Session #7:
In-Context Learning:

Tuesday, Sept 20

CSCI 601.771: Self-supervised Statistical Models

JOHNS HOPKINS UNIVERSITY

This paper is really good at ___ but fails to address ___

Rethinking the Role of Demonstrations: What makes In-context learning Work?

Flow of the presentation:–

- Background (MetaICL, Noisy, and direct inference in brief)
- What is the article trying to answer?
- Design of the Experiments (What is their approach?)
- Results
- Discussion and Conclusions

: Ammar and Karan

Background

- What is Meta learning?

	Meta-training	Inference
Task	C *meta-training* tasks	An unseen *target* task
Data given	Training examples $\mathcal{T}_i = \{(x_j^i, y_j^i)\}_{j=1}^{N_i},\ \forall i \in [1, C] \quad (N_i \gg k)$	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 $P(y_{k+1} \mid x_1, y_1, \cdots, x_k, y_k, x_{k+1})$	$\text{argmax}_{c \in \mathcal{C}} P(c \mid x_1, y_1, \cdots, x_k, y_k, x)$

Table 1: Overview of MetaICL (Section 3). MetaICL uses the same in-context learning setup at both meta-training and inference. At meta-training time, $k+1$ examples for a task is sampled, where the last example acts as the test example and the rest k examples act as the training examples. Inference is the same as typical in-context learning where k labeled examples are used to make a prediction for a test input.

: Ammar and Karan

Background

- What is Noisy/Channel Vs Direct Inference(what we normally do)?

3.3 Channel MetaICL

We introduce a noisy channel variant of MetaICL called Channel MetaICL, following Min et al. (2022). In the noisy channel model, $P(y|x)$ is reparameterized to $\frac{P(x|y)P(y)}{P(x)} \propto P(x|y)P(y)$. We follow Min et al. (2022) in using $P(y) = \frac{1}{|\mathcal{C}|}$ and modeling $P(x|y)$ which allows us to use the channel approach by simply flipping x_i and y_i. Specifically, at meta-training time, the model is given a concatenation of $y_1, x_1, \cdots, y_k, x_k, y_{k+1}$ and is trained to generate x_{k+1}. At inference, the model computes $\mathrm{argmax}_{c \in \mathcal{C}} P(x|y_1, x_1, \cdots, y_k, x_k, c)$.

: Ammar and Karan

What is article trying to answer?

- The article is trying to empirically find the the importance of demonstrations (conditions) for in-context learning

1. **The input-label mapping**, i.e., whether each input x_i is paired with a correct label y_i.
2. **The distribution of the input text**, i.e., the underlying distribution that $x_1 \dots x_k$ are from.
3. **The label space**, i.e., the space covered by $y_1 \dots y_k$.
4. **The format**—specifically, the use of input-label pairing as the format.[7]

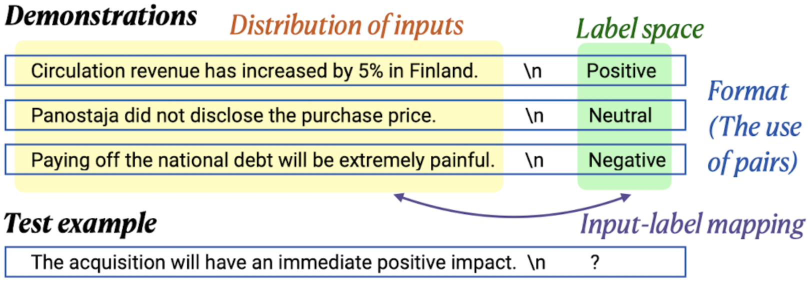


Figure 7: Four different aspects in the demonstrations: the input-label mapping, the distribution of the input text, the label space, and the use of input-label pairing as the format of the demonstrations.

: Ammar and Karan

Design of the Experiments

- 6 LMs (12 decoder-based models, which are a variation of the 6 LMs)
- Different number of tasks specific to experiment
- Accuracies averaged locally and grouped (classification and multi-choice)
- Main experiment Ideas

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‡	✗	✗

Demos w/ gold labels	(*Format* ✓ *Input distribution* ✓ *Label space* ✓ *Input-label mapping* ✓) Circulation revenue has increased by 5% in Finland and 4% in Sweden in 2008. \n positive Panostaja did not disclose the purchase price. \n neutral
Demos w/ random labels	(*Format* ✓ *Input distribution* ✓ *Label space* ✓ *Input-label mapping* ✗) Circulation revenue has increased by 5% in Finland and 4% in Sweden in 2008. \n neutral Panostaja did not disclose the purchase price. \n negative
OOD Demos w/ random labels	(*Format* ✓ *Input distribution* ✗ *Label space* ✓ *Input-label mapping* ✗) Colour-printed lithograph. Very good condition. Image size: 15 x 23 1/2 inches. \n neutral Many accompanying marketing claims of cannabis products are often well-meaning. \n negative
Demos w/ random English words	(*Format* ✓ *Input distribution* ✓ *Label space* ✗ *Input-label mapping* ✗) Circulation revenue has increased by 5% in Finland and 4% in Sweden in 2008. \n unanimity Panostaja did not disclose the purchase price. \n wave
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

