Data Collection through Large Language Model Prompts

References:
https://courses.grainger.illinois.edu/CS447/sp2023/Slides/Lecture27.pdf
https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1224/slides/cs224n-2022-lecture10-pretraining.pdf
https://web.stanford.edu/class/cs224u/slides/cs224u-incontextlearning-2023-handout.pdf
https://www.cs.princeton.edu/courses/archive/fall22/cos597G/

LLM Background

State of GPT (up to 10’)

https://www.youtube.com/watch?v=bZQun8Y4L2A

In-context Learning

- **In-context learning**:
  - A **frozen LM** performs a task only by conditioning on the prompt text.
  - Using text input of a pre-trained LM as a form of task specification
- **Few-shot in-context learning**: The prompt includes examples of the intended behavior
- **Zero-shot in-context learning**: The prompt includes no examples of the intended behavior

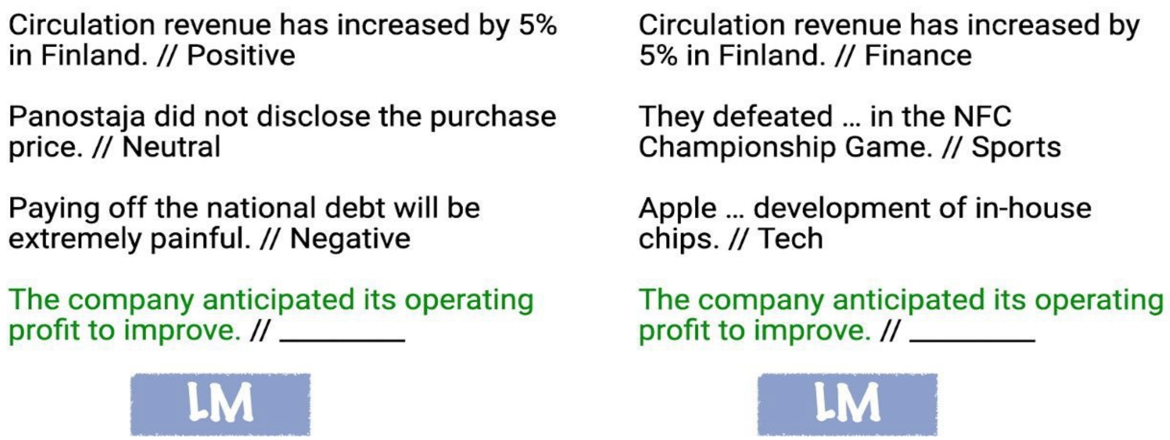
In-Context Learning

	No Prompt	Prompt
Zero-shot (0s)	skicts = sticks	Please unscramble the letters into a word, and write that word: skicts = sticks

1-shot
(1s)
	chiar = chair
	skicts = sticks
		Please unscramble the letters into a word, and write that word:
		chiar = chair
		skicts = sticks

Few-shot
(FS)
	chiar = chair
	[.]
	pciinc = picnic
	skicts = sticks
		Please unscramble the letters into a word, and write that word:
		chiar = chair
		[.]
		pciinc = picnic
		skicts = sticks

Mystery of 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. // ________

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

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

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

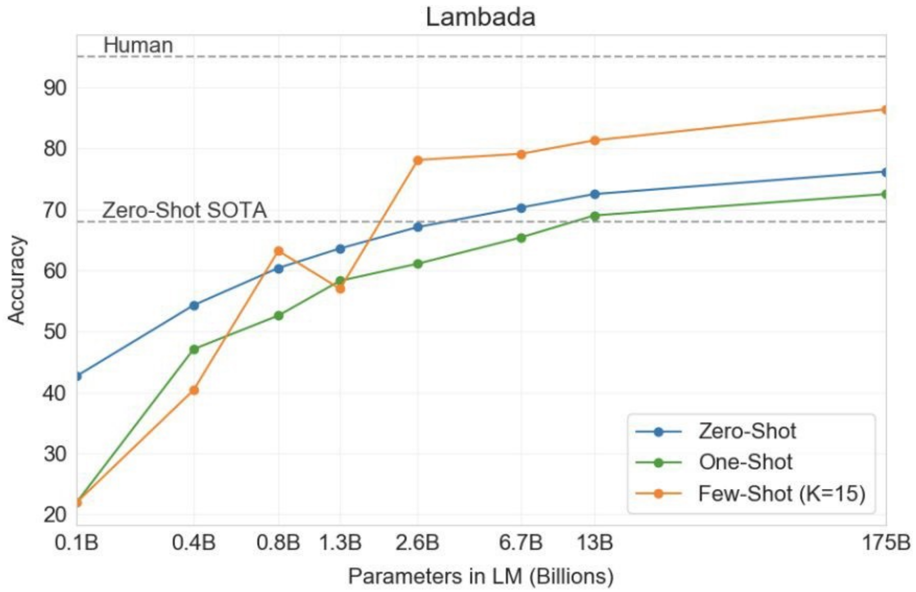
The company anticipated its operating profit to improve. // ________

- **No optimization of any parameters**!
  - Unlike conventional machine learning which fine tune parameters for specific tasks
- LM isn’t trained to learn from examples. **It is trained to do next token prediction**.
  - Unlike traditional meta-learning methods that train models to learn from examples
- There is seemingly **a mismatch between pretraining and in-context learning** (what we’re asking it to do).

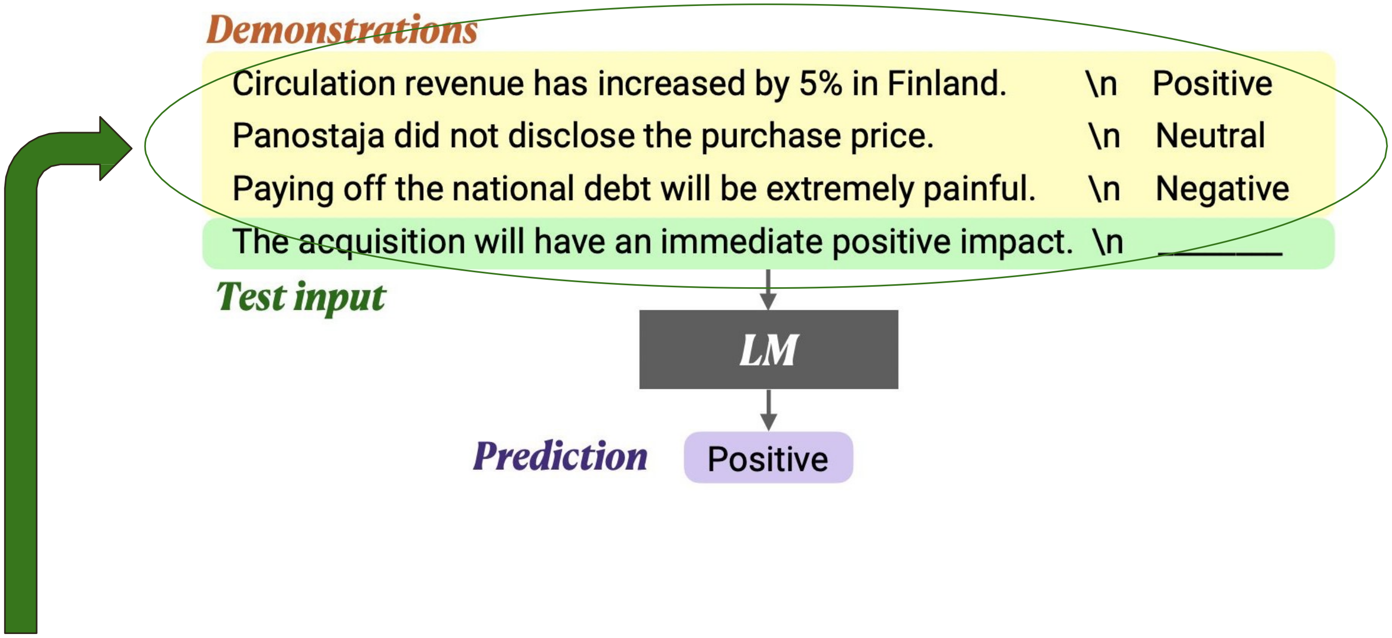
What Can In-Context Learning Do?

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

Works like magic!



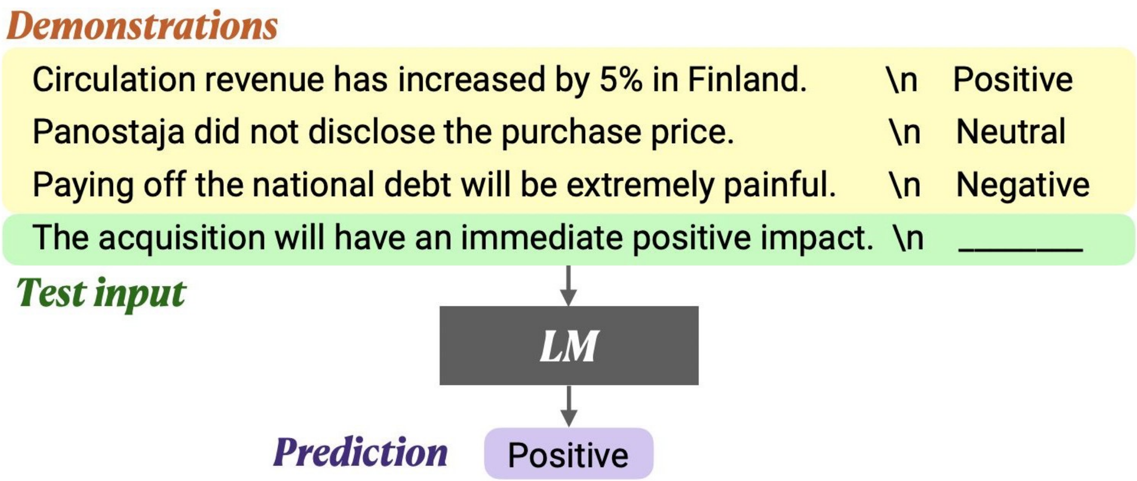
We don’t know how models in-context learn



