.3	**70.9** ±1.2
		Prompt	**75.4** ±0.8	69.7 ±0.3

Table 2: Mean and stddev of zero-shot domain transfer between two paraphrase detection tasks.

Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning[J]. arXiv preprint arXiv:2104.08691, 2021.

Interpretability:

Compute the nearest neighbors to each prompt token from the frozen model’ s vocabulary.
The top-5 nearest neighbors form tight semantic clusters.

{ Technology / technology / Technologies / technological / technologies }

{ entirely / completely / totally / altogether / 100% }

Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning[J]. arXiv preprint arXiv:2104.08691, 2021.

Prefix-Tuning: Optimizing Continuous Prompts for Generation

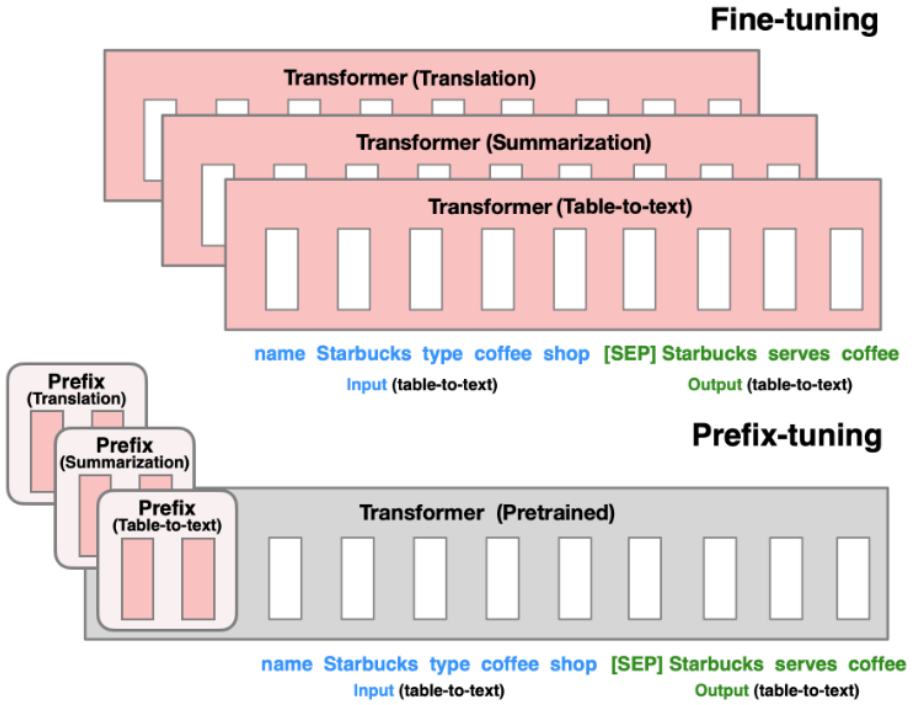
ACL 2021

In-context learning drawback：

Transformers can only condition on a bounded-length context (e.g., 2048 tokens for GPT-3), in-context learning is restricted to very small training sets.

Prefix-tuning：

a lightweight alternative to fine-tuning for natural language generation (NLG) tasks



Li X L, Liang P. Prefix-tuning: Optimizing continuous prompts for generation[J]. arXiv preprint arXiv:2101.00190, 2021.

Problem Statement：

Consider a conditional generation task where the input x is a context and the output y is a sequence of tokens. (eg, table-to-text, Summarization)

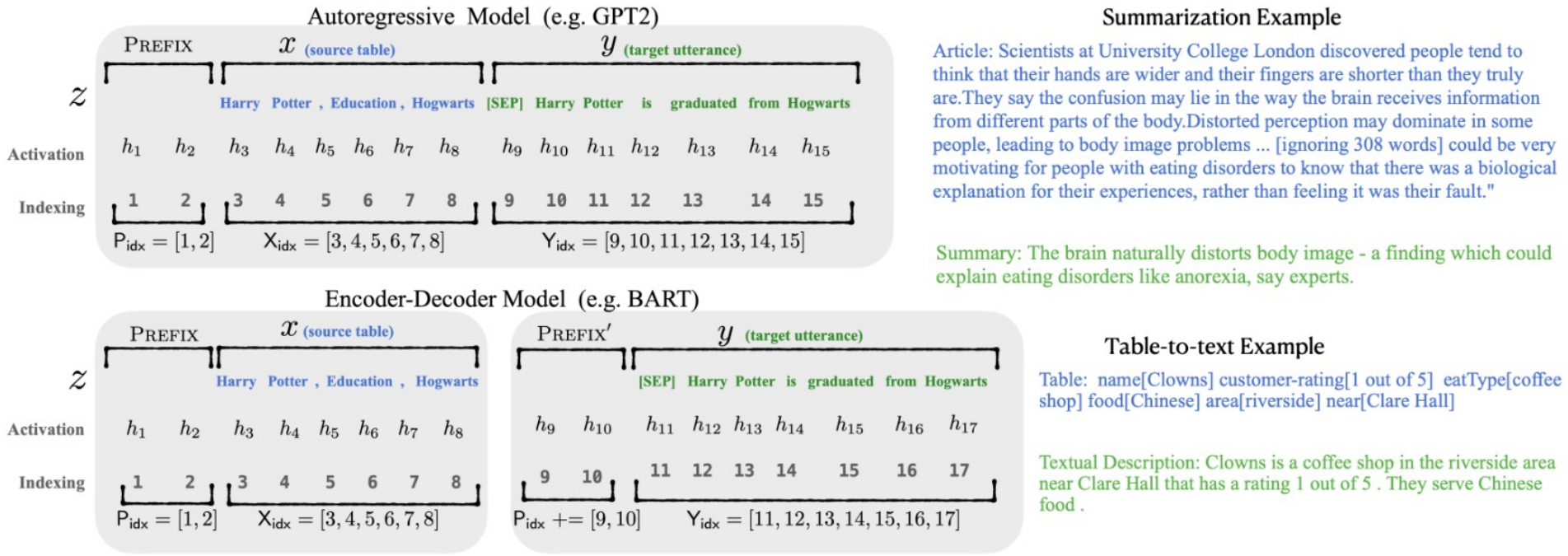


Figure 2: An annotated example of prefix-tuning using an autoregressive LM (top) and an encoder-decoder model (bottom). The prefix activations $\forall i \in \mathsf{P}_{\mathsf{idx}}, h_i$ are drawn from a trainable matrix P_θ. The remaining activations are computed by the Transformer.

