JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

ICL CIPHERS: Quantifying "Learning” in In-Context Learning via Substitution Ciphers

Zhouxiang Fang, Aayush Mishra, Muhan Gao, Anqi Liu, Daniel Khashabi

Presented by Zhouxiang Fang

What is “In-context Learning”?

What is “In-context Learning”?

Definition: In-context Learning (ICL) is an emerging feature of Large Language Models that allows them to identify patterns in demonstrations given as prompts and apply these patterns to similar tasks. [1]

1. Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

What is “In-context Learning”?

Definition: In-context Learning (ICL) is an emerging feature of Large Language Models that allows them to identify patterns in demonstrations given as prompts and apply these patterns to similar tasks. [1]

Changes of parameters are not needed when adapting to different tasks.

1. Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

What is "In-context Learning"?

Definition: In-context Learning (ICL) is an emerging feature of Large Language Models that allows them to identify patterns in demonstrations given as prompts and apply these patterns to similar tasks. [1]

Changes of parameters are not needed when adapting to different tasks.

Demonstrations

```
Input: I love my school! There is ...
Output: positive

Input: The sky doesn't looks so nice
Output: negative

    ⋮
```

1. Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

What is “In-context Learning”?

Definition: In-context Learning (ICL) is an emerging feature of Large Language Models that allows them to identify patterns in demonstrations given as prompts and apply these patterns to similar tasks. [1]

Changes of parameters are not needed when adapting to different tasks.

Demonstrations

```
Input: I love my school! There is ...
Output: positive

Input: The sky doesn’t looks so nice
Output: negative

    ⋮
```

Test case Input

```
Input: I don’t love looking at the sky
```

1. Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

What is “In-context Learning”?

Definition: In-context Learning (ICL) is an emerging feature of Large Language Models that allows them to identify patterns in demonstrations given as prompts and apply these patterns to similar tasks. [1]

Changes of parameters are not needed when adapting to different tasks.

Demonstrations

```
Input: I love my school! There is …
Output: positive

Input: The sky doesn’t looks so nice
Output: negative

    ⋮
```

Test case Input

```
Input: I don’t love looking at the sky
```

LLM

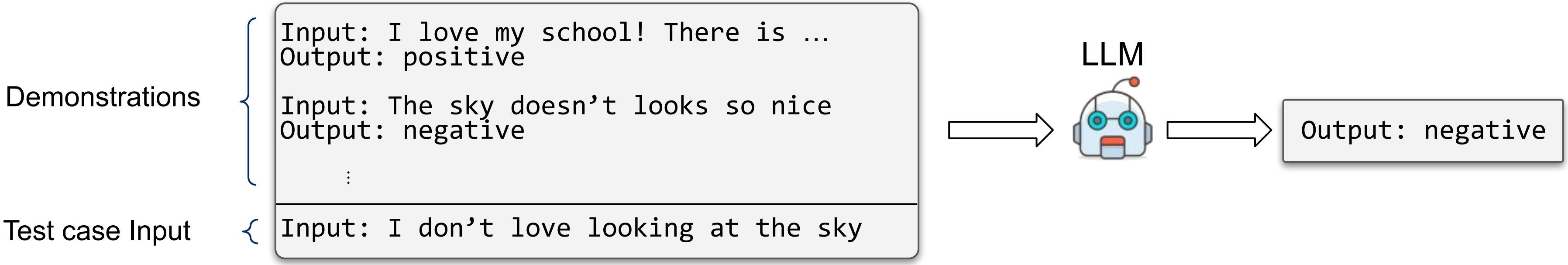
1. Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

What is “In-context Learning”?

Definition: In-context Learning (ICL) is an emerging feature of Large Language Models that allows them to identify patterns in demonstrations given as prompts and apply these patterns to similar tasks. [1]

Changes of parameters are not needed when adapting to different tasks.



1. Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

Motivation: Task Learning vs. Retrieval

Motivation: Task Learning vs. Retrieval

Literatures suggest that ICL operates in dual modes[1,2]

1.Wang, Xiaolei, et al. "Investigating the Pre-Training Dynamics of In-Context Learning: Task Recognition vs. Task Learning." *arXiv preprint arXiv:2406.14022* (2024).
2.Lin, Ziqian, and Kangwook Lee. "Dual operating modes of in-context learning." *Forty-first International Conference on Machine Learning*. 2024.

Motivation: Task Learning vs. Retrieval

Literatures suggest that ICL operates in dual modes[1,2]

- **Task Retrieve (TR):** Recall learned patterns from pre-training

- **Task Learning (TL):** Learning from demonstrations during inference time

1.Wang, Xiaolei, et al. "Investigating the Pre-Training Dynamics of In-Context Learning: Task Recognition vs. Task Learning." *arXiv preprint arXiv:2406.14022* (2024).
2.Lin, Ziqian, and Kangwook Lee. "Dual operating modes of in-context learning." *Forty-first International Conference on Machine Learning*. 2024.

Motivation: Task Learning vs. Retrieval

Literatures suggest that ICL operates in dual modes[1,2]

- **Task Retrieve (TR):** Recall learned patterns from pre-training

- **Task Learning (TL):** Learning from demonstrations during inference time

Challenge: It’s non-trivial to disentangle these two modes

1.Wang, Xiaolei, et al. "Investigating the Pre-Training Dynamics of In-Context Learning: Task Recognition vs. Task Learning." *arXiv preprint arXiv:2406.14022* (2024).
2.Lin, Ziqian, and Kangwook Lee. "Dual operating modes of in-context learning." *Forty-first International Conference on Machine Learning*. 2024.

Examples of TR and TL

Examples of TR and TL

Examples of TR and TL

Sentiment Classification

Input: The weather is so good!
Output: positive
Input: I don’t like that movie.
Output: negative
…

Examples of TR and TL

Sentiment Classification

Input: The weather is so good!
Output: positive
Input: I don’t like that movie.
Output: negative
…

TR

I’ve seen this task!
It’s sentiment classification.

Examples of TR and TL

Sentiment Classification

Input: The weather is so good!
Output: positive
Input: I don’t like that movie.
Output: negative
…

TR

I’ve seen this task!
It’s sentiment classification.

(a*b + 2025) mod (a+b) + 35

Input: 10 and 11
Output: 49
Input: 5 and 8
Output: 46
…

Examples of TR and TL

Sentiment Classification

Input: The weather is so good!
Output: positive
Input: I don’t like that movie.
Output: negative
…

TR

I’ve seen this task!
It’s sentiment classification.

(a*b + 2025) mod (a+b) + 35

Input: 10 and 11
Output: 49
Input: 5 and 8
Output: 46
…

TL

I never see this task.
Let me try to solve it by observing more demos…

Prior work: Disentangling TR/TL via label space manipulation[1]

1.Pan, Jane. *What in-context learning “learns” in-context: Disentangling task recognition and task learning*. MS thesis. Princeton University, 2023.

Prior work: Disentangling TR/TL via label space manipulation[1]

TR:

Randomize the labels

1.Pan, Jane. *What in-context learning "learns" in-context: Disentangling task recognition and task learning*. MS thesis. Princeton University, 2023.

Prior work: Disentangling TR/TL via label space manipulation[1]

TR:
Randomize the labels

```
Input: I love my school! There is ...
Output: positive

Input: The sky doesn’t look so nice
Output: negative

    ⋮
----------------------------------------
Input: I don’t love looking at the sky
```

1.Pan, Jane. *What in-context learning “learns” in-context: Disentangling task recognition and task learning*. MS thesis. Princeton University, 2023.

Prior work: Disentangling TR/TL via label space manipulation[1]

TR:
Randomize the labels

`Input: I love my school! There is …`
`Output:` ~~`positive`~~ **`positive`**

`Input: The sky doesn’t look so nice`
`Output:` ~~`negative`~~ **`positive`**

⋮

`Input: I don’t love looking at the sky`

1.Pan, Jane. *What in-context learning “learns” in-context: Disentangling task recognition and task learning*. MS thesis. Princeton University, 2023.

