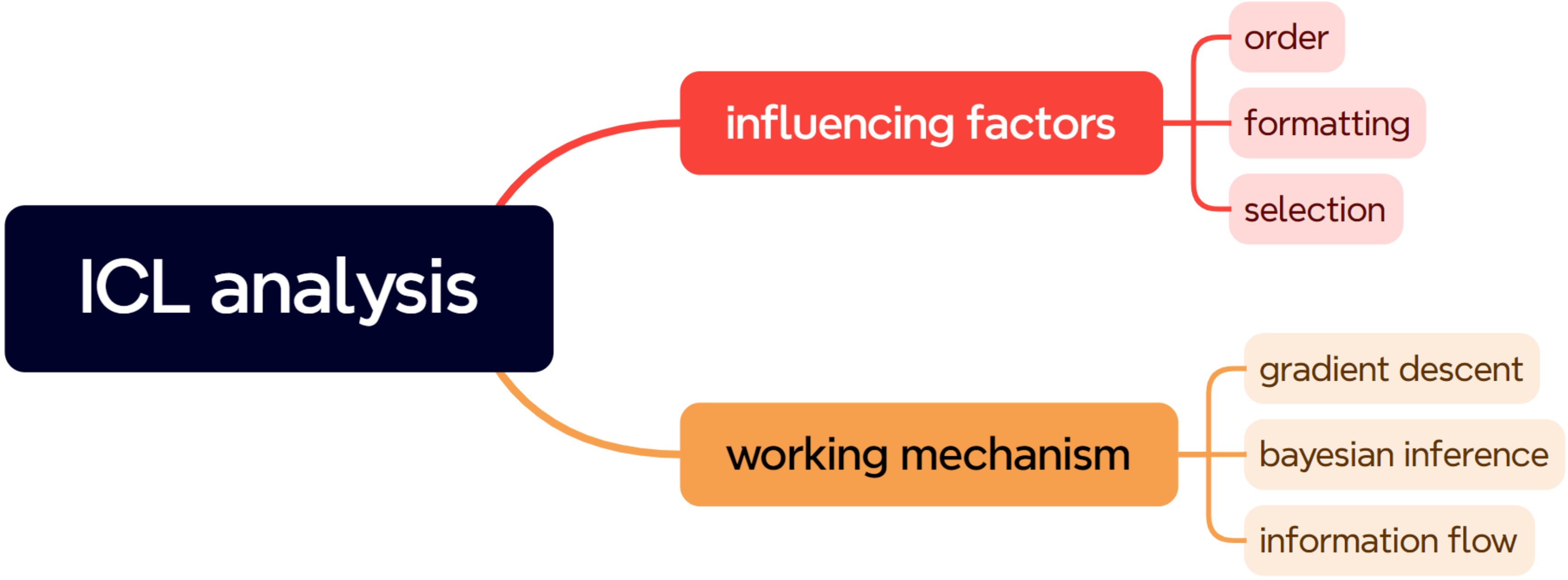
Label Words are Anchors: An Information Flow Perspective for Understanding In-Context Learning

Lean Wang[†,§], Lei Li[†], Damai Dai[†], Deli Chen[§],
Hao Zhou[§], Fandong Meng[§], Jie Zhou[§], Xu Sun[†]
[†]National Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University
[§]Pattern Recognition Center, WeChat AI, Tencent Inc., China
{lean,daidamai,xusun}@pku.edu.cn nlp.lilei@gmail.com
victorchen@deepseek.com {tuxzhou,fandongmeng,withtomzhou}@tencent.com

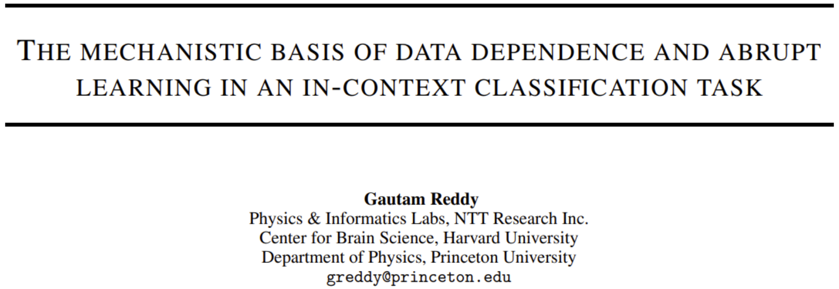
Relate Work



ICL analysis

- influencing factors
  - order
  - formatting
  - selection
- working mechanism
  - gradient descent
  - bayesian inference
  - information flow

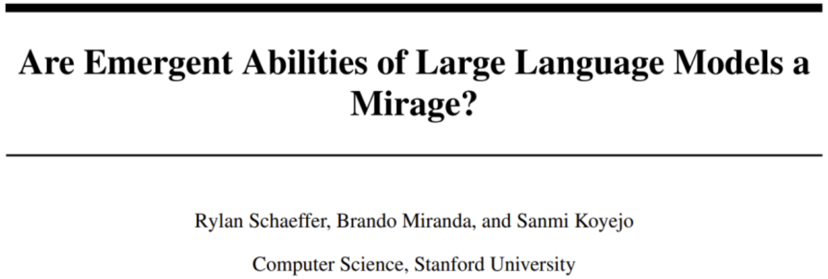
ICLR2024 score:5,8,10,10



THE MECHANISTIC BASIS OF DATA DEPENDENCE AND ABRUPT LEARNING IN AN IN-CONTEXT CLASSIFICATION TASK

Gautam Reddy
Physics & Informatics Labs, NTT Research Inc.
Center for Brain Science, Harvard University
Department of Physics, Princeton University
greddy@princeton.edu

