Label Words are Anchors

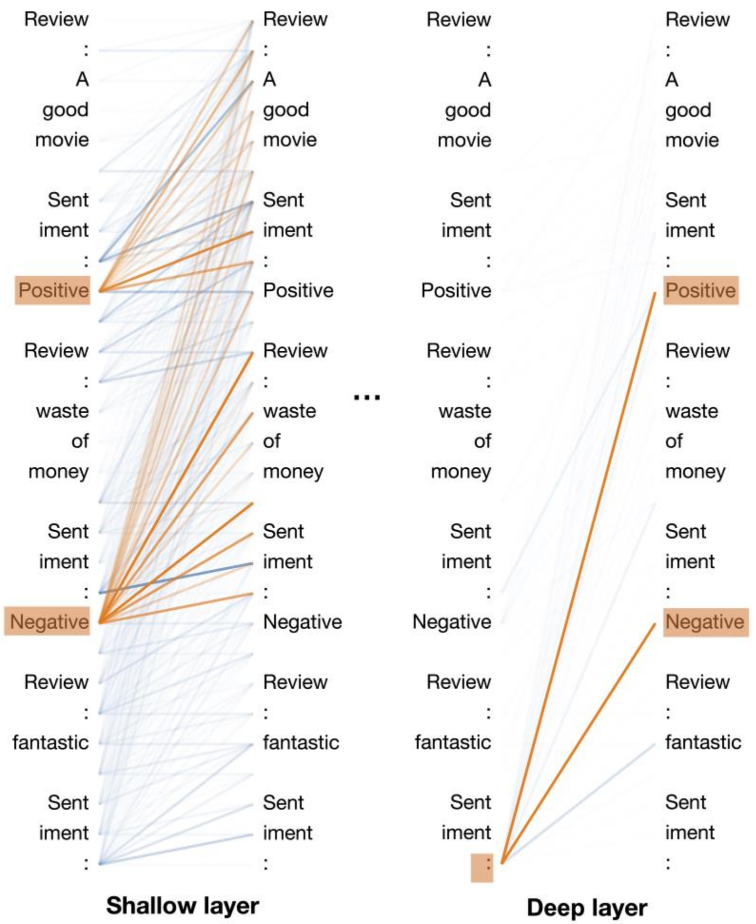
Meizhi Zhong

12.21

Label Words are Anchors

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



Motivation

How LLMs learn from the provided context?

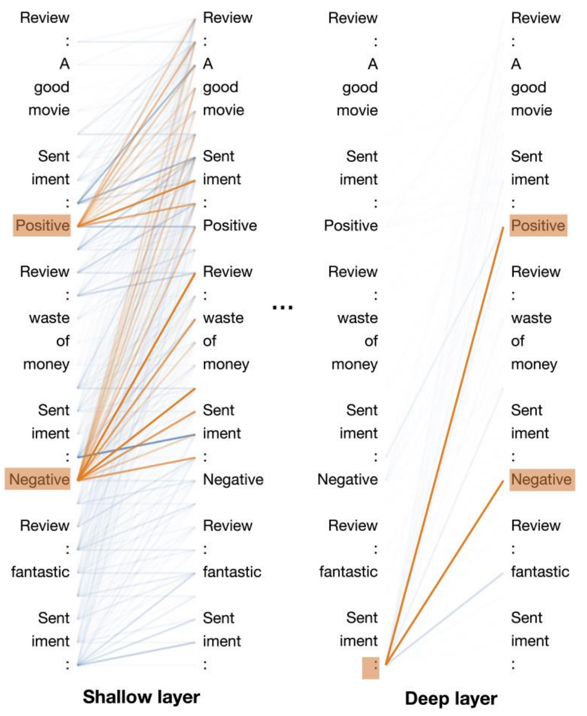


Figure 1: Visualization of the information flow in a GPT model performing ICL. The line depth reflects the significance of the information flow from the right word to the left. The flows involving label words are highlighted. Label words gather information from demonstrations in shallow layers, which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

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.

Significance of Information

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

$I_l(i, j)$ represents the significance of the information flow from the j-th word to the i-th word for ICL.

