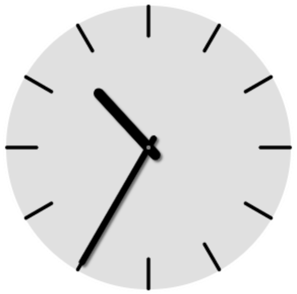
Few-shot Text Classification based on Pretrained Language Models: *An Unfinished* Research *Story*

Mohsen Tabasi


Final decision
Change direction
The first idea
Few-shot learning

Few-Shot Learning

A brief Introduction



Why Just a Few shots!?

- Supervised information are sometimes hard or impossible to acquire
- Large-scale data collection is laborious
- Humans are few-shot learners

This section (Introduction to few-shot learning) is derived from:

Wang, Yaqing, et al. "Generalizing from a few examples: A survey on few-shot learning." ACM Computing Surveys (CSUR) 53.3 (2020): 1-34.

Relevant Problems

- Weakly supervised learning
- Imbalanced learning
- Transfer learning
- Meta-learning

The Core Issue

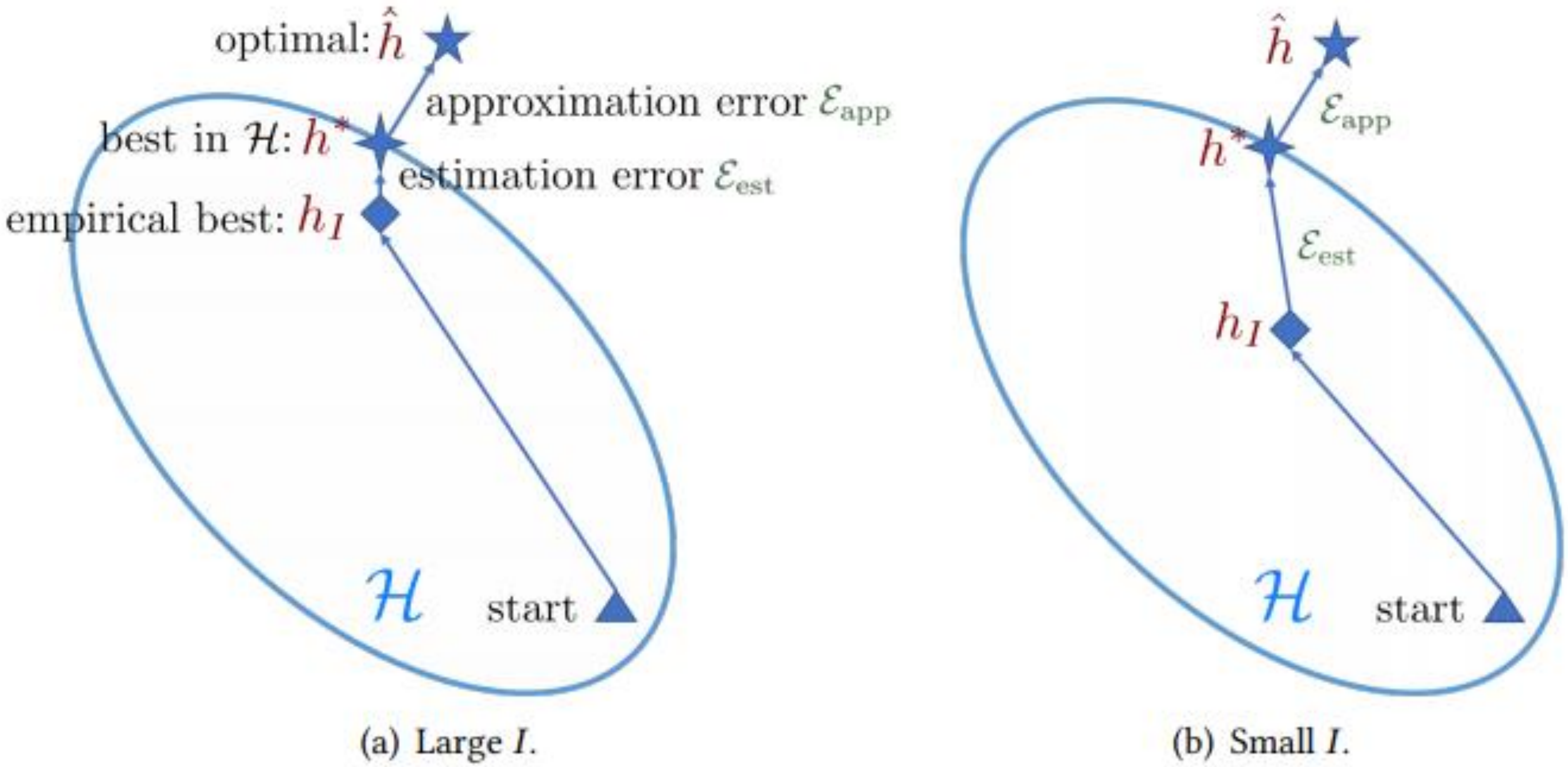


Fig. 1. Comparison of learning with sufficient and few training samples.

FSL Solutions

- Prior Knowledge is the key!