Prior work: Disentangling TR/TL via label space manipulation[1]

TR:
Randomize the labels

Input: I love my school! There is …
Output: ~~positive~~ **positive**

Input: The sky doesn’t look so nice
Output: ~~negative~~ **positive**

⋮

Input: I don’t love looking at the sky

LLM

1.Pan, Jane. *What in-context learning “learns” in-context: Disentangling task recognition and task learning*. MS thesis. Princeton University, 2023.

Prior work: Disentangling TR/TL via label space manipulation[1]

TR:
Randomize the labels

Input: I love my school! There is …
Output: ~~positive~~ **positive**

Input: The sky doesn’t look so nice
Output: ~~negative~~ **positive**

⋮

Input: I don’t love looking at the sky

LLM

Output: negative

1.Pan, Jane. *What in-context learning “learns” in-context: Disentangling task recognition and task learning*. MS thesis. Princeton University, 2023.

Prior work: Disentangling TR/TL via label space manipulation[1]

TR:
Randomize the labels

Input: I love my school! There is …
Output: ~~positive~~ **positive**

Input: The sky doesn’t look so nice
Output: ~~negative~~ **positive**

⋮

Input: I don’t love looking at the sky

LLM

Output: negative

TL:
Replacing the original labels with new labels

1.Pan, Jane. *What in-context learning “learns” in-context: Disentangling task recognition and task learning*. MS thesis. Princeton University, 2023.

Prior work: Disentangling TR/TL via label space manipulation[1]

TR:
Randomize the labels

```
Input: I love my school! There is ...
Output: ~~positive~~ positive

Input: The sky doesn't look so nice
Output: ~~negative~~ positive

    ⋮
-----------------------------------
Input: I don't love looking at the sky
```

LLM

```
Output: negative
```

TL:
Replacing the original labels with new labels

```
Input: I love my school! There is ...
Output: positive

Input: The sky doesn't look so nice
Output: negative

    ⋮
-----------------------------------
Input: I don't love looking at the sky
```

1.Pan, Jane. *What in-context learning "learns" in-context: Disentangling task recognition and task learning*. MS thesis. Princeton University, 2023.

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Prior work: Disentangling TR/TL via label space manipulation[1]

TR:
Randomize the labels

```
Input: I love my school! There is ...
Output: ~~positive~~ positive

Input: The sky doesn’t look so nice
Output: ~~negative~~ positive

⋮

Input: I don’t love looking at the sky
```

LLM

```
Output: negative
```

TL:
Replacing the original labels with new labels

```
Input: I love my school! There is ...
Output: ~~positive~~ *

Input: The sky doesn’t look so nice
Output: ~~negative~~ A

⋮

Input: I don’t love looking at the sky
```

1.Pan, Jane. *What in-context learning “learns” in-context: Disentangling task recognition and task learning*. MS thesis. Princeton University, 2023.

Prior work: Disentangling TR/TL via label space manipulation[1]

TR:
Randomize the labels

Input: I love my school! There is …
Output: ~~positive~~ **positive**

Input: The sky doesn’t look so nice
Output: ~~negative~~ **positive**

⋮

Input: I don’t love looking at the sky

LLM

Output: negative

TL:
Replacing the original labels with new labels

Input: I love my school! There is …
Output: ~~positive~~ *****

Input: The sky doesn’t look so nice
Output: ~~negative~~ **A**

⋮

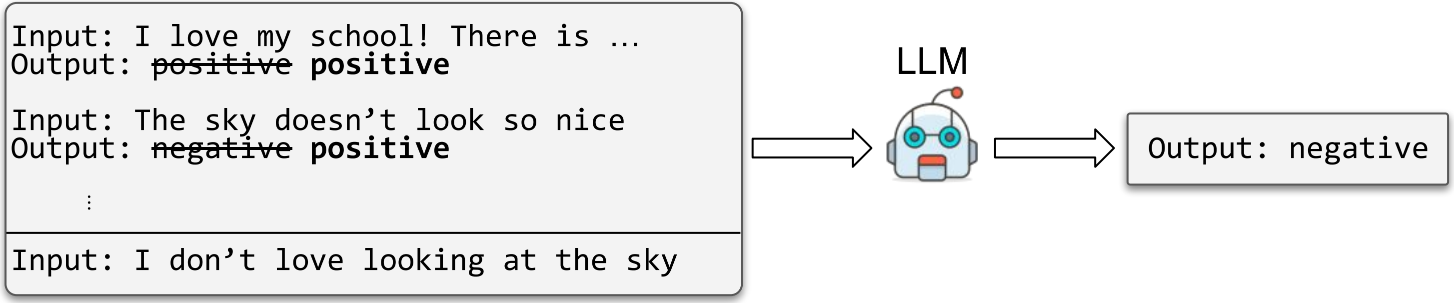
Input: I don’t love looking at the sky

LLM

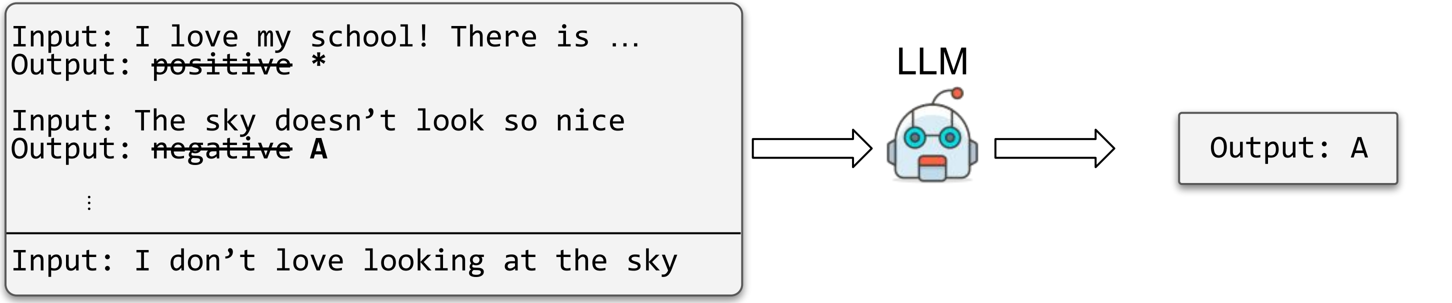
1.Pan, Jane. *What in-context learning “learns” in-context: Disentangling task recognition and task learning*. MS thesis. Princeton University, 2023.

Prior work: Disentangling TR/TL via label space manipulation[1]

TR:
Randomize the labels



TL:
Replacing the original labels with new labels



1.Pan, Jane. *What in-context learning "learns" in-context: Disentangling task recognition and task learning*. MS thesis. Princeton University, 2023.

High-level Idea and Solution

High-level Idea and Solution

High-level idea:
1. Create a task that is rather **unlikely to be included** during pre-training
2. See if LLMs can solve the task via ICL – **Evidence for TL**

High-level Idea and Solution

High-level idea:
1. Create a task that is rather **unlikely to be included** during pre-training
2. See if LLMs can solve the task via ICL – **Evidence for TL**

Solution:
ICL ciphers - a class of task reformulations based on substitution ciphers

Substitution Cipher: Tool to define "learning"

Substitution Cipher: Tool to define "learning"

What is a substitution cipher?
- It is a method of encrypting in which **units** of plaintext are replaced with the ciphertext, in a defined manner (mapping)

Substitution Cipher: Tool to define "learning"

What is a substitution cipher?
- It is a method of encrypting in which **units** of plaintext are replaced with the ciphertext, in a defined manner (mapping)

The unit could a single letters, pair of letters, subwords, words…

Substitution Cipher: Tool to define "learning"

What is a substitution cipher?
- It is a method of encrypting in which **units** of plaintext are replaced with the ciphertext, in a defined manner (mapping)

The unit could a single letters, pair of letters, subwords, words…

e.g. Caesar cipher is one kind of classic substitution cipher

ICL Ciphers: Big picture

ICL Ciphers: Big picture

What is a ICL cipher?
- A **token-level** substitution cipher, applied to ICL inputs