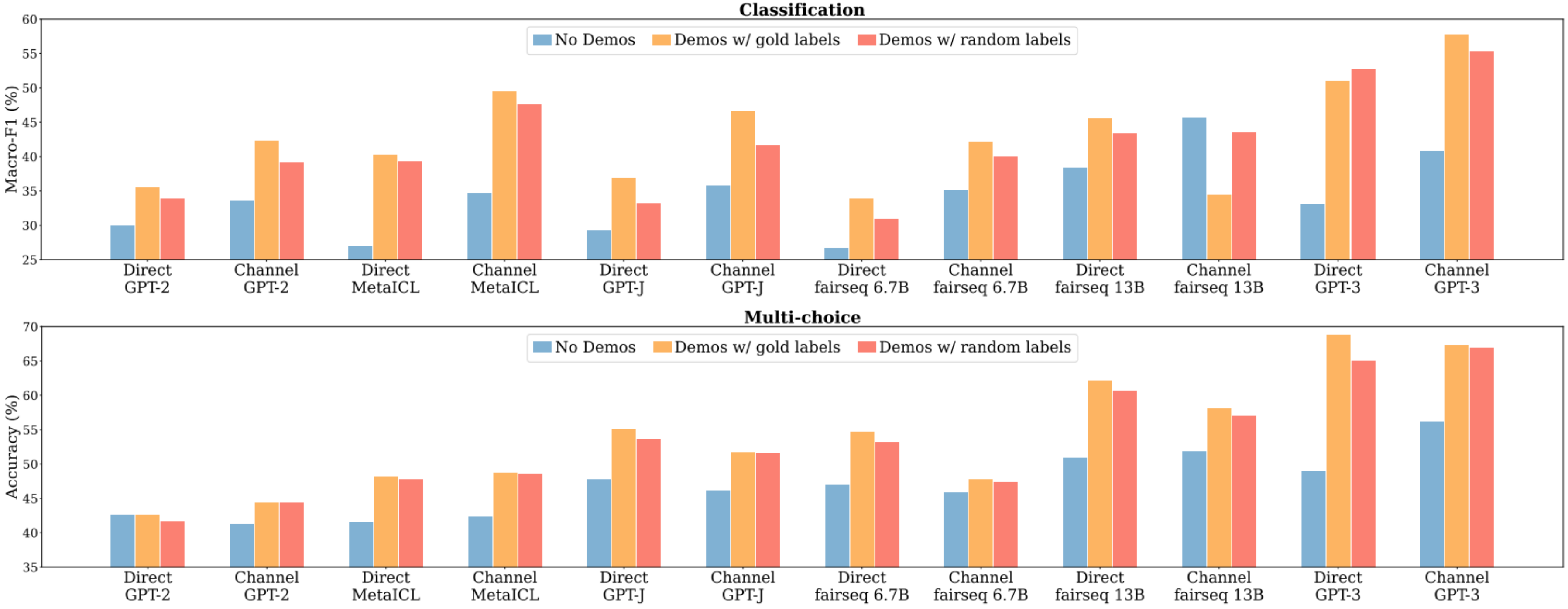
Table 4: Example demonstrations when using methods in Section 5. The financial_phrasebank dataset with $\mathcal{C}$ = {"positive", "neutral", "negative"} is used. Red text indicates the text is sampled from an external corpus; blue text indicates the labels are randomly sampled from the label set; purple text indicates a random English word.

