ICML
International Conference
On Machine Learning

Spotlight
Poster

# On the Power of Context-Enhanced Learning in LLMs

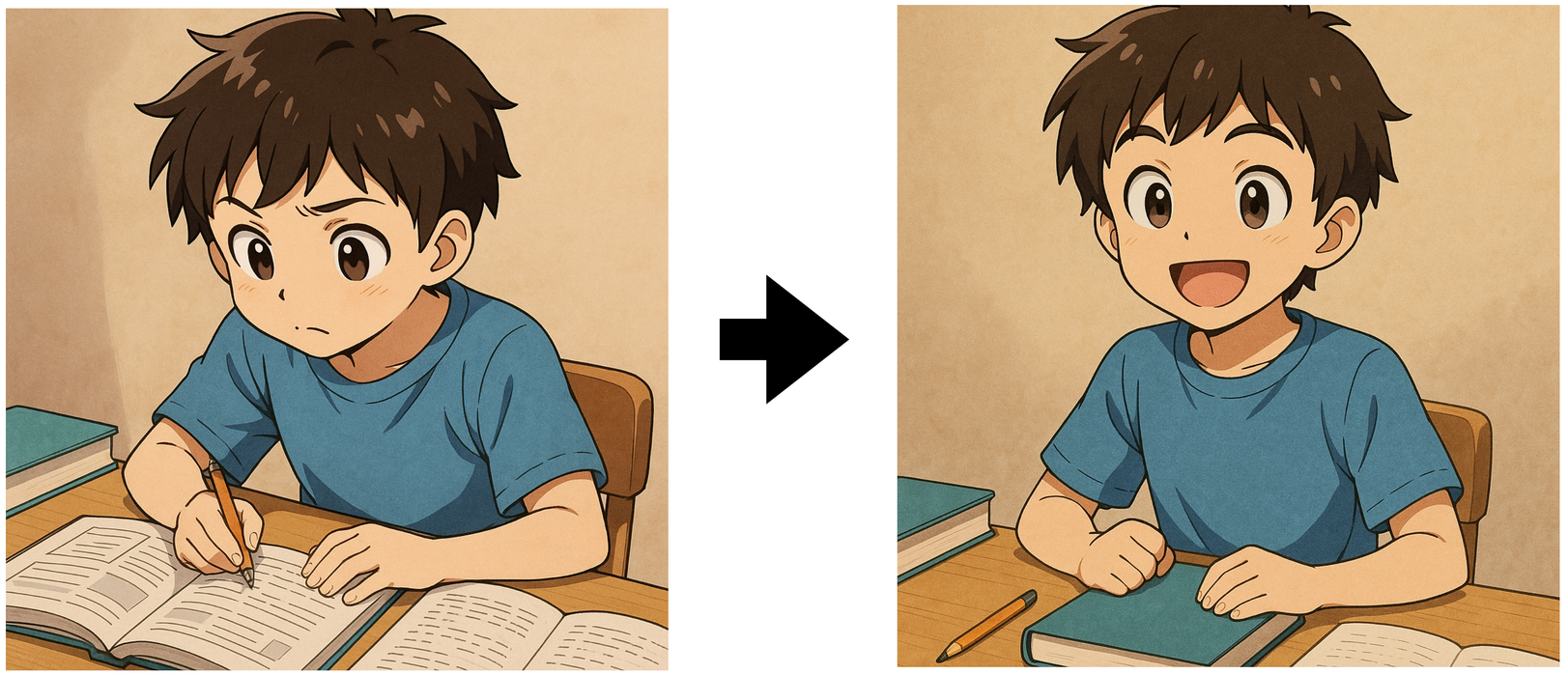
Xingyu Zhu*, Abhishek Panigrahi*, Sanjeev Arora

