

The Role of Demonstrations: What In-Context Learning Actually Needs

Sewon Min

University of Washington
sewon@cs.washington.edu



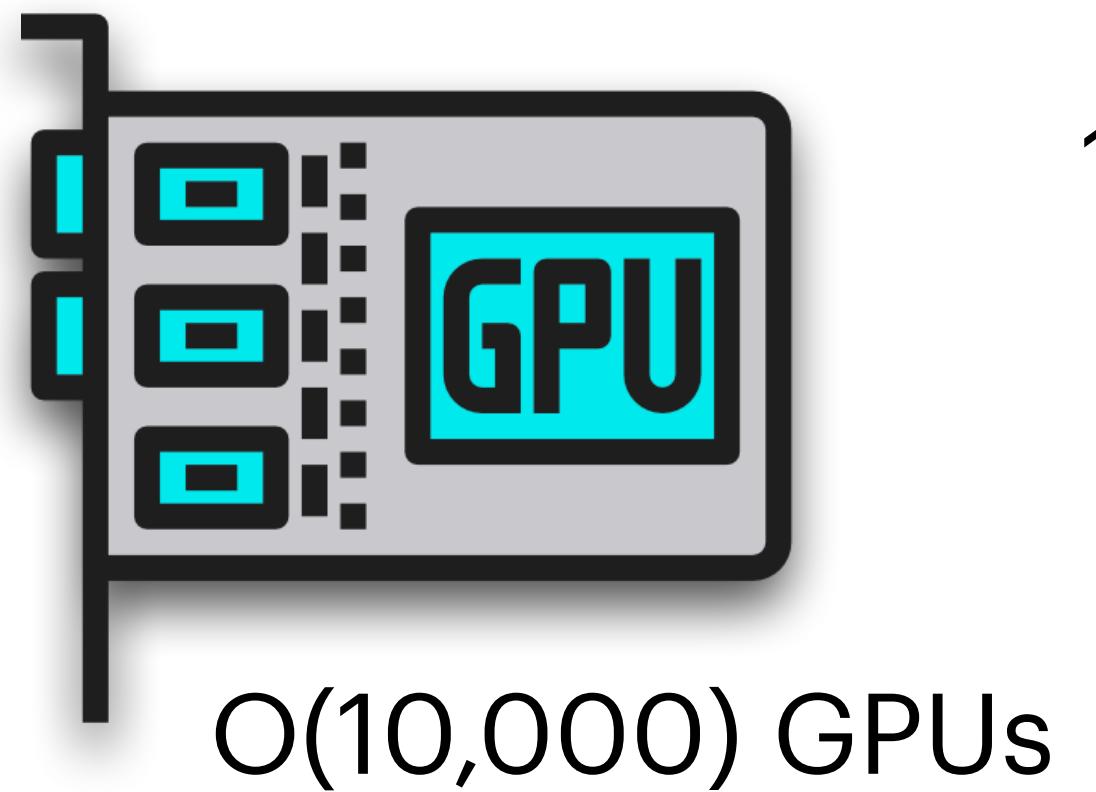
Junyeob Kim

Seoul National University
juny116@europa.snu.ac.kr

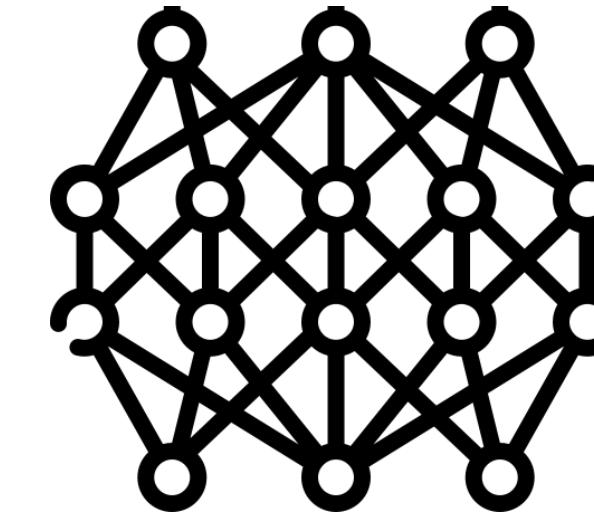




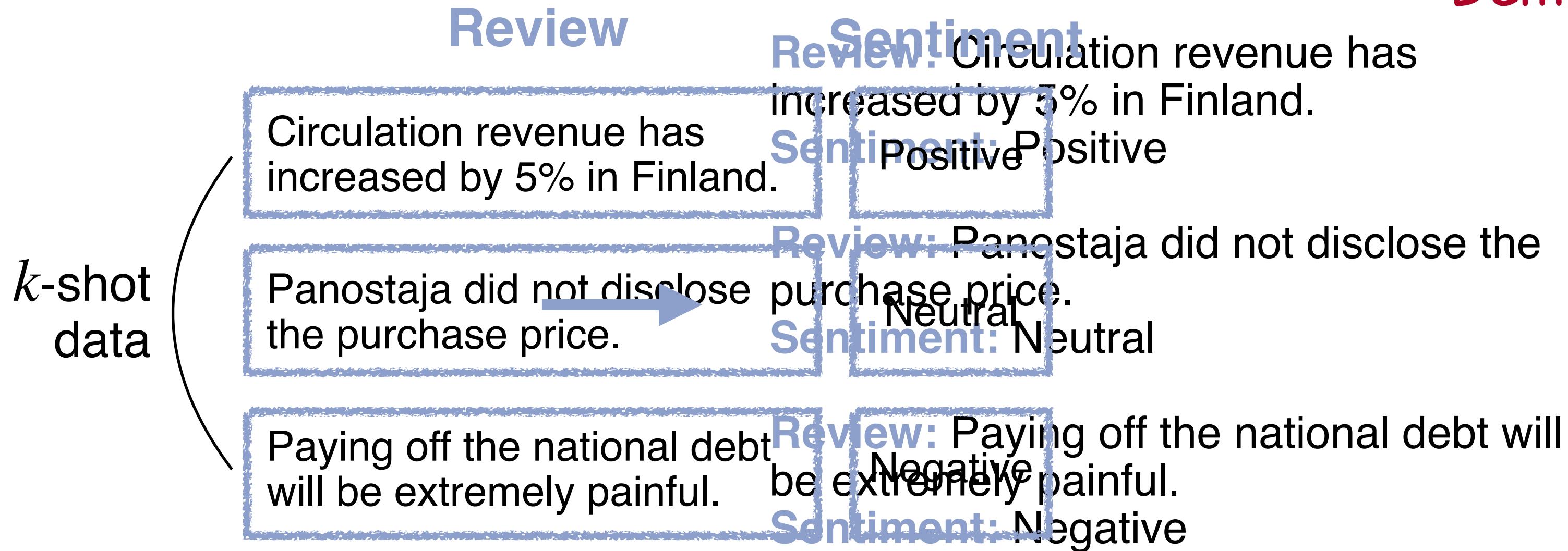
New!
In-Context
Learning



175-billion parameters



In-Context Learning (ICL)



In-Context Learning (ICL)

- Assumption: ICL learns a new task (input-label correlation) *on the fly* from the correct labeled dataset.
- Little understanding of how this is possible.

In-Context Learning (ICL)

- **Assumption:** ICL learns a new task (input-label correlation) *on the fly* from the correct labeled dataset.
- Little understanding of how this is possible.
- Adversarial Q: **How far can you get without the **correct****

Talk outline

Sewon: Ground truth demonstrations are **not required** for ICL.

Junyeob: ICL **do** need ground truth demonstrations, and **can learn** newly-defined tasks.

Sewon: These results **do not rebut** our paper:
they either do not affect our findings, or lack evidence

Sewon: Our general findings **hold more broadly**
(generation/reasoning tasks, alignment)

Junyeob: Yes, but in the end **small deterioration matters**

Junyeob & Sewon: Summary

Talk outline

Sewon: Ground truth demonstrations are **not required** for ICL.

Junyeob: ICL **do** need ground truth demonstrations, and **can learn** newly-defined tasks.

Sewon: These results **do not rebut** our paper:
they either do not affect our findings, or lack evidence

Sewon: Our general findings **hold more broadly**
(generation/reasoning tasks, alignment)

Junyeob: Yes, but in the end **small deterioration matters**

Junyeob & Sewon: Summary

ICL with random labels

Review: Circulation revenue has increased by 5% in Finland.

Sentiment: Positive

Review: Panostaja did not disclose the purchase price.

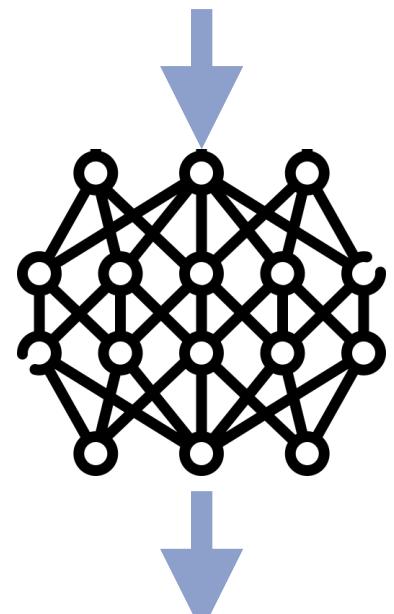
Sentiment: Neutral

Review: Paying off the national debt will be extremely painful.

Sentiment: Negative

Review: The company anticipated its operating profit to improve.

Sentiment:



Positive

Review: Circulation revenue has increased by 5% in Finland.

Sentiment: Neutral

Review: Panostaja did not disclose the purchase price.

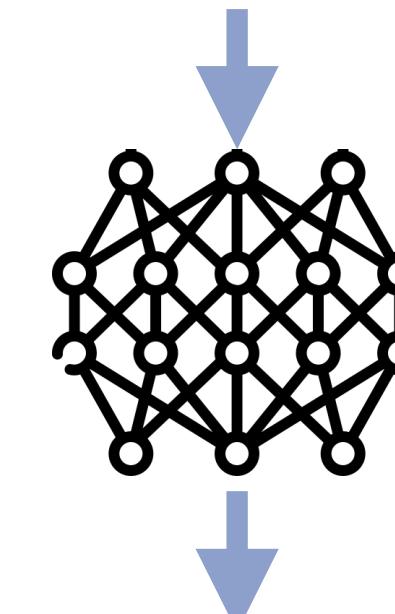
Sentiment: Negative

Review: Paying off the national debt will be extremely painful.

Sentiment: Positive

Review: The company anticipated its operating profit to improve.

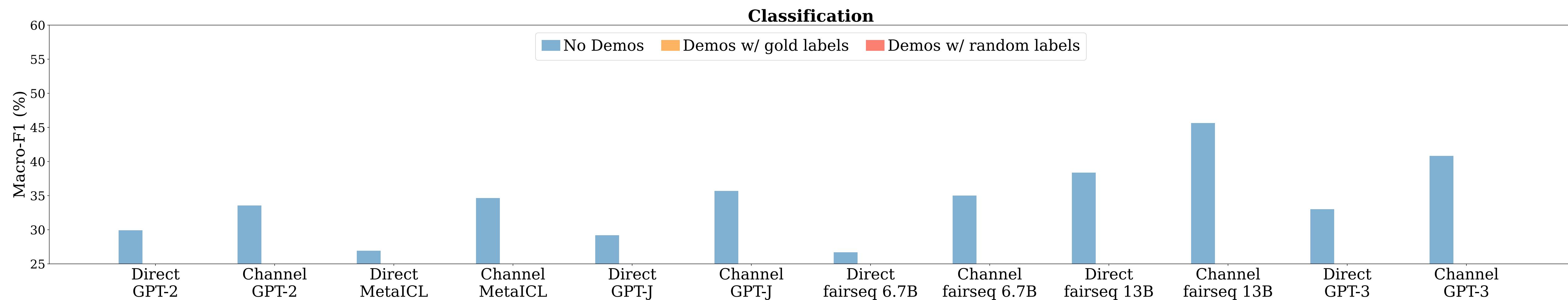
Sentiment:



Positive

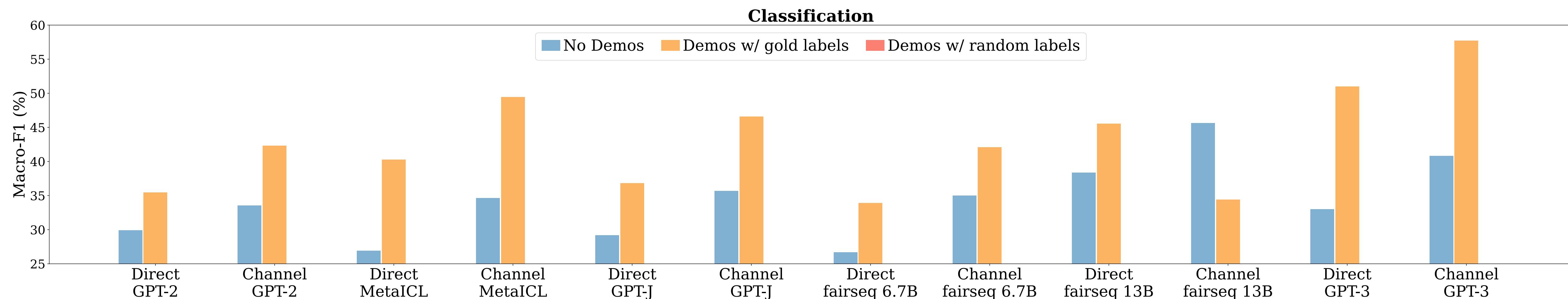
ICL with random labels

Averaged over 16 classification tasks



ICL with random labels

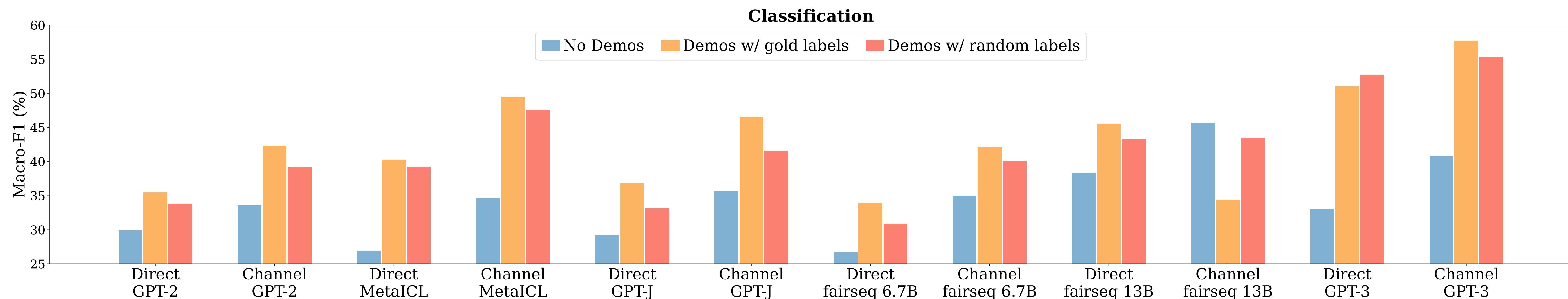
Averaged over 16 classification tasks



Significant improvements through demonstrations

ICL with random labels

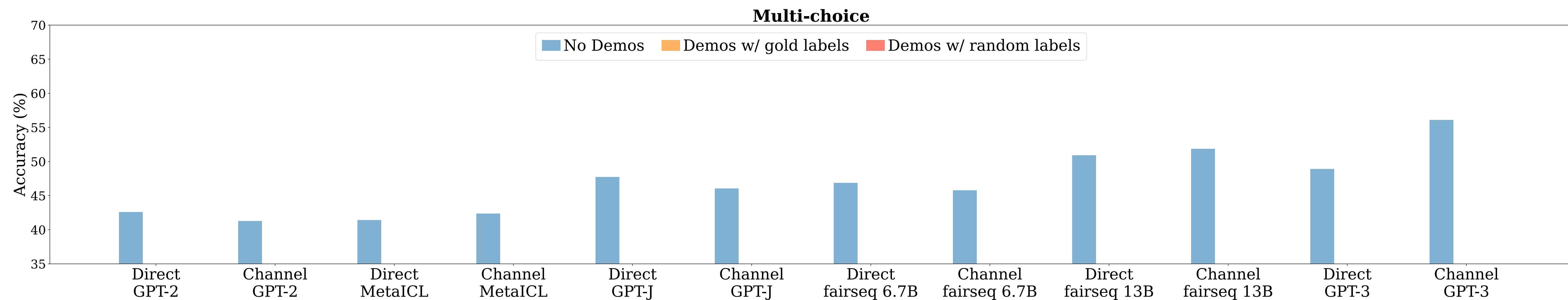
Averaged over 16 classification tasks



Replacing gold labels with random labels hurt performance **only by 2.6%** on average

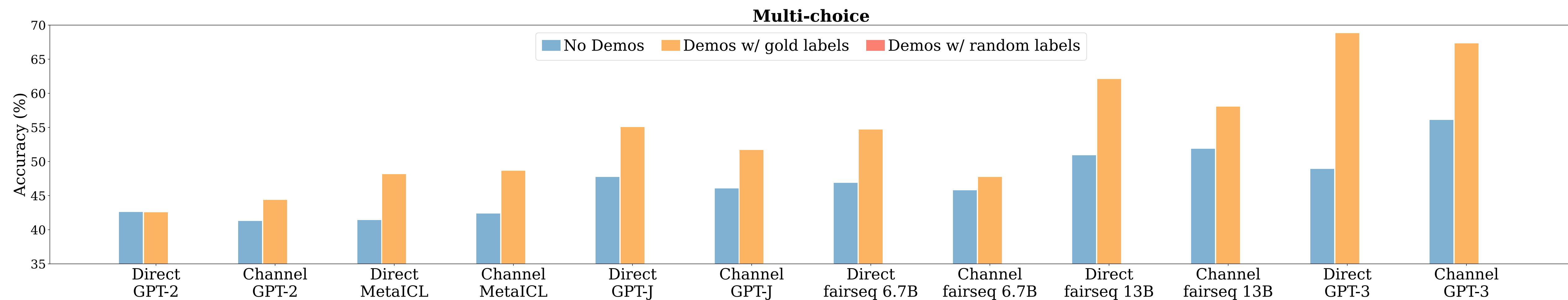
ICL with random labels

Averaged over 10 multi-choice tasks



ICL with random labels

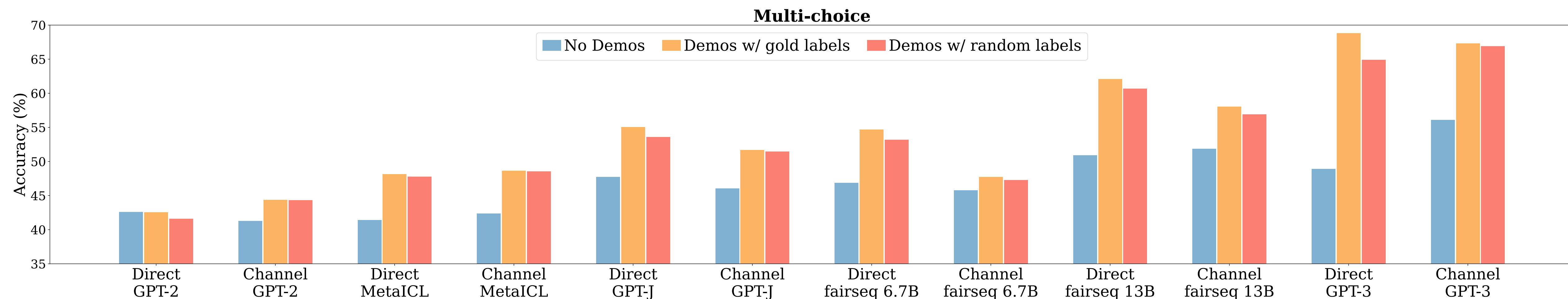
Averaged over 10 multi-choice tasks



Significant improvements through demonstrations

ICL with random labels

Averaged over 10 multi-choice tasks



Replacing gold labels with random labels hurt performance **only by 1.7%** on average

Summary

- Correct demonstrations are **not required**
- LMs are **not** learning a new task on the fly — they already knew the task (from **priors from training**)
 - Demonstrations activate such priors

•  Reviewed March 31, 2018 □ via mobile
S White Chocolate Banana Bread

Our group of high school students and a few parents wandered by Cafe Solstice on a Saturday morning. We had some time to kill before a music clinic at UW. We walked in together, but part of the group walked back out. I stayed to... [More](#)

Date of visit: March 2018

Helpful? 

•  Reviewed April 12, 2017
Really good coffee

In the land of great coffee shops, this one stands out. Not super fancy inside - its right near the college - it carries a fun ambience with great baked goods and really good coffee. Great service and a large area allowing a lot of... [More](#)

Date of visit: April 2017

Helpful? 

•  Reviewed November 11, 2017 □ via mobile
Great java, nice ambiance and friendly staff

In a city full of excellent coffee houses, this place is a great choice if you're seeking your caffeine fix next to UW :)). The barista brewed delicious beverages for us, and the staff were very courteous as well... we sampled some of their baked... [More](#)

Date of visit: November 2017

Helpful? 