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

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

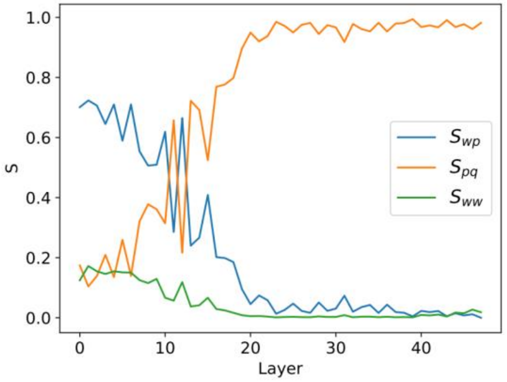
S_{pq}, the mean significance of information flow from label words to the target position:

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

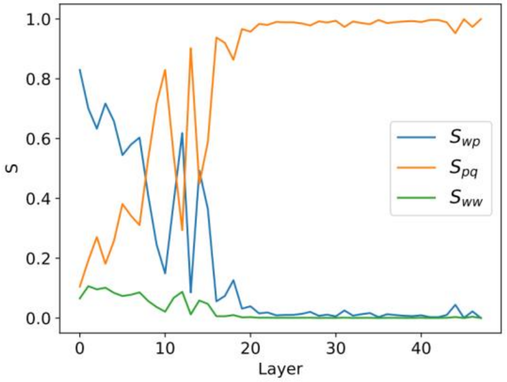
S_{ww}, the mean significance of the information flow amongst all words, excluding influences represented by S_{wp} and S_{pq} :

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

Label Words are Anchors: Hypothesis Motivated by Saliency Scores



(a) Results on the SST-2 dataset



(b) Results on the AGNews dataset

Figure 3: Relative sizes of S_{wp}, S_{pq}, and S_{ww} in different layers on SST-2 and AGNews. Results of other datasets can be found in Appendix B. Initially, S_{wp} occupies a significant proportion, but it gradually decays over layers, while S_{pq} becomes the dominant one.

Review: I dislike … Sentiment: Negative | Review: A good … Sentiment: Positive | Review: … Sentiment:

In shallow layers

(1) S_{pq}, the significance of the information flow from label words to targeted positions is low;

(2) S_{wp}, the information flow from the text part to label words is high;

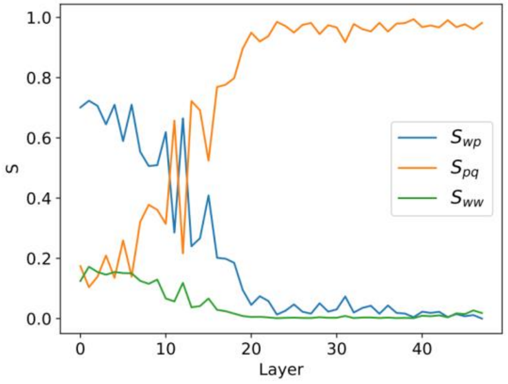
In deep layers

(1) S_{pq}, the significance of the information flow from label words to targeted positions is dominant;

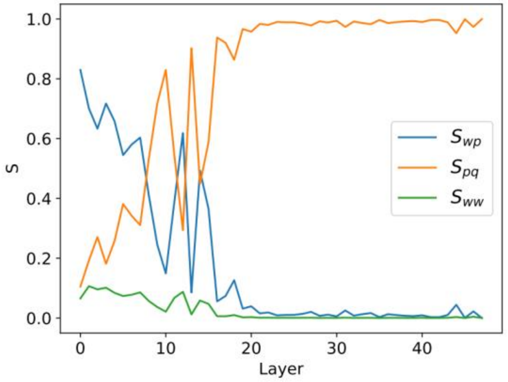
(2) S_{wp}, the information flow from the text part to label words is low;

S_{pq} and S_{wp} usually surpass S_{ww}, suggesting that interactions involving label words outweigh others.

Label Words are Anchors: Hypothesis Motivated by Saliency Scores



(a) Results on the SST-2 dataset



(b) Results on the AGNews dataset

Figure 3: Relative sizes of S_{wp}, S_{pq}, and S_{ww} in different layers on SST-2 and AGNews. Results of other datasets can be found in Appendix B. Initially, S_{wp} occupies a significant proportion, but it gradually decays over layers, while S_{pq} becomes the dominant one.