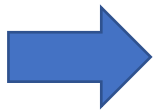
Li X L, Liang P. Prefix-tuning: Optimizing continuous prompts for generation[J]. arXiv preprint arXiv:2101.00190, 2021.

Prefix-Tuning：

Intuition: conditioning on a proper context can steer the LM without changing its parameters.

Idea: optimizing the activations of all the layers, not just the embedding layer.

$$h_i = \begin{cases} P_\theta[i,:], & \text{if } i \in \mathsf{P}_{\mathsf{idx}}, \\ \mathrm{LM}_\phi(z_i, h_{<i}), & \text{otherwise.} \end{cases} \quad (3)$$



$| P_{idx} | * \dim(h_i)$

Parametrization of P_θ：

$$P_\theta[i,:] = \mathrm{MLP}_\theta(P'_\theta[i,:])$$

Li X L, Liang P. Prefix-tuning: Optimizing continuous prompts for generation[J]. arXiv preprint arXiv:2101.00190, 2021.

Datasets：

	#examples	input length	output length
E2E	50K	28.5	27.8
WebNLG	22K	49.6	30.7
DART	82K	38.8	27.3
XSUM	225K	473.3	28.1

Main Result：

	E2E					WebNLG									DART					
	BLEU	NIST	MET	R-L	CIDEr	BLEU			MET			TER ↓			BLEU	MET	TER ↓	Mover	BERT	BLEURT
						S	U	A	S	U	A	S	U	A						
GPT-2$_{\text{MEDIUM}}$																				
FT-FULL	68.8	8.71	46.1	71.1	2.43	**64.7**	26.7	45.7	**0.46**	0.30	0.38	**0.33**	0.78	0.54	46.2	**0.39**	**0.46**	**0.50**	**0.94**	**0.39**
FT-TOP2	68.1	8.59	46.0	70.8	2.41	53.6	18.9	36.0	0.38	0.23	0.31	0.49	0.99	0.72	41.0	0.34	0.56	0.43	0.93	0.21
ADAPTER(3%)	68.9	8.71	46.1	71.3	**2.47**	60.5	**47.9**	54.8	0.43	**0.38**	**0.41**	0.35	**0.46**	**0.39**	45.2	0.38	**0.46**	**0.50**	**0.94**	**0.39**
ADAPTER(0.1%)	66.3	8.41	45.0	69.8	2.40	54.5	45.1	50.2	0.39	0.36	0.38	0.40	0.46	0.43	42.4	0.36	0.48	0.47	**0.94**	0.33
PREFIX(0.1%)	**70.3**	**8.82**	**46.3**	**72.1**	2.46	62.9	45.3	**55.0**	0.44	0.37	**0.41**	0.35	0.51	0.42	**46.4**	0.38	**0.46**	**0.50**	**0.94**	**0.39**
GPT-2$_{\text{LARGE}}$																				
FT-FULL	68.5	8.78	46.0	69.9	2.45	**65.3**	43.1	55.5	**0.46**	0.38	**0.42**	**0.33**	0.53	0.42	**47.0**	**0.39**	0.46	**0.51**	**0.94**	**0.40**
Prefix	**70.3**	**8.85**	**46.2**	**71.7**	**2.47**	63.4	**47.7**	**56.3**	0.45	**0.39**	**0.42**	0.34	**0.48**	**0.40**	46.7	**0.39**	**0.45**	**0.51**	**0.94**	**0.40**
SOTA	68.6	8.70	45.3	70.8	2.37	63.9	52.8	57.1	0.46	0.41	0.44	-	-	-	-	-	-	-	-	-

Table 2: Metrics (higher is better, except for TER) for table-to-text generation on E2E (left), WebNLG (middle) and DART (right). With only 0.1% parameters, Prefix-tuning outperforms other lightweight baselines and achieves a comparable performance with fine-tuning. The best score is boldfaced for both GPT-2$_{\text{MEDIUM}}$ and GPT-2$_{\text{LARGE}}$.

	R-1 ↑	R-2 ↑	R-L ↑
FT-FULL(Lewis et al., 2020)	45.14	22.27	37.25
PREFIX(2%)	43.80	20.93	36.05
PREFIX(0.1%)	42.92	20.03	35.05

Table 3: Performance of methods on the XSUM summarization dataset. Prefix-tuning slightly underperforms fine-tuning in the full-data regime.

Li X L, Liang P. Prefix-tuning: Optimizing continuous prompts for generation[J]. arXiv preprint arXiv:2101.00190, 2021.

Low-data Setting：

Source	name : The Eagle | type : coffee shop | food : Chinese | price : cheap | customer rating : average | area : riverside | family friendly : no | near : Burger King
Prefix (50)	The Eagle is a cheap Chinese coffee shop located near Burger King.
Prefix (100)	The Eagle is a cheap coffee shop located in the riverside near Burger King. It has average customer ratings.
Prefix (200)	The Eagle is a cheap Chinese coffee shop located in the riverside area near Burger King. It has average customer ratings.
Prefix (500)	The Eagle is a coffee shop that serves Chinese food. It is located in the riverside area near Burger King. It has an average customer rating and is not family friendly.
FT (50)	The Eagle coffee shop is located in the riverside area near Burger King.
FT (100)	The Eagle is a cheap coffee shop near Burger King in the riverside area. It has a low customer rating and is not family friendly.
FT (200)	The Eagle is a cheap Chinese coffee shop with a low customer rating. It is located near Burger King in the riverside area.
FT (500)	The Eagle is a cheap Chinese coffee shop with average customer ratings. It is located in the riverside area near Burger King.

