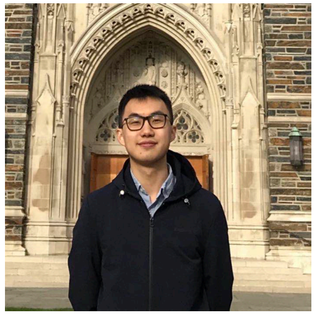
Meta-learning via Language Model In-context Tuning

Yanda Chen*, Ruiqi Zhong, Sheng Zha, George Karypis, He He


* Work done during summer internship at AWS AI.

Few-Shot Learning

- Quickly learns a ***new*** task with ***few*** labeled examples

Adapt

x_1: "I like the movie!", y_1 = Positive 😀

x_2: "Horrible movie!", y_2 = Negative 😔

Sentiment Classification

NEW!

Predict

x^{target}: "The movie is boring.", y^{target}: ?

x^{target}: "This movie is exciting!", y^{target}: ?

.....


Few-Shot Learning

- Why we care?
  - Save annotation efforts
  - Human-like AI

Credit: icons created by Freepik - Flaticon

LM Prompting for FSL

LM Prompting

Chemistry is the study of _____ —> matter and change.

Football is played by _____ —> two teams of eleven players.

Few-shot Learning

x_1: "I like the movie!", y_1 = Positive

x_2: "Horrible movie!", y_2 = Negative

x^{target}: "The movie is boring.", y^{target}: ?

$$I \circ x_1 \circ y_1 \circ x_2 \circ y_2 \circ x^{\text{target}} \quad \_\_\_\_\_ \text{—>} \hat{y}^{\text{target}}$$

What is the sentiment of this review? I like the movie! Positive. Horrible movie! Negative.

This movie is boring. _____ —> Negative

In-context Learning! (ICL)

Oversensitivity

- instruction wording (Schick and Schütze, 2021)

  "What is the sentiment of this review?" vs. "Sentiment of this review?"

- example ordering (Liu et al., 2021)

  $I \circ \square \circ \triangle \circ x^{\text{target}}$ vs. $I \circ \triangle \circ \square \circ x^{\text{target}}$

- example selection (Liu et al., 2021)

Root Cause

Training

LM Prompting

Objective Mismatch!

