Towards Understanding In-context Learning

Sam Liang, Kexin Jin

Princeton University

What is In-Context Learning?

Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. // _____

LM

Circulation revenue has increased by 5% in Finland. // Finance

They defeated … in the NFC Championship Game. // Sports

Apple … development of in-house chips. // Tech

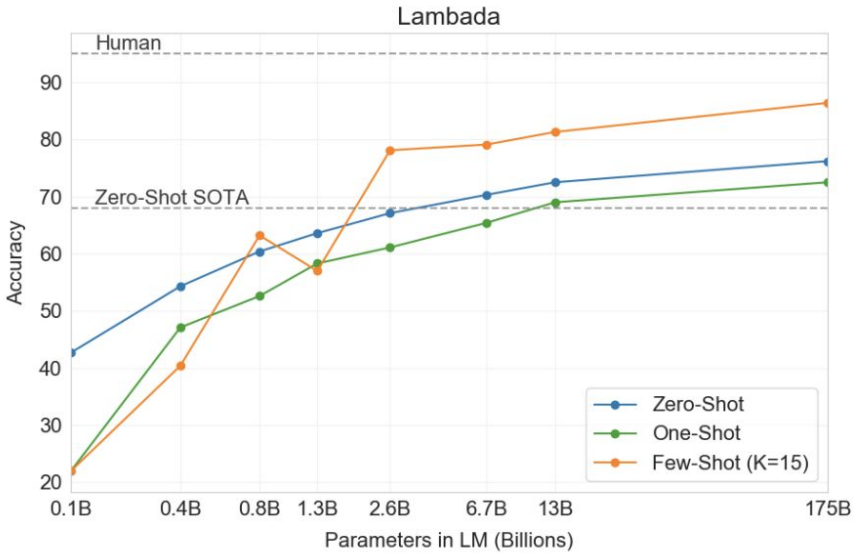
The company anticipated its operating profit to improve. // _____

LM

https://ai.stanford.edu/blog/understanding-incontext/

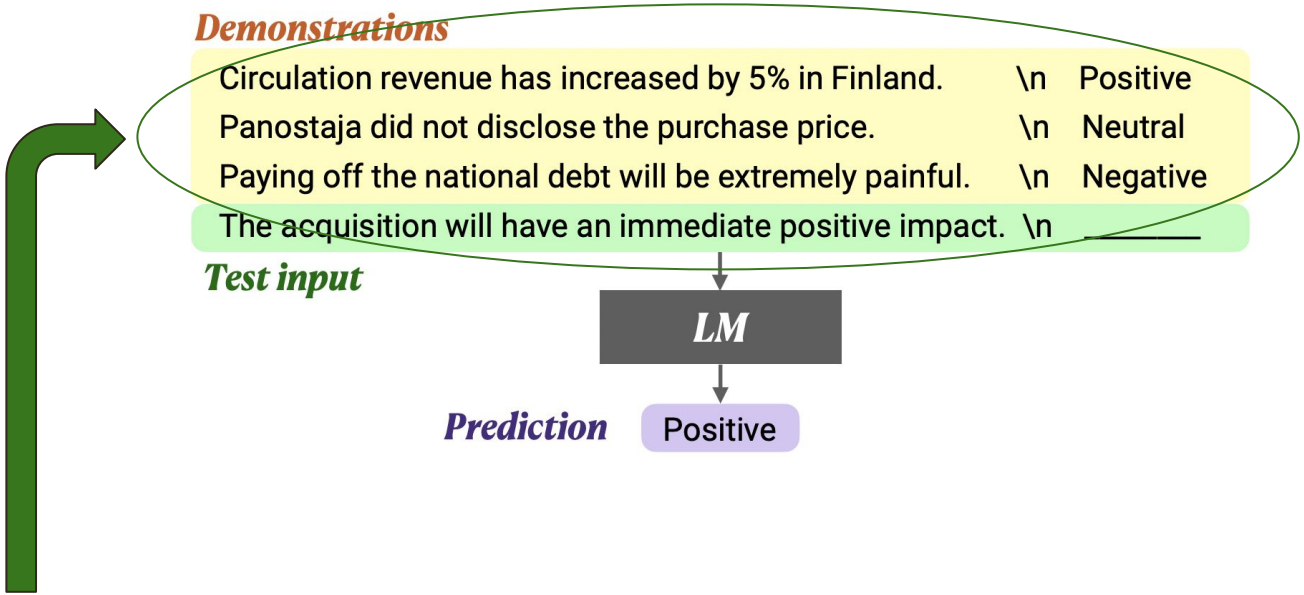
What Can In-Context Learning Do?

- No parameter tuning need
- Only need few examples for downstream tasks
- GPT-3 improved SOTA on LAMBADA by 18%!



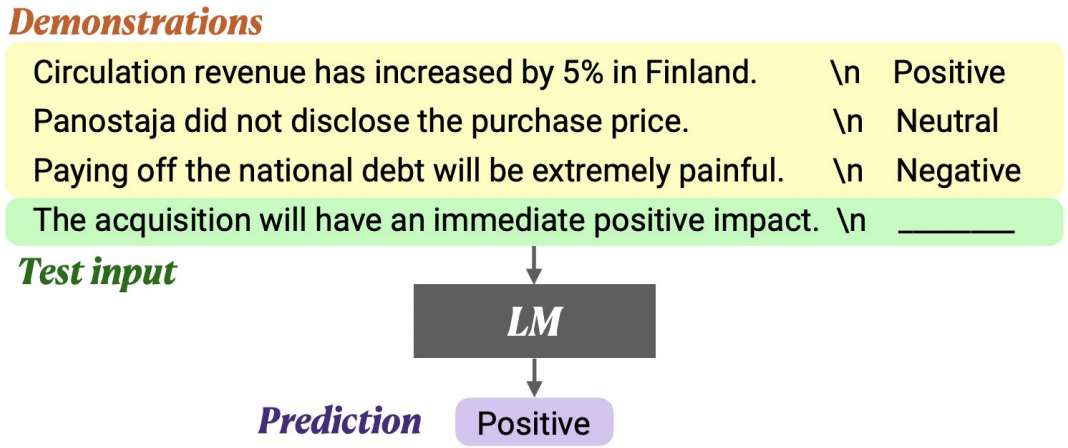
Works like magic!

We don’t know how models in-context learn



Learns to do a downstream task by conditioning on input-output examples

We don’t know how models in-context learn



No weight update and model is **not explicitly pre-trained to learn from examples**

How does it know what to do then?

Research Goals

Develop a mathematical framework for understanding how in-context learning emerges during pre-training

(Xie et al., 2022): An Explanation of In-context Learning as Implicit Bayesian Inference

Analyze empirically which aspects of the prompt affect downstream task performance

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

Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. // _____

Positive

Circulation revenue has increased by 5% in Finland. // Finance

They defeated … in the NFC Championship Game. // Sports

Apple … development of in-house chips. // Tech

The company anticipated its operating profit to improve. // _____

Finance

Model needs to figure out:

input distribution (financial or general news)

output distribution (Positive/Negative or topic)

input-output mapping (sentiment or topic classification)

formatting

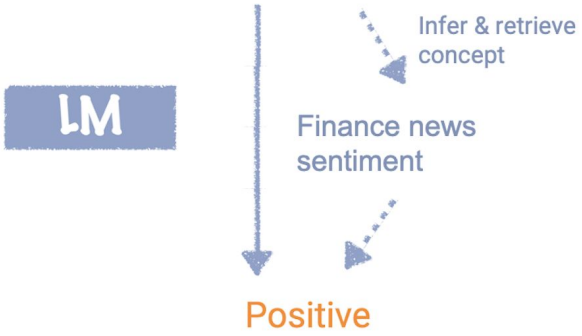
Concepts (long-term coherence)

Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. // _____

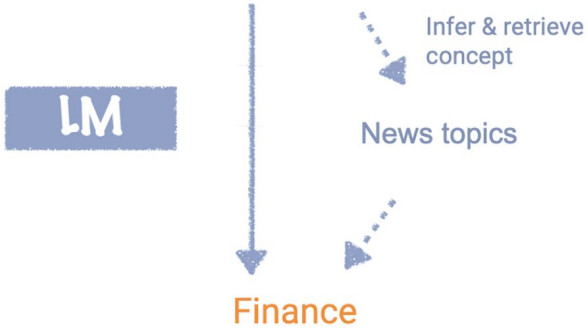


Circulation revenue has increased by 5% in Finland. // Finance

They defeated … in the NFC Championship Game. // Sports

Apple … development of in-house chips. // Tech

The company anticipated its operating profit to improve. // _____



A latent variable that contains various **document-level statistics**: a distribution of words, a format, a relation between sentences, and other semantic and syntactic relations in general.

An Explanation of In-context Learning as Implicit Bayesian Inference, Xie et al.

Hypothesis

Language model (LM) uses the in-context learning prompt to “locate” a previously learned concept to do the in-context learning task

Bayesian inference!

How does the LM learn to do Bayesian inference?

- **Pretrain**: To predict the next token during pre-training, the LM must infer the latent concept for the document using evidence from the previous sentences.

- **In-context learning**: If the LM also infers the prompt concept using demonstrations in the prompt, then in-context learning succeeds!

Mathematically…

Mixture of Hidden Markov Models (HMM)

Pretraining distribution

- Each pretraining document is a length T sequence sampled by

$$p(o_1, ..., o_T) = \int_{\theta \in \Theta} p(o_1, ..., o_T|\theta) p(\theta) d\theta,$$

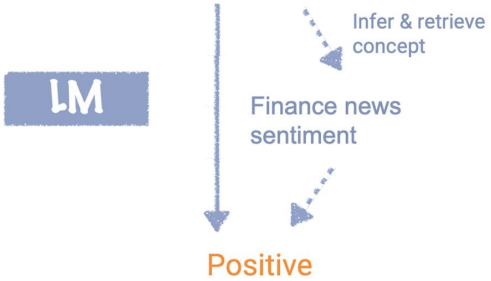
  where Θ is a family of concepts that defines a distribution over observed tokens o from a vocabulary $\mathcal{O}$.
- Assumption: $p(o_1, ..., o_T|\theta)$ is defined by a Hidden Markov Model (HMM). The concept θ determines the transition probability matrix of the HMM hidden states $h_1, ..., h_T$ from a hidden state set $\mathcal{H}$.

Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. // _____

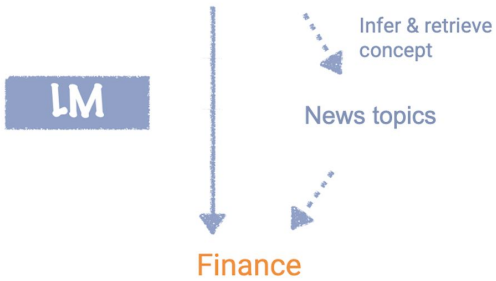


Circulation revenue has increased by 5% in Finland. // Finance

They defeated … in the NFC Championship Game. // Sports

Apple … development of in-house chips. // Tech

The company anticipated its operating profit to improve. // _____



If the pretraining data is a mixed of finance news sentiment task and news topics task, intuitively, we could say there are two concepts θ1 and θ2.

p(Paying off the national debt will be extremely painful) =

½ p(Paying off the national debt will be extremely painful | θ1)

+

½ p(Paying off the national debt will be extremely painful | θ2)

In-context learning

- Given $\theta^* \in \Theta$, the prompt is a concatenation of n independent demonstrations and 1 test input x_{test} that are all conditioned on θ^*
- The goal is to predict the test output y_{test} by predicting the next token.

