ScatterShot: Interactive In-context Example Curation for Text Transformation

Sherry Tongshuang Wu, *Carnegie Mellon University*
Hua Shen, *PennState University*
Daniel S. Weld, *University of Washington*
Jeffrey Heer, *University of Washington*
Marco Tulio Ribeiro, *Microsoft*

What is prompt-based learning with LLMs?

Encourages a **pre-trained** Large Language Model (LLM) to make **particular predictions** by providing a **"prompt"** specifying the task to be done.

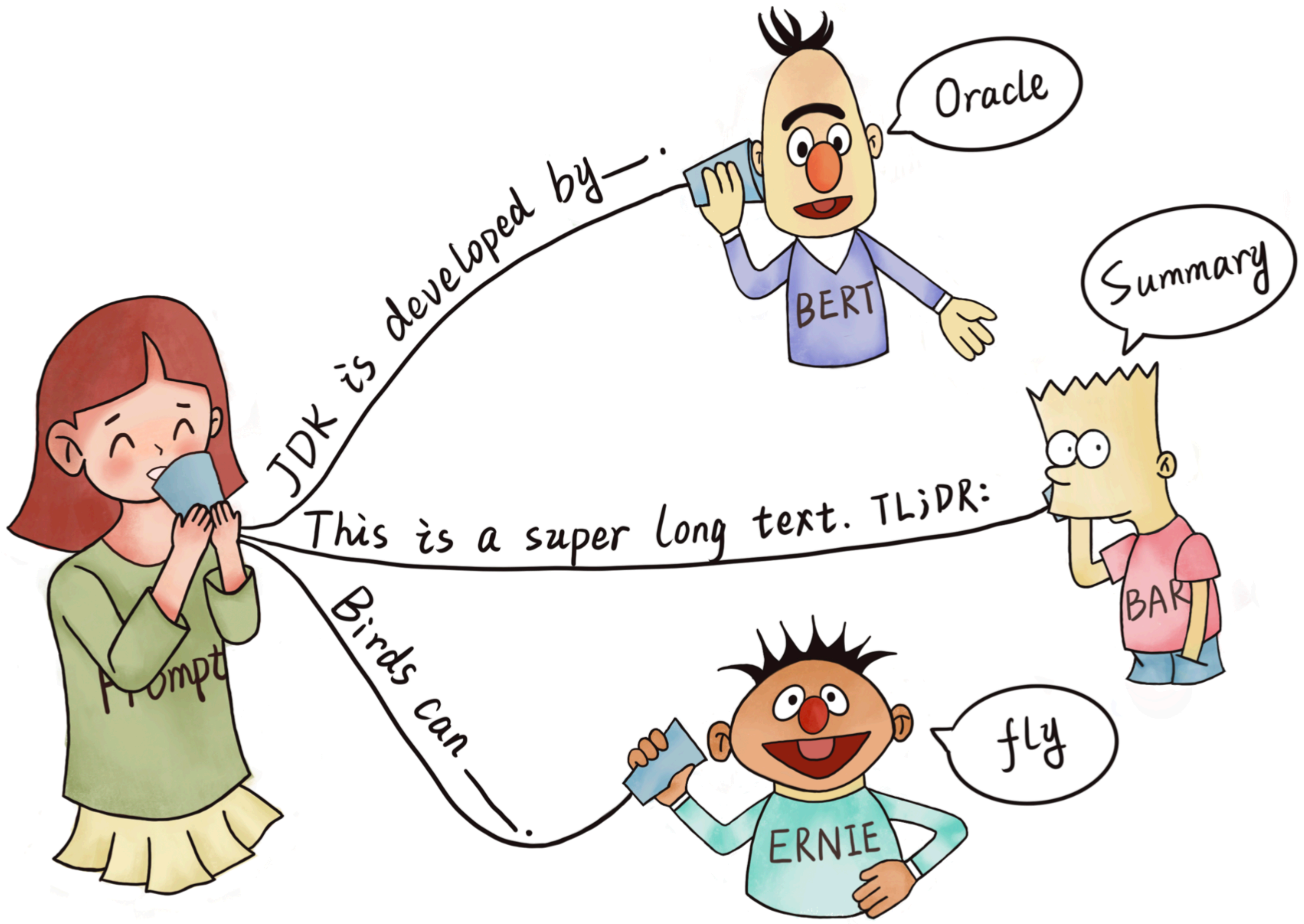


Figure from: Liu Pengfei, et al. "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing." *arXiv 2021*

What is prompt-based learning with LLMs?

Encourages a **pre-trained** Large Language Model (LLM) to make **particular predictions** by providing a **"prompt"** specifying the task to be done.