NIPS 2023 Outstanding Main Track Papers

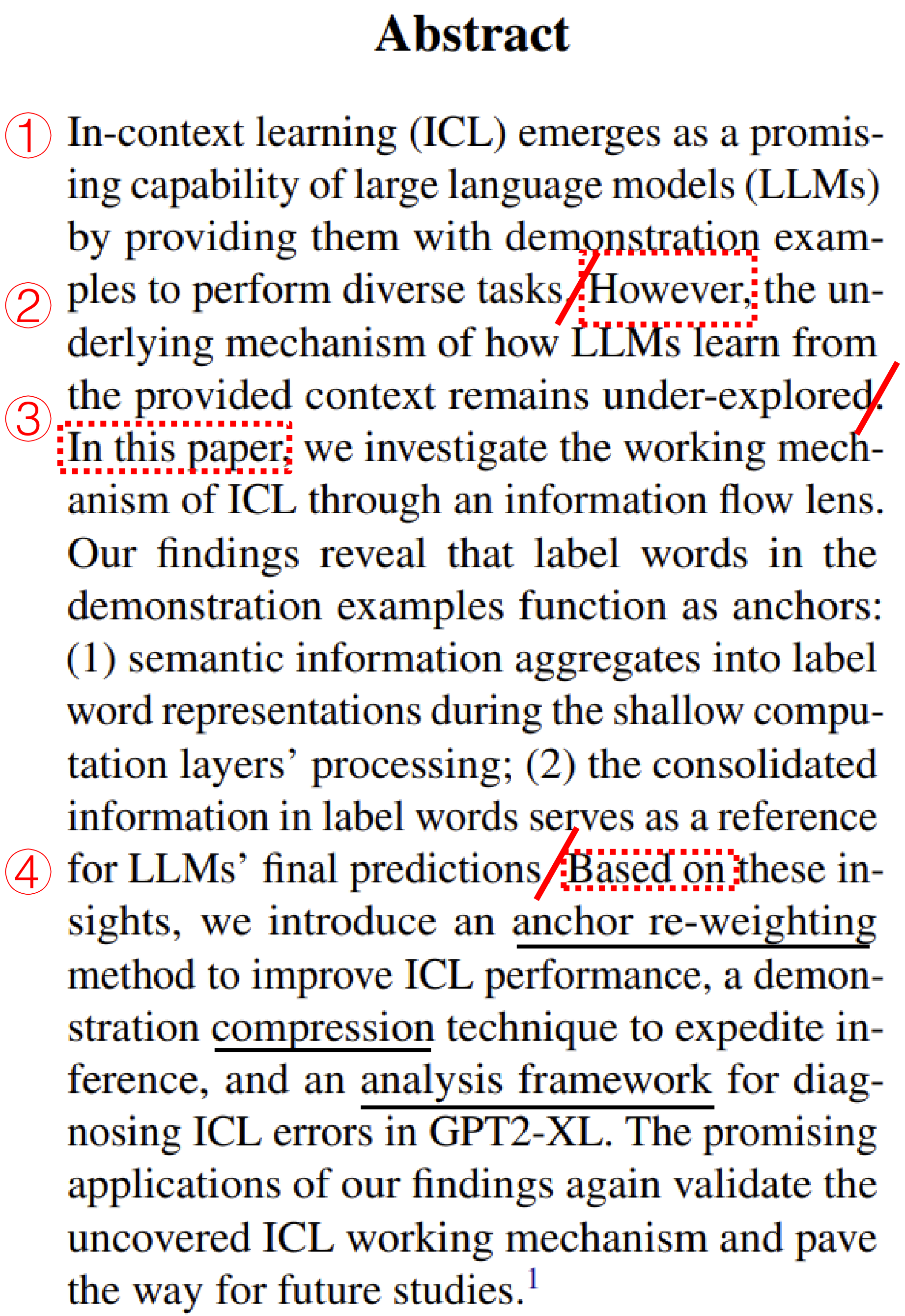


Are Emergent Abilities of Large Language Models a Mirage?

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo

Computer Science, Stanford University

Abstract:



Abstract

① In-context learning (ICL) emerges as a promising capability of large language models (LLMs) by providing them with demonstration examples to perform diverse tasks. ② However, the underlying mechanism of how LLMs learn from the provided context remains under-explored. ③ In this paper, we investigate the working mechanism of ICL through an information flow lens. Our findings reveal that label words in the demonstration examples function as anchors: (1) semantic information aggregates into label word representations during the shallow computation layers' processing; (2) the consolidated information in label words serves as a reference for LLMs' final predictions. ④ Based on these insights, we introduce an anchor re-weighting method to improve ICL performance, a demonstration compression technique to expedite inference, and an analysis framework for diagnosing ICL errors in GPT2-XL. The promising applications of our findings again validate the uncovered ICL working mechanism and pave the way for future studies.[1]

Motivation (Introduction)

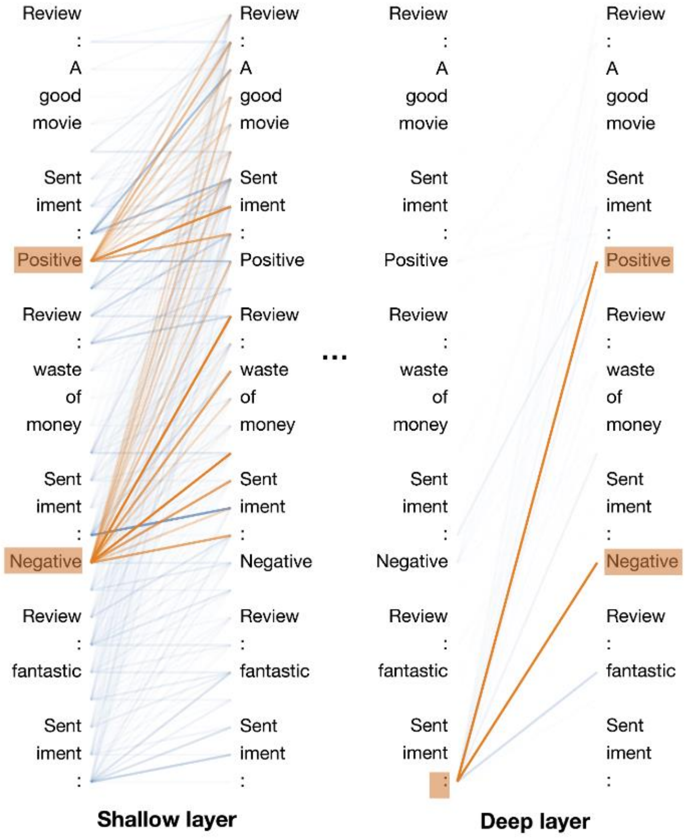
visualize the attention interactive pattern

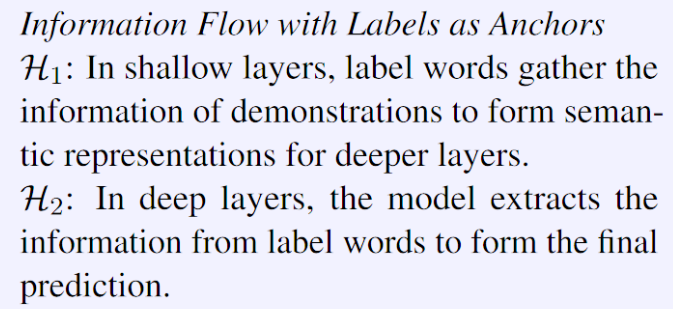
initial observation

three metrics to propose 2 hypothesis

two experiments to validate the hypothesis

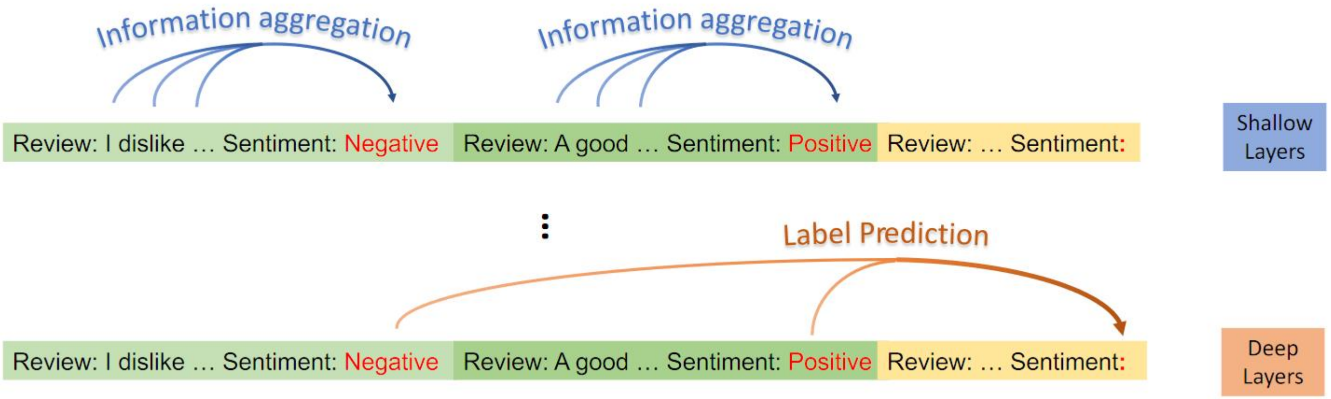
three approaches to enhance
ICL's effectiveness, efficiency, and interpretability





Information Flow with Labels as Anchors
$\mathcal{H}_1$: In shallow layers, label words gather the information of demonstrations to form semantic representations for deeper layers.
$\mathcal{H}_2$: In deep layers, the model extracts the information from label words to form the final prediction.