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
The acquisition will have an immediate positive impact. \n ______

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

We don’t know how models in-context learn



Demonstrations

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

The acquisition will have an immediate positive impact. \n ______

Test input

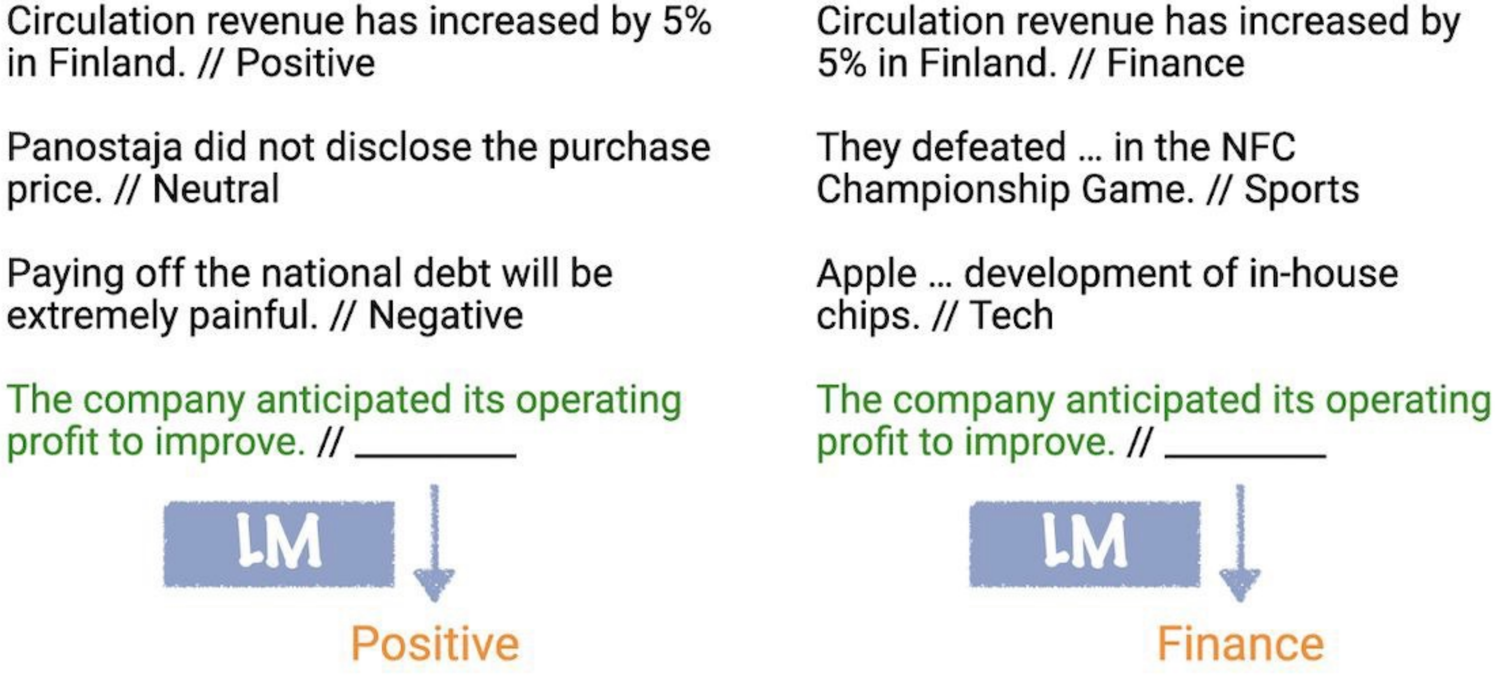
LM

Prediction Positive

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

How does it know what to do then?

[redacted]



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)

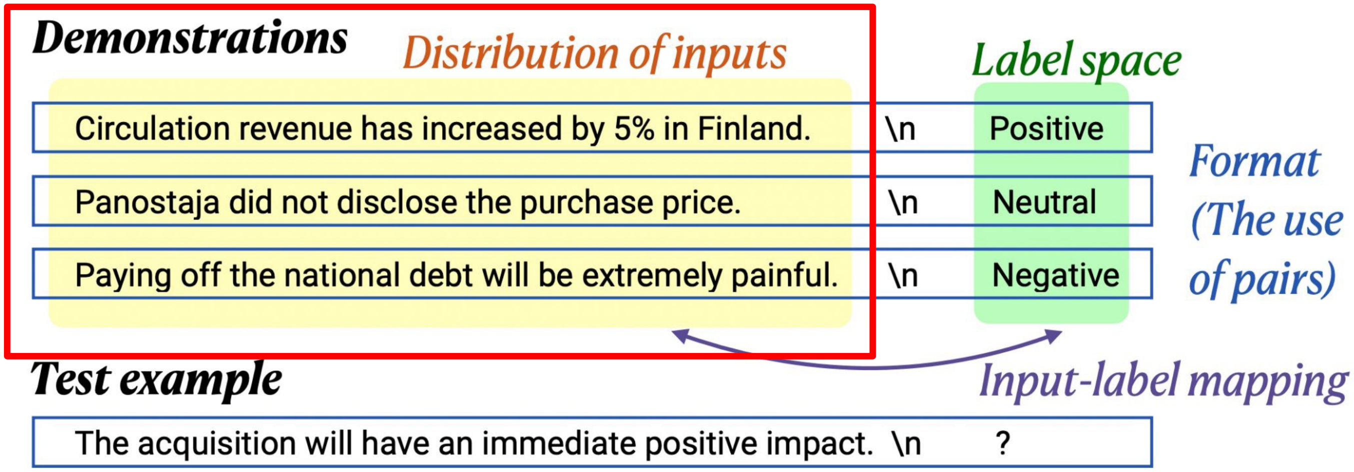
formatting

Which aspects of the prompt affect downstream task performance?

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

Rethinking the Role of Demonstrations: What makes In-context Learning Work?, Min et al., 2022

Distribution of Inputs



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

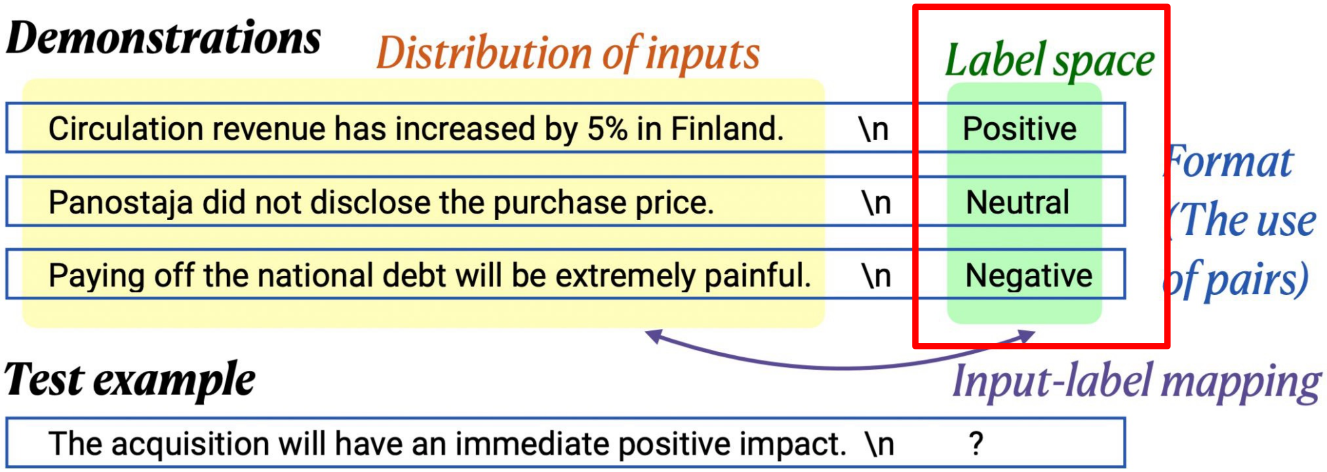
Paying off the national debt will be extremely painful. \n Negative

Format (The use of pairs)

Test example *Input-label mapping*

The acquisition will have an immediate positive impact. \n ?

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

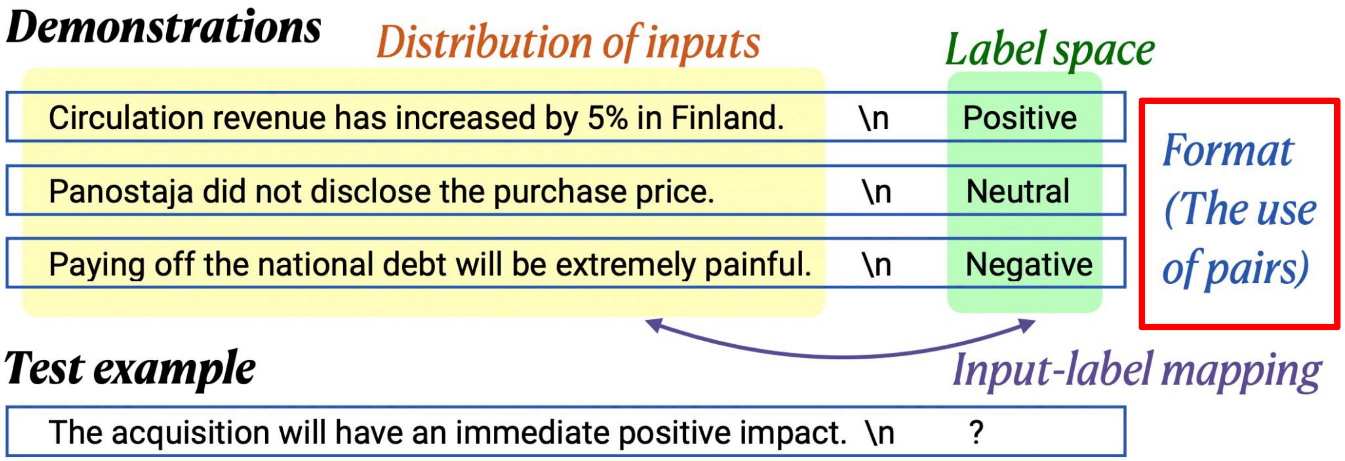
Format (The use of pairs)

Input-label mapping

Test example

The acquisition will have an immediate positive impact. \n ?

