What Makes In-Context Learning Work?

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

Presented by :

Group 4 資管碩一 楊鈺翎 / 資管碩一 林慧娟 / 資管三 郭大呈

Content

What is In-context learning?

What is In-context learning?

- "I love this movie!" => Positive
- "This is the worst day ever." => Negative
- "I am very happy today." => Positive

- "The food was terrible and cold." =>

The composition of ICL

simply conditioning(inference) on a few input-label pairs (demonstrations)

- **demonstrations (示例)**

- **input text (X)**
- **label (Y)**

input text (X)	label (Y)
1."I love this movie!"	Positive
2."This is the worst day ever."	Negative
3."It rains a lot today."	Neutral

- **input-label mapping**

- **k=3 the amount of pairs**
- **An input-label pairs**

- **test example**

"The food was terrible and cold.

LM : negative !

Comparing with Fine-tuning

- **Pros**
  - No parameter tuned / gradient updated
  - Much fewer examples
  - Higher versatile（泛用性強)
- **Cons**
  - Less accuracy in specific tasks

We don’t know how models in-context learn

Objective of the paper :

Analyze empirically which aspects of the prompt affect downstream task performance

Experimental Process & Result

Aspect 1

Is the ground truth input-label pairs matter?

Gold Labels vs Randon Labels

- Gold Labels

• input (X)	• label (Y)
1. "I love this movie!"	Positive
2. "This is the worst day ever."	Negative
3. "It rains a lot today."	Neutral

"The food was terrible and cold.

LM : negative !

- Random Labels

• input (X)	• label (Y)
1. "I love this movie!"	Neutral
2. "This is the worst day ever."	Positive
3. "It rains a lot today."	Negative

"The food was terrible and cold.

LM : negative !

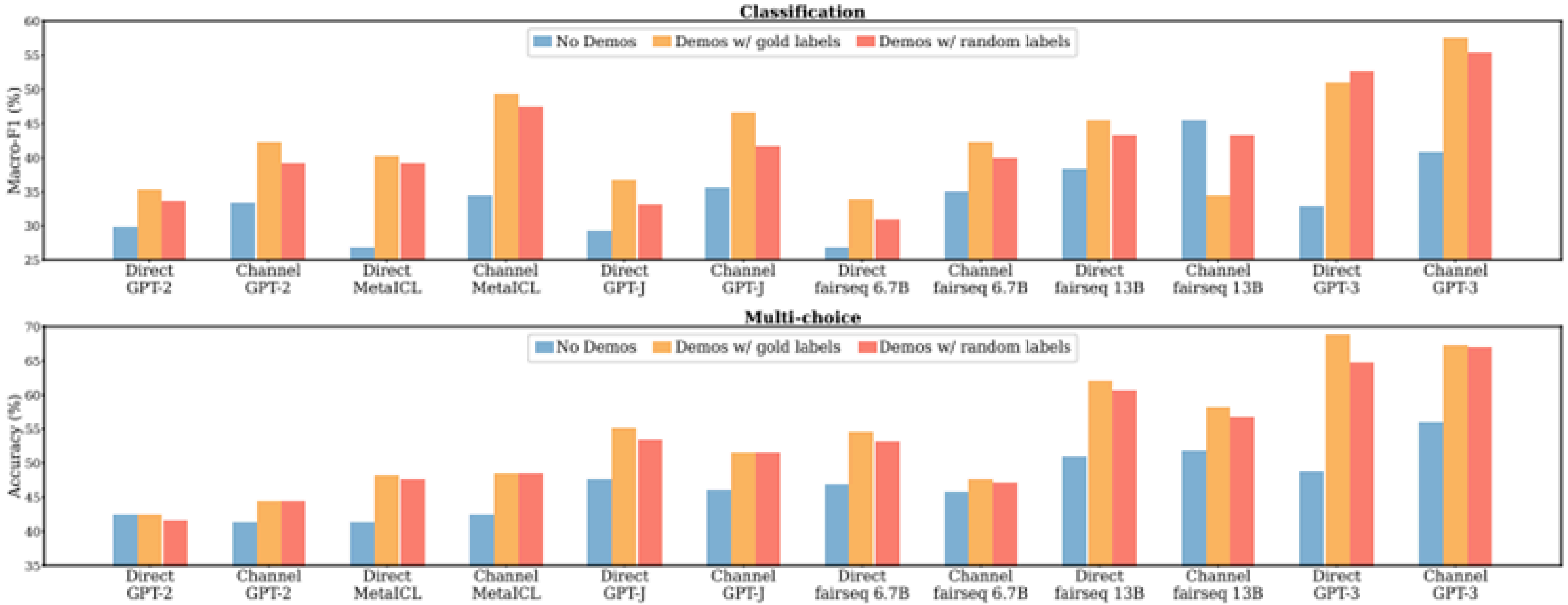
labels are randonly paired with input !