: Ammar and Karan

Results

✍ : Ammar and Karan

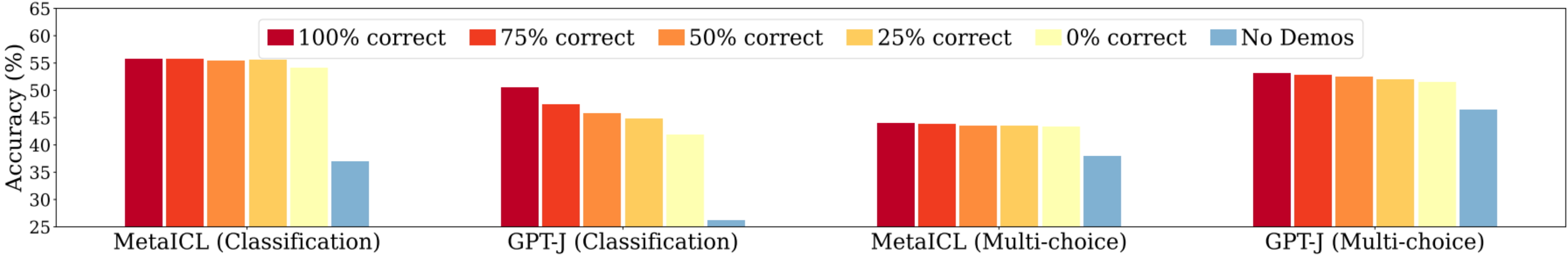
Ground Truth Matters Little



Replacing gold labels with random labels only marginally hurts performance

: Ammar and Karan

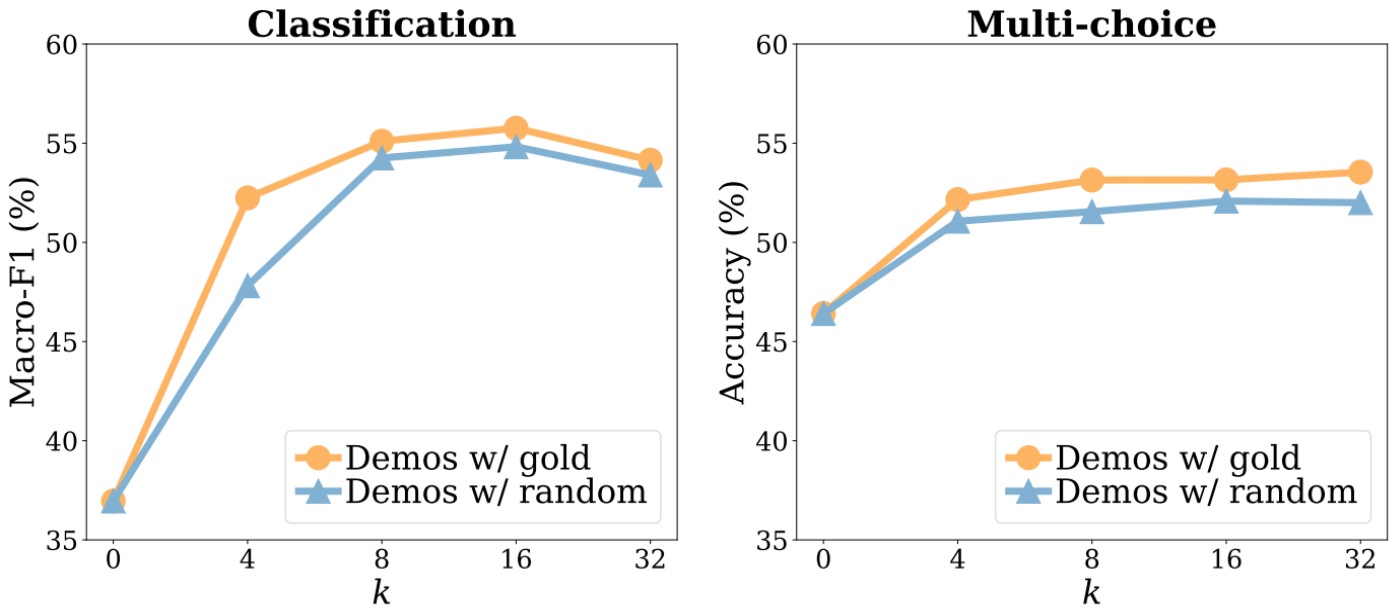
Does the Number of Correct Labels Matter?



Model performance is fairly insensitive to the number of correct labels in the demonstrations

: Ammar and Karan

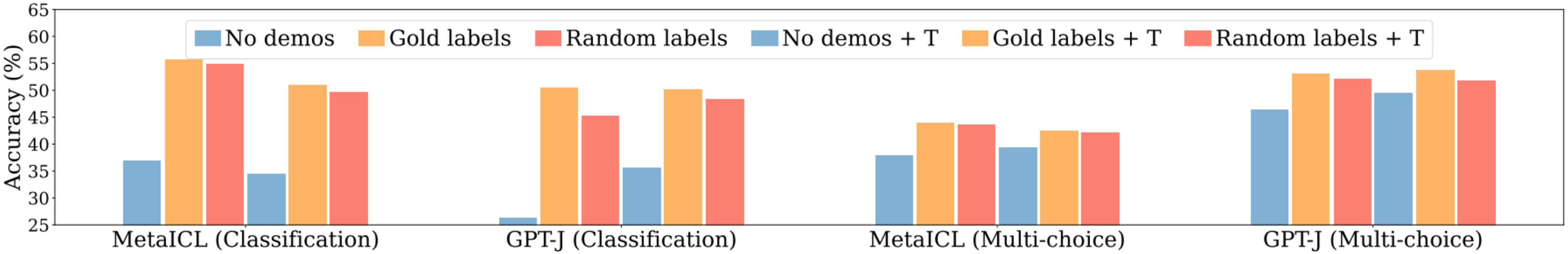
Does Varying K Matter?



