Paper Presentation

Renze Lou

03/01/2023

Pls feel free to interrupt me.

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

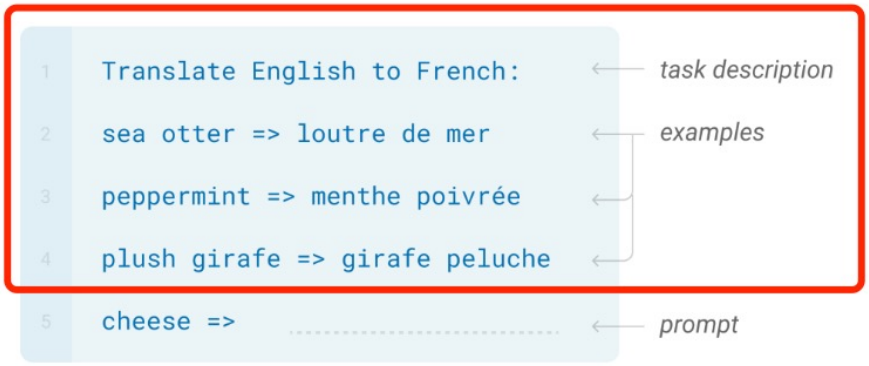
Sewon Min[1,2] **Xinxi Lyu**[1] **Ari Holtzman**[1] **Mikel Artetxe**[2]
Mike Lewis[2] **Hannaneh Hajishirzi**[1,3] **Luke Zettlemoyer**[1,2]
[1]University of Washington [2]Meta AI [3]Allen Institute for AI
{sewon,alrope,ahai,hannaneh,lsz}@cs.washington.edu
{artetxe,mikelewis}@fb.com

To clear up three questions:

- ***What*** is in-context learning?
- ***Which*** aspects contribute to performance?
- ***Why*** A instead of B?

What is in-context learning?

vs.



Supervised Fine-tuning (SFT)

- *Few-shot examples are used in **training**.*
- *Gradient update.*

In-context learning

- *Few-shot examples are used in **testing**.*
- ***No** gradient update.*

What is in-context learning?

GPT-3 has 1.3B parameters!

Fine-tuning such a **big guy** is **unaffordable** (for most of us)!

In-context learning is more **versatile** and **practical!**

Which aspects contribute to performance?

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 $x_1 ... x_k$ are from.
3. **The label space**, i.e., the space covered by $y_1 ... y_k$.
4. **The format**—specifically, the use of input-label pairing as the format.[7]