Testing

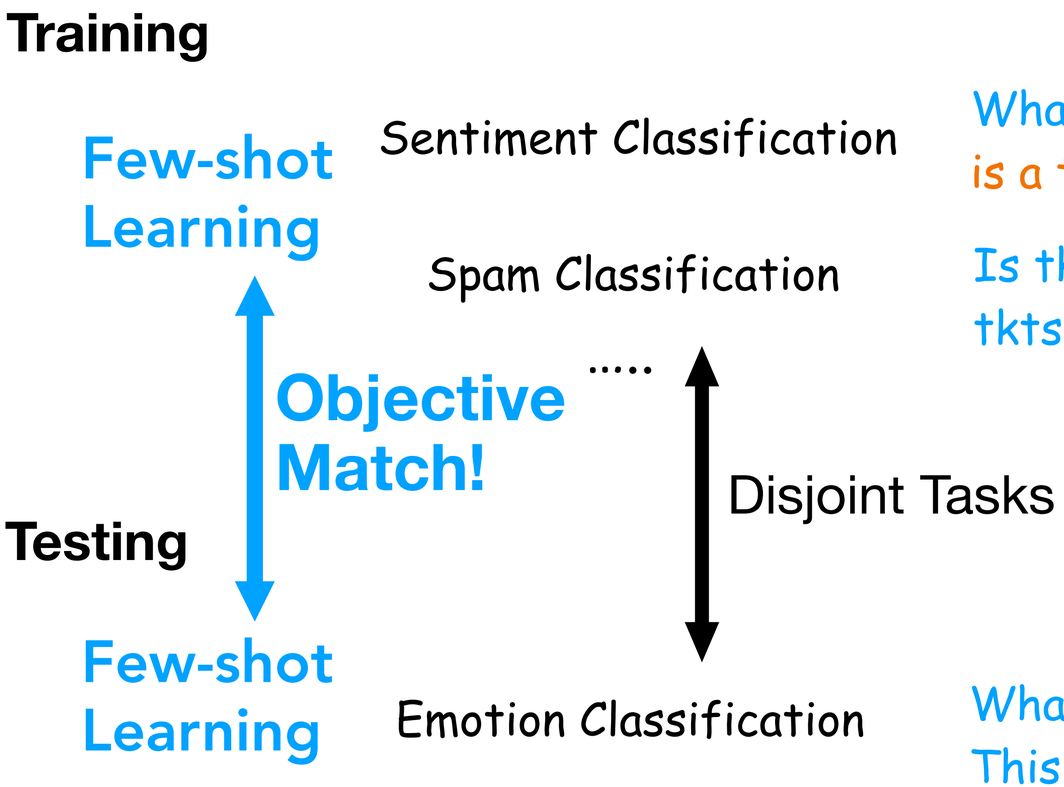
Few-shot Learning

Chemistry is the study of _____ —> matter and change.
Football is played by _____ —> two teams of eleven players.

What is the sentiment of this review? I like the movie! Positive.
This is a total waste of time. _____ —> Negative

In-context Tuning (ICT) META-LEARNING!

Fine-tune LMs to learn **in-context learning** on various tasks



What is the sentiment of this review? I like the movie! Positive. This is a total waste of time. _____ —> Negative

Is this text spam? Free entry in 2 a wkly comp to win FA Cup final tkts. Yes. XXXMobileMovieClub: click the WAP link. _____ —> Yes

What is the emotion of the text? This is so annoying! Anger. This is such an enjoyment. Happiness. I'm so sad. _____ —> ?

A Meta-learning Perspective

$(x_1, y_1), (x_2, y_2)$ — Adapt

$(x^{\text{target}}, ?)$ — Predict

MAML: fine-tune on $(x_1, y_1), (x_2, y_2)$ —> evaluate on x^{target}

(model weights updated with gradient descent)

In-context Tuning: $I \circ x_1 \circ y_1 \circ x_2 \circ y_2 \circ x^{\text{target}} \circ$ _____

(model weights frozen)

Datasets (LAMA)

- Relation: Subject —> Object

Relation = born in

Kandi Burruss —> Atlanta

Relation = capital of

Minsk —> Belarus

- Prediction accuracy
- ~30 different tasks

Datasets (BinaryClfs)

- ~200 binary classification tasks
    - sentiment classification
    - stance classification
    - spam classification
    - …
- AUC-ROC