Model performance does not increase much as k increases when k ≥ 8, both with gold labels and with random labels

: Ammar and Karan

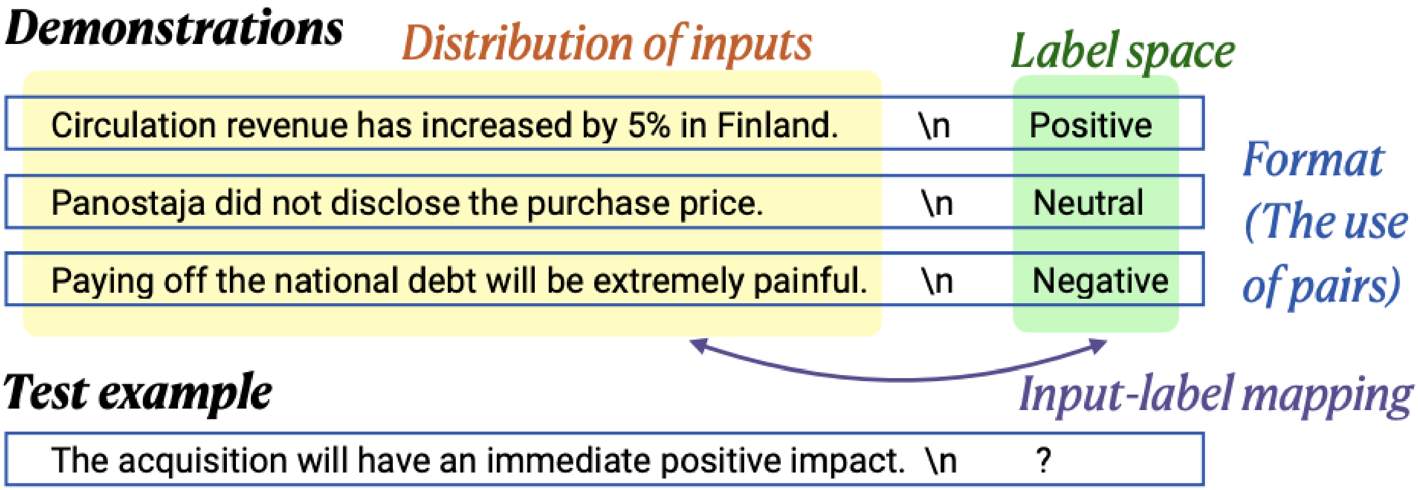
Is the result consistent with better templates?



The trend—replacing gold labels with random labels barely hurting performance—holds with manual templates

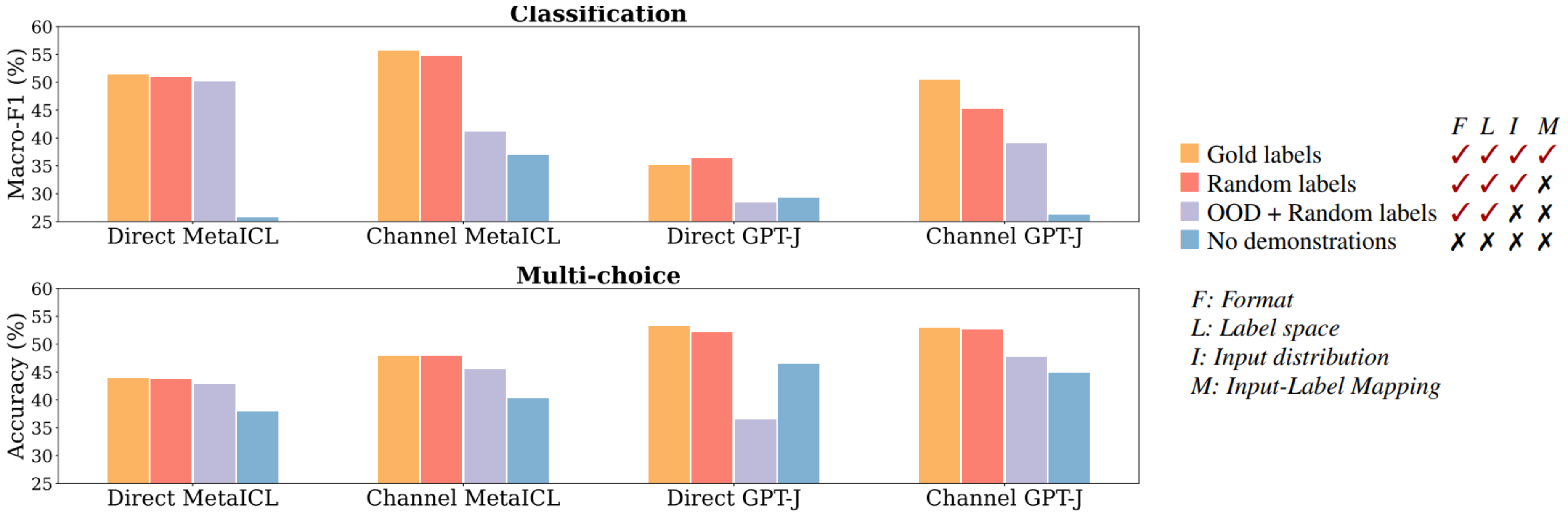
: Ammar and Karan

Why does in-context learning work?