Format



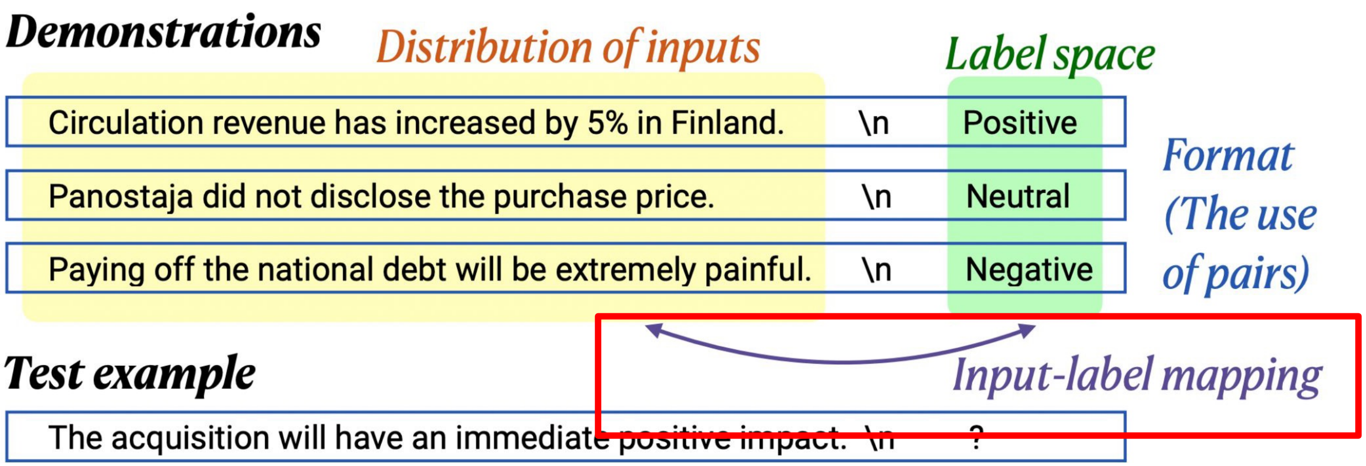
Circulation revenue has increased by 5% in Finland. \n Positive

Panostaja did not disclose the purchase price. \n Neutral

Paying off the national debt will be extremely painful. \n Negative

The acquisition will have an immediate positive impact. \n ?

Input-label Mapping



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

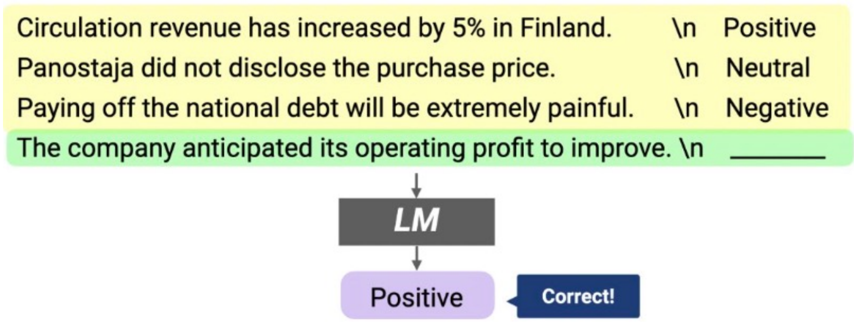
Format (The use of pairs)

Input-label mapping

Test example

The acquisition will have an immediate positive impact. \n ?

True Labels vs Random Labels



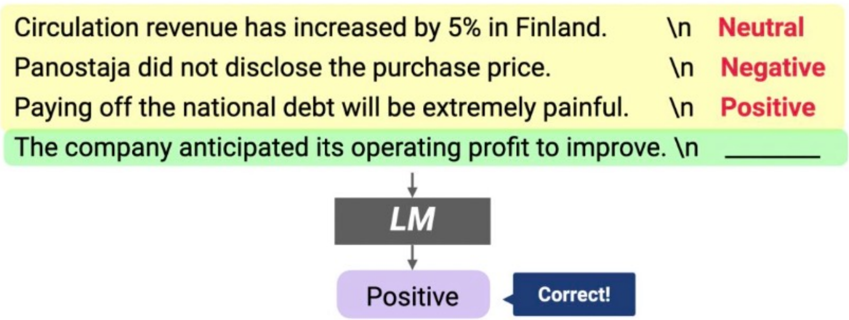
Circulation revenue has increased by 5% in Finland. \n Positive

Panostaja did not disclose the purchase price. \n Neutral

Paying off the national debt will be extremely painful. \n Negative

The company anticipated its operating profit to improve. \n ______

Prompt with true labels



Circulation revenue has increased by 5% in Finland. \n Neutral

Panostaja did not disclose the purchase price. \n Negative

Paying off the national debt will be extremely painful. \n Positive

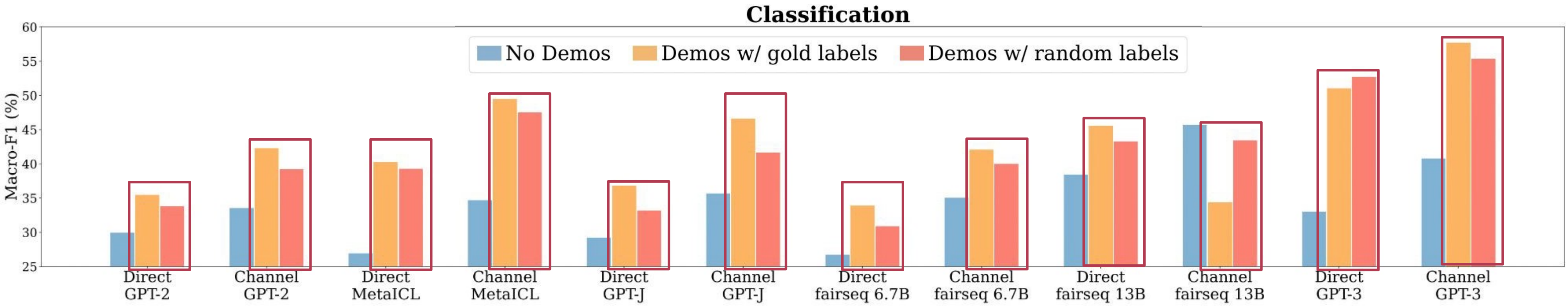
The company anticipated its operating profit to improve. \n ______

Prompt with random labels

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

Results

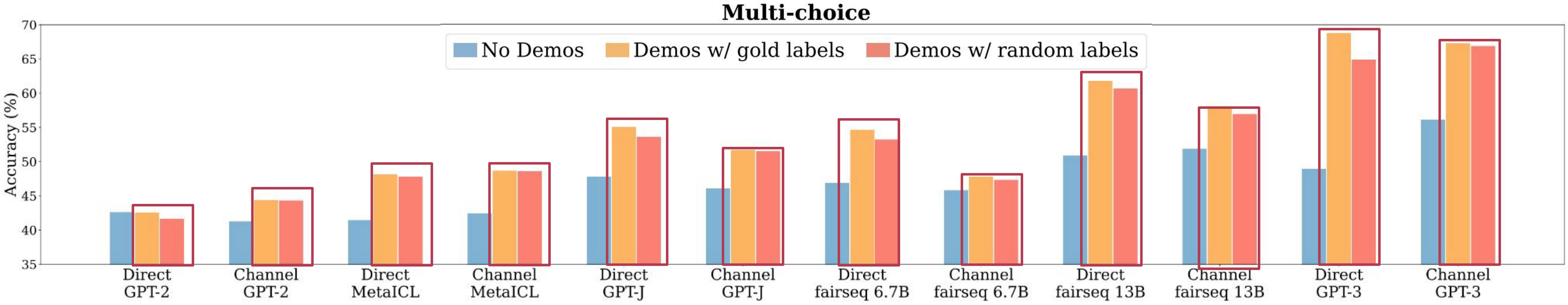
Warning: as the model has been updated, you may not replicate these results.



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)

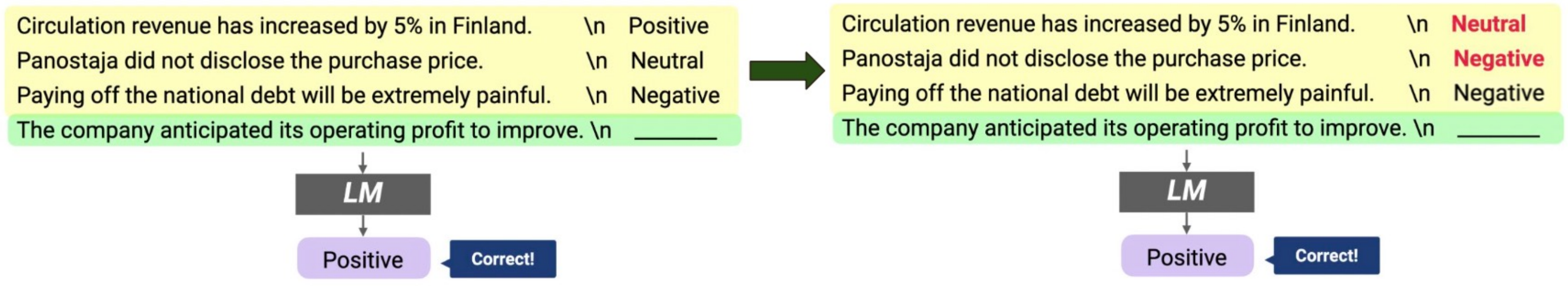
Models see small performance drop in the range of 0–5% absolute with random labels

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

Does the number of correct labels matter?