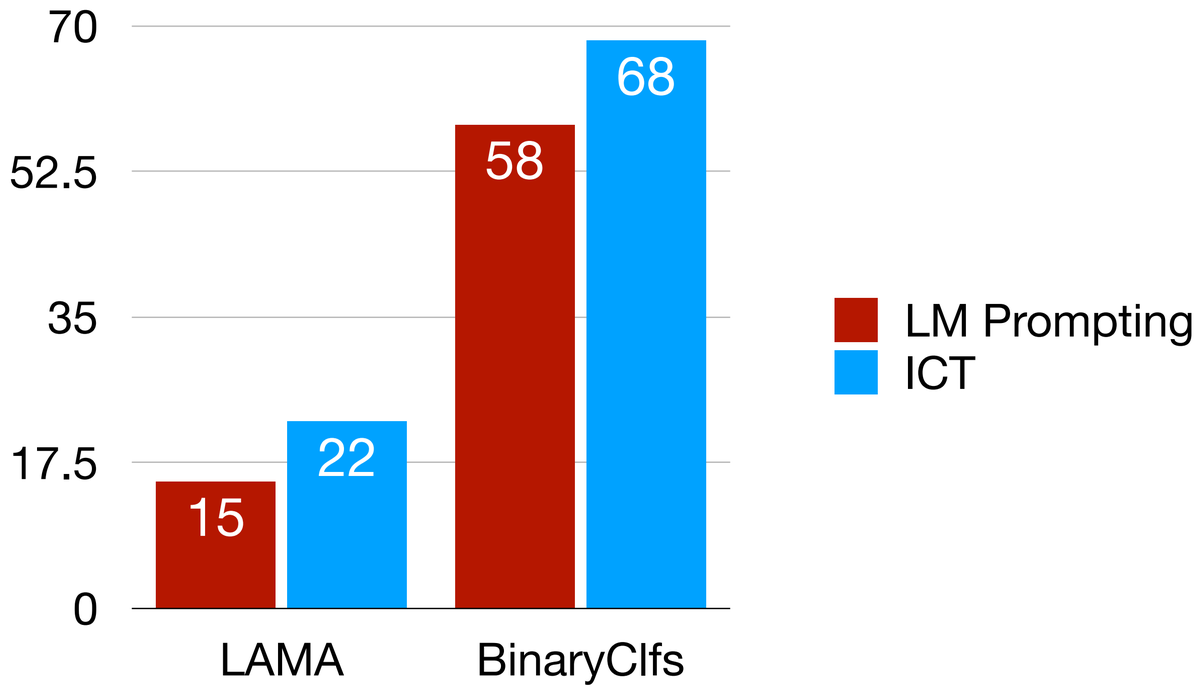
Models

- LAMA
  - BERT - BERT-Base [110M], BERT-Large [340M], DeBERTa-xLarge [900M]
- BinaryClfs
  - GPT2 - GPT2-Medium [345M], GPT2-Large [774M]


Accuracy

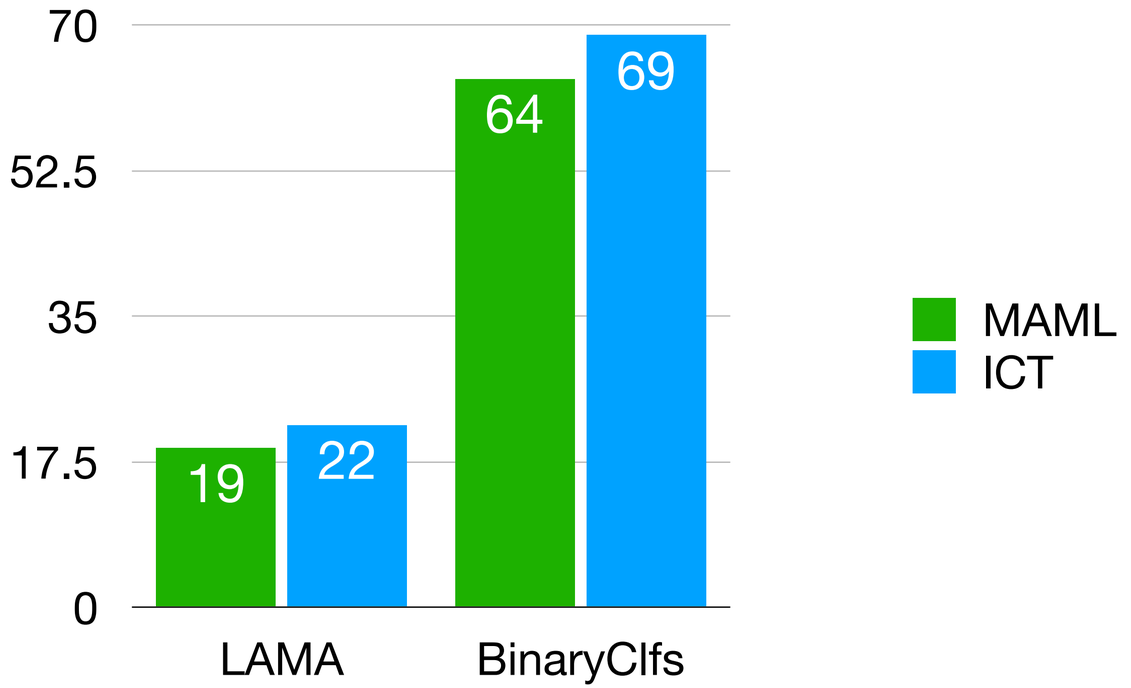
- **ICT** > LM Prompting?
- **ICT** > MAML?
- More few-shot examples —> Better **ICT**?