Hypothesis Motivated by Saliency Scores

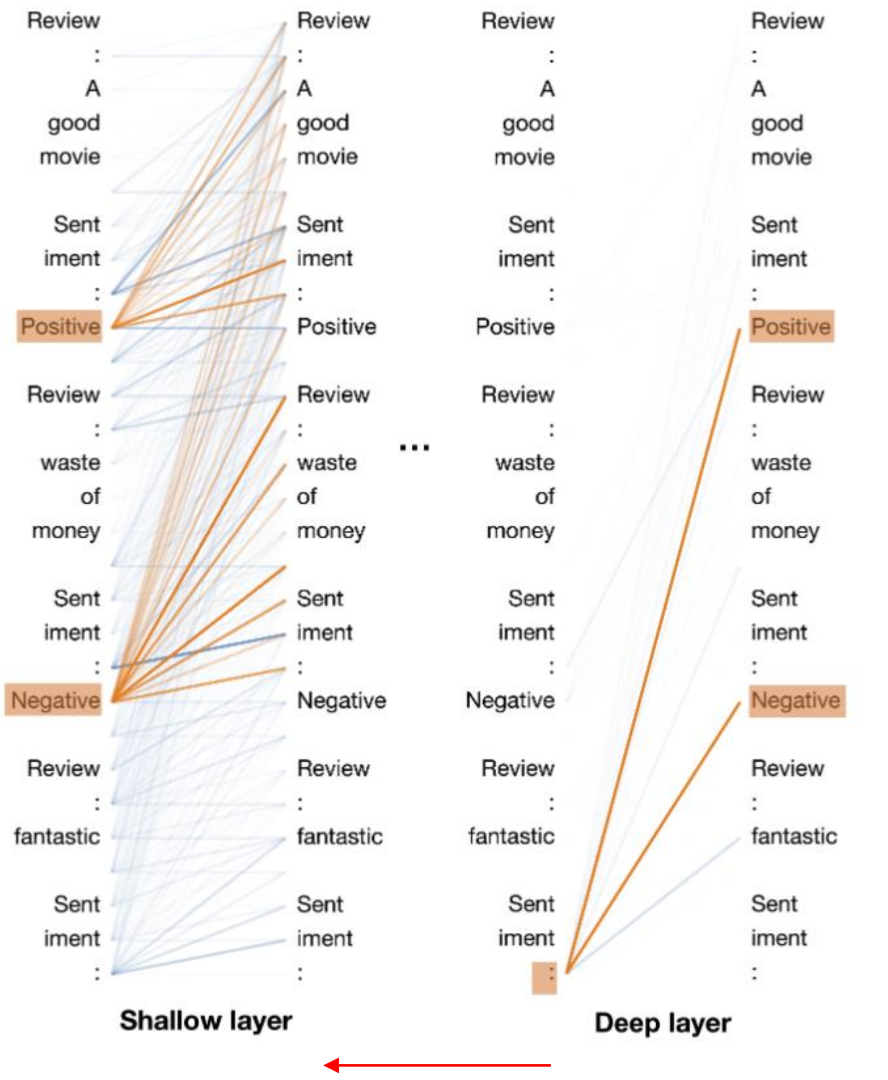


Information aggregation
Information aggregation
Review: I dislike … Sentiment: Negative
Review: A good … Sentiment: Positive
Review: … Sentiment:
Shallow Layers
⋮
Label Prediction
Review: I dislike … Sentiment: Negative
Review: A good … Sentiment: Positive
Review: … Sentiment:
Deep Layers

Figure 2: Illustration of our hypothesis. In shallow layers, label words gather information from demonstrations to form semantic representations for deeper processing, while deep layers extract and utilize this information from label words to formulate the final prediction.

$$I_l = \sum_h \left| A_{h,l}^{\top} \frac{\partial \mathcal{L}(x)}{\partial A_{h,l}} \right|$$

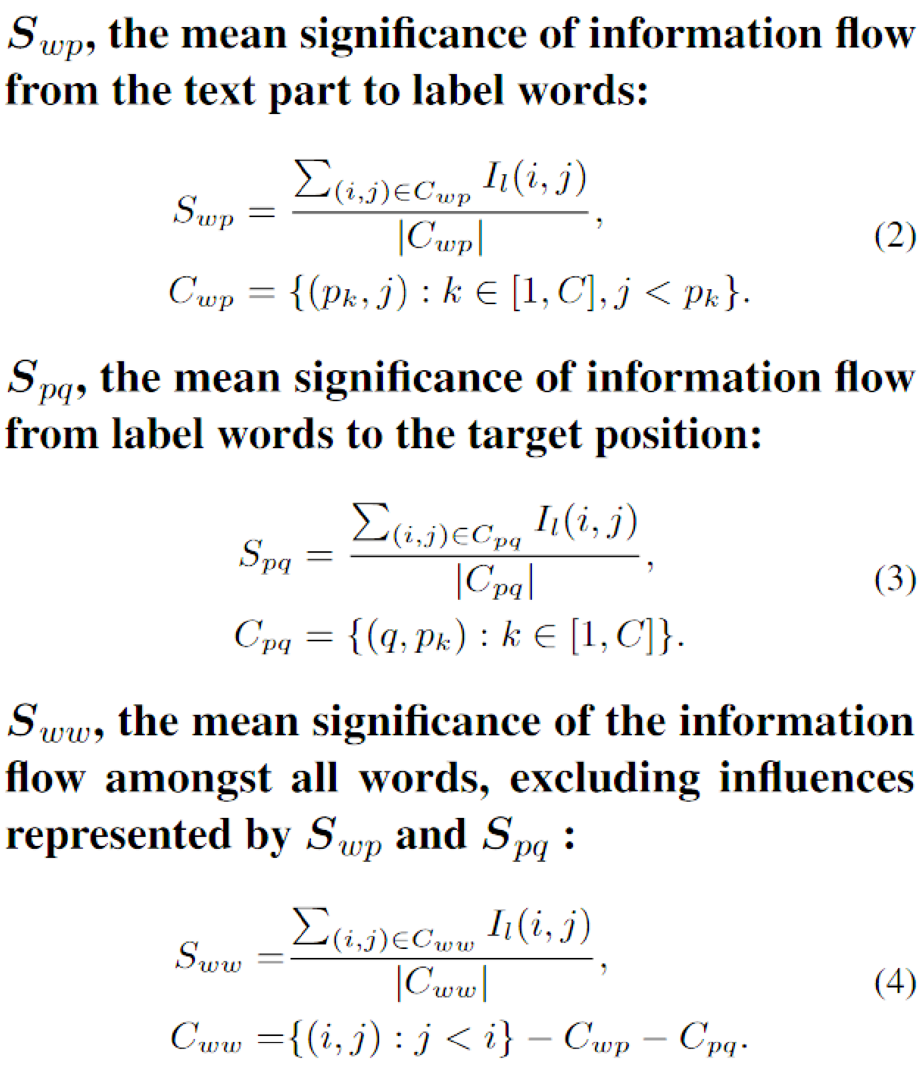
- $A_{h,l}$: value of the attention matrix of the h-th attention head in the l-th layer
- x: input   L(x): loss function(cross-entropy)
- = one stage Taylor expansion
- Why absolute value?
- Why saliency scores **not** attention score?



Review : A good movie Sent iment : Positive Review : waste of money Sent iment : Negative Review : fantastic Sent iment :
…
Shallow layer
Deep layer

Figure 1: Visualization of the information flow in a GPT model performing ICL. The line depth reflects the significance of the information flow from the right word to the left. The flows involving label words are highlighted. Label words gather information from demonstrations in shallow layers, which is then extracted in deep layers for final prediction.