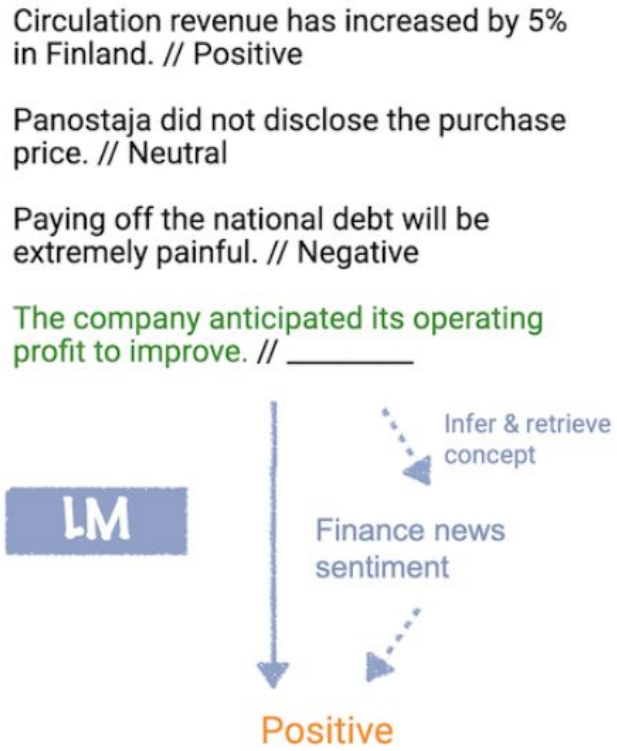
Prompt distribution

- For $i = 1, ..., n$, the i-th demonstration $O_i = [x_i, y_i]$, where x_i is an input token sequence and y_i is an output token.
- Each O_i is independently generated using

$$p(O_i | h_i^{start}, \theta^*),$$

  i.e. the pretraining distribution conditioned on a prompt concept θ^*.
- The prompt is a sequence of demonstrations S_n followed by the test example x_{test}:

$$\begin{aligned}[S_n, x_{test}] &= [O_1, o^{delim}, O_2, o^{delim}, ..., o^{delim}, O_n, x_{test}] \\ &= [x_1, y_1, o^{delim}, x_2, y_2, o^{delim}, ..., x_n, y_n, o^{delim}, x_{test}]\end{aligned}$$



[Circulation revenue has increased by 5% in Finland., **Positive**,

#,

Panostaja did not disclose the purchase price., **Neutral**,

#,

Paying off the national debt will be extremely painful., **Negative**,

#,

The company anticipated its operating profit to improve.]

Theorem (Xie et al., 2021)

Under some assumptions, as $n \to \infty$,

$$\arg\max_y p(y|S_n, x_{test}) \to \arg\max_y p_{prompt}(y|x_{test}).$$

- $p_{prompt} \sim p(\cdot|\theta^*)$
- The in-context predictor asymptotically achieves the optimal expected error
- More examples $\to$ More signals for Bayesian inference $\to$ Smaller error

Heuristic derivation

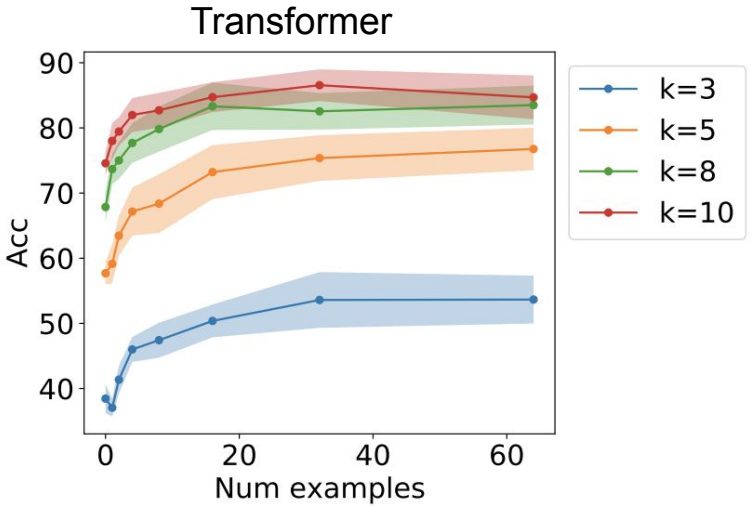
Heuristic derivation

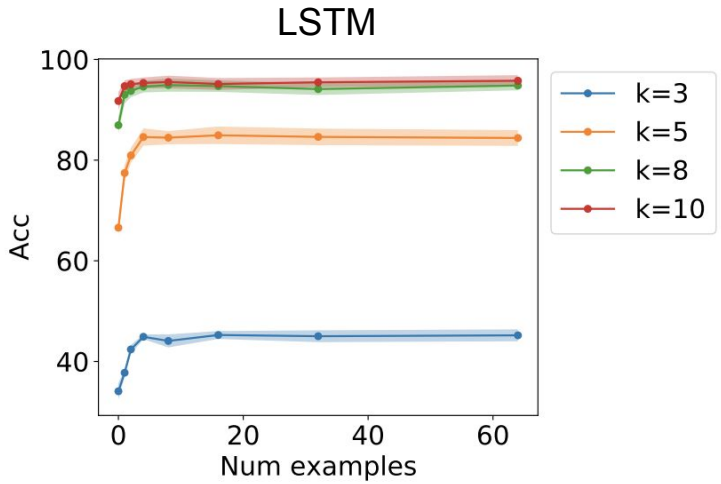
Heuristic derivation

Heuristic derivation

GINC

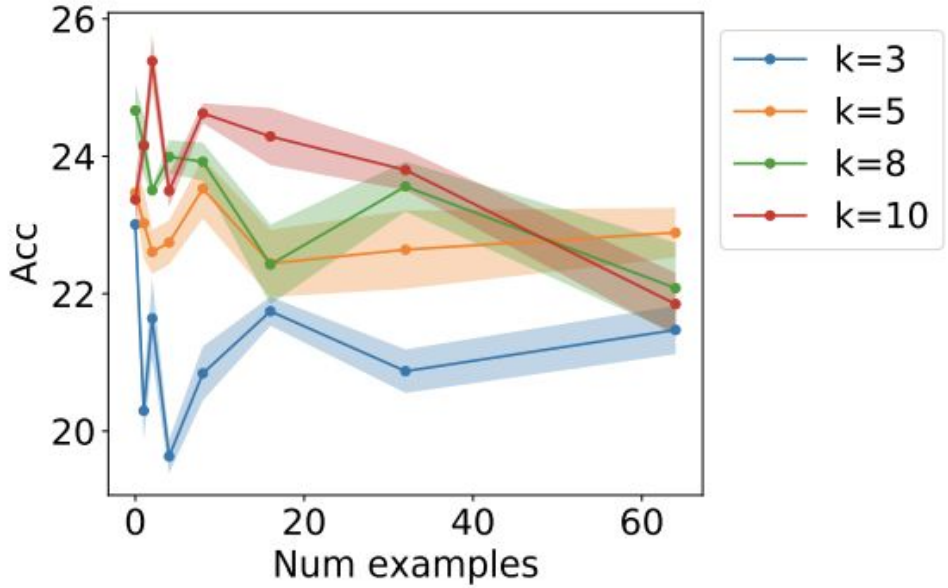
- A synthetic pretraining dataset and in-context learning testbed with the latent concept structure.
- **Pre-training:** a uniform mixture of HMMs over a family of 5 concepts, 1000 pre-training documents, ~10 million tokens in total
- **Prompts:** 0~64 training examples, example length k=3, 5, 8, 10
- GPT-2-based Transformers and LSTMs
- Vocabulary size: 50, 100, 150





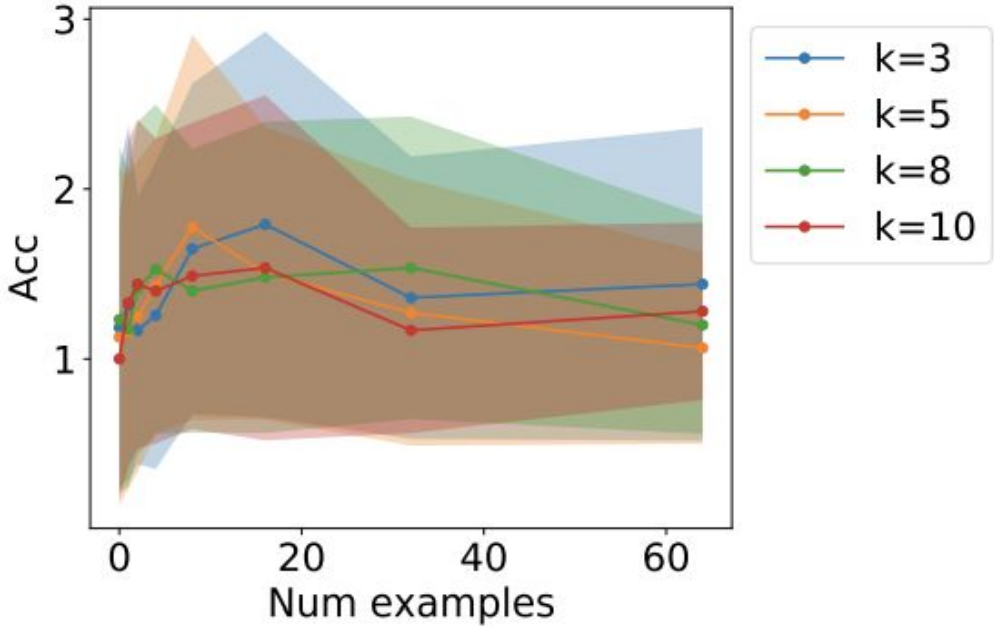
Accuracy increases with number of examples n and length of each example k, which is consistent with the theoretical results.

Is the HMMs assumption necessary?



When pretrained with only one concept, in-context learning fails.

Is the HMMs assumption necessary?



When the pretraining data has random transitions, in-context learning fails.

Is the HMMs assumption necessary?

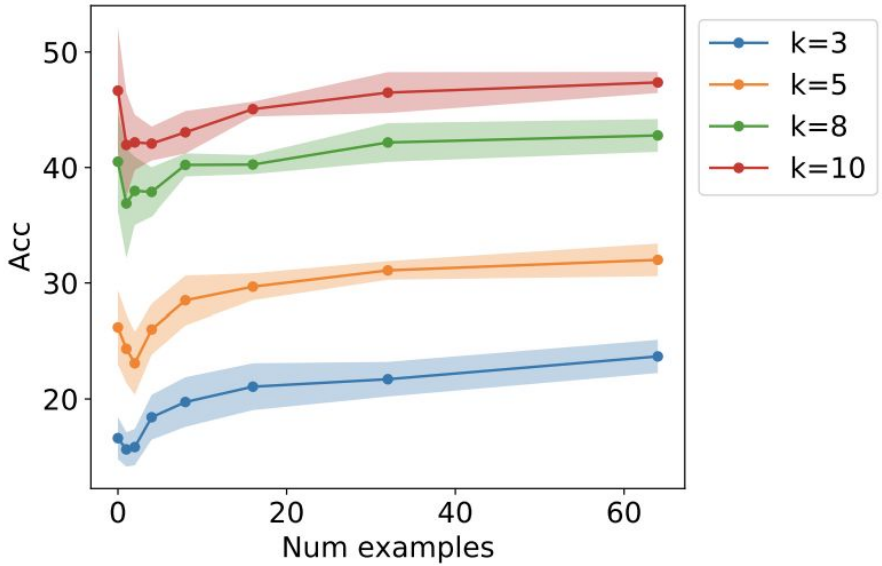
Yes, the mixture-of-concepts structure is important!

Q2. What assumptions do Xie et al., 2021 make about the pre-training distribution? Do you think this captures the characteristics of pre-training corpora used for LLMs?

1) A document is generated by first sampling a latent concept, and then the document is generated by conditioning on the latent concept, i.e. the pre-training distribution is a mixture of HMMs. They also assume that the pretraining data and LM are large enough that the LM fits the pretraining distribution exactly.