: Ammar and Karan

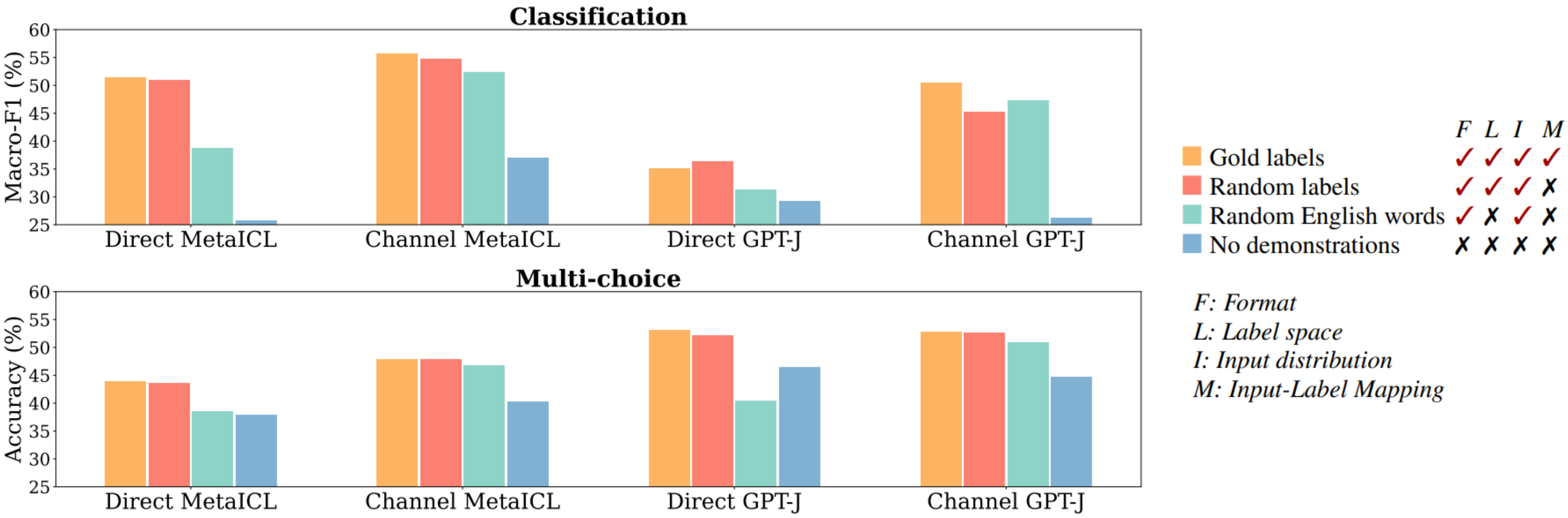
Impact of the distribution of the input text



In-distribution inputs in the demonstrations substantially contribute to performance gains

: Ammar and Karan

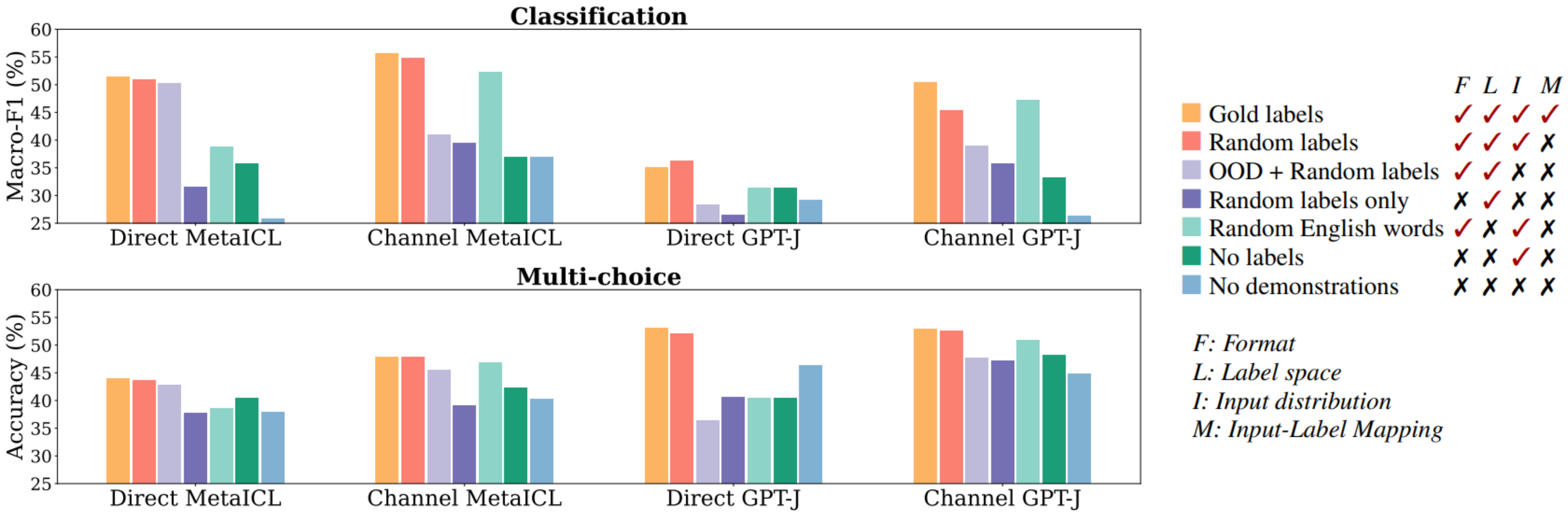
Impact of the label space



Conditioning on the label space seems to significantly contribute to performance gains only for direct models.