To draw a clearer picture of this phenomenon:



S_{wp}, the mean significance of information flow from the text part to label words:

$$S_{wp} = \frac{\sum_{(i,j)\in C_{wp}} I_l(i,j)}{|C_{wp}|}, \qquad (2)$$
$$C_{wp} = \{(p_k, j) : k \in [1, C], j < p_k\}.$$

S_{pq}, the mean significance of information flow from label words to the target position:

$$S_{pq} = \frac{\sum_{(i,j)\in C_{pq}} I_l(i,j)}{|C_{pq}|}, \qquad (3)$$
$$C_{pq} = \{(q, p_k) : k \in [1, C]\}.$$

S_{ww}, the mean significance of the information flow amongst all words, excluding influences represented by S_{wp} and S_{pq} :

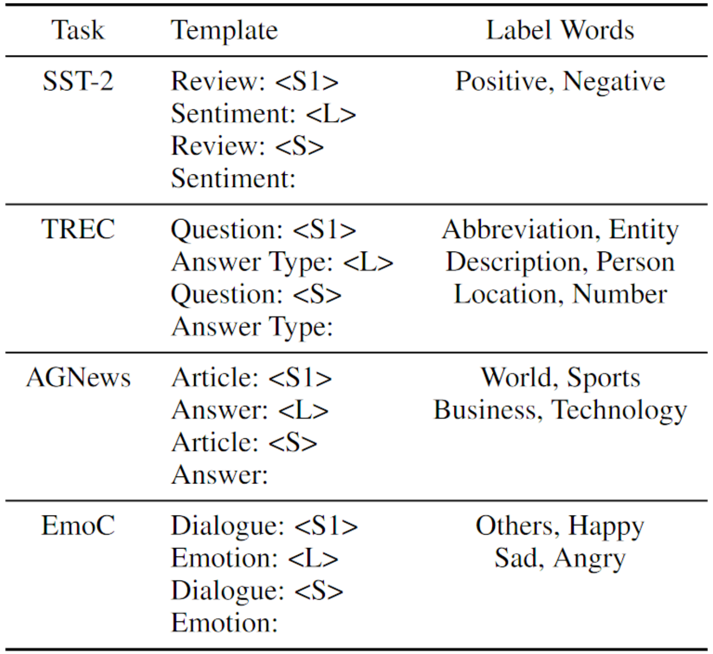
$$S_{ww} = \frac{\sum_{(i,j)\in C_{ww}} I_l(i,j)}{|C_{ww}|}, \qquad (4)$$
$$C_{ww} = \{(i,j) : j < i\} - C_{wp} - C_{pq}.$$

- p_i: label word
- q: target position (final token in the input)
- w: text part before label words in the demonstration

- S_{wp}: intensity of information aggregation onto label words
- S_{pq}: information extraction from label words for final decision-making
- S_{ww}: benchmark intensity

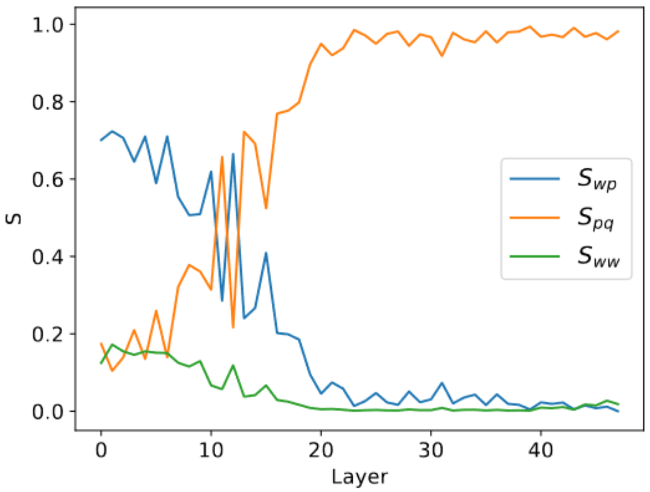
- GPT2-XL(1.5B) and GPT-J(6B)

- Text classification:

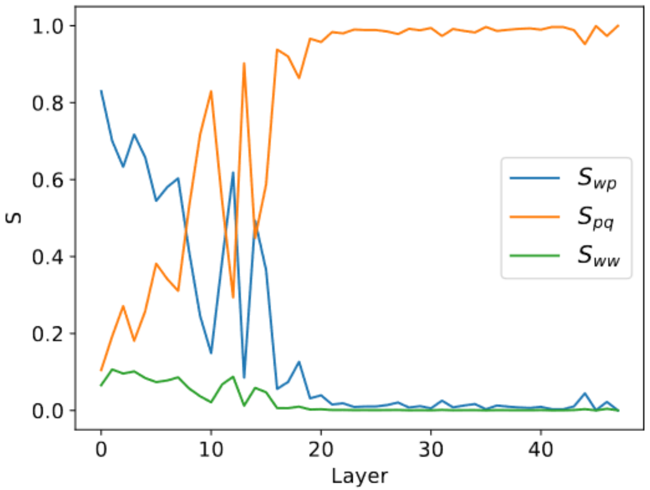


Task	Template	Label Words
SST-2	Review: <S1> Sentiment: <L> Review: <S> Sentiment:	Positive, Negative
TREC	Question: <S1> Answer Type: <L> Question: <S> Answer Type:	Abbreviation, Entity Description, Person Location, Number
AGNews	Article: <S1> Answer: <L> Article: <S> Answer:	World, Sports Business, Technology
EmoC	Dialogue: <S1> Emotion: <L> Dialogue: <S> Emotion:	Others, Happy Sad, Angry

Results:



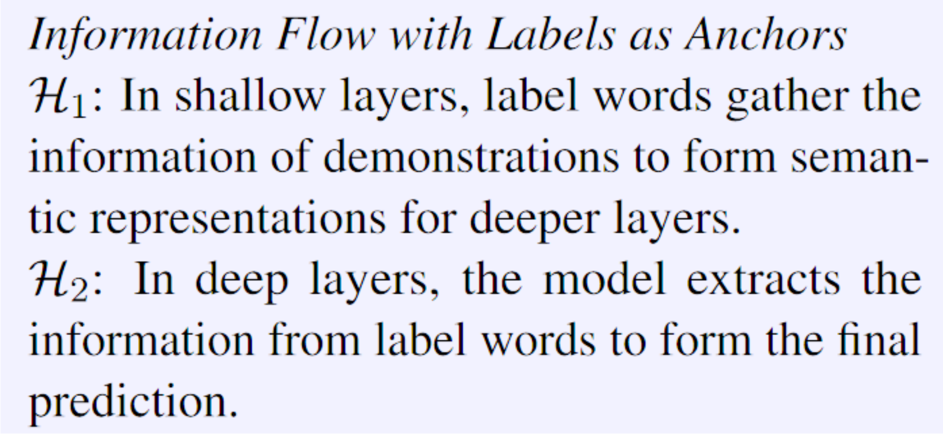
(a) Results on the SST-2 dataset



(b) Results on the AGNews dataset

- In shallow layer …
- In deeper layer …
- usually surpass…

Hypothesis:



Information Flow with Labels as Anchors
$\mathcal{H}_1$: In shallow layers, label words gather the information of demonstrations to form semantic representations for deeper layers.
$\mathcal{H}_2$: In deep layers, the model extracts the information from label words to form the final prediction.

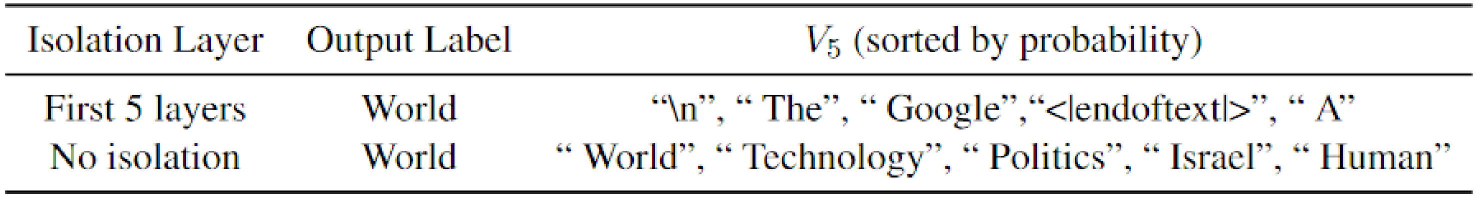
Shallow Layers: Information Aggregation :

Isolate label words by manipulating the attention matrix A to block the information flow to label words

$$A_l(p, i)(i < p) \text{ to } 0$$

Metrics

- **Label Loyalty**: consistency of output labels with and without isolation
- **Word Loyalty**: Jaccard similarity to compare the top-5 predicted words with and without isolation (why?)