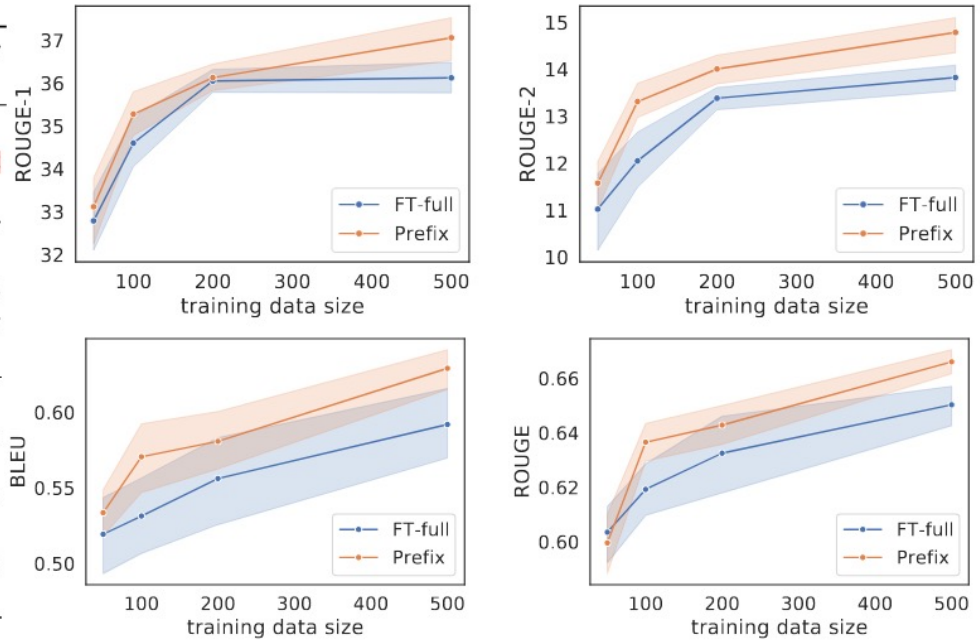


Figure 3: (Left) qualitative examples in lowdata settings. (Right) prefix-tuning (orange) outperforms fine-tuning (blue) in low-data regimes in addition to requiring many fewer parameters. The top two plots correspond to summarization, measured by ROUGE-1 and ROUGE-2. The bottom two plots correspond to table-to-text, measured by BLEU and ROUGE-L. The x-axis is the training size and the y-axis is the evaluation metric (higher is better).

Li X L, Liang P. Prefix-tuning: Optimizing continuous prompts for generation[J]. arXiv preprint arXiv:2101.00190, 2021.

Low-data Setting：

Source	name : The Eagle | type : coffee shop | food : Chinese | price : cheap | customer rating : average | area : riverside | family friendly : no | near : Burger King
Prefix (50)	The Eagle is a cheap Chinese coffee shop located near Burger King.
Prefix (100)	The Eagle is a cheap coffee shop located in the riverside near Burger King. It has average customer ratings.
Prefix (200)	The Eagle is a cheap Chinese coffee shop located in the riverside area near Burger King. It has average customer ratings.
Prefix (500)	The Eagle is a coffee shop that serves Chinese food. It is located in the riverside area near Burger King. It has an average customer rating and is not family friendly.
FT (50)	The Eagle coffee shop is located in the riverside area near Burger King.
FT (100)	The Eagle is a cheap coffee shop near Burger King in the riverside area. It has a low customer rating and is not family friendly.
FT (200)	The Eagle is a cheap Chinese coffee shop with a low customer rating. It is located near Burger King in the riverside area.
FT (500)	The Eagle is a cheap Chinese coffee shop with average customer ratings. It is located in the riverside area near Burger King.

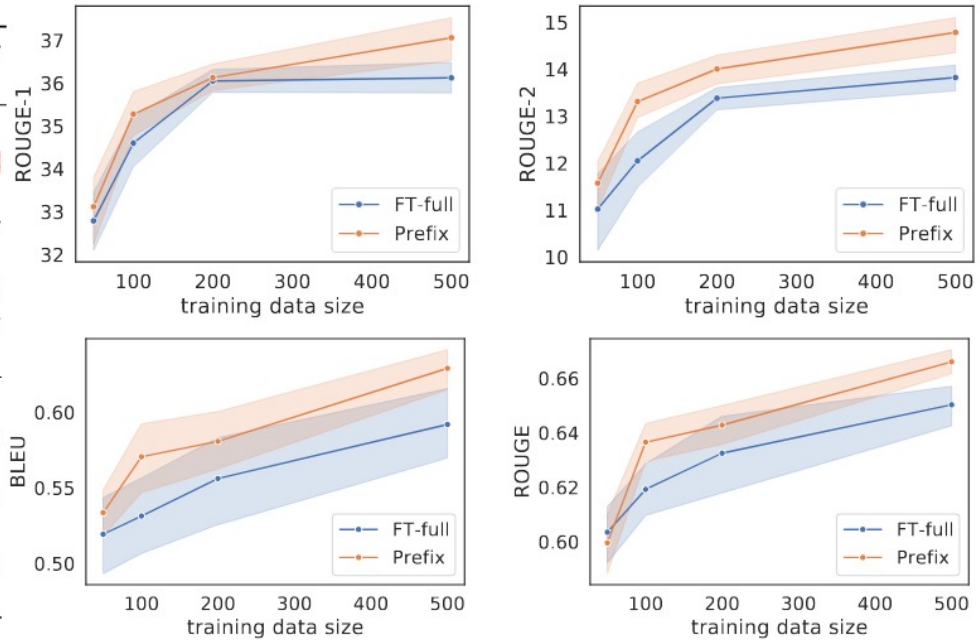


Figure 3: (Left) qualitative examples in lowdata settings. (Right) prefix-tuning (orange) outperforms fine-tuning (blue) in low-data regimes in addition to requiring many fewer parameters. The top two plots correspond to summarization, measured by ROUGE-1 and ROUGE-2. The bottom two plots correspond to table-to-text, measured by BLEU and ROUGE-L. The x-axis is the training size and the y-axis is the evaluation metric (higher is better).