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
The company anticipated its operating profit to improve. \n ______

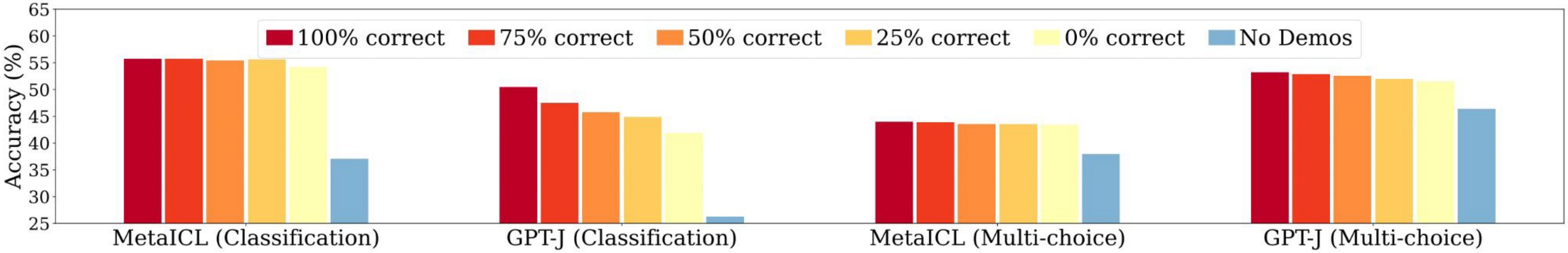
Prompt with all true labels

Circulation revenue has increased by 5% in Finland. \n Neutral
Panostaja did not disclose the purchase price. \n Negative
Paying off the national debt will be extremely painful. \n Negative
The company anticipated its operating profit to improve. \n ______

Prompt with one true label

1. Vary the number of correct labels in examples

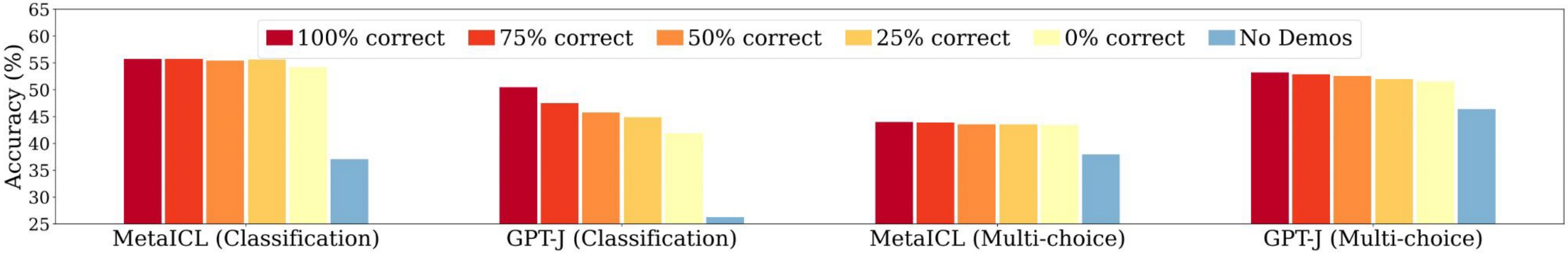
Results



Results of varying number of correct labels in the prompt

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

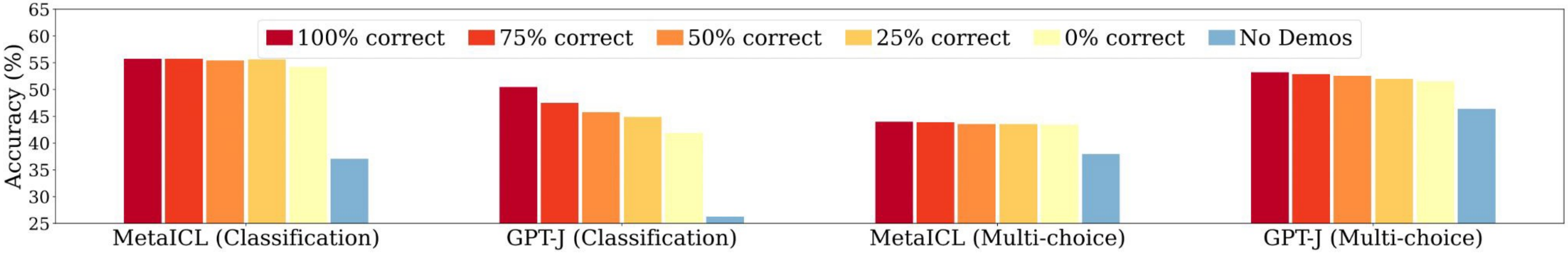
Results



Results of varying number of correct labels in the prompt

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

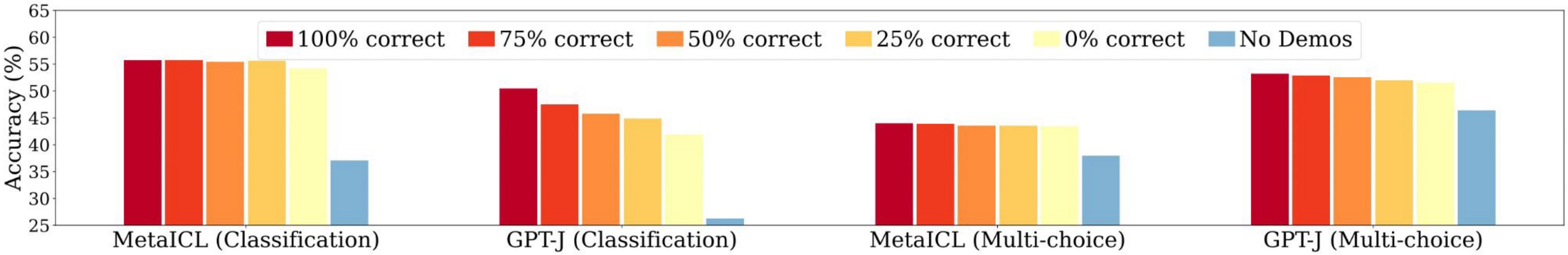
Results



Results of varying number of correct labels in the prompt

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

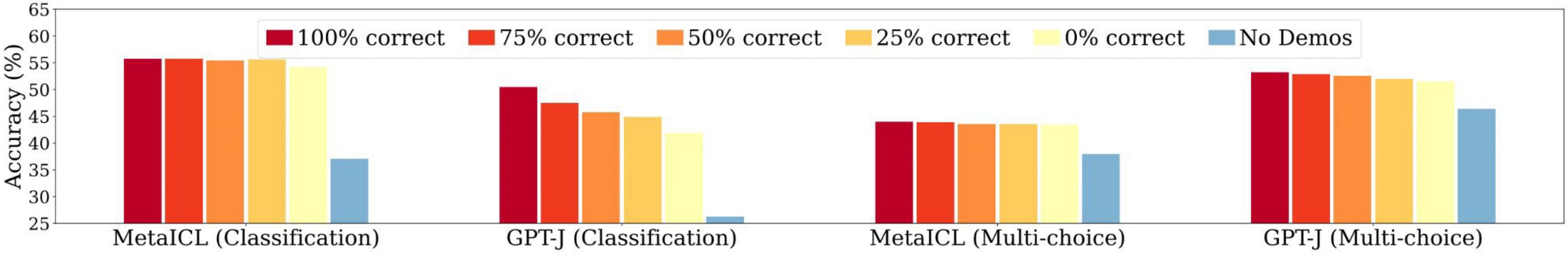
Results



Results of varying number of correct labels in the prompt

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

Results Takeaways

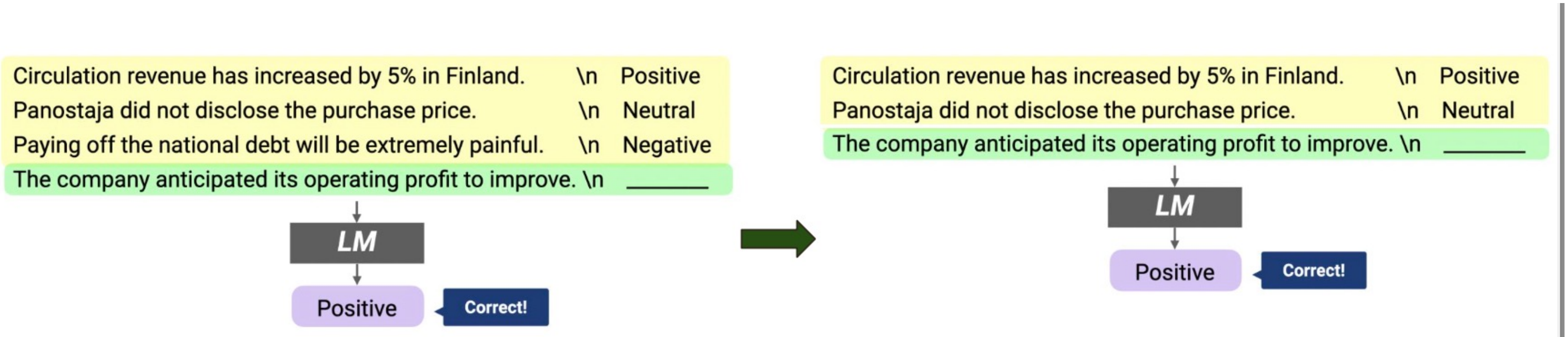


Results of varying number of correct labels in the prompt

Model performance is fairly insensitive to the number of correct labels

Using incorrect labels is better than no examples

Varying the Number of Examples



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
The company anticipated its operating profit to improve. \n ______