•  Reviewed April 2, 2017
The best you are ging to get on the Ave

Coffee is OK. The pastries are good. The sandwiches are so-so. There are not good choices for breakfast food on the Ave.

Date of visit: February 2017

Helpful? 

On,

n

Consistent findings from others



Riley Goodside
@goodside



Part 6: Multiple choice answers are random. As reported by [@sewon_min](#) et al., this almost no effect on correctness of answers vs. giving correct labels:
arxiv.org/abs/2202.12837

Q: How many legs do most dogs have?

- A) 0; B) 2; C) 4; D) 6; E) 8

Answer: E

Q: What is the square root of 16?

- A) 1; B) 4; C) 6; D) 8; E) 16

Answer: A

Q: Where do fish typically live?

- A) in the sky; B) in water; C) in the desert; D) in the forest; E) underground

Answer: E

Q: A car swerves suddenly on the road while beeping its horn. Why might this happen?

- A) the oil is low; B) the driver is tired; C) there is a animal in the road; D) the car is out of gas; E) the brakes are not working

Answer: C

Consistent findings from others



Riley Goodside
@goodside

Part 4: Answers are off by one. The model can be briefly confused by several highly consistent wrong examples but normal arithmetic usually wins. I can't get it to do all answers +1 for general, basic arithmetic without explicit instruction.

...

What is 11 + 1?

13

What is 11 + 2?

14

What is 11 + 5?

17

What is 11 + 6?

18

What is 6 + 11?

17

What is 11 + 9?

20

What is 11 + 8?

19

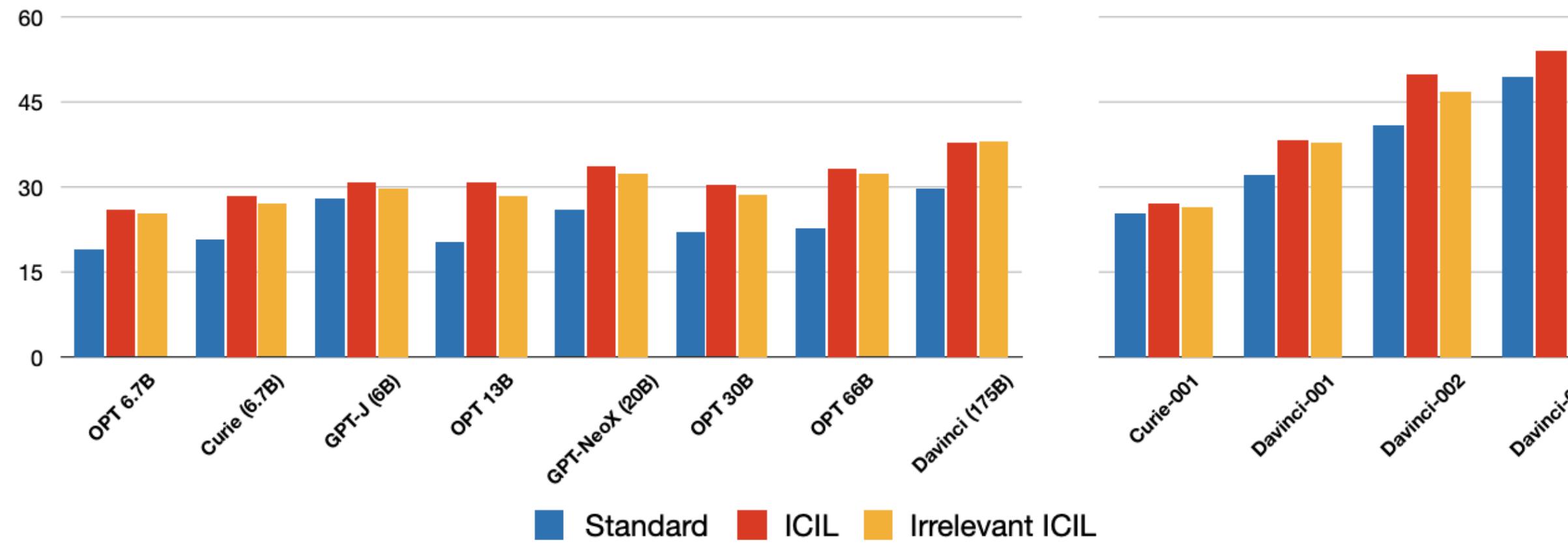
What is 11 + 7?

18

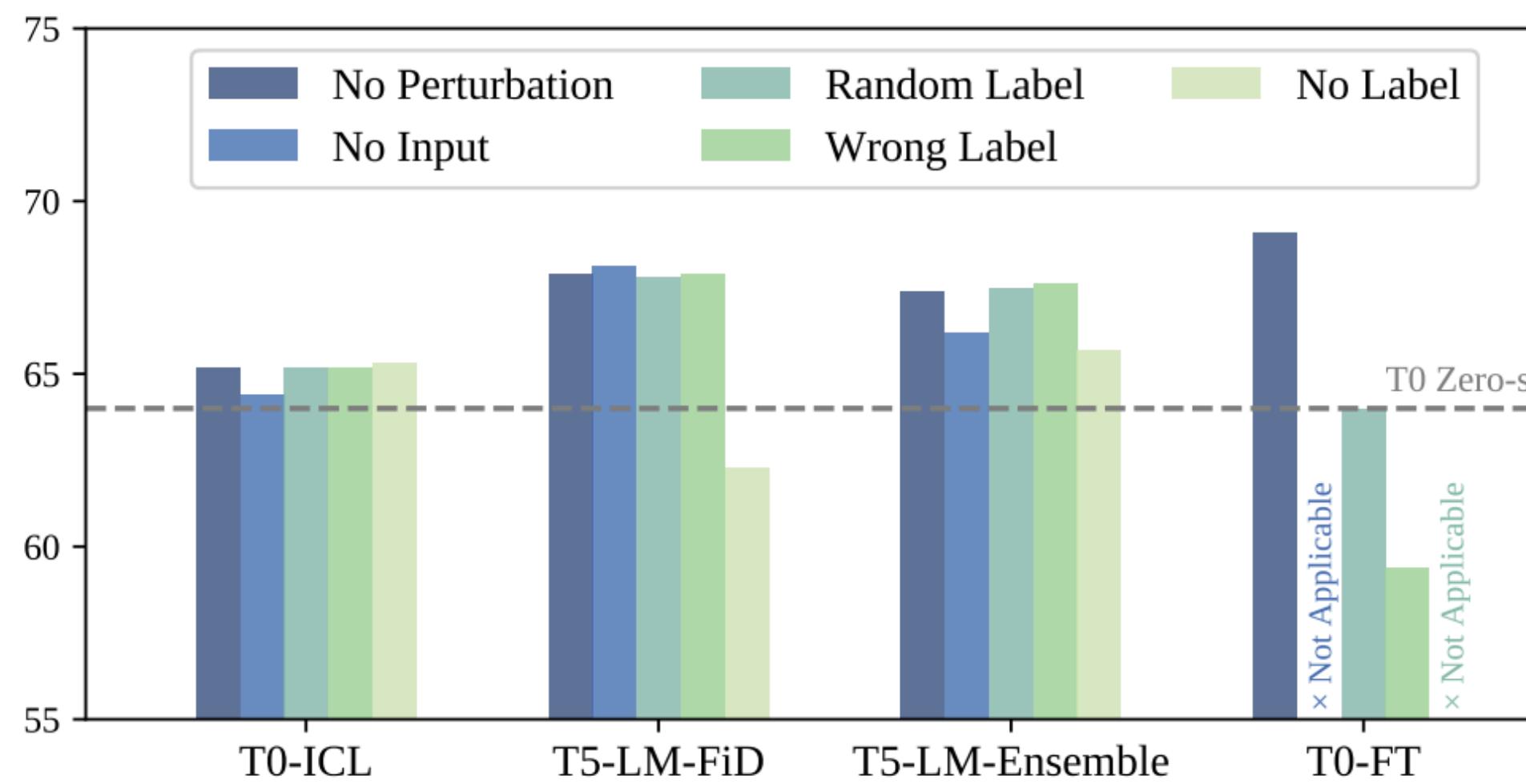
What is 11 + 6?

17

Consistent findings from others



Findings hold for cross-task ICL



Findings hold with different ICL methods

Talk outline

Sewon: Ground truth demonstrations are **not required** for ICL.

Junyeob: ICL **do** need ground truth demonstrations, and **can learn** newly-defined tasks.

Sewon: These results **do not rebut** our paper:
they either do not affect our findings, or lack evidence

Sewon: Our general findings **hold more broadly**
(generation/reasoning tasks, alignment)

Junyeob: Yes, but in the end **small deterioration matters**

Junyeob & Sewon: Summary

Ground-Truth Labels Matter: A Deeper Look into Input-Label Demonstrations

Kang Min Yoo*^{#†‡§}, **Junyeob Kim***[§], **Hyuhng Joon Kim**[§], **Hyunsoo Cho**[§],
Hwiyeol Jo[†], **Sang-Woo Lee**^{†‡¶}, **Sang-goo Lee**[§], **Taeuk Kim**^{#¶}

[§]Seoul National University, [†]NAVER AI Lab, [‡]NAVER CLOVA

[¶]Korea Advanced Institute of Science and Technology, [¶]Hanyang University
{juny116,heyjoonkim,johyunsoo,sglee}@europa.snu.ac.kr
{hwiyeol.jo,sang.woo.lee,kangmin.yoo}@navercorp.com
kimtaeuk@hanyang.ac.kr

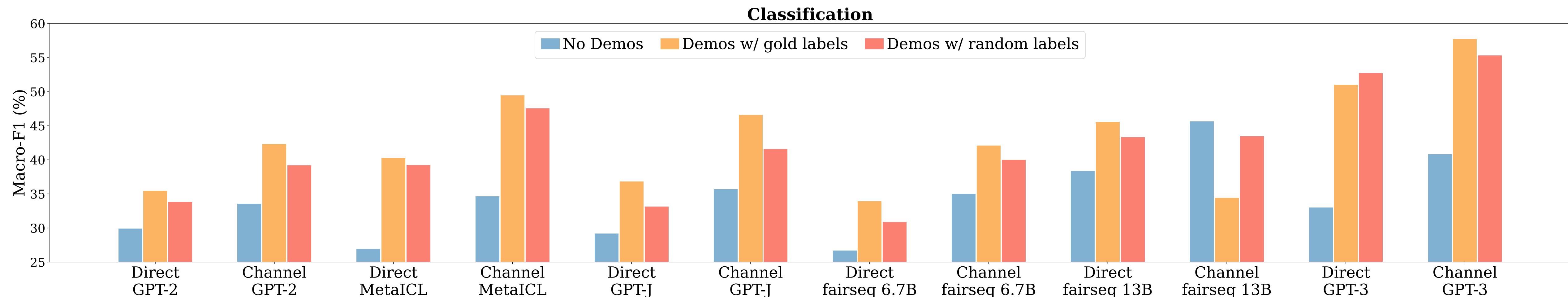
**“The impact of ground truth demonstrations [for in-context learning]
varies depending on the setting”**

“There is some evidence that in fact LLMs learn a new task on-the-fly”

Motivation

- Correct demonstrations are **not required**

→ The claim seem too strong and we need some quantification of the phenomenon



In almost all cases, the performance drops slightly in the case of random labels; can this be neglected?

Research Question

Does demonstration accuracy *really* matter?
If so when, and to what degrees?



What if we control task priors learned during pretraining?



Research Question

Does demonstration accuracy *really* matter?
If so when, and to what degrees?



What if we control task priors learned during pretraining?



ICL with corrupted labels

Review: Circulation revenue has increased by 5% in Finland.

Sentiment: Positive

Review: Panostaja did not disclose the purchase price.

Sentiment: Neutral

Review: Paying off the national debt will be extremely painful.

Sentiment: Negative

Review: The company anticipated its operating profit to improve.

Sentiment: Positive

Review: A financial crisis may break out.

Sentiment:

50% corruption



Review: Circulation revenue has increased by 5% in Finland.

Sentiment: **Neutral**

Review: Panostaja did not disclose the purchase price.

Sentiment: **Neutral**

Review: Paying off the national debt will be extremely painful.

Sentiment: **Positive**

Review: The company anticipated its operating profit to improve.

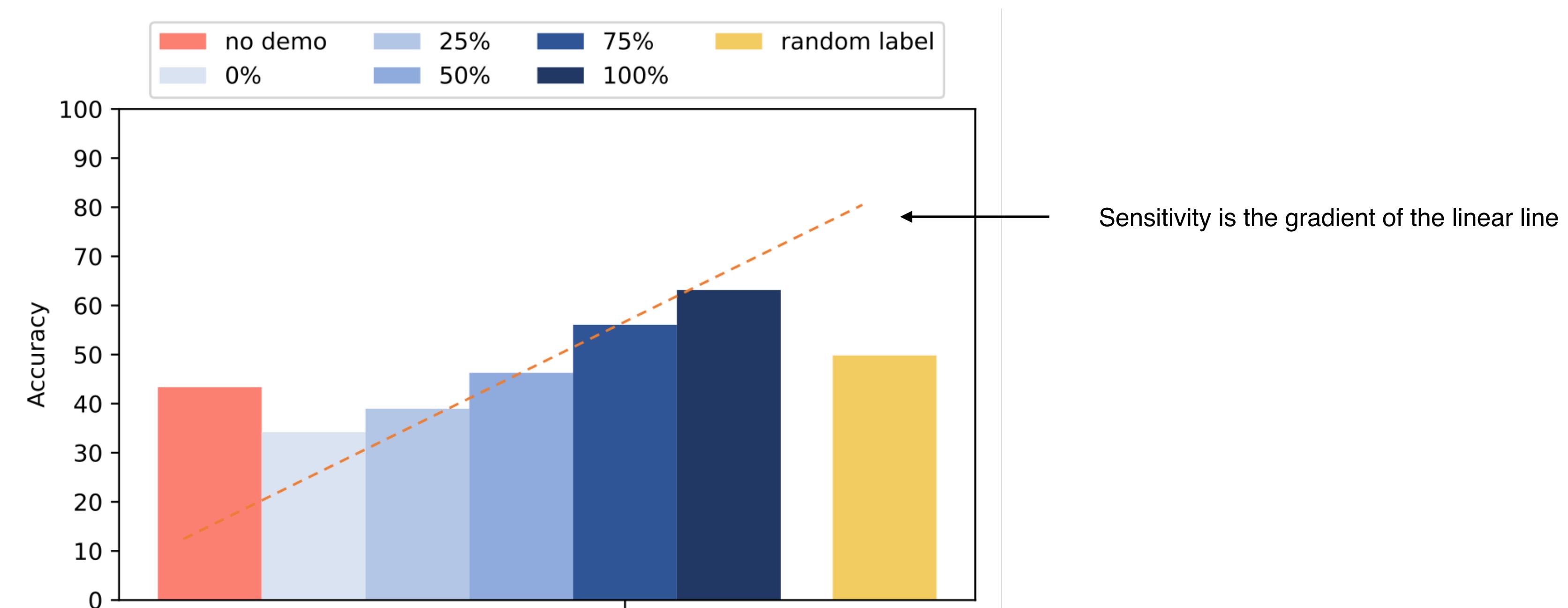
Sentiment: **Positive**

Review: A financial crisis may break out.

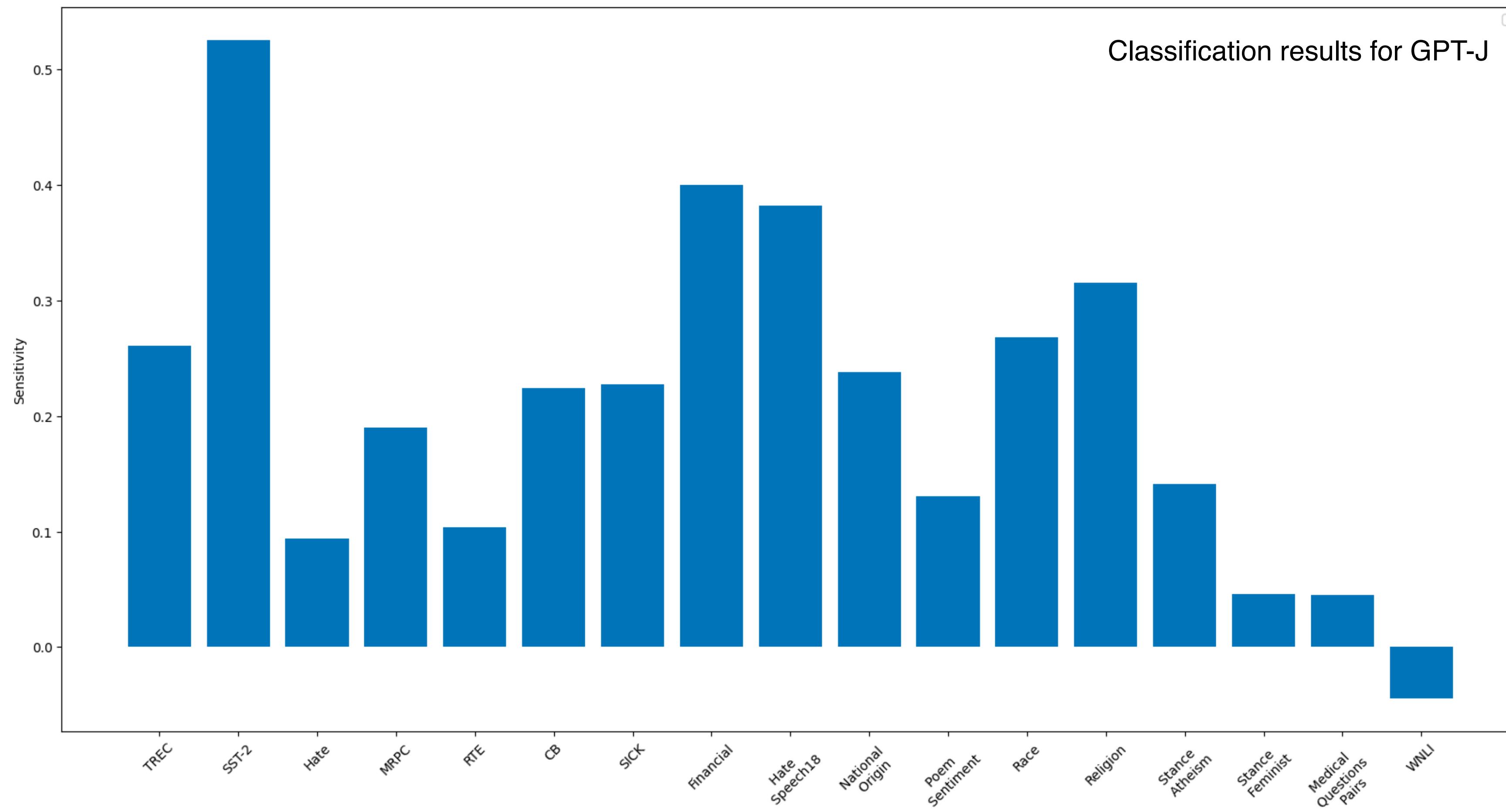
Sentiment:

Quantification

Label-correctness **sensitivity**: The degree of which the downstream classification performance changes when the model is subject to a fixed amount of label corruption.



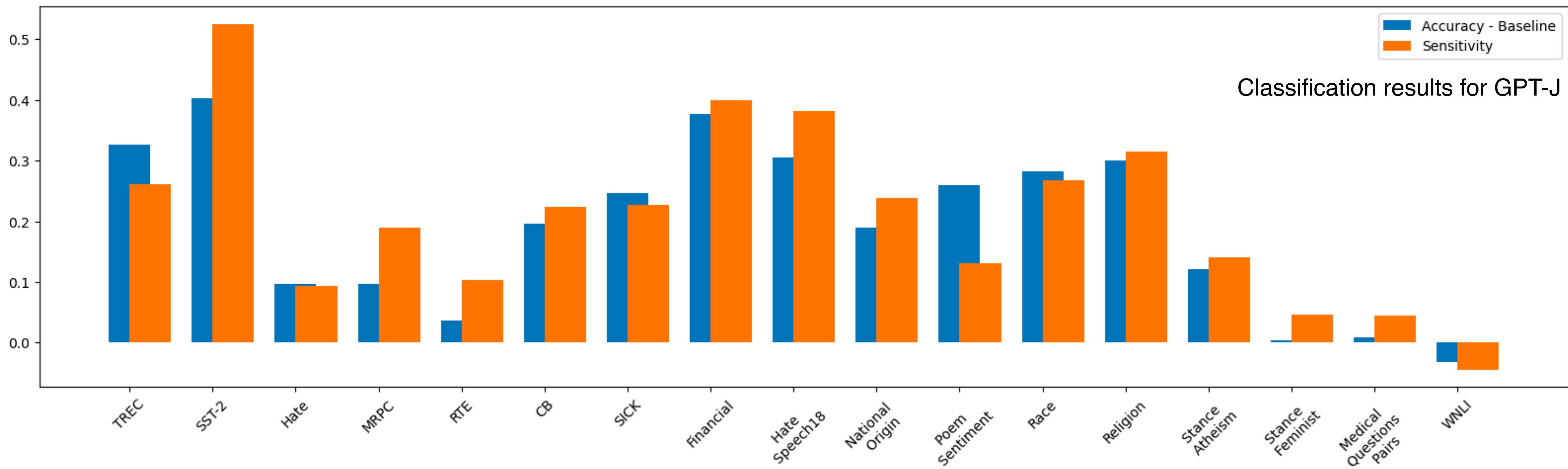
Task level analysis on sensitivity



Sensitivity varies across different tasks (AVG: 0.3)

→ It is not safe to conclude that the label correctness does not matter

When does the labels matter more?