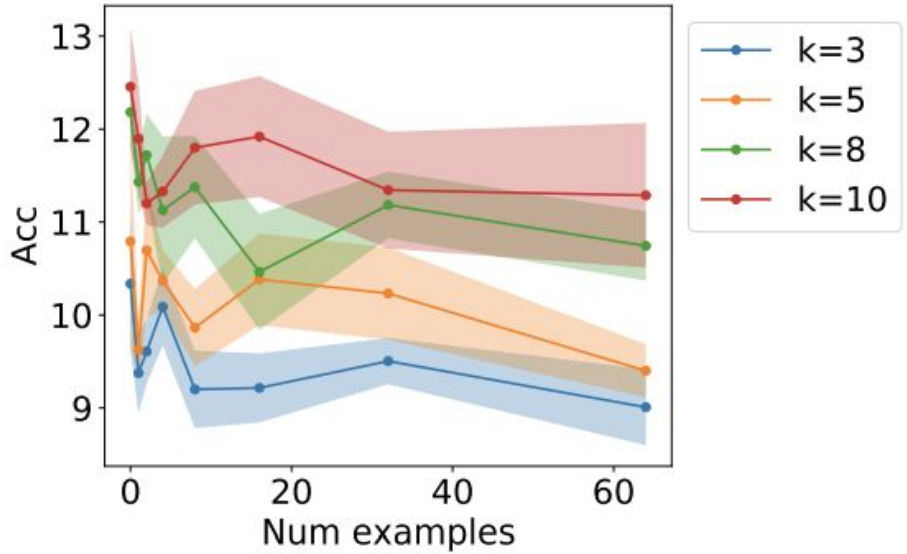
2) To some extent, yes. HMM is often used to model the distribution of languages. By introducing the notion of concepts and identify the transition probability for each HMM using concepts, they model the scenario where the data is from different HMMs, e.g. a mixture of different languages. Their ablation experiments verify this assumption. The second assumption is approximately true, and a possible future work can study the case when there is a gap between the pre-trained LM and the true pretraining distribution.

Zero-shot vs One-shot



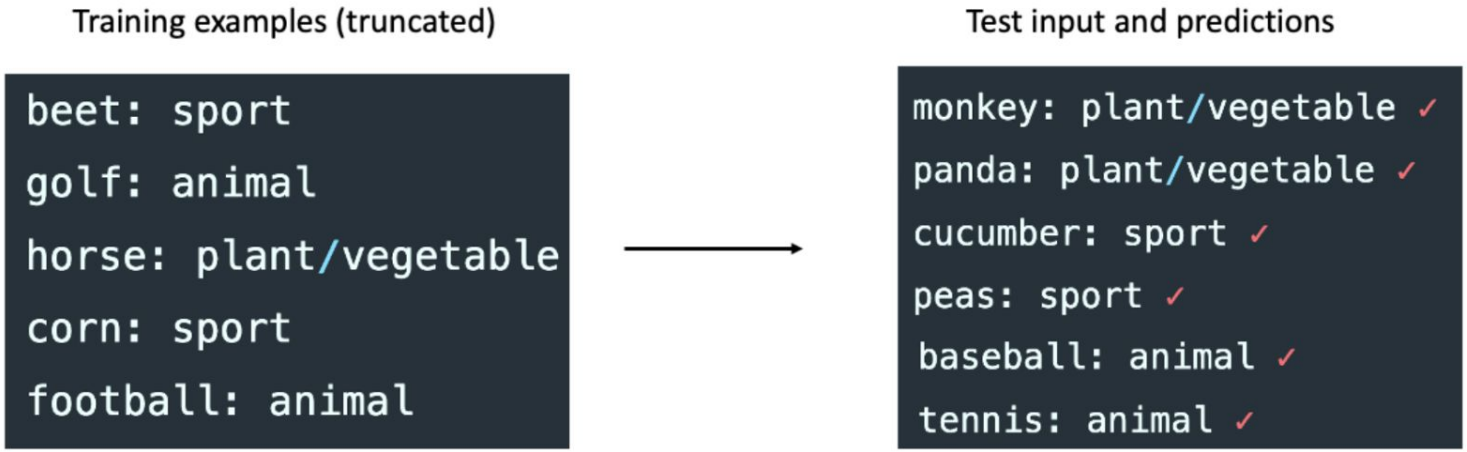
- In some settings, few-shot accuracy is initially worse than zero-shot accuracy, but can recover with more examples.
- Mirroring the behavior of GPT-3 on some datasets such as LAMBADA, HellaSwag, PhysicalQA, RACE-m, CoQA/SAT
- Especially because the transition probabilities in GINC are **lower entropy**

Unseen Concepts



When prompts are from random unseen concepts, in-context learning fails to extrapolate.

However…



An example synthetic task with unusual semantics that GPT-3 can successfully learn. A modified figure from Rong.

https://ai.stanford.edu/blog/in-context-learning/

Research Goals

Develop a mathematical framework for understanding how in-context learning emerges during pre-training

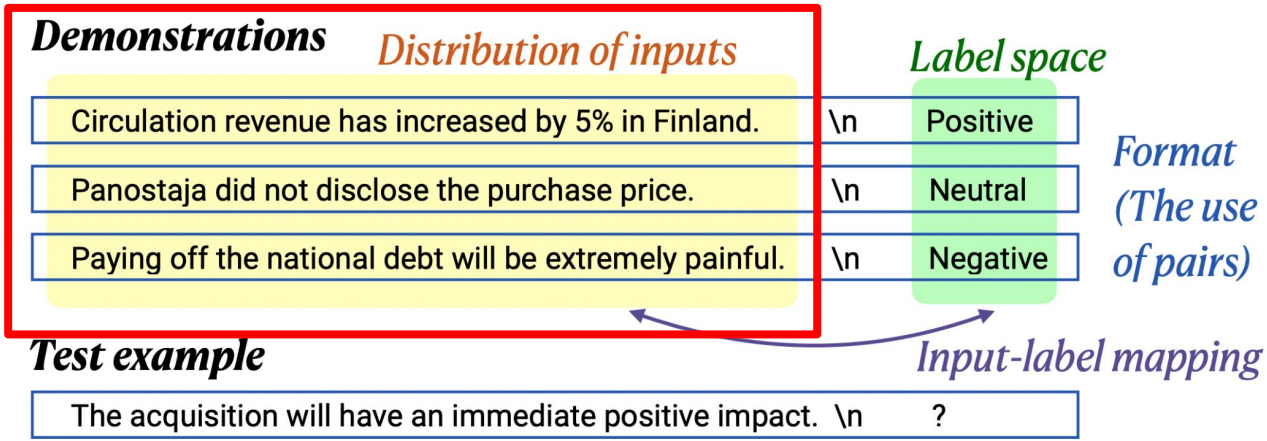
(Xie et al., 2022): An Explanation of In-context Learning as Implicit Bayesian Inference

Analyze empirically which aspects of the prompt affect downstream task performance

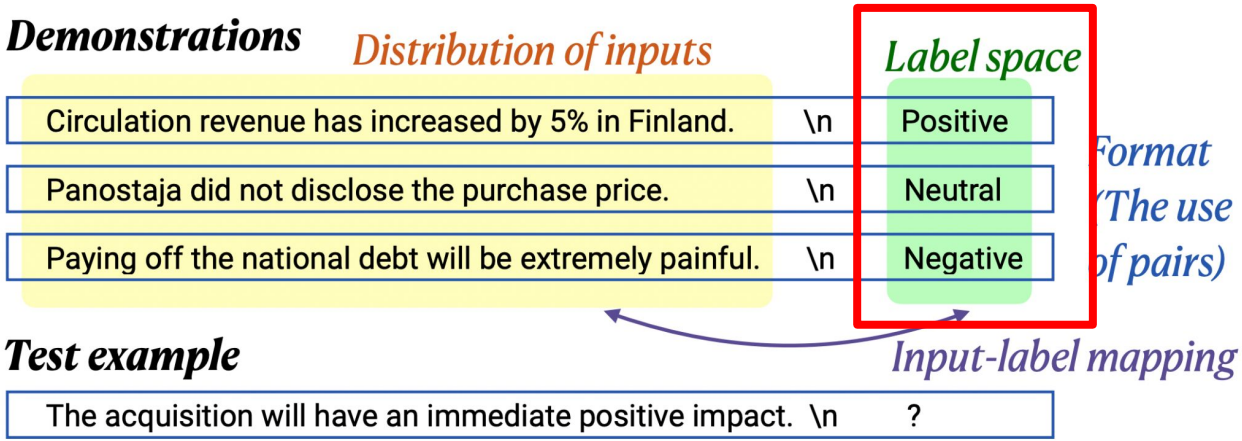
(Min et al., 2022): Rethinking the Role of Demonstrations: What Makes In-Context Learning Work

We break the prompt into four parts that provide signal to the model

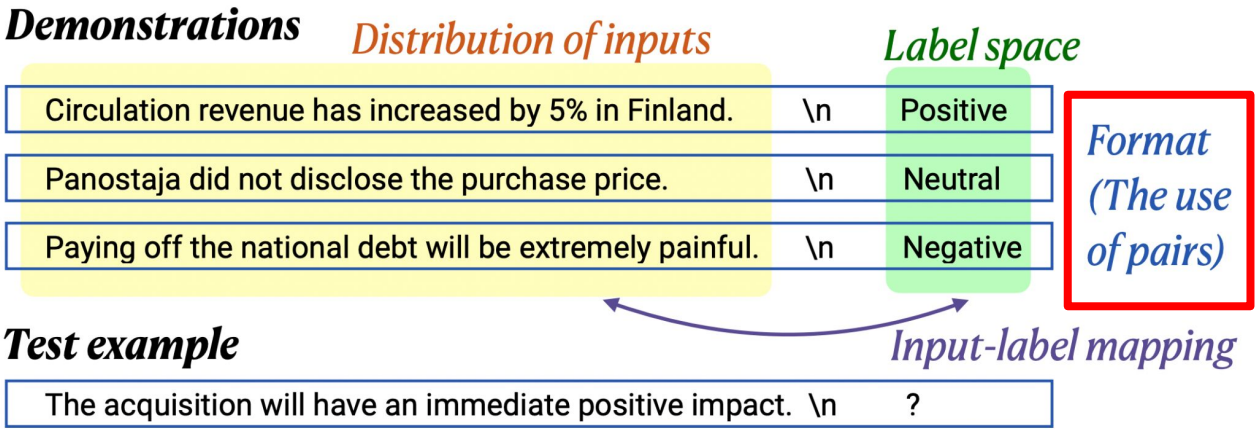
Distribution of Inputs



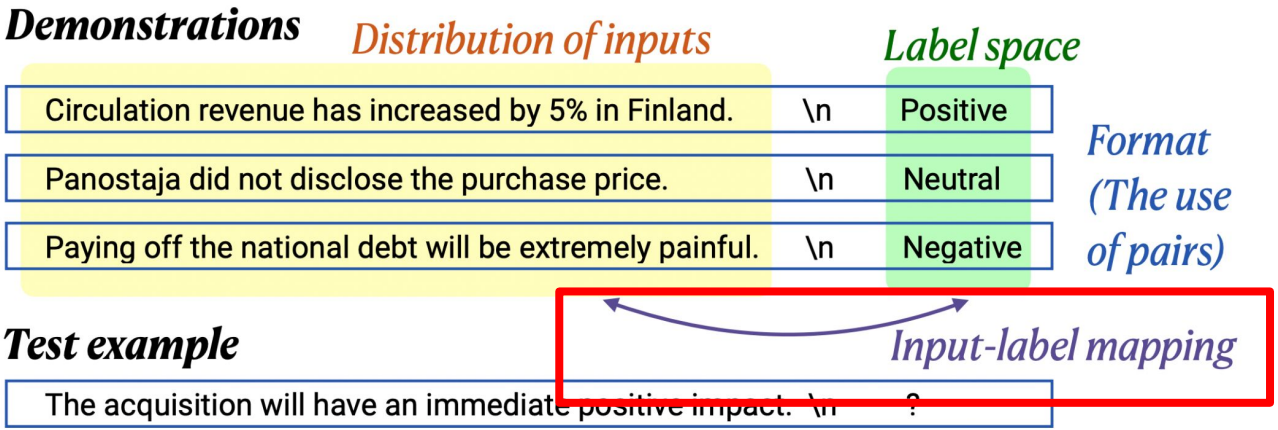
Label Space


Demonstrations
Distribution of inputs
Label space
Circulation revenue has increased by 5% in Finland. \n Positive
Panostaja did not disclose the purchase price. \n Neutral
Paying off the national debt will be extremely painful. \n Negative
Format
(The use
of pairs)
Test example
Input-label mapping
The acquisition will have an immediate positive impact. \n ?