Review: I dislike … Sentiment: Negative Review: A good … Sentiment: Positive Review: … Sentiment:

In shallow layers

(1) S_{pq}, the significance of the information flow from label words to targeted positions is low;

(2) S_{wp}, the information flow from the text part to label words is high;

In deep layers

(1) S_{pq}, the significance of the information flow from label words to targeted positions is dominant;

(2) S_{wp}, the information flow from the text part to label words is low;

S_{pq} and S_{wp} usually surpass S_{ww}, suggesting that interactions involving label words outweigh others.

Shallow Layers: Information Aggregation

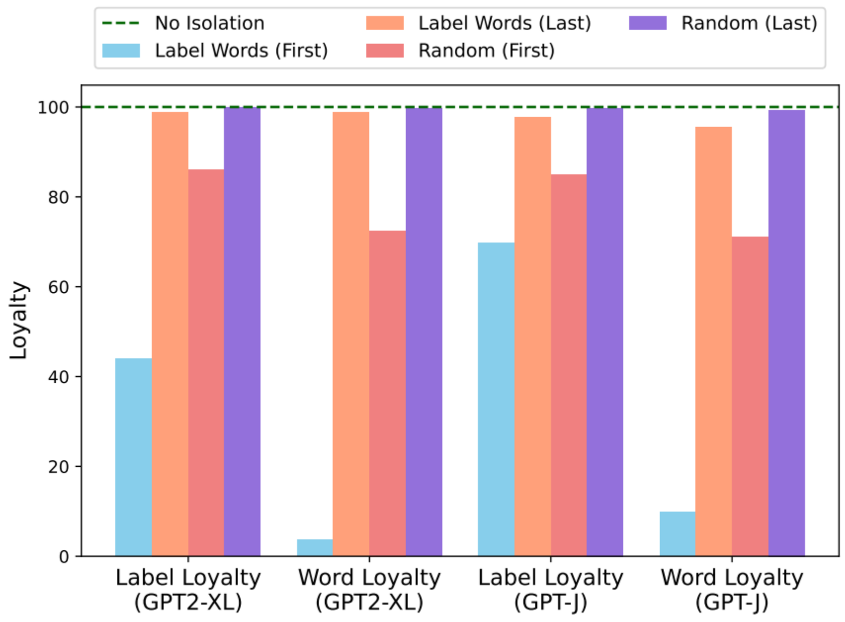


Figure 4: The impact of isolating label words versus randomly isolating non-label words within the first or last 5 layers. Isolating label words within the first 5 layers exerts the most substantial impact, highlighting the importance of shallow-layer information aggregation via label words.

Loyalty

cally, we set $A_l(p,i)(i < p)$ to 0 in the attention matrix A_l of the l-th layer, where p represents label words and i represents preceding words.

(1) Label Loyalty: measures the consistency of output labels with and without isolation.

(2) Word Loyalty: employs the Jaccard similarity to compare the top-5 predicted words /w and /wo isolation, capturing more subtle model output alterations