: Ammar and Karan

Impact of the use of input-label pairing



Keeping the format of the input-label pairs is the key.

: Ammar and Karan

Discussion and Conclusion

- Three main conclusions
  - Accuracy gains are mainly coming from independent specification of the input space and the label space
  - The models can still retain up to 95% of performance gains by using either the inputs only or the label set only if the right format is used
  - meta-training with an in-context learning objective magnifies these trends (as it forces the model to exploit simpler details in the demonstrations)

Discussion and Conclusion

- Discussions
  - Does the model learn at Test time? (Not strictly, it learns a simple map for a task)
  - Capacity of LMs. (demonstrations are for task location and the intrinsic ability to perform the task is obtained at pretraining time)
  - Just using random input-label pairs as demonstrations improve performance. Does this mean the model has a higher zero shot learning capacity then we thought?

: Ammar and Karan

Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?

Sewon Min[1,2] Xinxi Lyu[1] Ari Holtzman[1] Mikel Artetxe[2]
Mike Lewis[2] Hannaneh Hajishirzi[1,3] Luke Zettlemoyer[1,2]
[1]University of Washington [2]Meta AI [3]Allen Institute for AI
{sewon,alrope,ahai,hannaneh,lsz}@cs.washington.edu
{artetxe,mikelewis}@fb.com

A review of the Paper

: Ayo Ajayi, Elisee Djapa, Yongrui Qi

Paper Summary

- How in-context learning works
- Demonstration
  - Label space
  - Input distribution
  - Sequence format
- Main results show that input label mapping is not important to in-context learning.

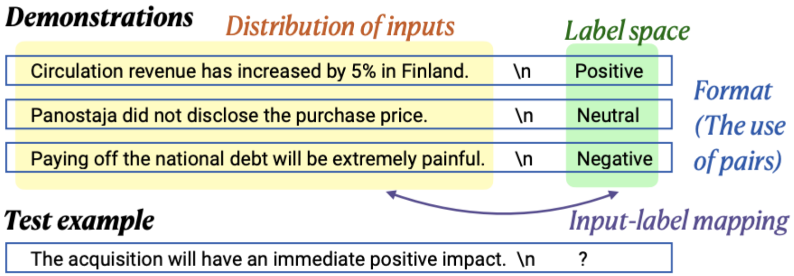


Figure 7: Four different aspects in the demonstrations: the input-label mapping, the distribution of the input text, the label space, and the use of input-label pairing as the format of the demonstrations.

Positives

- ✓ Empirical paper
  - ✓ Explains how and why in-context learning works.
    - ✓ Could be useful for anyone trying to incorporate in-context learning within their LM.
  - ✓ Lots of background research.
- ✓ Explains how and why datasets were chosen.
  - ✓ Diverse datasets (covers topics on science, social media, and finance) supported by GLUE and SuperGLUE benchmarks.
- ✓ Low resource data sets, good for reproducibility (size of datasets)
- ✓ Capacity of Large Model

Negatives

- ✓ Why use decoder-only models?
- ✓ Lack of explanation in methodology: why direct and channel methods?
- ✓ What exactly is meta learning and its impact?
- ✓ How does demonstration contribute to models?