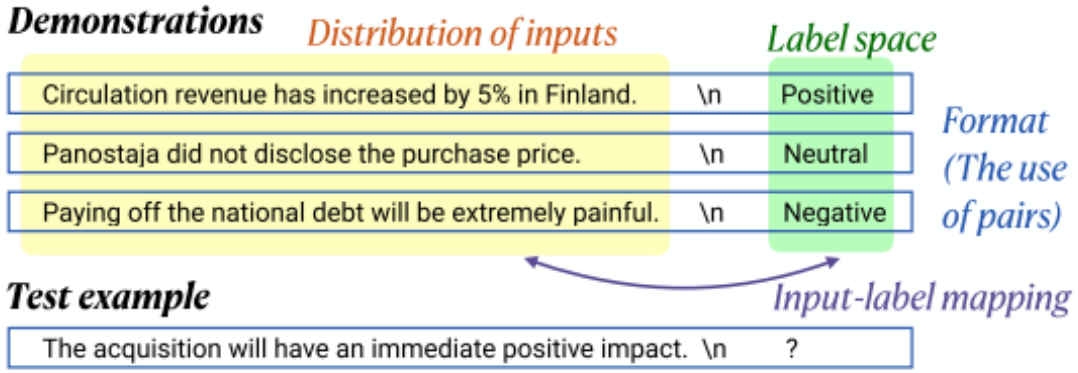
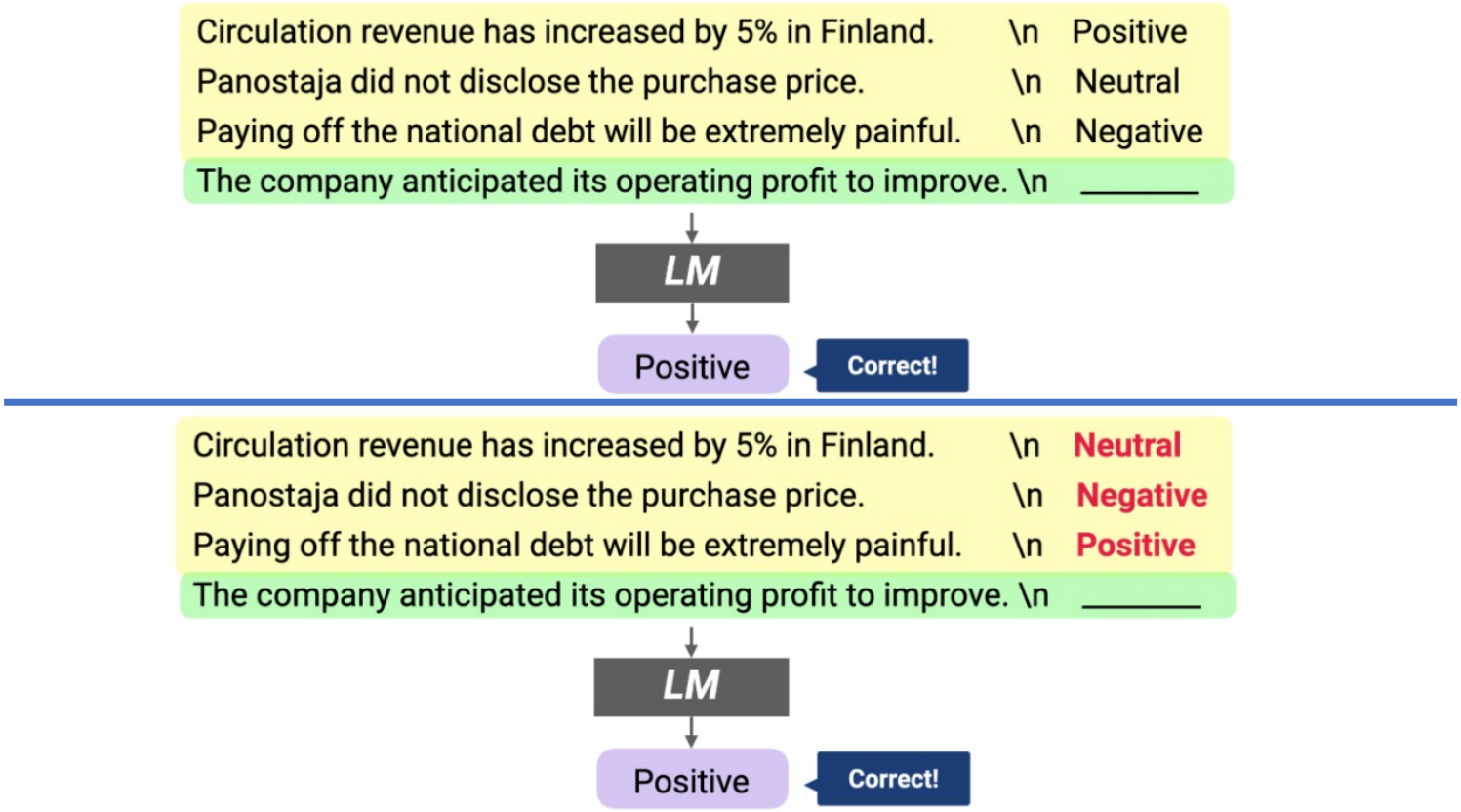


Figure 7: Four different aspects in the demonstrations: the input-label mapping, the distribution of the input text, the label space, and the use of input-label pairing as the format of the demonstrations.