Prompt Design

In-context Learning

Prompt Search

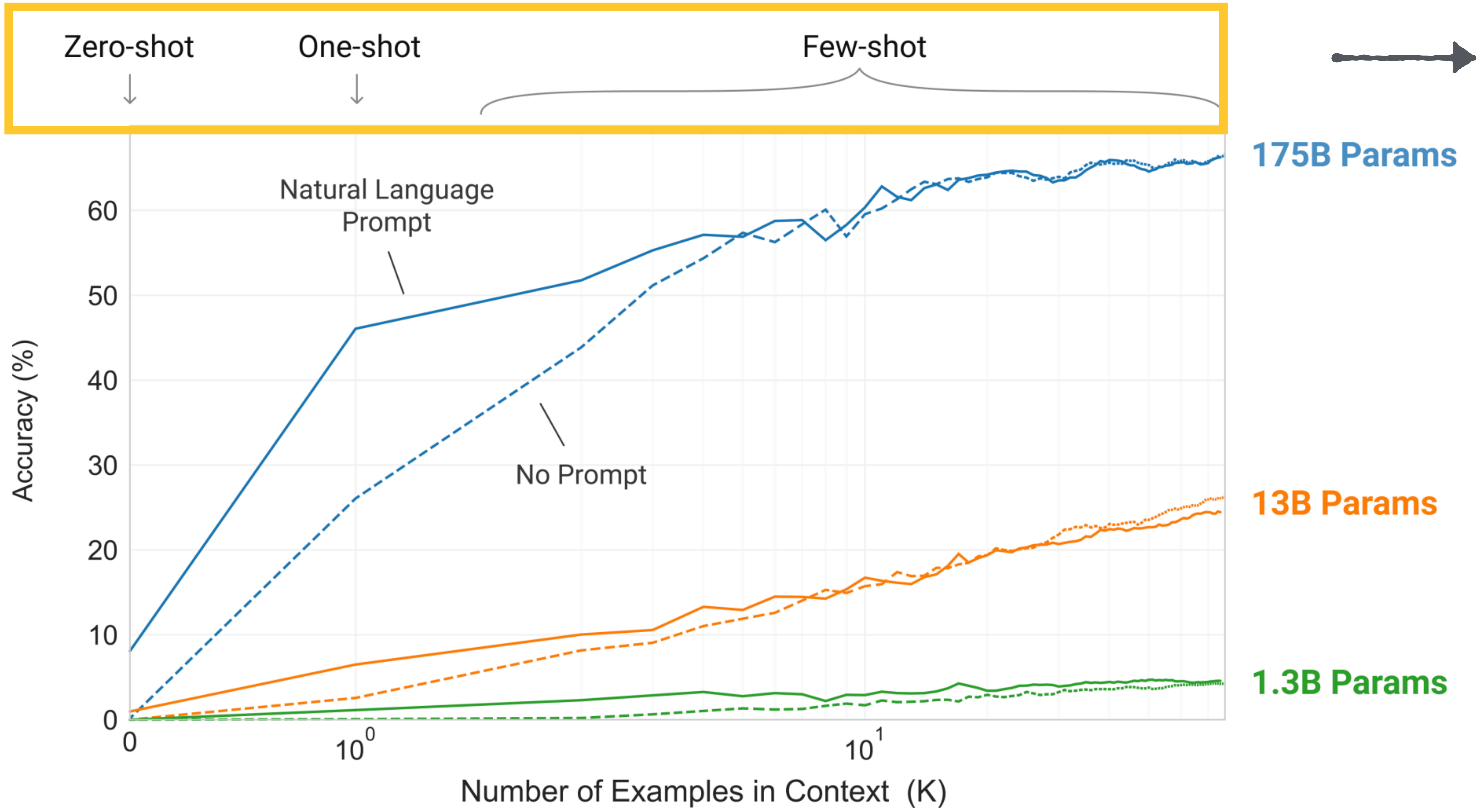
P* tuning

LM + P* tuning

What is in-context learning?

The input to the model describes a new task with some possible examples, **in natural language**.

Effective on **very large models** (173B GPT-3)



How many examples ?

Brown, Tom, et al. "Language Models are Few-Shot Learners." *NeurIPS 2020.*

In-context learning: Prompt types

Zero-shot
Natural language descriptions only

Prompt	
`Find the nationality of people:`	*Task description*
`Marie Curie =>`	*Task*

One-shot
Description + one example

Prompt	
`Find the nationality of people:`	*Task description*
`Albert Einstein => German`	*Example*
`Marie Curie =>`	*Task*

Few-shot
Description + a few example (3-100)
[5-10 is most common]

