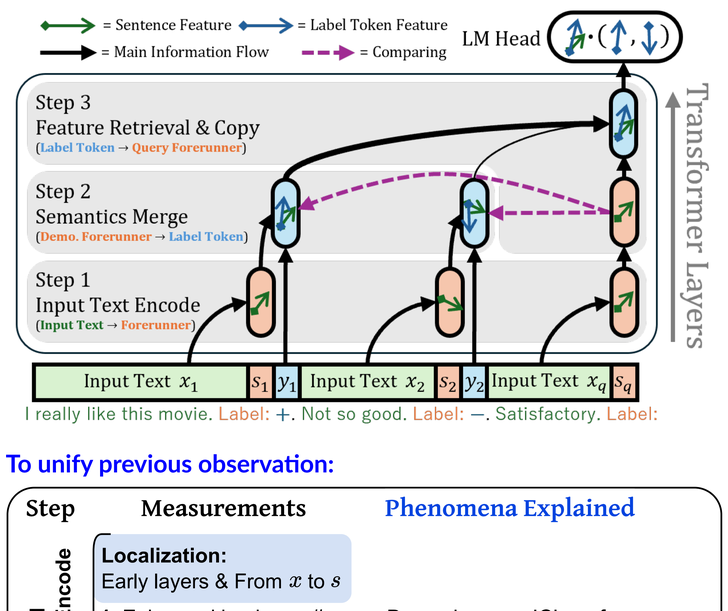
Revisiting In-context Learning Inference Circuit in Large Language Models

Hakaze Cho[1] Mariko Kato[1] Yoshihiro Sakai[1] Naoya Inoue[1,2] yfzhao@jaist.ac.jp
[1]Japan Advanced Institute of Science and Technology [2]RIKEN

Background

Some scattered phenomena have been observed, but no unified explanation.

Abstract



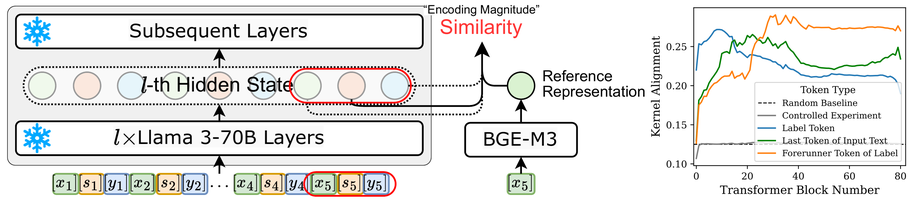
To unify previous observation:

Step	Measurements	Phenomena Explained
Step 1 Input Text Encode	**Localization:** Early layers & From x to s	
	1. Enhanced by demo. #	→ Demo. improve ICL performance
	2. Easier sentence = better	→ PPL-ICL[Gonen+23], ICCL[Liu+24]
	3. Biased to position	→ Location Bias[Zhao+21]
	4. Informative for task	→ Hidden Calibration[Cho+25]
Step 2 Semantics Merge	**Localization:** Later step 1 & From s to y	
	Label noise correction are conducted by semantic arithmetic	→ ICL are Robust against Label Noise[Min+22]; → Larger LM are Less Robust[Wei+23]
Step 3 Feature Ret. & Copy	**Localization:** Later step 2 & From y to y	
	1. Correct in Minority	→ Frequency Bias[Zhao+21]
	2. Task-specific Attn. Heads	
	3. Attn. Assignment Saturate	→ Demo. Saturation[Agarwal+24]

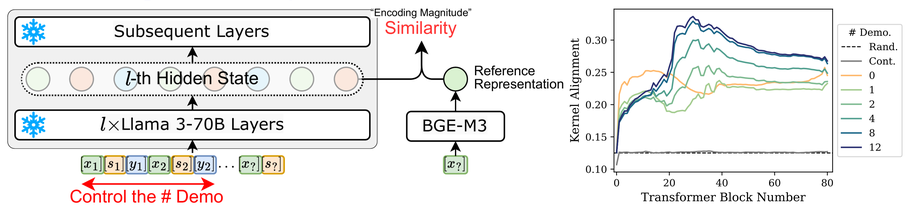
This work is supported by The Nakajima Foundation Start-Up Support Grant. The authors would like to thank Mr. Lihao Liu at the Beijing Institute of Technology, and Ms. Yunpin Li at the Capital Normal University for proofreading.