FSL Solutions

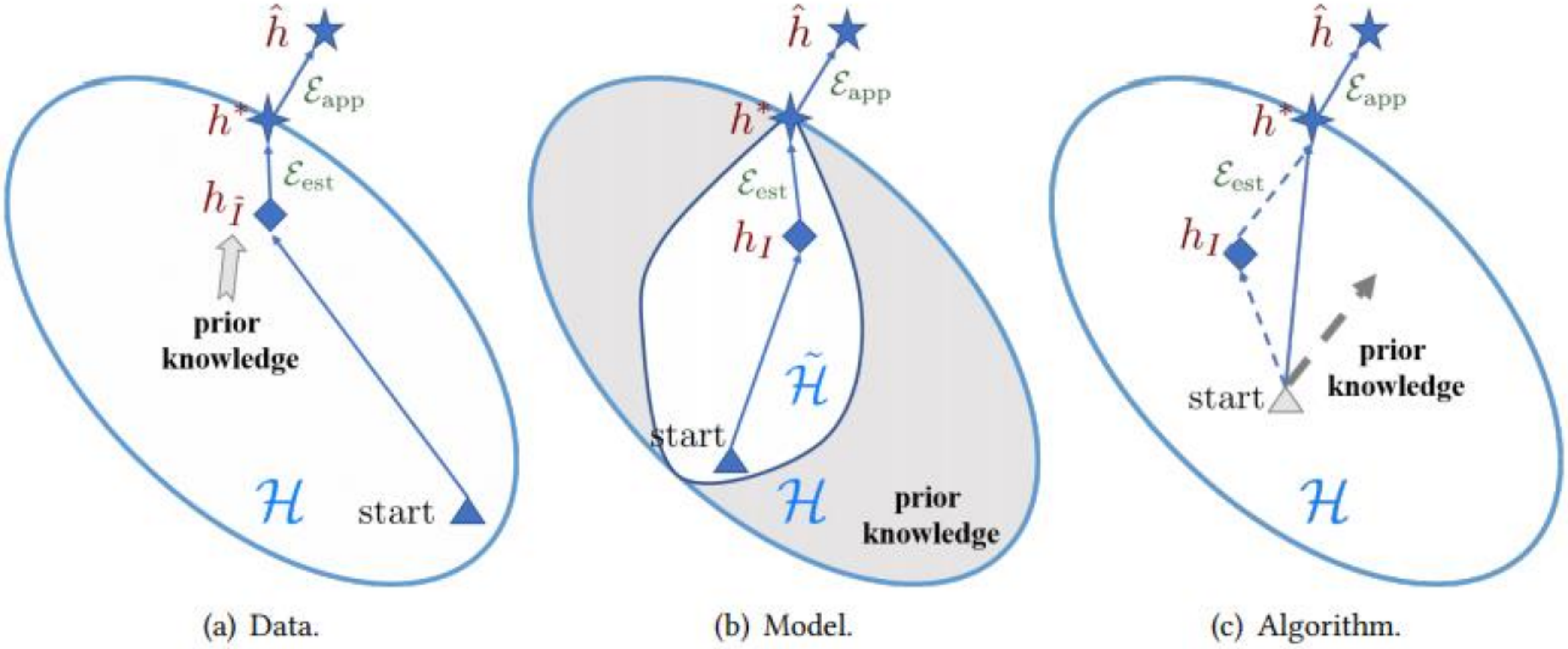
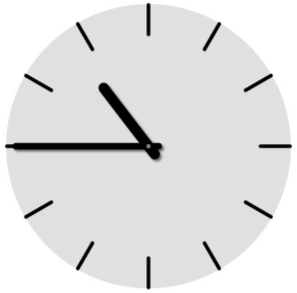


Fig. 2. Different perspectives on how FSL methods solve the few-shot problem.

The First Idea

Towards few-shot text classification



Playing with MLM

- Lets go to colab!

Few-shot w/ Cloze Questions

- Add a fixed pattern with a single [MASK] token to the input text
- Take BERT embeddings or LM probs for the [MASK] as features
- Train a linear classifier on few examples

First Experiments, First Results

- Very promising in sentiment analysis (SST-2)
  - In comparison to Fine-tuning, Using [CLS] token embedding
- Not so impressive for language Inference (MNLI)
  - Not so intuitive patterns, Or maybe the model lacks knowledge!

- Special adaptation for Word-in-Context task
  - On par with fine-tuning approach