Isolation Layer	Output Label	V_5 (sorted by probability)
First 5 layers	World	“\n”, “ The”, “ Google”,“<|endoftext|>”, “ A”
No isolation	World	“ World”, “ Technology”, “ Politics”, “ Israel”, “ Human”

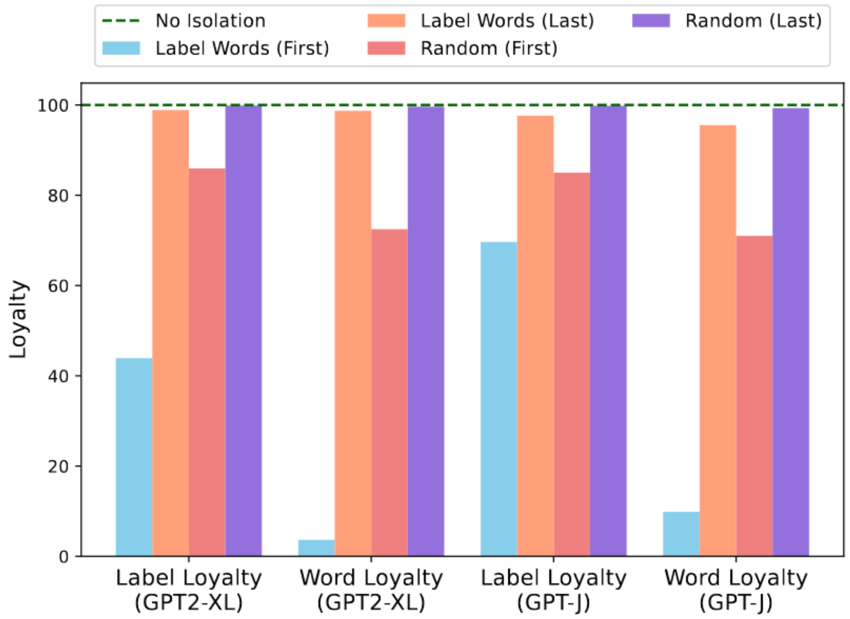


Figure 4: The impact of isolating label words versus randomly isolating non-label words within the first or last 5 layers. Isolating label words within the first 5 layers exerts the most substantial impact, highlighting the importance of shallow-layer information aggregation via label words.

More experiments:

- variable numbers of layers
- ICL with semantically unrelated labels

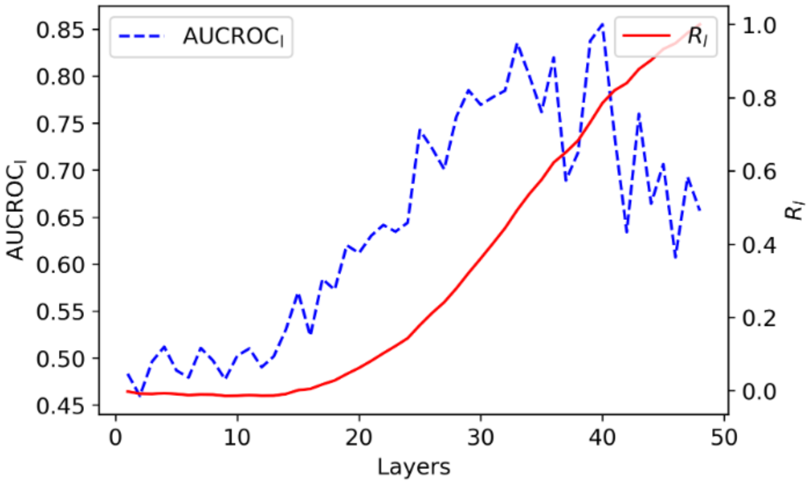
Deep Layers: Information Extraction

Utilize the **AUC-ROC** score to quantify correlation between $A_l(q, pi)$ and model prediction.

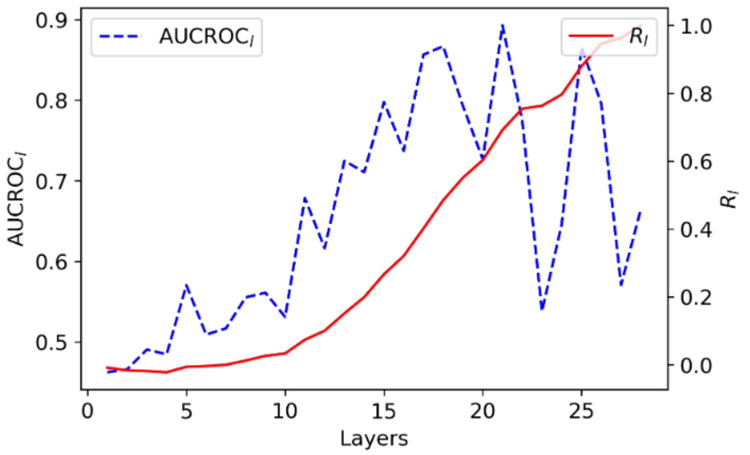
reason

- $A_l(q, pi)$ might differ from the probability of the model outputting label i by a constant factor
- reducing disturbances caused by class imbalance

$$R_l = \frac{\sum_{i=1}^{l}(\text{AUCROC}_i - 0.5)}{\sum_{i=1}^{N}(\text{AUCROC}_i - 0.5)}.$$



(a) GPT2-XL (total 48 layers).



(b) GPT-J (total 28 layers).

Analysis:

- deep layers approaches 0.8
- R significant increase in middle and deep layers

Applications:

- Effectiveness: Anchor Re-weighting
- Efficiency: Anchor-Only Context Compression
- Interpretability: Anchor Distances for Error Diagnosis

classification loss: $\boldsymbol{\beta}^{\star} = \arg\min_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{X}_{train}, \boldsymbol{Y}_{train})$.

1、Effectiveness: Anchor Re-weighting

We can view the attention module as a classifier $\boldsymbol{f}$,

$$\begin{aligned}&\Pr_{\boldsymbol{f}}(Y = i|X = x)\\ \approx& A(q, p_i)\\ =&\frac{\exp(\mathbf{q}_q \mathbf{k}_{p_i}^T/\sqrt{d})}{\sum_{j=1}^{N}\exp(\mathbf{q}_q \mathbf{k}_j^T/\sqrt{d})}.\end{aligned} \tag{6}$$

By setting $\mathbf{q}_q/\sqrt{d} = \hat{\mathbf{x}}$ and $\mathbf{k}_{p_i} - \mathbf{k}_{p_C} = \boldsymbol{\beta}_i$, we deduce:

$$\log\frac{\Pr_{\boldsymbol{f}}(Y = i|X = x)}{\Pr_{\boldsymbol{f}}(Y = C|X = x)} = \boldsymbol{\beta}_i^T \hat{\mathbf{x}}. \tag{7}$$

This approximates a logistic regression model where:

$$\log\frac{\Pr_{\boldsymbol{f}}(Y = i|X = x)}{\Pr_{\boldsymbol{f}}(Y = C|X = x)} = \beta_0^i + \boldsymbol{\beta}_i^T \mathbf{x}. \tag{8}$$

In this equation, β_0^i and $\boldsymbol{\beta}_i^T$ are parameters that can be learned, while $\mathbf{x}$ is the input feature.