The input-label mapping.

Let's break the input-label correspondence.



The prompt with ground truth outputs (top) and the prompt with random outputs (bottom).

The input-label mapping.

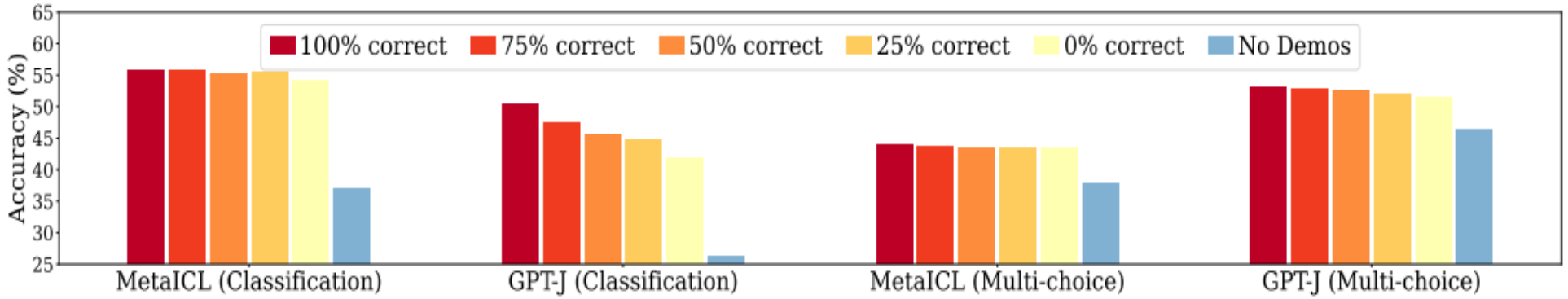


Figure 4: Results with varying number of correct labels in the demonstrations. Channel and Direct used for classification and multi-choice, respectively. Performance with no demonstrations (blue) is reported as a reference.

Takeway#1:

The **input-label mapping** is not necessarily required.

The input distribution.

Let's use the input from another corpus.

The input distribution.