Li X L, Liang P. Prefix-tuning: Optimizing continuous prompts for generation[J]. arXiv preprint arXiv:2101.00190, 2021.

Impact of Prefix Length / Embedding only / Infix / Initialization :

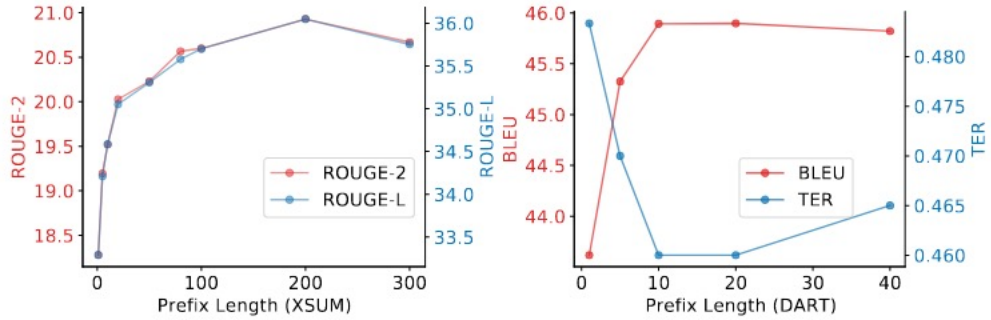


Figure 4: Prefix length vs. performance on summerization (left) and table-to-text (right). Performance in-

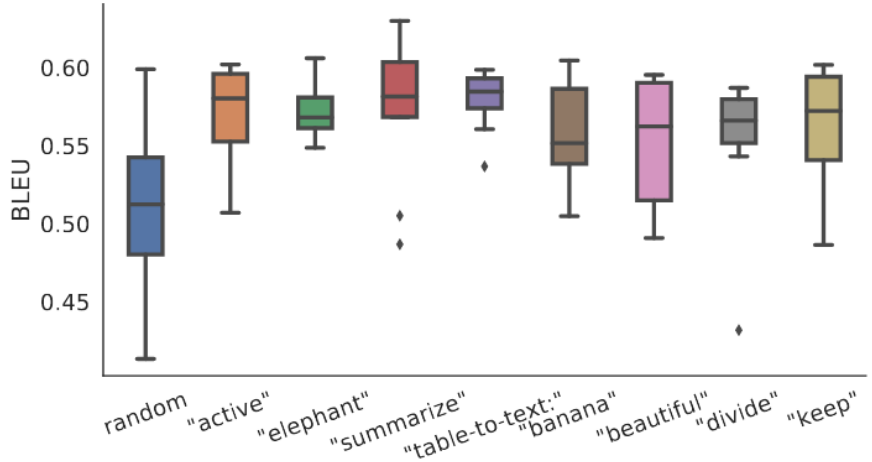


Figure 5: Initializing the prefix with activations of real words significantly outperforms random initialization, in low-data settings.

	E2E				
	BLEU	NIST	MET	ROUGE	CIDEr
PREFIX	70.3	8.82	46.3	72.1	2.46
Embedding-only: EMB-{PrefixLength}					
EMB-1	48.1	3.33	32.1	60.2	1.10
EMB-10	62.2	6.70	38.6	66.4	1.75
EMB-20	61.9	7.11	39.3	65.6	1.85
Infix-tuning: INFIX-{PrefixLength}					
INFIX-1	67.9	8.63	45.8	69.4	2.42
INFIX-10	67.2	8.48	45.8	69.9	2.40
INFIX-20	66.7	8.47	45.8	70.0	2.42

Table 5: Intrinsic evaluation of Embedding-only (§7.2) and Infixing (§7.3). Both Embedding-only ablation and Infix-tuning underperforms full prefix-tuning.

Li X L, Liang P. Prefix-tuning: Optimizing continuous prompts for generation[J]. arXiv preprint arXiv:2101.00190, 2021.

P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks

ACL 2022

Empirical finding :

Properly optimized prompt tuning can be comparable to fine-tuning universally across various model scales and NLU tasks.

Can be viewed as an optimized and adapted implementation of Deep Prompt Tuning(Soft prompt & Prefix tuning)

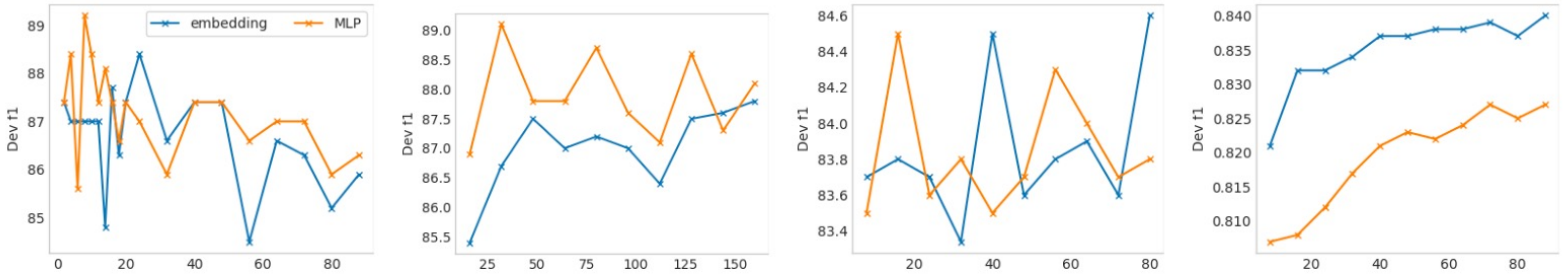
Figure 1: Average scores on RTE, BoolQ and CB of SuperGLUE dev. With 0.1% task-specific parameters, P-tuning v2 can match fine-tuning across wide scales of pre-trained models, while Lester et al. (2021) & P-tuning can make it conditionally at 10B scale.