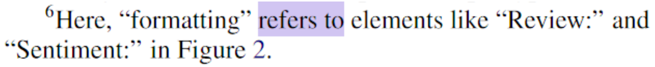
$$\boxed{\hat{A}(q, p_i) = \exp(\beta_0^i) A(q, p_i)}$$

$$\begin{aligned}&\text{Attention}_l^h(Q, K, V) = \hat{A}_l^h V,\\ &A_l^h = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right),\\ &\hat{A}_l^h(k, j) = \begin{cases}\exp(\beta_{0,lh}^i) A_l^h(k, j), & \text{if } k = q, j = p_i\\ A_l^h(k, j), & \text{otherwise}\end{cases}\end{aligned}$$

Results:

Method	SST-2	TREC	AGNews	EmoC	Average
Vanilla In-Context Learning (1-shot per class)	61.28	57.56	73.32	15.44	51.90
Vanilla In-Context Learning (5-shot per class)	64.75	60.40	52.52	9.80	46.87
Anchor Re-weighting (1-shot per class)	**90.07**	**60.92**	**81.94**	**41.64**	**68.64**

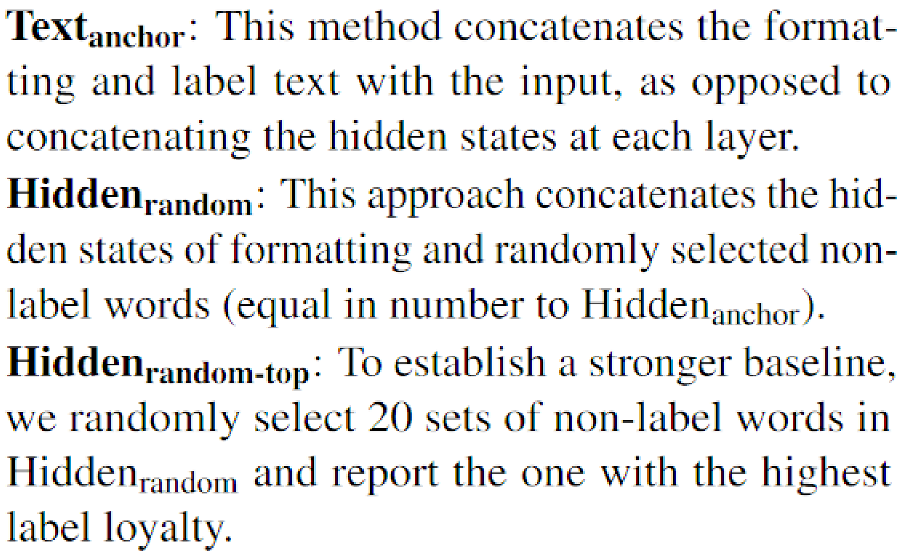
Table 1: The effect after adding parameter β_0^i. For AGNews, due to the length limit, we only use three demonstrations per class. Our Anchor Re-weighting method achieves the best performance overall tasks.



[6]Here, “formatting” refers to elements like “Review:” and “Sentiment:” in Figure 2.

2、 Efficiency: Anchor-Only Context Compression

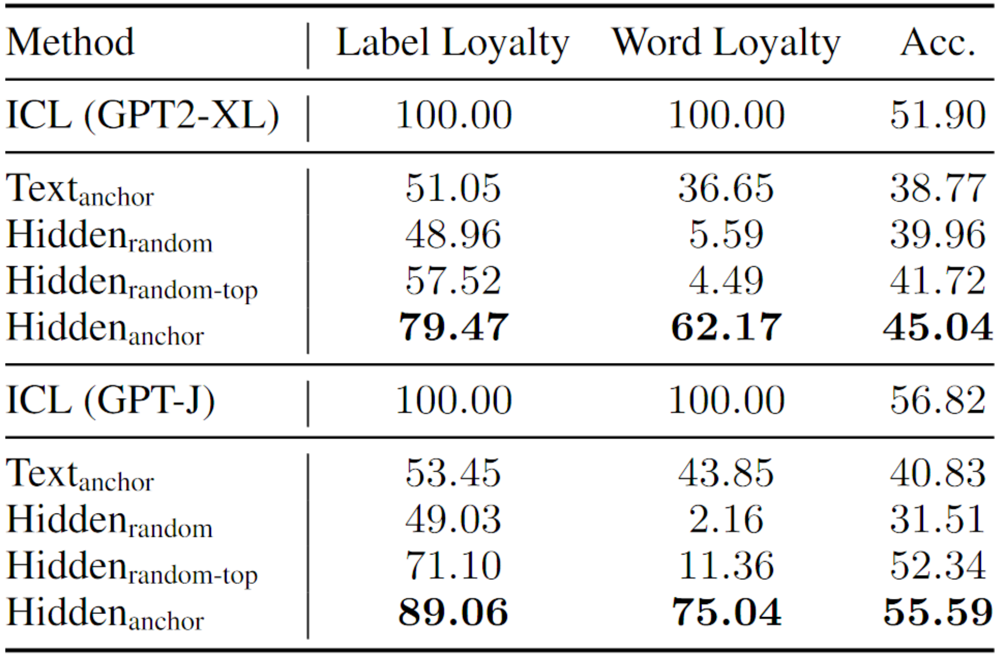
- concatenate hidden states $h_l^1 \ldots h_l^C$ at the front in each layer during inference ✗
- **Hidden$_{\text{anchor}}$**: amalgamate the hidden states of both the formatting and the label words



Text$_{\text{anchor}}$: This method concatenates the formatting and label text with the input, as opposed to concatenating the hidden states at each layer.

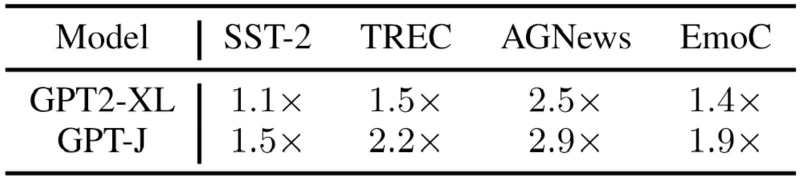
Hidden$_{\text{random}}$: This approach concatenates the hidden states of formatting and randomly selected non-label words (equal in number to Hidden$_{\text{anchor}}$).

Hidden$_{\text{random-top}}$: To establish a stronger baseline, we randomly select 20 sets of non-label words in Hidden$_{\text{random}}$ and report the one with the highest label loyalty.



Method	Label Loyalty	Word Loyalty	Acc.
ICL (GPT2-XL)	100.00	100.00	51.90
Text$_{\text{anchor}}$	51.05	36.65	38.77
Hidden$_{\text{random}}$	48.96	5.59	39.96
Hidden$_{\text{random-top}}$	57.52	4.49	41.72
Hidden$_{\text{anchor}}$	**79.47**	**62.17**	**45.04**
ICL (GPT-J)	100.00	100.00	56.82
Text$_{\text{anchor}}$	53.45	43.85	40.83
Hidden$_{\text{random}}$	49.03	2.16	31.51
Hidden$_{\text{random-top}}$	71.10	11.36	52.34
Hidden$_{\text{anchor}}$	**89.06**	**75.04**	**55.59**

Table 2: Results of different compression methods on GPT2-XL and GPT-J (averaged over SST-2, TREC, AGNews, and EmoC). Acc. denotes accuracy. The best results are shown in bold. Our method achieves the best compression performance.



Model	SST-2	TREC	AGNews	EmoC
GPT2-XL	1.1×	1.5×	2.5×	1.4×
GPT-J	1.5×	2.2×	2.9×	1.9×

Table 3: Acceleration ratios of the Hidden$_{\text{anchor}}$ method.

3、 Interpretability: Anchor Distances for Error Diagnosis

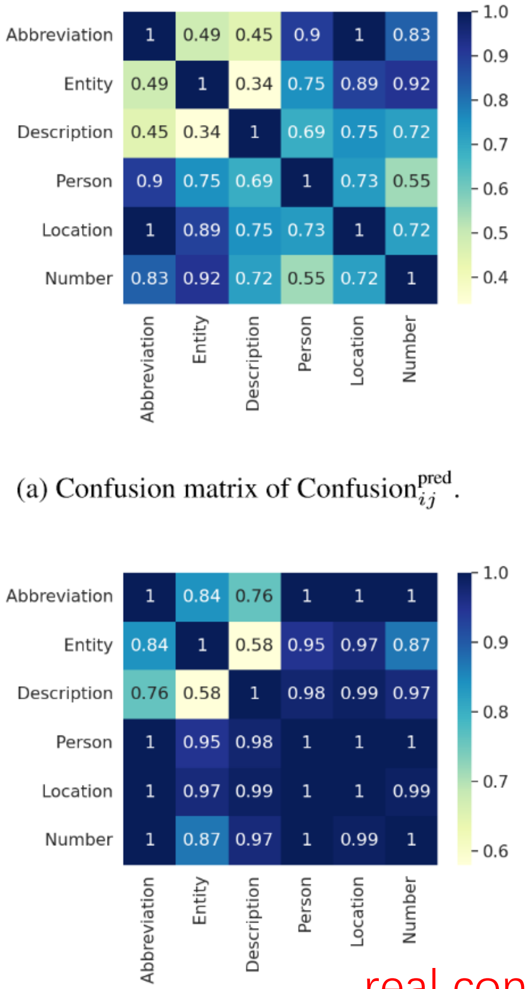
a strong correlation between the model output and $A_l(q, pi)$