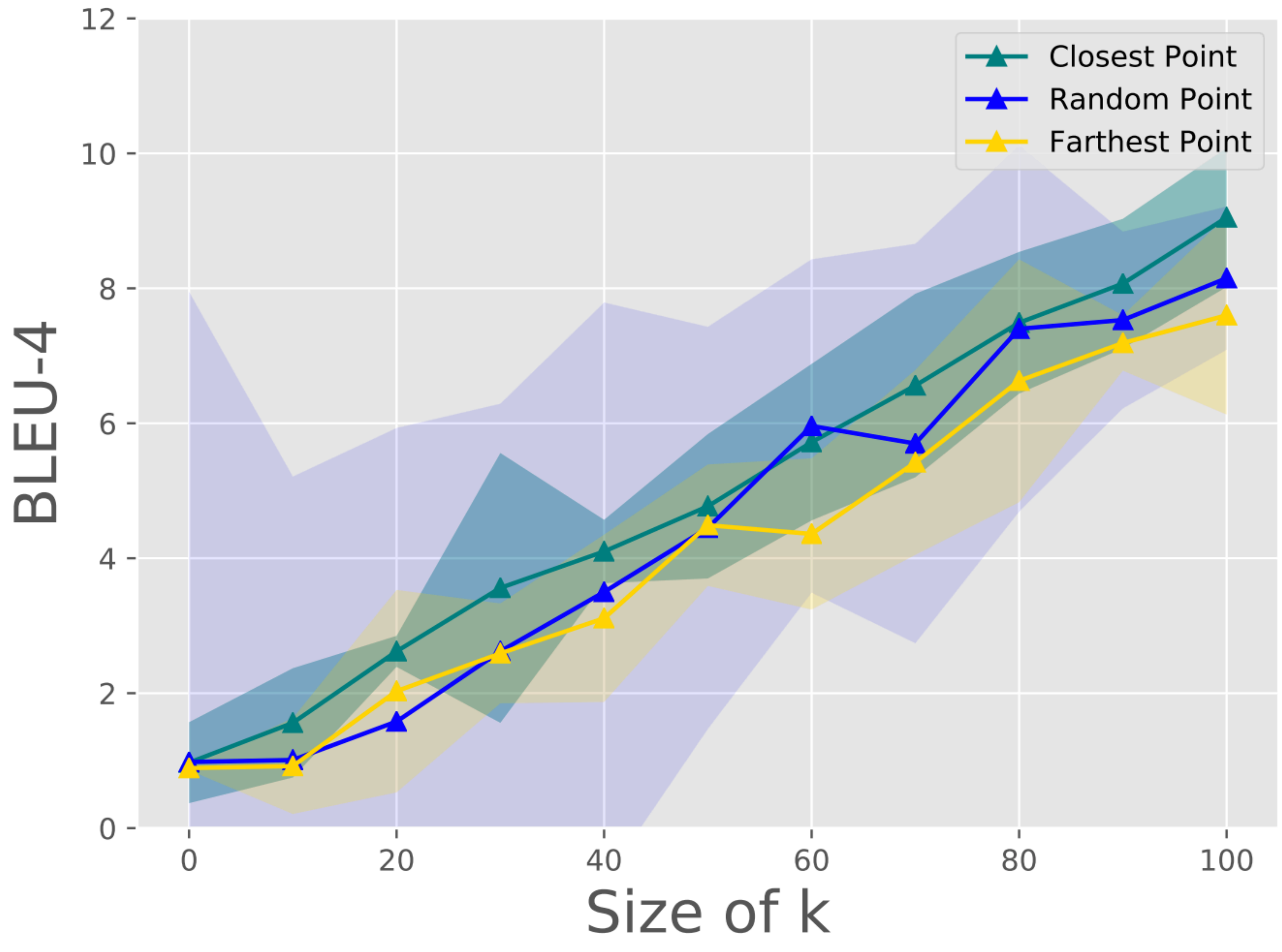
Prompt	
`Find the nationality of people:`	*Task description*
`Albert Einstein => German`	*Examples*
`Alan Turing => English`	
`Mahatma Gandhi => Indian`	
`Marie Curie =>`	*Task*

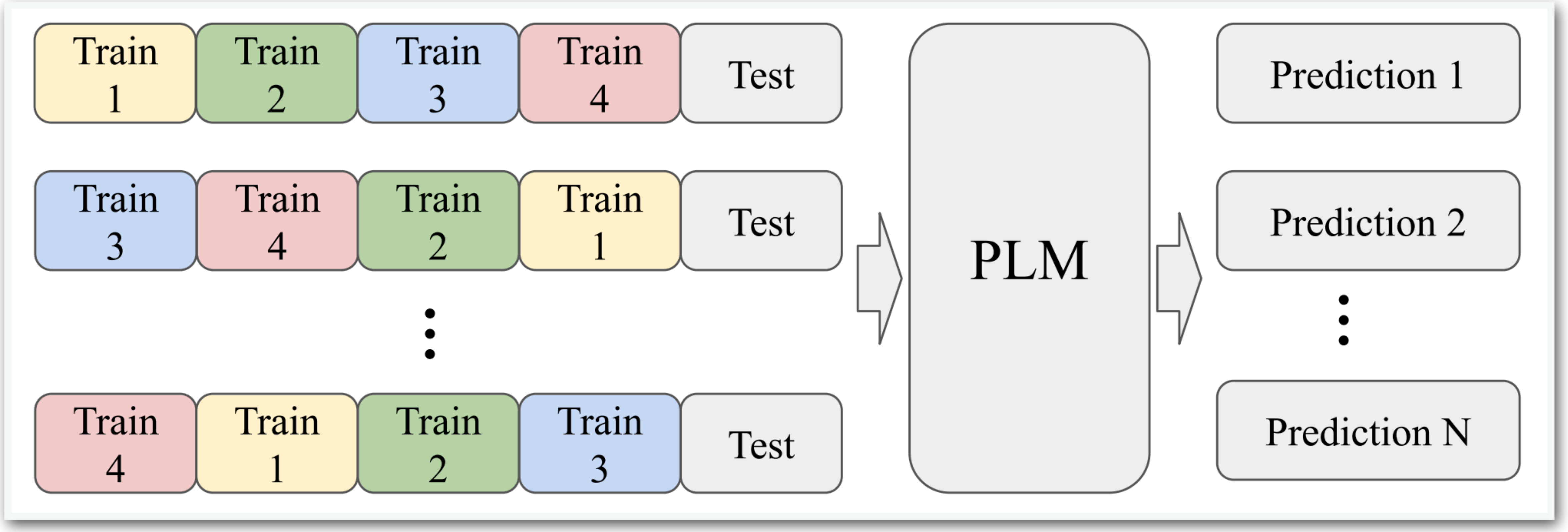
How to make ?