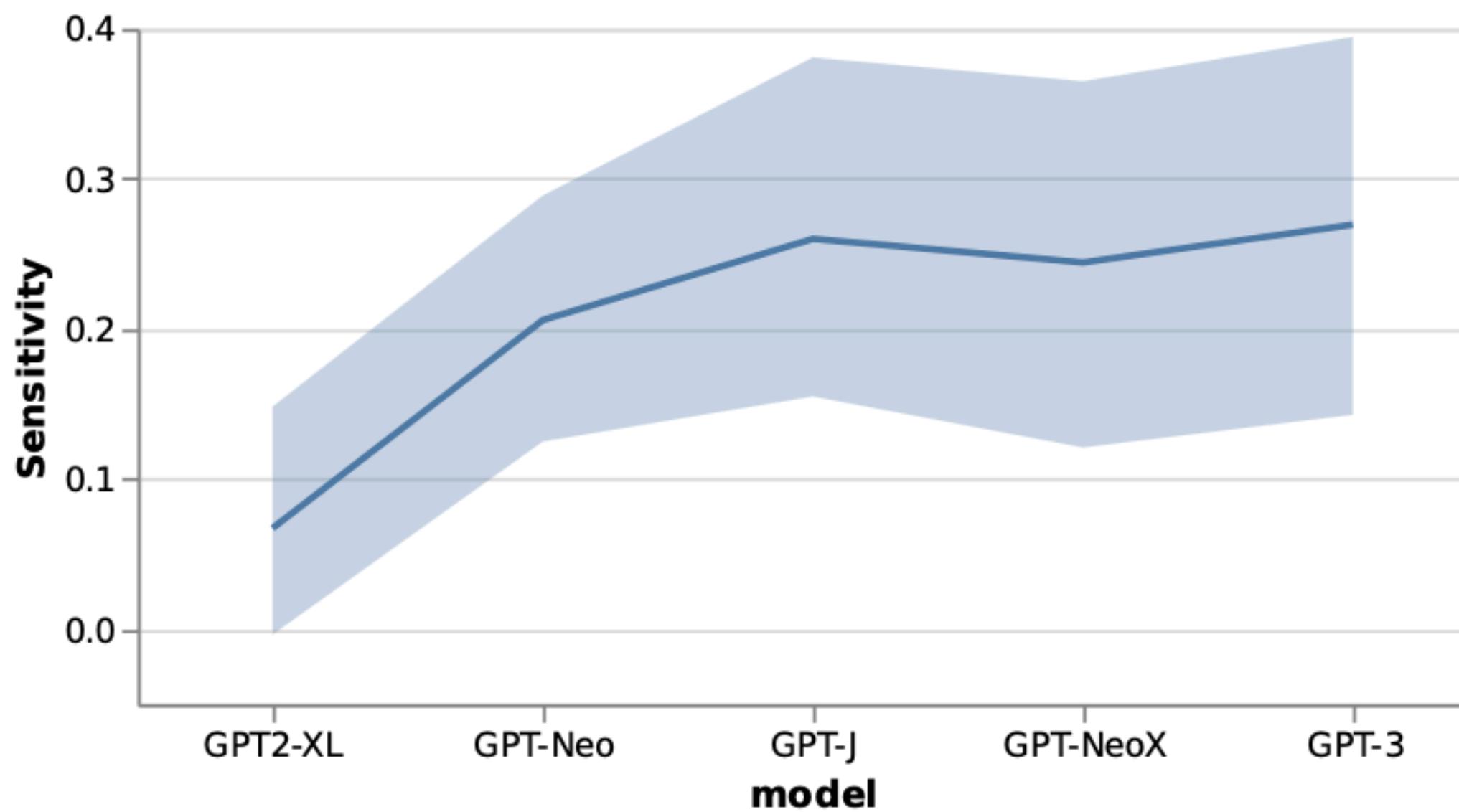


Sensitivity seem higher when the tasks are relatively easy (for the model)

→ The capabilities of the model might be related to the sensitivity

Model size matters

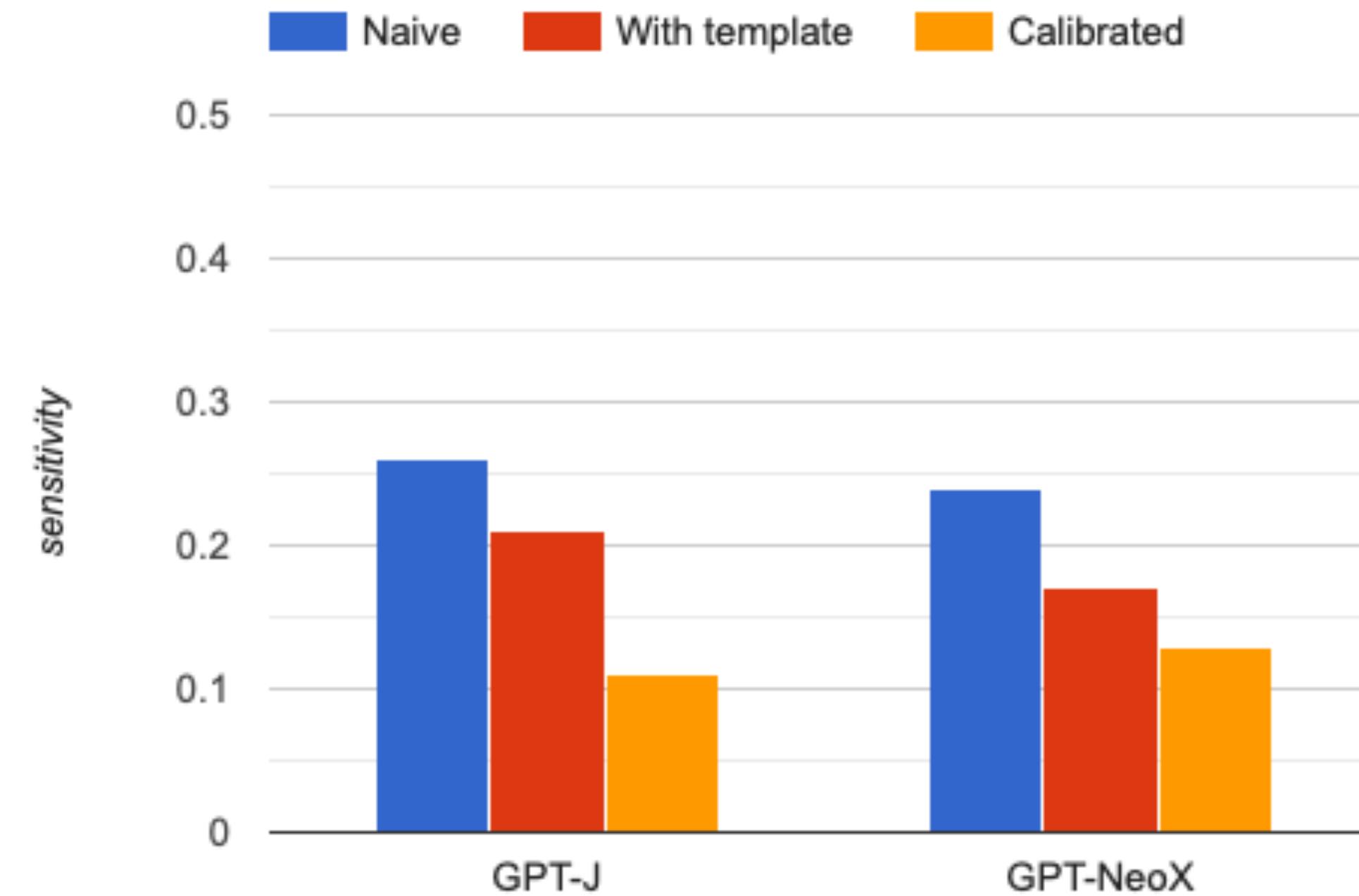
If the model capability seem relevant, does the size of the model matter?



In general, **larger models** seem to have **larger sensitivity**

→ We will discuss more in following sections (Wei et al.)

Prompting method matters



Using template and applying noisy channel inference all **reduces** the sensitivity
(even though the accuracy was higher)

→ Seem to contradict the previous finding that relative accuracy matters

Why does prompting methods reduce sensitivity?

We hypothesize that

- Templates and calibrating induce the model to be more reliant on prior knowledge

Review: Circulation revenue has increased by 5% in Finland.
Sentiment: Positive

In other words, it is less sensitive to the context

Review: Panostaja did not disclose the purchase price.

Sentiment: Neutral

With template

(taken from Zhao et al. 2021)

Review: Paying off the national debt will be extremely painful.

Sentiment: Negative

→ Models assume that it is sentiment analysis from the word “Review” and “Sentiment”, which leads to neglecting the validity of the label.

Research Question

Does demonstration accuracy *really* matter?
If so when, and to what degrees?



What if we control task priors learned during pretraining?



Prior-free label

Review: Circulation revenue has increased by 5% in Finland.

Sentiment: Positive

Review: Panostaja did not disclose the purchase price.

Sentiment: Neutral

Review: Paying off the national debt will be extremely painful.

Sentiment: Negative

Newly defined labels



Review: The company anticipated its operating profit to improve.

Sentiment: Positive

Review: A financial crisis may break out.

Sentiment:

Review: Circulation revenue has increased by 5% in Finland.

Sentiment: I

Review: Panostaja did not disclose the purchase price.

Sentiment: J

Review: Paying off the national debt will be extremely painful.

Sentiment: K

Review: The company anticipated its operating profit to improve.

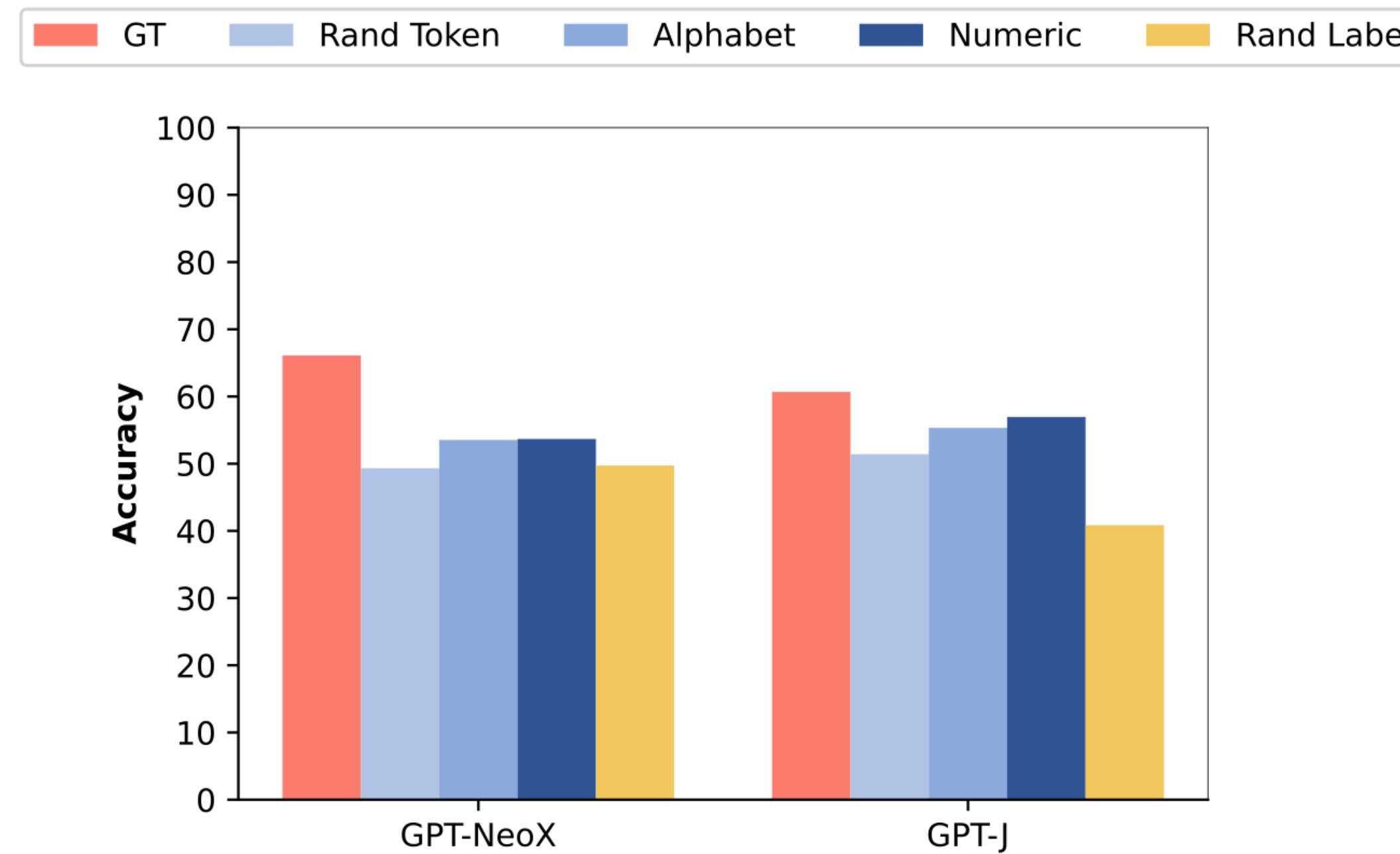
Sentiment: I

Review: A financial crisis may break out.

Sentiment:

If the correctness of labels doesn't matter, models won't perform newly-defined labels at all.

Can LLMs perform without label priors?



Prior-free label which need to learn new task on the fly outperforms random label

Summary

- The **sensitivity varies** across different tasks
 - There seems to be a correlation between model capability and sensitivity.
- Even when we remove the task priors of the labels the LMs still performs the downstream task.
- The results indicates that the demonstration works as both **inducer of prior knowledge** and **training samples for model**, depending on the setting, the importance of certain role may change.

Follow-up Work

Google Research

March 9, 2023

LARGER LANGUAGE MODELS DO IN-CONTEXT LEARNING DIFFERENTLY

Jerry Wei^{1,2,*}

Jason Wei¹

Yi Tay¹

Dustin Tran¹

Albert Webson^{1,3,*}

Yifeng Lu¹

Xinyun Chen¹

Hanxiao Liu¹

Da Huang¹

Denny Zhou¹

Tengyu Ma^{1,2,†}

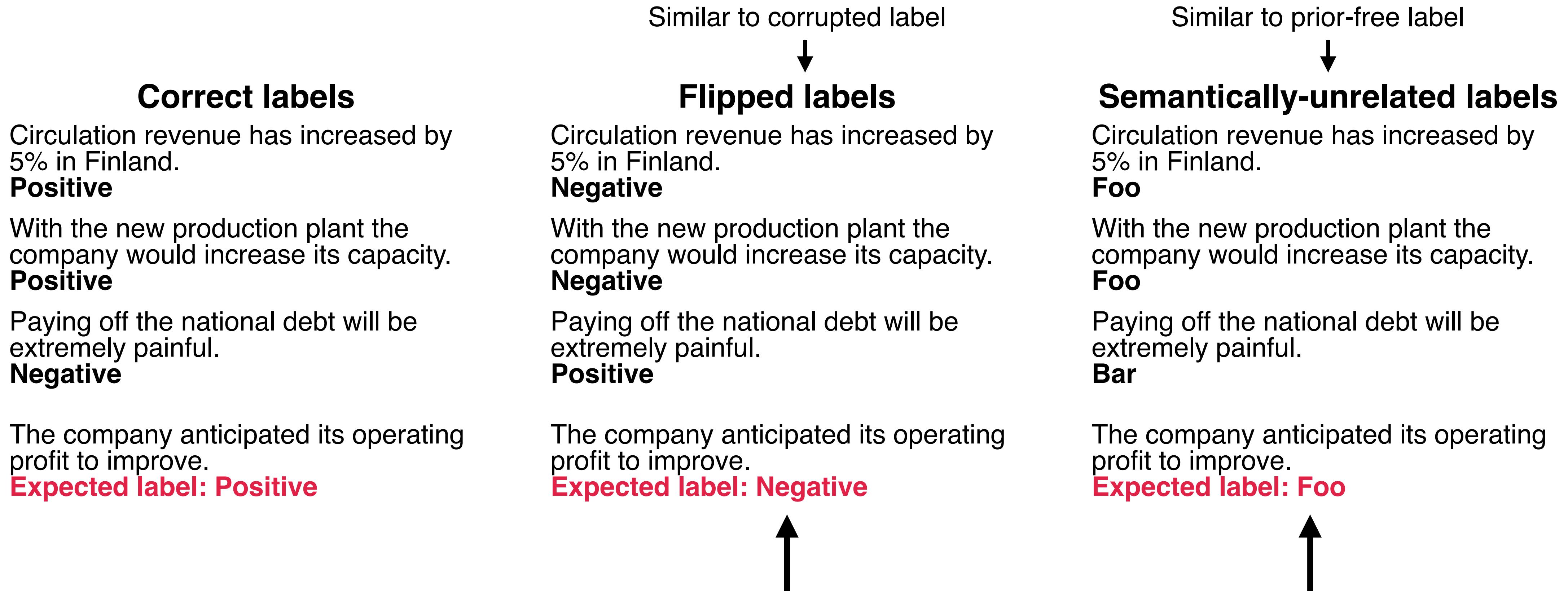
¹ Google Research, Brain Team

² Stanford University

³ Brown University

Our interpretation of the work

Research Question



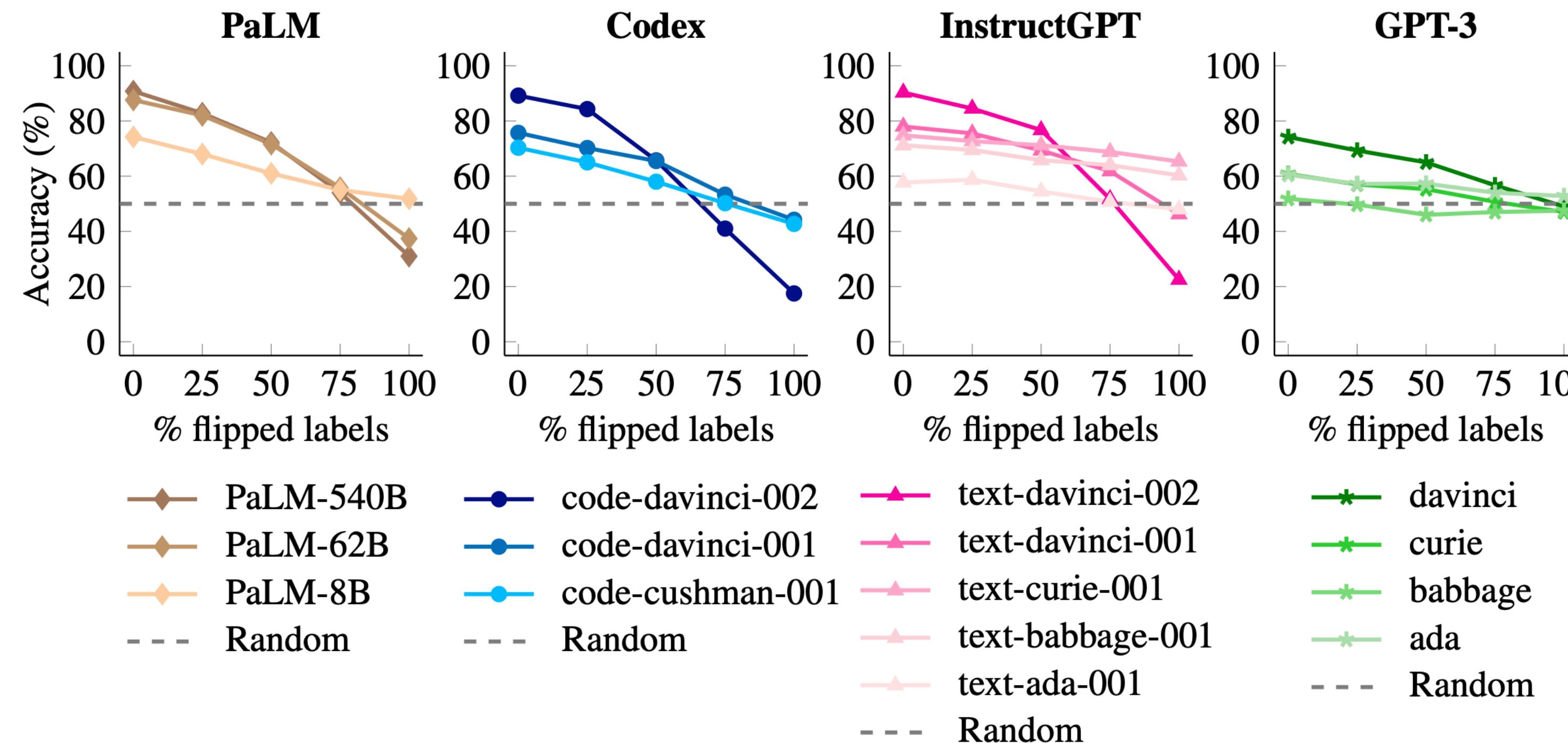
This may indicate the LMs are capable of learning new tasks on the fly.