ICL Ciphers: Big picture

What is a ICL cipher?
- A **token-level** substitution cipher, applied to ICL inputs

Two types of ICL cipher
- Bijective Cipher: A **reversible** framework, where a **bijective mapping** between original tokens and encoded tokens is maintained
- Non-bijective Cipher: An **non-reversible** framework, where such bijective mapping doesn't exist

ICL Ciphers: Big picture

What is a ICL cipher?
- A **token-level** substitution cipher, applied to ICL inputs

Two types of ICL cipher
- Bijective Cipher: A **reversible** framework, where a **bijective mapping** between original tokens and encoded tokens is maintained
- Non-bijective Cipher: An **non-reversible** framework, where such bijective mapping doesn't exist

We use the **performance gap** between Bijective and Non-bijective Cipher to **quantify TL**

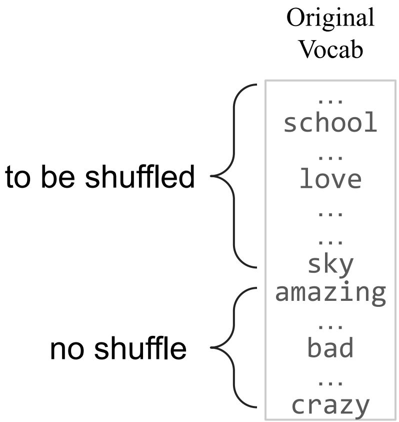
ICL Ciphers: Bijective cipher details

Original
Vocab

...
school
...
love
...
...
sky
amazing
...
bad
...
crazy

ICL Ciphers: Bijective cipher details

Bijective Cipher:
1. Pick a shuffle rate **r** from 0 to 1, then choose **r * Vocab size** tokens to be shuffled

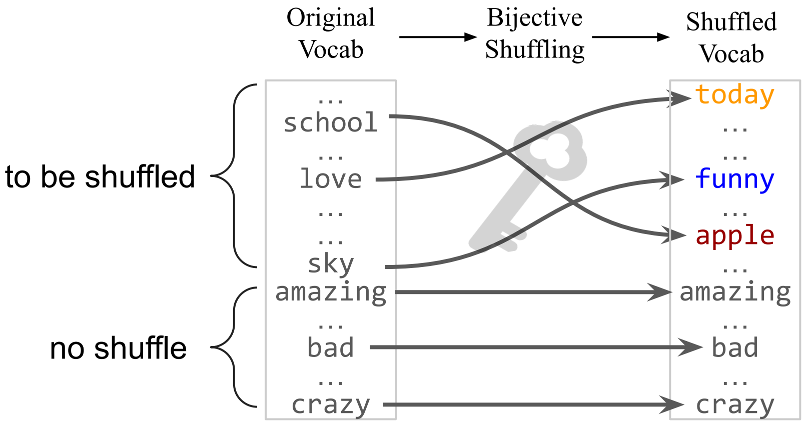


ICL Ciphers: Bijective cipher details

Bijective Cipher:

1. Pick a shuffle rate **r** from 0 to 1, then choose **r * Vocab size** tokens to be shuffled

2. Shuffle chosen tokens and create a shuffled vocab, which maintains a bijective mapping with the original vocab



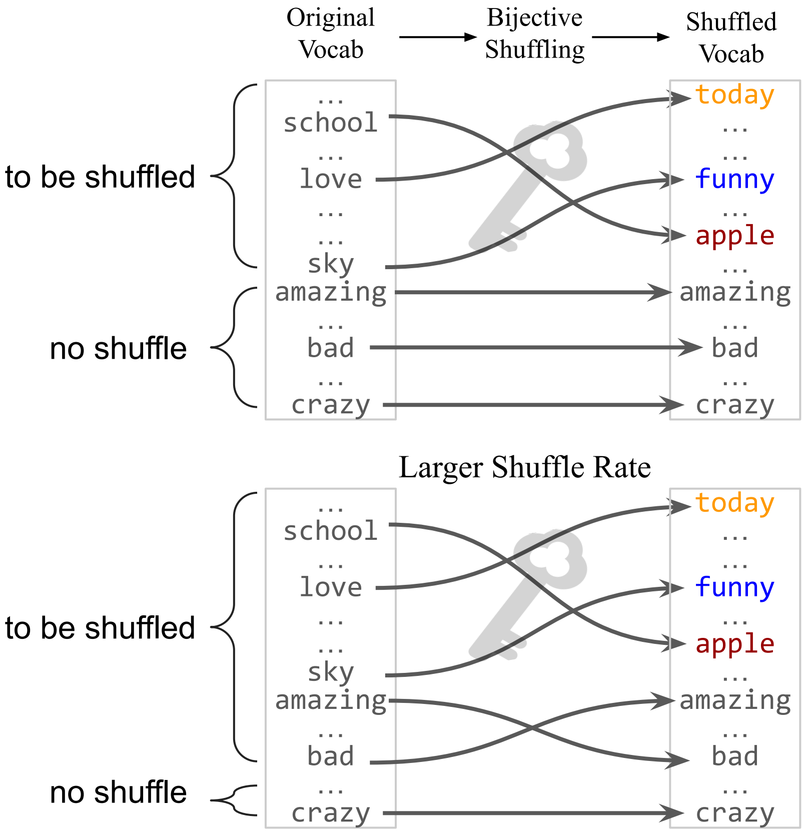
ICL Ciphers: Bijective cipher details

Bijective Cipher:

1. Pick a shuffle rate **r** from 0 to 1, then choose **r * Vocab size** tokens to be shuffled

2. Shuffle chosen tokens and create a shuffled vocab, which maintains a bijective mapping with the original vocab

Larger shuffle rate means more tokens are shuffled

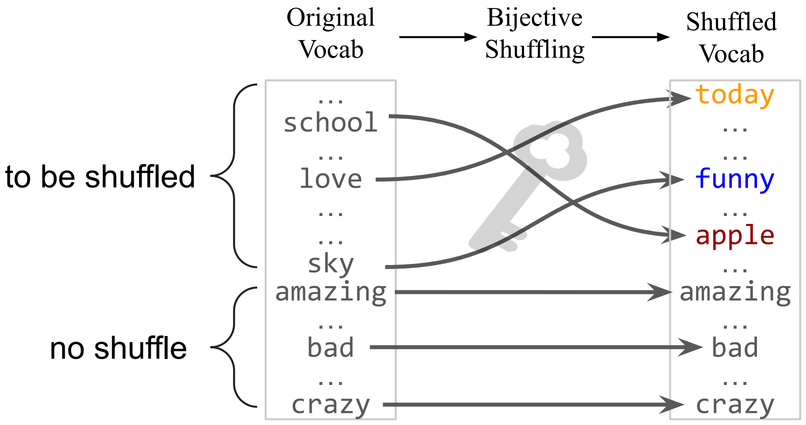


ICL Ciphers: Bijective cipher details

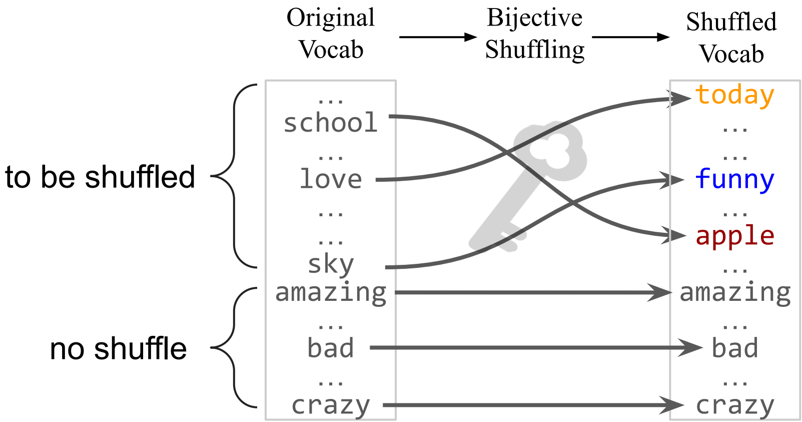
Bijective Cipher:

1. Pick a shuffle rate **r** from 0 to 1, then choose **r * Vocab size** tokens to be shuffled

2. Shuffle chosen tokens and create a shuffled vocab, which maintains a bijective mapping with the original vocab



ICL Ciphers: Bijective cipher details



Bijective Cipher:

1. Pick a shuffle rate **r** from 0 to 1, then choose **r * Vocab size** tokens to be shuffled

2. Shuffle chosen tokens and create a shuffled vocab, which maintains a bijective mapping with the original vocab

3.Replace tokens in the input text according to the bijective mapping

In-Context Learning

```
Input: I love my school! There is
...
Output: positive

Input: The sky doesn't look so nice
Output: negative

    ⋮

Input: I don't love looking at the
sky
```

Ciphered In-Context Learning

```
Input: I love my school! There is
...
Output: positive

Input: The sky doesn't look so nice
Output: negative

    ⋮

Input: I don't love looking at the
sky
```

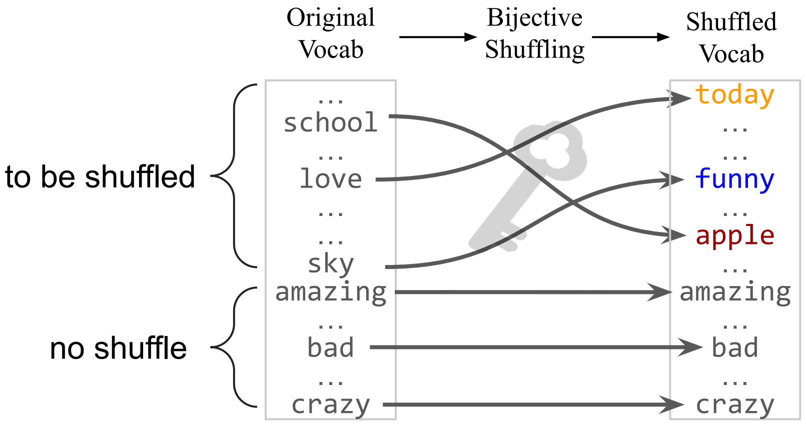
ICL Ciphers: Bijective cipher details

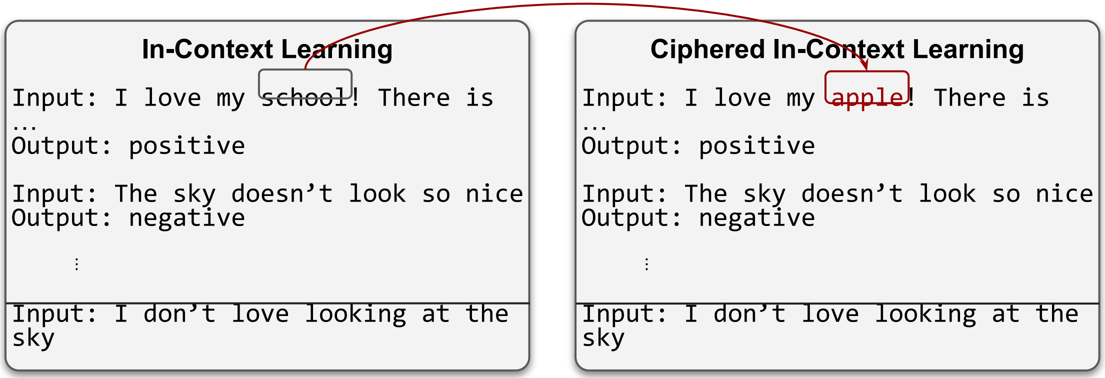
Bijective Cipher:

1. Pick a shuffle rate **r** from 0 to 1, then choose **r * Vocab size** tokens to be shuffled

2. Shuffle chosen tokens and create a shuffled vocab, which maintains a bijective mapping with the original vocab

3.Replace tokens in the input text according to the bijective mapping





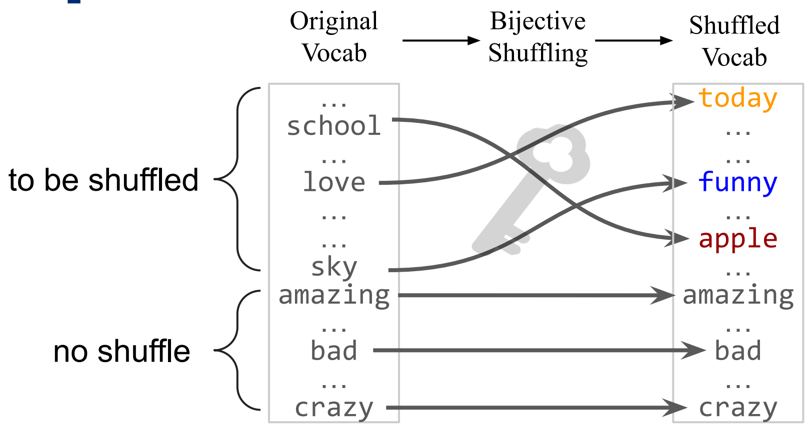
ICL Ciphers: Bijective cipher details

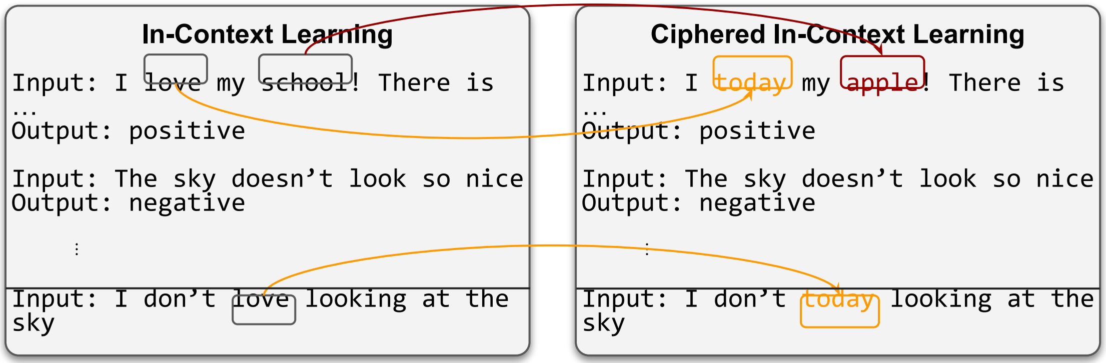
Bijective Cipher:

1. Pick a shuffle rate **r** from 0 to 1, then choose **r * Vocab size** tokens to be shuffled

2. Shuffle chosen tokens and create a shuffled vocab, which maintains a bijective mapping with the original vocab

3.Replace tokens in the input text according to the bijective mapping





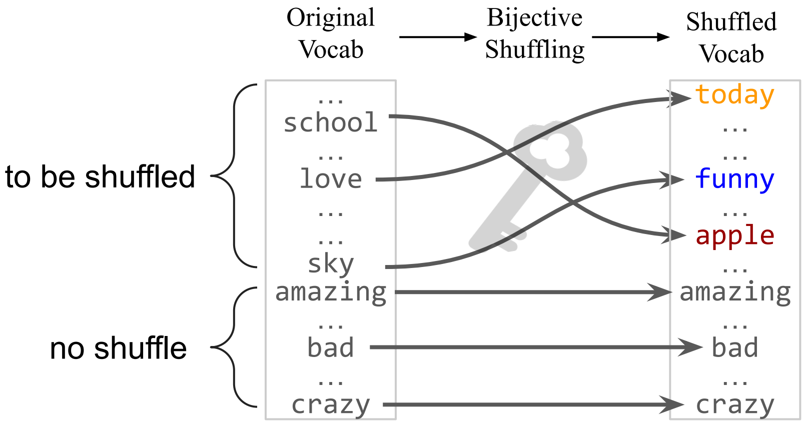
ICL Ciphers: Bijective cipher details

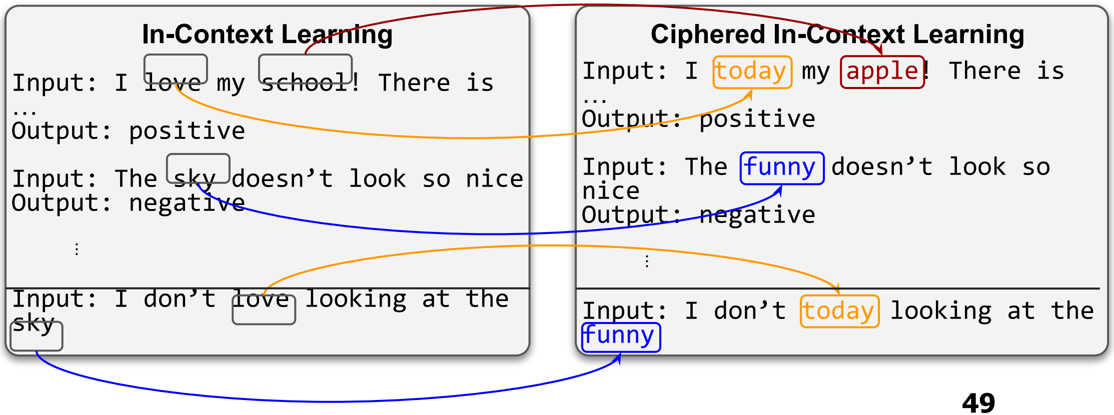
Bijective Cipher:

1. Pick a shuffle rate **r** from 0 to 1, then choose **r * Vocab size** tokens to be shuffled

2. Shuffle chosen tokens and create a shuffled vocab, which maintains a bijective mapping with the original vocab

3.Replace tokens in the input text according to the bijective mapping





ICL Ciphers: Non-bijective cipher details

ICL Ciphers: Non-bijective cipher details

Non-bijective Cipher:

1. Conduct replacing on the same tokens as bijective cipher

ICL Ciphers: Non-bijective cipher details

Non-bijective Cipher:

1. Conduct replacing on the same tokens as bijective cipher

In-Context Learning

```
Input: I love my school! There is …
Output: positive

Input: The sky doesn’t look so nice
Output: negative

     ⋮
```

```
Input: I don’t love looking at the sky
```

ICL Ciphers: Non-bijective cipher details

Non-bijective Cipher:

1. Conduct replacing on the same tokens as bijective cipher

In-Context Learning

```
Input: I love my school! There is ...
Output: positive

Input: The sky doesn’t look so nice
Output: negative

    ⋮

Input: I don’t love looking at the sky
```

ICL Ciphers: Non-bijective cipher details

Non-bijective Cipher:

1. Conduct replacing on the same tokens as bijective cipher

2. For each token that should be ciphered (replaced), replace it with a randomly selected token

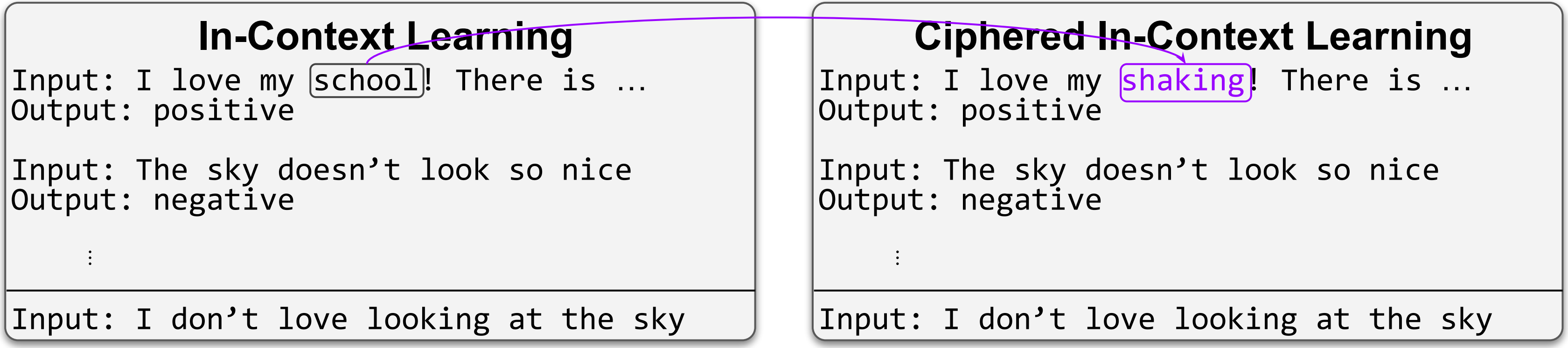
In-Context Learning	Ciphered In-Context Learning
Input: I love my school! There is … Output: positive Input: The sky doesn't look so nice Output: negative ⋮	Input: I love my school! There is … Output: positive Input: The sky doesn't look so nice Output: negative ⋮
Input: I don't love looking at the sky	Input: I don't love looking at the sky

ICL Ciphers: Non-bijective cipher details

Non-bijective Cipher:

1. Conduct replacing on the same tokens as bijective cipher

2. For each token that should be ciphered (replaced), replace it with a randomly selected token

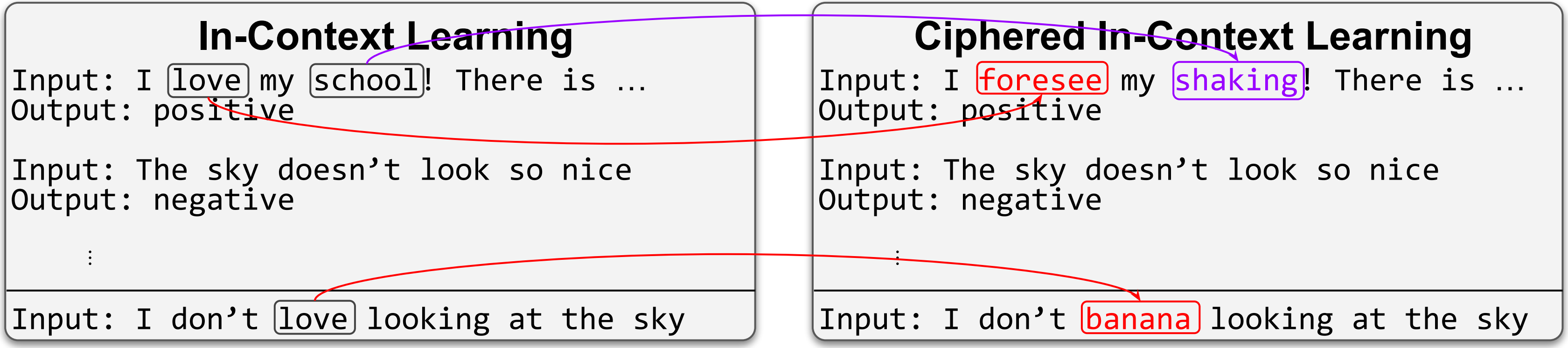


ICL Ciphers: Non-bijective cipher details

Non-bijective Cipher:

1. Conduct replacing on the same tokens as bijective cipher

2. For each token that should be ciphered (replaced), replace it with a randomly selected token

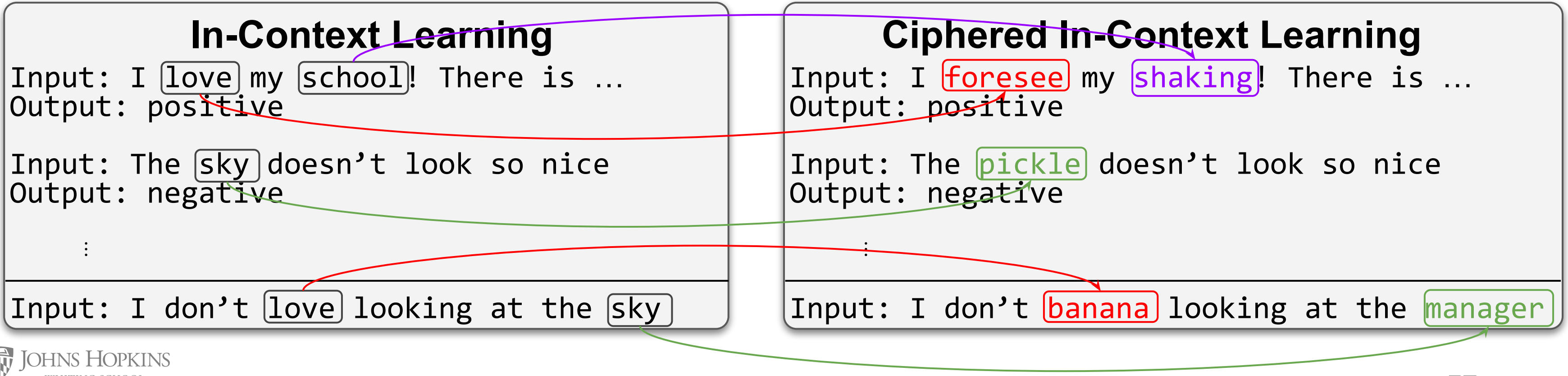


ICL Ciphers: Non-bijective cipher details

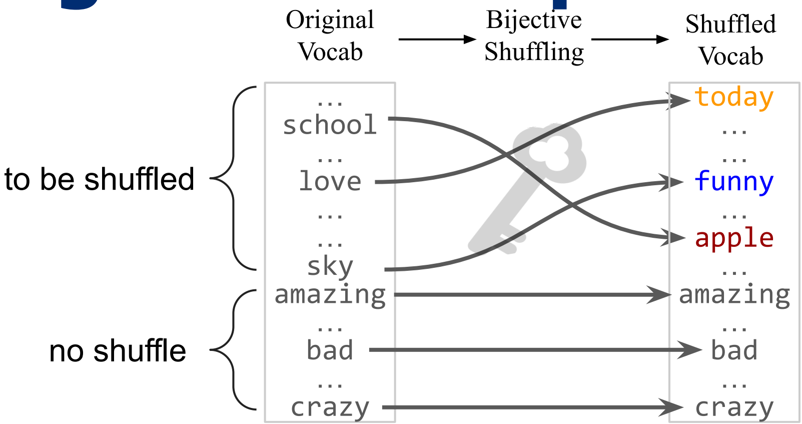
Non-bijective Cipher:

1. Conduct replacing on the same tokens as bijective cipher

2. For each token that should be ciphered (replaced), replace it with a randomly selected token



Quantifying "Task Learning" via ICL Ciphers

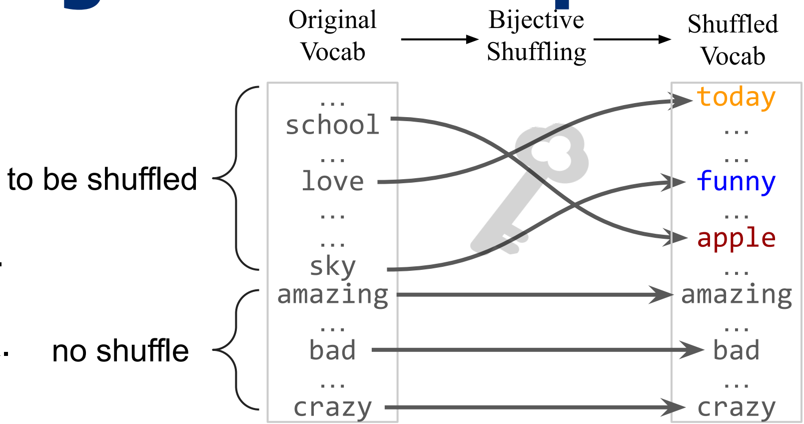


Quantifying “Task Learning” via ICL Ciphers

Original (plain) text: I **love** my cats and they **love** me back.

Bijective Cipher: I **today** my cats and they **today** me back.

Non-bijective Cipher: I **pickle** my cats and they **share** me back.

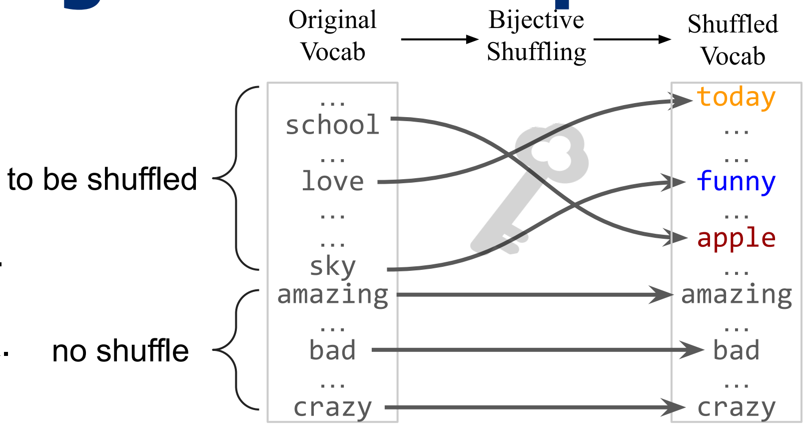


Quantifying “Task Learning” via ICL Ciphers

Original (plain) text: I **love** my cats and they **love** me back.

Bijective Cipher: I **today** my cats and they **today** me back.

Non-bijective Cipher: I **pickle** my cats and they **share** me back.



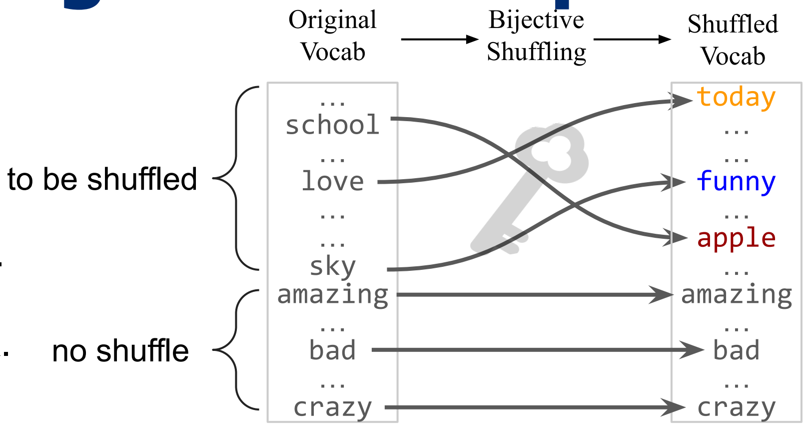
Working Hypothesis:
The ***bijective mappings*** between original and shuffled vocabs ensures bijective ciphers are **reversible and learnable**, while non-bijective ciphers are **non-reversible and unlearnable** as their replacement process don’t follow the bijective mapping

Quantifying "Task Learning" via ICL Ciphers

Original (plain) text: I **love** my cats and they **love** me back.

Bijective Cipher: I **today** my cats and they **today** me back.

Non-bijective Cipher: I **pickle** my cats and they **share** me back.