PRINCETON
UNIVERSITY

PRINCETON
Language + Intelligence

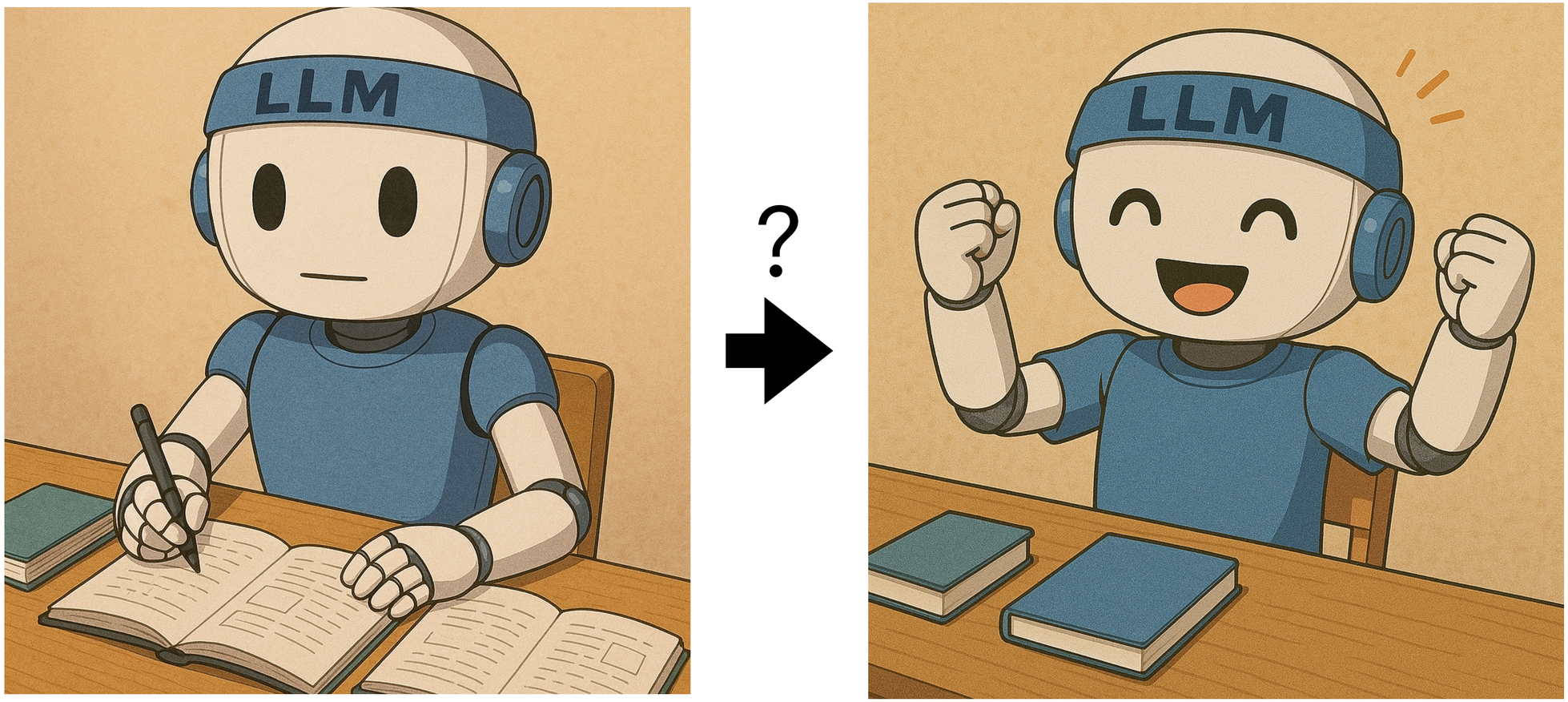
# Motivation

- When studying for assignments, kids have textbooks open in front of them.
- The textbooks in context benefit their internalization of knowledge, but they usually do not verbatim memorize the textbooks.



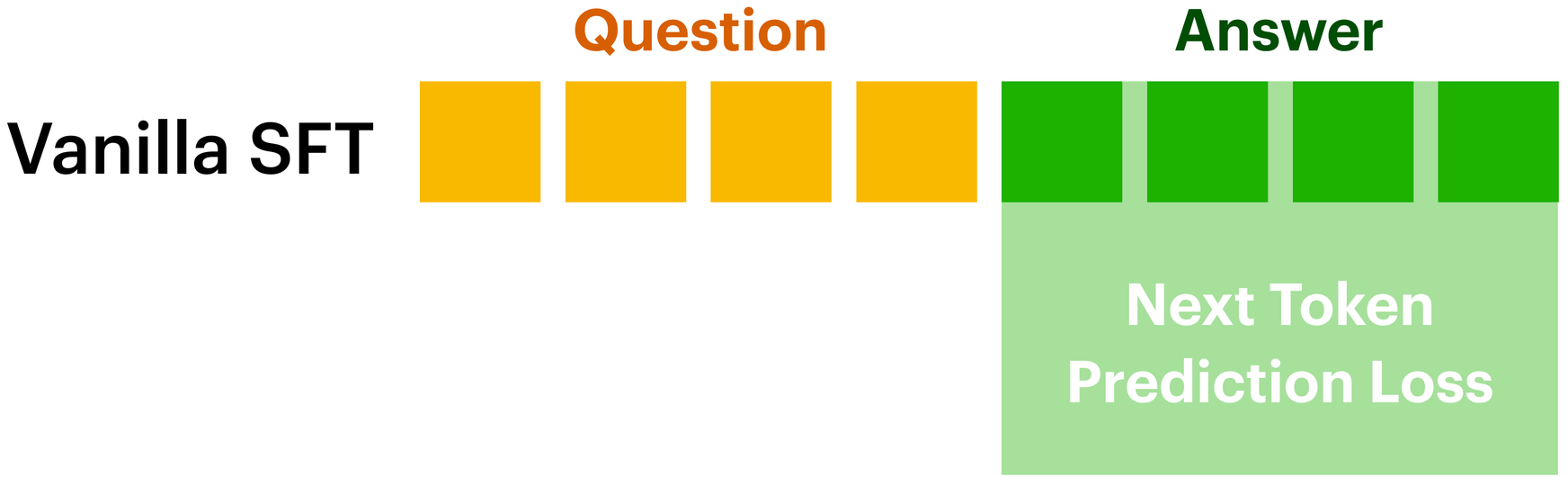
# Motivation

- Does similar behavior exist for the training of LLMs?