Shallow Layers: Information Aggregation

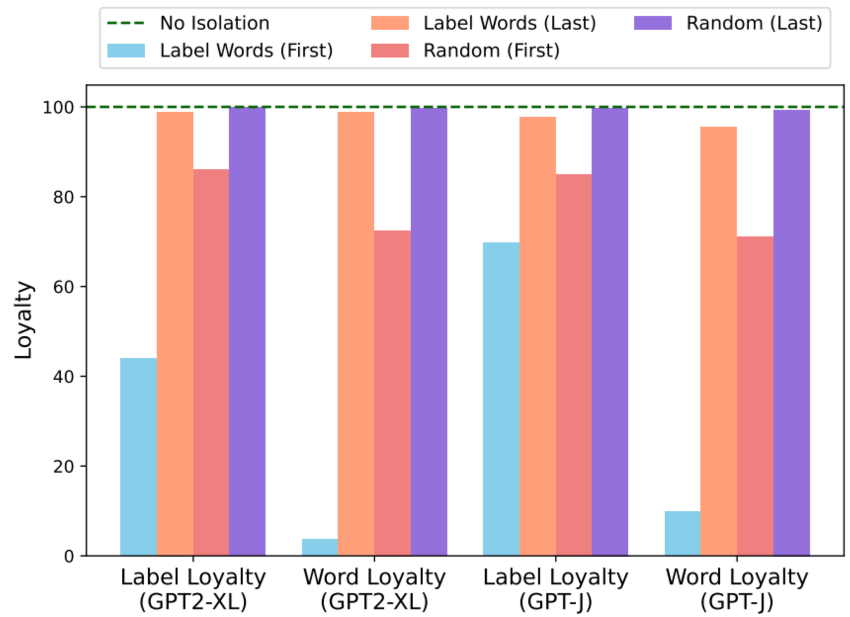


Figure 4: The impact of isolating label words versus randomly isolating non-label words within the first or last 5 layers. Isolating label words within the first 5 layers exerts the most substantial impact, highlighting the importance of shallow-layer information aggregation via label words.

Loyalty

cally, we set $A_l(p, i)(i < p)$ to 0 in the attention matrix A_l of the l-th layer, where p represents label words and i represents preceding words.

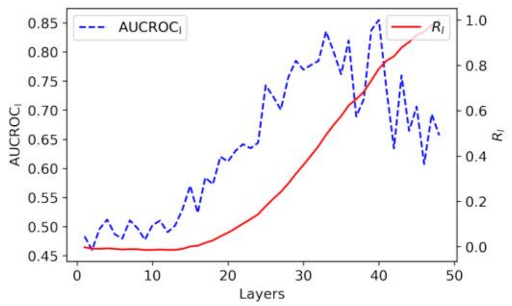
(1) Label Loyalty: measures the consistency of output labels with and without isolation.

(2) Word Loyalty: employs the Jaccard similarity to compare the top-5 predicted words /w and /wo isolation, capturing more subtle model output alterations

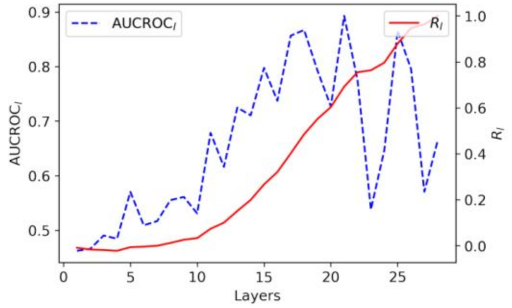
- **notable influence** on the model's behavior when label words are isolated within the first 5 layers
- **Influence becomes inconsequential** within the last 5 layers.

- Emphasizes the **superiority** of **label** words over non-label words.

Deep Layers: Information Extraction



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



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

Figure 5: AUCROC$_l$ and R_l of each layer in GPT models. The result is averaged over SST-2, TREC, AGNews, and Emoc. AUCROC$_l$ reaches 0.8 in deep layers, and R_l increases mainly in the middle and later layers.

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

- R_ℓ quantify the accumulated contribution of the first ℓ layers to model prediction

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

- AUC-ROC得分越接近1.0，表示模型性能越好，能够更好地区分正例和负例。如果AUC-ROC得分接近0.5，则模型性能较差，类似于随机猜测。

Hypothesis

- Shallow layers: assemble information from demonstrations via label words to form semantic representations.

- Deep layers: the aforementioned aggregated information on label words is then extracted to form the final prediction.

- Label words are Anchors

Anchor Re-weighting: Method(prefermance)

a strong correlation between the model's output category and the attention distribution

$$\boxed{\left(A\left(q, p_1\right), \ldots, A\left(q, p_C\right)\right)}$$

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

$$\log \frac{\operatorname{Pr}_{\boldsymbol{f}}(Y=i \mid X=x)}{\operatorname{Pr}_{\boldsymbol{f}}(Y=C \mid X=x)} = \boldsymbol{\beta}_i^T \hat{\mathbf{x}}.$$

$$\log \frac{\operatorname{Pr}_{\boldsymbol{f}}(Y=i \mid X=x)}{\operatorname{Pr}_{\boldsymbol{f}}(Y=C \mid X=x)} = \beta_0^i + \boldsymbol{\beta}_i^T \mathbf{x}.$$