Two ciphers conduct replacement on same tokens. Therefore, the performance gap between them only comes from the way of ciphering, which reflects how much the LLM learns about the bijective mapping — **Quantified evidence of TL!**

Note that LLMs are not required to fully "deciphering" the ciphers, but only need to (internally) capture related information or attributes (e.g. sentiment) of the ciphered tokens that help solve the reformulated tasks. — **The reformulated tasks are different!**

Results: Consistent gaps across datasets and models

Model → / Dataset (shuffle rate) ↓	Cipher	20-shot			
		Llama3.1	Qwen2.5	Olmo	Gemma2
SST-2 ($r = 0.5$)	NON-BIJECTIVE	58.3	69.0	67.7	70.5
	BIJECTIVE	63.1 (+4.8 ↑)*	73.5 (+4.5 ↑)*	72.7 (+5.0 ↑)*	74.2 (+3.7 ↑)*

Green: Gain Red: Loss *: Statistically significant

Results: Consistent gaps across datasets and models

Model → / Dataset (shuffle rate) ↓	Cipher	20-shot			
		Llama3.1	Qwen2.5	Olmo	Gemma2
Amazon ($r = 0.6$)	NON-BIJECTIVE	64.7	72.6	77.2	80.8
	BIJECTIVE	72.3 (+7.6 ↑)*	77.9 (+5.3 ↑)*	80.2 (+3.0 ↑)*	85.0 (+4.2 ↑)*

Green: Gain Red: Loss *: Statistically significant

Results: Consistent gaps across datasets and models

Model →	Cipher	20-shot			
Dataset (shuffle rate) ↓		Llama3.1	Qwen2.5	Olmo	Gemma2
HellaSwag ($r = 0.3$)	NON-BIJECTIVE	29.7	52.8	25.9	37.1
	BIJECTIVE	31.9 (+2.2 ↑)*	62.3 (+9.5 ↑)*	26.1 (+0.2 ↑)*	36.6 (-0.5 ↓)

Green: Gain Red: Loss *: Statistically significant

Results: Consistent gaps across datasets and models

Model → / Dataset (shuffle rate) ↓	Cipher	20-shot Llama3.1	20-shot Qwen2.5	20-shot Olmo	20-shot Gemma2
WinoGrande ($r = 0.1$)	NON-BIJECTIVE BIJECTIVE	53.7 55.5 (+1.8 ↑)*	61.3 62.5 (+1.2 ↑)	53.4 53.1 (-0.3 ↓)	63.5 63.5 (+0.0 ↑)

Green: Gain   Red: Loss   *: Statistically significant

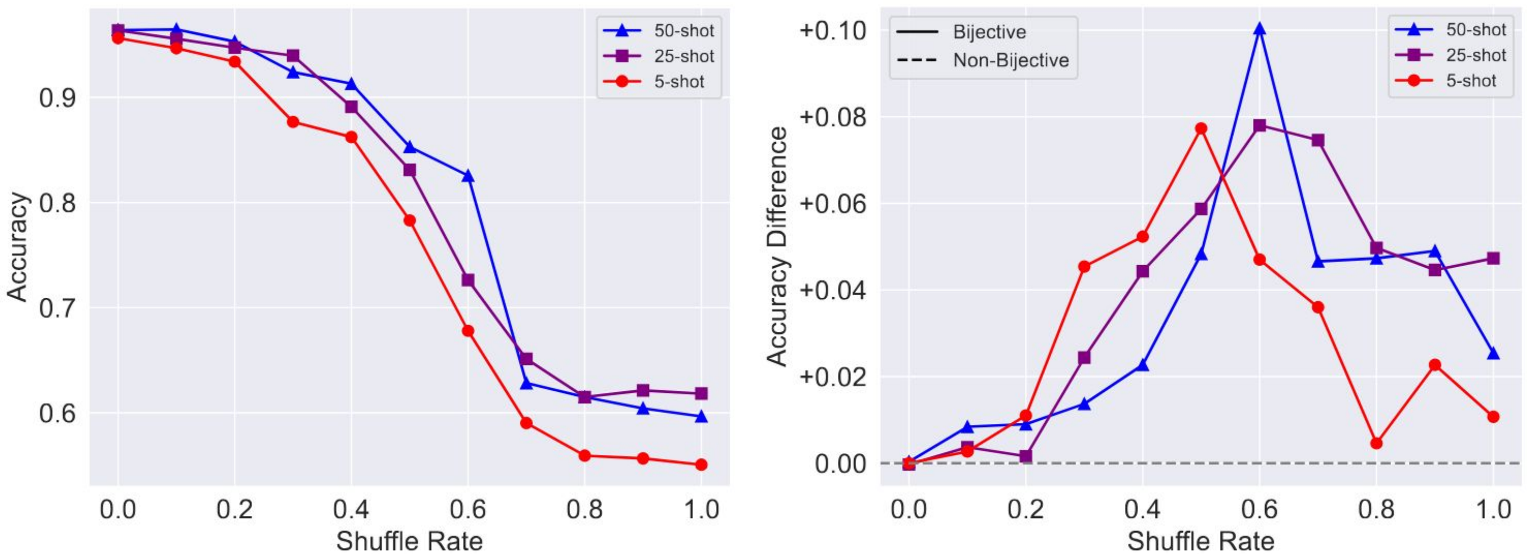
Results: Consistent gaps across datasets and models

Model → / Dataset (shuffle rate) ↓	Cipher	20-shot Llama3.1	20-shot Qwen2.5	20-shot Olmo	20-shot Gemma2
SST-2 ($r = 0.5$)	NON-BIJECTIVE	58.3	69.0	67.7	70.5
	BIJECTIVE	63.1 (+4.8 ↑)*	73.5 (+4.5 ↑)*	72.7 (+5.0 ↑)*	74.2 (+3.7 ↑)*
Amazon ($r = 0.6$)	NON-BIJECTIVE	64.7	72.6	77.2	80.8
	BIJECTIVE	72.3 (+7.6 ↑)*	77.9 (+5.3 ↑)*	80.2 (+3.0 ↑)*	85.0 (+4.2 ↑)*
HellaSwag ($r = 0.3$)	NON-BIJECTIVE	29.7	52.8	25.9	37.1
	BIJECTIVE	31.9 (+2.2 ↑)*	62.3 (+9.5 ↑)*	26.1 (+0.2 ↑)*	36.6 (-0.5 ↓)
WinoGrande ($r = 0.1$)	NON-BIJECTIVE	53.7	61.3	53.4	63.5
	BIJECTIVE	55.5 (+1.8 ↑)*	62.5 (+1.2 ↑)	53.1 (-0.3 ↓)	63.5 (+0.0 ↑)

Green: Gain   Red: Loss   *: Statistically significant

Consistent **gaps** between bijective and non-bijective ciphers across different **models and datasets**, demonstrating LLMs are able to solve the bijective cipher via ICL — evidence for TL.

Results: Performance decreases as shuffle rate increase



As shuffle rate increases, the accuracy drops but remain above random.
The gap first increases to a certain peak (r=0.6) then decreases

Results: Gaps increase with more demonstrations

Shots → / Dataset (shuffle rate)↓	Cipher	Model: Llama 3.1 8B 5-shot	10-shot	15-shot	20-shot	25-shot	50-shot
SST-2 ($r = 0.5$)	NON-BIJECTIVE	56.9	59.5	58.6	58.3	62.6	58.4
	BIJECTIVE	59.5 (+2.6 ↑)*	61.0 (+1.5 ↑)	60.8 (+2.2 ↑)	63.1 (+4.8 ↑)*	65.4 (+2.8 ↑)*	64.9 (+6.5 ↑)*

Green: Gain Red: Loss *: Statistically significant

Results: Gaps increase with more demonstrations

Shots → / Dataset (shuffle rate)↓	Cipher	Model: Llama 3.1 8B					
		5-shot	10-shot	15-shot	20-shot	25-shot	50-shot
Amazon ($r = 0.6$)	NON-BIJECTIVE	63.1	61.8	68.1	64.7	64.8	72.5
	BIJECTIVE	67.8 (+4.7 ↑)*	67.6 (+5.8 ↑)*	74.5 (+6.4 ↑)*	72.3 (+7.6 ↑)*	72.6 (+7.8 ↑)*	82.6 (+10.1 ↑)*

