Pre-training
&
In-Context Learning

Shansan Gong & Lei Li

2023/09

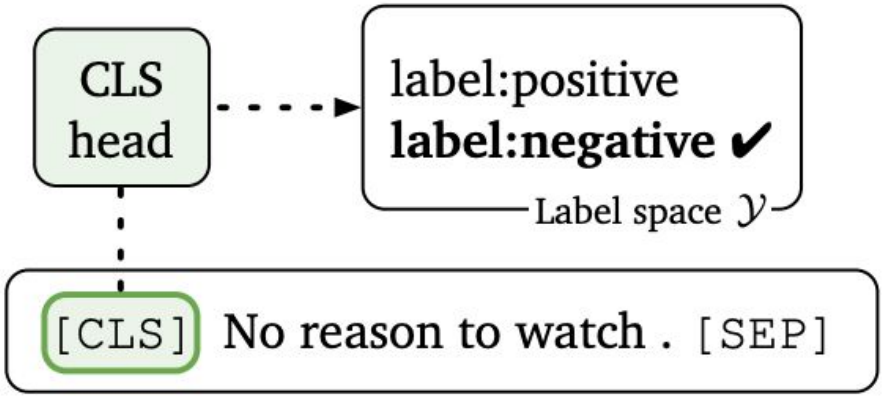
{hisansas, nlp.lilei}@gmail.com

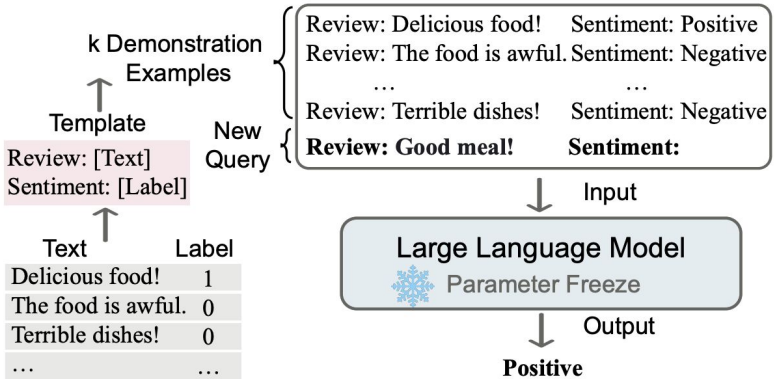
In-Context Learning

Outline

- **What is In-context Learning?**
  - **Characteristics**
  - **Ingredients**
- **Dive into the Demonstration world**
  - **Selection, Ordering, Formatting**
  - **Instablity of ICL**
  - **The role of demonstraion**
- **Chain-of-Thought Prompting for Complex Reasoning Tasks**
  - **Few-shot CoT**
  - **Zero-shot CoT**
- **Discussion & QA**
  - **Recent Trend**
  - **Useful Resource**

What is In-Context Learning (ICL) ?





Fine-tuning v.s. In-Context Learning

Making Pre-trained Language Models Better Few-shot Learners, Gao et al., ACL 2021

Characteristics of ICL

- **Emergence**: LMs with sufficient scale (empirically > 6B), trained with sufficient language tokens (10+ tokens / parameter).

- **Gradient-free Learning**: LMs inductively learn from the input context, make predictions without any updates on the parameters.

- **Human-friendly Interface**: natural language tokens prompting provides an interpretable interface for human to communicate with LMs.

Key Ingredients of ICL

- **Scoring Function**: How to obtain (stable) results from LM predictions?
  - *Calibrate Before Use: Improving Few-Shot Performance of Language Models*
  - *Noisy Channel Language Model Prompting for Few-Shot Text Classification*

- **Model Warmup:** How to adapt LMs for ICL?
  - *MetaICL: Learning to Learn In Context*

- **Demonstration Construction**: How to construct suitable in-context examples?
  - *Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?*
  - *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*

Instablity of Vanilla ICL

Demonstration in ICL

- Example Selection

Review: Good Taste! Sentiment: Positive
Review: No service at all. Sentiment: Negative
Review: Amazing dishes! Sentiment: _____

v.s

Review: Good Taste! Sentiment: Positive
Review: Amazing dishes! Sentiment: Positive
Review: No service at all. Sentiment: _____

- Example Ordering (Permutation)

Review: Good Taste! Sentiment: Positive
Review: No service at all. Sentiment: Negative
Review: Amazing dishes! Sentiment: _____

v.s

Review: No service at all. Sentiment: Negative
Review: Good Taste! Sentiment: Positive
Review: Amazing dishes! Sentiment: _____

- Example Formatting

Review: Good Taste! Sentiment: Positive
Review: No service at all. Sentiment: Negative
Review: Amazing dishes! Sentiment: _____

v.s

Q: What's the sentiment of "Good Taste!" A: Positive
Q: What's the sentiment of "No service at all." A: Negative
Q: What's the sentiment of "Amazing dishes!" A: _____

Demonstration Selection Matters

Trial	1	2	3	4	5
Accuracy	94.6	95.0	95.8	93.9	86.9

Table 1: Results of GPT-3 on the task of sentiment analysis on the SST-2 dataset. Five different in-context examples are randomly selected from the training set.

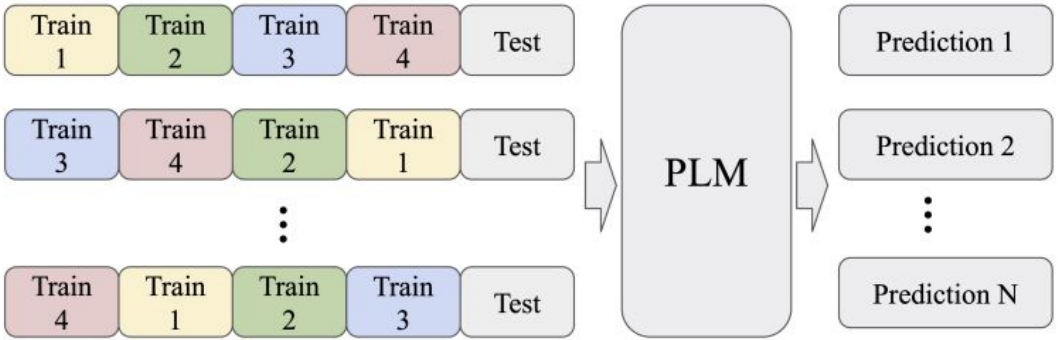
Commonly Selection Criterion:

- Sentence Embedding Similarity
- Perplexity
- Diversity
- …

Overall, there a lots of strategies to be explored.

What Makes Good In-Context Examples for GPT-3?, Liu et al., 2021

Demonstration Ordering Matters, too!



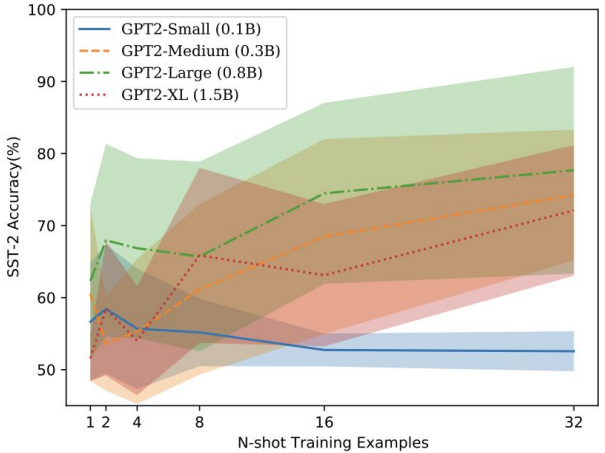


Figure 3: Order sensitivity using different numbers of training samples.

Different orders have great impact (variance in the right plot) on the performance.

Demonstration Sensitivity Aggreated

Calibrate Before Use: Improving Few-Shot Performance of Language Models

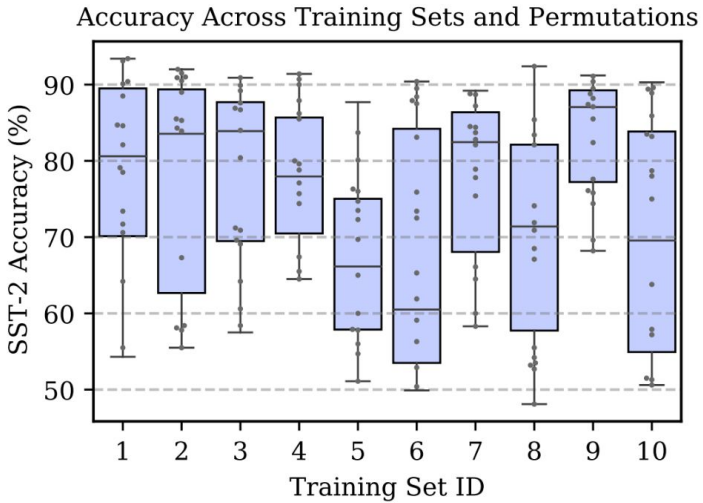


Figure 2. There is high variance in GPT-3's accuracy as we change the prompt's **training examples**, as well as the **permutation** of the examples. Here, we select ten different sets of four SST-2 training examples. For each set of examples, we vary their permutation and plot GPT-3 2.7B's accuracy for each permutation (and its quartiles).