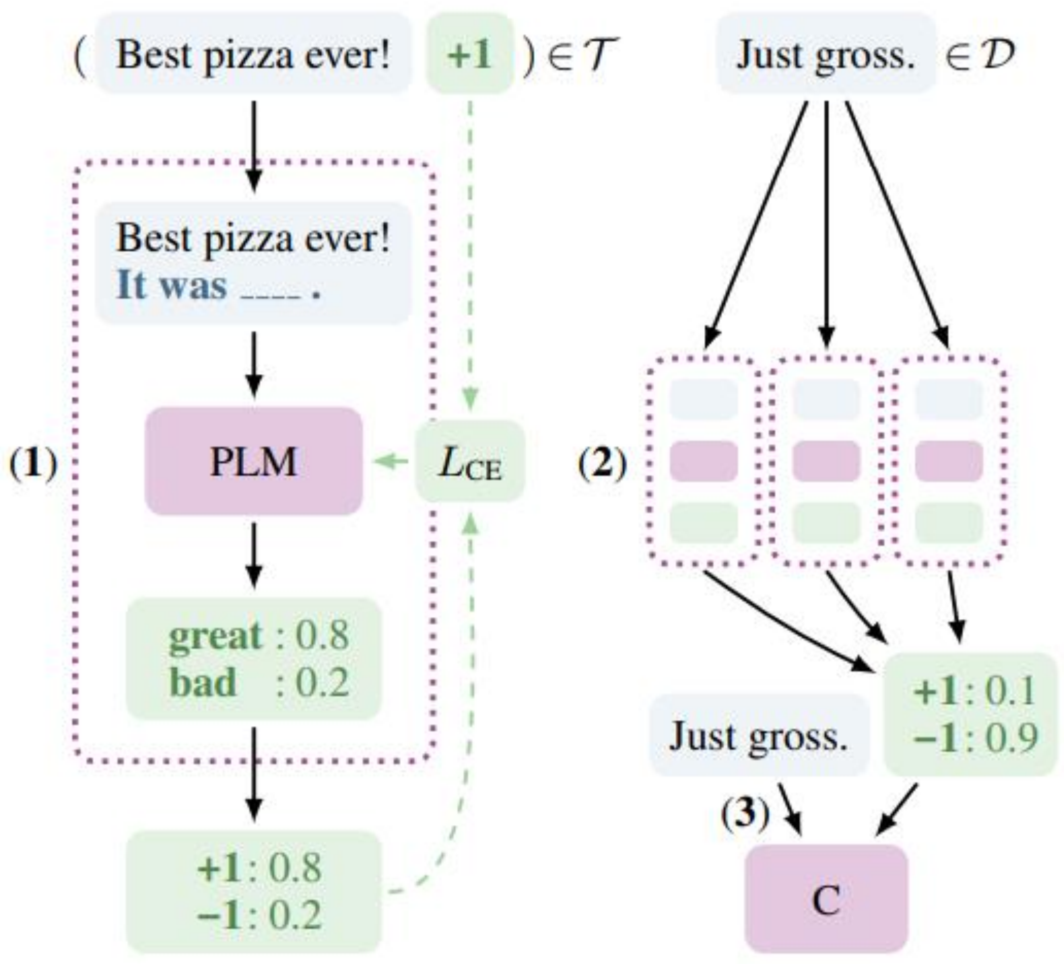
Facing a Bitter Reality ☹

- A random paper search led to an AWESOME paper titled:

"Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference"

- It is accepted at EACL 2021 as we talk…

Pattern Exploiting Training (PET)



Schick, Timo, and Hinrich Schütze. "Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference." *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2021.

Pattern Exploiting Training (PET)

Line	Examples	Method	Yelp	AG's	Yahoo	MNLI (m/mm)
1		unsupervised (avg)	33.8 ±9.6	69.5 ±7.2	44.0 ±9.1	39.1 ±4.3 / 39.8 ±5.1
2	$\|\mathcal{T}\| = 0$	unsupervised (max)	40.8 ±0.0	79.4 ±0.0	56.4 ±0.0	43.8 ±0.0 / 45.0 ±0.0
3		iPET	**56.7** ±0.2	**87.5** ±0.1	**70.7** ±0.1	**53.6** ±0.1 / **54.2** ±0.1
4		supervised	21.1 ±1.6	25.0 ±0.1	10.1 ±0.1	34.2 ±2.1 / 34.1 ±2.0
5	$\|\mathcal{T}\| = 10$	PET	52.9 ±0.1	87.5 ±0.0	63.8 ±0.2	41.8 ±0.1 / 41.5 ±0.2
6		iPET	**57.6** ±0.0	**89.3** ±0.1	**70.7** ±0.1	**43.2** ±0.0 / **45.7** ±0.1
7		supervised	44.8 ±2.7	82.1 ±2.5	52.5 ±3.1	45.6 ±1.8 / 47.6 ±2.4
8	$\|\mathcal{T}\| = 50$	PET	60.0 ±0.1	86.3 ±0.0	66.2 ±0.1	63.9 ±0.0 / 64.2 ±0.0
9		iPET	**60.7** ±0.1	**88.4** ±0.1	**69.7** ±0.0	**67.4** ±0.3 / **68.3** ±0.3
10		supervised	53.0 ±3.1	86.0 ±0.7	62.9 ±0.9	47.9 ±2.8 / 51.2 ±2.6
11	$\|\mathcal{T}\| = 100$	PET	61.9 ±0.0	88.3 ±0.1	69.2 ±0.0	74.7 ±0.3 / 75.9 ±0.4
12		iPET	**62.9** ±0.0	**89.6** ±0.1	**71.2** ±0.1	**78.4** ±0.7 / **78.6** ±0.5
13	$\|\mathcal{T}\| = 1000$	supervised	63.0 ±0.5	**86.9** ±0.4	70.5 ±0.3	73.1 ±0.2 / 74.8 ±0.3
14		PET	**64.8** ±0.1	**86.9** ±0.2	**72.7** ±0.0	**85.3** ±0.2 / **85.5** ±0.4

Table 1: Average accuracy and standard deviation for RoBERTa (large) on Yelp, AG's News, Yahoo and MNLI (m:matched/mm:mismatched) for five training set sizes $|\mathcal{T}|$.

GPT-3 as a few-shot learner

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

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

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

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

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

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

Brown, Tom B., et al. "Language models are few-shot learners." arXiv preprint arXiv:2005.14165 (2020).

GPT-3 as a few-shot learner

Context →	The bet, which won him dinner for four, was regarding the existence and mass of the top quark, an elementary particle discovered in 1995. question: The Top Quark is the last of six flavors of quarks predicted by the standard model theory of particle physics. True or False? answer:
Target Completion →	False

Figure G.31: Formatted dataset example for RTE

Context →	An outfitter provided everything needed for the safari. Before his first walking holiday, he went to a specialist outfitter to buy some boots. question: Is the word 'outfitter' used in the same way in the two sentences above? answer:
Target Completion →	no

Figure G.32: Formatted dataset example for WiC

PET strikes again!