How will diverse (extremely) large models react?

diverse (extremely) large models react?

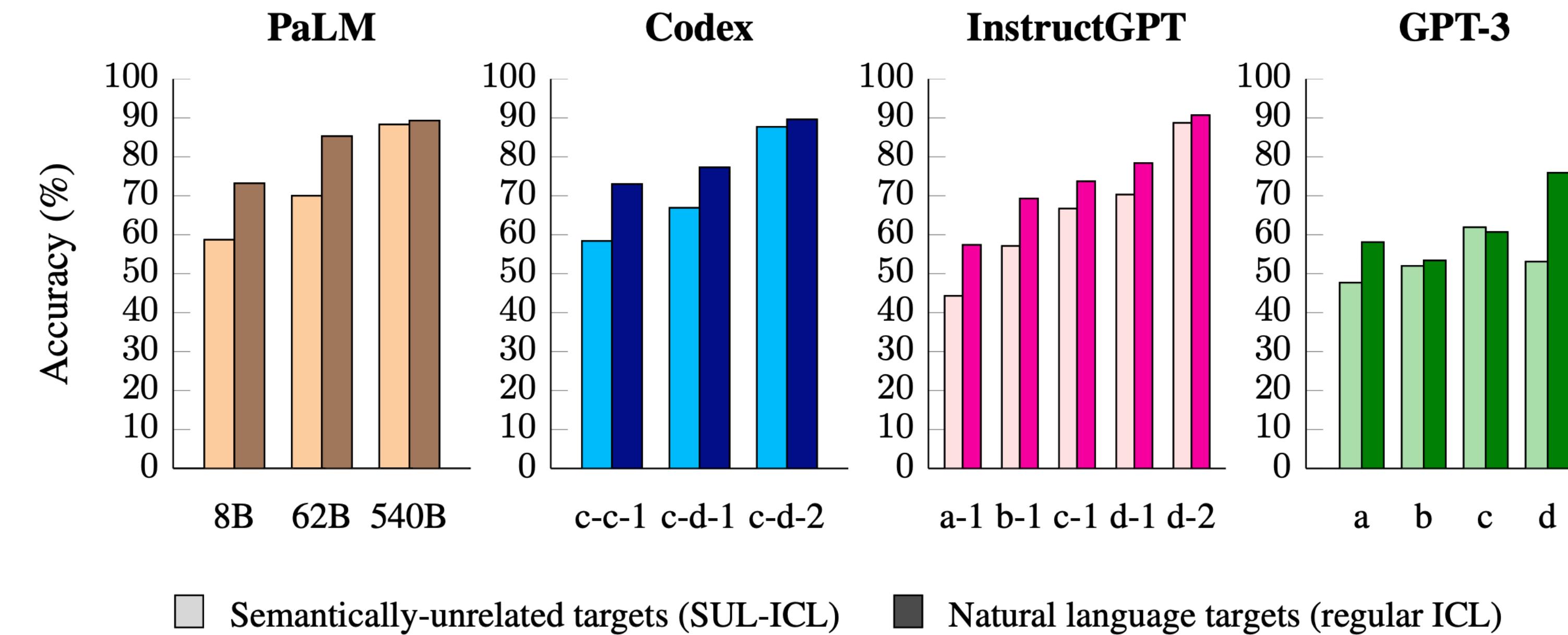
training prior.

Are the semantic priors overridden?



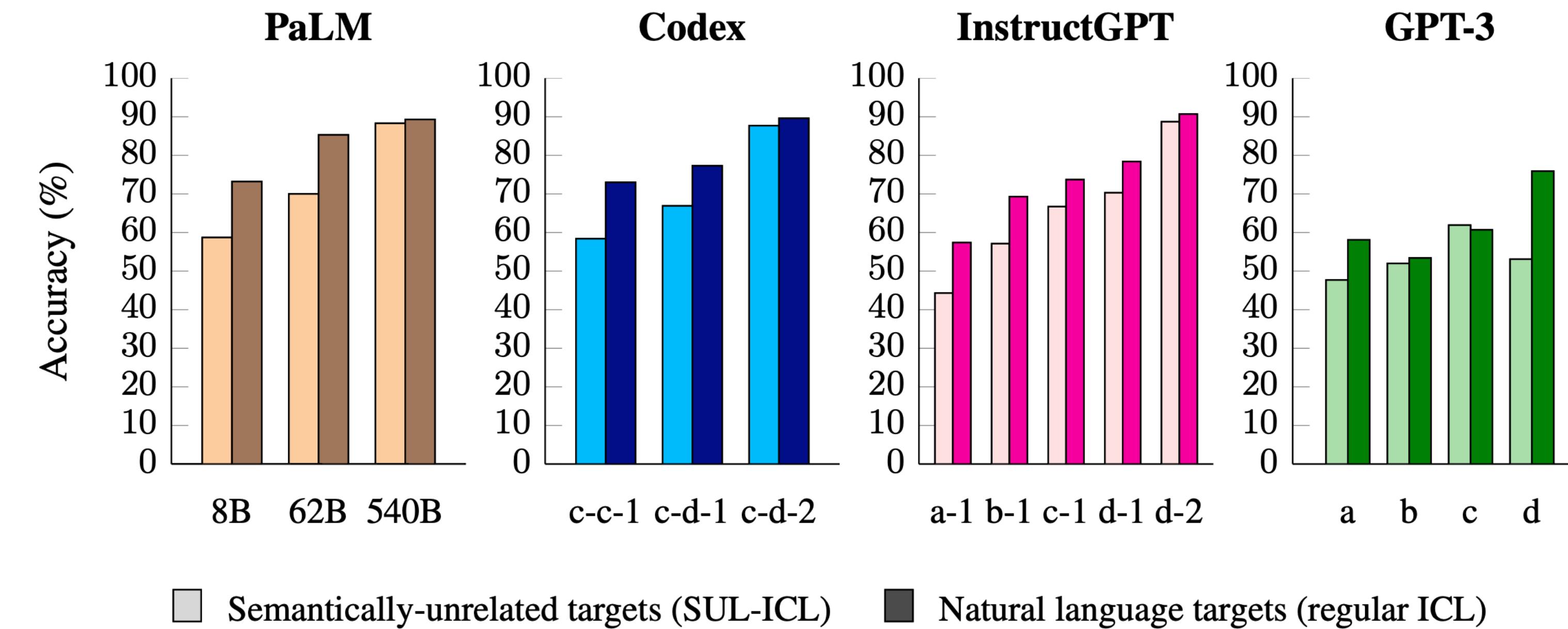
The larger models show greater ability to degrade using semantic priors if flipped label increased.

In-context learning with semantically unrelated labels



The ability to do tasks defined with SUL emerges with scale.

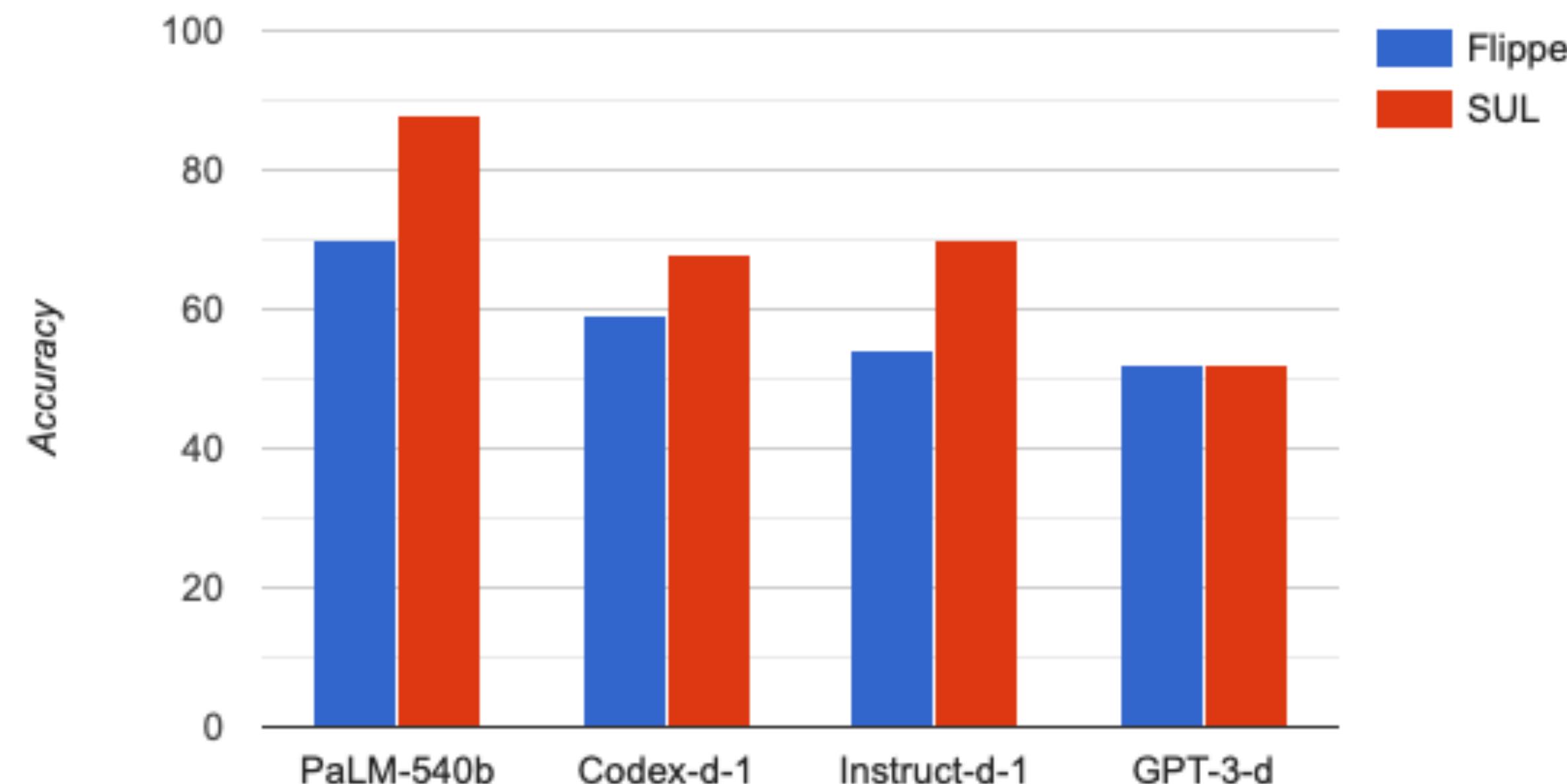
In-context learning with semantically unrelated labels



When using smaller models SUL results in a greater performance drop from using natural language targets

→ The result indicates that a smaller model heavily rely on priors, whereas a larger model benefits from the input-label mapping.

Flipped vs SUL



ICL with SUL outperforms ICL with flipped labels

The result indicates that when the impact of priors is minimized,
the impact of label correctness grows

Thoughts

1. Larger models tends to do better on overriding semantic priors. However, they do not fully override the prior.
2. The ability to do tasks defined with semantically unrelated labels seem to emerge with scale.
3. The impact of label correctness grows when the labels do not induce the task specific priors.

demonstration works as both **inducer of prior knowledge** and **training samples for model**, depending on the setting, the importance of certain role may change.

Revisiting the finding of Min et al.

- Correct demonstrations are **not required**
 - Not always, it depends on various factors, such as the model size, prompt template, etc.
- LMs are **not** learning new task on the fly — they already know the task (from **priors from training**).
 - It does seem to learn new tasks as SUL labels outperformed baseline.
- **Not every nonsense demonstration works:** the correct input distribution, the correct label space, and the correct format
 - We agree that these factors greatly matters
 - However, we argue that the correctness of label is also one of the core component

Talk outline

Sewon: Ground truth demonstrations are **not required** for ICL.

Junyeob: ICL **do** need ground truth demonstrations, and **can learn** newly-defined tasks.

Sewon: These results **do not rebut** our paper:
they either do not affect our findings, or lack evidence

Sewon: Our general findings **hold more broadly**
(generation/reasoning tasks, alignment)

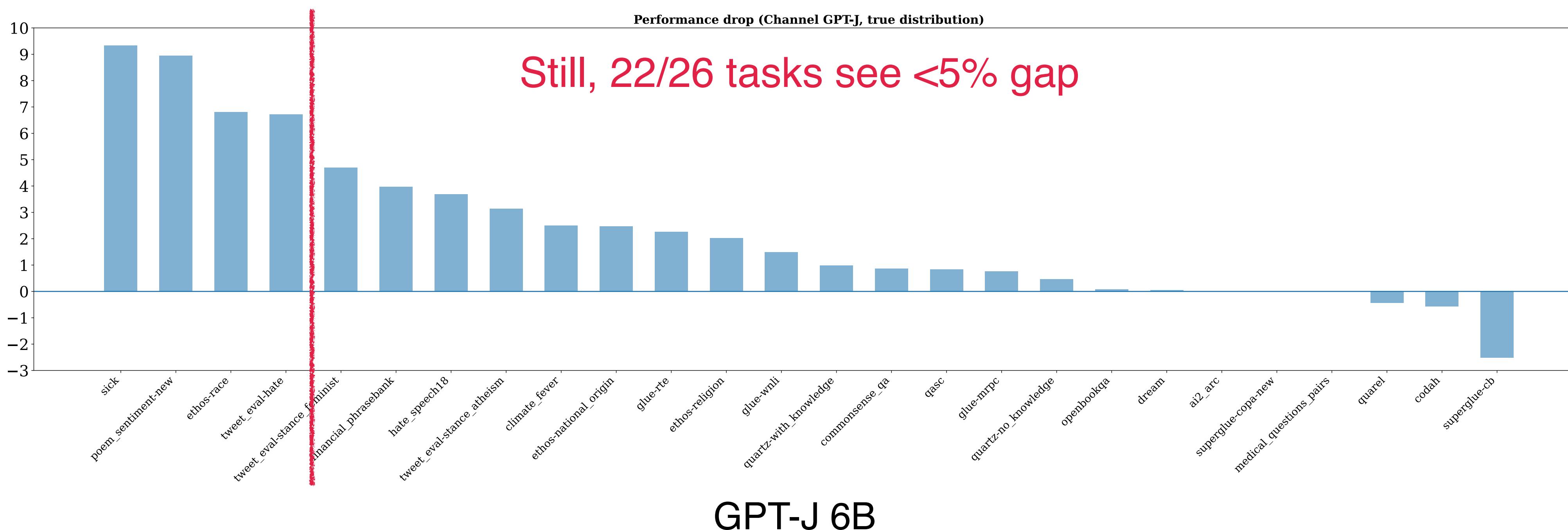
Junyeob: Yes, but in the end **small deterioration matters**

Junyeob & Sewon: Summary

Summary of Yoo et al.

Whether or not you need correct demonstration depends:

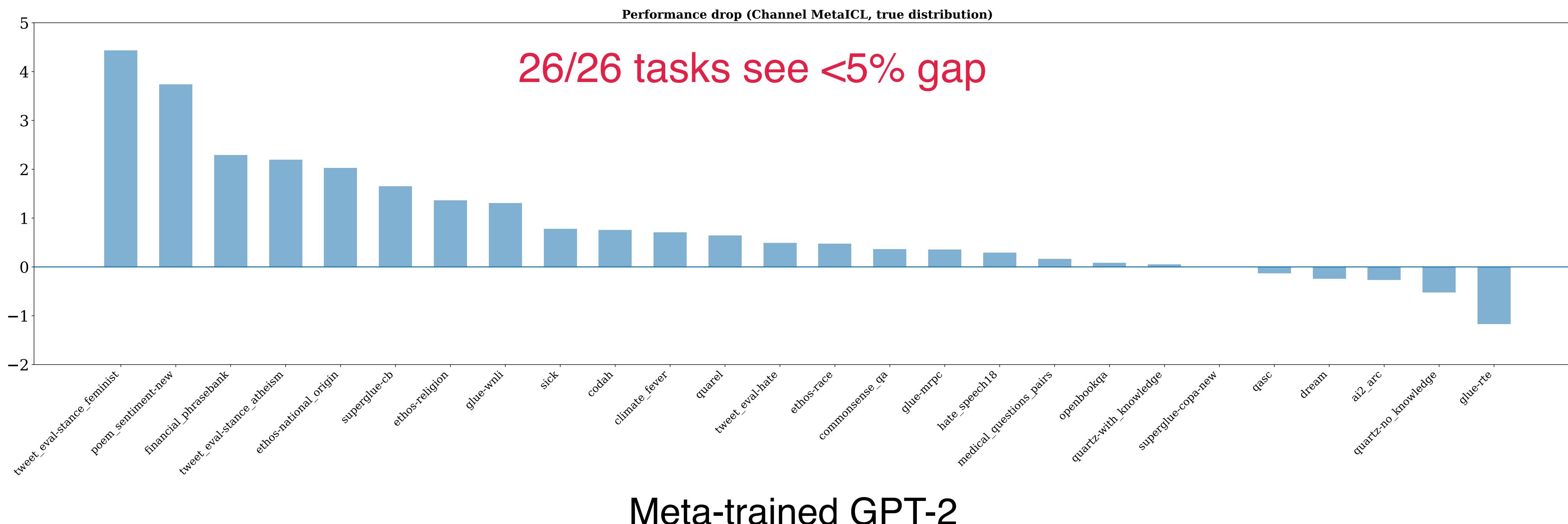
- Task → Consistent to our paper
- LM size (e.g., the findings do not hold for the largest GPT-3)
- Prompting method



Summary of Yoo et al.

Whether or not you need correct demonstration depends:

- Task → *Consistent to our paper*
- LM size (e.g., the findings do not hold for the largest GPT-3)
- Prompting method



Summary of Yoo et al.

More discussion on this will be in the next part

Whether or not you need correct demonstration depends:

- Task
 - LM size (e.g., the findings do not hold for the largest GPT-3)
 - Prompting method
-
- In most settings, the findings hold (with careful prompting)
 - The best learner still does not need correct labels

GPT-3 Experiments

Circulation revenue has increased by 5% in Finland.

Positive

Panostaja did not disclose the purchase price.

Neutral

No template

Paying off the national debt will be extremely painful.

Negative

Review: Circulation revenue has increased by 5% in Finland.

Sentiment: Positive

Review: Panostaja did not disclose the purchase price.

Sentiment: Neutral

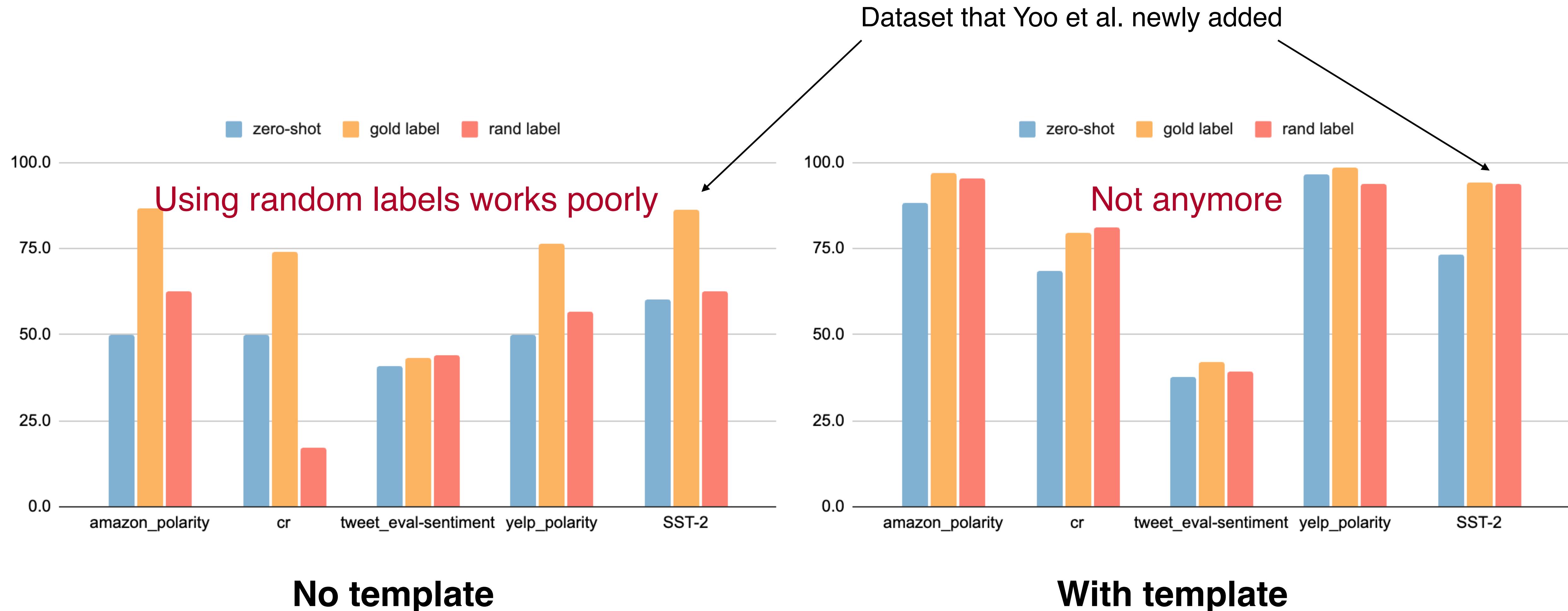
With template

(taken from Zhao et al. 2021)

Review: Paying off the national debt will be extremely painful.