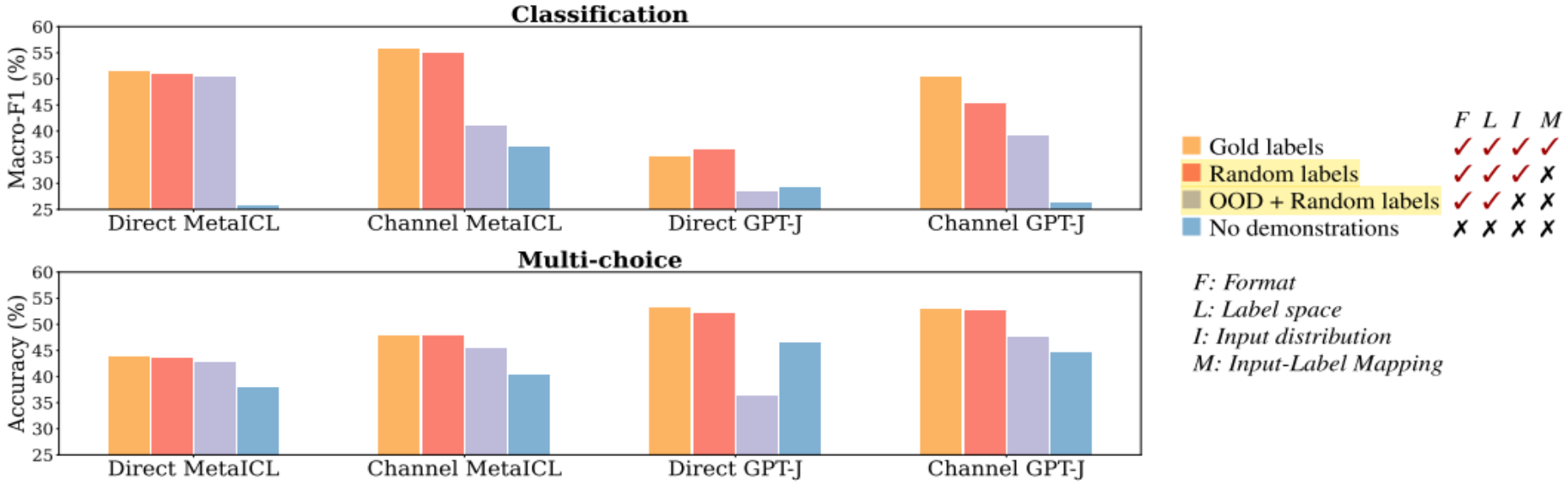


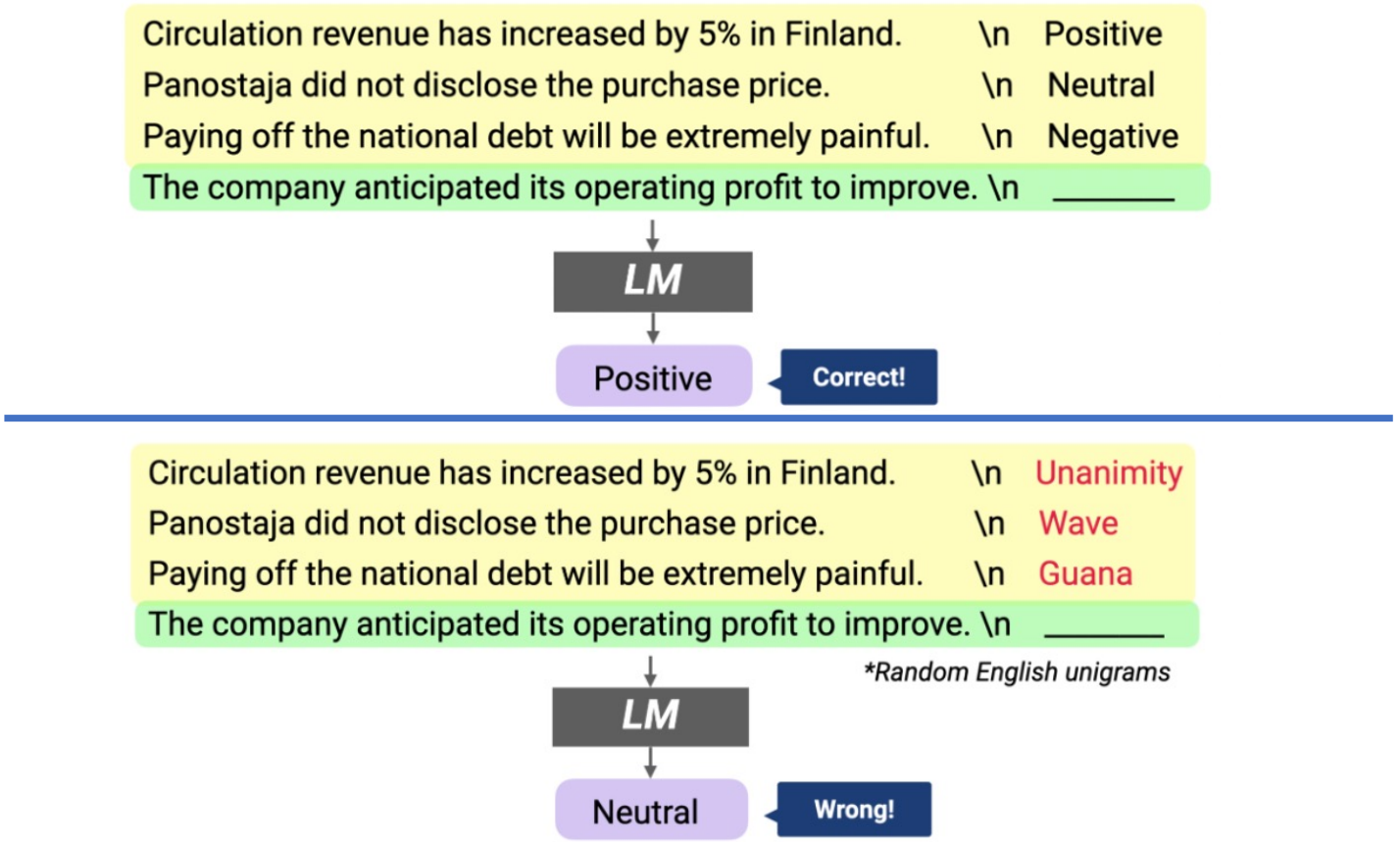
Figure 8: Impact of the distribution of the inputs. Evaluated in classification (top) and multi-choice (bottom). The impact of the distribution of the input text can be measured by comparing ■ and ■. The gap is substantial, with an exception in Direct MetaICL (discussion in Section 5.1).