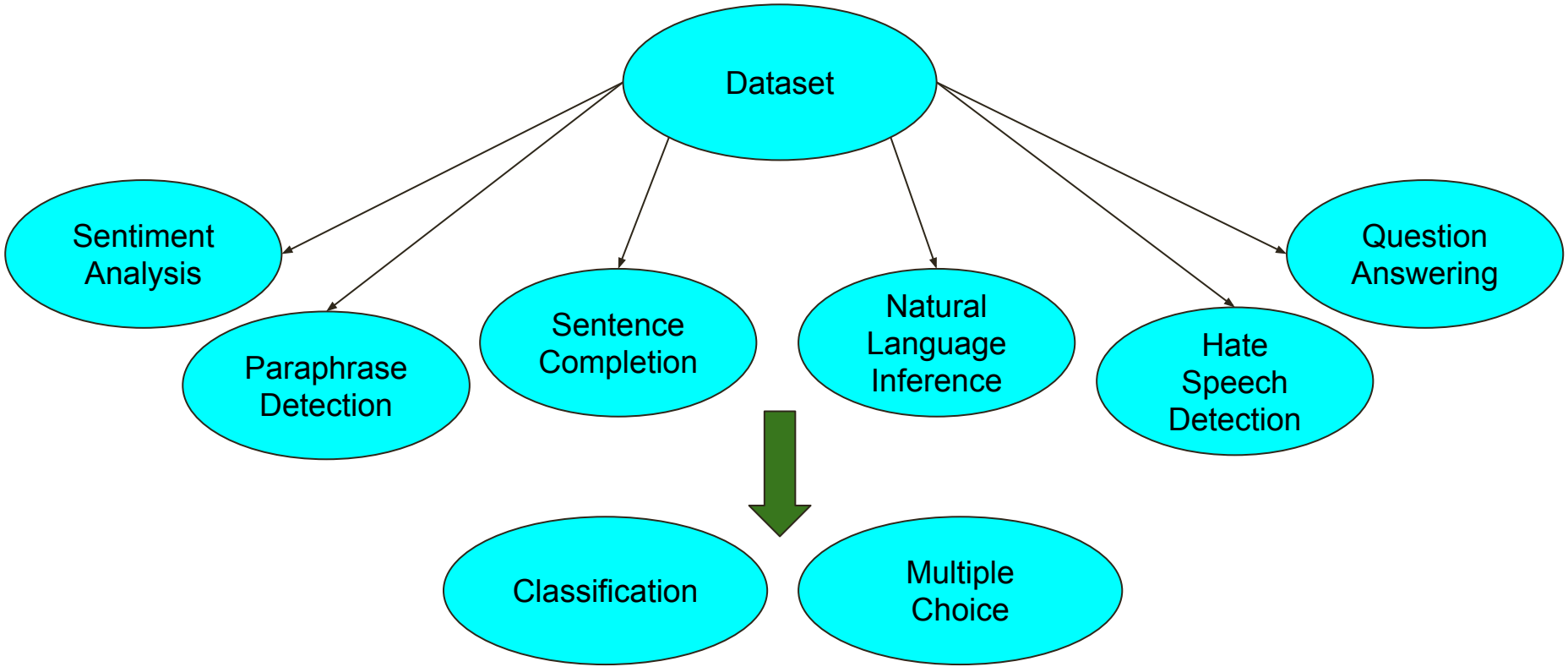
Format


Demonstrations
Distribution of inputs
Label space
Circulation revenue has increased by 5% in Finland. \n Positive
Panostaja did not disclose the purchase price. \n Neutral
Paying off the national debt will be extremely painful. \n Negative
Format (The use of pairs)
Test example
Input-label mapping
The acquisition will have an immediate positive impact. \n ?

Input-label Mapping



Experiment Setup



Datasets

Classification Tasks
Task category: Sentiment analysis
financial_phrasebank
poem_sentiment
Task category: Paraphrase detection
medical_questions_pairs
glue-mrpc
Task category: Natural language inference
glue-wnli
climate_fever
Task category: Question answering
quarel
openbookqa

Some of the datasets used for classification

Multiple Choice Tasks
Task category: Hate speech detection
hate_speech18
ethos-national_origin
Task category: Sentence completion
codah
superglue-copa

Some of the datasets used for multiple choice

Evaluation Methodology

- Metrics
  - Classification: **Macro-F1**
  - Multiple Choice: **Accuracy**
- Compute per-dataset average across seeds, and report **macro-average over datasets**.

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‡	✗	✗

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‡	✗	✗

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‡	✗	✗

- GPT-2 Large tuned on a set of tasks to learn how to in-context learn

(Min et al., 2022): MetaICL: Learning to Learn In Context

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‡	✗	✗

- An open-source autoregressive model that is an alternative to GPT-3

https://github.com/kingoflolz/mesh-transformer-jax

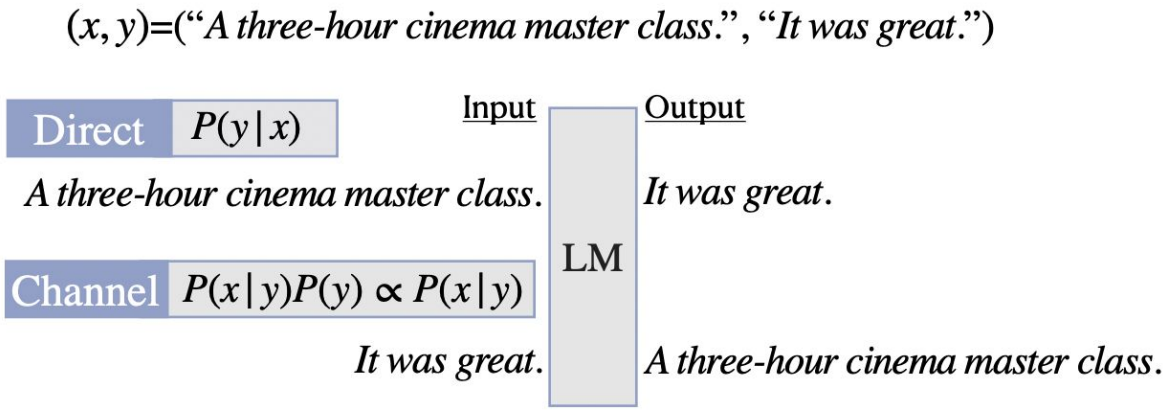
Models

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‡	✗	✗

- Autoregressive transformer model developed by Meta similar in size and architecture to GPT-3

(Artetxe et al., 2021) Efficient Large Scale Language Modeling with Mixtures of Experts

Direct vs Channel Models

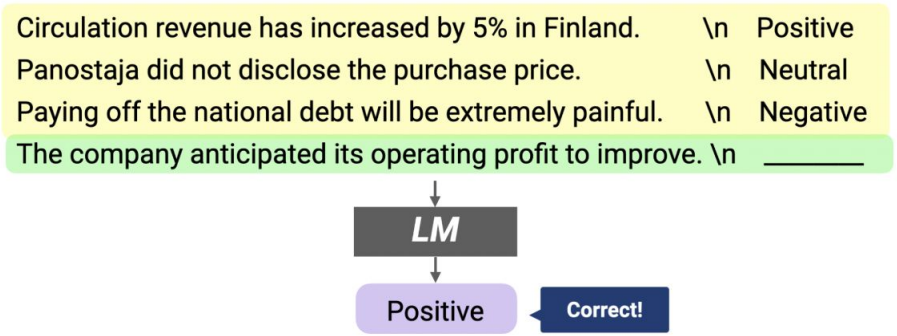


P(y|x) ⟶ P(x|y)

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

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

True Labels vs Random Labels



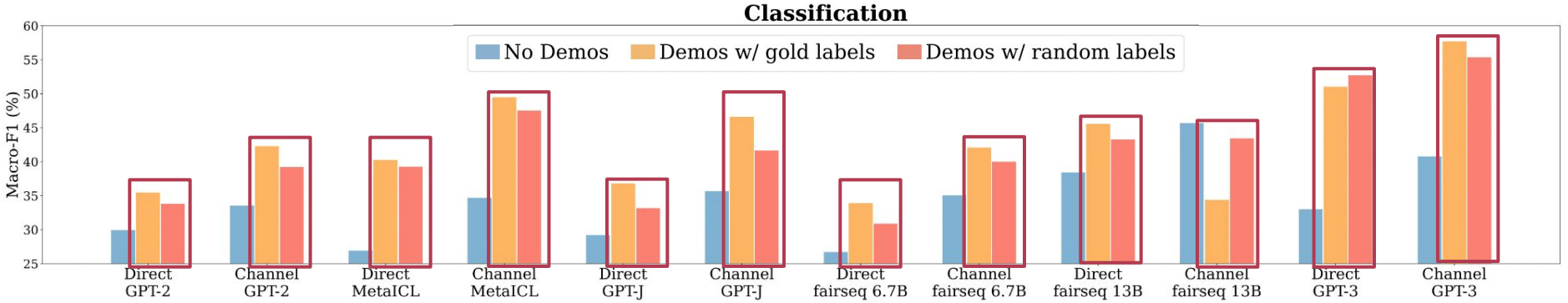
Prompt with true labels

Prompt with random labels

1. Randomly sample a label from the correct label space
2. Assign the label to the example

https://ai.stanford.edu/blog/understanding-incontext/

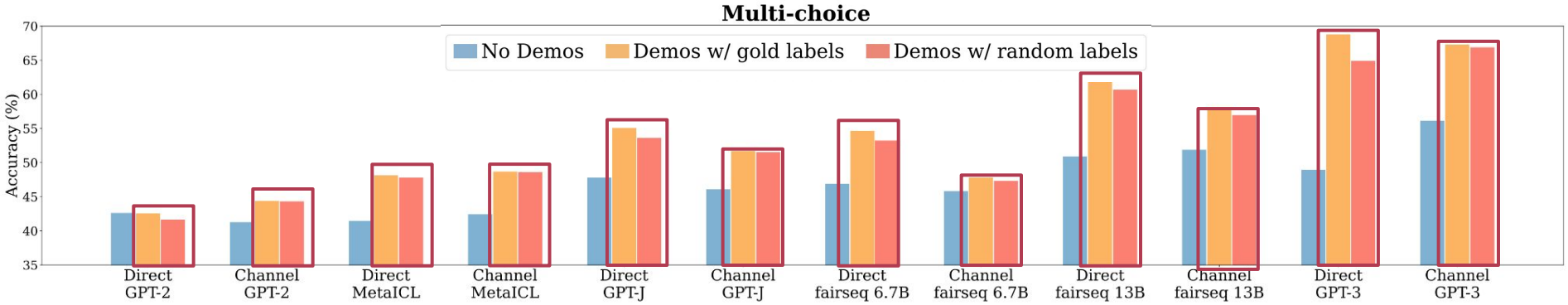
Results



Comparisons between no-examples (blue), examples with ground truth outputs (yellow) and examples with random outputs (red)

Models see small performance drop in the range of 0–5% absolute with random labels

Results



Comparisons between no-examples (blue), examples with ground truth outputs (yellow) and examples with random outputs (red)

Models see small performance drop in the range of 0–5% absolute with random labels

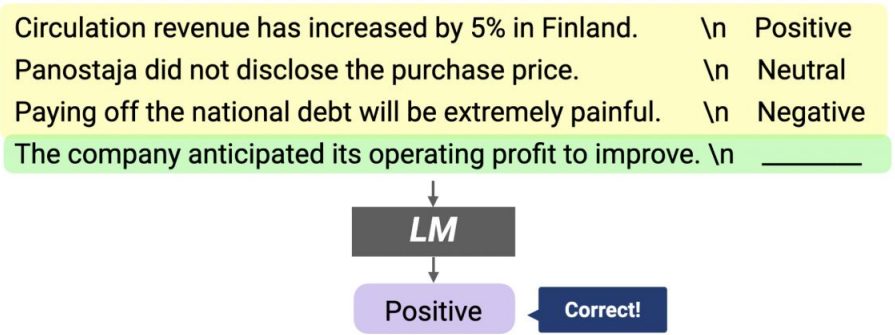
Results Takeaways

Ground truth input-label mapping in the prompt is not as important as we thought

Model is not recovering the expected input-label correspondence for the task from the input-label pairings

Is this result consistent in other setups?