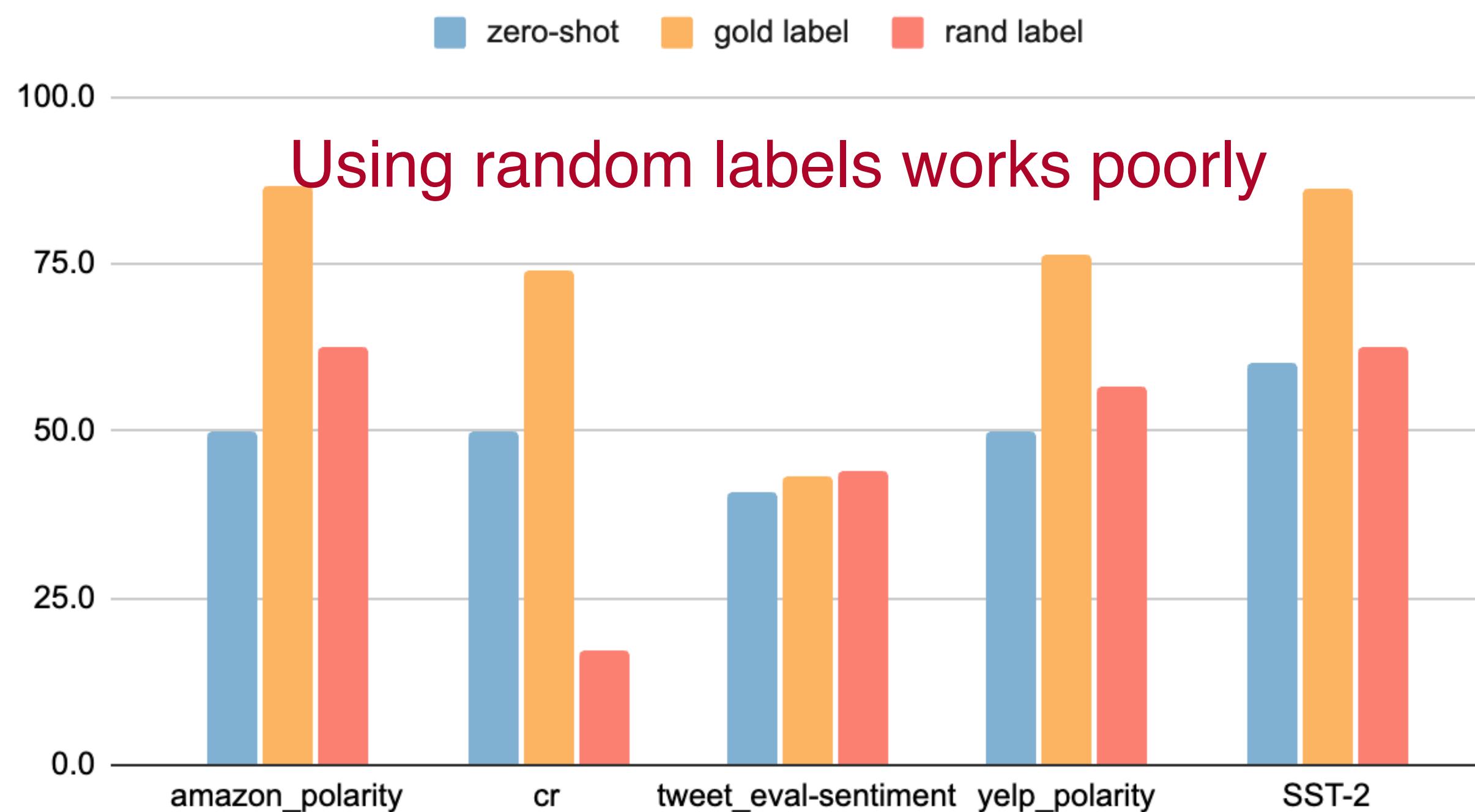
Sentiment: Negative

GPT-3 Experiments

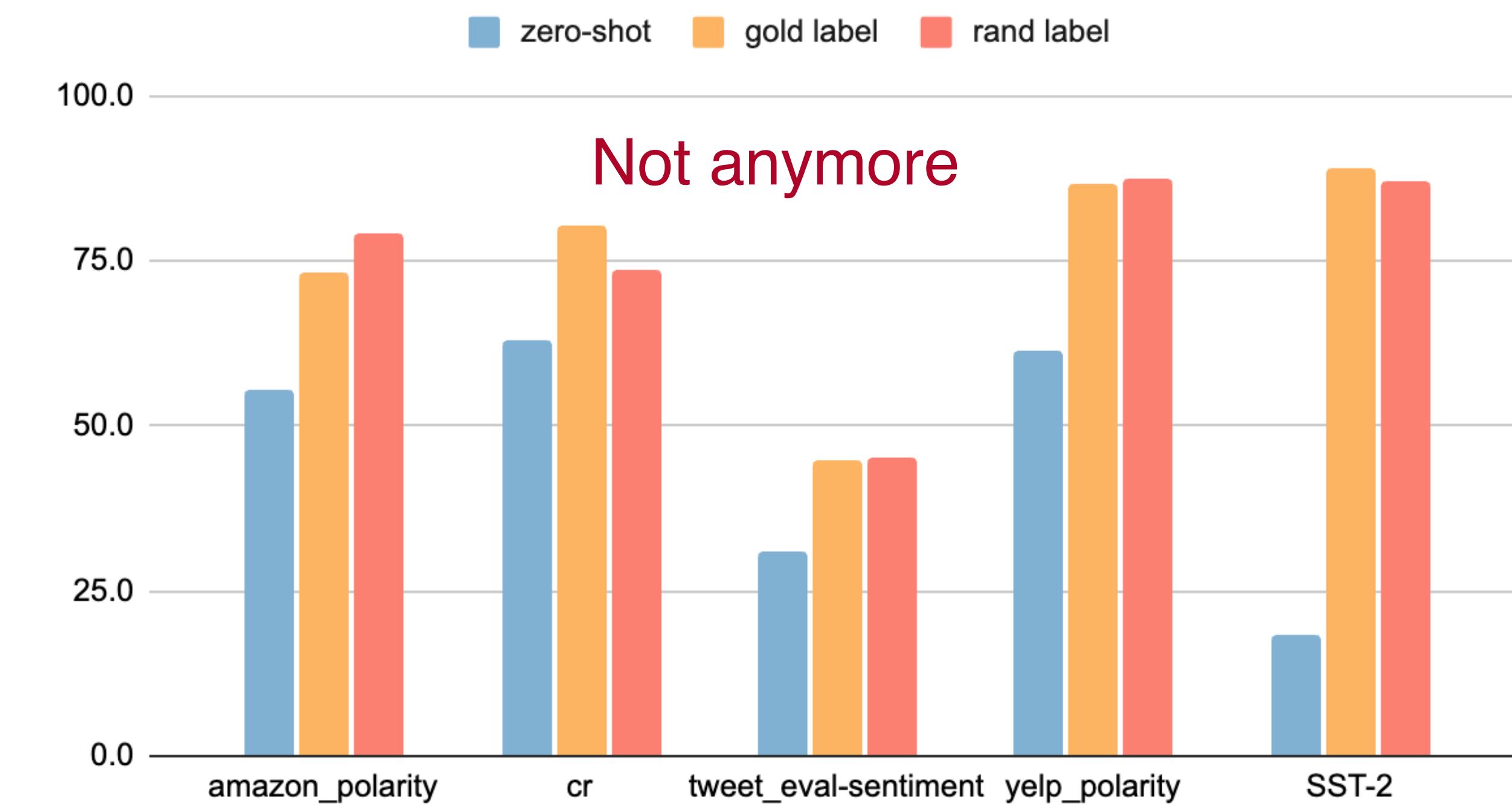


GPT-3 (davinci). Additional results not in the paper.

GPT-3 Experiments



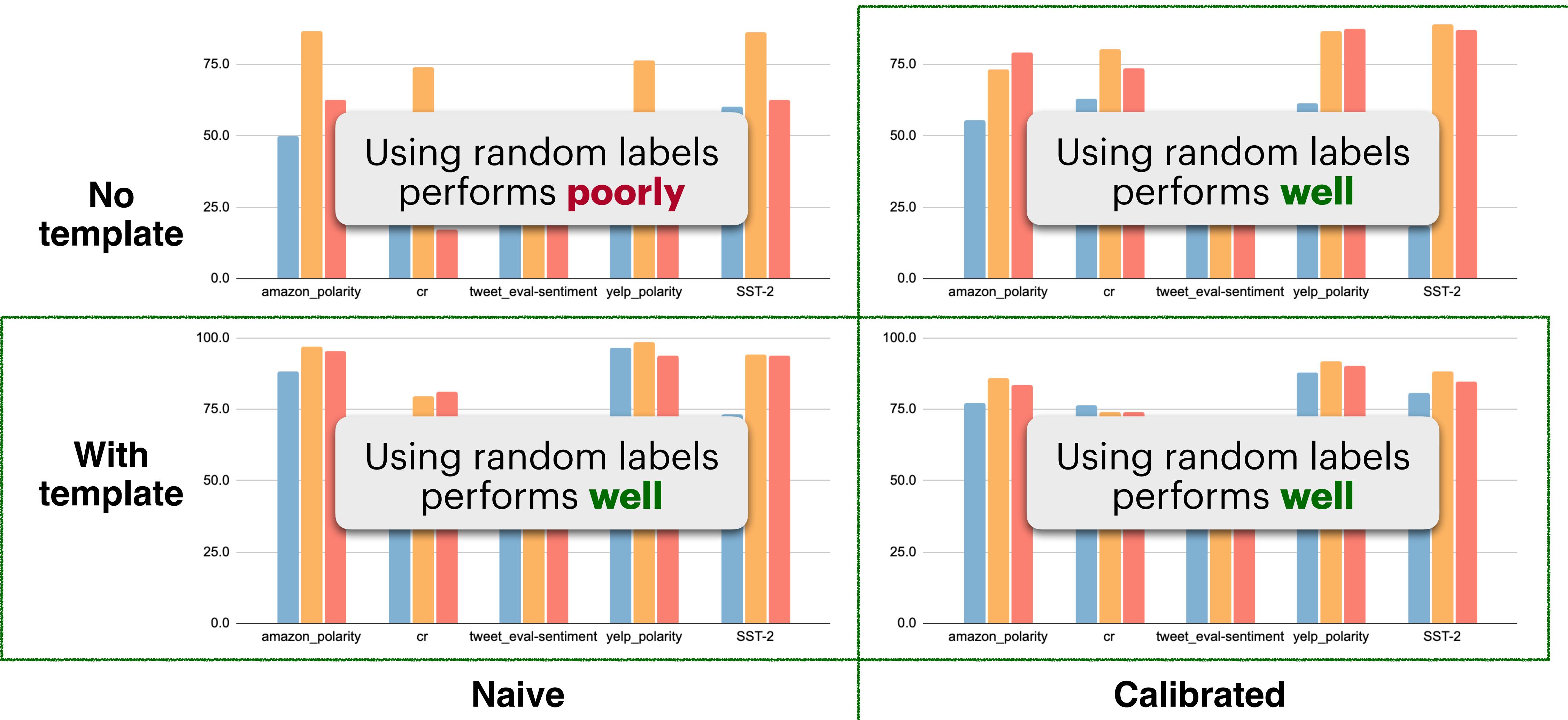
Naive



Calibrated
(Channel prompting)

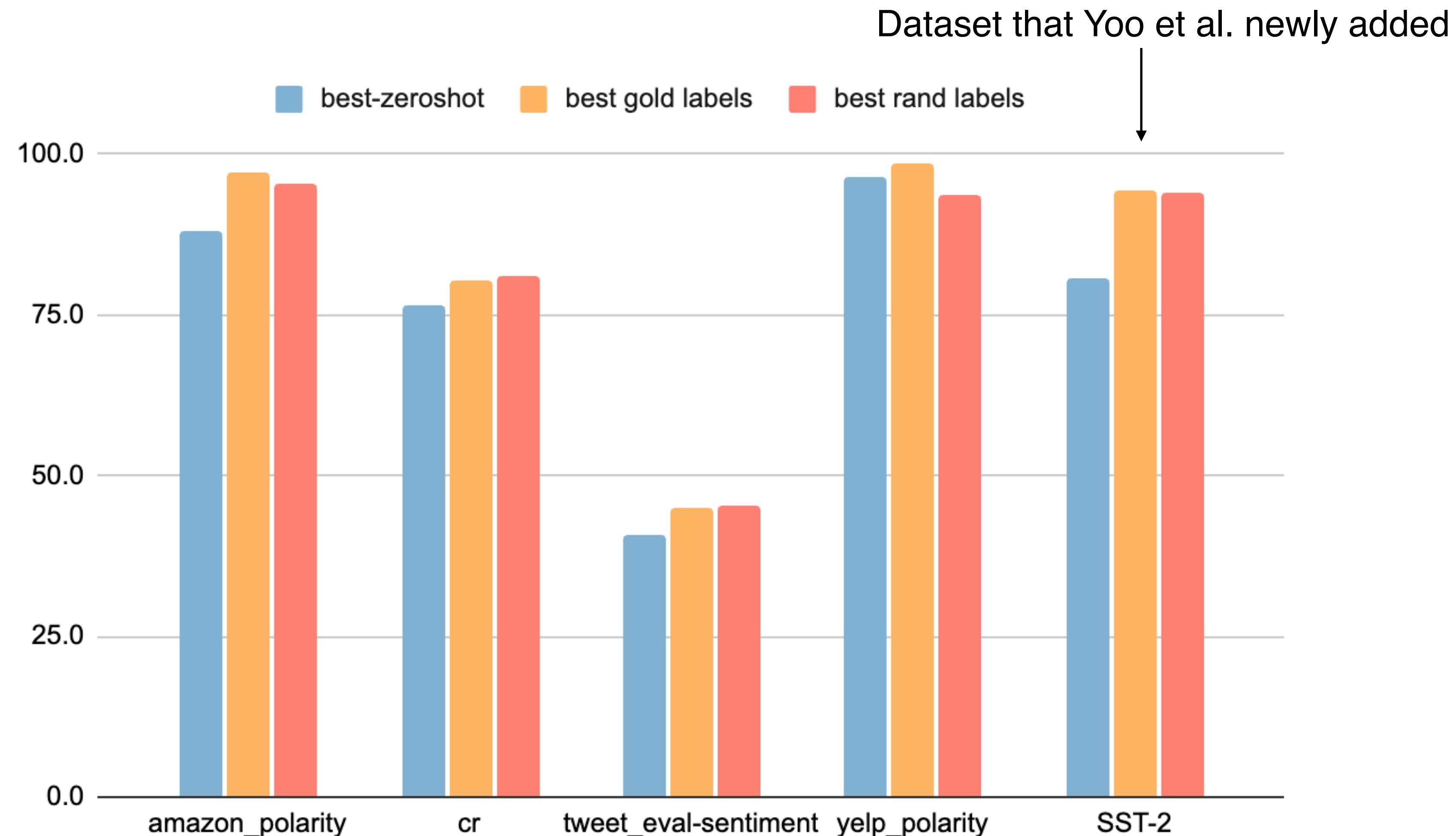
GPT-3 (davinci). No template. Additional results not in the paper.

Summary



Additional results not in the paper.

The best learner is not sensitive



Additional results not in the paper.

Summary

- Yoo et al: There are setups (e.g. not using a template, not using a prompting method) that ICL with random labels performs poorly.
- Our thoughts:
 - Valuable to find ICL w/ random labels is more sensitive than standard ICL.
 - This doesn't affect "**ICL doesn't require correctness**" as **ICL with random labels works with careful prompting**.
 - The best learner **still does not need** correct demonstrations.

Moving on to Wei et al.

- Our findings:
 - Correct demonstrations are not required when tasks align with priors from training
- New questions:
 - What if tasks do not align with priors from training?
 - Learning a new task vs. priors from training

Moving on to Wei et al.

ABSTRACT

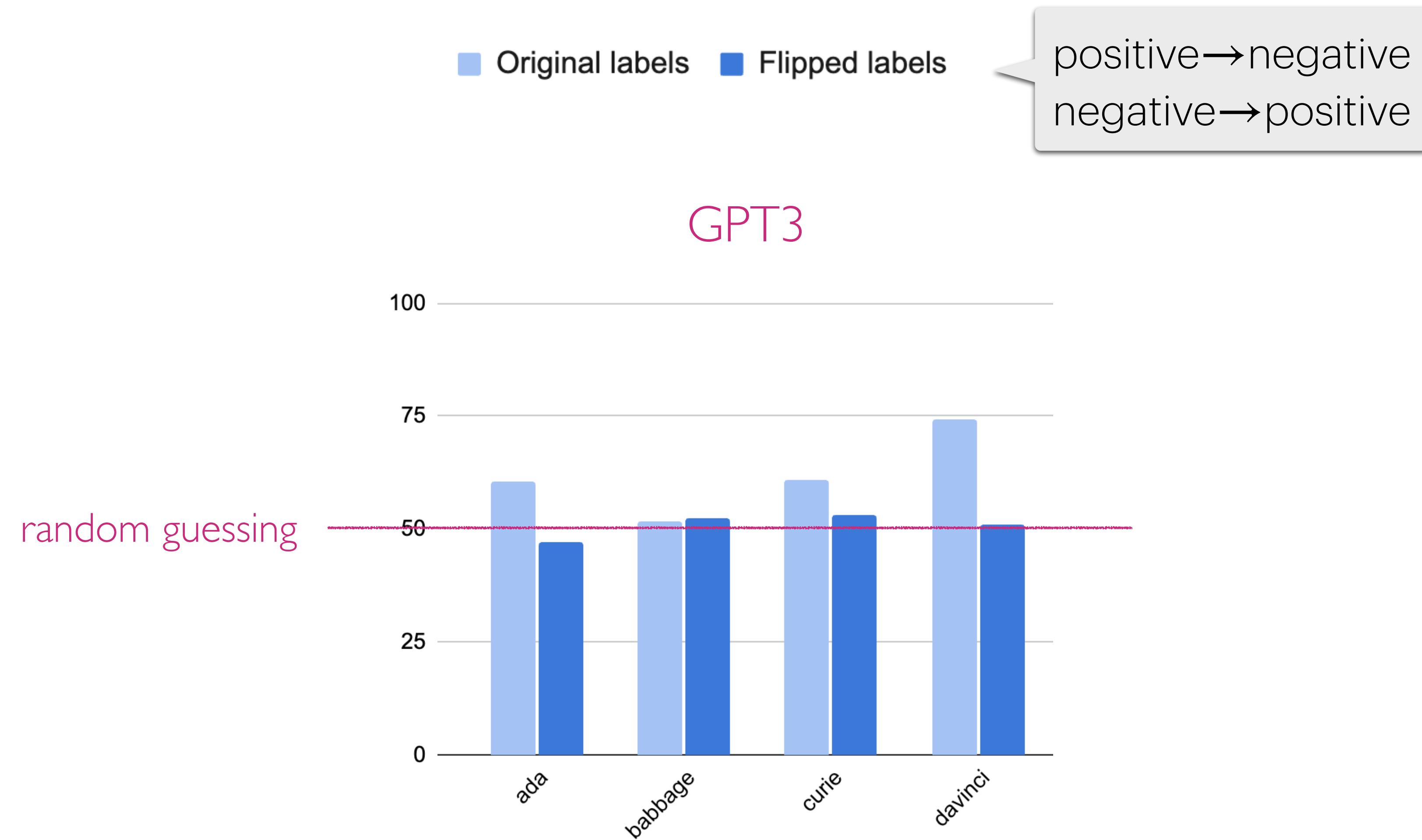
Large models can override semantic priors

We study how in-context learning (ICL) ~~priors versus input–label mappings~~. We emergent ability of model scale.