Brown, Tom, et al. "Language Models are Few-Shot Learners." *NeurIPS 2020.*

Challenge: which sets of examples?

Let's assume users are given a training data set to choose prompt examples.





Different few-shot **example sets** lead to very different results.

Different ordering of the same set also lead to different results!!

Chang, Ernie, et al. "On training instance selection for few-shot neural text generation." arXiv preprint arXiv:2107.03176 (2021).Lu, Yao, et al. "Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity." arXiv preprint arXiv:2104.08786 (2021).

Challenge: when "enough" examples?

```
Find the nationality of people:
Albert Einstein => German
...
Alan Turing => English
...
Mahatma Gandhi => Indian
Marie Curie =>
```

Task description

Examples

***k* = ?**

Task

When do the prompt examples enough to make good prediction?

The model **performs better** when the test input is **similar** to some **training input**.
But it's **hard to get coverage** in 30 examples.

Liu, Jiachang, et al. "What Makes Good In-Context Examples for GPT-3?." arXiv preprint arXiv:2101.06804 (2021).

Research objectives

*We present, **ScatterShot**, to help users interactively and iteratively find high-quality demonstrative examples to build effective in-context functions.*

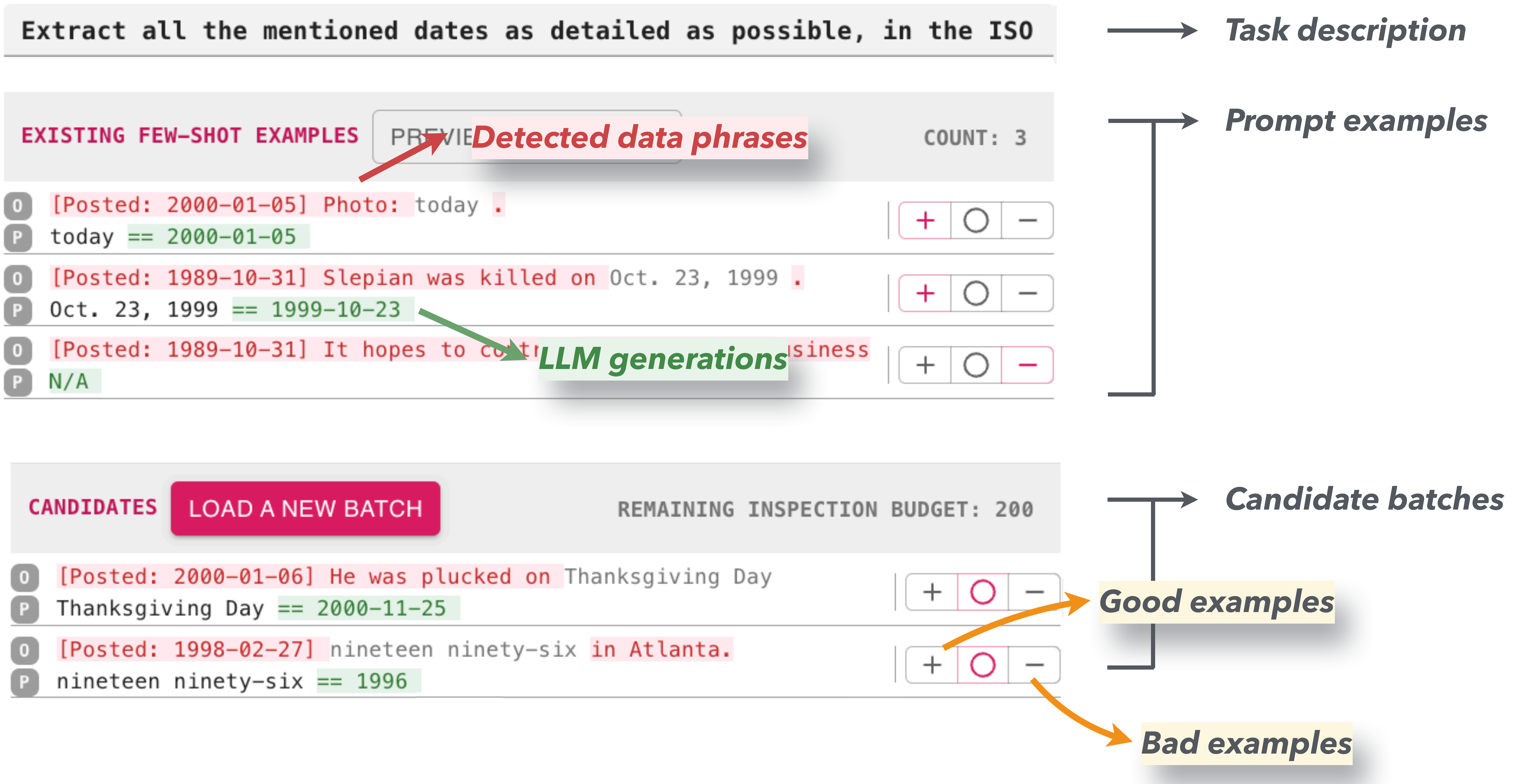