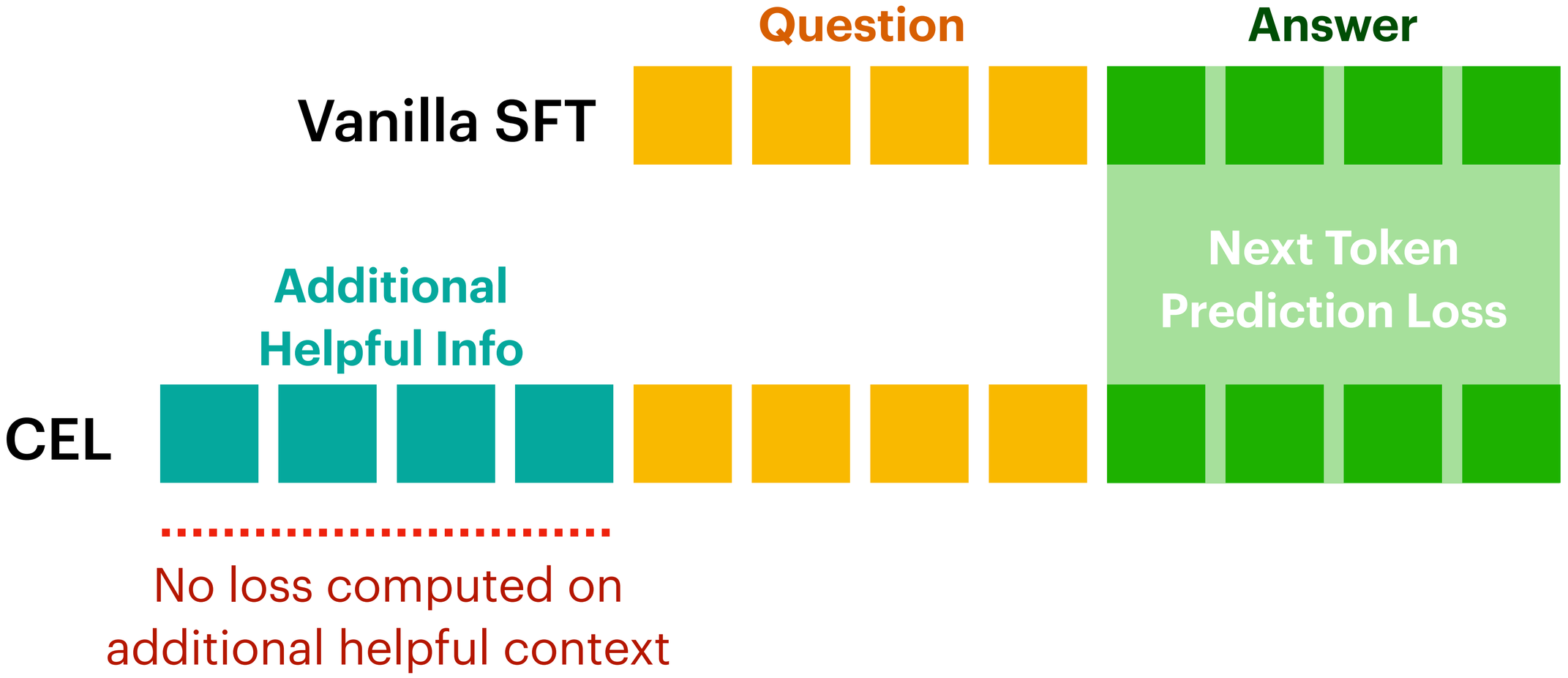


To what extent can LLM training benefit from **in-context** information with **no gradients computed on them**?