~~with flipped labels now that overriding semantic priors is an emergent ability of model scale. While small language models ignore flipped labels presented in-context and thus rely primarily on semantic priors from pretraining, large models can override semantic priors when presented with in-context exemplars that contradict priors, despite the stronger semantic priors that larger models may hold.~~ We next study *semantically-unrelated label ICL* (SUL-ICL), in which labels are semantically unrelated to their inputs (e.g., foo/bar instead of negative/positive), thereby forcing language models to learn the input–label mappings shown in in-context exemplars in order to perform the task. ~~The ability to do SUL-ICL also emerges primarily with scale, and large-enough language models can even perform linear classification in a SUL-ICL setting. Finally, we evaluate instruction-tuned models and find that instruction tuning st and the capacity to learn input–label ma~~

The ability to do [tasks defined with semantically unrelated labels] emerges primarily with scale

Overriding priors



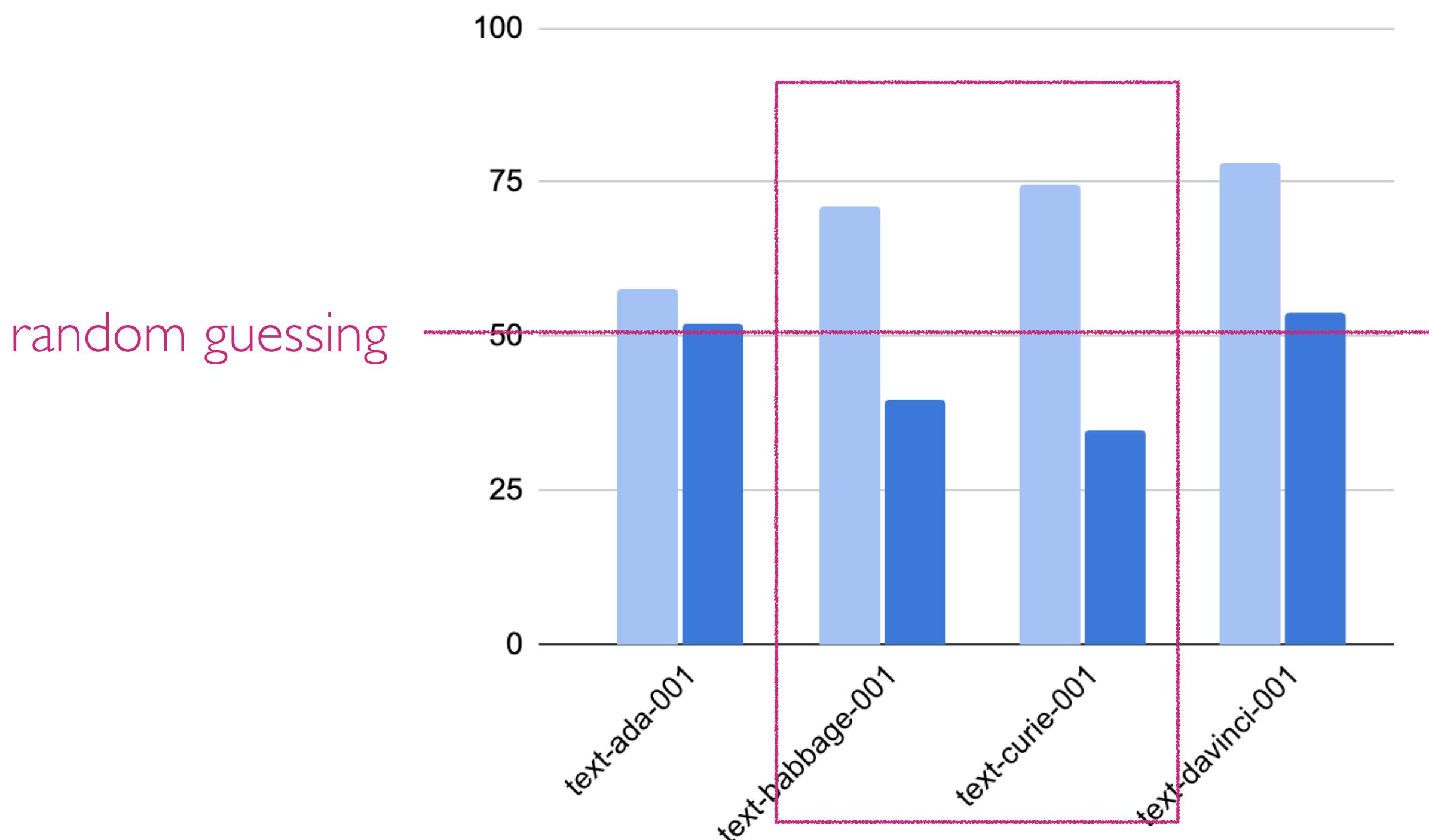
Graph redrawn by the speaker, numbers obtained from the authors of the original paper

Overriding priors

Original labels Flipped labels

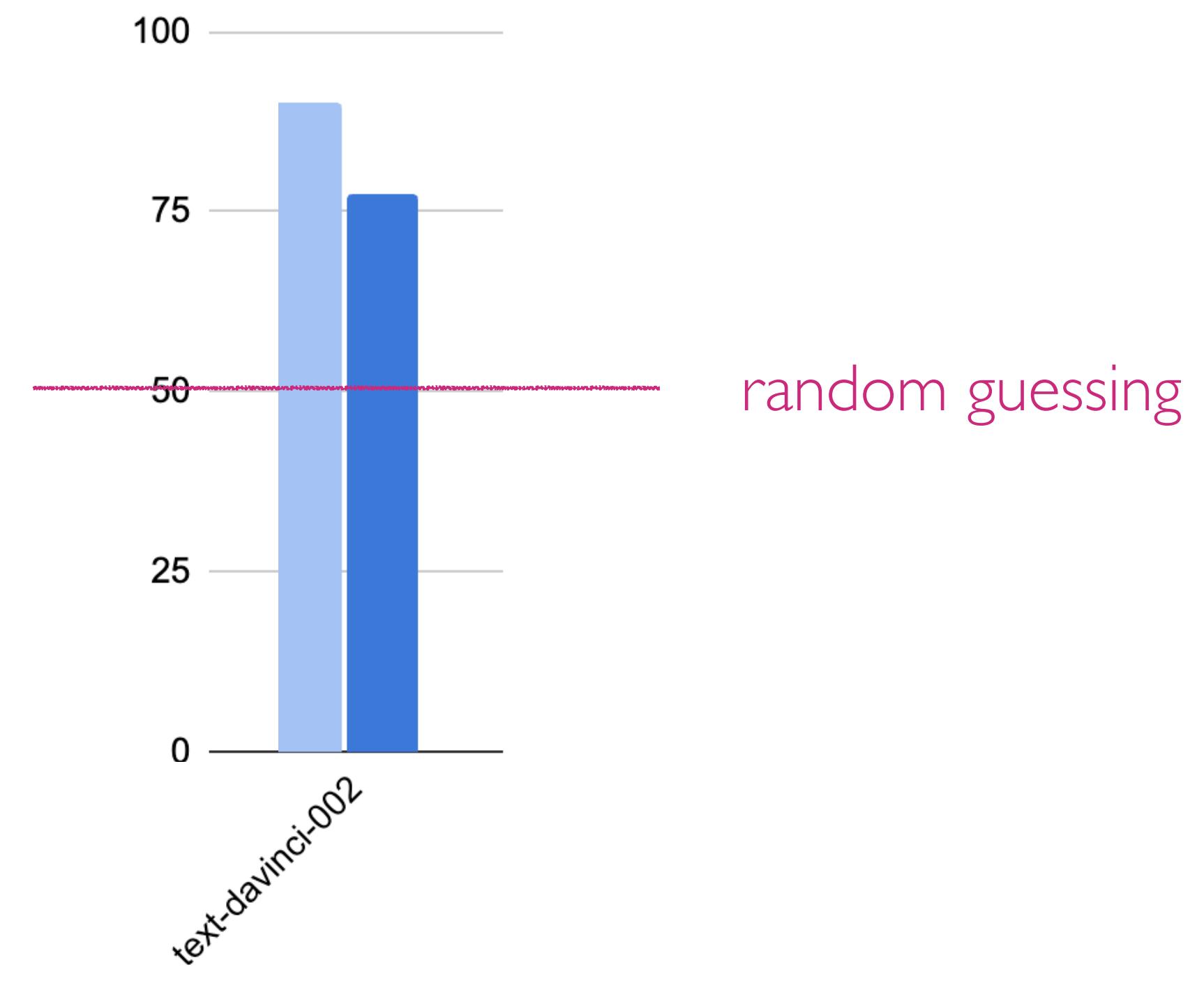
positive → negative
negative → positive

Instruct GPT3 001



rather do the original task

Instruct GPT3 002
(released after our paper)



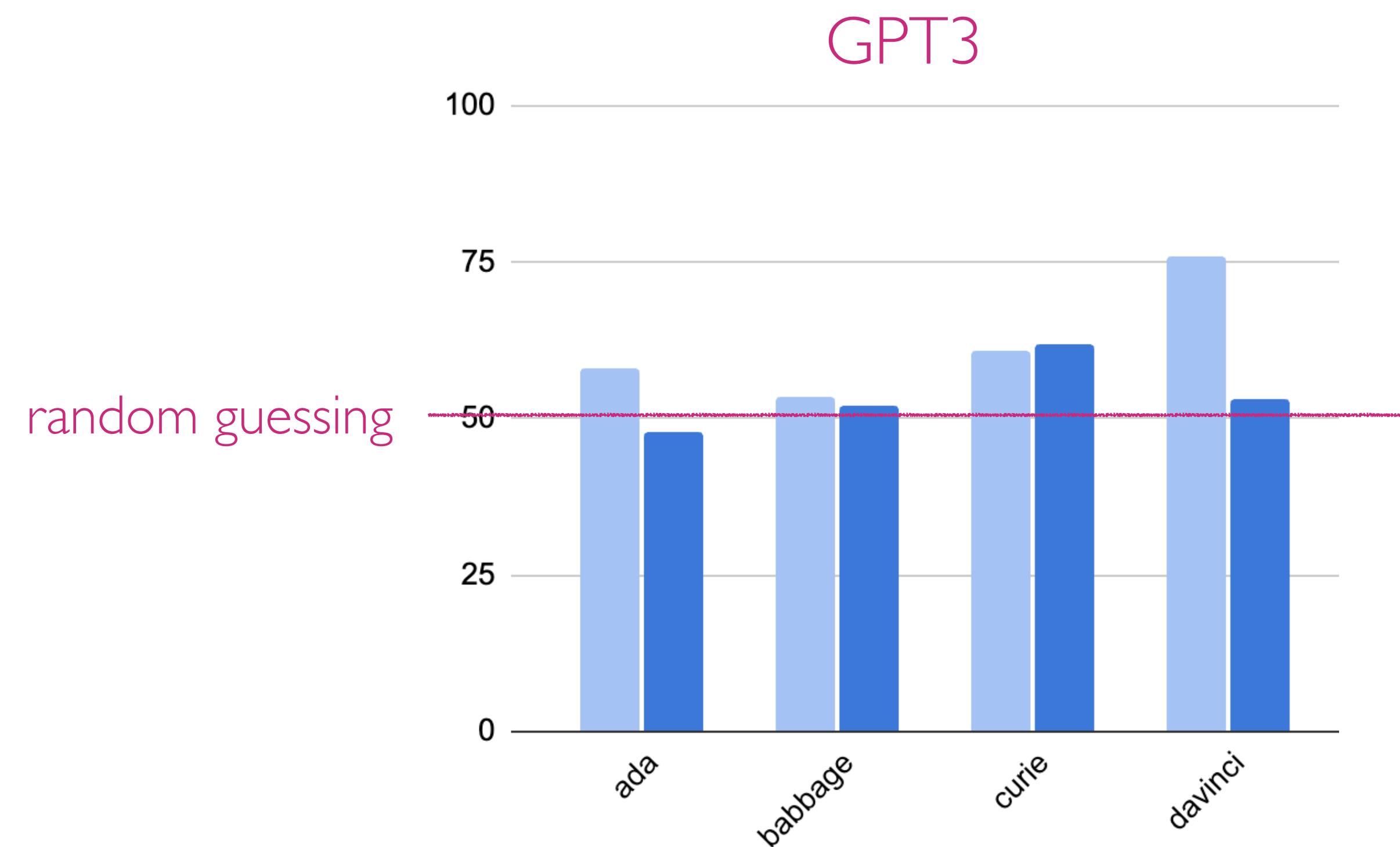
random guessing

Graph redrawn by the speaker, numbers obtained from the authors of the original paper

Semantically unrelated labels

■ Original labels ■ Semantically-unrelated labels

positive → foo
negative → bar



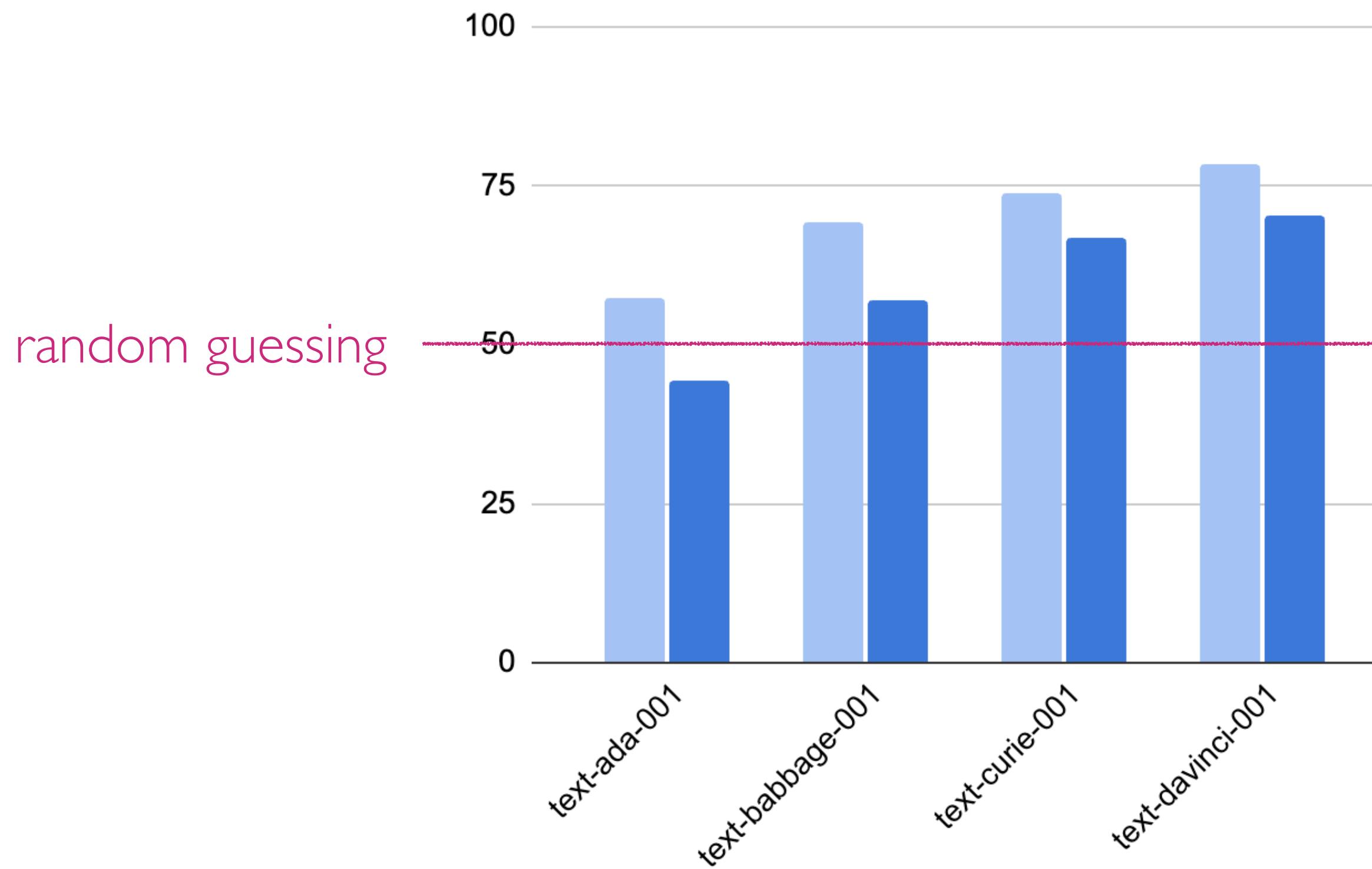
Graph redrawn by the speaker, numbers obtained from the authors of the original paper

(Unknown) instruction-tuning helps

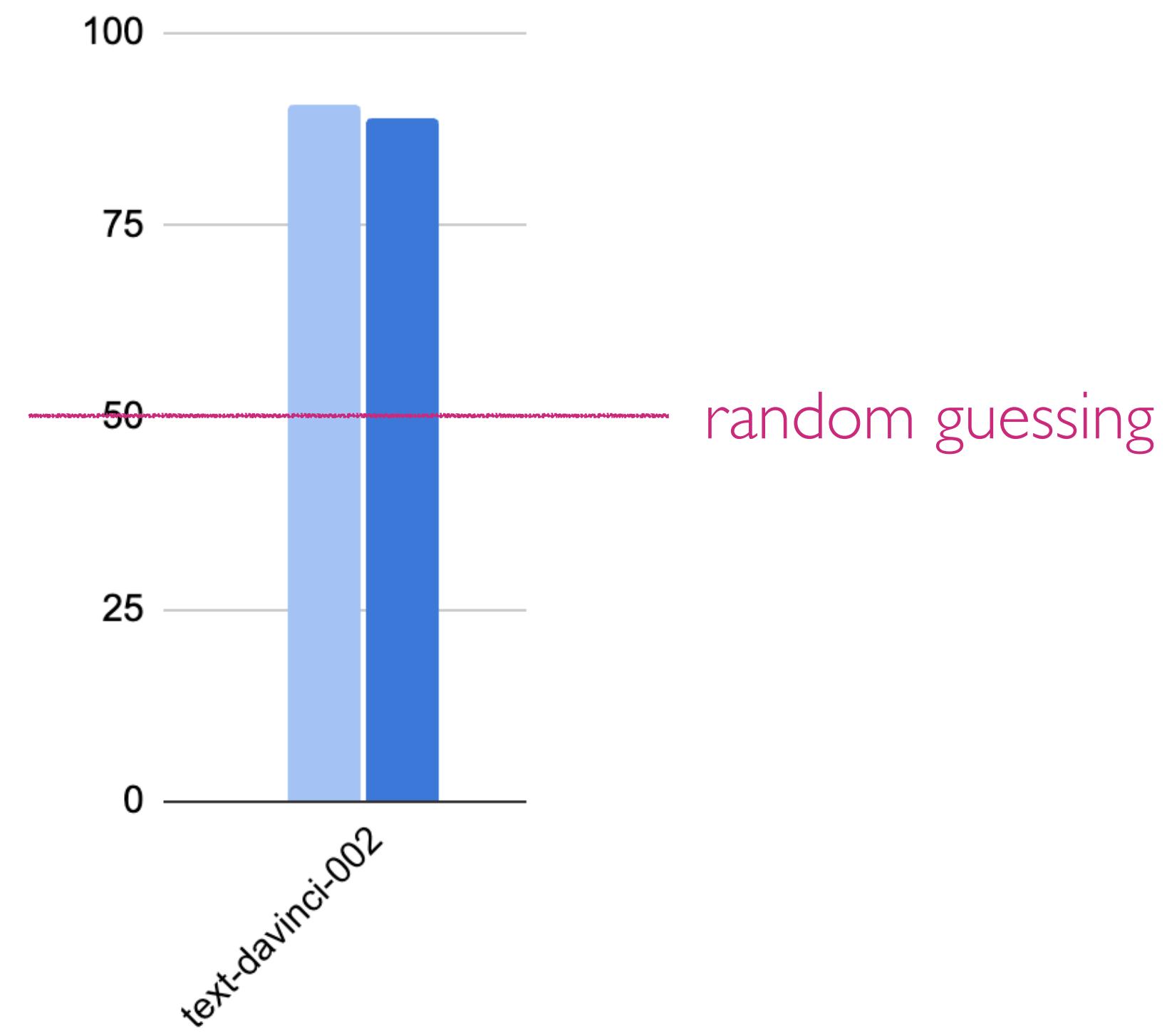
■ Original labels ■ Semantically-unrelated labels

positive → foo
negative → bar

Instruct GPT3 001



Instruct GPT3 002
(released after our paper)