Scattershot principles

1. Handle common patterns → Help the user discover previously unexplored patterns.

2. Not neglect unusual ones → Help the user prioritize the most informative examples.

3. Cost effective → Minimize annotation cost.

User interface



How can we use the least examples to cover most prompt patterns?

Scattershot algorithm

Input-output pairs, iteration 1 to $i-1$

[Posted: 1998-02-27] **nineteen ninety-six** in Atlanta
nineteen ninety-six == 1996

[Posted: 2000-01-05] Photo: on **today** .
today == 2000-01-05

[Posted: 2000-01-06] He was plucked on **Thanksgiving** Day.
Thanksgiving == 1999-11-25

A Existing prompt examples

Scattershot algorithm

Input-output pairs, iteration 1 to $i-1$

[Posted: 1998-02-27] **nineteen ninety-six** in Atlanta
nineteen ninety-six == 1996

A

[Posted: 2000-01-05] Photo: on **today** .
today == 2000-01-05

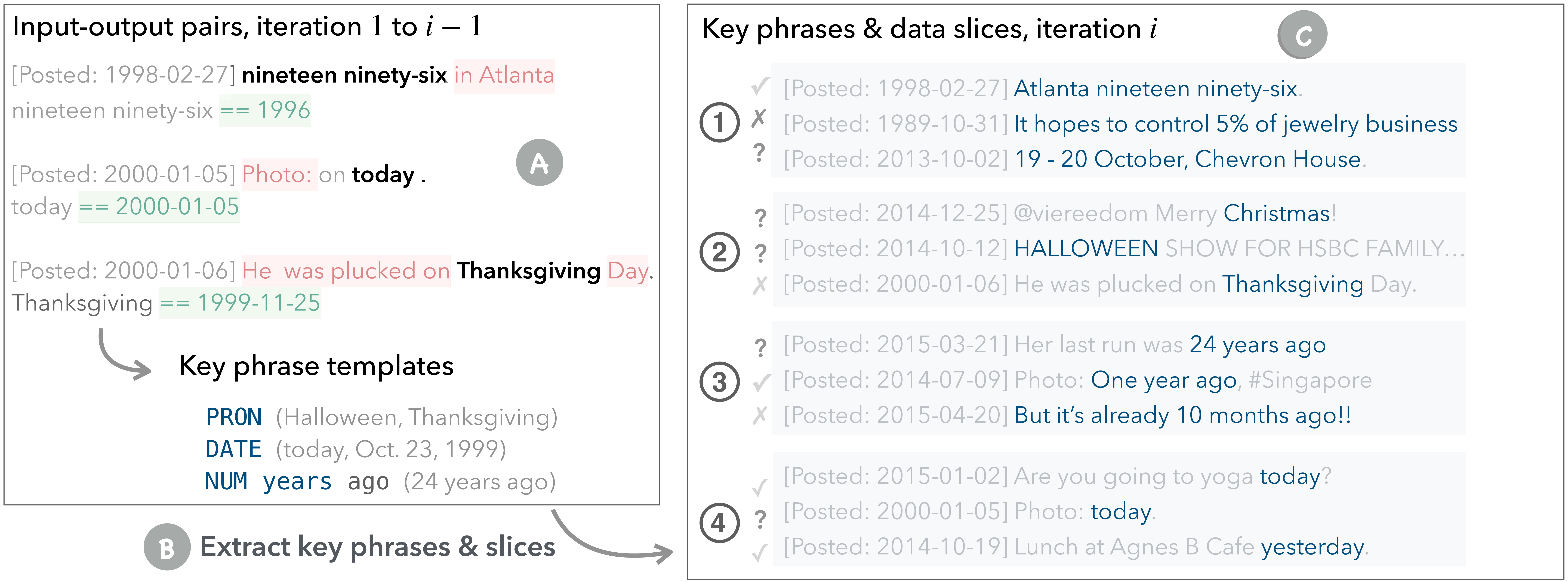
[Posted: 2000-01-06] He was plucked on **Thanksgiving** Day.
Thanksgiving == 1999-11-25

Key phrase templates

`PRON` (Halloween, Thanksgiving)
`DATE` (today, Oct. 23, 1999)
`NUM years ago` (24 years ago)

B Extract key phrases & slices

Slice-baed Sampling



Prioritize sampled examples

Input-output pairs, iteration 1 to $i-1$

[Posted: 1998-02-27] **nineteen ninety-six** in Atlanta
nineteen ninety-six == 1996