# Context-Enhanced Learning (CEL)

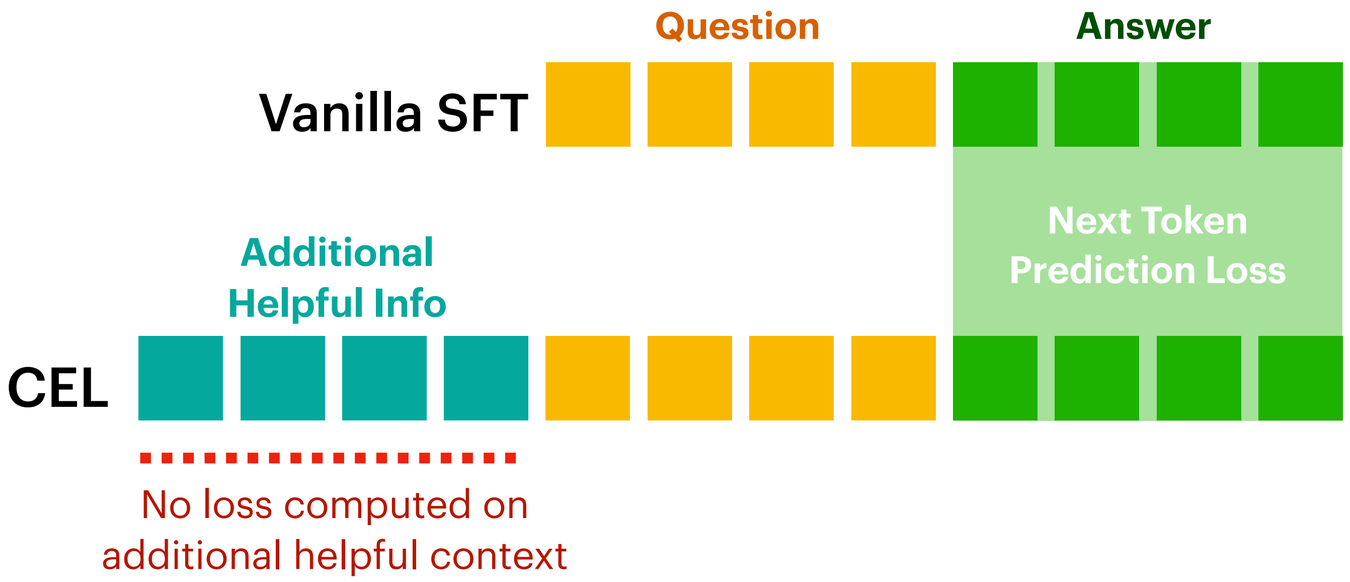


# Context-Enhanced Learning (CEL)




The additional helpful context is **only presented in training**

# Context-Enhanced Learning (CEL)




- Examples of helpful context:
  - Few-shot examples[1]
  - URLs of the source[2]
  - Thought-provoking guidance[3]
- The helpful context can depend on the input and training step - in-context curriculum

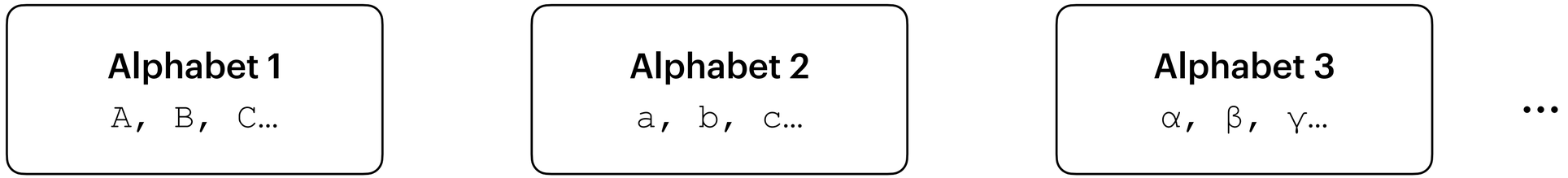
[1] Liao, Huanxuan, et al. "SKIntern: Internalizing Symbolic Knowledge for Distilling Better CoT Capabilities into Small Language Models."
[2] Gao, Tianyu, et al. "Metadata Conditioning Accelerates Language Model Pre-training.
[3] Choi, Younwoo, et al. "Teaching LLMs How to Learn with Contextual Fine-Tuning."

# Main Results

1. CEL (Context-enhanced learning) can be **exponentially more sample efficient** vs. vanilla SFT
2. Model **requires certain ICL capability** to benefit from CEL

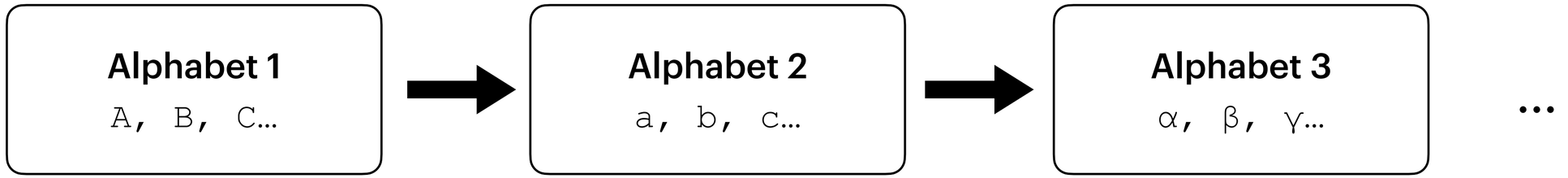
# Synthetic Testbed: Multi-Layer Translation (MLT)

- $d + 1$ alphabets / languages



# Synthetic Testbed: Multi-Layer Translation (MLT)

- $d + 1$ alphabets / languages
- $d$ phrasebooks to translate between 2-tuples in consecutive alphabets