	Model	Params (M)	BoolQ Acc.	CB Acc. / F1	COPA Acc.	RTE Acc.	WiC Acc.	WSC Acc.	MultiRC EM / F1a	ReCoRD Acc. / F1	Avg –
dev	GPT-3 Small	125	43.1	42.9 / 26.1	67.0	52.3	49.8	58.7	6.1 / 45.0	69.8 / 70.7	50.1
	GPT-3 Med	350	60.6	58.9 / 40.4	64.0	48.4	55.0	60.6	11.8 / 55.9	77.2 / 77.9	56.2
	GPT-3 Large	760	62.0	53.6 / 32.6	72.0	46.9	53.0	54.8	16.8 / 64.2	81.3 / 82.1	56.8
	GPT-3 XL	1,300	64.1	69.6 / 48.3	77.0	50.9	53.0	49.0	20.8 / 65.4	83.1 / 84.0	60.0
	GPT-3 2.7B	2,700	70.3	67.9 / 45.7	83.0	56.3	51.6	62.5	24.7 / 69.5	86.6 / 87.5	64.3
	GPT-3 6.7B	6,700	70.0	60.7 / 44.6	83.0	49.5	53.1	67.3	23.8 / 66.4	87.9 / 88.8	63.6
	GPT-3 13B	13,000	70.2	66.1 / 46.0	86.0	60.6	51.1	75.0	25.0 / 69.3	88.9 / 89.8	66.9
	GPT-3	175,000	77.5	82.1 / 57.2	92.0	72.9	**55.3**	75.0	32.5 / 74.8	**89.0 / 90.1**	73.2
	PET	223	79.4	85.1 / 59.4	**95.0**	69.8	52.4	**80.1**	**37.9 / 77.3**	86.0 / 86.5	74.1
	iPET	223	**80.6**	**92.9 / 92.4**	**95.0**	**74.0**	52.2	**80.1**	33.0 / 74.0	86.0 / 86.5	**76.8**
test	GPT-3	175,000	76.4	75.6 / 52.0	**92.0**	69.0	49.4	80.1	30.5 / 75.4	**90.2 / 91.1**	71.8
	PET	223	79.1	87.2 / 60.2	90.8	67.2	**50.7**	**88.4**	**36.4 / 76.6**	85.4 / 85.9	74.0
	iPET	223	**81.2**	**88.8 / 79.9**	90.8	**70.8**	49.3	**88.4**	31.7 / 74.1	85.4 / 85.9	**75.4**
	SotA	11,000	*91.2*	*93.9 / 96.8*	*94.8*	*92.5*	*76.9*	*93.8*	*88.1 / 63.3*	*94.1 / 93.4*	*89.3*

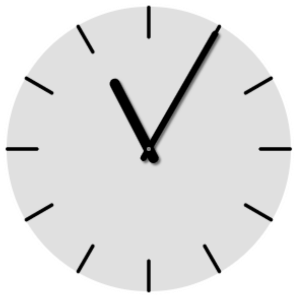
Table 1: Results on SuperGLUE for GPT-3 primed with 32 randomly selected examples and for PET / iPET with ALBERT-xxlarge-v2 after training on FewGLUE. State-of-the-art results when using the regular, full size training sets for all tasks (Raffel et al., 2020) are shown in italics.

Recap

- PET (iPET) leaves no room for few-shot performance improvement!
- Although PET was published several month before in arxiv, It was neither accepted in any conference, nor being referenced by other works.
- This direction seems to be still intact in some aspects…
- Which aspects?!?!

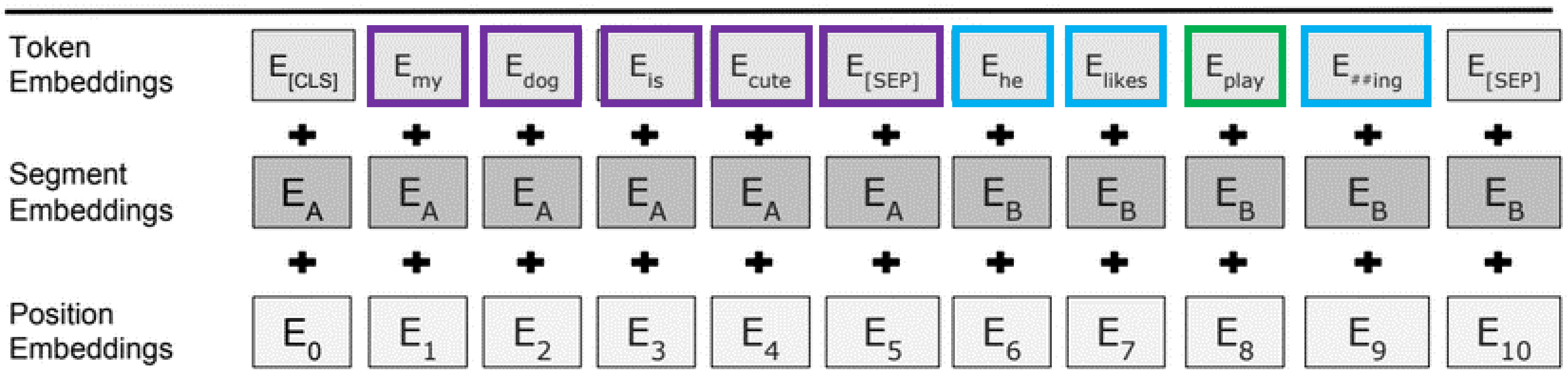
Heroic Exercise! ☺

Not leaving this so easily…



Learn the Pattern

- As PET seems to get the most out of cloze questions, we can search for best possible pattern
  - Choose a pattern template, e.g. [sentence] [PAD] [PAD] [MASK] [PAD]
  - Learn an embedding vector for each [PAD] token
  - Set nearest in-vocab word for each position as the final pattern



Learn the Pattern

- As PET seems to get the most out of cloze questions, we can search for best possible pattern
  - Choose a pattern template, e.g. [sentence] [PAD] [PAD] [MASK] [PAD]
  - Learn an embedding vector for each [PAD] token
  - Set nearest in-vocab word for each position as the final pattern
- Failed! Why?
- Improvement when starting from a valid pattern!