[Posted: 2000-01-05] Photo: on **today** .
today == 2000-01-05

[Posted: 2000-01-06] He was plucked on **Thanksgiving** Day.
Thanksgiving == 1999-11-25

A

Key phrase templates

PRON (Halloween, Thanksgiving)
DATE (today, Oct. 23, 1999)
NUM years ago (24 years ago)

B Extract key phrases & slices

Key phrases & data slices, iteration i

C

①	✓	[Posted: 1998-02-27] Atlanta nineteen ninety-six.	n=449 m=10 k=4 μ=4.82
	✗	[Posted: 1989-10-31] It hopes to control 5% of jewelry business	
	?	[Posted: 2013-10-02] 19 - 20 October, Chevron House.	
②	?	[Posted: 2014-12-25] @viereedom Merry Christmas!	n=19 m=2 k=0 μ=4.34
	?	[Posted: 2014-10-12] HALLOWEEN SHOW FOR HSBC FAMILY…	
	✗	[Posted: 2000-01-06] He was plucked on Thanksgiving Day.	
③	?	[Posted: 2015-03-21] Her last run was 24 years ago	n=31 m=5 k=1 μ=3.61
	✓	[Posted: 2014-07-09] Photo: One year ago, #Singapore	
	✗	[Posted: 2015-04-20] But it's already 10 months ago!!	
④	✓	[Posted: 2015-01-02] Are you going to yoga today?	n=113 m=3 k=3 μ=1.14
	?	[Posted: 2000-01-05] Photo: today.	
	✓	[Posted: 2014-10-19] Lunch at Agnes B Cafe yesterday.	

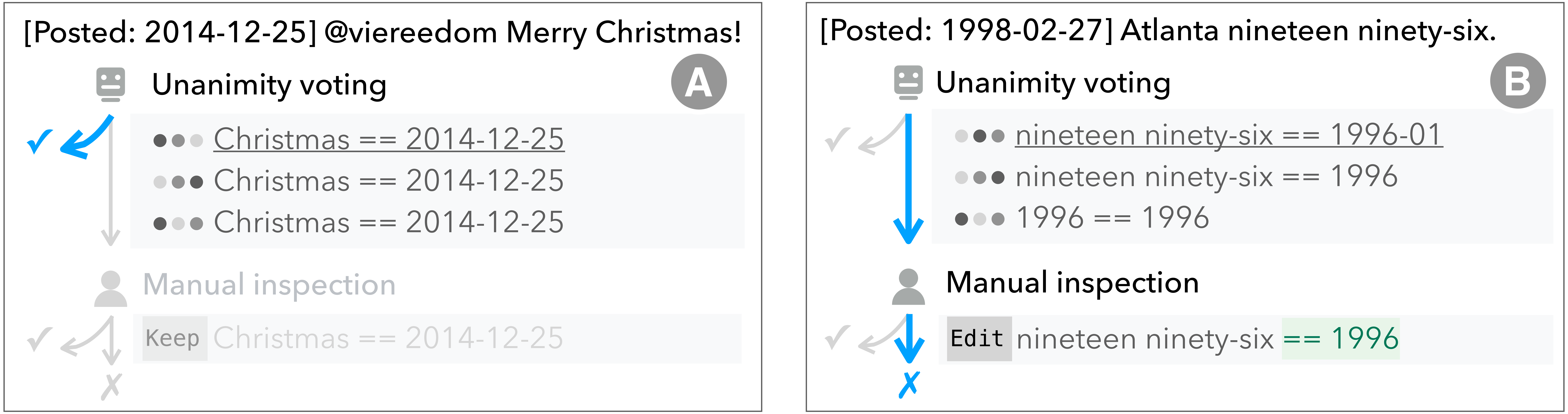
Prioritize similar data that has **low performance**, are **large**, and slices that have **not been** sampled many times.

$$\mu_{i,c} = \underbrace{\left(1 - \frac{k}{m}\right)}_{\text{Error rate}} \cdot \underbrace{\ln n}_{\text{Size}} + \underbrace{\sqrt{\frac{\ln t}{m}}}_{\text{Sample Rarity}}$$

Slice c has n examples, m are labeled in previous iterations.
Out of m, the current function is correct on k.

How to handle no ground truth labels?

We estimate function quality by re-ordering stability.



Scattershot evaluation

Task & Datasets

1. Simulation Experiment
   - Simulate the labeling process
2. Within-subject User Study
   - 10 person evaluation
   - QA-pair rewriting task

Temporal Expression Extraction

```
O [Posted: 2000-01-05] Photo: today .
P today == 2000-01-05

O [Posted: 1989-10-31] Slepian was killed on Oct. 23, 1999 .
P Oct. 23, 1999 == 1999-10-23

O [Posted: 1989-10-31] It hopes to control 5% of jewelry business
P N/A
```

Question-Answer Pair Rewriting

```
O Q: Where are the buildings? A: in distance
P Q: Are the buildings in distance? A: yes

O Q: Why is it dark? A: twilight
P Q: Is it dark because of the twilight? A: yes

O Q: Is the water warm or cold? A: cold
P Q: Is the water cold? A: yes
```