Does the number of correct labels matter?



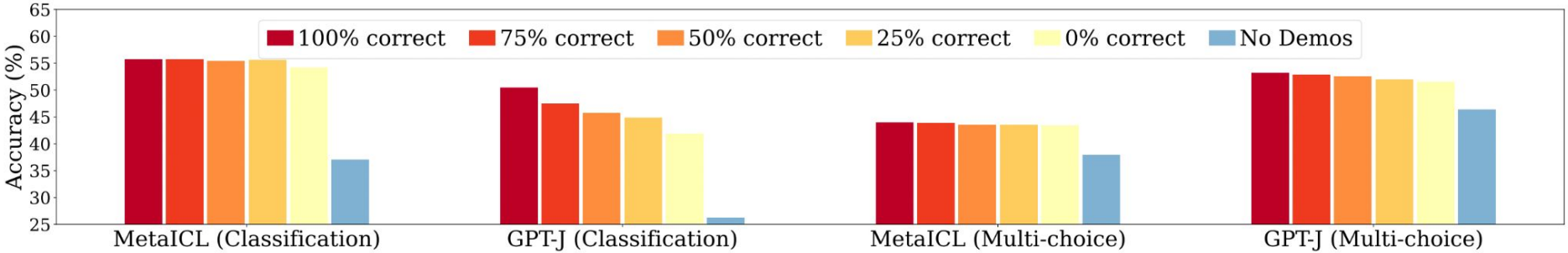
Prompt with all true labels

Prompt with one true label

1. Vary the number of correct labels in examples

https://ai.stanford.edu/blog/understanding-incontext/

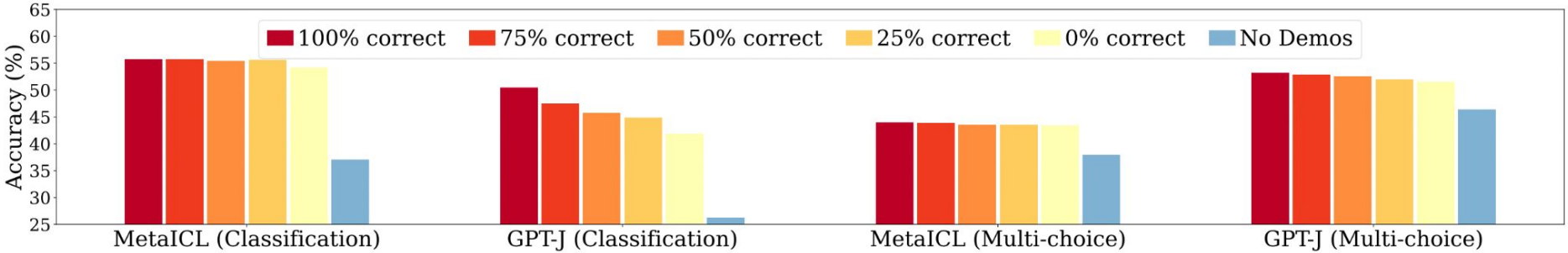
Results



Results of varying number of correct labels in the prompt

Using all incorrect labels preserve **92%** of improvements from using all correct labels

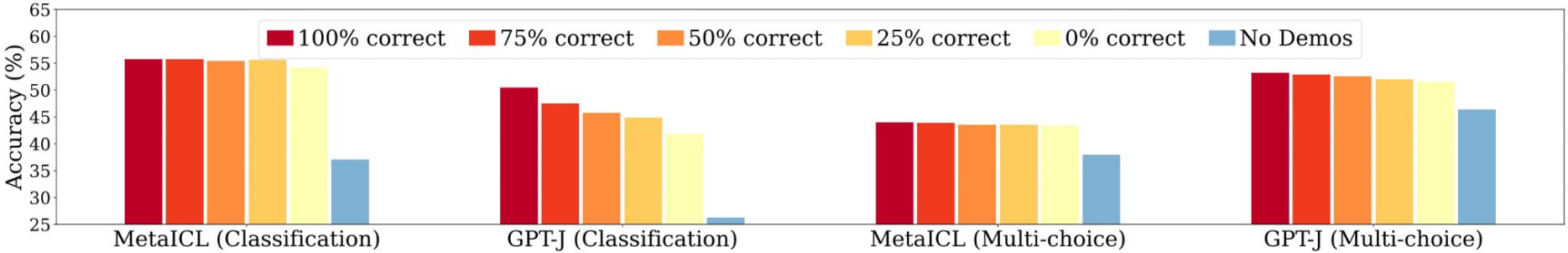
Results



Results of varying number of correct labels in the prompt

Using all incorrect labels preserves **100%** of improvements from using all correct labels

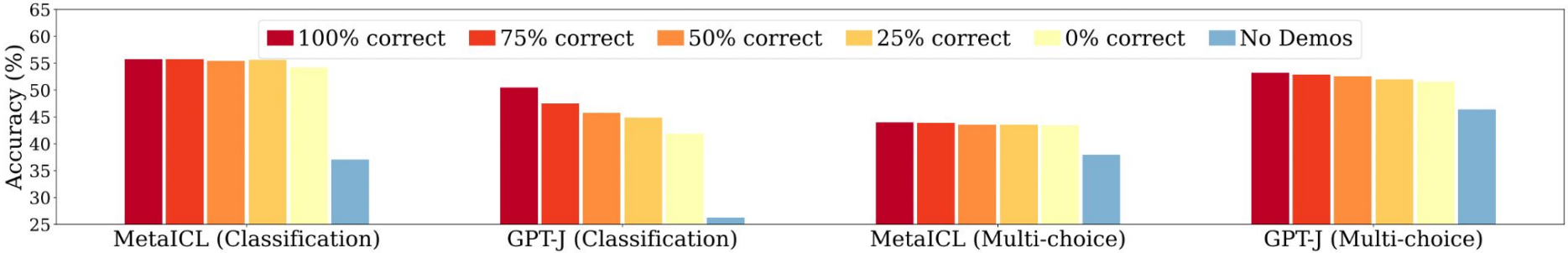
Results



Results of varying number of correct labels in the prompt

Using all incorrect labels preserves **97%** of improvements from using all correct labels

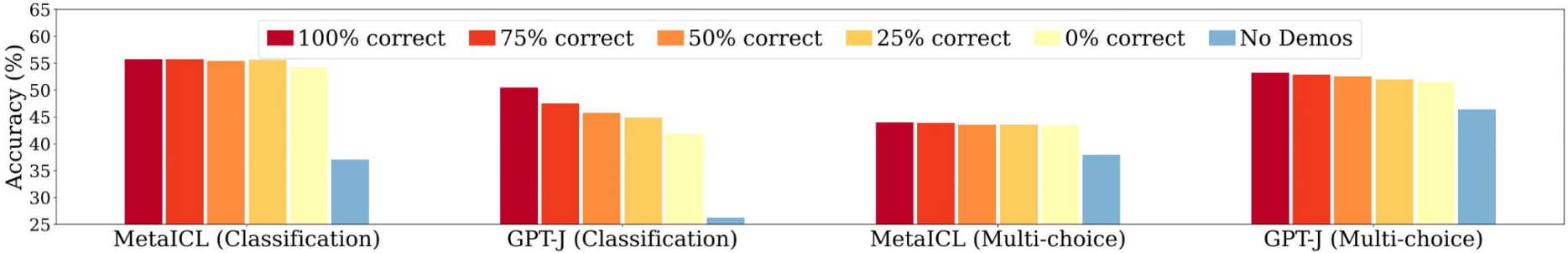
Results



Results of varying number of correct labels in the prompt

Performance **depends more** on number of correct labels

Results Takeaways

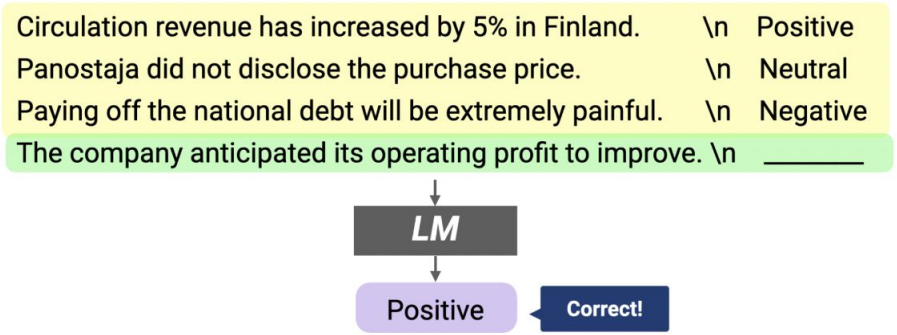


Results of varying number of correct labels in the prompt

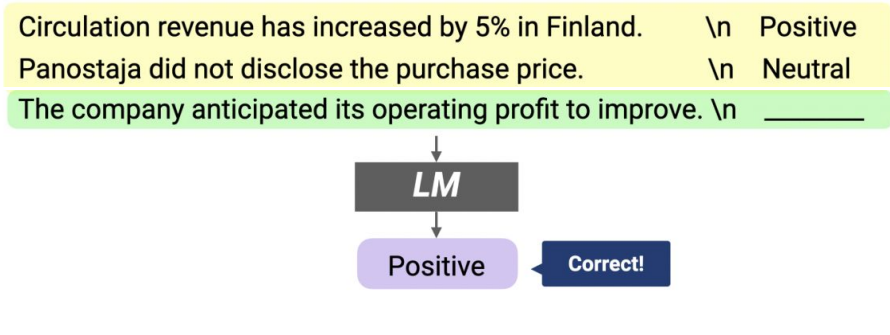
Model performance is fairly insensitive to the number of correct labels

Using incorrect labels is better than no examples

Varying the Number of Examples



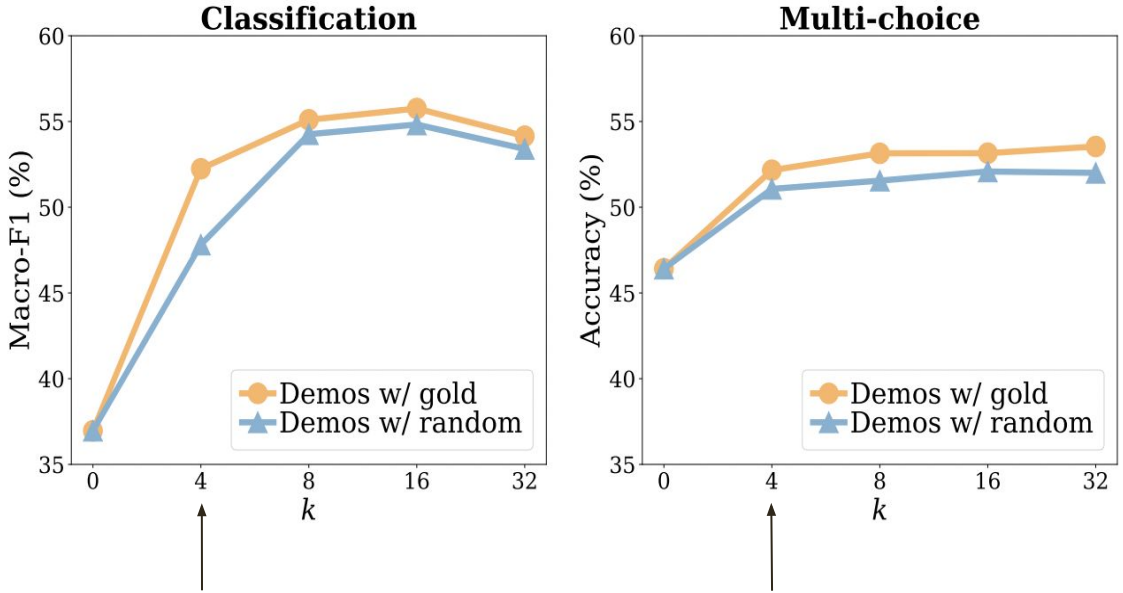
Prompt with three examples



Prompt with two examples

Measure whether the results of using **random labels** is consistent across **differing number of examples**