Learn the Input (Not Few-shot Only)

- DeepDream

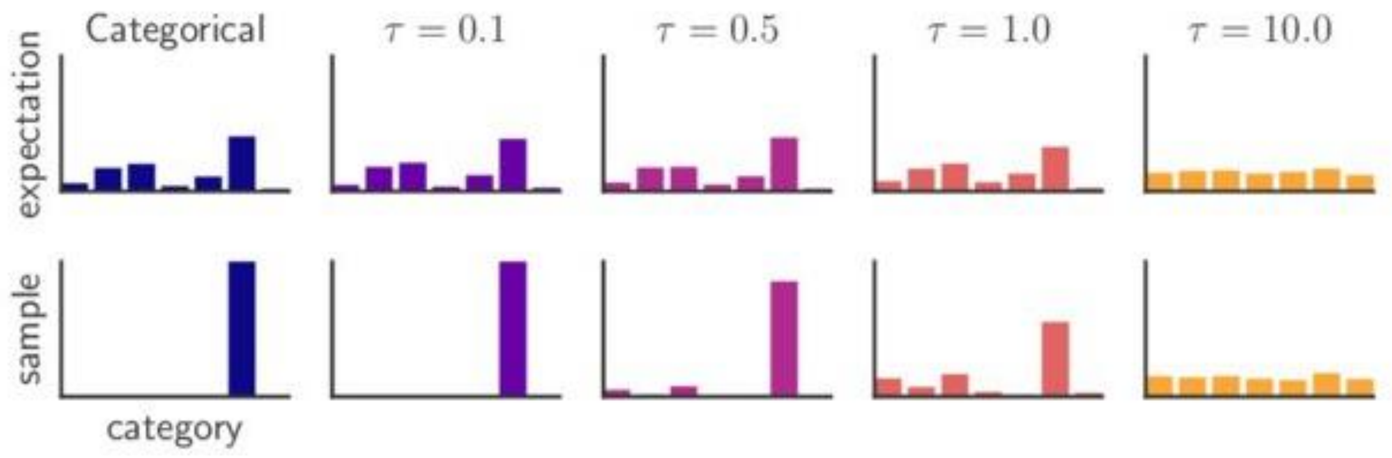
Learn the Input (Not Few-shot Only)

- If we can find an input text which satisfies a given objective, we can move towards…

- Text Dream
- Model Interpretation
- Adversarial Attack

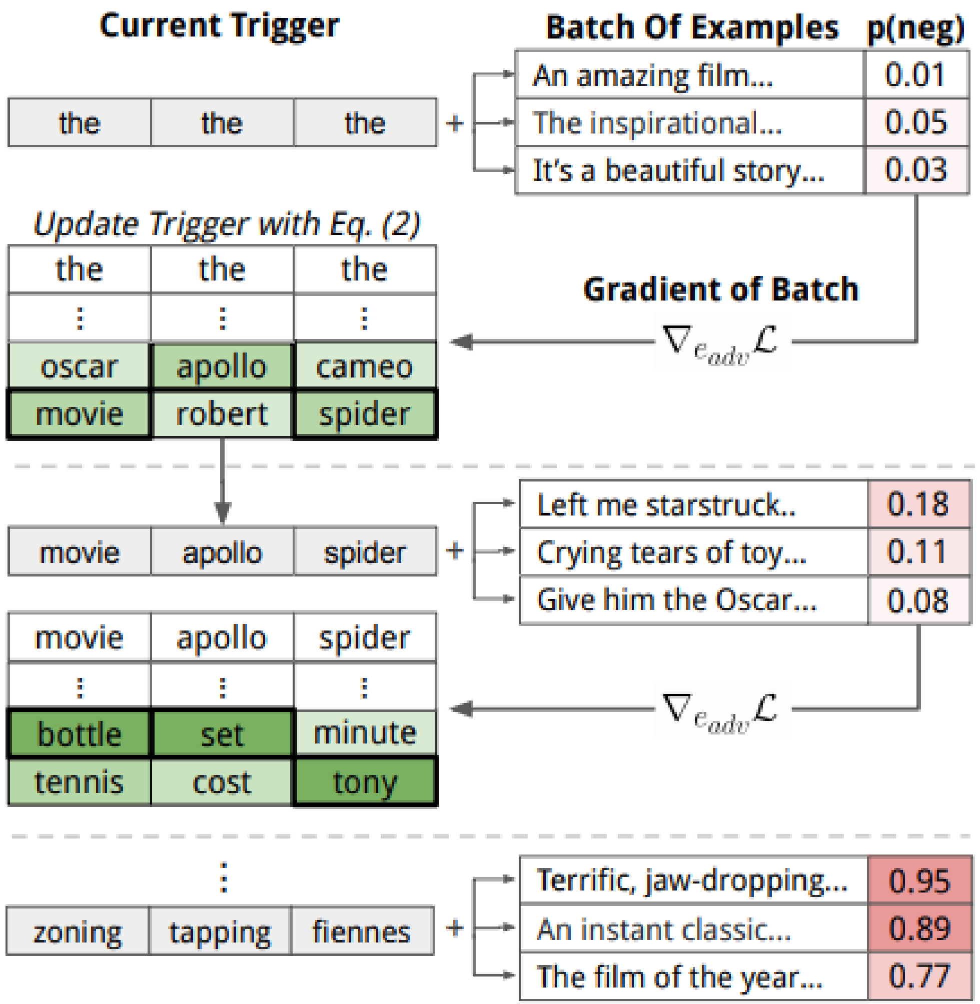
Improve Input Search method

- Learn weights of a Gumbel Softmax instead of embedding vectors



- Beam Search
  - The most promising search method, which let us return to learning patterns for few-shot text classification