Step 1: Input Text Encode

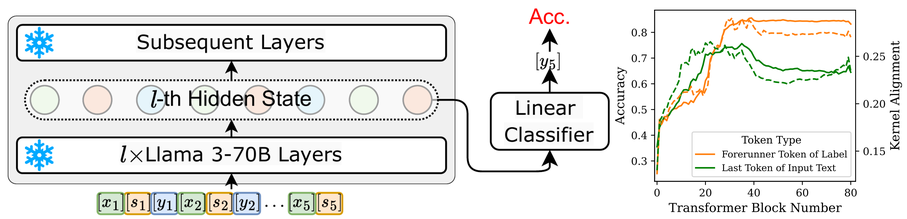
1.1. Localization: Which Token and Layers? → Early layers & From x to s



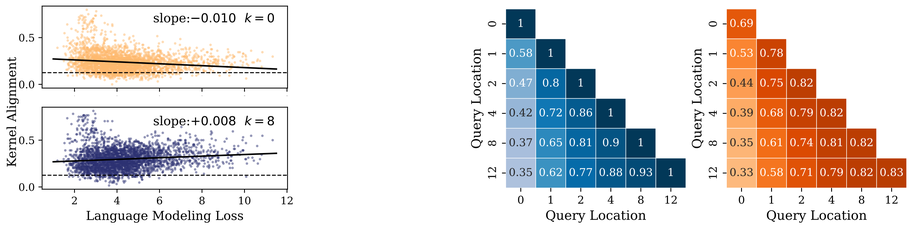
1.2. Encoding Enhanced by Demonstrations



1.3. Encoding is Informative to Downstream Task



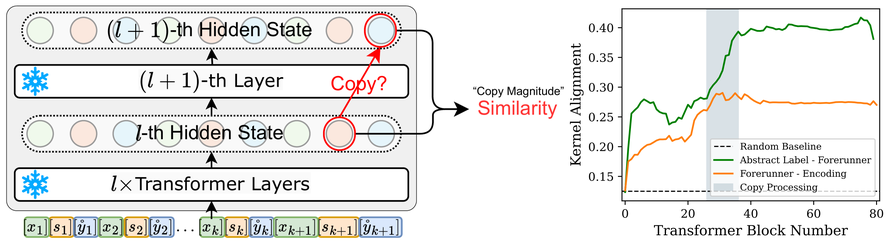
1.4. Encoding Weakened by CLM Loss; More Demonstrations Restore Them



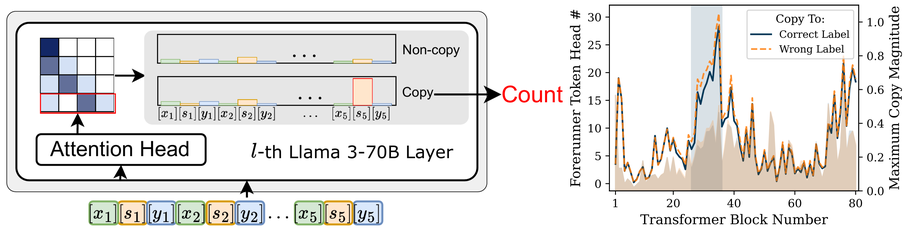
1.5. Encoding is Biased towards Demonstration Index ↑

Step 2: Semantics Merge

2.1. Localization: Which Layers? → Later than Step 1 & From s to y

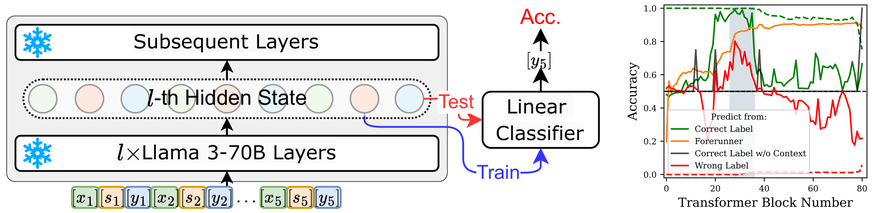


2.2. Why Robust to Noise? Consistent Copy Magnitude on Correct / Wrong y



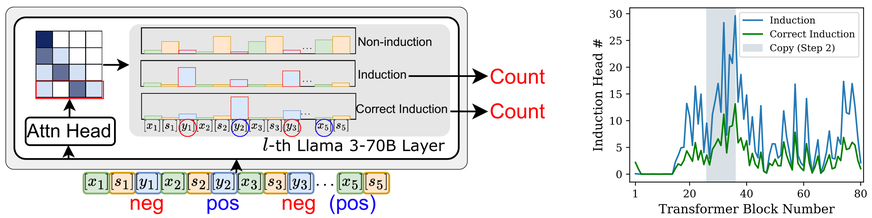
(*Continue* Step 2: Semantics Merge)