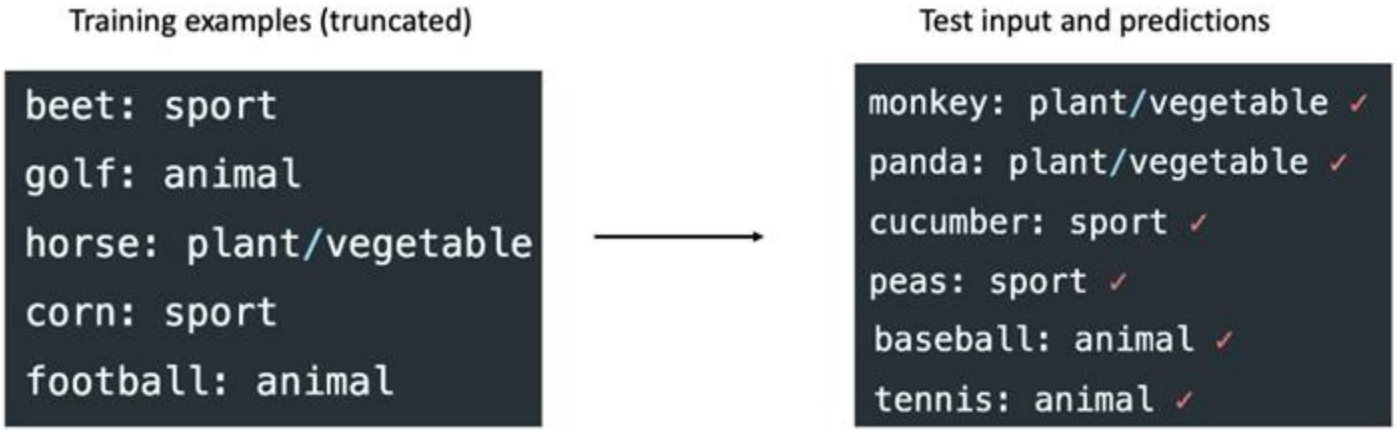
Visionary: Conclusion

- In-context learning doesn't depend on the association between input and annotation, but the ability to activate pre-trained models by presenting them in the form of data.

- Proved that model does not rely on the ground truth input-label mapping as much as we thought (replacing the labels in demonstrations with random labels barely hurts performance)

: Haoyue, Fadil

Unnatural In-context learning

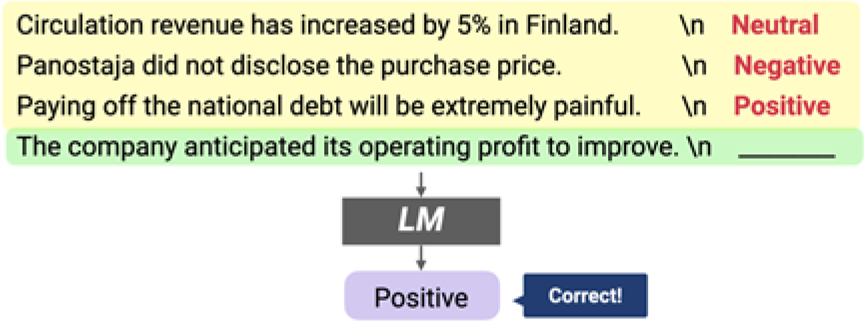


Model is “localizing” or “retrieving” concepts that it has learned during pretraining, thus it can perform unnatural/unseen synthetic task with in-context learning.

: Haoyue, Fadil

Zero-shot performance improvement

- Possible to achieve nearly k-shot performance without using any labeled data!
- You can simply pairing each unlabeled input with a **random label** and using it as the demonstrations.
- Raise zero-shot baseline level up to 20% absolute in classification and up to 15% absolute in multi-choice tasks.

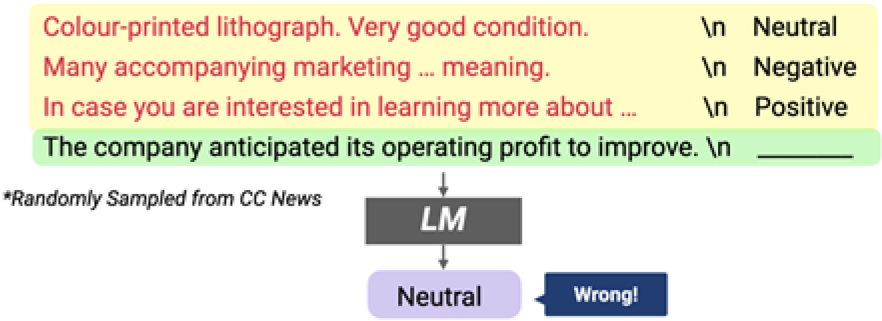


: Haoyue, Fadil

Watch Out!

- Prompts randomly sampled from an **external corpus are detriment** to the model
- Need to care about the choice of demonstrations. (**same/close input distribution**)