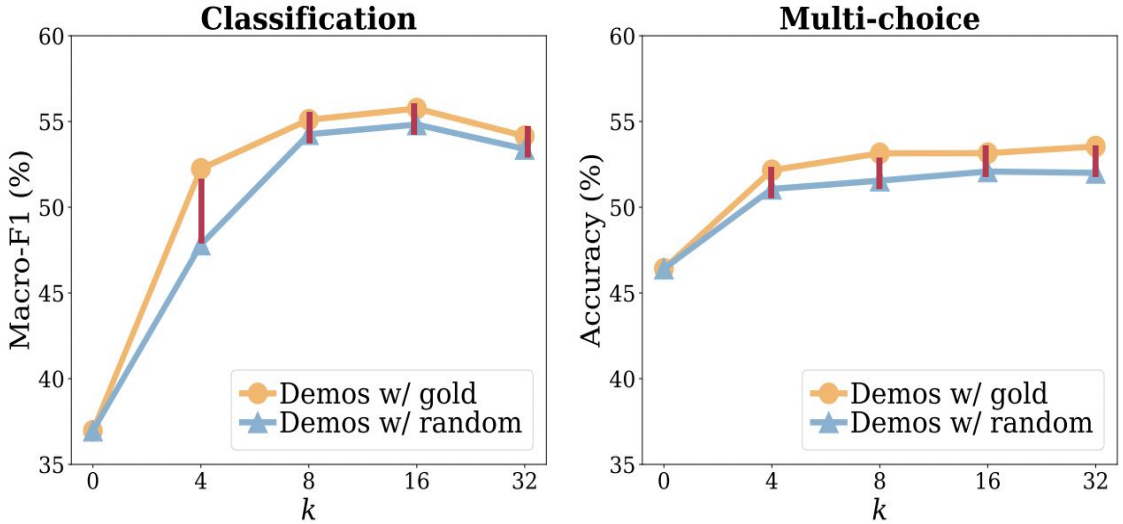
Results



Ablations on varying numbers of examples (k) in the prompt.

Using **small number** of examples with **random labels** is better than **no examples**

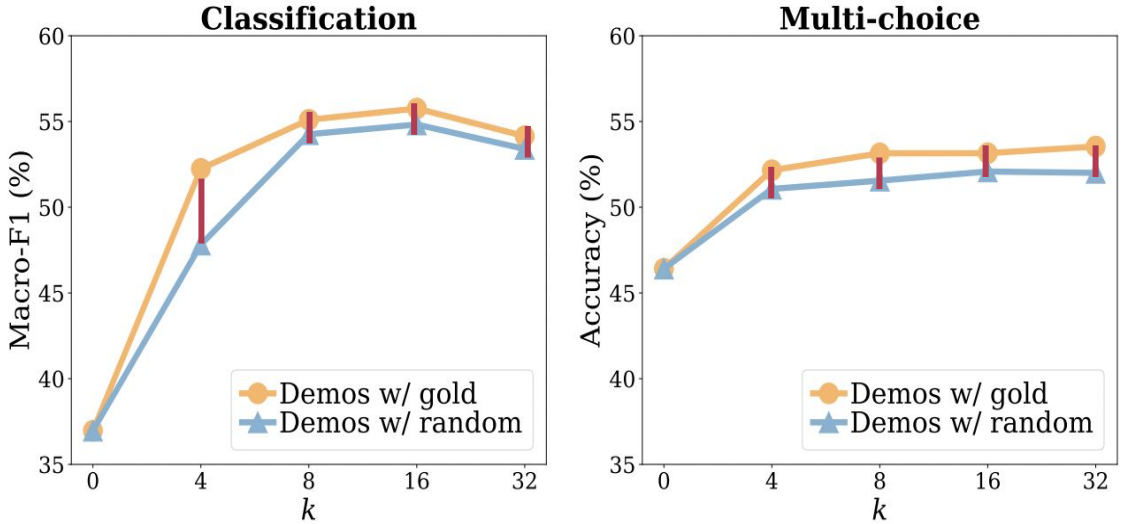
Results



Ablations on varying numbers of examples (k) in the prompt.

Performance drop from using gold labels to using random labels is **consistently small** across varying k, ranging from 0.8–1.6%

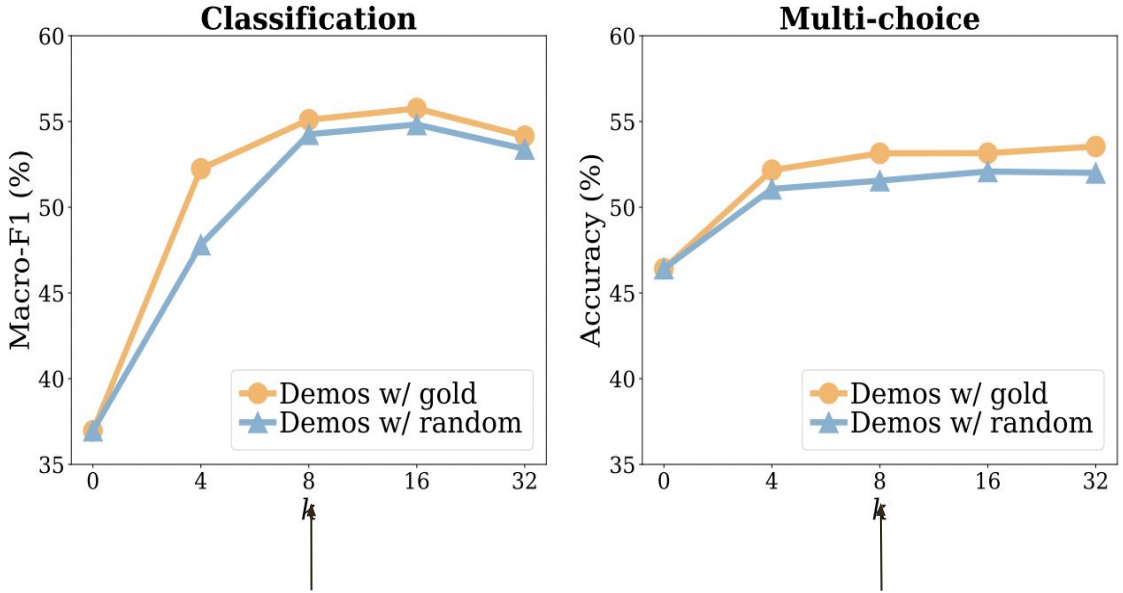
Results Takeaways



Ablations on varying numbers of examples (k) in the prompt.

Performance differences of random labels is consistent across number of examples

Results Takeaways



Ablations on varying numbers of examples (k) in the prompt.

More examples even with random labels improves model performance except beyond a threshold

Using Better Templates

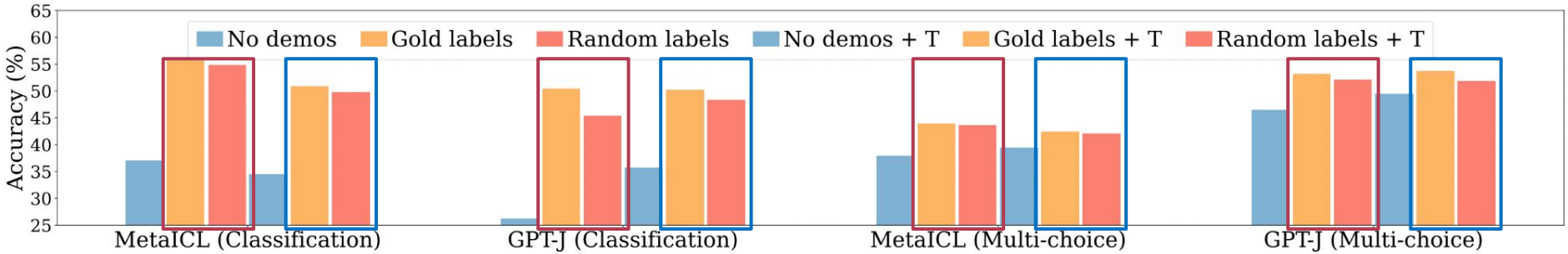
Dataset	Type	Example
Tweet_eval-hate	Minimal	The Truth about #Immigration \n {hate|non-hate}
	Manual	Tweet: The Truth about #Immigration \n Sentiment: {against|favor}

Example of minimal and manual templates

- Minimal templates follow a conversion procedure (**dataset-agnostic**)
- Manual templates are written in a **dataset-specific** manner

Measure whether the results of using **random labels** is consistent when using **manual templates**

Results



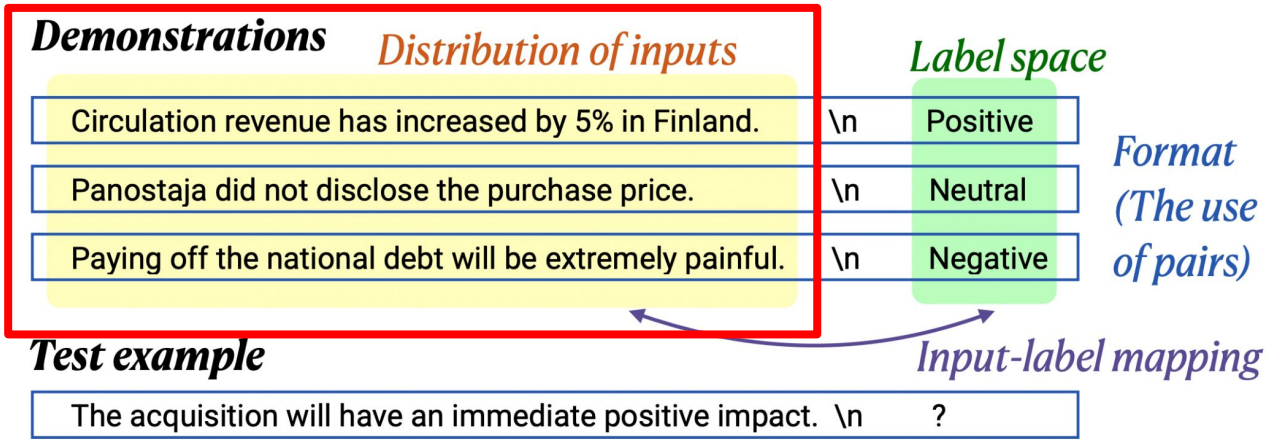
Results with minimal templates and manual templates. '+T' indicates that manual templates are used.