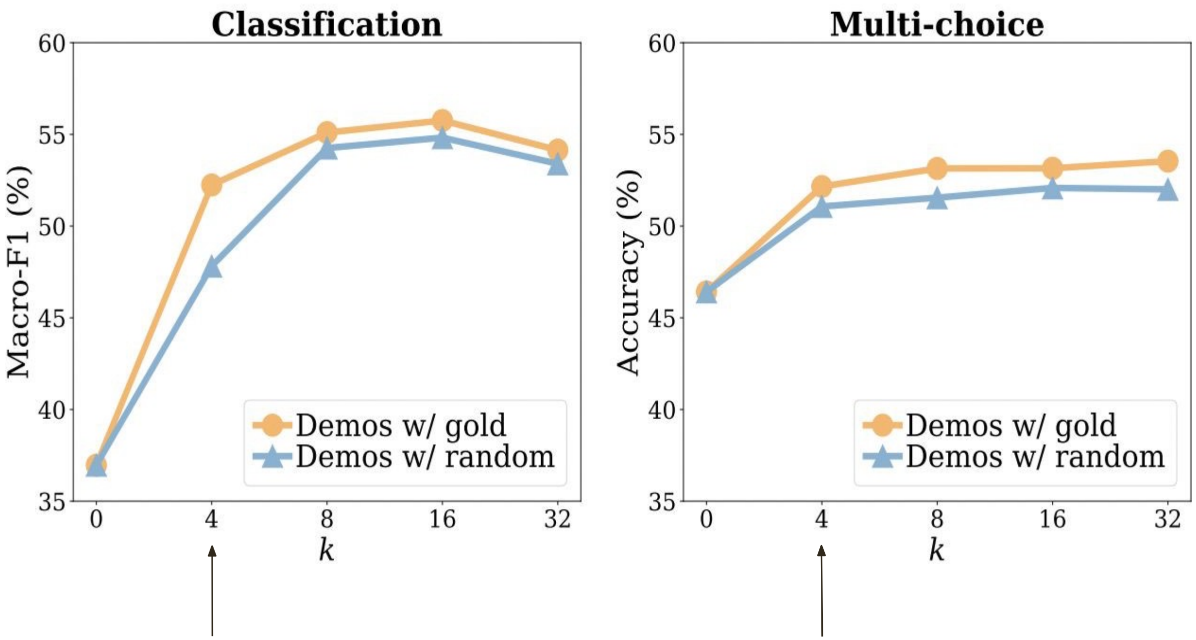
Prompt with three examples

Circulation revenue has increased by 5% in Finland. \n Positive
Panostaja did not disclose the purchase price. \n Neutral
The company anticipated its operating profit to improve. \n ______