Experimental method:

- using following 3 types of prompt
  1. Demonstrations w/ gold labels (correct label)
  2. Demonstrations w/ random labels (incorrect label)
  3. No demonstrations (zero-shot)

Result

1.demonstrations significantly improves the performance
2.random labels only marginally hurts performance (0–5%)



Takeaways

Is the ground truth input-label pairs matter?

- the ground truth input-label pairs are not necessary to achieve performance gains
- Using incorrect labels is better than no examples

Aspect 2

Is the distribution of the input text matter?

What Is the distribution of input ?

- in-distribution

  - input (X) social media!

    1. "I love this movie!"
    2. "This is the worst day ever."

  - label (Y)

    Positive
    Negative

- test example social media!

  "The food was terrible and cold.

- OOD (Out-of-Distribution)

  - input (X) financial statements

    1. "The stock prices soared after the company's earnings report!"
    2. "Investors are concerned about the ongoing market volatility."

  - label (Y)

    Positive
    Negative

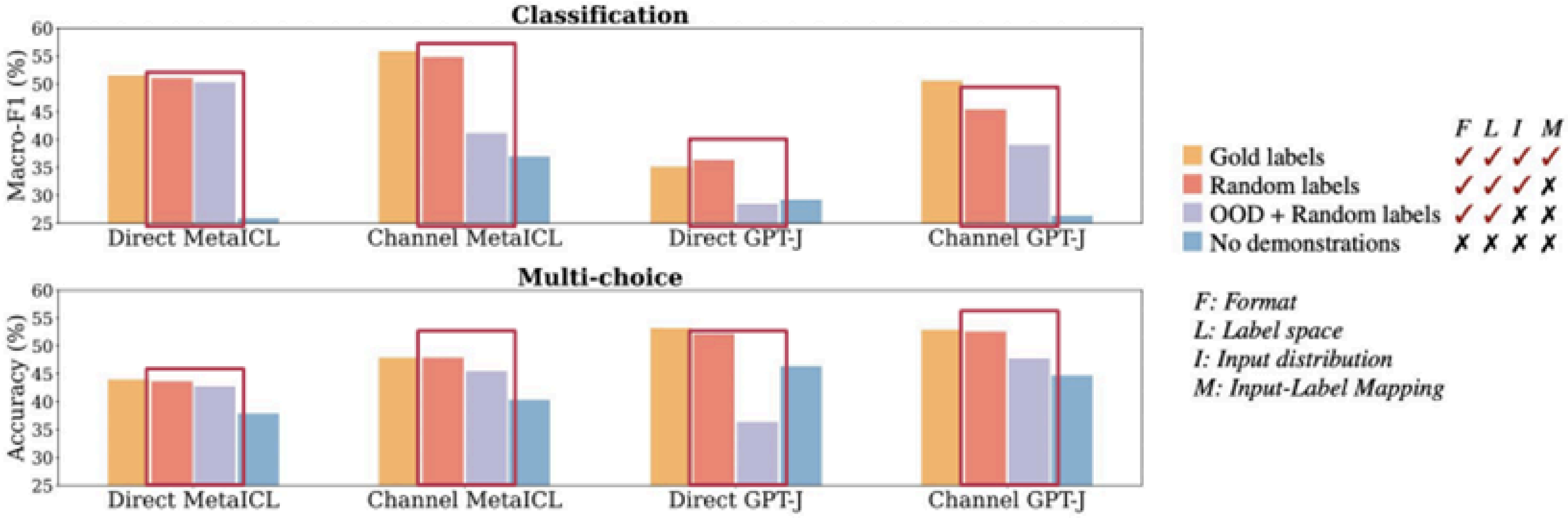
- test example social media!

  "The food was terrible and cold.

- OOD: input sentences are randomly sampled from an external corpus

Result

performance decreases significantly (up to 16%)



Results of using out-of-distribution input sentences

Takeaways

Is better templates matter?

- in-distribution inputs substantially increase performance

Aspect 3

Is label space matter?

What Is label space?

- input (X)

1. "I love this movie!"
2. "This is the worst day ever."
3. "It rains a lot today."