Green: Gain Red: Loss *: Statistically significant

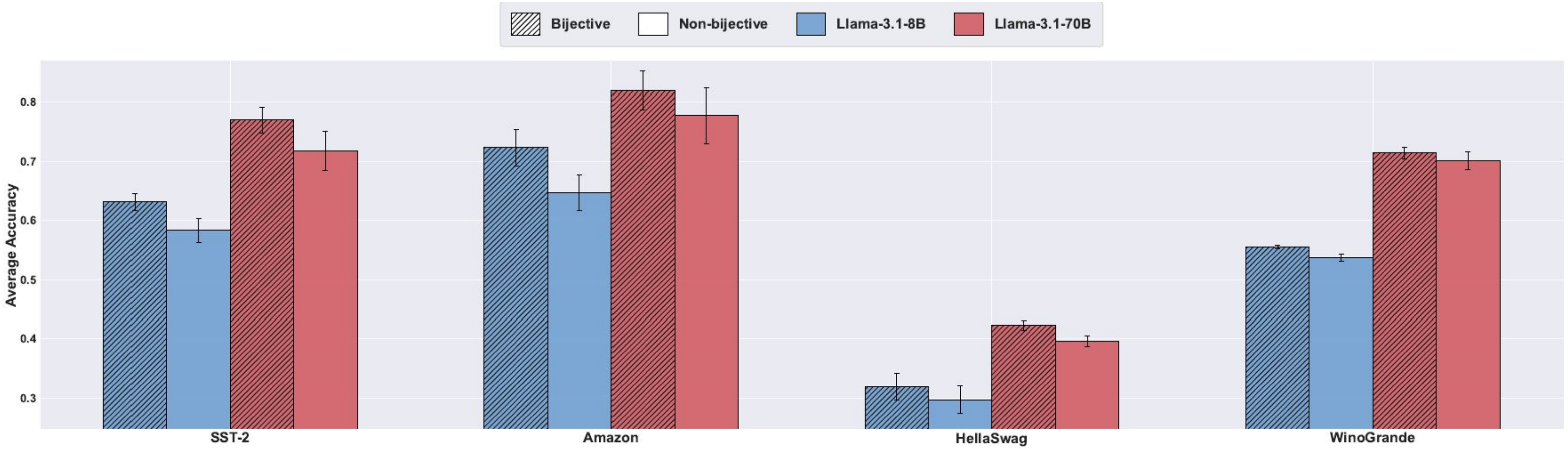
Results: Gaps increase with more demonstrations

Shots → / Dataset (shuffle rate)↓	Cipher	Model: Llama 3.1 8B					
		5-shot	10-shot	15-shot	20-shot	25-shot	50-shot
SST-2 ($r = 0.5$)	NON-BIJECTIVE	56.9	59.5	58.6	58.3	62.6	58.4
	BIJECTIVE	59.5 (+2.6 ↑)*	61.0 (+1.5 ↑)	60.8 (+2.2 ↑)	63.1 (+4.8 ↑)*	65.4 (+2.8 ↑)*	64.9 (+6.5 ↑)*
Amazon ($r = 0.6$)	NON-BIJECTIVE	63.1	61.8	68.1	64.7	64.8	72.5
	BIJECTIVE	67.8 (+4.7 ↑)*	67.6 (+5.8 ↑)*	74.5 (+6.4 ↑)*	72.3 (+7.6 ↑)*	72.6 (+7.8 ↑)*	82.6 (+10.1 ↑)*
HellaSwag ($r = 0.3$)	NON-BIJECTIVE	31.7	29.7	30.7	29.7	30.9	33.1
	BIJECTIVE	34.2 (+2.5 ↑)*	31.7 (+2.0 ↑)	34.1 (+3.4 ↑)*	31.9 (+2.2 ↑)*	31.6 (+0.7 ↑)	33.9 (+0.8 ↑)
WinoGrande ($r = 0.1$)	NON-BIJECTIVE	54.9	53.2	53.7	53.7	53.3	54.3
	BIJECTIVE	56.3 (+1.4 ↑)	53.8 (+0.6 ↑)*	54.2 (+0.5 ↑)*	55.5 (+1.8 ↑)*	54.6 (+1.3 ↑)*	55.5 (+1.2 ↑)*

Green: Gain Red: Loss *: Statistically significant

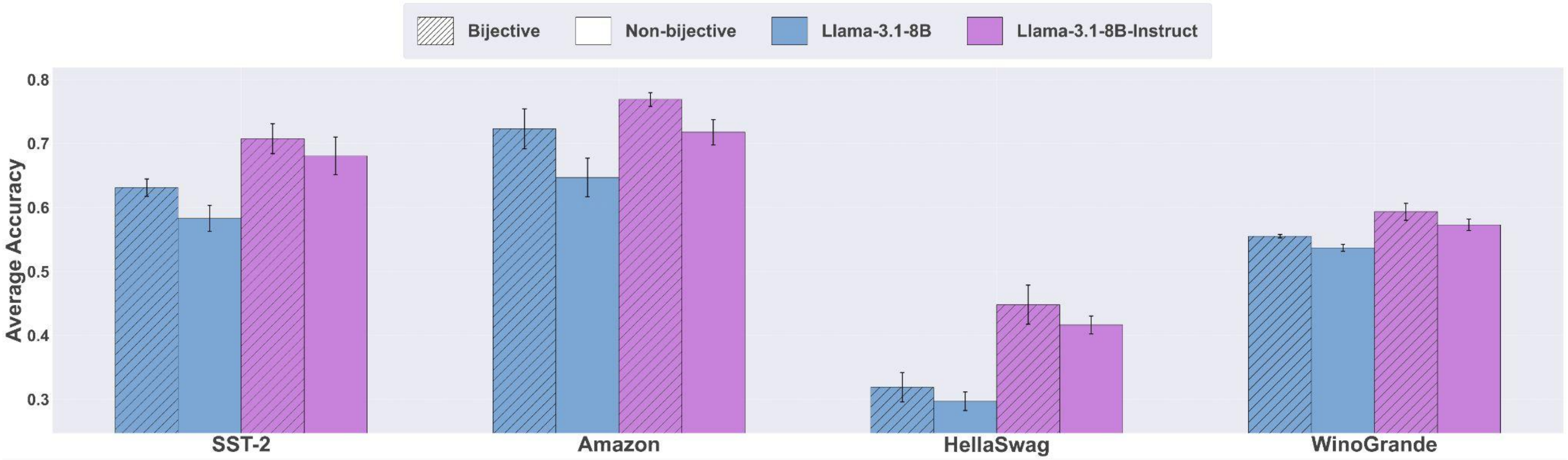
The gap between bijective and non-bijective ciphers increases as few-shot number grows

Results: Larger model has on-par gaps with smaller model



The gaps between bijective and non-bijective ciphers exist in **larger model**, on-par with smaller model

Results: Mixed gap difference between aligned/pretrained models



The aligned model has better absolute performance.
The differences in gaps for pretrained and aligned models are mixed.

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Conclusion

Conclusion

- **Our motivation:**
  ICL has dual operating modes — TR and TL, which are non-trivial to disentangle

Conclusion

- **Our motivation:**
  ICL has dual operating modes — TR and TL, which are non-trivial to disentangle
- **Our solution:**
  We propose a new class of task reformulations — *ICL ciphers*, which is very unlikely to be included in pretraining. We use the gaps between bijective and non-bijective ciphers to quantify TL.

Conclusion

- **Our motivation:**
  ICL has dual operating modes — TR and TL, which are non-trivial to disentangle
- **Our solution:**
  We propose a new class of task reformulations — *ICL ciphers*, which is very unlikely to be included in pretraining. We use the gaps between bijective and non-bijective ciphers to quantify TL.
- **Our findings:**
  Gaps exist across models and datasets — Evidence for TL

Conclusion

- **Our motivation:**
  ICL has dual operating modes — TR and TL, which are non-trivial to disentangle
- **Our solution:**
  We propose a new class of task reformulations — *ICL ciphers*, which is very unlikely to be included in pretraining. We use the gaps between bijective and non-bijective ciphers to quantify TL.
- **Our findings:**
  Gaps exist across models and datasets — Evidence for TL
- **Future work:**
  1. More models/datasets/interpretability analysis
  2. Different levels of ciphering (e.g. word)