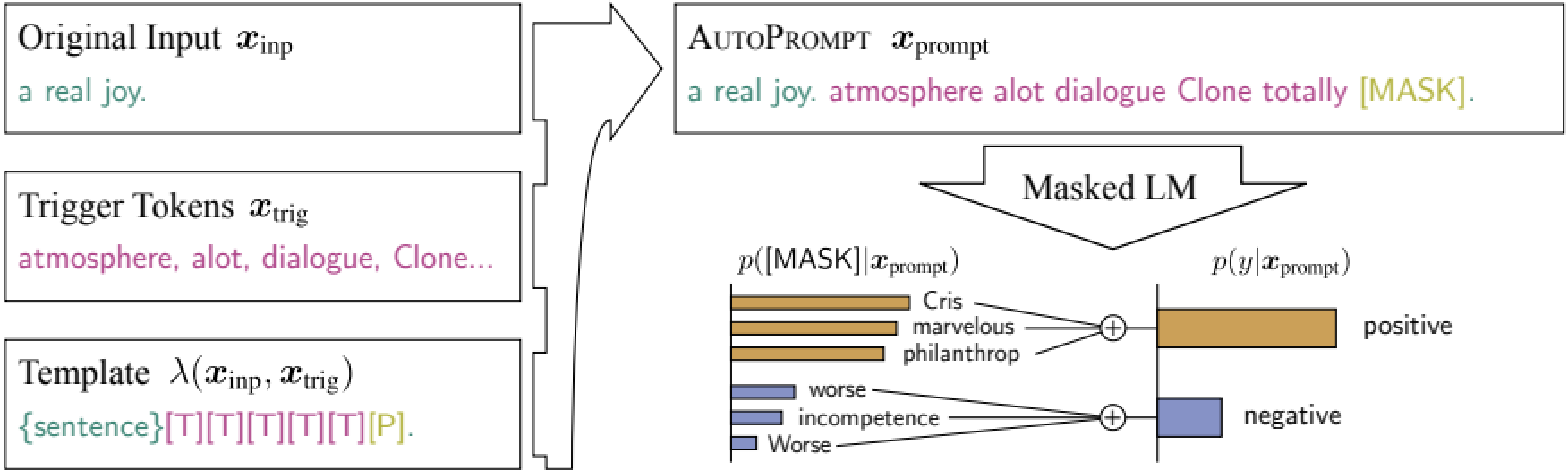
Improve Input Search method



Facing another Bitter Reality!

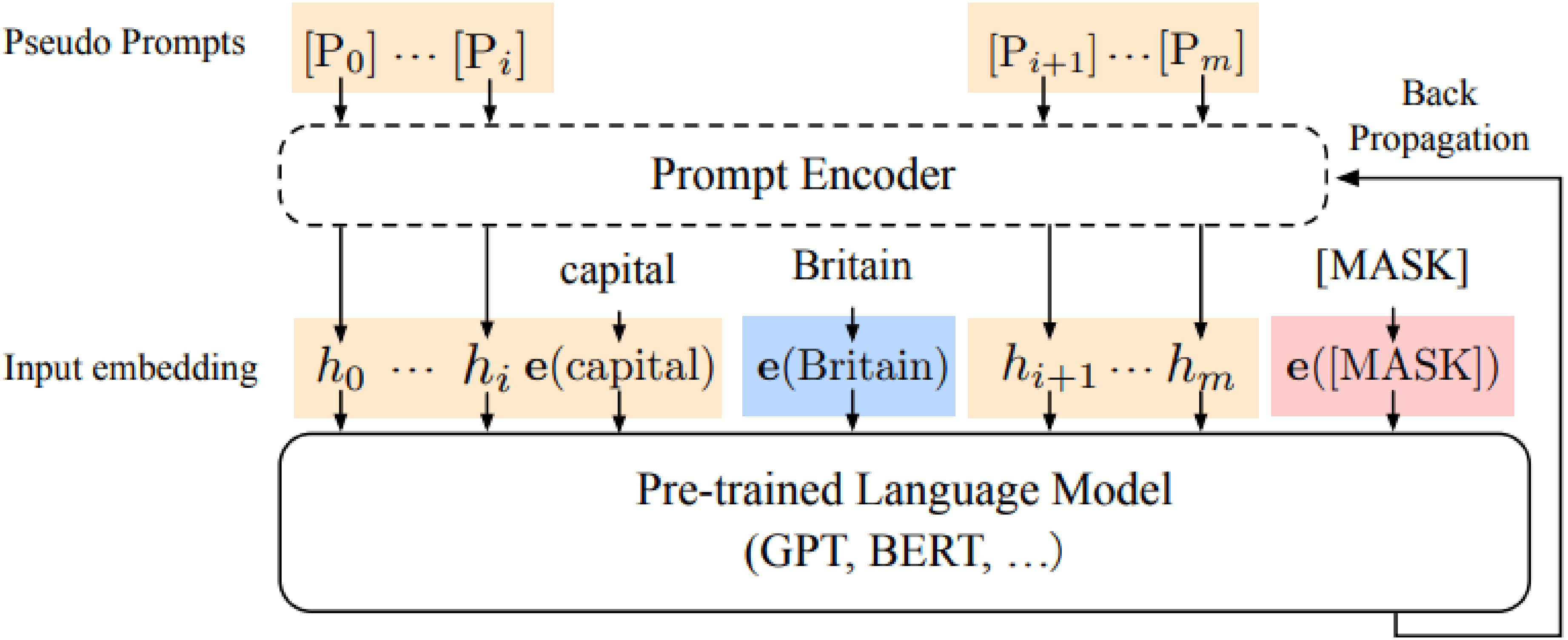
- While we were patiently looking for a promising pattern search method…
- Few-shot text classification using cloze questions (or prompts) has become a (rather small) trend…