**Phrasebook $\pi_1$**

BC → eb DE → hd

FG → af HA → gc

…

**Phrasebook $\pi_2$**

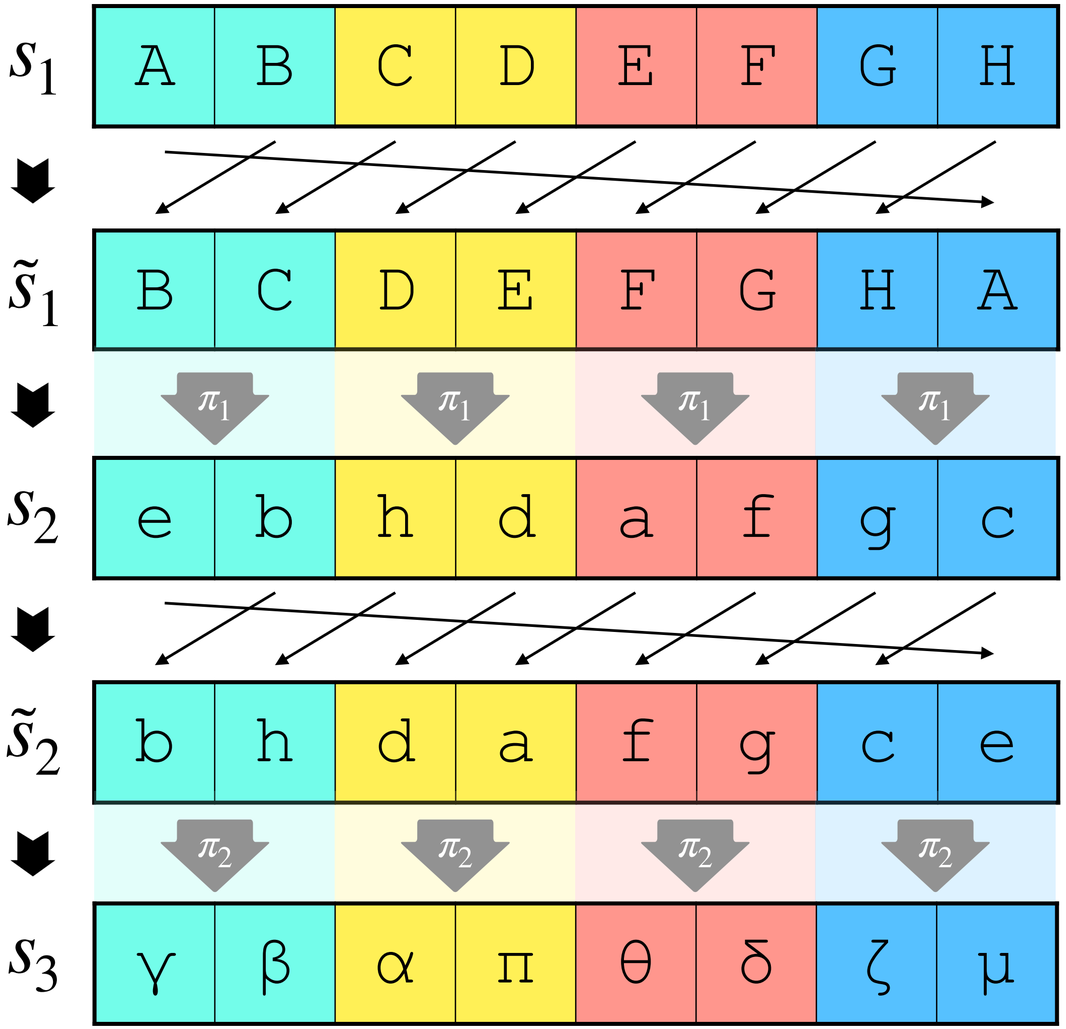
bh → γβ da → απ

fg → θδ ce → ζμ

…

…

# Multi-Layer Translation (MLT)




**Circular Shift**

**Phrasebook $\pi_1$**

BC → eb DE → hd

FG → af HA → gc

…

**Circular Shift**

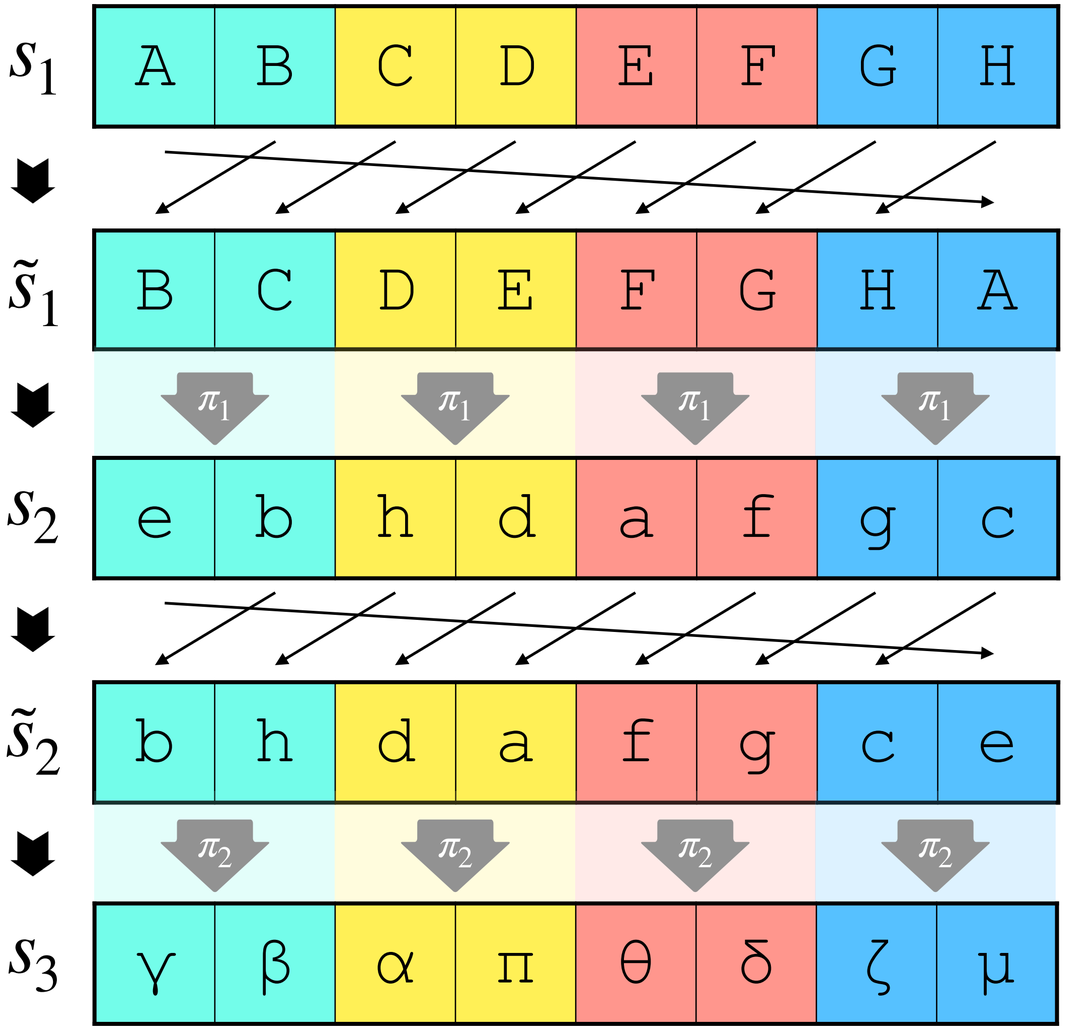
**Phrasebook $\pi_2$**

bh → γβ da → απ

fg → θδ ce → ζμ

…

$$s_{d+1} = f_{\pi_d} \circ f_{\circlearrowright} \circ f_{\pi_{d-1}} \circ f_{\circlearrowright} \circ \cdots \circ f_{\pi_1} \circ f_{\circlearrowright}(s_1)$$




**Circular Shift**

**Phrasebook $\pi_1$**

BC → eb DE → hd

FG → af HA → gc

…

**Circular Shift**

**Phrasebook $\pi_2$**

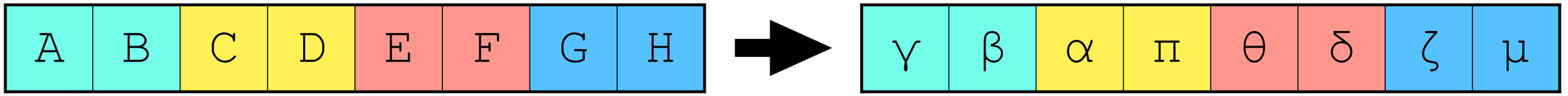
bh → γβ da → απ

fg → θδ ce → ζμ

…

# Multi-Layer Translation (MLT)

- Task: Fix $d$ phrasebooks $(\pi_1^*, \pi_2^*, \ldots, \pi_d^*)$, learn the translation process defined by the phrasebooks via (input, output) pairs **without intermediate steps.**




**SFT**

**Prompt:** Your are translating a sequence from language 1 to language 3. The input in language 1 is “A B C D E F G H”.

**Answer:** The output in language 3 is “γ β α π θ δ ζ μ”.

# Learning MLT from Input-Output is Provably Hard

- The family of $d$-depth MLT has SQ-dimension $e^{\Omega(d)}$ under uniformly random inputs.
- **Vanilla SFT (without intermediate steps) require $e^{\Omega(d)}$ samples.**

**SFT**

**Prompt:** Your are translating a sequence from language 1 to language 3. The input in language 1 is “A B C D E F G H”.

**Answer:** The output in language 3 is “γ β α π θ δ ζ μ”.

# Context-Enhanced Learning for MLT

**CEL**

**Prompt:** Your are translating a sequence from language 1 to language 3.

**Use phrasebook 1: B C → e b, F G → a f, D E → h d, H A → g c and phrasebook 2: b h → γ β, f g → θ δ, d a → α π, c e → ζ μ.**

The input in language 1 is "**A B C D E F G H**".