Prompt with two examples

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

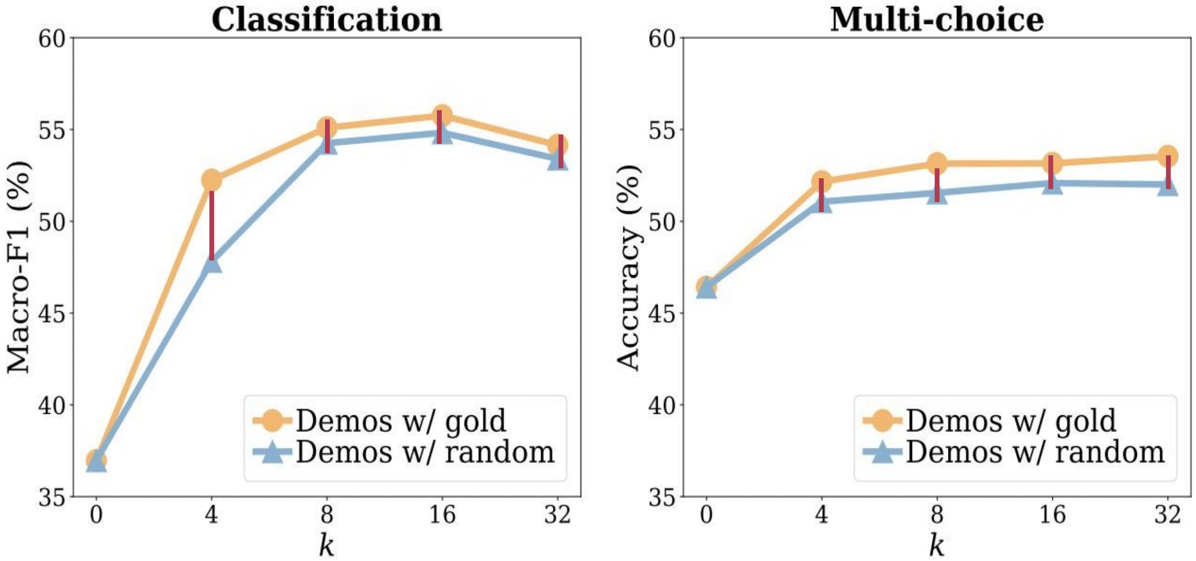
Results



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

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

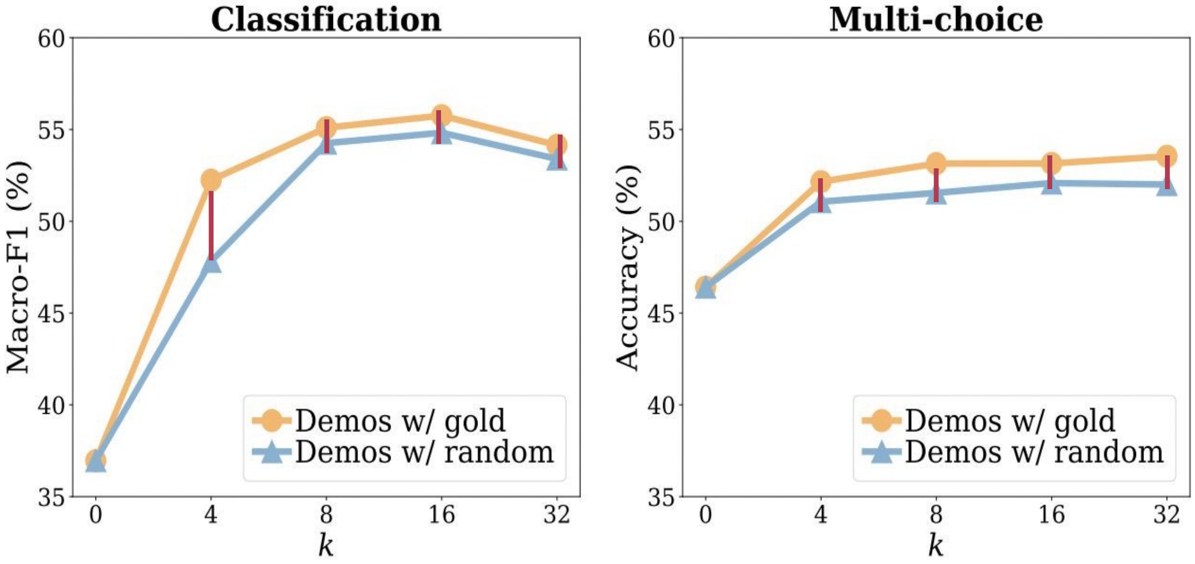
Results



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%

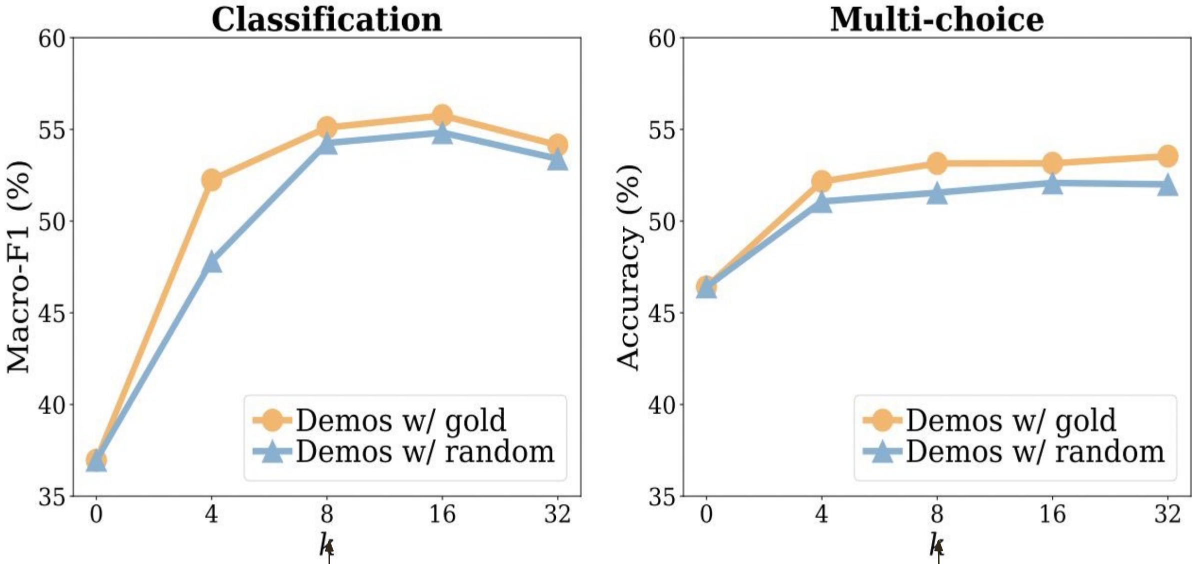
Results Takeaways



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

Performance differences of random labels is consistent across number of examples

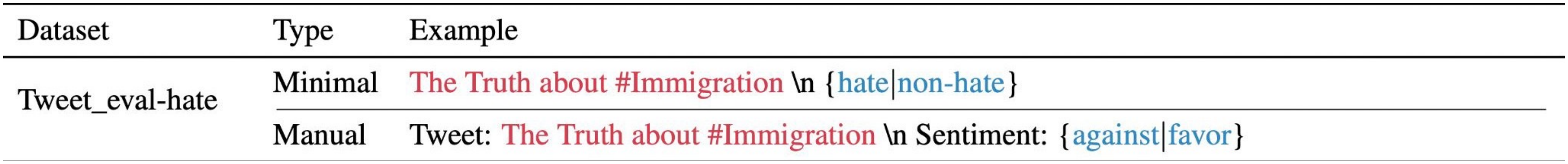
Results Takeaways



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



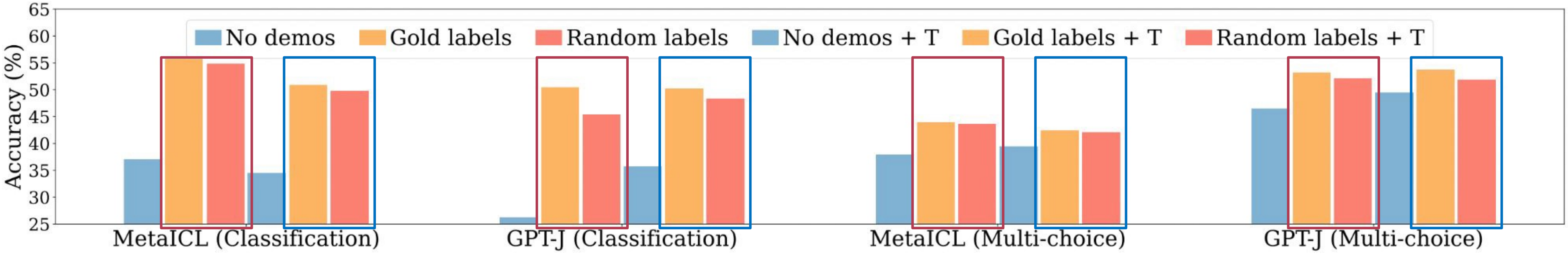
Dataset	Type	Example
Tweet_eval-hate	Minimal	The Truth about #Immigration \n {hate|non-hate}
	Manual	Tweet: The Truth about #Immigration \n Sentiment: {against|favor}

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



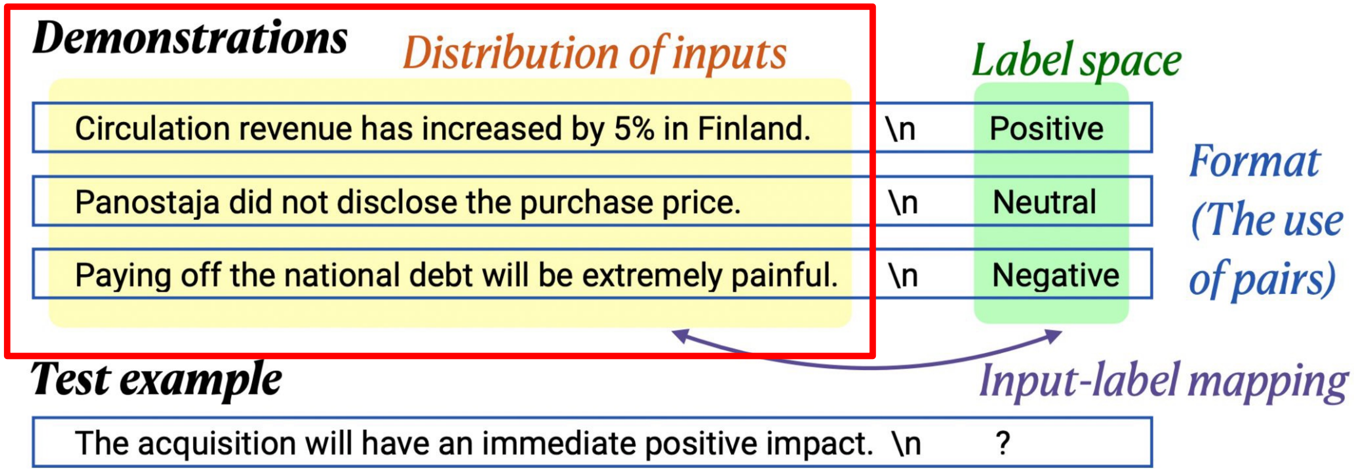
Results with minimal templates and manual templates. '+T' indicates that manual templates are used.

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



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)

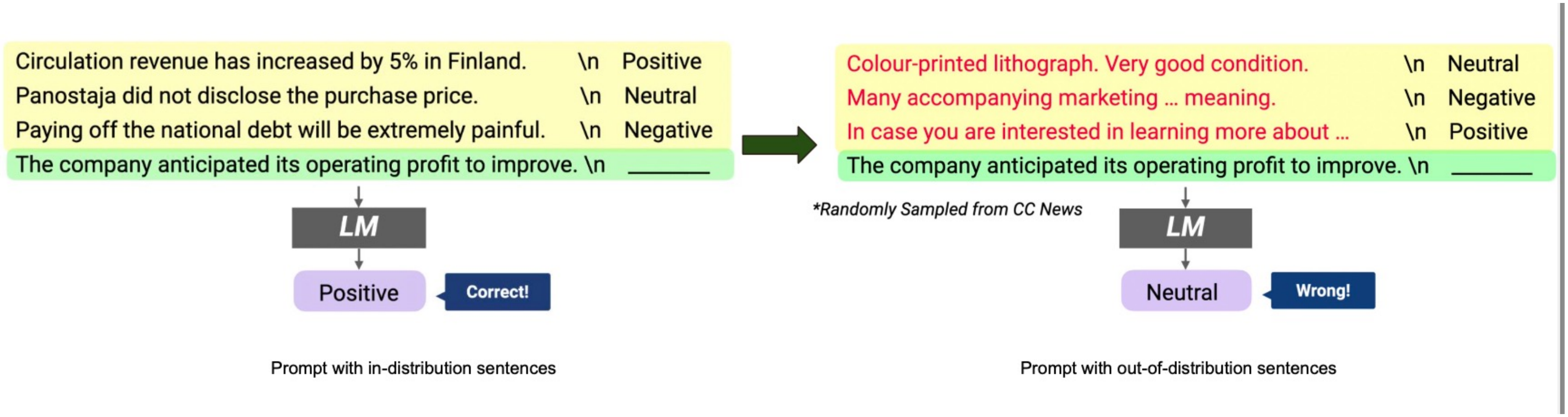
Input-label mapping

Test example

The acquisition will have an immediate positive impact. \n ?

Evaluate the importance of the distribution of inputs

Using out-of-distribution input text



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
The company anticipated its operating profit to improve. \n ______

Prompt with in-distribution sentences

Colour-printed lithograph. Very good condition. \n Neutral
Many accompanying marketing … meaning. \n Negative
In case you are interested in learning more about … \n Positive
The company anticipated its operating profit to improve. \n ______