AutoPrompt



Shin, Taylor, et al. "Eliciting Knowledge from Language Models Using Automatically Generated Prompts." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.

P-Tuning



Liu, Xiao, et al. "GPT Understands, Too." arXiv preprint arXiv:2103.10385 (2021).

Recap 2

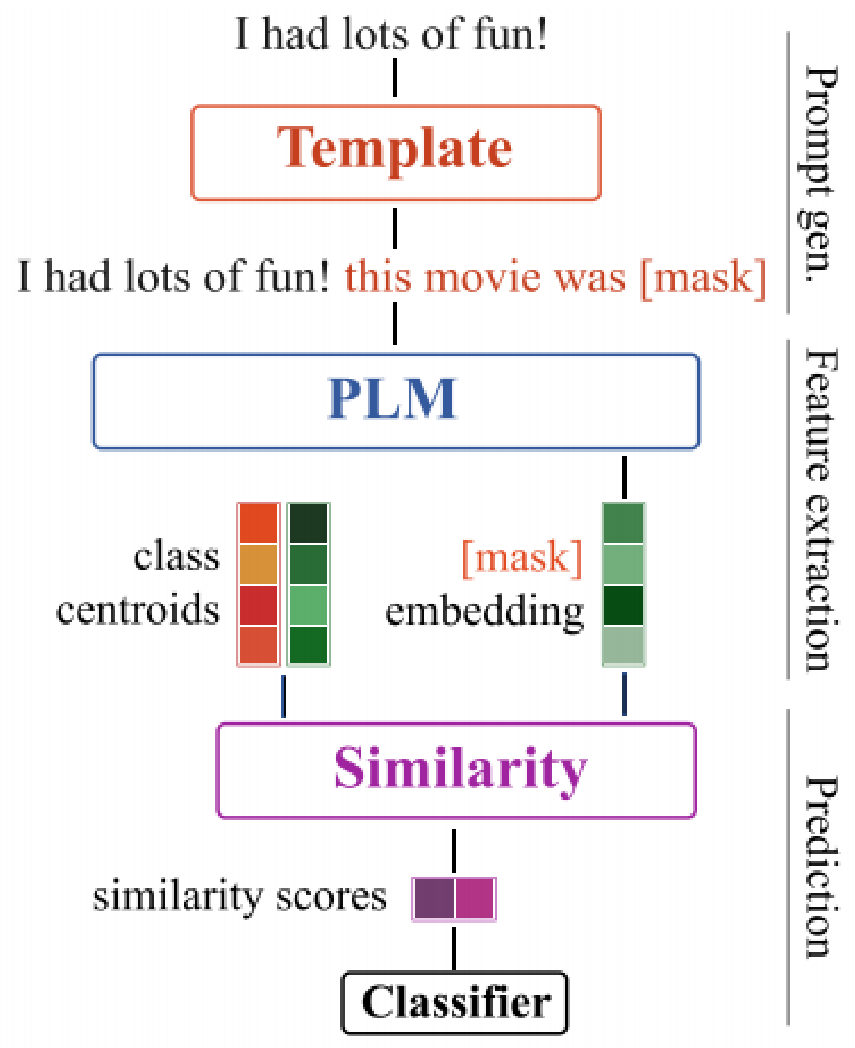
- We were one (or more) steps behind a new trend, in which we could be pioneers!
- Some of the results we simply skip, may become the main idea of other articles…

The Final Decision

How to avoid falling behind?

Publishing Our Findings

"Exploiting Language Model Prompts using Similarity Measures: A Case Study on the Word-in-Context Task"



Publishing Our Findings

Method	WiC	
	dev	**test**
Random Baseline	50.0	50.0
Fine-tuned RoBERTa-Large	-	69.9
GPT3 few-shot	55.3	49.4
PET (ALBERT-xxlarge-v2)	52.4	50.7
P-Tuning (GPT2-medium)	56.3	-
SP-cosine	60.9	63.6
SP-Spearman	70.2	70.2
SP-RBO	66.6	63.4
SP-RBO w/ stem	**71.1**	**70.9**

Table 1: Accuracy scores for Word-in-Context task. SP models are based on RoBERTa-Large.

Publishing Our Findings

Method	SST-2	SICK-E	
		standard	balanced
Majority	50.0	56.7	33.3
FT BERT	93.5	86.7	84.0
AutoPrompt	85.2	-	-
SP-cosine	89.1	74.3	76.0
SP-Spearman	91.1	73.6	76.2
SP-RBO	88.7	71.4	74.2
AutoPrompt*	91.4	65.0	69.3
SP-cosine*	90.0	50.1	55.9
SP-Spearman*	90.4	45.0	51.0
SP-RBO*	90.6	53.2	55.9

Table 2: Test set accuracy on SST-2 and SICK-E tasks. Methods marked with * use the template found by AutoPrompt (Shin et al., 2020) while other prompt-based methods use manual templates. SP and AutoPrompt methods are based on RoBERTa-Large.