Random labels still minimally hurt performance with manual templates

The prompt provides evidence for the model to locate the concepts learned during pre-training

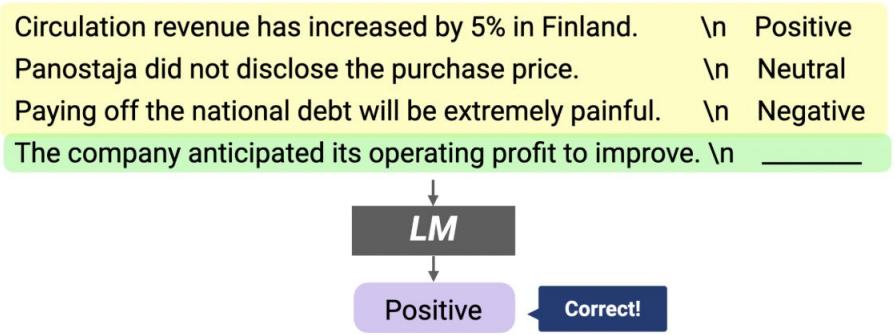
- Random input-label mapping **increases noise** but the **other components of the prompt** allow the model to perform Bayesian inference by **providing signals**

Distribution of Inputs

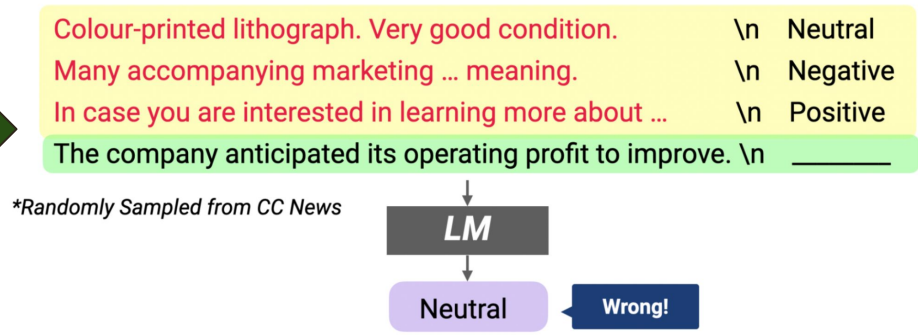


Evaluate the importance of the distribution of inputs

Using out-of-distribution input text



Prompt with in-distribution sentences

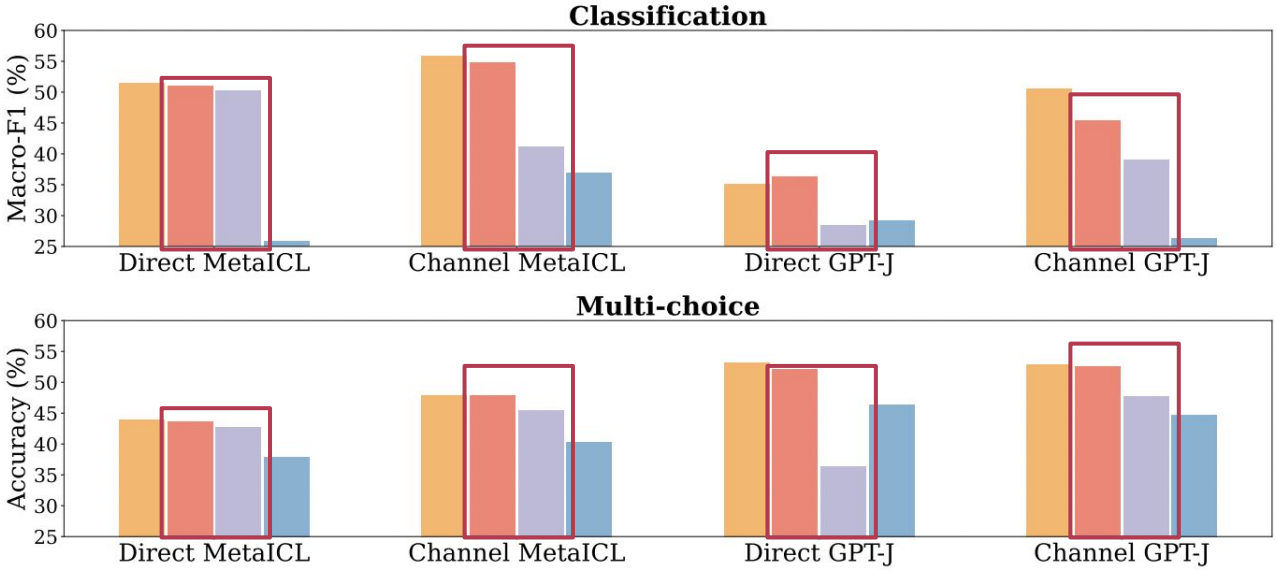


Prompt with out-of-distribution sentences

Input sentences are randomly sampled from an **external corpus**, replacing the input from the downstream task training data

https://ai.stanford.edu/blog/understanding-incontext/

Seeing in-distribution inputs improves performance



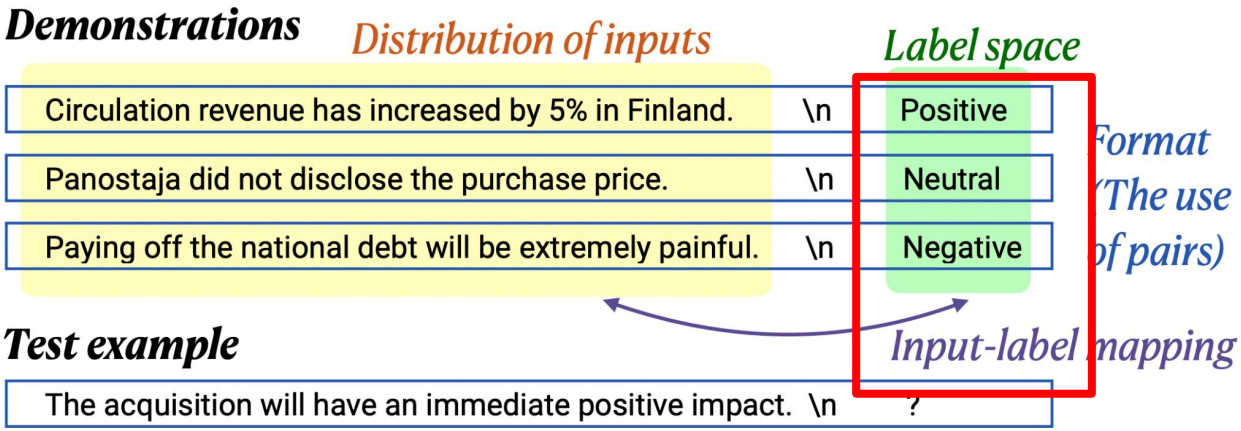
	F	L	I	M
Gold labels	✓	✓	✓	✓
Random labels	✓	✓	✓	✗
OOD + Random labels	✓	✓	✗	✗
No demonstrations	✗	✗	✗	✗

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

Results of using out-of-distribution input sentences

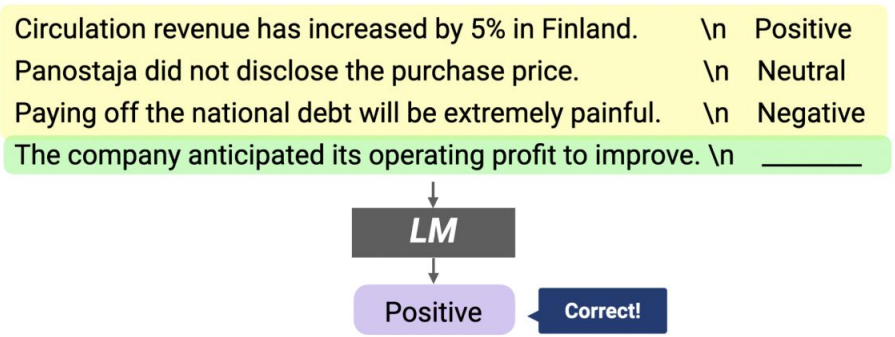
Random sentences result in performance **decreases of up to 16% absolute** compared to using inputs from training data

Label Space

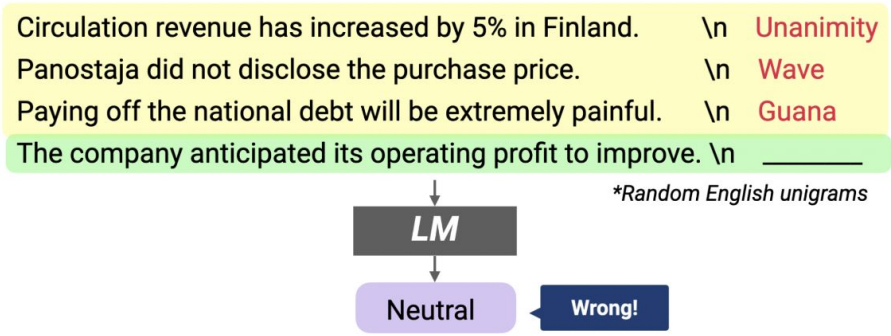


Evaluate the importance of the label space

Using random labels from an incorrect label space



Prompt with true labels

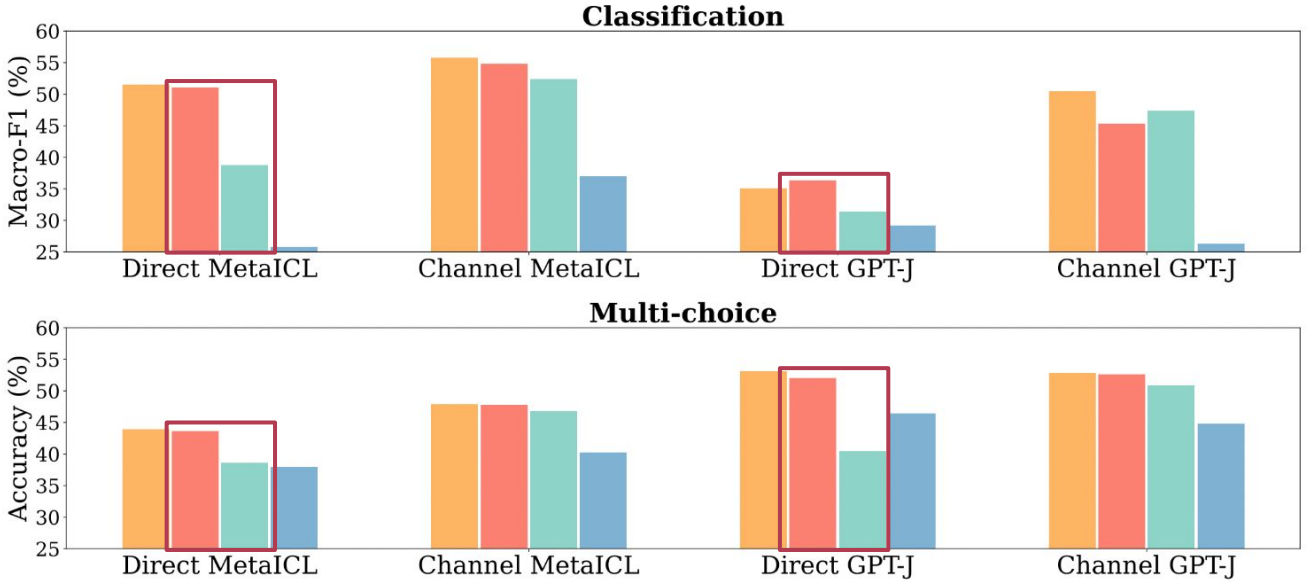


Prompt with random English words as labels

1. Sample a random subset of English words with same size as set of truth labels
2. Labels are replaced with words randomly drawn from this subset