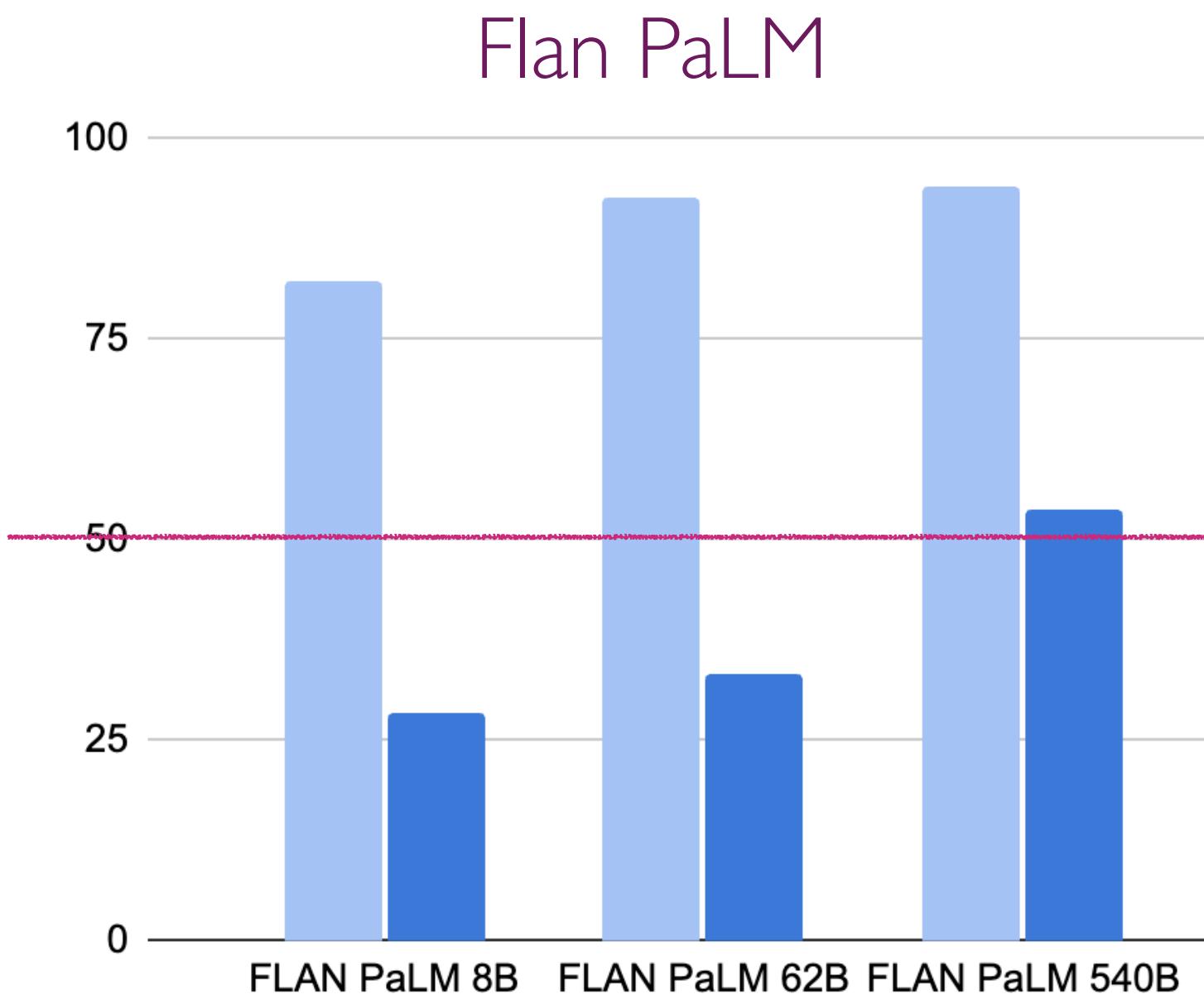


Graph redrawn by the speaker, numbers obtained from the authors of the original paper

Known instruction tuning doesn't

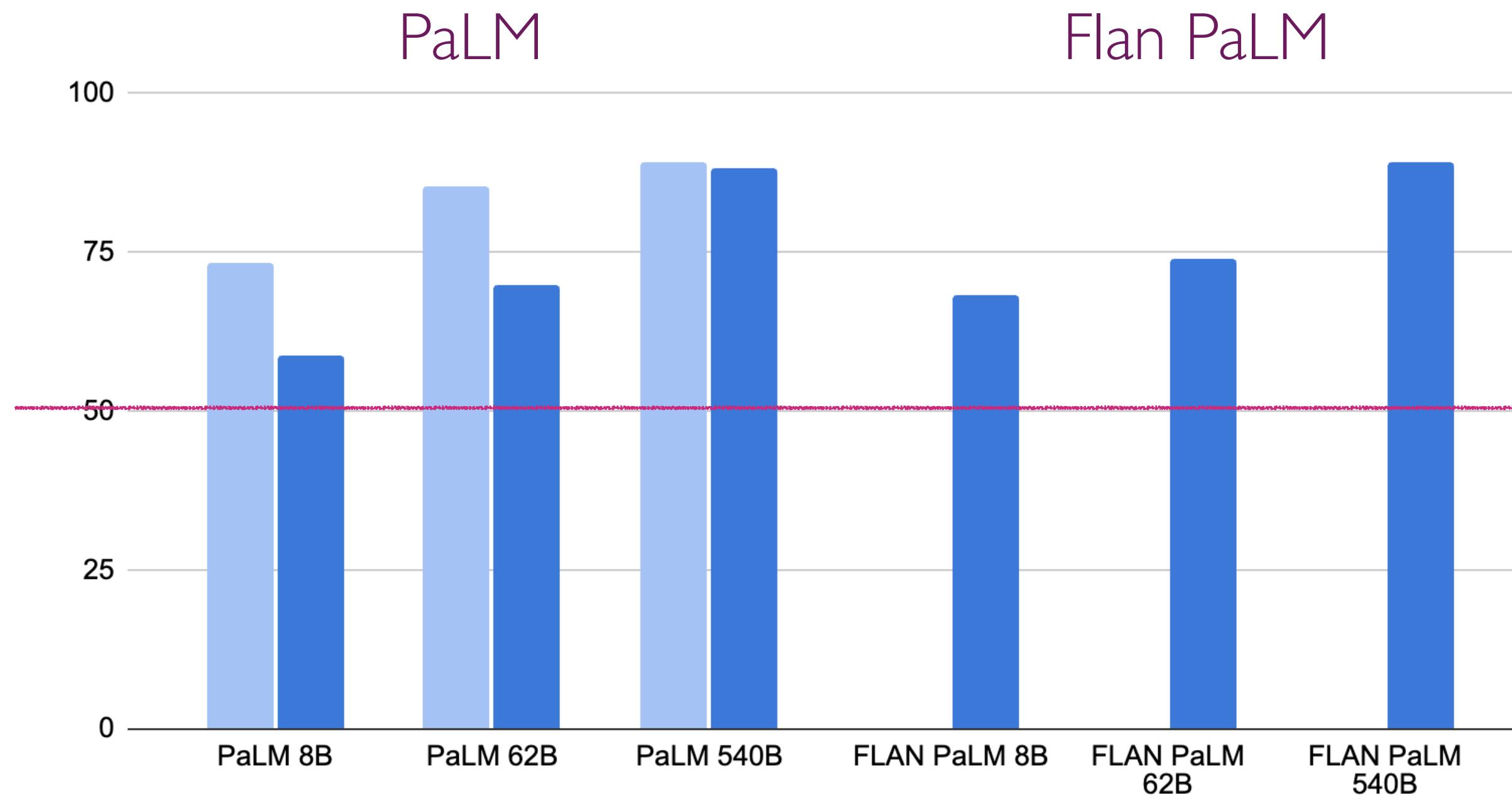
positive → negative
negative → positive

Original labels Flipped labels



positive → foo
negative → bar

Original labels Semantically-unrelated labels



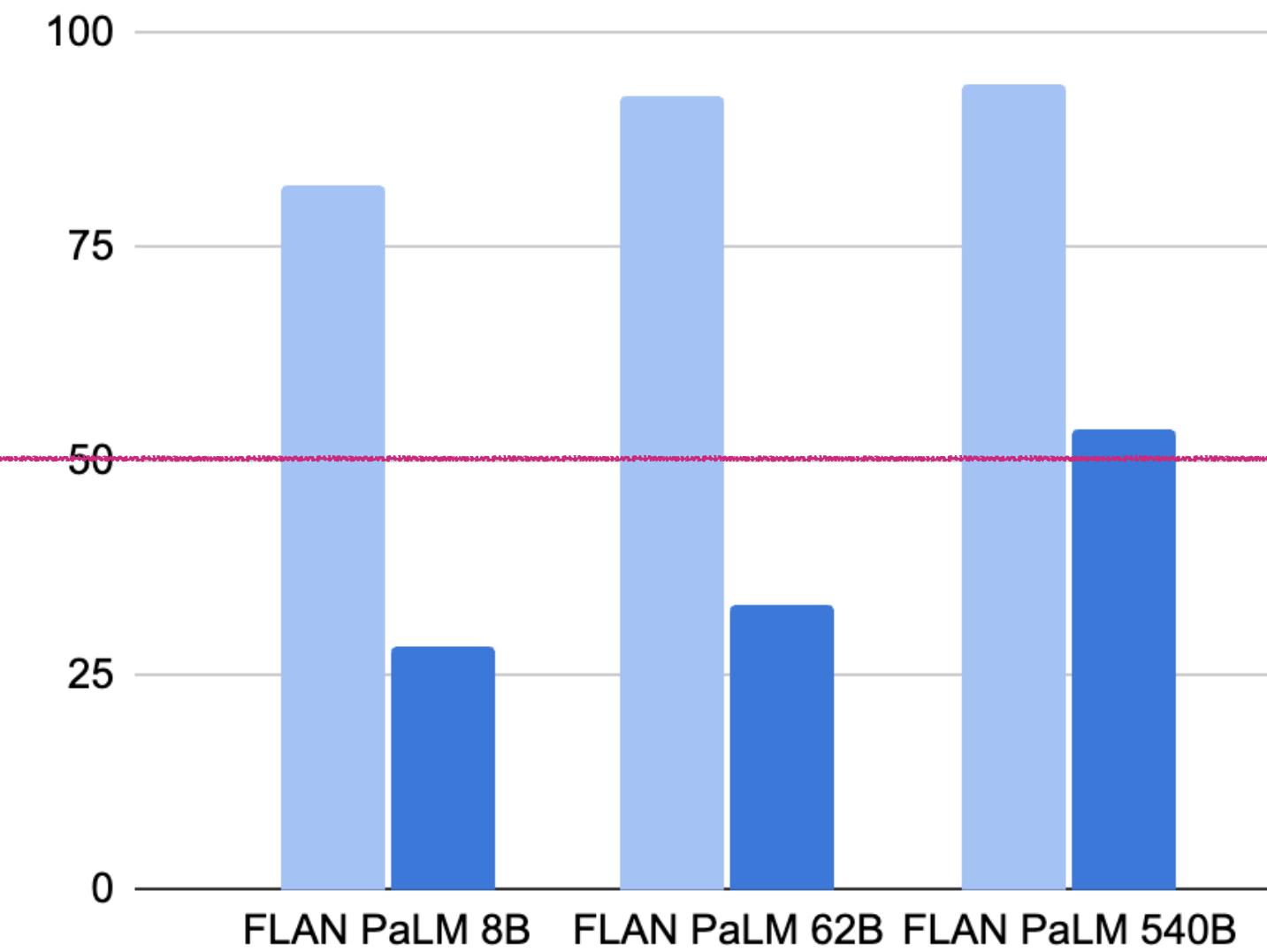
6 datasets, one set of semantically-unrelated labels

Known instruction tuning doesn't

positive→negative
negative→positive

Original labels Flipped labels

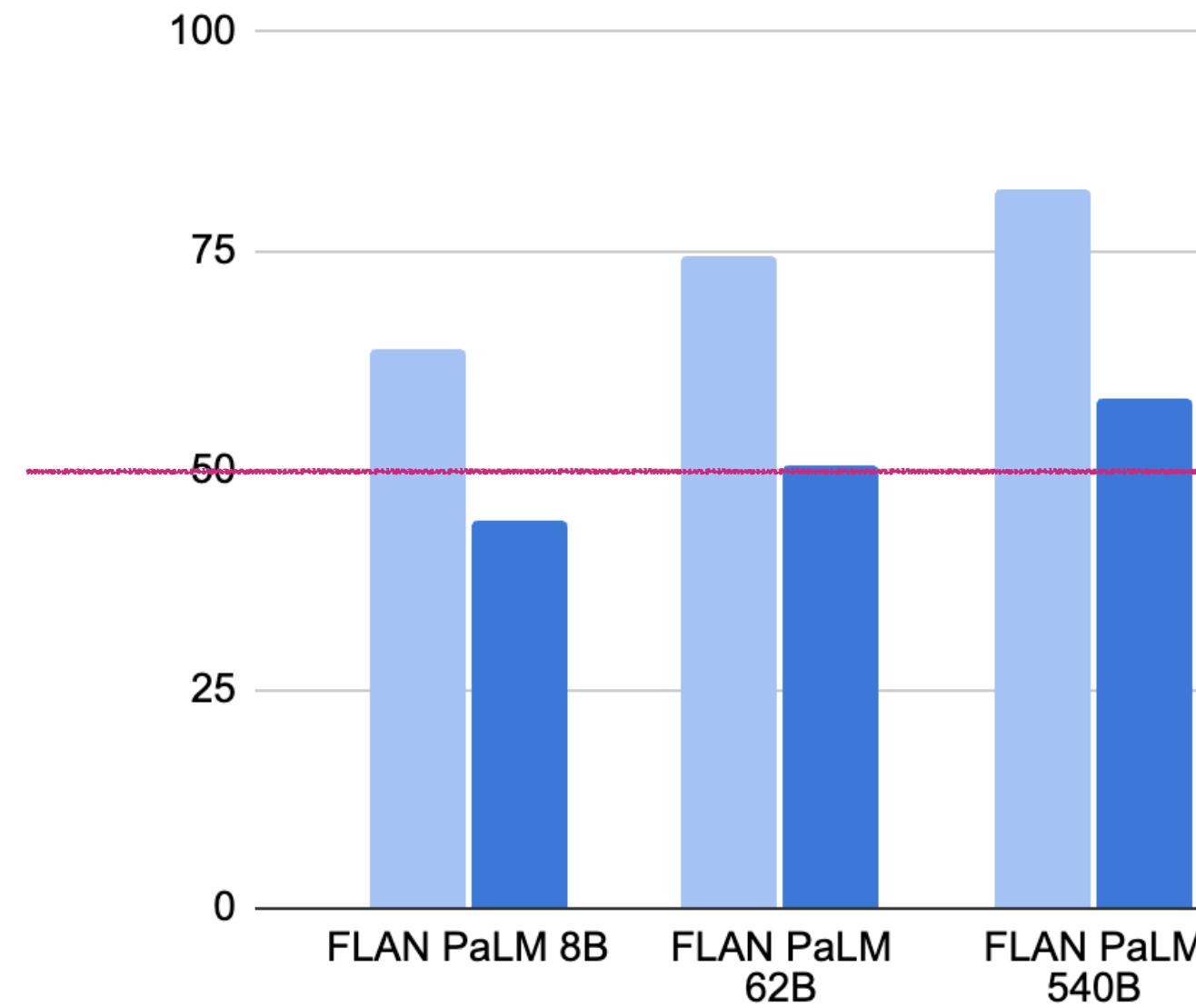
Flan PaLM



positive→foo
negative→bar

Original labels Semantically-unrelated labels

Flan PaLM



11 datasets, three sets of semantically-unrelated labels

Summary

1. Overriding semantic priors emerges with scale?
 - No models perform the task with flipped labels better than random.
 2. Performing tasks defined with semantically unrelated labels emerges with scale?
 - Mixed evidence, but hard to say LMs can do the task.
- Our primary focus is on raw LMs.
 - Even with instruction-tuned models, results are mixed.

Summary

1. Overriding semantic priors emerges with scale?
 - No models perform the task with flipped labels better than random.
Instruction-tuning makes it worse (except 002).
 2. Performing tasks defined with semantically unrelated labels emerges with scale?
 - Mixed evidence, but hard to say LMs can do the task. **Private instruction-tuning helps, but known instruction-tuning doesn't.**
- **Our primary focus is on raw LMs.**
 - **Even with instruction-tuned models, results are mixed.**

Summary & thoughts

**Learning unseen tasks isn't reliable enough, but is often possible.
When is it possible, and when is it not?**

Circulation revenue has increased by 5% in Finland.

Bar

With the new production plant the company would increase its capacity.

Bar

Paying off the national debt will be extremely

Foo

The company anticipated its operating profit to improve.

Expected label: Bar

Sentiment classification

+

**Mapping
(positive→foo, negative→bar)**

Compositional generalization, rather than learning a brand new task?

Talk outline

Sewon: Ground truth demonstrations are **not required** for ICL.

Junyeob: ICL **do** need ground truth demonstrations, and **can learn** newly-defined tasks.

Sewon: These results **do not rebut** our paper:
they either do not affect our findings, or lack evidence

Sewon: Our general findings **hold more broadly**
(generation/reasoning tasks, alignment)

Junyeob: Yes, but in the end **small deterioration matters**

Junyeob & Sewon: Summary

Research question

Classification/Multi-choice tasks 

What about generation/reasoning tasks? 

Chain-of-thought

Demonstrations

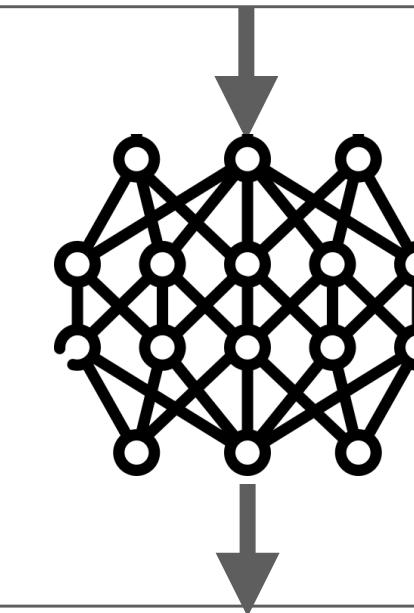
Test input

Question: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left?

Answer: Originally, Leah had 32 chocolates and her sister had 42. So in total they had $32 + 42 = 74$. After eating 35, they had $74 - 35 = 39$ pieces left in total. The answer is 39.

Question: Julie is reading a 120-page book. Yesterday, she was able to read 12 pages and today, she read twice as many pages as yesterday. How many pages are left?

Answer:



Julie is reading a 120-page book. She read 12 pages yesterday and 24 pages today. So she read a total of $12 + 24 = 36$ pages. She has $120 - 36 = 84$ pages left. The answer is 84.



Chain-of-thought (random)

Demonstrations
w/ random output

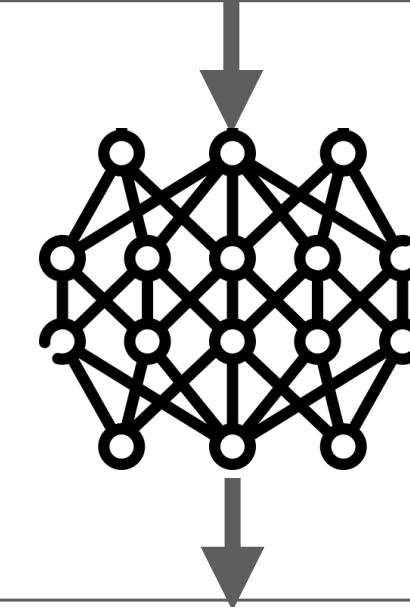
Test input

Question: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left?

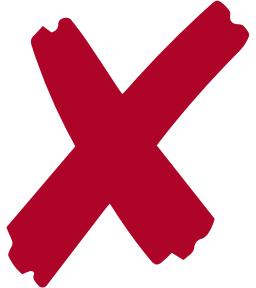
Answer: The popsicles cost $0.25 * 20 = 5$ dollars. The ice cream bars cost $0.5 * 4 = 2$ dollars.
He paid $5 + 2 = 7$ dollars. The answer is 7.

Question: Julie is reading a 120-page book. Yesterday, she was able to read 12 pages and today, she read twice as many pages as yesterday. How many pages are left?

Answer:



Luisa earned \$480. The math award costs \$15 each. $\$15 + \$45 = \$60$. The film costs \$20. $\$480 - \$60 + \$20 = \385 . The number of lessons is $100 - 15 + 45 - 60 + 20 = 10$. Luisa earned \$385 in 10 lessons. The answer is \$3850.



Chain-of-thought (incorrect)

Demonstrations
w/ incorrect output

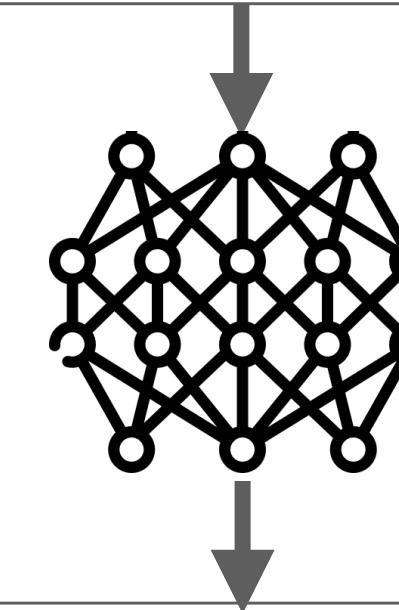
Test input

Question: Leah had 32 chocolates and her sister had 42. If they ate 35, how many pieces do they have left?

Answer: Originally, Leah had 32 chocolates and her sister had 42. So her sister had $42 - 32 = 10$ more than Leah has. After eating 35, they have $10 + 35 = 45$ in total. The answer is 45.

Question: Julie is reading a 120-page book. Yesterday, she was able to read 12 pages and today, she read twice as many pages as yesterday. How many pages are left?

Answer:



Yesterday, Julie read 12 pages. Today, she read $12 * 2 = 24$ pages. So she read a total of $12 + 24 = 36$ pages. Now she needs to read $120 - 36 = 84$ more pages. The answer is 84.