Does ICT improve ICL accuracy?



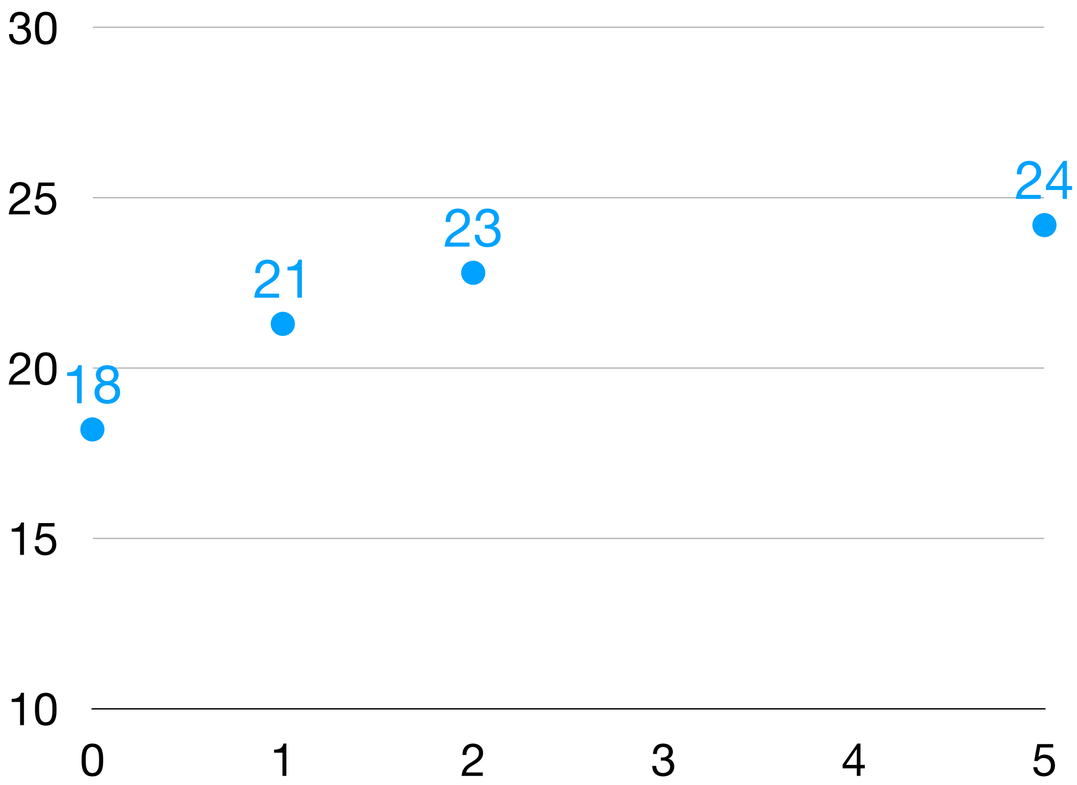
Aligning Train / Test objectives improves few-shot ICL.

How does ICT compare to MAML?



ICT benefits from the inductive bias of LMs to do pattern matching.

Are more few-shot examples better?



ICT effectively uses few-shot examples for task adaptation.