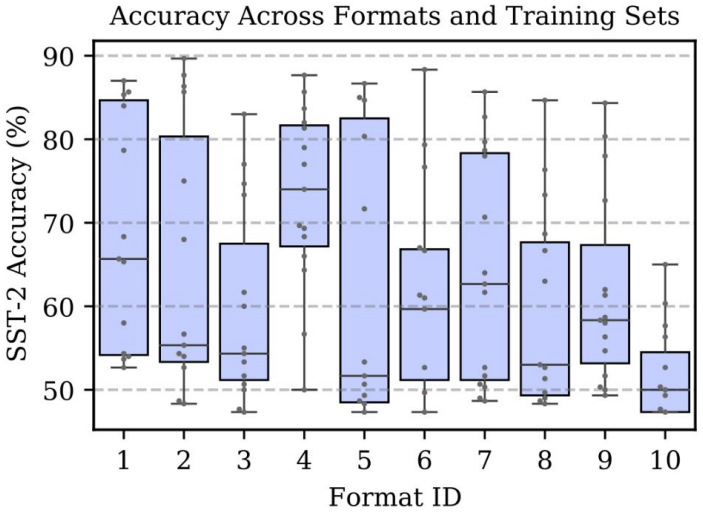
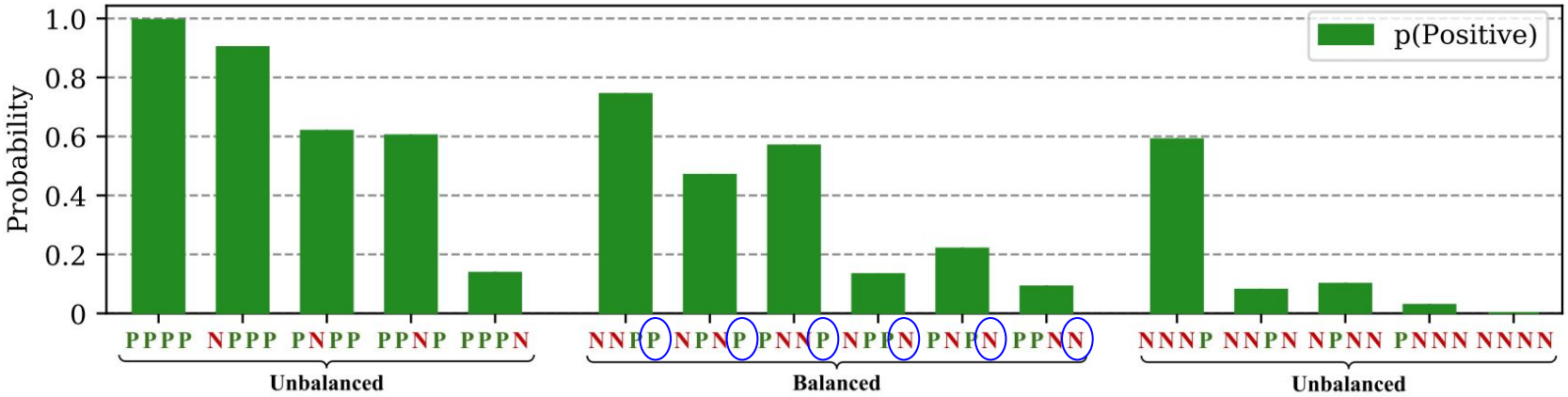


Figure 3. There is high variance in GPT-3's accuracy as we change the **prompt format**. In this figure, we use ten different prompt formats for SST-2. For each format, we plot GPT-3 2.7B's accuracy for different sets of four training examples, along with the quartiles.

Calibrate Before Use: Improving Few-Shot Performance of Language Models, Zhao et al., ICML 2021

Biases in ICL: Label & Recency



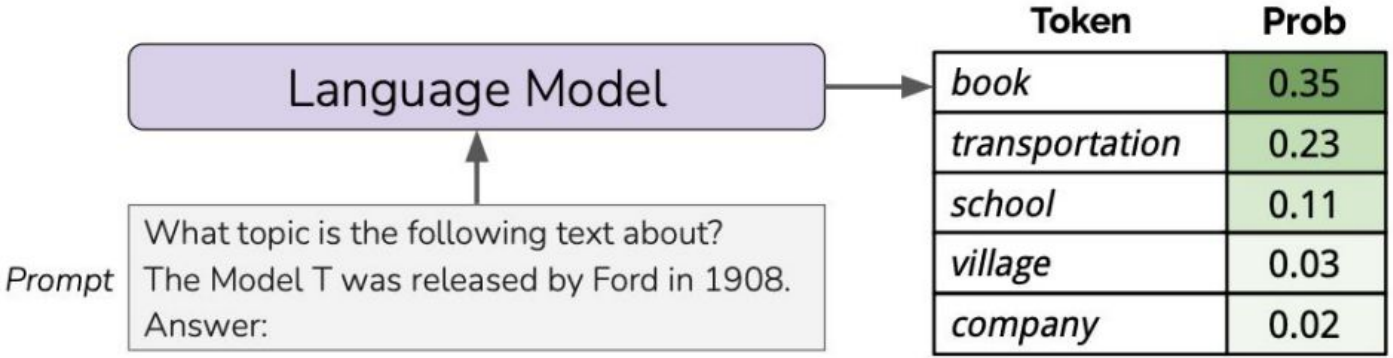
Model prefers to predict positive when the majority labels is "P/Positive"

Model is heavily biased towards the most recent label

Model prefers to predict positive when the majority labels is "N/Negative"

Calibrate Before Use: Improving Few-Shot Performance of Language Models, Zhao et al., ICML 2021

Biases in ICL: Common Token Bias



	Token	Web (%)	Label (%)	Prediction (%)
❌	*book*	0.026	9	29
✅	*transportation*	0.0000006	9	4

Model is biased towards the incorrect frequent token "book" even when both "book" and "transportation" are equally likely labels in the dataset.

https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec08.pdf

Solution: Contextualization Calibration

Step 1: Estimate the bias via Content-free test input

Input: Subpar acting. Sentiment: Negative
Input: Beautiful film. Sentiment: Positive
Input: N/A Sentiment:

Positive: 0.65
Negative: 0.35

Step 2: Counter the bias via affine transformation

bias term

$$\hat{\mathbf{q}} = \text{softmax}(\mathbf{W}\hat{\mathbf{p}} + \mathbf{b}),$$

calibrated prob

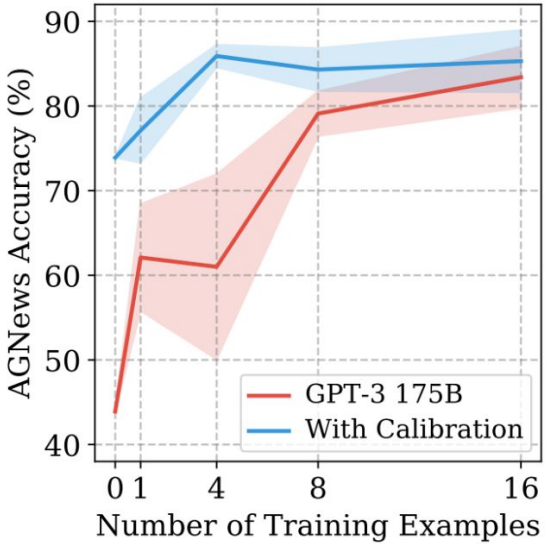
original prob

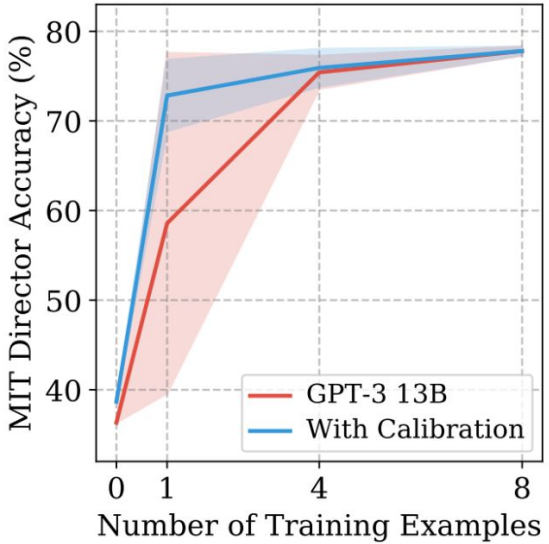
$$\mathbf{W} = \begin{bmatrix} \frac{1}{0.65} & 0 \\ 0 & \frac{1}{0.35} \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

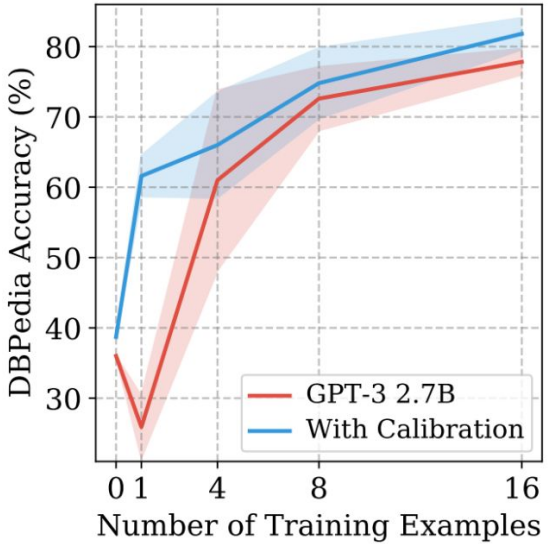
W: diagonal matrix
b: bias set to zeros

https://www.cs.princeton.edu/courses/archive/fall22/cos597G/lectures/lec08.pdf

Results After Calibration







Reduces variance across training sets and permutations, for all three tested GPT-3 models

Calibrate Before Use: Improving Few-Shot Performance of Language Models, Zhao et al., ICML 2021

Alternative: Noisy Channel Model

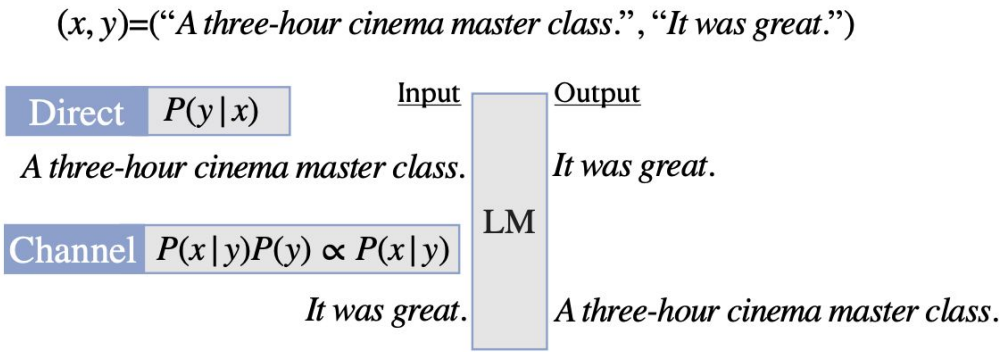


Figure 1: An illustration of the direct model and the channel model for language model prompting in the sentiment analysis task.