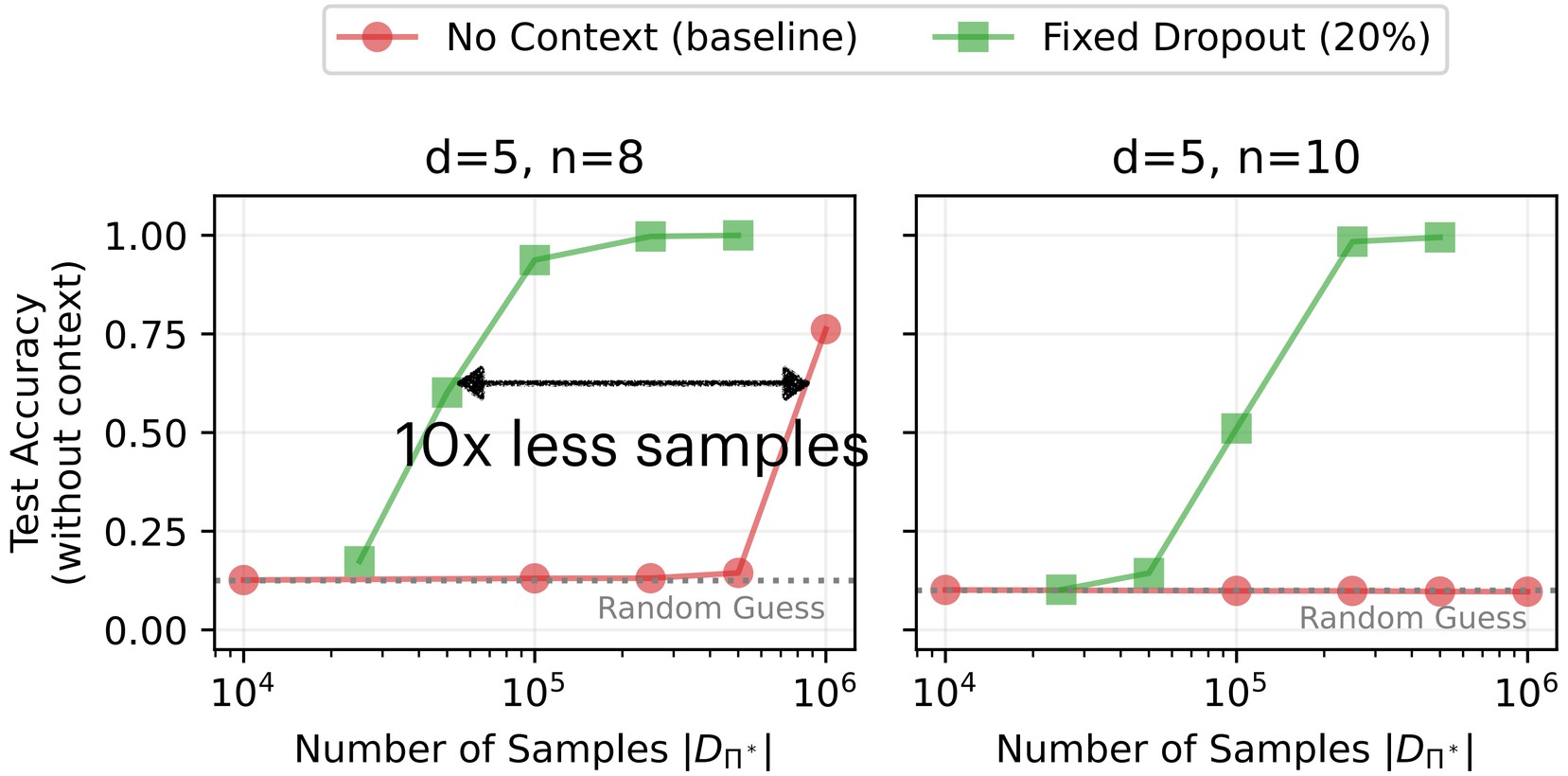
**Answer:** The output in language 3 is "**γ β α π θ δ ζ μ**".

**Two-stage training**:

- Stage 1: Teach the model how to read phrasebook rules in context.
- Stage 2: Provide useful phrasebooks in context and apply random dropout.