Takeway#2:

The **input distribution** is necessary.

The label space.

Let's use some random words as labels.



The label space.

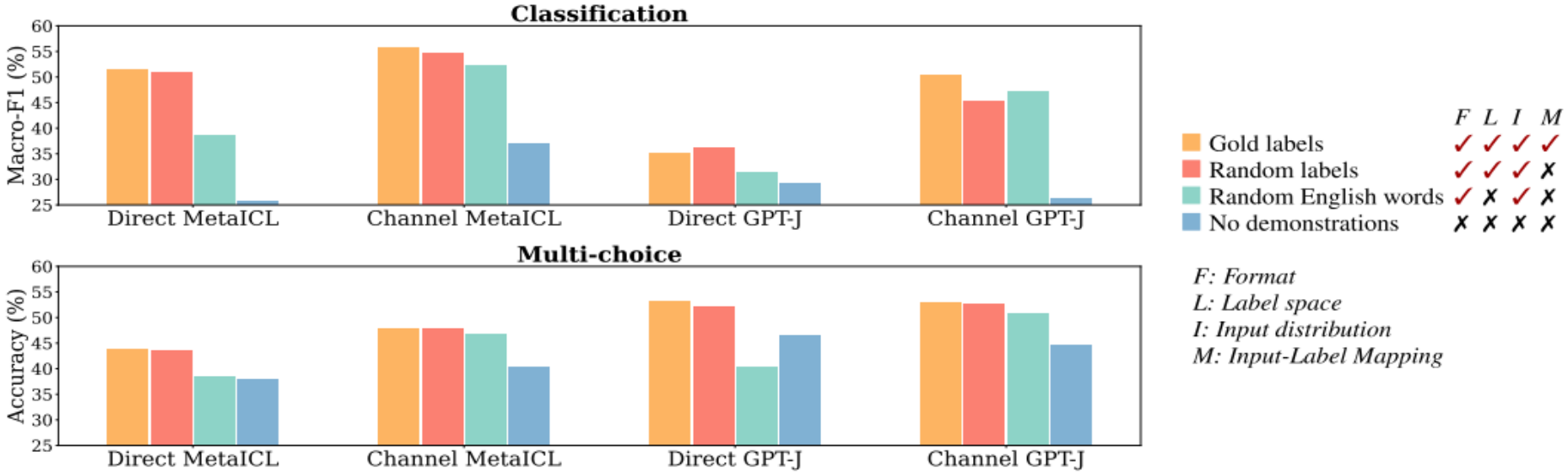


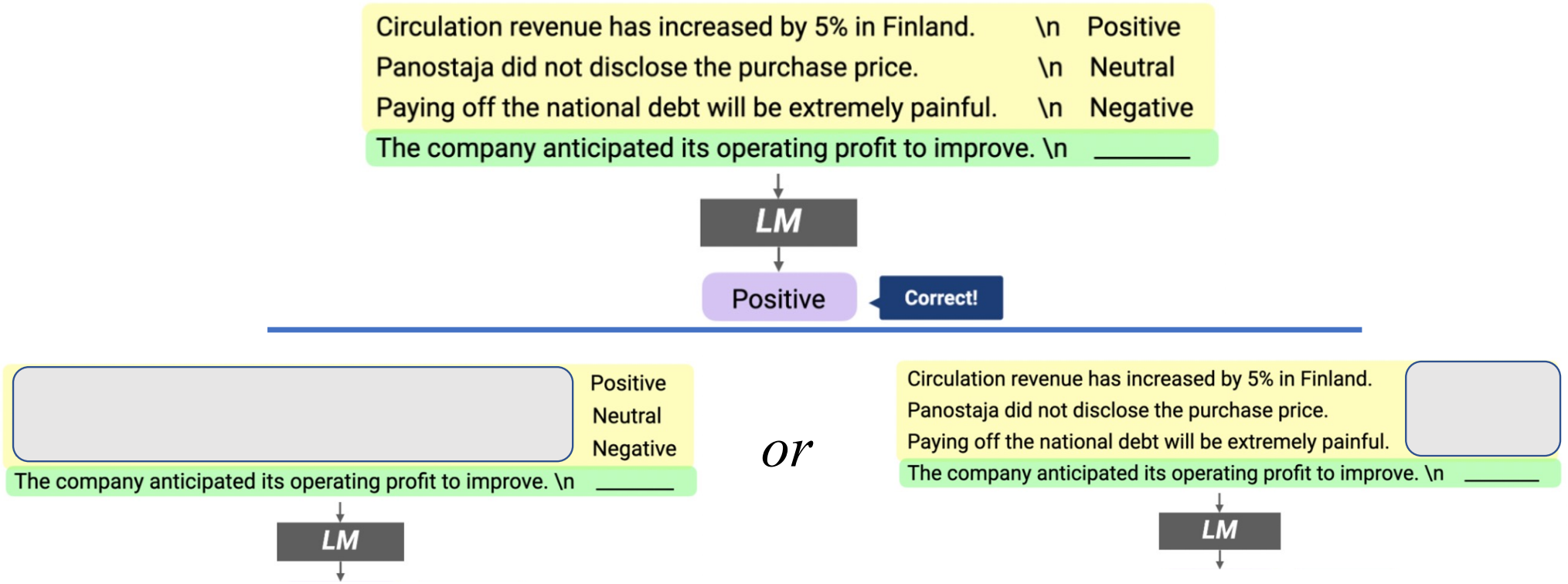
Figure 9: Impact of the label space. Evaluated in classification (top) and multi-choice (bottom). The impact o the label space can be measured by comparing ■ and ■. The gap is significant in the direct models but not in the channel models (discussion in Section 5.2).

Takeway#3:

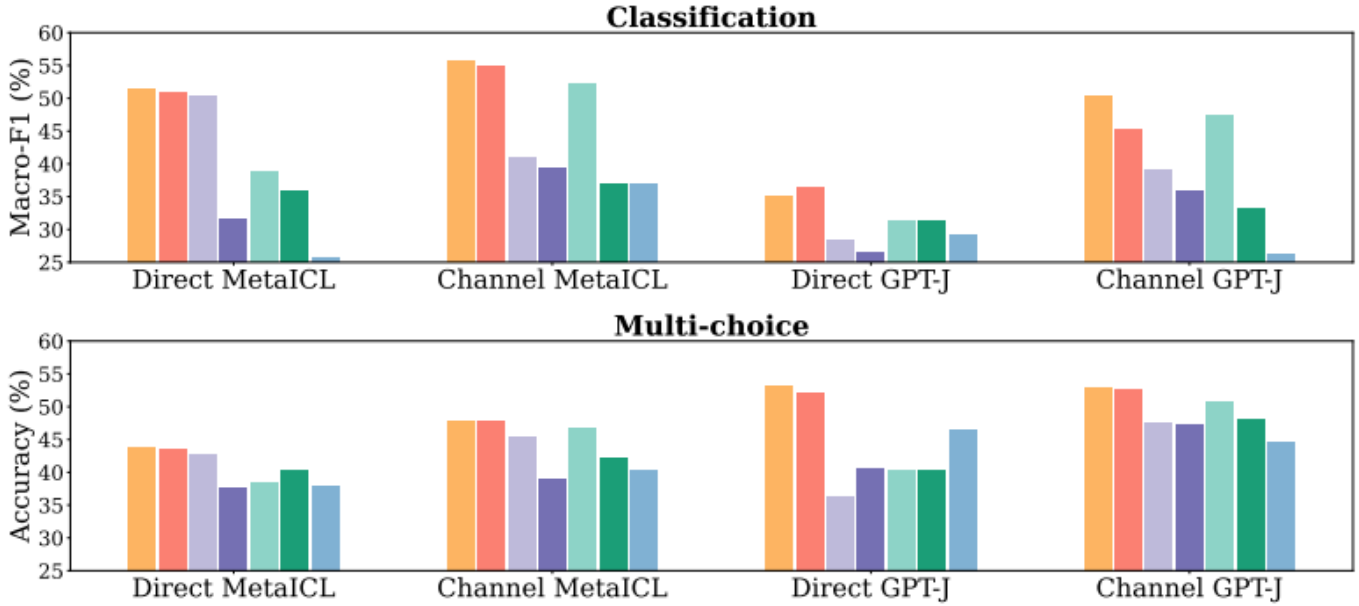
The **label space** is necessary.

The input-label pairing format.

Let's remove inputs or labels.



The input-label pairing format.



	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

Takeway#4:

The **format** is necessary.

To conclude:

✗ 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 $x_1 ... x_k$ are from.

√ 3. **The label space**, i.e., the space covered by $y_1 ... y_k$.

√ 4. **The format**—specifically, the use of input-label pairing as the format.[7]

Most important!

Why A instead of B?

The Model is lazy!

- No gradient updates (**no actual "learning")** ==> simply **exploit** shallow patterns.
- The format is likely easier to exploit.
- The input-label mapping is likely harder to exploit.

Highly rely on the pre-training knowledge.

- In-context learning fulfills the model's pretraining objective (e.g., *next token prediction*).
- **Just a query**! The essence of in-context learning is to making a good query.
- **Domain adaption**. Quickly activate the domain-specific knowledge from LMs.

Some limitations?

- How to explain some other aspects, e.g., the **order** of demonstrations[1]?
- How about **Encoder-decoder LMs**, e.g., T5[2]?
- …

[1] Zhao, Zihao, et al. "Calibrate before use: Improving few-shot performance of language models." International Conference on Machine Learning. PMLR, 2021.
[2] Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." The Journal of Machine Learning Research 21.1 (2020): 5485-5551.

References

- [1] Zhao, Zihao, et al. "Calibrate before use: Improving few-shot performance of language models." International Conference on Machine Learning. PMLR, 2021.
- [2] Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." The Journal of Machine Learning Research 21.1 (2020): 5485-5551.
- [3] “GPT-3: In-Context Few-Shot Learner (2020) - KiKaBeN,” KiKaBeN - Smart Tech Information: From Concept to Coding, Jan. 04, 2023. https://kikaben.com/gpt-3-in-context-few-shot-learner-2020/ (accessed Mar. 01, 2023).
- [4] M. Xie, “How does in-context learning work? A framework for understanding the differences from traditional supervised learning,” SAIL Blog, Aug. 2022. http://ai.stanford.edu/blog/understanding-incontext/ (accessed Mar. 01, 2023).
- [5] Min, Sewon, et al. "Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?." arXiv preprint arXiv:2202.12837 (2022).

Thanks!

Feel free to reach out.