Liu, Xiao, et al. "P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2022.

Optimization and Implementation :

(1) Reparameterization

(2) Prompt Length



(a) NLI: RTE (b) NER: CoNLL04 (c) MQA: BoolQ (d) SRL: CoNLL12

(3) Multi-task Learning

(4) Classification Head

Method	Task	Re-param.	Deep PT	Multi-task	No verb.
P-tuning (Liu et al., 2021)	KP NLU	LSTM	-	-	-
PROMPTTUNING (Lester et al., 2021)	NLU	-	-	✓	-
Prefix Tuning (Li and Liang, 2021)	NLG	MLP	✓	-	-
SOFT PROMPTS (Qin and Eisner, 2021)	KP	-	✓	-	-
P-tuning v2 (Ours)	NLU SeqTag	*(depends)*	✓	✓	✓

Table 1: Conceptual comparison between P-tuning v2 and existing Prompt Tuning approaches (KP: Knowledge Probe; SeqTag: Sequence Tagging; Re-param.: Reparameterization; No verb.: No verbalizer).

Liu, Xiao, et al. "P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2022.

Result：

	#Size	BoolQ			CB			COPA			MultiRC (F1a)		
		FT	PT	PT-2	FT	PT	PT-2	FT	PT	PT-2	FT	PT	PT-2
$\text{BERT}_{\text{large}}$	335M	**77.7**	67.2	75.8	**94.6**	80.4	**94.6**	69.0	55.0	**73.0**	70.5	59.6	**70.6**
$\text{RoBERTa}_{\text{large}}$	355M	**86.9**	62.3	84.8	98.2	71.4	**100**	**94.0**	63.0	93.0	**85.7**	59.9	82.5
$\text{GLM}_{\text{xlarge}}$	2B	**88.3**	79.7	87.0	**96.4**	76.4	**96.4**	**93.0**	92.0	91.0	84.1	77.5	**84.4**
$\text{GLM}_{\text{xxlarge}}$	10B	88.7	**88.8**	**88.8**	**98.7**	98.2	96.4	**98.0**	**98.0**	**98.0**	**88.1**	86.1	**88.1**

	#Size	ReCoRD (F1)			RTE			WiC			WSC		
		FT	PT	PT-2	FT	PT	PT-2	FT	PT	PT-2	FT	PT	PT-2
$\text{BERT}_{\text{large}}$	335M	70.6	44.2	**72.8**	70.4	53.5	**78.3**	74.9	63.0	**75.1**	**68.3**	64.4	**68.3**
$\text{RoBERTa}_{\text{large}}$	355M	89.0	46.3	**89.3**	86.6	58.8	**89.5**	**75.6**	56.9	73.4	63.5	**64.4**	63.5
$\text{GLM}_{\text{xlarge}}$	2B	91.8	82.7	**91.9**	**90.3**	85.6	**90.3**	**74.1**	71.0	72.0	**95.2**	87.5	92.3
$\text{GLM}_{\text{xxlarge}}$	10B	**94.4**	87.8	92.5	**93.1**	89.9	**93.1**	**75.7**	71.8	74.0	**95.2**	94.2	93.3

Table 2: Results on SuperGLUE development set. P-tuning v2 significantly surpasses P-tuning & Lester et al. (2021) on models smaller than 10B, and matches the performance of fine-tuning across different model scales. (FT: fine-tuning; PT: Lester et al. (2021) & P-tuning; PT-2: P-tuning v2; **bold**: the best; underline: the second best).

Liu, Xiao, et al. "P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2022.

Result：

	#Size	CoNLL03				OntoNotes 5.0				CoNLL04			
		FT	PT	PT-2	MPT-2	FT	PT	PT-2	MPT-2	FT	PT	PT-2	MPT-2
$\text{BERT}_{\text{large}}$	335M	**92.8**	81.9	90.2	91.0	**89.2**	74.6	86.4	86.3	85.6	73.6	84.5	**86.6**
$\text{RoBERTa}_{\text{large}}$	355M	92.6	86.1	**92.8**	**92.8**	**89.8**	80.8	**89.8**	**89.8**	88.8	76.2	88.4	**90.6**
$\text{DeBERTa}_{\text{xlarge}}$	750M	**93.1**	90.2	**93.1**	**93.1**	90.4	85.1	90.4	**90.5**	89.1	82.4	86.5	**90.1**

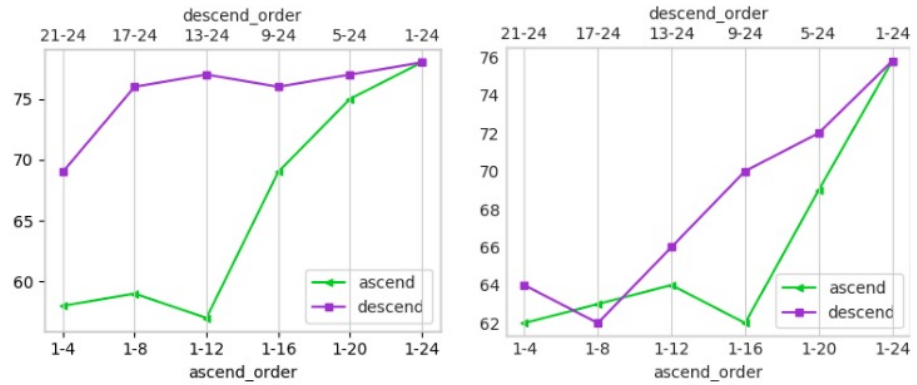
	#Size	SQuAD 1.1 dev (EM / F1)								SQuAD 2.0 dev (EM / F1)							
		FT		PT		PT-2		MPT-2		FT		PT		PT-2		MPT-2	
$\text{BERT}_{\text{large}}$	335M	**84.2**	**91.1**	1.0	8.5	77.8	86.0	82.3	89.6	**78.7**	**81.9**	50.2	50.2	69.7	73.5	72.7	75.9
$\text{RoBERTa}_{\text{large}}$	355M	**88.9**	**94.6**	1.2	12.0	88.5	94.4	88.0	94.1	**86.5**	**89.4**	50.2	50.2	82.1	85.5	83.4	86.7
$\text{DeBERTa}_{\text{xlarge}}$	750M	90.1	95.5	2.4	19.0	**90.4**	**95.7**	89.6	95.4	88.3	91.1	50.2	50.2	**88.4**	**91.1**	88.1	90.8