Publishing Our Findings

Prompt1 (Top-5 words)	Prompt2 (Top-5 words)	Prediction	Ground Truth
The drawing or — of water from the well. (use, extraction, taking, pumping, consumption)	He did complicated pen-and-ink drawings or — like medieval miniatures. (paintings, sculptures, something, more, looked)	Not matched	Not matched
The body or — of the car was badly rusted. (trunk, roof, chassis, frame, grill)	Administrative body or —. (agency, institution, government, commission, equivalent)	Not matched	Not matched
The main body of the sound or — ran parallel to the coast. (river, bay, sea, ocean, channel)	He strained to hear the faint sounds or —. (voices, footsteps, whispers, conversations, cries)	Not matched	Not matched
He could not conceal his hostility or —. (anger, disgust, irritation, contempt, frustration)	He could no longer contain his hostility or —. (anger, rage, frustration, aggression, disgust)	Matched	Not matched
There was a blockage or — in the sewer, so we called out the plumber. (something, leak, obstruction, defect, overflow)	We had to call a plumber to clear out the blockage or — in the drainpipe. (debris, obstruction, water, leak, crack)	Matched	Not matched
She used to wait or — down at the Dew Drop Inn. (sit, work, sleep, gamble, wash)	Wait or — here until your car arrives. (sit, stand, park, wait, stay)	Matched	Not matched

Continue Exploring

- Use generative LMs (GPT-2)
- Get rid of a fixed pattern and single mask token
- Generate class descriptors with custom beam search decoding

Continue Exploring

```
pattern: " remake",   prob: 0.0027, diff: 0.0642, val_prob: 0.0047, val_diff: 0.2756
pattern: " sequel",   prob: 0.0044, diff: 0.0951, val_prob: 0.0071, val_diff: 0.2633
pattern: " parody",   prob: 0.0046, diff: 0.0972, val_prob: 0.0050, val_diff: 0.1256
pattern: " disaster", prob: 0.0029, diff: 0.1826, val_prob: 0.0026, val_diff: 0.1197
pattern: " horror",   prob: 0.0022, diff: 0.1016, val_prob: 0.0025, val_diff: 0.1169
pattern: " complete", prob: 0.0065, diff: 0.2084, val_prob: 0.0064, val_diff: 0.0964
pattern: " sad",      prob: 0.0029, diff: 0.1036, val_prob: 0.0025, val_diff: 0.0891
pattern: " failure",  prob: 0.0016, diff: 0.1340, val_prob: 0.0012, val_diff: 0.0886
LENGTH=2: 100% 64/64 [01:02<00:00, 1.03it/s]

pattern: " complete failure", prob: 0.0347, diff: 0.5473, val_prob: 0.0313, val_diff: 0.5793
pattern: " complete waste",  prob: 0.0260, diff: 0.7117, val_prob: 0.0257, val_diff: 0.4531
pattern: " total failure",   prob: 0.0393, diff: 0.3501, val_prob: 0.0357, val_diff: 0.5599
pattern: " total waste",     prob: 0.0181, diff: 0.5776, val_prob: 0.0191, val_diff: 0.5343
pattern: " shoddy",          prob: 0.1251, diff: 0.5867, val_prob: 0.1267, val_diff: 0.6957
pattern: " sad example",     prob: 0.0188, diff: 0.0536, val_prob: 0.0180, val_diff: 0.3511
pattern: " total disaster",  prob: 0.0340, diff: 0.3309, val_prob: 0.0352, val_diff: 0.4283
pattern: " terrible example", prob: 0.0223, diff: 0.0064, val_prob: 0.0222, val_diff: 0.3898
LENGTH=3: 100% 64/64 [00:40<00:00, 1.57it/s]

pattern: " complete failure to",    prob: 0.0262, diff: 0.0580, val_prob: 0.0220, val_diff: 0.2777
pattern: " total failure to",       prob: 0.0209, diff: 0.0687, val_prob: 0.0177, val_diff: 0.3208
pattern: " failure to address",     prob: 0.0048, diff: 0.0549, val_prob: 0.0051, val_diff: 0.2569
pattern: " shoddy remake",          prob: 0.0132, diff: 0.0204, val_prob: 0.0165, val_diff: 0.1716
pattern: " failure to understand",  prob: 0.0191, diff: 0.0469, val_prob: 0.0164, val_diff: 0.2287
pattern: " complete failure by",    prob: 0.0048, diff: 0.0067, val_prob: 0.0047, val_diff: 0.0908
```

Questions?

References

- Wang, Yaqing, et al. "Generalizing from a few examples: A survey on few-shot learning." ACM Computing Surveys (CSUR) 53.3 (2020): 1-34.
- Schick, Timo, and Hinrich Schütze. "Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference." *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2021.
- Brown, Tom B., et al. "Language models are few-shot learners." arXiv preprint arXiv:2005.14165 (2020).
- Schick, Timo, and Hinrich Schütze. "It’s Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners." Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021.
- Shin, Taylor, et al. "Eliciting Knowledge from Language Models Using Automatically Generated Prompts." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020.
- Liu, Xiao, et al. "GPT Understands, Too." arXiv preprint arXiv:2103.10385 (2021).