$$A(q, p_i) = \frac{\exp(\mathbf{q}_q \mathbf{k}_{p_i}^T / \sqrt{d})}{\sum_{j=1}^{N} \exp(\mathbf{q}_q \mathbf{k}_j^T / \sqrt{d})}.$$

the key vectors k for label words p_i and p_k be similar, $A_l(q, pi)$ and $A_l(q, pk)$ will also likely be similar

$$\text{Confusion}_{ij}^{\text{pred}} = \frac{\|\hat{\mathbf{k}_{\mathbf{p_i}}} - \hat{\mathbf{k}_{\mathbf{p_j}}}\|}{\max_{s \neq t} \|\hat{\mathbf{k}_{\mathbf{p_s}}} - \hat{\mathbf{k}_{\mathbf{p_t}}}\|},$$

considering the distribution of query vectors q_q, employ a PCA-like method to extract the components of the key vectors along the directions with significant variations in q_q



	Abbreviation	Entity	Description	Person	Location	Number
Abbreviation	1	0.49	0.45	0.9	1	0.83
Entity	0.49	1	0.34	0.75	0.89	0.92
Description	0.45	0.34	1	0.69	0.75	0.72
Person	0.9	0.75	0.69	1	0.73	0.55
Location	1	0.89	0.75	0.73	1	0.72
Number	0.83	0.92	0.72	0.55	0.72	1

(a) Confusion matrix of $\text{Confusion}_{ij}^{\text{pred}}$.

	Abbreviation	Entity	Description	Person	Location	Number
Abbreviation	1	0.84	0.76	1	1	1
Entity	0.84	1	0.58	0.95	0.97	0.87
Description	0.76	0.58	1	0.98	0.99	0.97
Person	1	0.95	0.98	1	1	1
Location	1	0.97	0.99	1	1	0.99
Number	1	0.87	0.97	1	0.99	1

real confusion?

(b) Confusion matrix of Confusion_{ij}.

Figure 6: Predicted and real confusion matrix on TREC. We set undefined diagonals to 1 for better visualization. The heatmaps display similarity in confusing category pairs, particularly in lighter-colored blocks.

Conclusion:

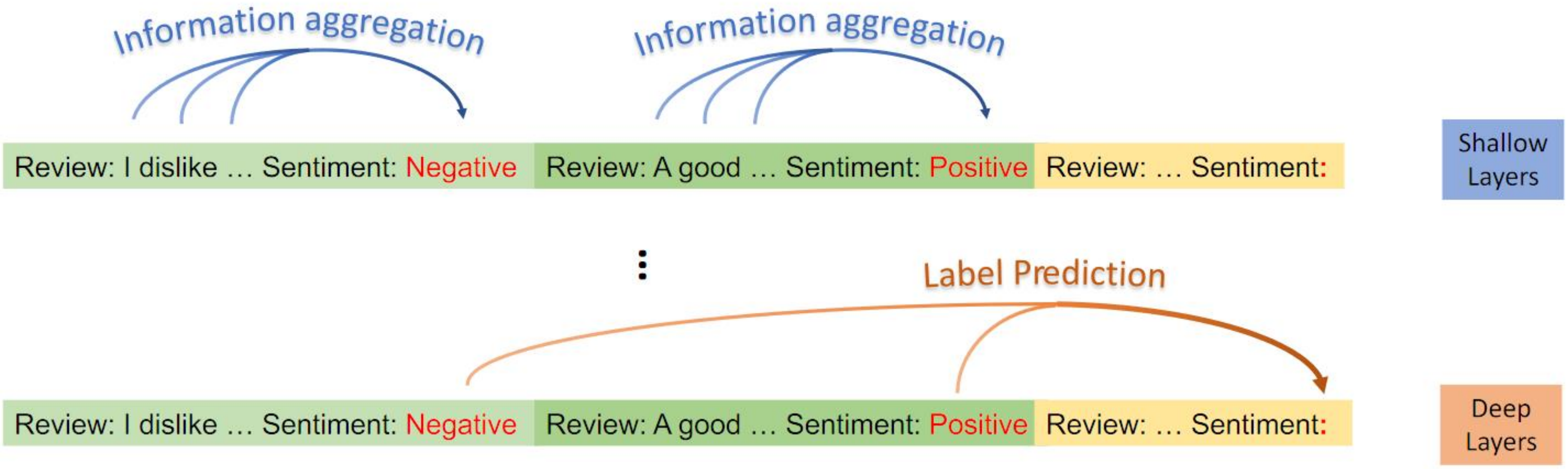

Information aggregation
Information aggregation
Review: I dislike … Sentiment: Negative
Review: A good … Sentiment: Positive
Review: … Sentiment:
Shallow Layers
Label Prediction
Review: I dislike … Sentiment: Negative
Review: A good … Sentiment: Positive
Review: … Sentiment:
Deep Layers

Figure 2: Illustration of our hypothesis. In shallow layers, label words gather information from demonstrations to form semantic representations for deeper processing, while deep layers extract and utilize this information from label words to formulate the final prediction.

- three metrics to propose 2 hypothesis
- two experiments to validate the hypothesis
- three approaches to enhance ICL's effectiveness, efficiency, and interpretability

validate anchor hypothesis and show the significance of anchors in ICL

Limitation:

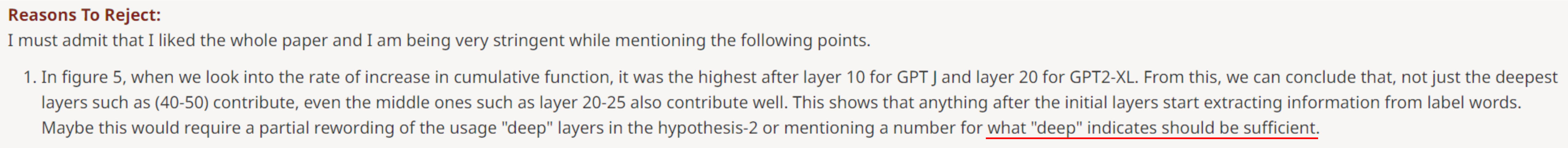
- task: classification
- other ICL paradigms (CoT)
- hardware constraints (**对计算资源受限的工作 容忍度提高了？)**

Question for you:

- **若你是审稿人，能够提出怎样的问题?**
  - **design for formatting**

- **Best paper 给你的启发?**
  - **分析、实验类的论文参考、分析问题的角度**
  - **大：文章的整体构思； 小：实验的设计与阐述**

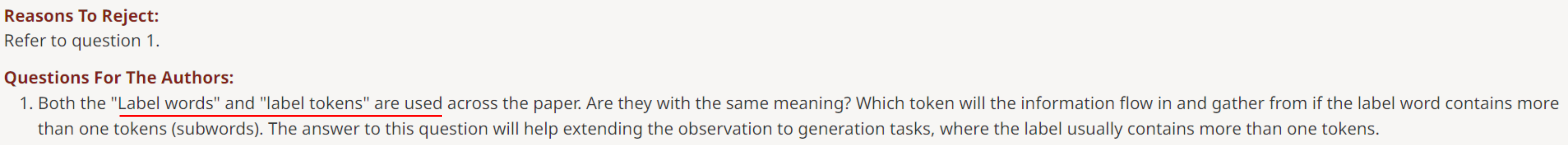
Reviewer:



Reasons To Reject:

I must admit that I liked the whole paper and I am being very stringent while mentioning the following points.