2.3. Why Robust to Noise? Feature Enhance / Neutralize on Correct / Wrong y



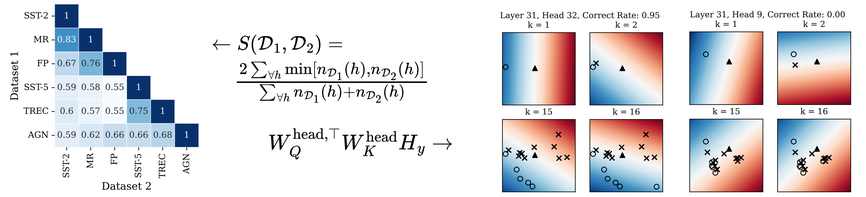
Step 3: Feature Retrieval and Copy (Induction Heads)

3.1. Localization: Which Layers? → Later than Step 2 & From y to s



3.2. Correct Retrieval is in Minority ↑

3.3. Some Induction Heads are Intrinsic; Some are Evoked by Task



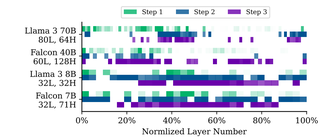
Ablation Study

The proposed 3 steps are essential

#	Attention Disconnected Key → Query	Affected Layers Ratio (from layer 1) 25%	50%	75%	100%
1	**None** (4-shot baseline)	±0 (Acc. 68.55)			
	– Step1: Input Text Encode –				
2	**Demo. Texts** x_i → **Forerunner** s_i	−4.98 −0.89 ± 0.00	−15.82 −1.19 ± 0.02	−23.43 −3.29 ± 1.87	−30.60 −1.61 ± 0.01
3	**Query Texts** x_q → **Forerunner** s_q	−13.87 −0.16 ± 0.00	−21.10 −0.08 ± 0.00	−24.74 −0.47 ± 0.04	−28.38 −0.55 ± 0.00
	– Step2: Semantics Merge –				
4	**Demo. Forerunner** s_i → **Label** y_i	−2.24 −0.00 ± 0.00	−3.45 −0.18 ± 0.00	−3.39 −0.10 ± 0.04	−3.42 −0.18 ± 0.01
	– Step3: Feature Retrieval & Copy –				
5	**Label** y_i → **Query Forerunner** s_q	−5.14 +0.03 ± 0.00	−10.03 −0.08 ± 0.00	−11.36 +0.00 ± 0.00	−10.22 −0.06 ± 0.00
	Reference Value				
6	**Zero-shot**	−17.90 (Acc. 50.65)			
7	**Random Prediction**	−36.05 (Acc. 32.50)			

Bypass Mechanism

Parallel Circuits



Shortcut Induction($s \rightarrow s$)

Attention Disconnected (Vi)	25%	50%	75%	100%
Forerunner s_{i-1} → **Forerunner** s_{eq}	−1.30 −0.00	−0.75 −0.16	−0.78 −0.16	−1.56 −0.71

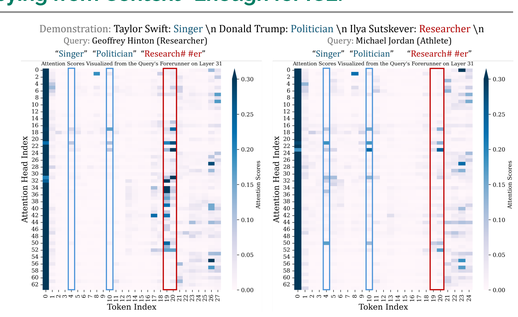
Direct Decoding (early layers → LM head)

Limitations: Is "Copying from Context" Enough for ICL?

Induction-based ICL Explanation

Taylor Swift → Singer
Donald Trump → Politician
Ian Goodfellow → Researcher
Geoffrey Hinton → Researcher
1. Search Similar Representations
2. Copy the Co-responding Label

Issue: Generalization to Unseen Labels
Taylor Swift → Singer
Donald Trump → Politician
Geoffrey Hinton → Researcher



[Gonen+23] Demystifying prompts in language models via perplexity estimation.
[Liu+24] Let's learn step by step: Enhancing in-context learning ability with curriculum learning.
[Zhao+21] Calibrate before use: Improving few-shot performance of language models.
[Cho+25] Token-based Decision Criteria Are Suboptimal in In-context Learning.
[Min+22] Rethinking the role of demonstrations: What makes in-context learning work?
[Wei+23] Larger language models do in-context learning differently.
[Agarwal+23] Many-shot in-context learning.