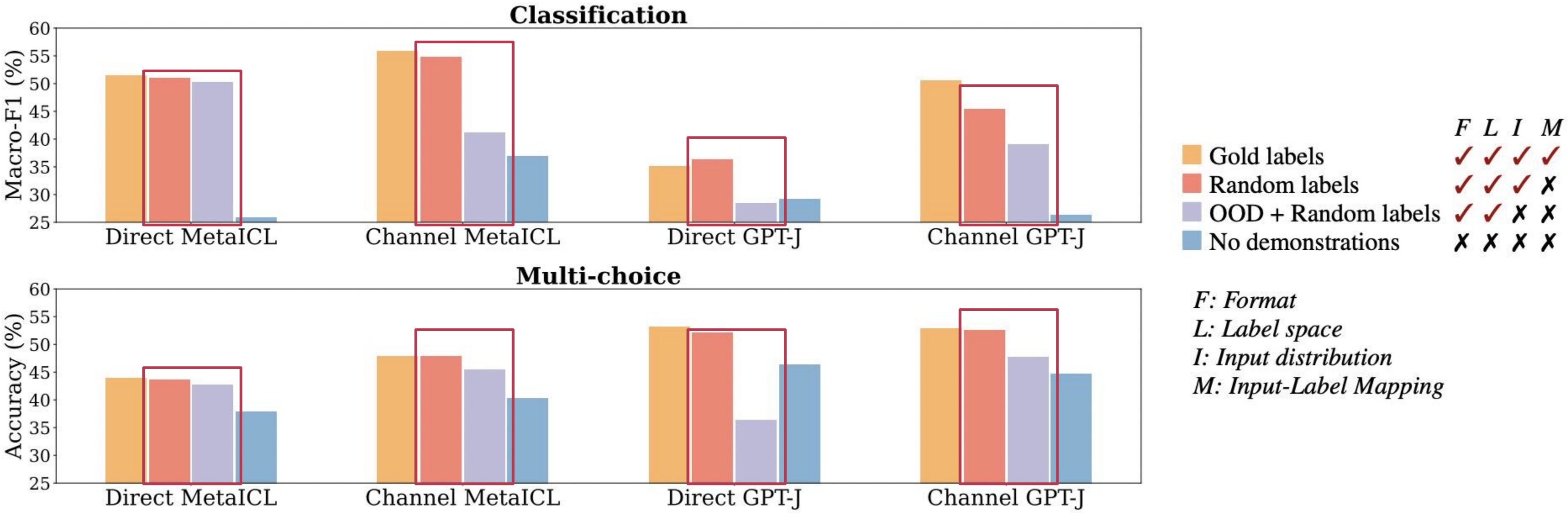
*Randomly Sampled from CC News

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/

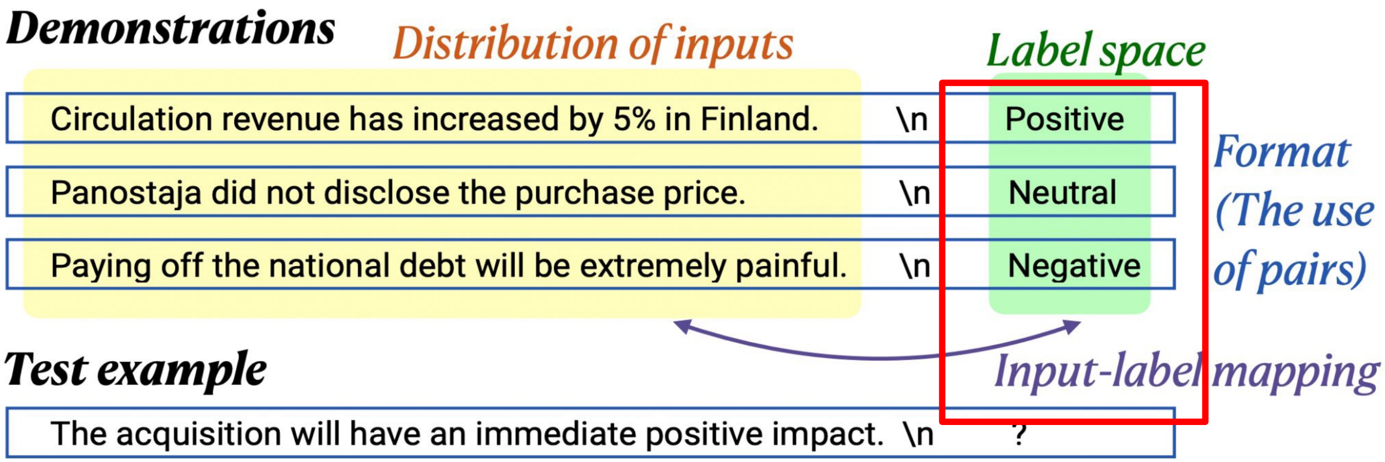
Seeing in-distribution inputs improves performance



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



Demonstrations

Circulation revenue has increased by 5% in Finland. \n Positive

Panostaja did not disclose the purchase price. \n Neutral

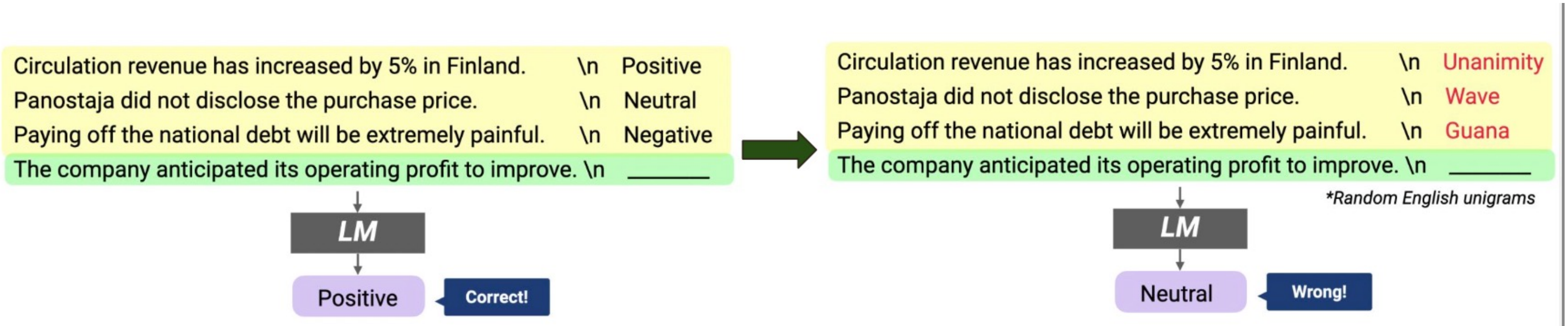
Paying off the national debt will be extremely painful. \n Negative

Test example

The acquisition will have an immediate positive impact. \n ?

Evaluate the importance of the label space

Using random labels from an incorrect 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