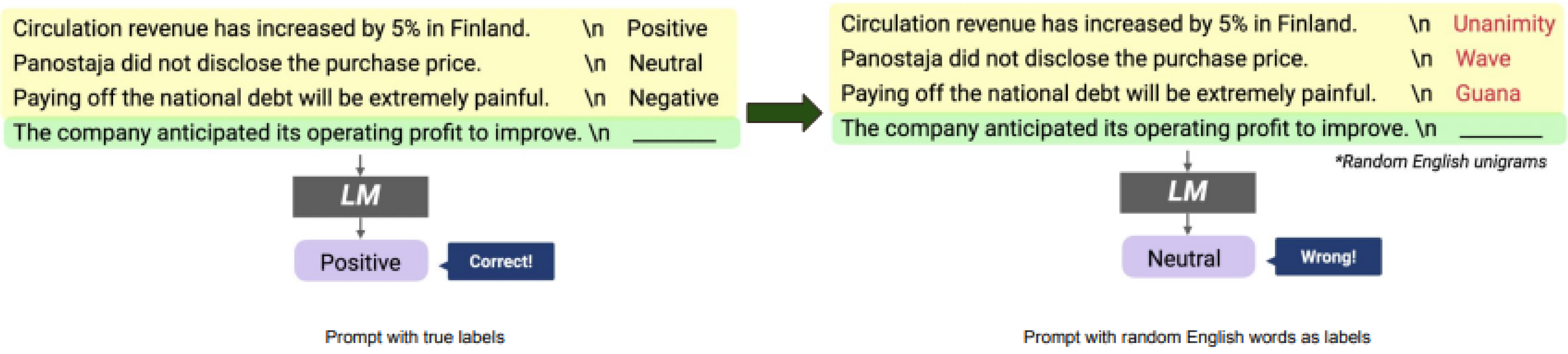
- label (Y)

Positive
Negative
Neutral

- All possible label(Y) options in a specific task
- Sentiment analysis : {positive, negative}
- Classification problem: {finance, science, politic}

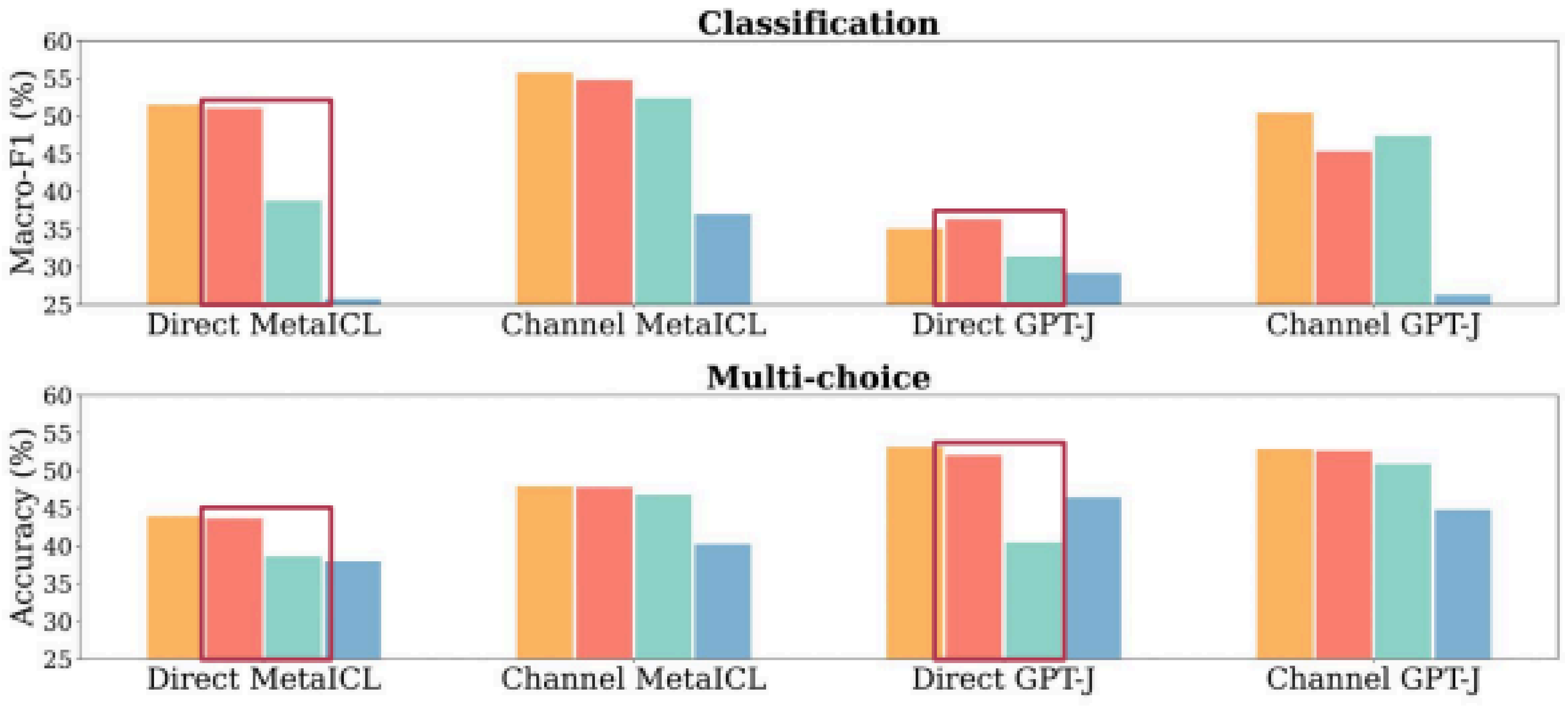
Experimental method:

- Using random labels from an incorrect label space
  1. Demonstrations w/ gold labels (correct label)
  2. Demonstrations w/ random labels (incorrect label)
  3. Demonstrations w/ incorrect label space
  4. No demonstrations (zero-shot)



Result

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



Results of using random English words as labels

Takeaways

Is label space matter?

- correct label space substantially increase performance in direct models

Aspect 4

Is format of prompt matter?

Input-label pairing

- Remove labels(Y)

- input (X)

1. "I love this movie!"
2. "This is the worst day ever."
3. "It rains a lot today."

"The food was terrible and cold.

LM : negative !

- Remove input text(X)

- label (Y)

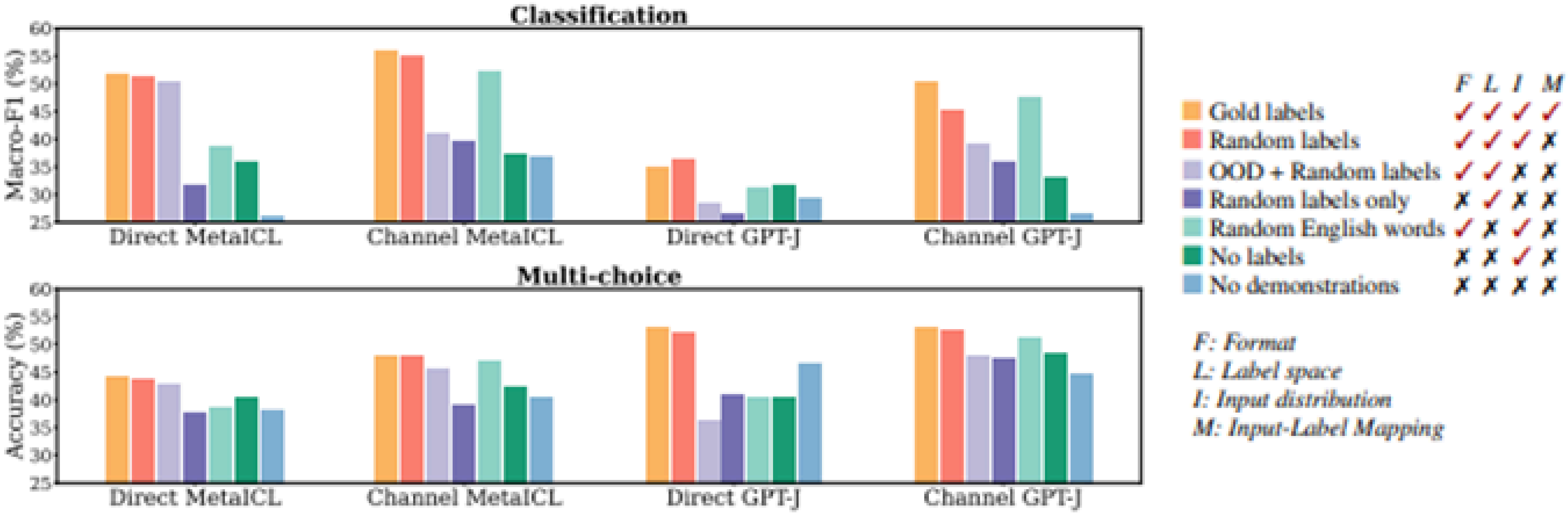
Positive
Negative
Neutral

"The food was terrible and cold.

LM : negative !

Result

- Not using the input-label format decreases performance
- Using OOD inputs and random English words as labels is better than only keeping one part of the format or having no demonstrations



Takeaways

Is the format matter?

- The foramt of the prompt significantly increase performance

Conclusion

What Makes ICL Work?

- the vital factors:

1. in-distribution input txet
2. label space
3. Input-label pairing format

- not really matter factor:

1. the number of correct labels

Q / A

Questions

- Provide one advantage of In-context Learning over Fine-tuning
- When we do not know the answer (label) of the input text, randomly assigning it a label within the correct label space will result in better or worse performance compared to not using demonstrations?