1 Simulation performance

Temporal

Conditions	Extraction			Normalization		
	F1	Precision	Recall	F1	Precision	Recall
Random	73.2 ± 4.0	74.0 ± 3.8	72.9 ± 4.1	66.8 ± 3.2	67.3 ± 3.3	67.0 ± 3.1
ScatterShot	**75.0 ± 2.9**	**75.6± 2.8**	**74.7 ± 2.9**	**70.9 ± 3.4****	**71.3 ± 3.5***	**71.2 ± 3.2****

QA-Pair

Conditions	ROUGE-L	BLEU-4
Rule-based	78.4	66.7
Random	74.3 ± 3.9	65.4 ± 3.5
ScatterShot	**80.0 ± 3.5***	**69.1 ± 3.1***

The significant improvements, measured by the student's **t-test** are marked with ***: p<0.05**, and ****: p<0.01**.

Quantitative Results:

✓ Compared with the Random condition, **ScatterShot outperformed** the baseline on **all metrics**.

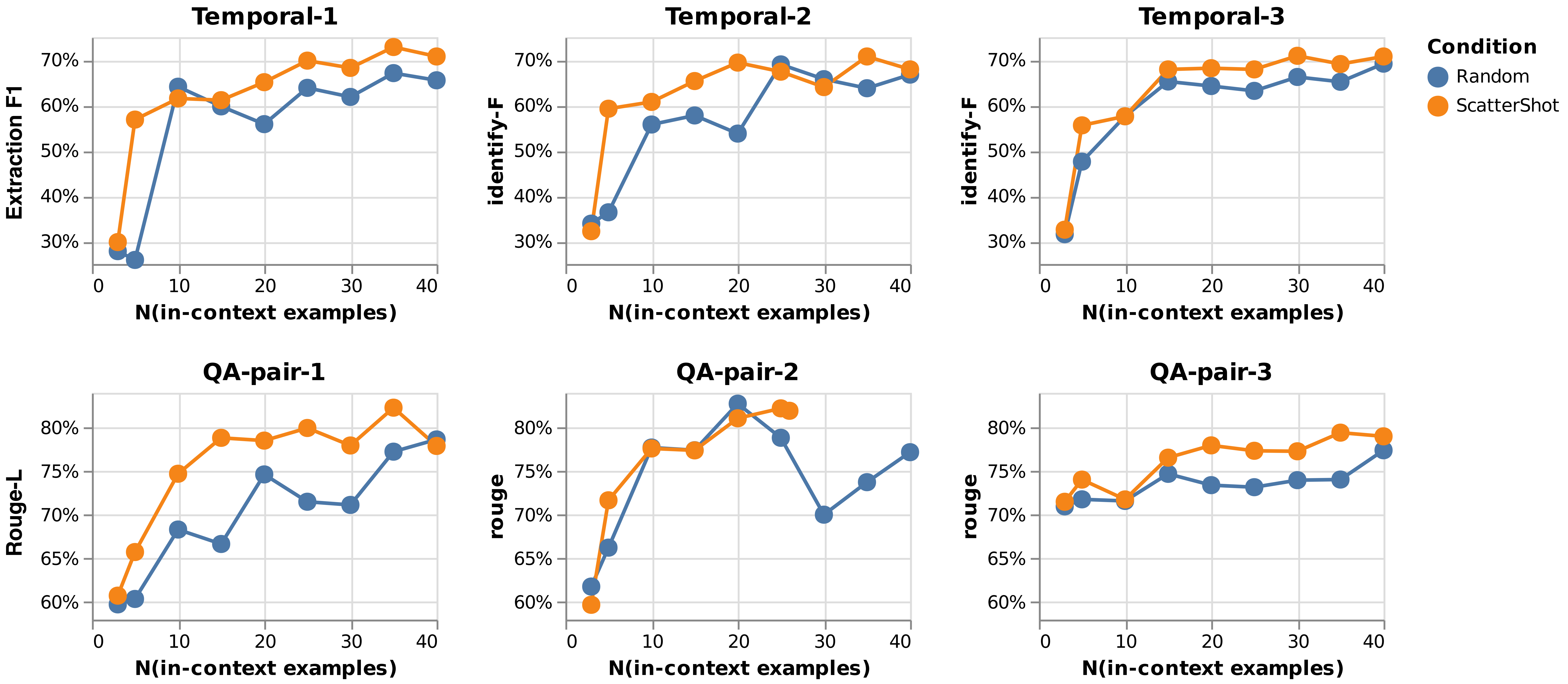
Example outputs

Coverage: Transforms more forms of inputs.		
Input	Q: Are there more girls or boys? A: equal	Q: How many hairs does the sheep in front have? A: infinite
Rule-based	✗ *(No generation)*	✗ *(No generation)*
Random	✓ Q: Are the girls and boys equal in number? A: yes	✗ N/A
ScatterShot	✗ Q: Are the girls and boys equal? A: yes	✓ Q: Does the sheep in front have infinite hairs? A: yes
Fluency: Generates outputs that sound natural.		
Input	Q: What make is the phone? A: vtech	Q: What does the woman have on her face? A: headband
Rule-based	✗ Q: Make is the phone vtech? A: yes	✗ Q: Does the woman have on her face headband? A: yes
Random	✓ Q: Is the phone a vtech? A: yes	✓ Q: Does the woman have a headband on her face? A: yes
ScatterShot	✓ Q: Is the phone a vtech? A: yes	✓ Q: Does the woman have a headband on her face? A: yes
Correctness: Produces desired outputs (the new question-pair are logically equivalent to the original pair).		
Input	Q: What monument are they next to? A: unknown	Q: What type of motorcycle is in the picture? A: mountain
Rule-based	✗ Q: Are they next to unknown? A: yes	✗ Q: Is the mountain in the picture ? A: yes
Random	✗ Q: Is the monument unknown? A: yes	✗ Q: Is the mountain type of motorcycle in the picture ? A: yes
ScatterShot	✓ Q: Are they next to an unknown monument? A: yes	✓ Q: Is the motorcycle in the picture a mountain bike? A: yes

Compared with the **Random** condition, and a **Rule-based** system:

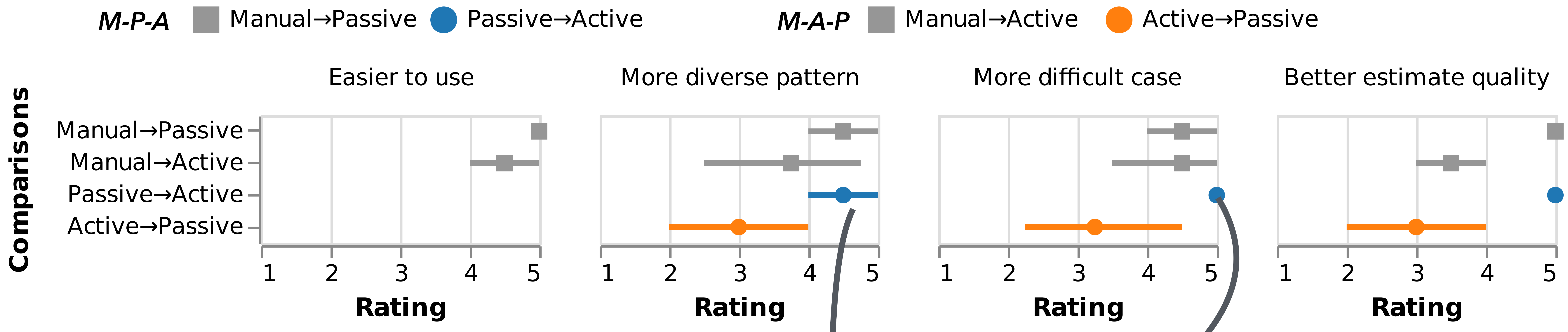
***ScatterShot** functions tend to have better **coverage**, **fluency**, and **correctness**.*

Performance trajectory w.r.t. examples