	#Size	CoNLL12				CoNLL05 WSJ				CoNLL05 Brown			
		FT	PT	PT-2	MPT-2	FT	PT	PT-2	MPT-2	FT	PT	PT-2	MPT-2
$\text{BERT}_{\text{large}}$	335M	84.9	64.5.	83.2	**85.1**	**88.5**	76.0	86.3	**88.5**	82.7	70.0	80.7	**83.1**
$\text{RoBERTa}_{\text{large}}$	355M	**86.5**	67.2	84.6	86.2	**90.2**	76.8	89.2	90.0	85.6	70.7	84.3	**85.7**
$\text{DeBERTa}_{\text{xlarge}}$	750M	86.5	74.1	85.7	**87.1**	**91.2**	82.3	90.6	**91.2**	86.9	77.7	86.3	**87.0**

Table 3: Results on Named Entity Recognition (NER), Question Answering (Extractive QA), and Semantic Role Labeling (SRL). All metrics in NER and SRL are micro-f1 score. (FT: fine-tuning; PT: P-tuning & Lester et al. (2021); PT-2: P-tuning v2; MPT-2: Multi-task P-tuning v2; **bold**: the best; underline: the second best).

Liu, Xiao, et al. "P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2022.

Prompt depth：



(a) RTE (b) BoolQ

Figure 3: Ablation study on prompt depth using BERT-large. "[x-y]" refers to the layer-interval we add continuous prompts (e.g., "21-24" means we are add prompts to transformer layers from 21 to 24). Same amount of continuous prompts added to deeper transformer layers (i.e., more close to the output layer) can yield a better performance than those added to beginning layers.

Liu, Xiao, et al. "P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2022.

Generated Knowledge Prompting for Commonsense Reasoning

ACL 2022

Generated knowledge prompting :

Incorporating external knowledge benefits commonsense reasoning while maintaining the flexibility of pretrained sequence models.

External knowledge:

(1) knowledge bases (coverage ?)
(2) Retrieval based method (flexibility ?)
(3) benefits of external knowledge may wash away as the underlying models increase in size and are pretrained on ever larger amounts of raw text

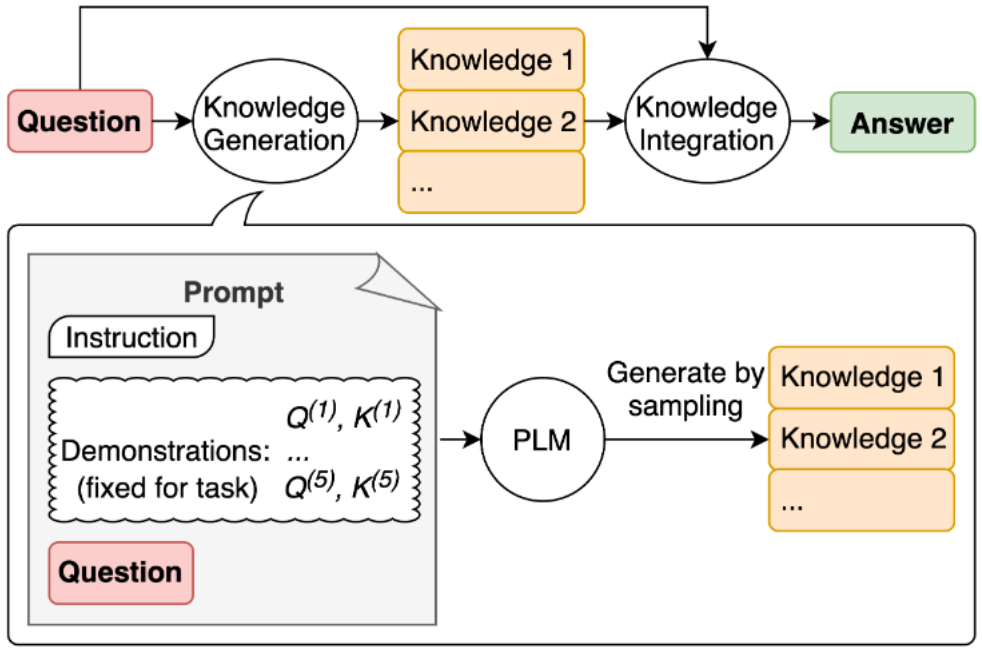


Figure 1: Generated knowledge prompting involves (i) using few-shot demonstrations to generate question-related knowledge statements from a language model; (ii) using a second language model to make predictions with each knowledge statement, then selecting the highest-confidence prediction.