~~Oversensitivity~~ **ICT is much less sensitive!**

- instruction wording (Schick and Schütze, 2021)

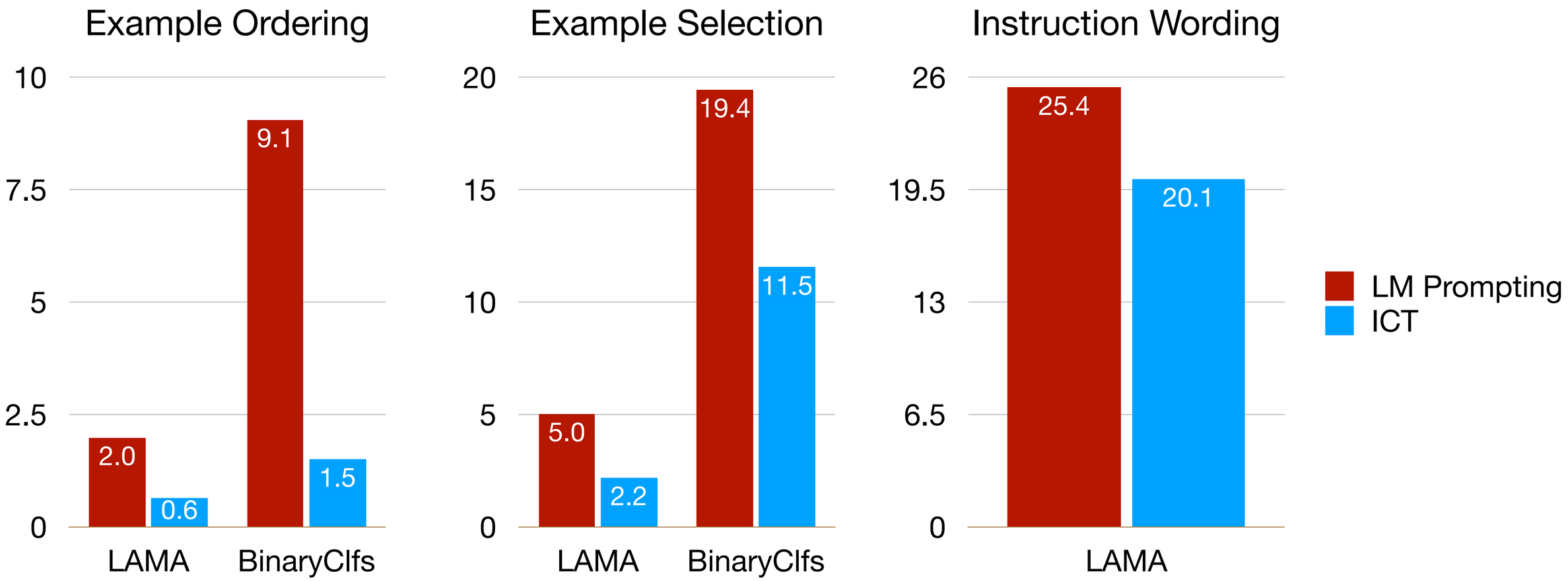
  "What is the sentiment of this review?" vs. "Sentiment of this review?"

- example ordering (Liu et al., 2021)

  $I \circ \square \circ \triangle \circ x^{\text{target}}$ vs. $I \circ \triangle \circ \square \circ x^{\text{target}}$

- example selection (Liu et al., 2021)

Sensitivity (lower better)



In-context Tuning reduces sensitivity significantly.

Conclusion & Takeaways

- We propose In-context Tuning (ICT) for few-shot learning.
    - A meta-learning approach
    - Task adaptation: in-context learning (no gradient update)
- Accuracy: **ICT** > LM Prompting & MAML
- Sensitivity: **ICT** is significantly less sensitive than LM Prompting


Future Directions

- Meta-learning for robustness
  - Distribution shift, rare subgroups, adversarial attacks
- Understanding in-context learning
  - Why it works?
  - Is in-context learning more robust to distribution shift?
  - Can we combine in-context learning with fine-tuning?

Paper: https://aclanthology.org/2022.acl-long.53.pdf

Code: https://github.com/yandachen/In-context-Tuning

Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, He He