We evaluate the **held-out test set** every time we add five more examples to the in-context bucket until the stop condition is satisfied.
ScatterShot tends to frequently **outperform** Random, and tends to **have better performance**

② User Study Performance



Active learning is effective for humans (More holistic view)!

I went through several rounds of pretty similar examples in Step 2 (Random), thinking the function is behaving quite decently, and didn't realize the function needed more diverse and edge cases until I reached Step 3.

Performance of user created function

Condition	Step 1	Step 2	Step 3
M-R-S	/ (59.3)	+17.4 (74.7)	**+3.2** (77.8)
M-S-R	/ (61.8)	**+18.1** (75.4)	-0.4 (74.9)

(a) ROUGE-L

R -> S
S -> R

Condition	Step 1	→ Step 2	→ Step 3
M-R-S	/ (63.9)	**+10.1** (74.0)	**+3.1** (76.9)
M-S-R	/ (65.3)	+8.9 (74.2)	-0.6 (73.6)

(b) BLEU-4

+/- : represents the **average performance change** compared to the prior step, (number) are the absolute performance.
M-R-S: users build in-context functions using methods of "Manual - Random - ScatterShot" in sequence.
M-S-R: users use "Manual - ScatterShot - Random" methods in sequence.

> ***M-R-S*** *users were able to keep* ***adding useful examples****,*
> *whereas* ***M-S-R*** *users* ***decreased*** *the function performance by 0.6 in Step 3 (ScatterShot -> Random), indicating that these efforts were wasted.*

What's more?

✓ Slice-based sampling can increase **data space coverage**

✗ Random sampling performs less

✓ Interacting with the latest function for users is essential for in-context learning.

✓ Human-AI collaborative labeling for building better functions results in better quality and better task definition.

Takeaways

ScatterShot *helps users find informative input examples in the unlabeled data,* ***improves*** *the annotator's awareness and handling of diverse patterns, and ultimately, the in-context function performance.*

The full user study instructions, and the detailed exit survey, are at:

Github: https: //github.com/tongshuangwu/scattershot

Thank You!





Sherry Tongshuang Wu
sherryw@cs.cmu.edu
@tongshuangwu

Hua Shen
huashen218@psu.edu
@huashen218

Daniel S. Weld
weld@cs.uw.edu

Jeffrey Heer
jheer@cs.uw.edu

Marco Tulio Ribeiro
marcotcr@microsoft.com