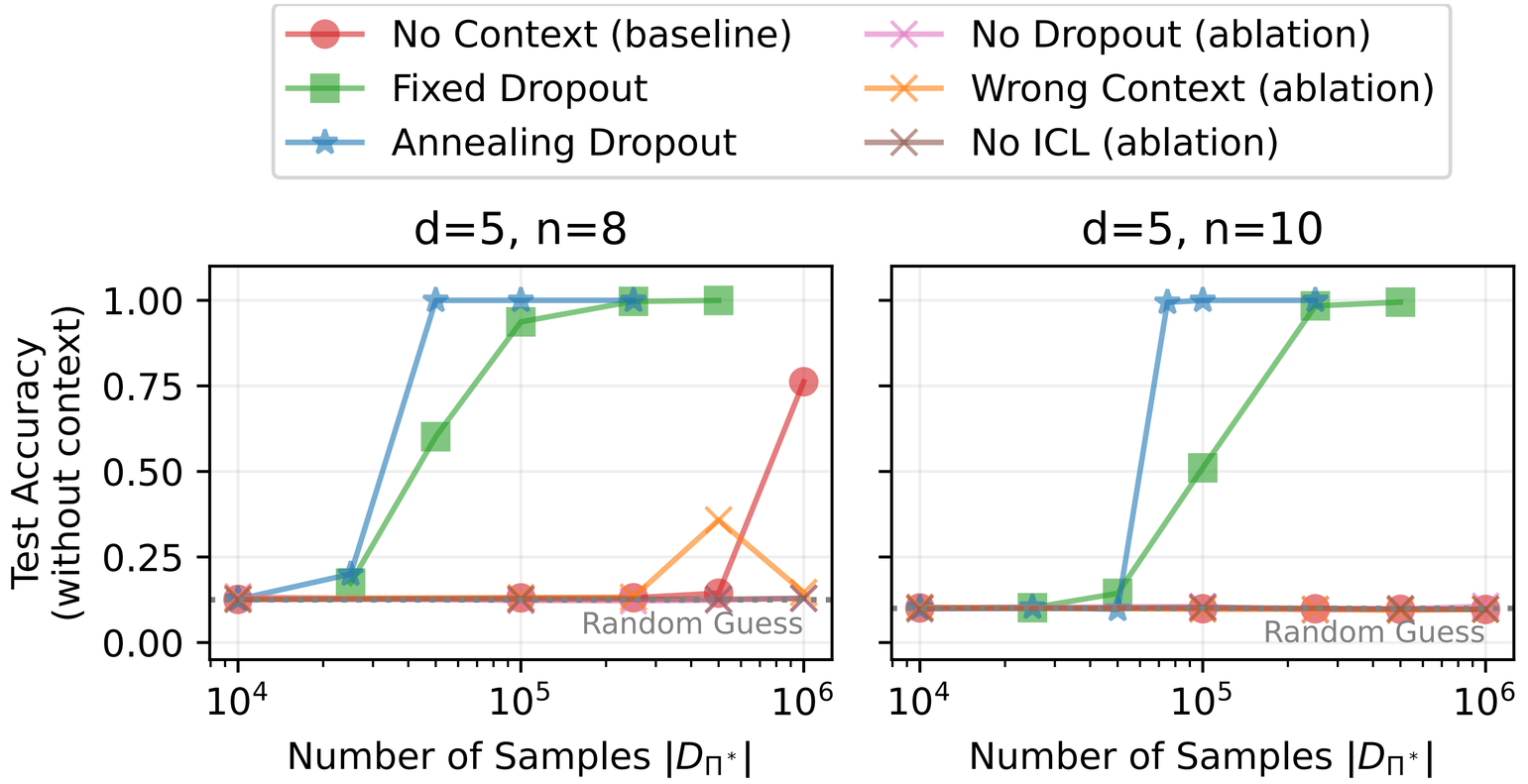
# Sample Efficiency of CEL on Learning MLT

- Experiment on Llama 3.2 3B, d = 5, with 8 or 10 characters per alphabet.



# Sample Efficiency of CEL on Learning MLT

- Experiment on Llama 3.2 3B, d = 5, with 8 or 10 characters per vocab.




- Better dropout curriculum leads for even faster learning.
- Dropout is required during training.
- ICL capability is required before training.

# Sample Complexity Separation for Toy Model

**Lower Bound: SFT**

There exists an $d$-step MLT instance such that any gradient-based training algorithm with uniformly random input-output pairs requires at least $e^{\Omega(d)}$ gradient updates.

**Upper Bound: CEL**

There exists an layer-wise search algorithm*, accompanied with a dropout curriculum on the context information, that can perfectly learn any $d$-step MLT task with $\mathcal{O}(d \log d)$ samples.

* Layer-wise GD results provable for $d = 2$, full parameter GD results empirically verified.

# Stop by our poster session @ 11:00AM - 1:30PM, July 15!

We thank Yun Cheng, Simon Park, Tianyu Gao, Yihe Dong, Zixuan Wang, Haoyu Zhao, and Bingbin Liu for discussions, suggestions, and proof-reading at various stages of the work.

# References


- Gao, Tianyu, et al. "Metadata Conditioning Accelerates Language Model Pre-training.
- Liao, Huanxuan, et al. "SKIntern: Internalizing Symbolic Knowledge for Distilling Better CoT Capabilities into Small Language Models."
- Choi, Younwoo, et al. "Teaching LLMs How to Learn with Contextual Fine-Tuning.”
- Deng, Y., Choi, Y., and Shieber, S. From explicit cot to implicit cot: Learning to internalize cot step by step. (2024)