Channel models compute the conditional probability of the input given the output,

P(y | x) -> P (x | y)

Noisy Channel Language Model Prompting for Few-Shot Text Classification, Min et al. ACL 2022

Model Warmup for ICL

Model Warmup

There still exists **gap between pre-training and in-context learning**.

Narrowing the Gap: Model Warmup.

- MetaICL: Train LMs on ICL tasks, then generalize to unseen tasks.

	Meta-training	Inference
Task	C *meta-training* tasks	An unseen *target* task
Data given	Training examples $\mathcal{T}_i = \{(x_j^i, y_j^i)\}_{j=1}^{N_i}$, $\forall i \in [1, C]$ $(N_i \gg k)$	Training examples $(x_1, y_1), \cdots, (x_k, y_k)$, Test input x
Objective	For each iteration, 1. Sample task $i \in [1, C]$ 2. Sample $k+1$ examples from $\mathcal{T}_i$: $(x_1, y_1), \cdots, (x_{k+1}, y_{k+1})$ 3. Maximize $P(y_{k+1} \vert x_1, y_1, \cdots, x_k, y_k, x_{k+1})$	$\text{argmax}_{c \in \mathcal{C}} P(c \vert x_1, y_1, \cdots, x_k, y_k, x)$

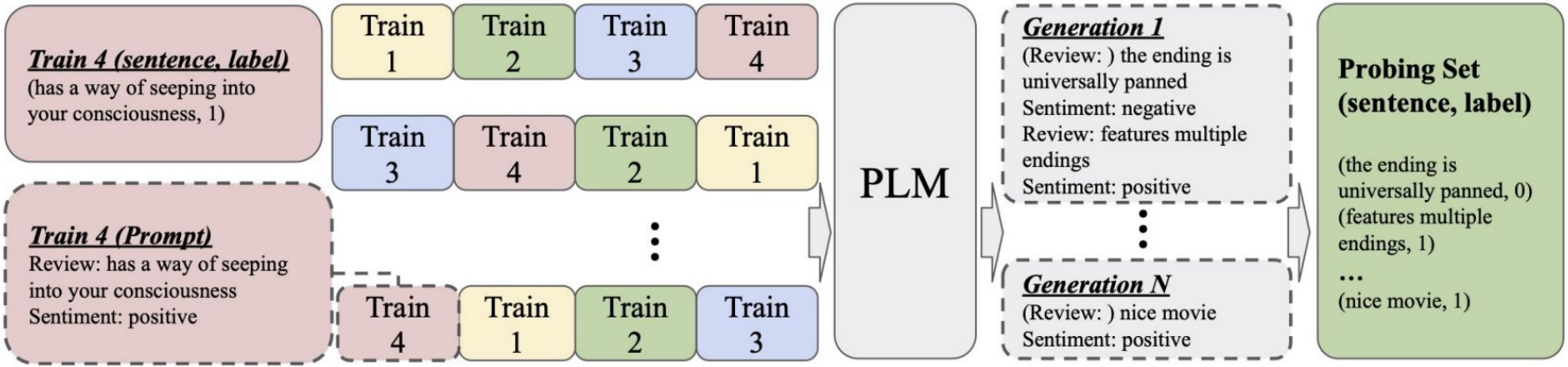
Model Warmup

	Method	HR→LR	Class →Class	non-Class →Class	QA →QA	non-QA →QA	non-NLI →NLI	non-Para →Para
				All target tasks				
GPT-2 Large 774M	Channel In-context	43.1/38.5	46.3/40.3	46.3/40.3	40.8/38.1	40.8/38.1	39.9/34.8	45.4/40.9
	MetaICL	43.3/41.7	43.4/39.9	38.1/31.8	46.0/44.8	38.5/36.8	49.0/44.8	33.1/33.1
	Channel MetaICL	**49.1**/46.8	50.7/48.0	50.6/48.1	44.9/43.5	42.1/40.8	**54.6**/51.9	**52.2**/50.3
	GPT-J Channel In-context	48.6/44.4	**51.5**/47.0	**51.5**/47.0	**47.0**/45.2	**47.0**/45.2	47.2/41.7	51.0/47.5
				Target tasks in unseen domains				
GPT-2 Large 774M	Channel In-context	39.6/33.6	39.6/33.6	39.6/33.6	44.7/40.6	44.7/40.6	40.4/35.7	44.1/36.8
	MetaICL	35.3/32.7	32.3/29.3	28.1/25.1	**69.9**/68.1	48.3/47.2	**80.1**/77.2	34.0/34.0
	Channel MetaICL	**47.7**/44.7	41.9/37.8	**48.0**/45.2	57.9/56.6	47.2/45.0	62.0/57.3	51.0/49.9
	GPT-J Channel In-context	42.8/38.4	**42.8**/38.4	42.8/38.4	55.7/54.4	**55.7**/54.4	51.1/40.4	**52.0**/46.5

Comparison between GPT-2 Large based models with raw LM baselines based on GPT-J which consists of 6B parameters.

MetaICL, despite being 8x smaller, outperforms or matches GPT-J baselines

Demonstration Ordering



Generate a probing dataset, and estimate the global and local entropy of each candidate demonstration order.

Optimal: the global labels and local prediction entropy is uniformly distributed.

Demonstration Formatting: Instruction Generation

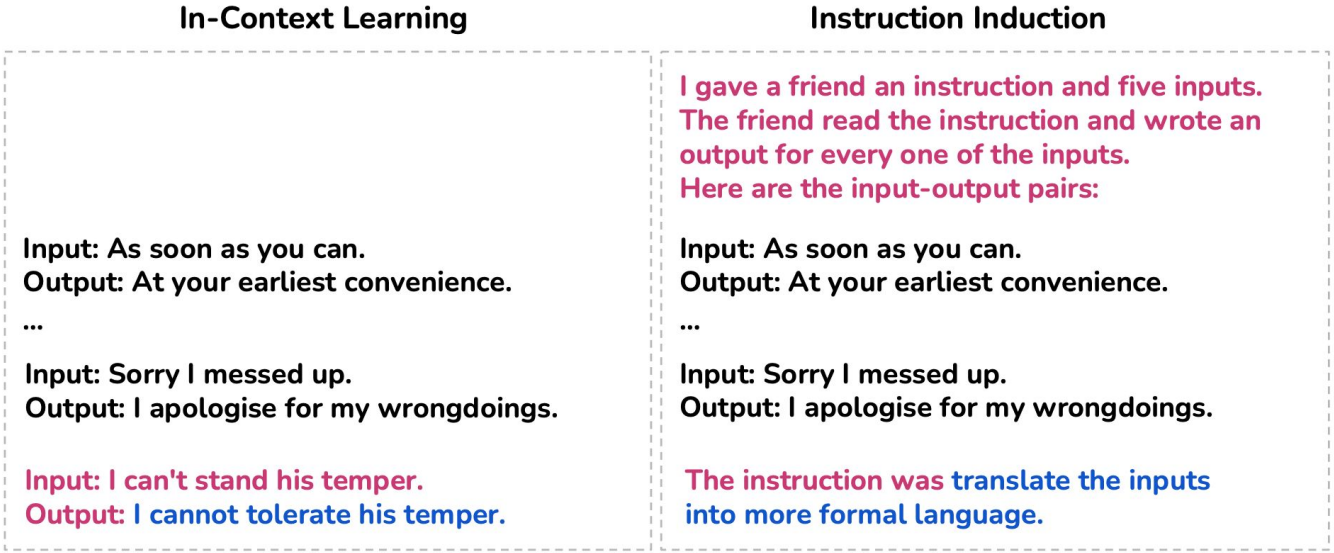
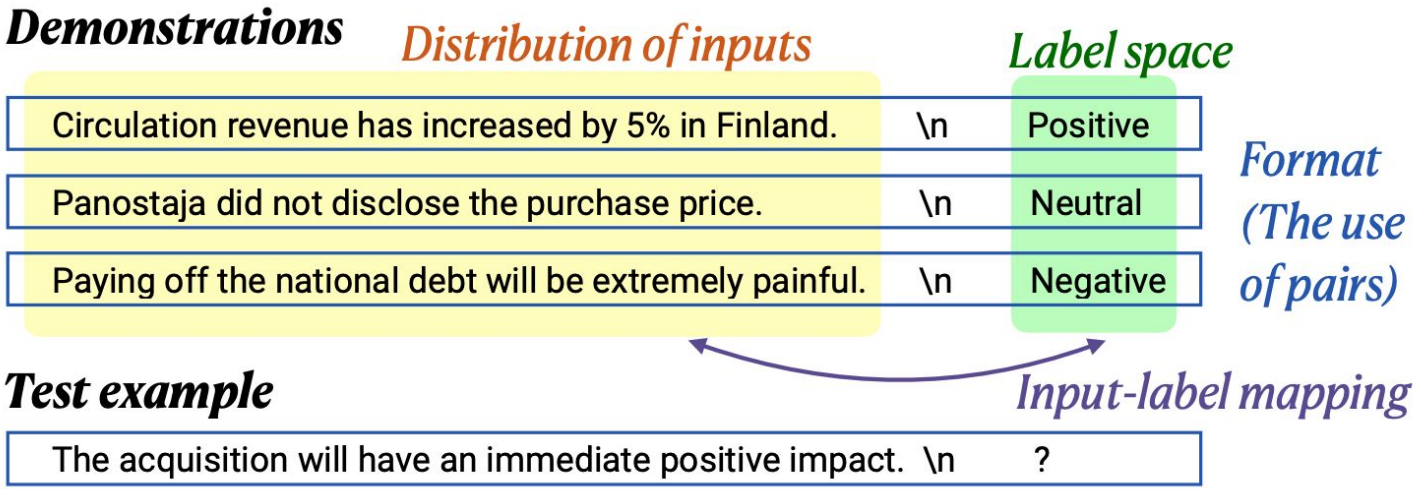


Figure 1: An example of instruction induction for the task of formality style transfer. *Left:* the standard in-context learning setting; given five demonstrations, complete the sixth. *Right:* instruction induction; the language model is prompted to generate a natural language instruction that describes the demonstrations. Model completions are in blue, prompt templates are in pink.