https://ai.stanford.edu/blog/understanding-incontext/

Seeing correct label space is important



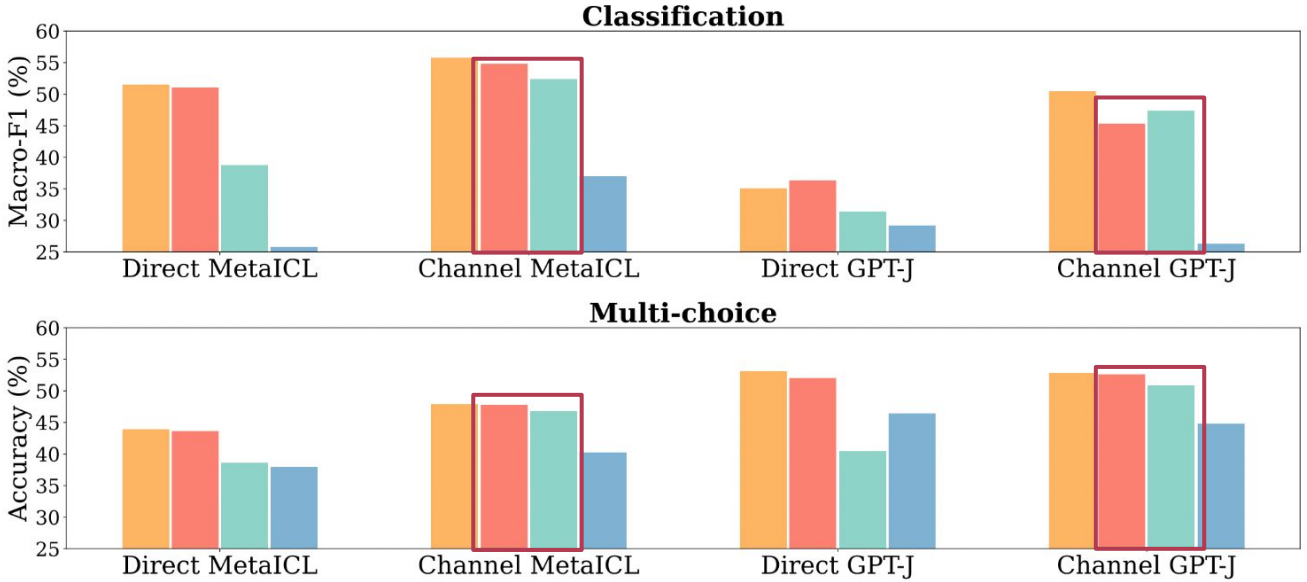
	F	L	I	M
Gold labels	✓	✓	✓	✓
Random labels	✓	✓	✓	✗
Random English words	✓	✗	✓	✗
No demonstrations	✗	✗	✗	✗

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

Results of using random English words as labels

Labels not in the correct label space result in **performance decreases of up to 16% absolute** in **direct models**

Seeing correct label space is important



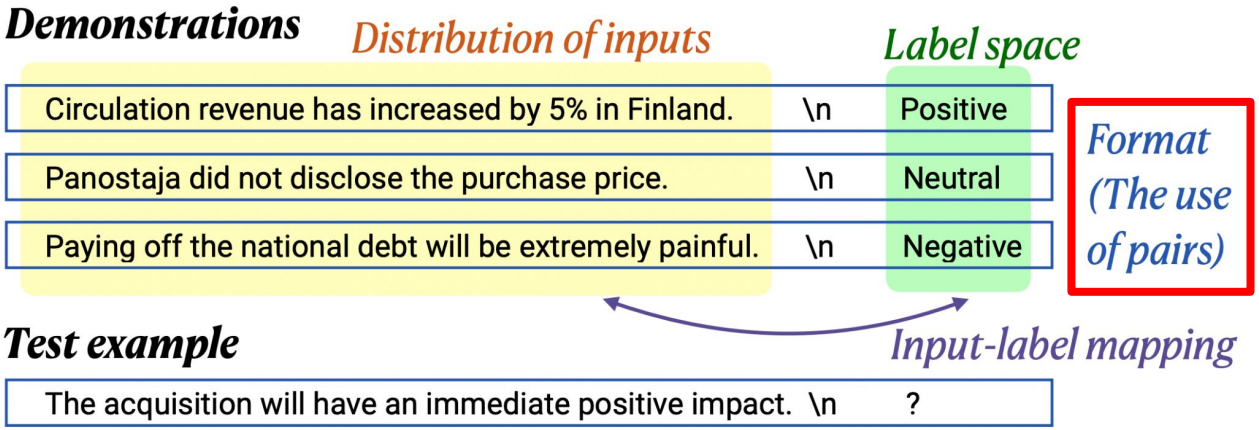
	F	L	I	M
Gold labels	✓	✓	✓	✓
Random labels	✓	✓	✓	✗
Random English words	✓	✗	✓	✗
No demonstrations	✗	✗	✗	✗

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

Results of using random English words as labels

Labels not in the correct label space result in **performance decreases of up to 2% absolute** in **channel models**

Format



Evaluate the importance of pairing an input sentence with a label

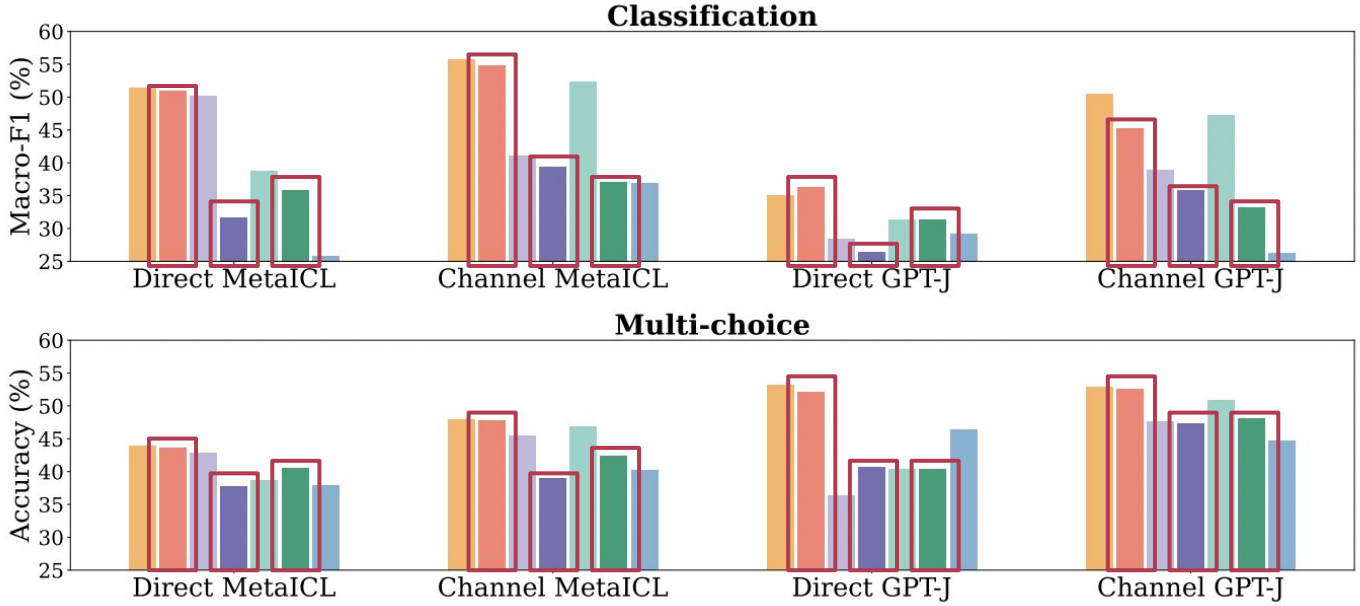
Changing the input-label format

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

Examples with only inputs (top) and only labels (bottom)

Feed in examples with **no labels** and **with labels only**

Keeping the input-label format for demonstrations is vital for performance



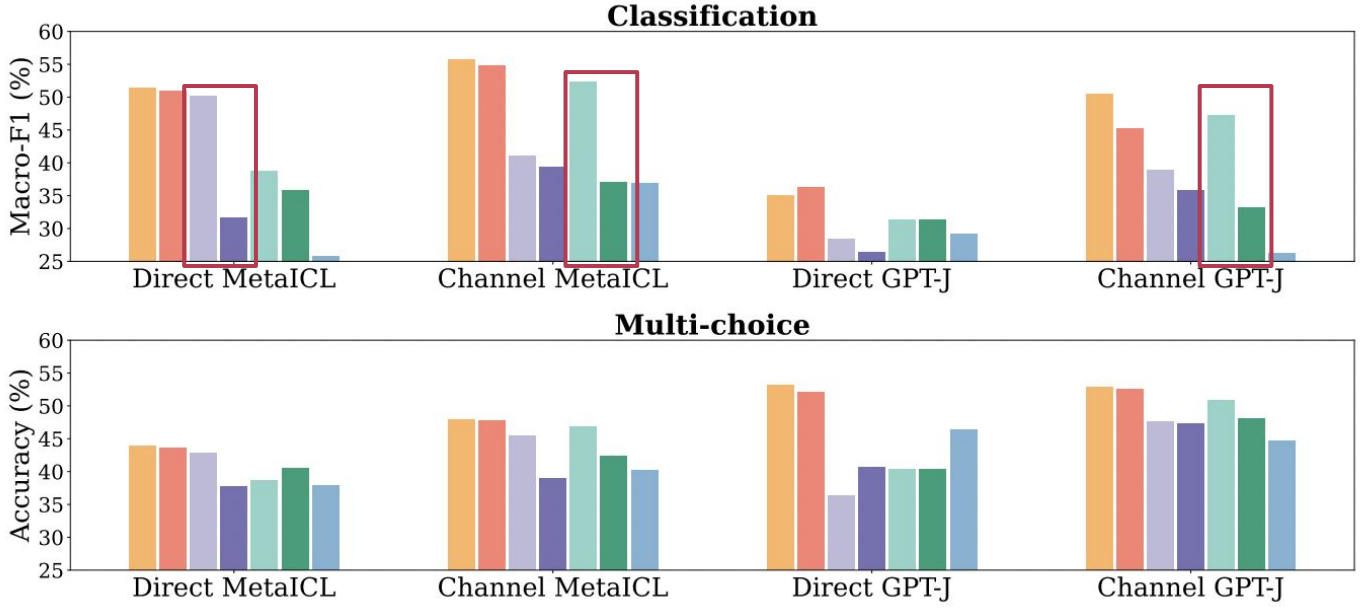
	F	L	I	M
Gold labels	✓	✓	✓	✓
Random labels	✓	✓	✓	✗
OOD + Random labels	✓	✓	✗	✗
Random labels only	✗	✓	✗	✗
Random English words	✓	✗	✓	✗
No labels	✗	✗	✓	✗
No demonstrations	✗	✗	✗	✗

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

Results of feeding in only inputs and only labels

Not using the input-label format **decreases performance**

Keeping the input-label format for demonstrations is vital for performance



	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

Results of feeding in only inputs and only labels

Using **out-of-distribution inputs** and **random English words** as labels is better than only keeping **one part of the format** or having no demonstrations

Q1. What are the most surprising findings to you in Min et al., 2022? How would this change our current understanding of in-context learning?

Having correct input-output pairs do not matter as much as long as we know the correct label space. Retaining the format (input-output pairs) whether by using (OOD + random labels) or (in-distribution sentences + random English words) also decently improves performance. This means that in-context learning actually has a higher zero-shot performance than we thought.

Future work

Possible projects for this class

- Understanding model performance on **“unseen” tasks**, i.e. the out-of-distribution case, where θ* is not in Θ
- Capturing effects from **model architecture** and training. How to include the model scale in this framework?
- Extending the framework to incorporate **task descriptions** as part of the prompts
- **Understanding pre-training data** for in-context learning. Is there a critical subset of data from which in-context learning emerges?
- **Variable length demonstrations**, i.e. k is different in each example

Questions

Contact info: saml@princeton.edu
kexinj@math.princeton.edu

Q3. We learned that the output space is very important for the success of in-context learning (e.g., the set of labels for classification tasks). However, Min et al 2022 mainly focus on classification and multiple-choice tasks and there are many other NLP tasks that the output space is much larger and harder to be in the "same format", such as open-domain QA, summarization or semantic parsing (the output can be a complex logical form). Can we design similar experiments as in Min et al 2022 for these tasks too (and how)? Can we design better ways of improving the in-context learning's performance following the findings we learned?