1. In figure 5, when we look into the rate of increase in cumulative function, it was the highest after layer 10 for GPT J and layer 20 for GPT2-XL. From this, we can conclude that, not just the deepest layers such as (40-50) contribute, even the middle ones such as layer 20-25 also contribute well. This shows that anything after the initial layers start extracting information from label words. Maybe this would require a partial rewording of the usage "deep" layers in the hypothesis-2 or mentioning a number for what "deep" indicates should be sufficient.

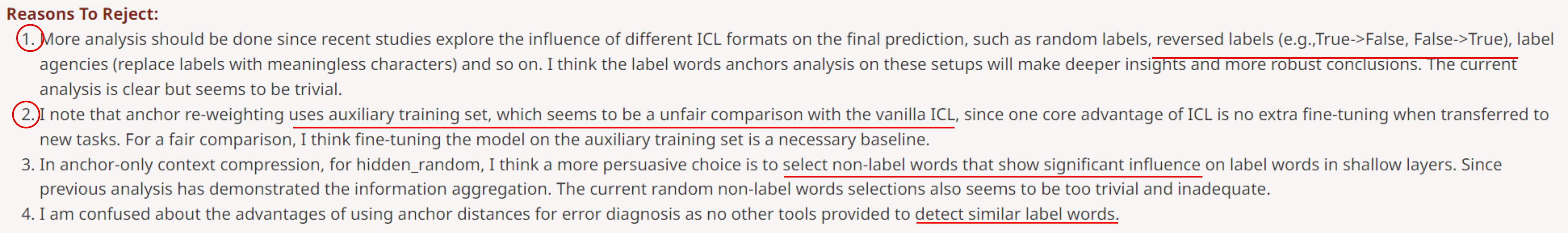


Reasons To Reject:

Refer to question 1.

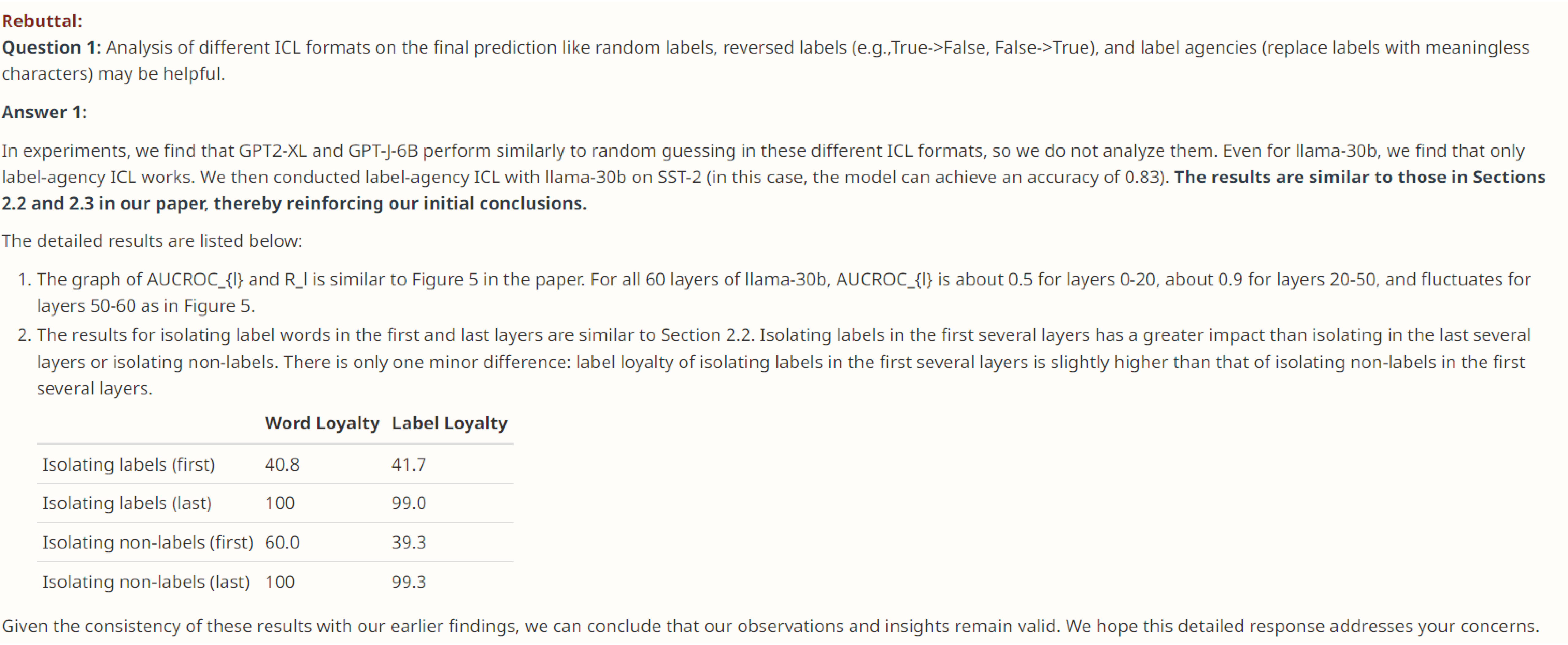
Questions For The Authors:

1. Both the "Label words" and "label tokens" are used across the paper. Are they with the same meaning? Which token will the information flow in and gather from if the label word contains more than one tokens (subwords). The answer to this question will help extending the observation to generation tasks, where the label usually contains more than one tokens.



Reasons To Reject:

1. More analysis should be done since recent studies explore the influence of different ICL formats on the final prediction, such as random labels, reversed labels (e.g.,True->False, False->True), label agencies (replace labels with meaningless characters) and so on. I think the label words anchors analysis on these setups will make deeper insights and more robust conclusions. The current analysis is clear but seems to be trivial.
2. I note that anchor re-weighting uses auxiliary training set, which seems to be a unfair comparison with the vanilla ICL, since one core advantage of ICL is no extra fine-tuning when transferred to new tasks. For a fair comparison, I think fine-tuning the model on the auxiliary training set is a necessary baseline.
3. In anchor-only context compression, for hidden_random, I think a more persuasive choice is to select non-label words that show significant influence on label words in shallow layers. Since previous analysis has demonstrated the information aggregation. The current random non-label words selections also seems to be too trivial and inadequate.
4. I am confused about the advantages of using anchor distances for error diagnosis as no other tools provided to detect similar label words.



Rebuttal:

Question 1: Analysis of different ICL formats on the final prediction like random labels, reversed labels (e.g.,True->False, False->True), and label agencies (replace labels with meaningless characters) may be helpful.

Answer 1:

In experiments, we find that GPT2-XL and GPT-J-6B perform similarly to random guessing in these different ICL formats, so we do not analyze them. Even for llama-30b, we find that only label-agency ICL works. We then conducted label-agency ICL with llama-30b on SST-2 (in this case, the model can achieve an accuracy of 0.83). **The results are similar to those in Sections 2.2 and 2.3 in our paper, thereby reinforcing our initial conclusions.**

The detailed results are listed below:

1. The graph of AUCROC_{l} and R_l is similar to Figure 5 in the paper. For all 60 layers of llama-30b, AUCROC_{l} is about 0.5 for layers 0-20, about 0.9 for layers 20-50, and fluctuates for layers 50-60 as in Figure 5.
2. The results for isolating label words in the first and last layers are similar to Section 2.2. Isolating labels in the first several layers has a greater impact than isolating in the last several layers or isolating non-labels. There is only one minor difference: label loyalty of isolating labels in the first several layers is slightly higher than that of isolating non-labels in the first several layers.

	Word Loyalty	Label Loyalty
Isolating labels (first)	40.8	41.7
Isolating labels (last)	100	99.0
Isolating non-labels (first)	60.0	39.3
Isolating non-labels (last)	100	99.3

Given the consistency of these results with our earlier findings, we can conclude that our observations and insights remain valid. We hope this detailed response addresses your concerns.

ICL中的标签正确性对于结果似乎影响不大？

Thank You!