Liu J, Liu A, Lu X, et al. Generated knowledge prompting for commonsense reasoning[J]. arXiv preprint arXiv:2110.08387, 2021

Dataset	Question / *Knowledge*	Prediction	Score
NumerSense	**the word children means [M] or more kids.** *The word child means one kid.*	one two	0.37 | 0.35 0.91
CSQA	**She was always helping at the senior center, it brought her what?** *People who help others are usually happier.*	feel better happiness	0.97 | 0.02 0.98
CSQA2	**Part of golf is trying to get a higher point total than others.** *The player with the lowest score wins.*	yes no	1.00 | 0.00 1.00
QASC	**Sponges eat primarily** *Sponges eat bacteria and other tiny organisms.*	cartilage krill and plankton	0.95 | 0.00 0.99

Table 1: Examples where prompting with generated knowledge rectifies model prediction. Each section shows the correct answer in green, the incorrect answer in red, and the prediction scores from the inference model that only sees the question (top) and the same model that sees the question prompted with the given knowledge (bottom).

(1) Knowledge generation: $K_q = \{k_m : k_m \sim p_G(k|q), m = 1 \dots M\},$

(2) Knowledge integration: $\hat{a} = \arg\max_{a \in A_q} p_I(a|q, K_q)$

(3) Selected Knowledge : $\hat{k} = k_{\hat{m}}$ where $\hat{m} = \arg\max_{0 \le m \le M} \max_{a \in A_q} p_I(a|q_m).$

Liu J, Liu A, Lu X, et al. Generated knowledge prompting for commonsense reasoning[J]. arXiv preprint arXiv:2110.08387, 2021

Knowledge generation:

The prompt consists of an instruction, a few demonstrations that are fixed for each task, and a new-question placeholder.

Task	NumerSense	QASC
Prompt	Generate some numerical facts about objects. Examples:	Generate some knowledge about the input. Examples:
	Input: **penguins have <mask> wings.** Knowledge: *Birds have two wings. Penguin is a kind of bird.*	Input: **What type of water formation is formed by clouds?** Knowledge: *Clouds are made of water vapor.*
	...	...
	Input: **a typical human being has <mask> limbs.** Knowledge: *Human has two arms and two legs.*	Input: **The process by which genes are passed is** Knowledge: *Genes are passed from parent to offspring.*
	Input: **{question}** Knowledge:	Input: **{question}** Knowledge:

Table 2: Prompts for knowledge generation for two of our tasks, NumerSense and QASC. The prompt consists of an instruction, five demonstrations of question-knowledge pairs, and a new question placeholder. For full prompts on all the tasks we evaluate on, see Appendix A.2.

Statement: Birds have two wings. Penguin is a kind of bird.

Question: Penguins have <mask> wings

Liu J, Liu A, Lu X, et al. Generated knowledge prompting for commonsense reasoning[J]. arXiv preprint arXiv:2110.08387, 2021

Results:

		A			B_1	B_2	C		D_1		D_2	
	Dataset	**NumerSense**			**CSQA**	**CSQA**	**CSQA2**		**QASC**		**QASC**	
	Inference Model	T5-11b			T5-11b	UQA-11b-ft	Unicorn-ft		T5-11b		UQA-11b-ft	
		dev	$\text{test}_{\text{core}}$	test_{all}	dev	dev	dev	test	dev	test	dev	test
Knowledge Gen.	($\varnothing$) Vanilla baseline	67.5	70.23	64.05	39.89	85.18	69.9	70.2†	48.16	44.89	81.75	76.74
	(R) Random sentences	68.5	–	–	21.79	**85.42**	70.37	–	49.35	–	82.18	–
	(C) Context sentences	70.5	–	–	42.51	85.34	70.92	–	55.83	–	82.61	–
	(T) Template-based	–	–	–	45.37	–	–	–	–	–	–	–
	(IR) Retrieval-based	–	70.41	65.10**	–	–	**74.0**	**73.3**††	**76.89**	–	**90.06**	–
	(A) Answers	73.0	–	–	51.84	84.93	69.22	–	52.48	–	81.53	–
	(K) Ours	**78.0**	**79.24**	**72.47**	**47.26**	85.34	72.37	73.03	58.32	55.00	84.02	80.33
	prev. SOTA (no IR)	–	72.61	66.18*	–	79.1 (test)#	69.9	70.2†	–	–	81.75	76.74‡
	Few-shot GPT-3 Infer.	60.5	–	–	–	71.58	53.80	–	–	–	66.09	–

Table 3: Experimental results of applying different knowledge generation methods on various tasks and inference models. T5-11b is the zero-shot inference model, whereas other inference models are finetuned based on T5-11b. We **bold** the best and underline the second best numbers. Previous SOTA and retrieval-based methods are also based on the inference model in their corresponding column: * T5-11b 1.1 +digits (Submission by ISI Waltham); ** T5-11b + IR (Yan, 2021); # UQA-11b-ft (Khashabi et al., 2020) (SOTA of single-model methods without referencing ConceptNet); † Unicorn-ft (Talmor et al., 2021); †† Unicorn-ft + Google snippets (Talmor et al., 2021); ‡ UQA-11b-ft (Khashabi et al., 2020).

Liu J, Liu A, Lu X, et al. Generated knowledge prompting for commonsense reasoning[J]. arXiv preprint arXiv:2110.08387, 2021

Results:

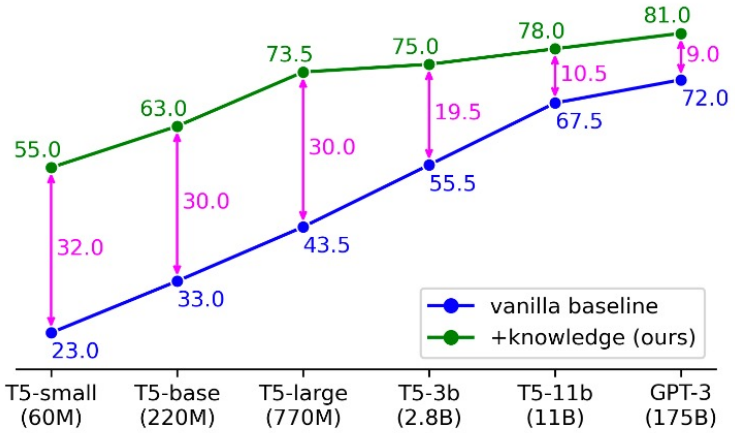


Figure 3: Improvement on top of different sizes of inference model (Numersense dev set).