The company anticipated its operating profit to improve. \n ______

Circulation revenue has increased by 5% in Finland. \n Unanimity

Panostaja did not disclose the purchase price. \n Wave

Paying off the national debt will be extremely painful. \n Guana

The company anticipated its operating profit to improve. \n ______

**Random English unigrams*

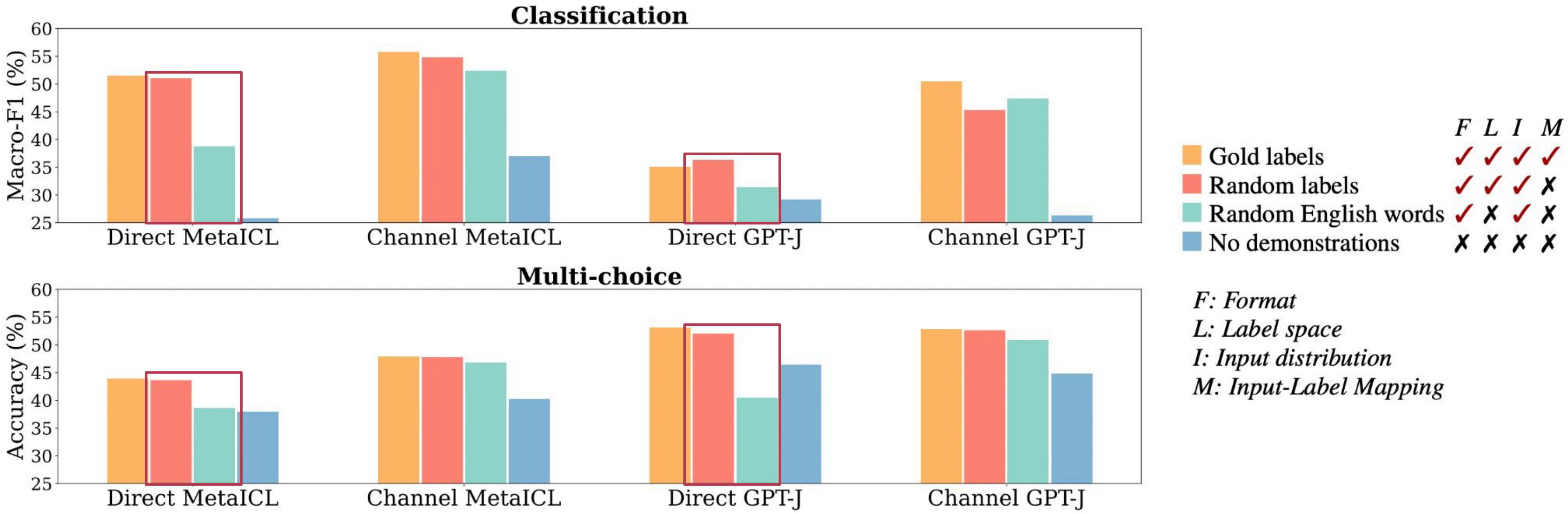
Prompt with true labels

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/

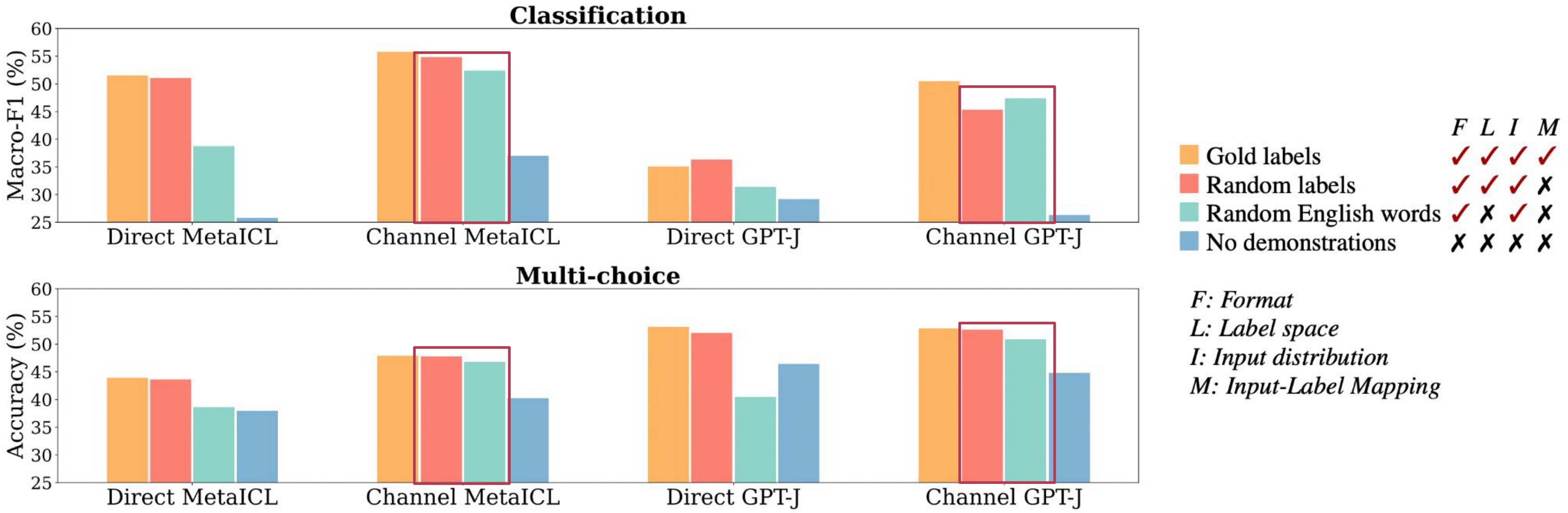
Seeing correct label space is important



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**

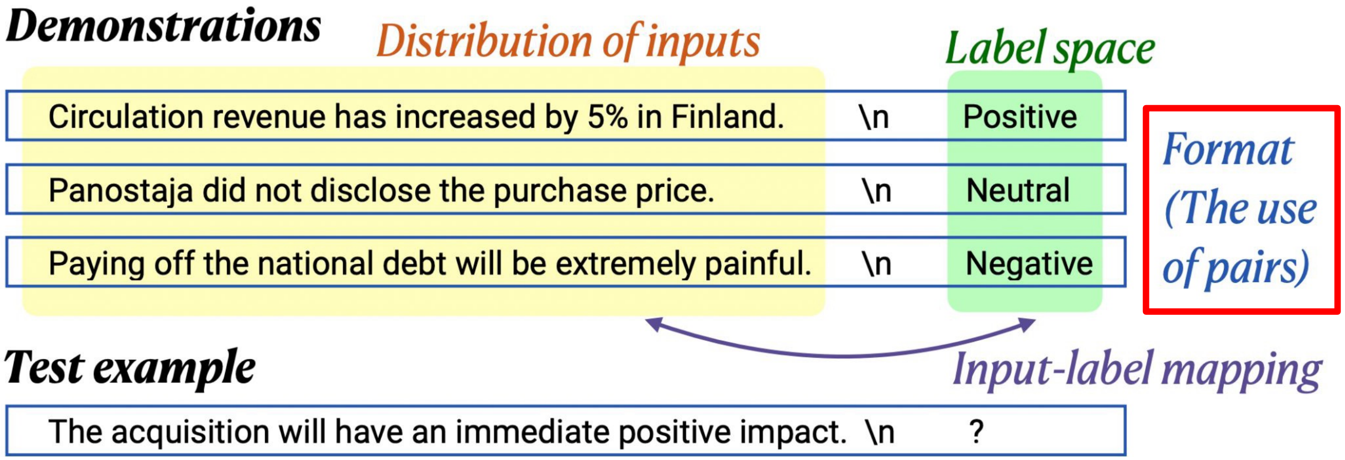
Seeing correct label space is important



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



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)

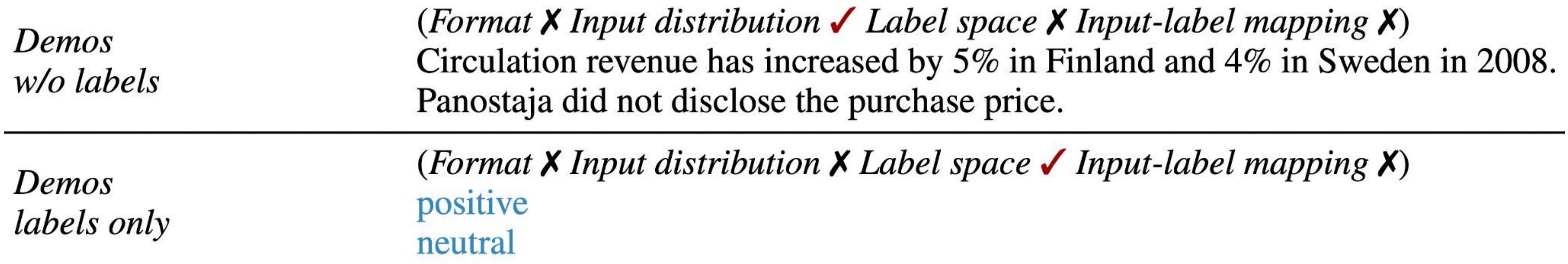
Test example

Input-label mapping

The acquisition will have an immediate positive impact. \n ?

Evaluate the importance of pairing an input sentence with a label

Changing the input-label format



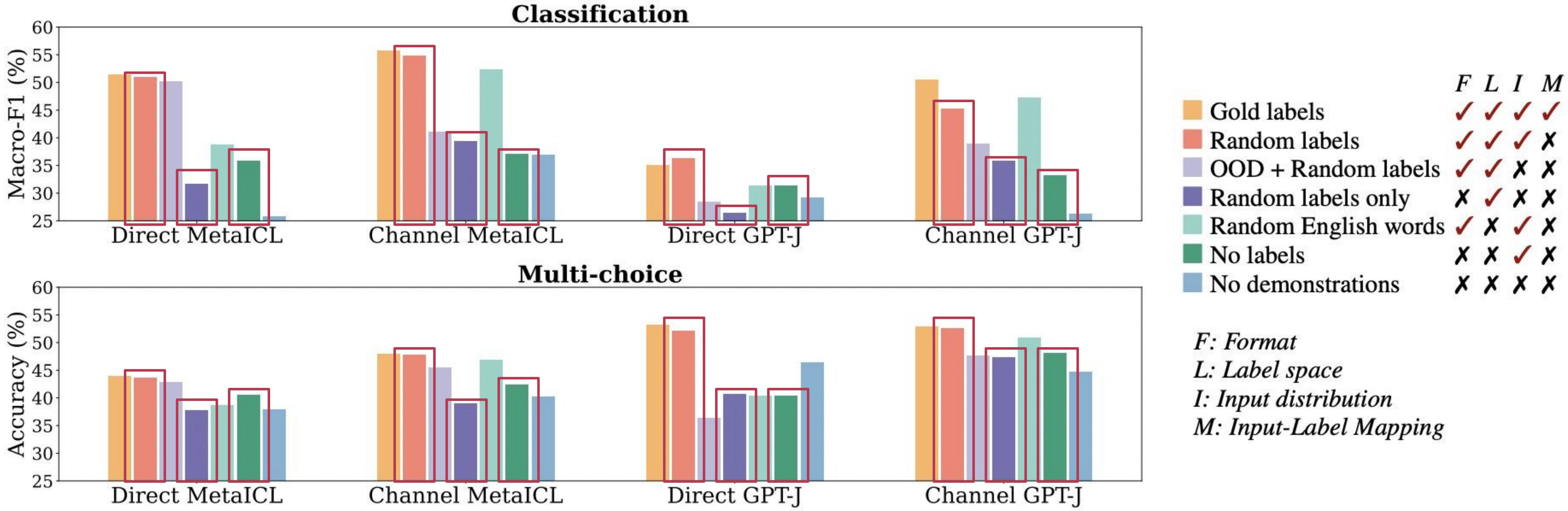
Demos w/o labels	(*Format* ✗ *Input distribution* ✓ *Label space* ✗ *Input-label mapping* ✗) Circulation revenue has increased by 5% in Finland and 4% in Sweden in 2008. Panostaja did not disclose the purchase price.
Demos labels only	(*Format* ✗ *Input distribution* ✗ *Label space* ✓ *Input-label mapping*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

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

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

[redacted]

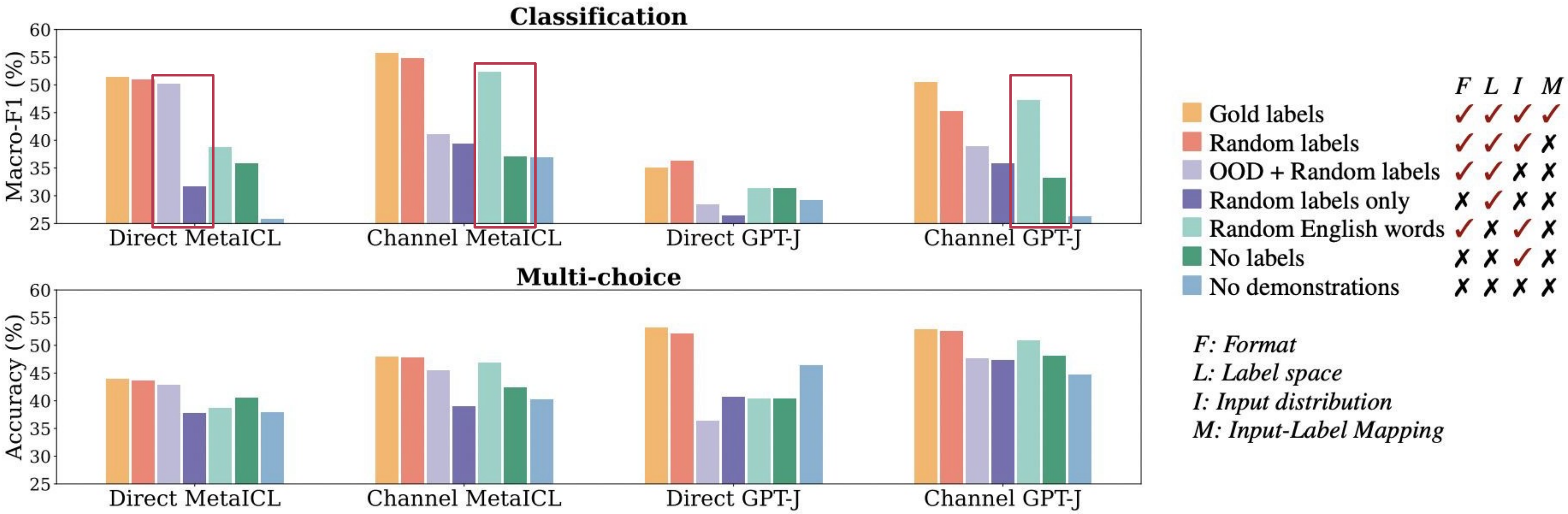
Keeping the input-label format for demonstrations is vital fo



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



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

What are the most surprising findings?

- 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.