Explicit tell the language model what to do, thus can saving the demonstration examples → Instruction Tuning (Flan-T5 series)

The Role of Demonstration in ICL

Four Aspects of Demonstration



Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?, Min et al. EMNLP 2022

Experimental Setup

Models

Model	# Params	Public	Meta-trained
GPT-2 Large	774M	✓	✗
MetaICL	774M	✓	✓
GPT-J	6B	✓	✗
fairseq 6.7B†	6.7B	✓	✗
fairseq 13B†	13B	✓	✗
GPT-3	175B‡	✗	✗

Table 1: A list of LMs used in the experiments: GPT-2 (Radford et al., 2019), MetaICL (Min et al., 2021b), GPT-J (Wang and Komatsuzaki, 2021), fairseq LMs (Artetxe et al., 2021) and GPT-3 (Brown et al., 2020). 'Public' indicates whether the model weights are public; 'Meta-trained' indicates whether the model is meta-trained with an in-context learning objective. †We use dense models in Artetxe et al. (2021) and refer them as fairseq LMs for convenience. ‡We use the Davinci API (the *base* version, not the *instruct* version)

Datasets

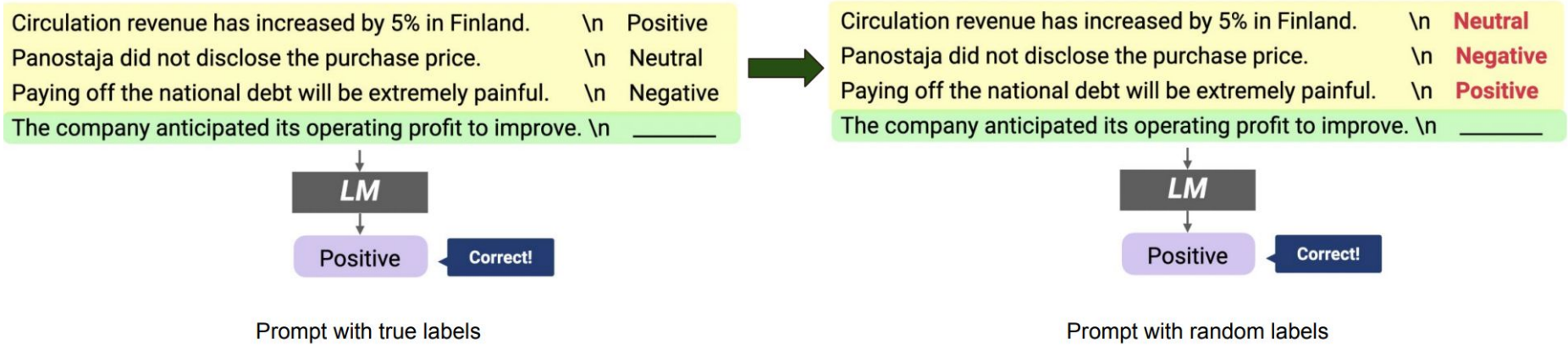
Dataset	# Train	# Eval
Task category: Sentiment analysis		
financial_phrasebank	1,811	453
poem_sentiment	892	105
Task category: Paraphrase detection		
medical_questions_pairs	2,438	610
glue-mrpc	3,668	408
Task category: Natural language inference		
glue-wnli	635	71
climate_fever	1,228	307
glue-rte	2,490	277
superglue-cb	250	56
sick	4,439	495
Task category: Hate speech detection		
hate_speech18	8,562	2,141
ethos-national_origin	346	87
ethos-race	346	87
ethos-religion	346	87
tweet_eval-hate	8,993	999
tweet_eval-stance_atheism	461	52
tweet_eval-stance_feminist	597	67
Task category: Question answering		
quarel	1,941	278
openbookqa	4,957	500
qasc	8,134	926
commonsense_qa	9,741	1,221
ai2_arc	1,119	299
Task category: Sentence completion		
codah	1665	556
superglue-copa	400	100
dream	6116	2040
quartz-with_knowledge	2696	384
quartz-no_knowledge	2696	384

k = 16 examples for demonstartion chosen randomly 5 times.

MacroF1 for classification tasks and Accuracy for multichoice tasks are reported,

A verage per-dataset over seeds, and then report macro-average over datasets.

RQ1: Does the input-label mapping matter?

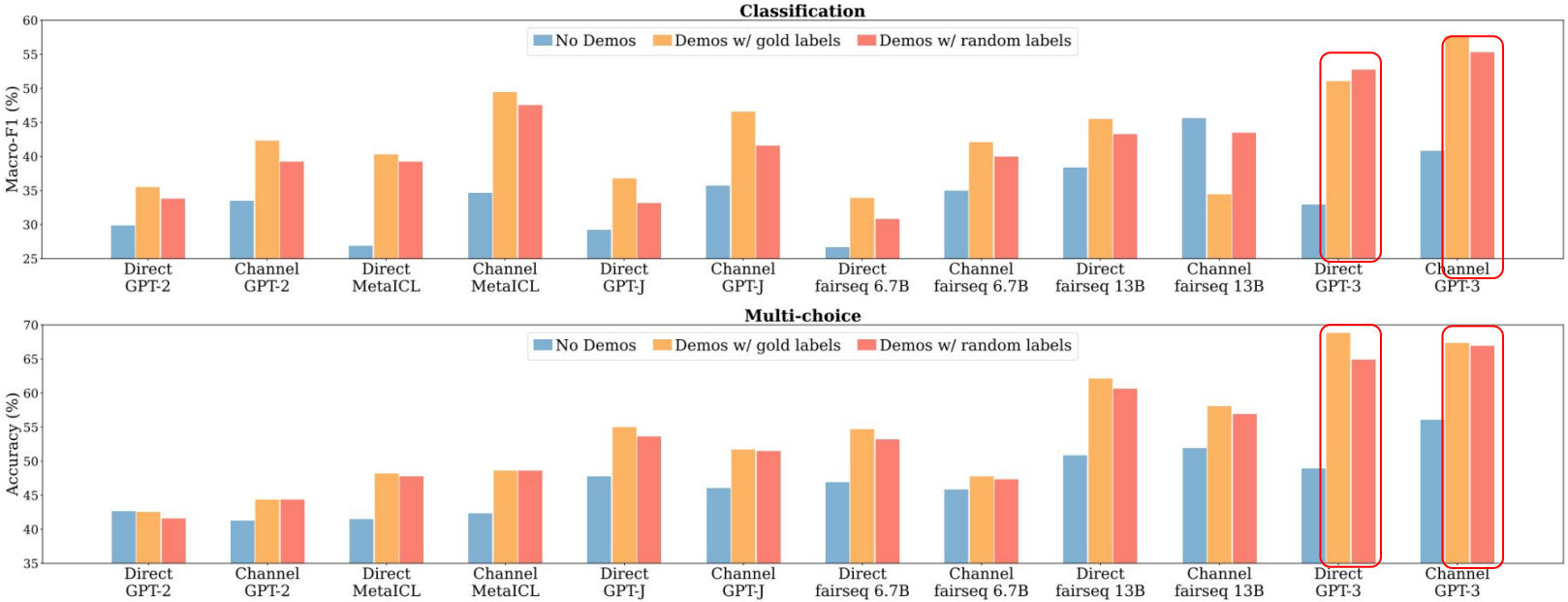


Randomly sample a label from the label space, and assign it to the demonstraion example.

Break the input-label mapping

Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?, Min et al. EMNLP 2022

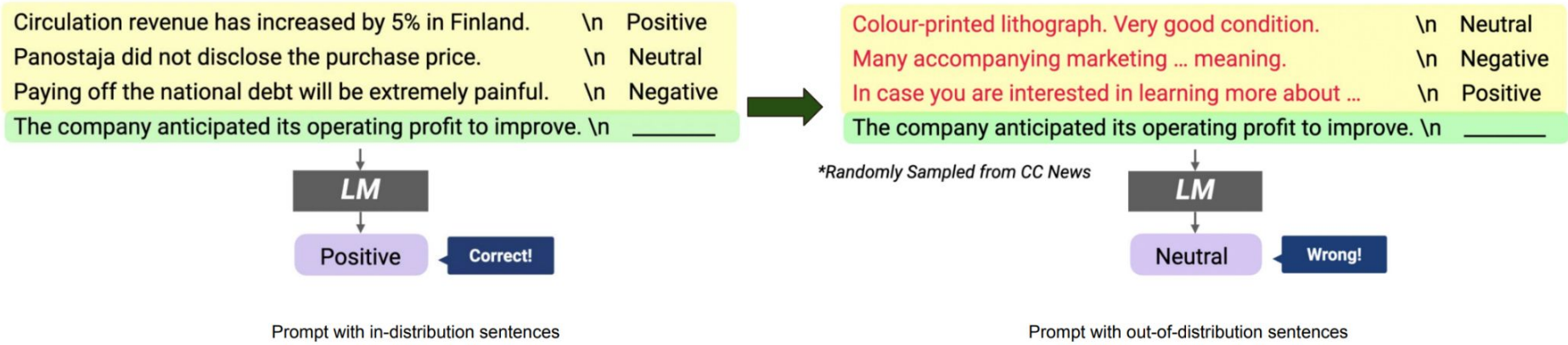
A1: Correct input-label mapping in the prompt is not as important as we thought



Model performance with random labels is very close to performance with gold labels, i.e., 0% - 5% absolute drop.

Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?, Min et al. EMNLP 2022

RQ2: Does the inputs distribution matter?

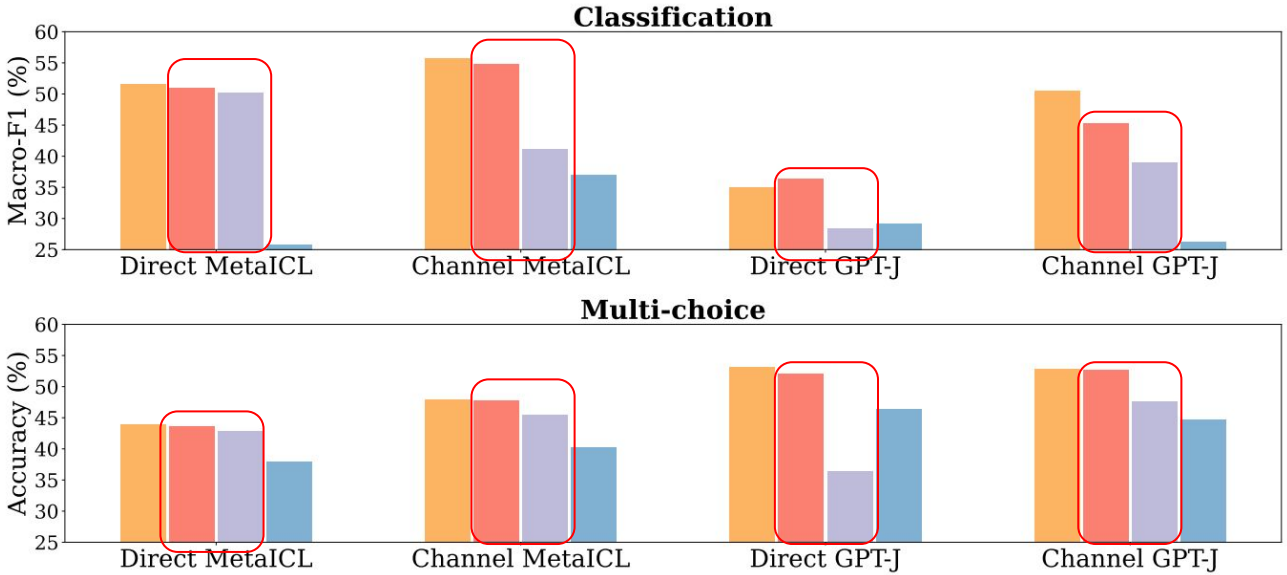


Inputs text are sampled from an external corpus, i.e., CC News.

Introducing a distributional gap between in-context demonstrations and the test example.

Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?, Min et al. EMNLP 2022

A2: Yes, significantly.



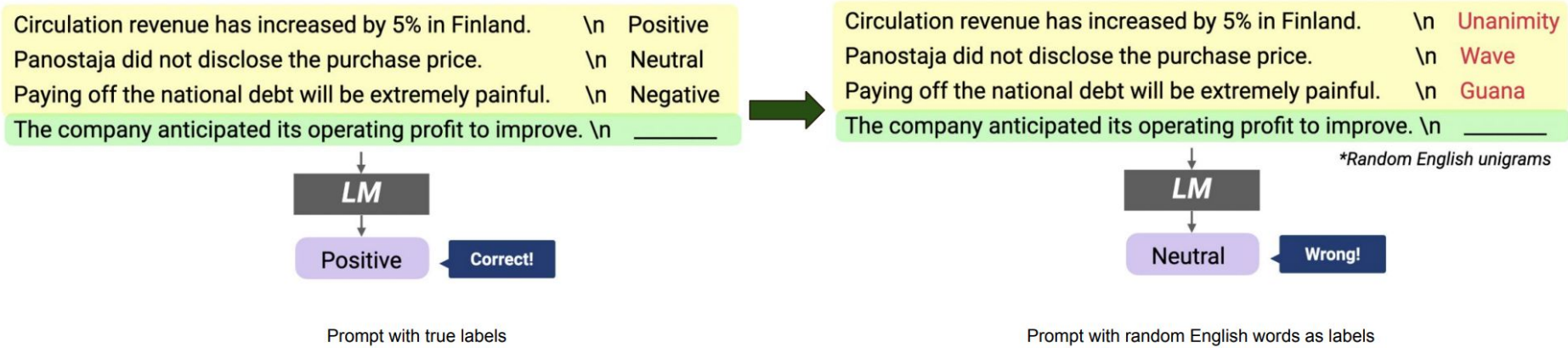
	F	*L*	*I*	*M*
Gold labels	✓	✓	✓	✓
Random labels	✓	✓	✓	✗
OOD + Random labels	✓	✓	✗	✗
No demonstrations	✗	✗	✗	✗

F: Format
L: Label space
I: Input distribution
M: Input-Label Mapping

Using out-of-distribution inputs instead of the inputs from the training data significantly drops the performance, when Channel MetaICL, Direct GPT-J or Channel GPT-J are used, both in classification and multichoice, by 3–16% in absolute.

Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?, Min et al. EMNLP 2022

RQ3: Does the label space matter?

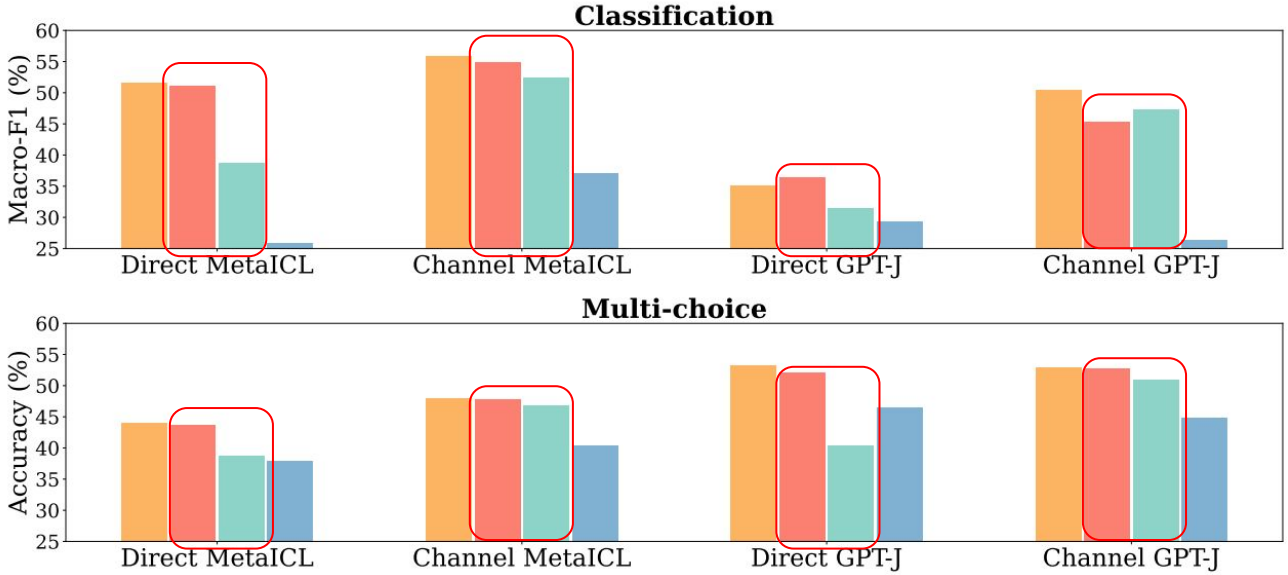


Labels are replaced by random English unigrams sampled.

Introducing a distributional gap between in-context demonstrations and the test example.

Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?, Min et al. EMNLP 2022

A3: Seeing correct label space is important.



	F	L	I	M
Gold labels	✓	✓	✓	✓
Random labels	✓	✓	✓	✗
Random English words	✓	✗	✓	✗
No demonstrations	✗	✗	✗	✗

F: Format
L: Label space
I: Input distribution
M: Input-Label Mapping

For Direct Models: performance decreases of up to 16% absolute.
For Channel Models: performance decreases of up to 2% absolute.

Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?, Min et al. EMNLP 2022

RQ4: Does the format matter?

Format refers to the usage of the input-label pair.