External Corpus



Random Unigrams

: Haoyue, Fadil

Complementary Work - Instruction following model

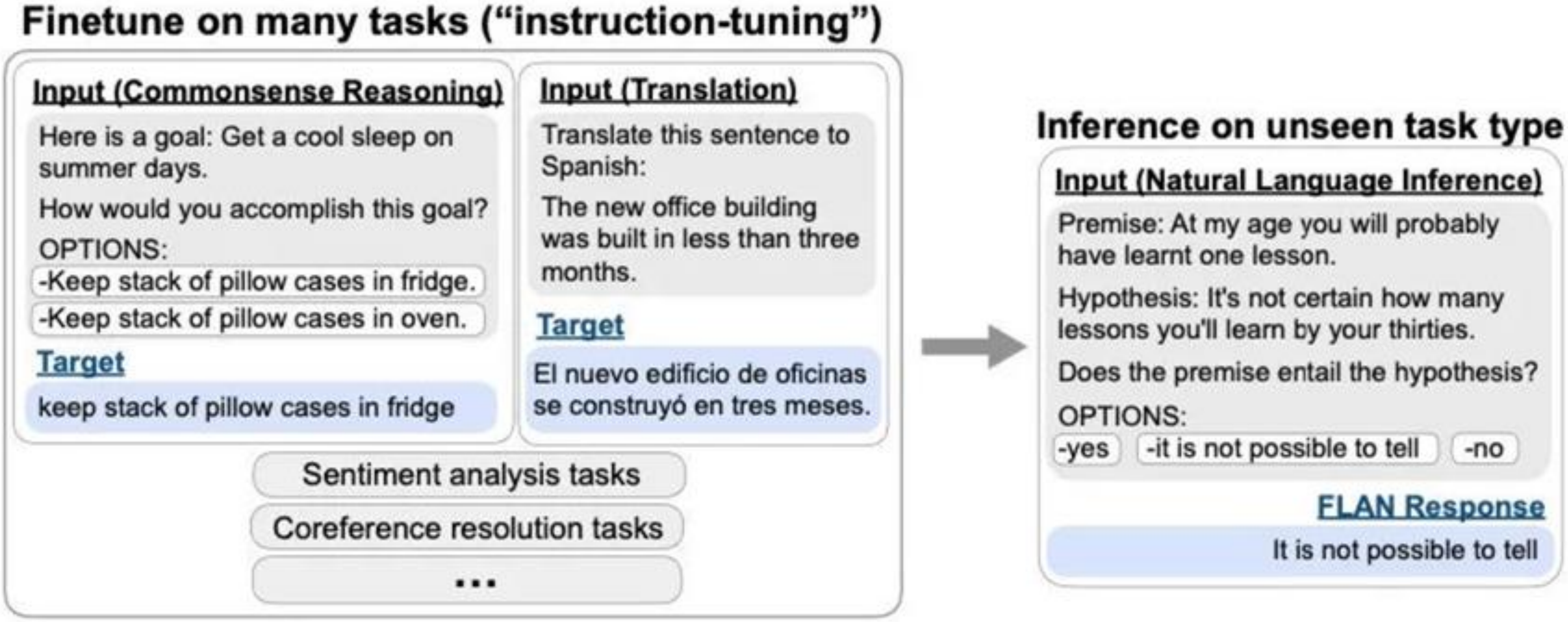
- *Multitask Prompted Training Enables Zero-Shot Task Generalization*, 2022 ICLR

- A total of 171 multitasking datasets were collected and a total of 1939 prompts were created, with an average of 11.3 prompts per dataset.

- Multi-task learning based on datasets containing **instruction prompt** (prompt form is more like an explicit command/instruction)

: Haoyue, Fadil

Complementary Work - Instruction following model



: Haoyue, Fadil

Complementary Work - Instruction following model

Why important?

The demonstrations and instructions largely have the same role to LMs, and the author hypothesizes that the findings hold for instruction-following models.

Why hypothesize is true?

The instructions prompt the model to recover the capacity it already has, but do not supervise the model to learn novel task semantics.

: Haoyue, Fadil

Future Work

- Investigate how does model scale, training objective, and architecture affect the model behavior during in-context learning.

The author also post other related works, experiments are overlapped

- *Noisy Channel Language Model Prompting for Few-Shot Text Classification*
- *MetaICL: Learning to Learn In Context*

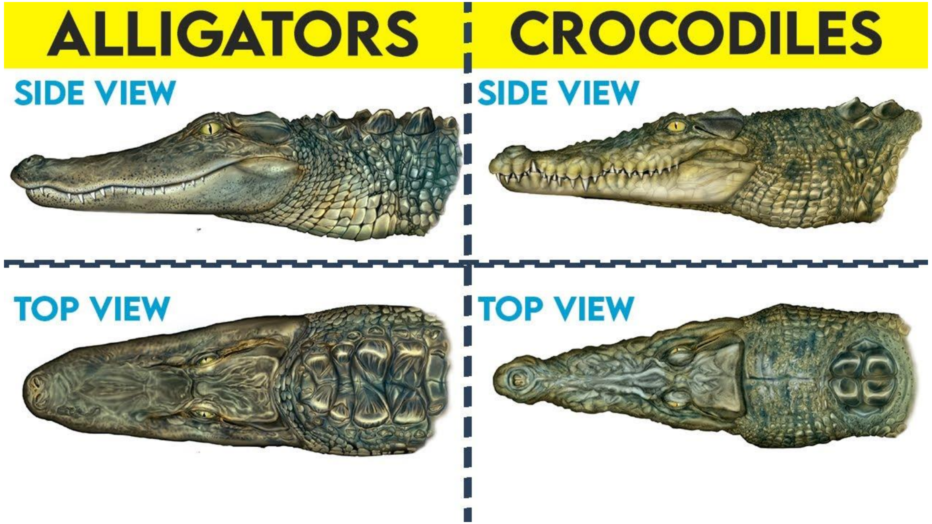
: Haoyue, Fadil

Using the methodology for FSL (And perhaps OSL?)

- Computer Vision
- Robotics
- Audio processing

: Haoyue, Fadil

Alligator vs Crocodile



What is this?