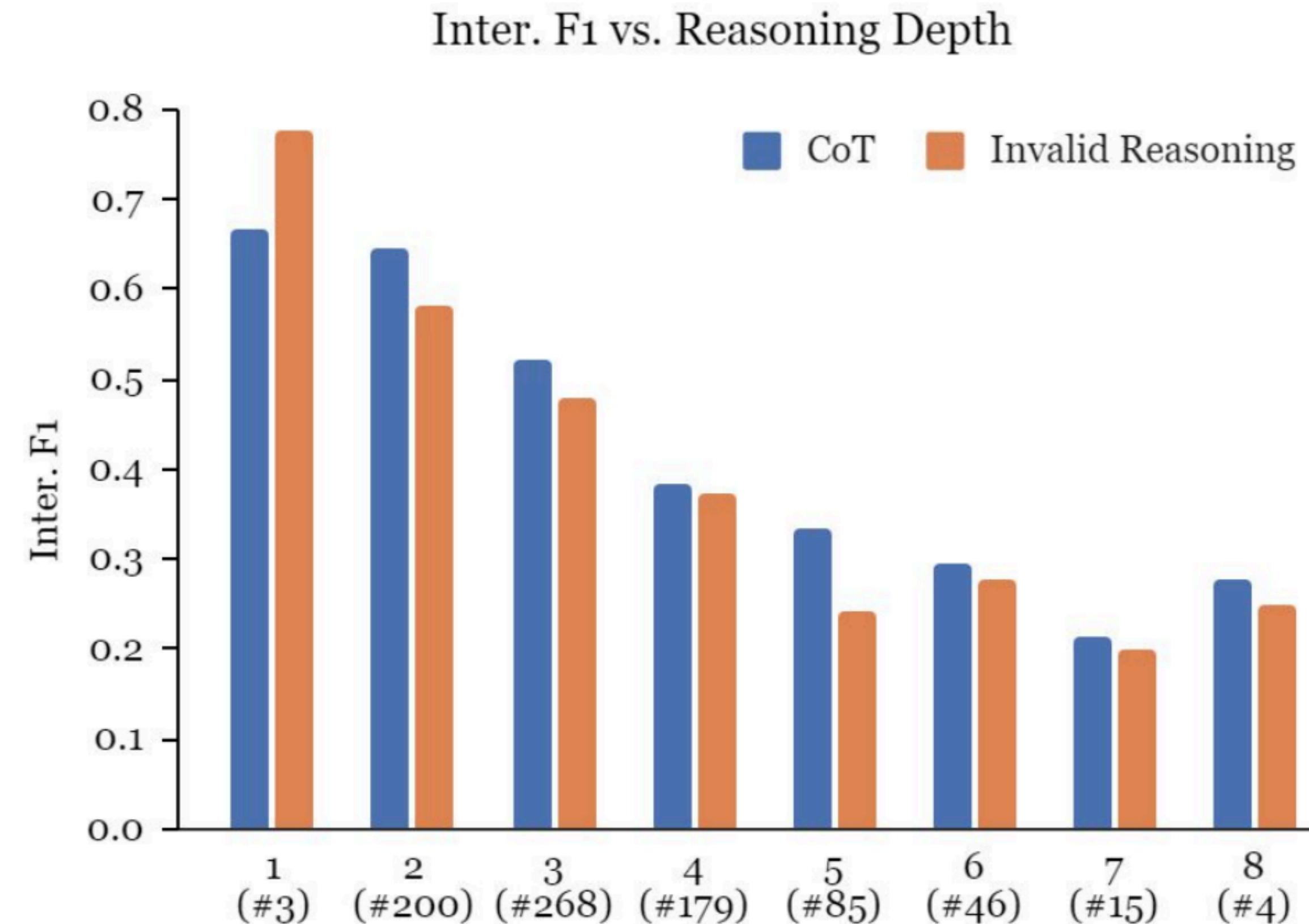
Chain-of-thought (correct vs. incorrect)

Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters

Boshi Wang¹ Sewon Min² Xiang Deng¹ Jiaming Shen³ You Wu³
Luke Zettlemoyer² Huan Sun¹

¹The Ohio State University ²University of Washington ³Google Research
{wang.13930, deng.595, sun.397}@osu.edu
{sewon, lsz}@cs.washington.edu, {jmshen, wuyou}@google.com

Chain-of-thought (correct vs. incorrect)



GSM8K dataset, InstructGPT 175B (text-davinci-002)

Consistent results with text-davinci-003, PaLM, and Flan-PaLM (see the paper)

Invalid Logic Model Prompts

Rylan Schaeffer, KAIST

Language models can sometimes generate invalid logic reasoning. For example, consider the following Chain-of-Thought (CoT) prompt:

```
\textit{why} such problems are hard to solve.
```

Thought (CoT) prompts are designed to replace the reasoning part of a problem. Criticizing them can lead to wrong conclusions. To resolve this, we can use logically valid prompts:

```
\textit{invalid} reasoning is often wrong, but it can also discover that some reasoning is correct. This is because valid reasoning are not always the only way to solve a problem.
```

The screenshot shows a tweet from Rylan Schaeffer (@RylanSchaeffer) with a reply from samar-khanna. The tweet content is as follows:

J₇ simplifies to not true . In Boolean logic, not true is false. So the answer is False.

1 474 4 474 ...

Rylan Schaeffer
@RylanSchaeffer

In doing so, we discovered that some of the original CoT prompts were actually already logically invalid! For instance, here, numbers in the below BBH CoT are incorrectly replaced but still reach the correct answer.

5/N

samar-khanna commented on Mar 8 • edited

Thank you for your very interesting paper and for sharing the code base! I think there might be a few typos in the cot-prompts that might be worth investigating:

On line 11 of [multistep_arithmetic_two.txt](#), should this line:

Then, the final equation is A * B = -41 * -3 = (-61) * (-3) = 123. So the answer is 123.

instead be

Then, the final equation is A * B = -41 * -3 = (-41) * (-3) = 123. So the answer is 123.

1

Reasoning in Language

language models can sometimes generate invalid logic reasoning. However, [invalid] Chain-of-Thought prompts that editing CoT typically doesn't harm raw meaningful performance gains as and that the logically reasoning prompts. We states beyond logically

Research question

In-Context Learning



Supervised fine-tuning (alignment)?



LIMA: Less Is More for Alignment

“almost all knowledge in large language models is learned during pretraining”

Gargi Ghosh^μ Mike Lewis^μ Luke Zettlemoyer^μ Omer Levy^μ

^μ Meta AI

^π Carnegie Mellon University

^λ University of Southern California

^τ Tel Aviv University

The False Promise of Imitating Proprietary LLMs

“imitation models are adept at mimicking ChatGPT’s style but not its factuality”

Young Geng
UC Berkeley
young.geng@berkeley.edu

Hao Liu
UC Berkeley
hao.liu@berkeley.edu

Pabbeel Zelener
UC Berkeley
pabbeel@berkeley.edu

Sergey Levine
UC Berkeley
svlevine@berkeley.edu

Dawn Song
UC Berkeley
dawnsong@berkeley.edu

THE UNLOCKING SPELL ON BASE LLMs: RETHINKING ALIGNMENT VIA IN-CONTEXT LEARNING

“alignment tuning primarily learns to adopt the language style [...] and knowledge [...] predominantly comes from the base LLMs themselves”

♦ Allen Institute for Artificial Intelligence ♦ University of Washington
✉ yuchenl@allenai.org

(posted two days ago)

Talk outline

Sewon: Ground truth demonstrations are **not required** for ICL.

Junyeob: ICL **do** need ground truth demonstrations, and **can learn** newly-defined tasks.

Sewon: These results **do not rebut** our paper:
they either do not affect our findings, or lack evidence

Sewon: Our general findings **hold more broadly**
(generation/reasoning tasks, alignment)

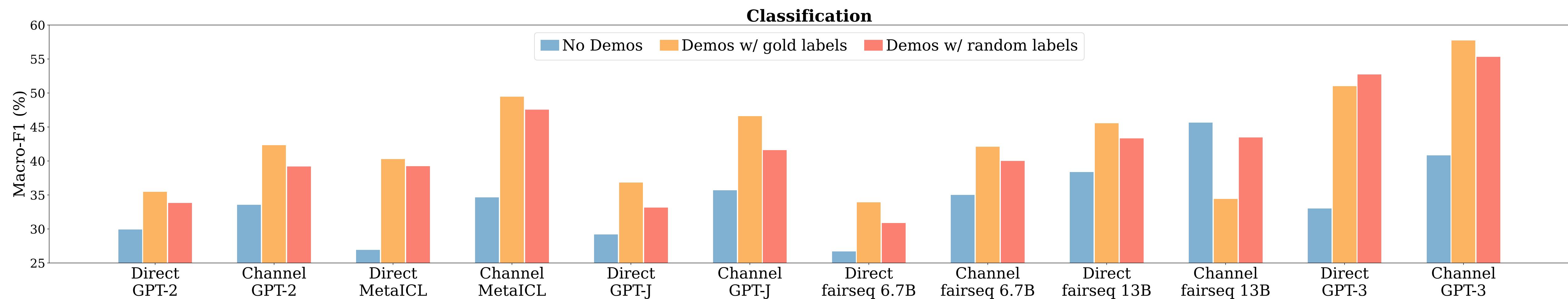
Junyeob: Yes, but in the end **small deterioration matters**

Junyeob & Sewon: Summary

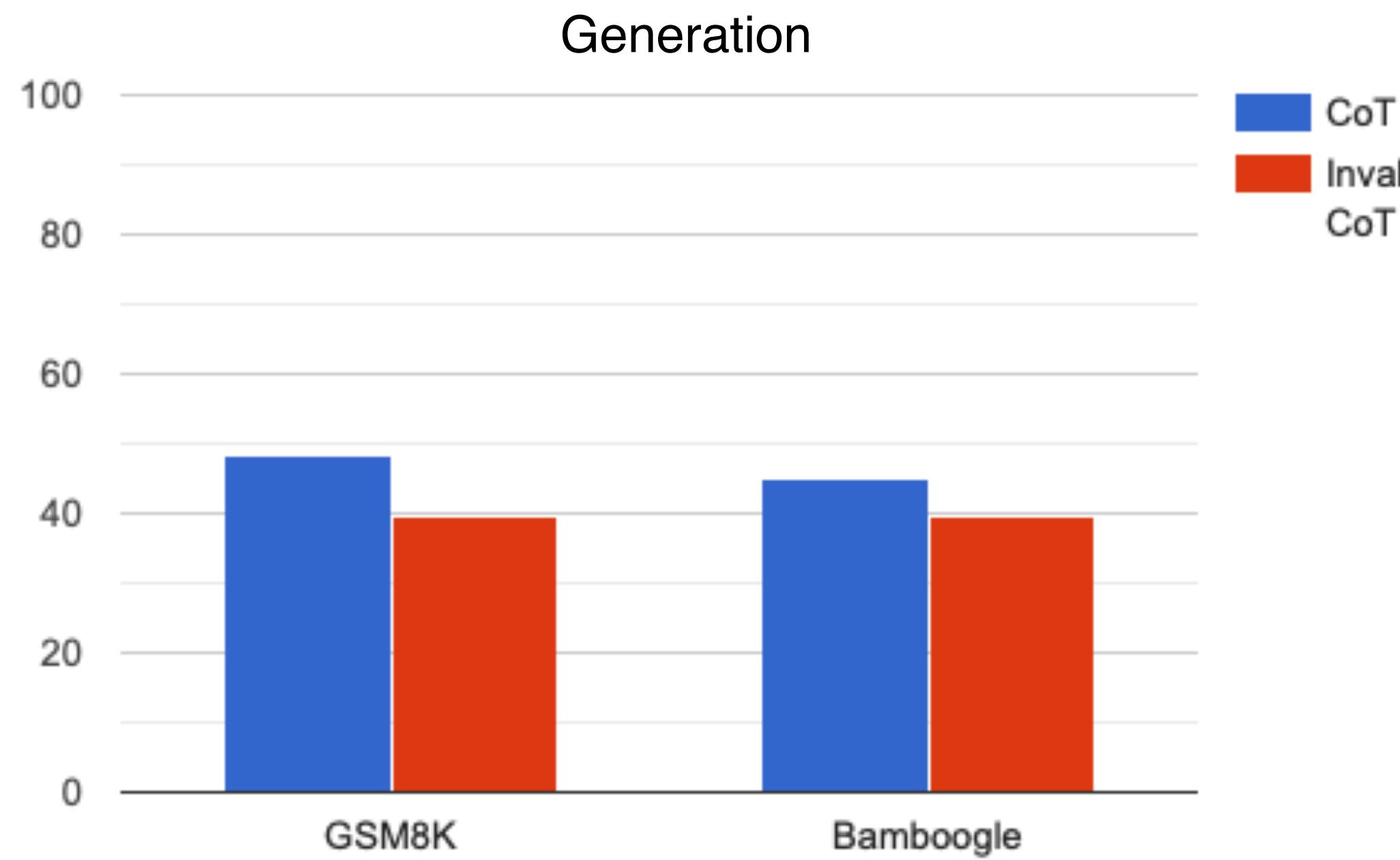
What we agree on

- **Generation/reasoning tasks:**
 - Incorrect CoT can achieve performance ***close*** to original CoT
 - **Fine-tuning/alignment:**
 - Pre-training taught most things
-  It does not directly contradict the fact that LLMs are capable of new task on the fly.

Consistent detoration in performance



Consistent deterioration in performance



GSM8K evaluated performance using accuracy, Bamboogle assessed it using F1 (text-davinci-002)

Wang et al. 2022. "Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters"

In the end, application is what matters

- In industrial applications, even small deterioration in task-solving performance may be intolerable (e.g., medical domain)
 - In the extreme case of zero performance deterioration tolerance, only cases where **the GTL sensitivity** is zero (i.e. no change in performance *at all* no matter the input-label mapping accuracy) can be considered "GTL does not matter"
- Key Questions in the Industrial Perspective
 - *How critical is the performance deterioration due to in-context label noises?*
 - *What is the cost of ensuring label quality of prompt demonstrations? What is the budget?*

The screenshot shows a web browser window with the URL clova-x.naver.com. The page displays a conversation interface with a sidebar for new messages and a main area for the conversation history.

Conversation History:

- + 새 대화
- ▽ 대화 리스트 >
 - What is the newton meth...
 - 근의 공식을 공부하는 중학생 ...
 - 한국형 LLM에 대한 특강을 진...
 - 아래 문장 번역이 어색한지 봐...
 - 질문: 학적 결합력이 강하며 열...
 - Hafnia 기반 ferroelectric 물...
- ▽ 디스커버리 >
 - ⚡ 자율주행 자동차에 관한 영...

Text Input:

What is the newton method for finding squared r

Response:

The Newton method is an iterative algorithm for finding better approximations to the roots (or zeroes) of a function. It is named after the English mathematician Sir Isaac Newton, who described it in his book "Mathematical Principles of Natural Philosophy" in 1687.

The Newton method for finding the square root of a number x is written as:

```
x_{n+1} = x_n - f(x_n) / f'(x_n)
```

where:

- x_n is the current approximation to the square root of x .
- $f(x)$ is the function we want to find the root of, $f(x) = x^2 - x$.
- $f'(x)$ is the derivative of $f(x)$, which is $f'(x) = 2x$.

To apply the Newton method, we start with an initial guess x_0 for the square root of x , and then iteratively refine our guess until we reach a satisfactory approximation. The iteration continues until the difference between successive approximations is sufficiently small.

자유롭게 대화해 보세요.

CLOVA X는 부정확하거나 불쾌감을 주는 정보를 제공할 수 있으며, 이는 NAVER

Talk outline

Sewon: Ground truth demonstrations are **not required** for ICL.

Junyeob: ICL **do** need ground truth demonstrations, and **can learn** newly-defined tasks.

Sewon: These results **do not rebut** our paper:
they either do not affect our findings, or lack evidence

Sewon: Our general findings **hold more broadly**
(generation/reasoning tasks, alignment)

Junyeob: Yes, but in the end **small deterioration matters**

Junyeob & Sewon: Summary

Junyeob's summary

Min et al. claimed ground truth demonstrations are **not required** for ICL, but **we found** that this is partially true

Although the impact may vary across setups, **the correctness of labels is one of the core components** of successful ICL

Our interpretation of Wei et al. is that the fact that it is easier to learn new task than overriding priors with ICL reinforces our findings

Although ICL with invalid demonstrations seems to work in general, a **small deterioration is not negligible** in real-world applications

Sewon's summary

Correct demonstrations are **not required** for ICL –
ICL activates training priors rather than **learning new tasks**

Yoo et al. claimed ICL sometimes uses correct demonstrations, but

1. Correct demonstrations are **still not required** with careful prompting
2. The best learner still **does not need** correct demonstrations

Our interpretation of Wei et al. is that the claim “ICL overriding semantic priors is an emergent feature” lacks evidence

Overall findings generalize to other settings such as
generation tasks & alignment

Science of Language Models



Review: Circulation revenue has increased by 5% in Finland.

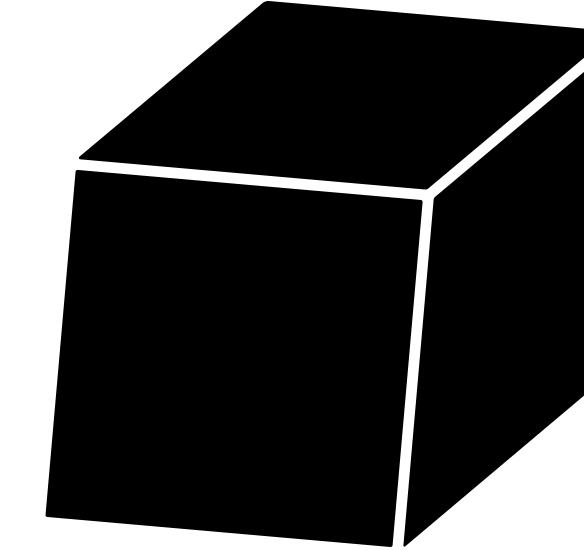
Sentiment: Positive

Review: Paying off the national debt will be extremely painful.

Sentiment: Negative

Review: The company anticipated its operating profit to improve.

Sentiment:



Positive

Science of Language Models



GPT-3



Review: Circulation revenue has increased by 5% in Finland.

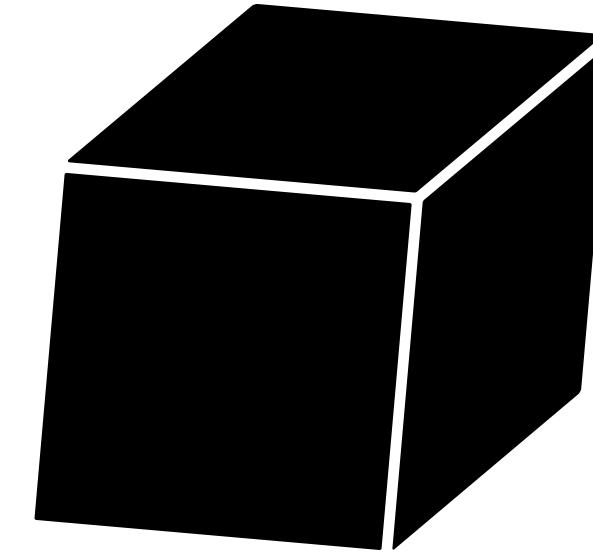
Sentiment: Negative

Review: Paying off the national debt will be extremely painful.

Sentiment: Neutral

Review: The company anticipated its operating profit to improve.

Sentiment:



Positive

?!
?

Do Prompt-Based Models Really Understand the Meaning of Their Prompts?

Albert Webson^{1,2} and Ellie Pavlick¹

{albert_webson, ellie_pavlick}@brown.edu

¹Department of Computer Science, Brown University

²Department of Philosophy, Brown University

Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm

Laria Reynolds
moire@knc.ai

Kyle McDonell
kyle@knc.ai

More analysis on what contributes to ICL

Rethinking the Role of Scale for In-Context Learning: An Interpretability-based Case Study at 66 Billion Scale

Hritik Bansal^{1*} Karthik Gopalakrishnan^{2†} Saket Dingliwal² Sravan Bodapati²
Katrín Kirchhoff² Dan Roth²

¹University of California, Los Angeles

hbansal@cs.ucla.edu

²AWS AI Labs

{karthgop, skdin, sravanb, katrinki, drot}@amazon.com

Other unexpected behaviors of LMs

Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting

Miles Turpin,^{1,2} Julian Michael,¹ Ethan Perez,^{1,3} Samuel R. Bowman^{1,3}

¹NYU Alignment Research Group, ²Cohere, ³Anthropic

miles.turpin@nyu.edu

Can In-context Learners Learn a Reasoning Concept from Demonstrations?

Michal Štefánik and Marek Kadlčík

Faculty of Informatics, Masaryk University, Czech Republic

{stefanik.m, kadlcik}@mail.muni.cz

Measuring Inductive Biases of In-Context Learning with Underspecified Demonstrations

Chenglei Si^{1*} Dan Friedman^{2*}

Nitish Joshi³ Shi Feng⁴ Danqi Chen² He He³

¹University of Maryland

²Princeton University

³New York University

⁴University of Chicago

clsi@umd.edu, dfriedman@cs.princeton.edu

... and more!

Thank you for listening

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



Ground-Truth Labels Matter: A Deeper Look into Input-Label Demonstrations



Contact:



sewon@cs.washington.edu / @sewon_min



juny116@europa.snu.ac.kr / kangmin.yoo@navercorp.com

Please leave feedback at:

<https://tinyurl.com/sewon-min-talk>