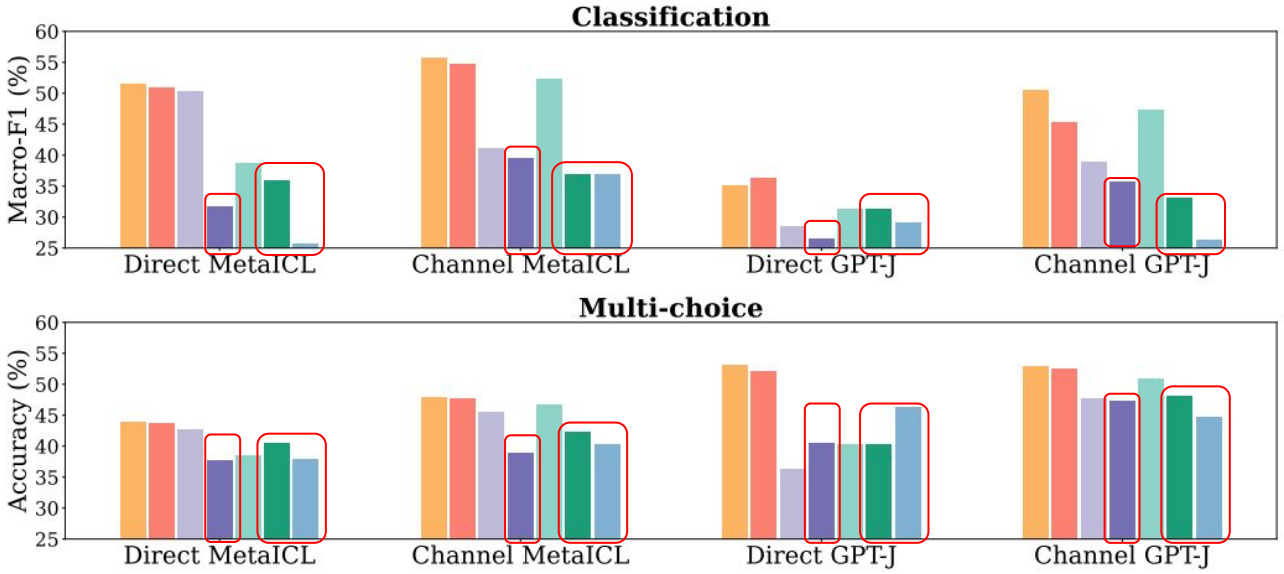
Demos w/o labels	(*Format* ✗ *Input distribution* ✓ *Label space* ✗ *Input-label mapping* ✗) Circulation revenue has increased by 5% in Finland and 4% in Sweden in 2008. Panostaja did not disclose the purchase price.
Demos labels only	(*Format* ✗ *Input distribution* ✗ *Label space* ✓ *Input-label mapping* ✗) positive neutral

Feed in examples with no labels and with labels only

This is somewhat non-sense as the LMs do not know what to do without specified input-output.

Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?, Min et al. EMNLP 2022

A4: Removing format ~= No Demonstration



	F	L	I	M
Gold labels	✓	✓	✓	✓
Random labels	✓	✓	✓	✗
OOD + Random labels	✓	✓	✗	✗
Random labels only	✗	✓	✗	✗
Random English words	✓	✗	✓	✗
No labels	✗	✗	✓	✗
No demonstrations	✗	✗	✗	✗

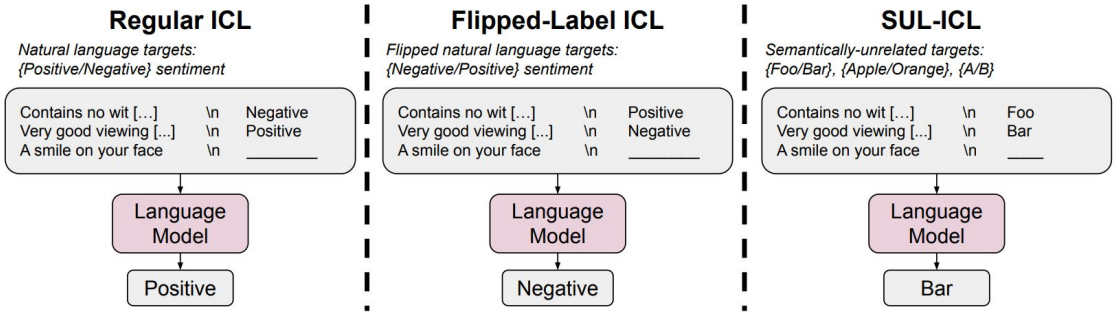
F: Format
L: Label space
I: Input distribution
M: Input-Label Mapping

Not using the input-label format decreases performance.

Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?, Min et al. EMNLP 2022

Take away for Demonstration Role in ICL

- ICL still works well when outputs in the prompt are replaced with random outputs.
  - *Recent study suggests that large language models can even perform flipped-label ICL.*



- gains are mainly coming from independent specification of the input space and the label space

Larger language models do in-context learning differently, Wei et al. 2023

Beyond input-label pair format

There are challenging reasoning tasks.

Q: If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

A: The answer is **5**

Arithmetic Reasoning (AR)
(+ − ×÷...)

Q: Take the last letters of the words in "Elon Musk" and concatenate them

A: The answer is **nk.**

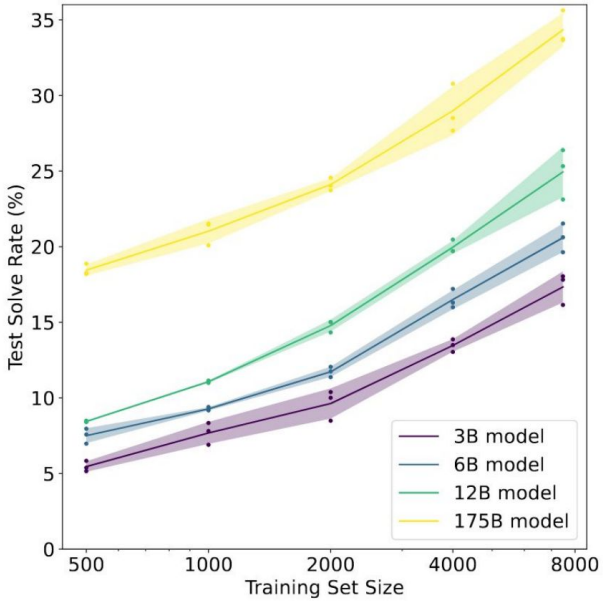
Symbolic Reasoning (SR)

Q: What home entertainment equipment requires cable? Answer Choices: (a) radio shack (b) substation (c) television (d) cabinet

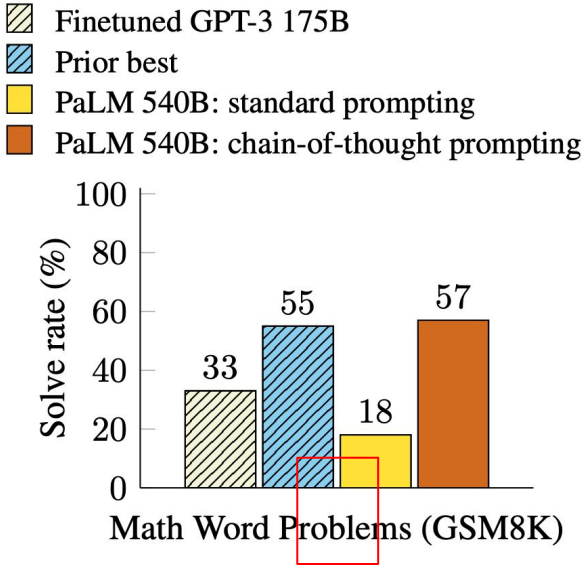
A: The answer is **(c).**

Commonsense Reasoning (CR)

Fine-tuning & ICL performance on GSM8K (arithmetic)



Results of fine-tune GPT-3 175B

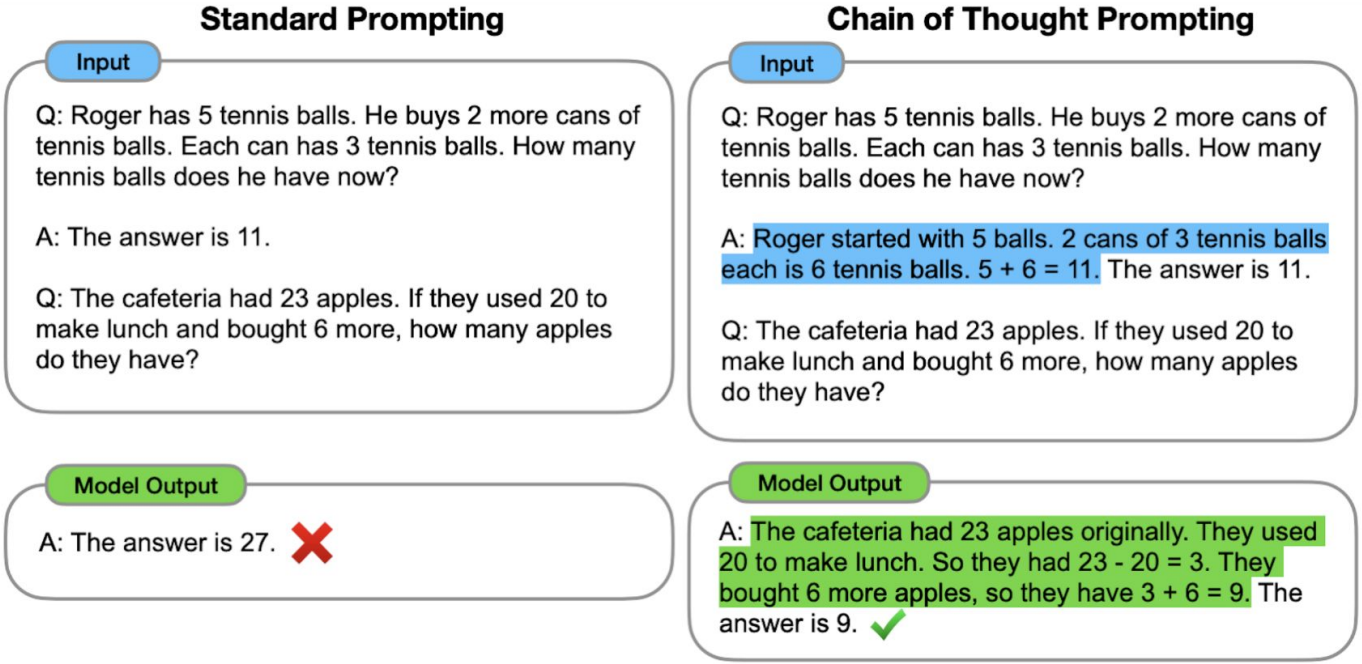


ICL with PaLM540B

Both Struggle! -> Scaling model size does not help for reasoning tasks.

Is there something wrong?