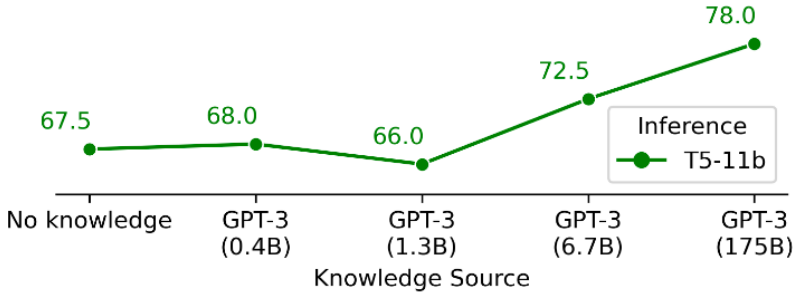


Figure 4: Improvement by different sizes of knowledge generation model (Numersense dev set, T5-11b inference model).

Liu J, Liu A, Lu X, et al. Generated knowledge prompting for commonsense reasoning[J]. arXiv preprint arXiv:2110.08387, 2021

Analysis:

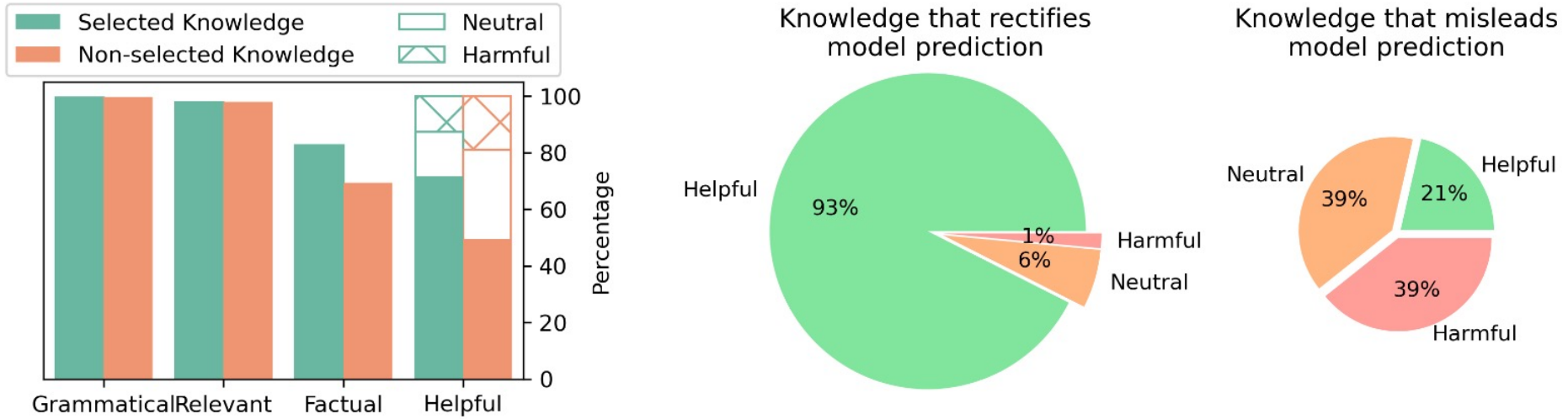


Figure 5: Human evaluation of generated knowledge. **Left:** Percentage of good knowledge statements along each axis. **Right:** Agreement between human and machine on helpfulness of selected knowledge.

Liu J, Liu A, Lu X, et al. Generated knowledge prompting for commonsense reasoning[J]. arXiv preprint arXiv:2110.08387, 2021

Analysis:

Dataset	Question / *Knowledge*	Prediction	Score	Reasoning
NumerSense	**clams have evolved to have [M] shells.** *Clams have a bivalve shell.*	no two	0.37 | 0.18 0.89	Commonsense Paraphrasing
NumerSense	**an easel can have [M] or four legs.** *A tripod is a kind of easel.*	two three	0.45 | 0.45 0.46	Commonsense Induction
CSQA	**Where does a heifer's master live?** *The master of a heifer is a farmer.*	slaughter house farm house	0.89 | 0.01 0.92	Commonsense Deduction
CSQA	**Aside from water and nourishment what does your dog need?** *Dogs need attention and affection.*	walked lots of attention	0.55 | 0.04 0.91	Commonsense Elimination
CSQA	**I did not need a servant. I was not a what?** *People who have servants are rich.*	in charge rich person	0.47 | 0.32 0.99	Commonsense Abduction
CSQA2	**Part of golf is trying to get a higher point total than others.** *The player with the lowest score wins.*	yes no	1.00 | 0.00 1.00	Commonsense Negation
CSQA2	**Eighth plus eight is smaller than fifteen.** *Eighth plus eight is sixteen, which is larger than fifteen.*	yes no	0.97 | 0.03 1.00	Commonsense Numerical
QASC	**[M] is used for transportation.** *Bicycles are used for transportation.*	plastic boats	0.41 | 0.12 0.74	Commonsense Analogy

Table 5: More examples where prompting with generated knowledge reduces the reasoning type and rectifies the prediction. The first row of each section is the original question and the inference results associated with it; the second row is a model-generated knowledge statement that prompts the inference model. We show correct answers in green, incorrect answers in red, and their corresponding scores assigned by the inference model.

Liu J, Liu A, Lu X, et al. Generated knowledge prompting for commonsense reasoning[J]. arXiv preprint arXiv:2110.08387, 2021

My work follows GKP:

(1) 将prompt tuning和deep prompt tuning的方式用在生成模型上生成知识（T5，GPT2）。

(2) 扩充数据集，产生多方面的知识。

(3) 进行多任务训练。