Possible Solution: Chain-of-Thought (CoT) Prompting

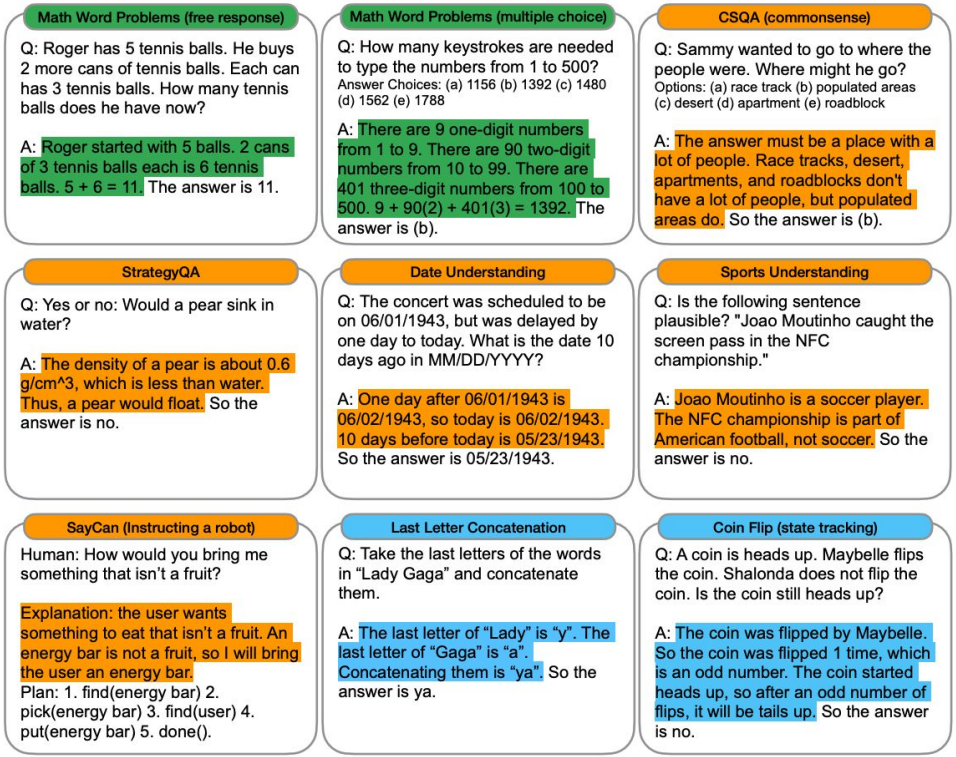


Definition:
A chain of thought is a series of intermediate natural language reasoning steps that lead to the final output.

<Input, Output> → <Input, Chain-of-Thought, Output>

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, Wei et al. NeurIPS 2022

Few-shot CoT

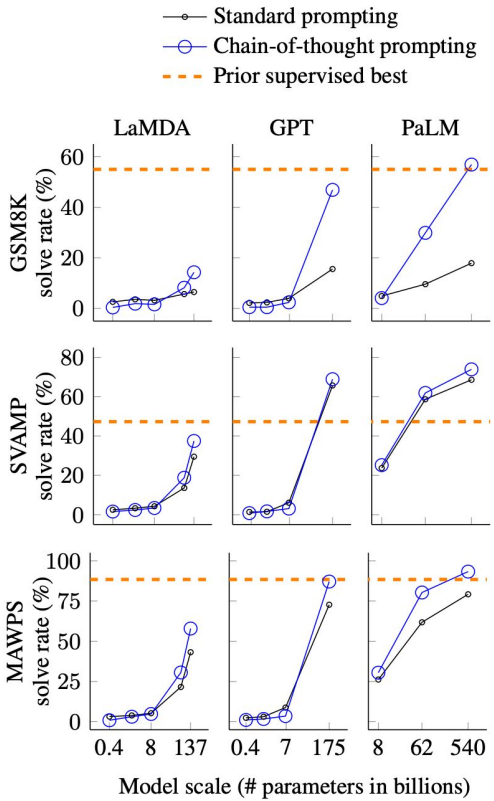


Manually crafted 8 Chain-of-Thought examples for various datasets.

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, Wei et al. NeurIPS 2022

Results

- Evaluated LMs:
  - Instruct GPT-3 (ada 350M, babbage 1.3B, curie 6.7B, and davinci 175B)
    (Ouyang et al. 2022)
  - PaLM (8B, 62B, 540B, google only)
    (Chowdhery et al., 2022)
  - LaMDA (422M, 2B, 8B, 68B, 137B)
    (Thoppilan et al., 2022)

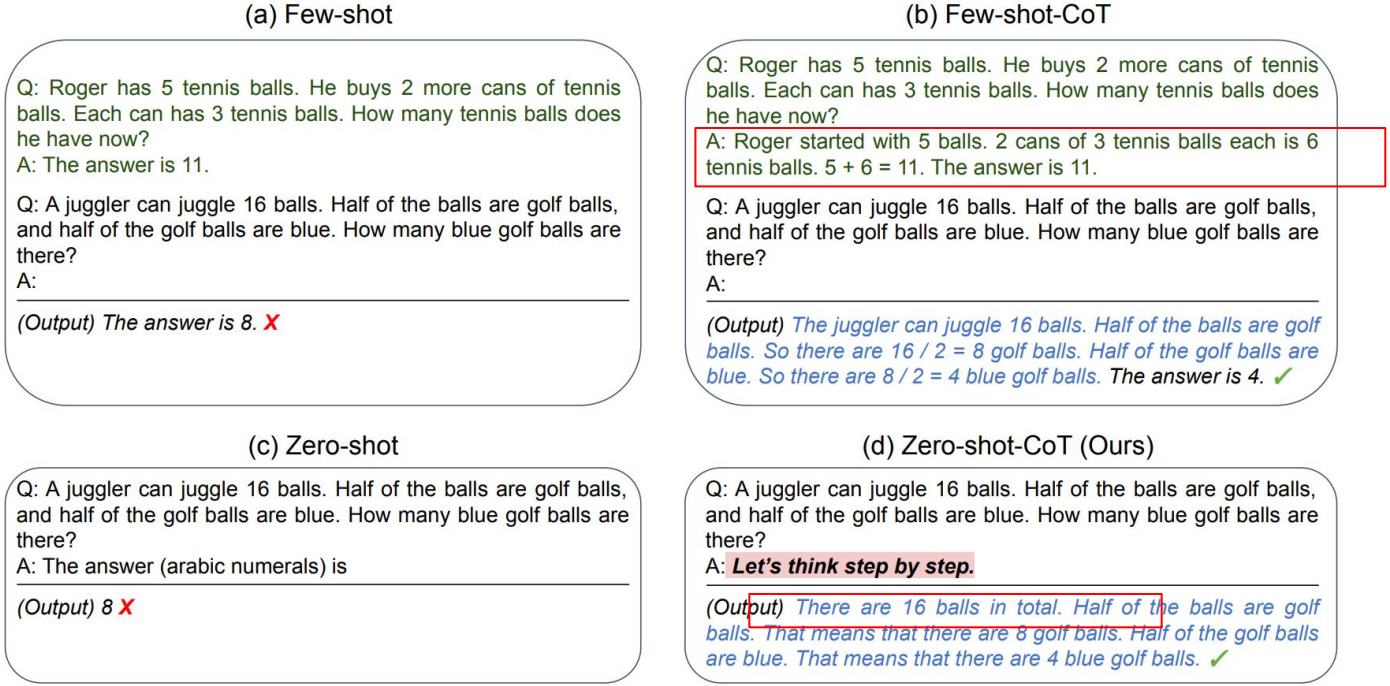


Key Findings:

- CoT only emerges with large language mosels (> 100B ?)
- Improvement is more pronounced in complex tasks.
- CoT even achieves comparable results with supervised methods.

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, Wei et al. NeurIPS 2022

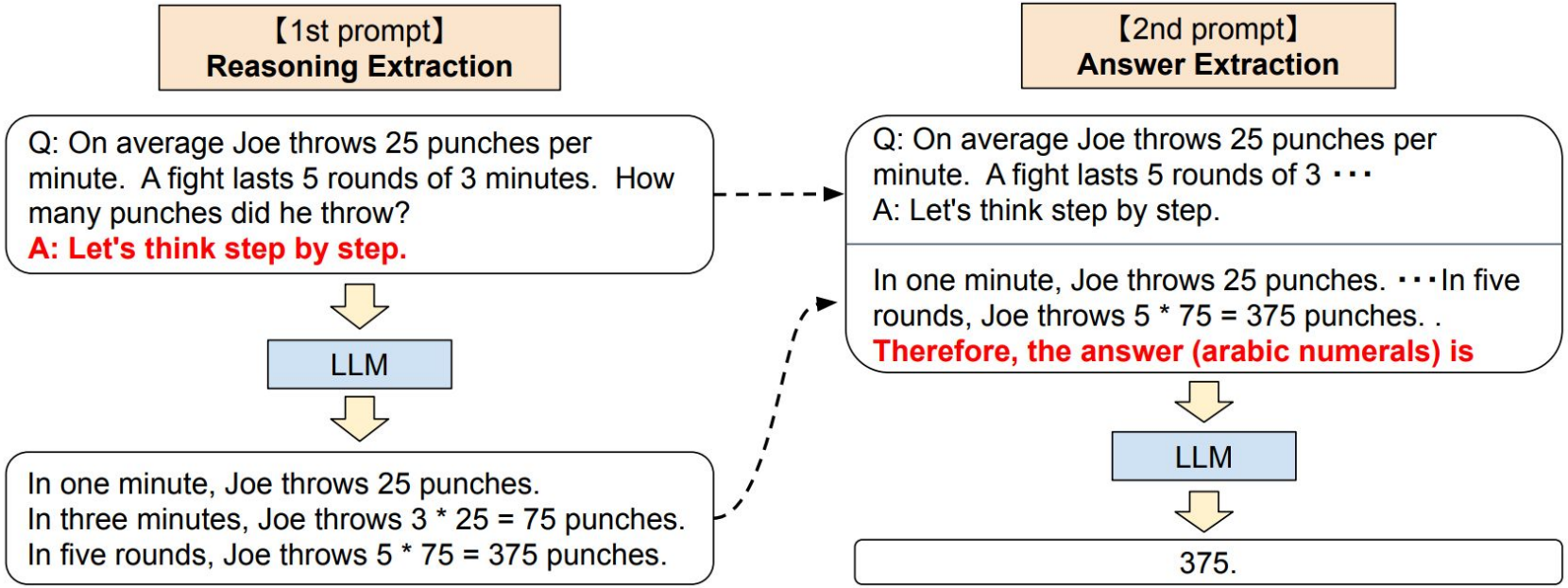
Zero-shot CoT Prompting



No demonstrated CoT anymore!
Just prompting LLMs to think step by step.

Large Language Models are Zero-Shot Reasoners, Kojima et al. NeurIPS 2022

Zero-shot CoT Prompting



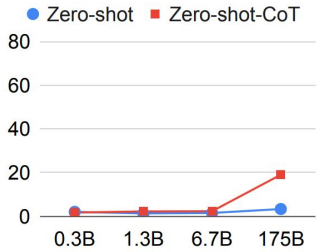
Generate Rationales first, then incorporate the generated CoT for eliciting the answer.

Large Language Models are Zero-Shot Reasoners, Kojima et al. NeurIPS 2022

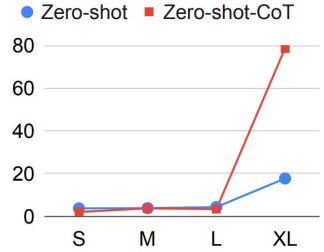
Results of Zero-Shot CoT

A voting mechanism for integrating answers of multiple sampled reasoning path.

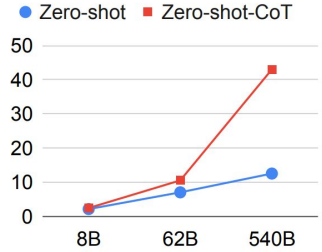
	MultiArith	GSM8K
PaLM 540B: Zero-Shot	**25.5**	**12.5**
PaLM 540B: Zero-Shot-CoT	**66.1**	**43.0**
PaLM 540B: Zero-Shot-CoT + self consistency	**89.0**	**70.1**
PaLM 540B: Few-Shot [Wei et al., 2022]	-	17.9
PaLM 540B: Few-Shot-CoT [Wei et al., 2022]	-	56.9
PaLM 540B: Few-Shot-CoT + self consistency [Wang et al., 2022]	-	74.4



(a) MultiArith on Original GPT-3



(b) MultiArith on Instruct GPT-3



(c) GMS8K on PaLM

Key Findings:

- Zero-shot CoT is effective, significantly boost the performance of zero-shot prompting.
- Also, Zero-shot CoT is an emergent ability of LLMs.

Large Language Models are Zero-Shot Reasoners, Kojima et al. NeurIPS 2022

Take away for CoT prompting

- A promising solution for challenging reasoning tasks:
  - <input, output> -> <input, intermediate steps, output>

- CoT emerges with model scale.

- LLMs can generate high-quality rationales with magic prompting words.

Personal Experience: How to Prompt GPT-series Better

- **1. Use the latest model, GPT-4 if possible.**
- **2. Be specific.**

  Write a short poem.

  v.s.

  Write a short inspiring poem about OpenAI, focusing on the recent DALL-E product launch (DALL-E is a text to image ML model) in the style of a {famous poet}

- **3. Use ICL to specify the desired outputs.**
- **4. Instead of just saying what not to do, say what to do instead**

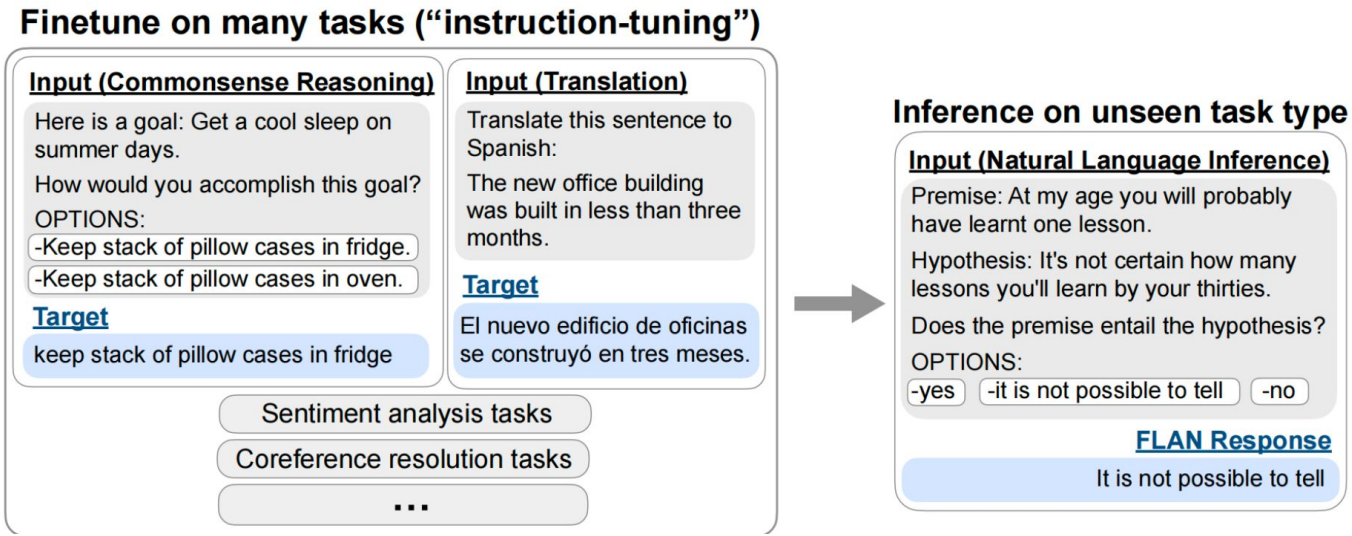
[Best practices for prompt engineering with OpenAI API]

Recent trends in ICL

- Better selection techniques & retriever.
  - *Self-Adaptive In-Context Learning: An Information Compression Perspective for In-Context Example Selection and Ordering, Wu et al. ACL 2023*
  - *Finding Supporting Examples for In-Context Learning, Li et al. 2023*
  - *Compositional Exemplars for In-context Learning, Ye et al. ICML 2023*
  - *Unified Demonstration Retriever for In-Context Learning, Li et al. ACL 2023*
  - ....

Recent trends in ICL

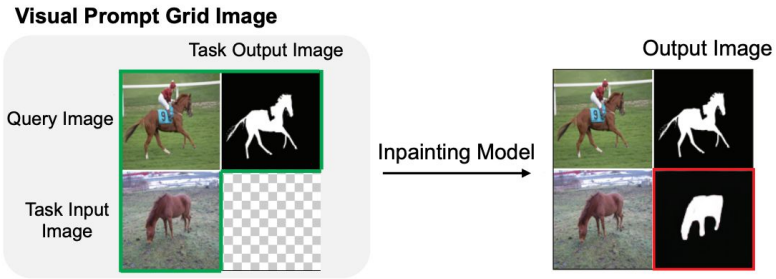
- Switch to Instruction Tuning gradually.

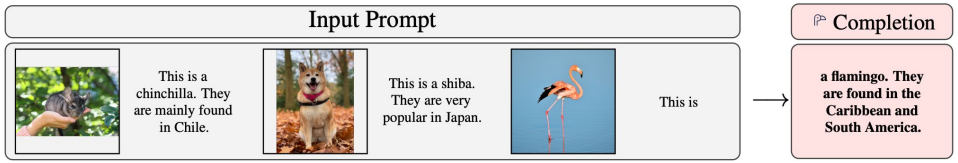


SFT can be seen as a generalized instruction tuning.

Recent trends in ICL

- Expand to other modality (e.g., vision & vision + language)





- Interpretability of In-context Learning & CoT
  - *Why Can GPT Learn In-Context? Language Models Secretly Perform Gradient Descent as Meta-Optimizers, Dai et al. Findings of ACL 2023*
  - *In-context Learning and Induction Heads, Olsson et al. 2022*
  - *Label Words are Anchors: An Information Flow Perspective for Understanding In-Context Learning, Wang et al. 2023*

Useful Resources

A Survey on In-context Learning

Qingxiu Dong[1], Lei Li[1], Damai Dai[1], Ce Zheng[1], Zhiyong Wu[2], Baobao Chang[1], Xu Sun[1], Jingjing Xu[2], Lei Li[3] and Zhifang Sui[1]
[1] MOE Key Lab of Computational Linguistics, School of Computer Science, Peking University
[2] Shanghai AI Lab [3] University of California, Santa Barbara

Survey Paper

Paper List

Openicl

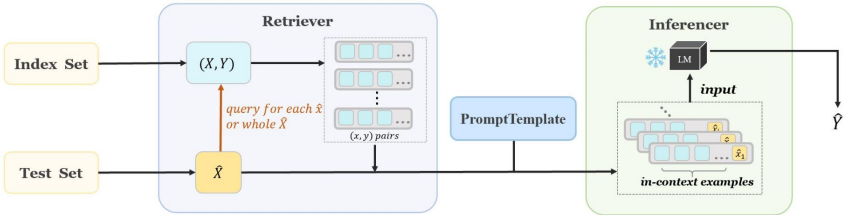
An Open-Source Framework for In-context Learning

Overview • Installation • Paper • Examples • Docs • Citation

version 0.1.8

Overview

OpenICL provides an easy interface for in-context learning, with many state-of-the-art retrieval and inference methods built in to facilitate systematic comparison of LMs and fast research prototyping. Users can easily incorporate different retrieval and inference methods, as well as different prompt instructions into their workflow.



https://github.com/Shark-NLP/OpenICL

Questions

- Q1: Why does in-context learning only appear in large language models?
- Q2: Does CoT prompt reasoning? or just more computation used?
- Q3: What application would ICL & CoT enable in other areas?