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

To train the re-weighting vector $\boldsymbol{\beta} = \{\beta_0^i\}$, we utilize an auxiliary training set $(\boldsymbol{X}_{train}, \boldsymbol{Y}_{train})$. Here, we perform ICL with normal demonstrations and optimize $\boldsymbol{\beta}$ with respect to the classification loss $\mathcal{L}$ on $(\boldsymbol{X}_{train}, \boldsymbol{Y}_{train})$:

$$\boldsymbol{\beta}^\star = \arg\min_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{X}_{train}, \boldsymbol{Y}_{train}). \qquad (10)$$

Anchor Re-weighting: 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.

Adding more demonstrations for vanilla ICL may not bring a stable accuracy boost due to the potential noise introduced

This shortens the input context and thus brings (almost) no extra cost to the inference speed.

Anchor-Only Context Compression: Method (Speed)

- a context compression technique that reduces the full demonstration to anchor hidden states for accelerating ICL inference.

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.

(2022b). As a solution, we amalgamate the hidden states of both the formatting and the label words, a method we've termed **Hidden$_{\text{anchor}}$**.

Anchor-Only Context Compression

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

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

$\text{Text}_{\text{anchor}}$: This method concatenates the formatting and label text with the input, as opposed to concatenating the hidden states at each layer.
$\text{Hidden}_{\text{random}}$: This approach concatenates the hidden states of formatting and randomly selected non-label words (equal in number to $\text{Hidden}_{\text{anchor}}$).
$\text{Hidden}_{\text{random-top}}$: To establish a stronger baseline, we randomly select 20 sets of non-label words in $\text{Hidden}_{\text{random}}$ and report the one with the highest label loyalty.

(2022b). As a solution, we amalgamate the hidden states of both the formatting and the label words, a method we've termed **$\text{Hidden}_{\text{anchor}}$**.

Here, "formatting" refers to elements like "Review:" and "Sentiment:"

only leads to a 1.5 accuracy drop

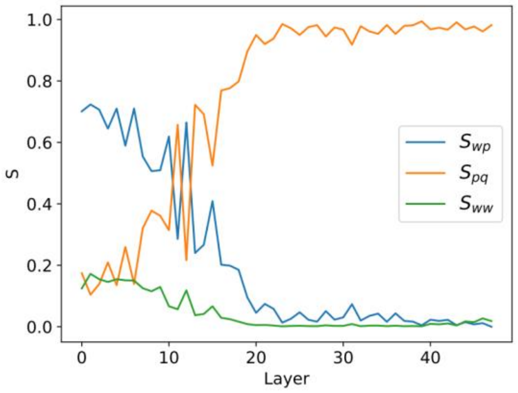
Efficiency improvements

- speed-up ratio ranges from 1.1× to 2.9×
- the acceleration effect is more pronounced in the GPT-J model compared to GPT2-XL, demonstrating its great potential to apply to larger language models.

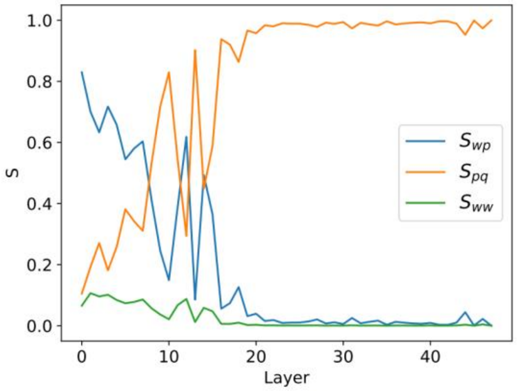
Model	SST-2	TREC	AGNews	EmoC
GPT2-XL	1.1×	1.5×	2.5×	1.4×
GPT-J	1.5×	2.2×	2.9×	1.9×

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

Thanks



(a) Results on the SST-2 dataset



(b) Results on the AGNews dataset

Figure 3: Relative sizes of S_{wp}, S_{pq}, and S_{ww} in different layers on SST-2 and AGNews. Results of other datasets can be found in Appendix B. Initially, S_{wp} occupies a significant proportion, but it gradually decays over layers, while S_{pq} becomes the dominant one.

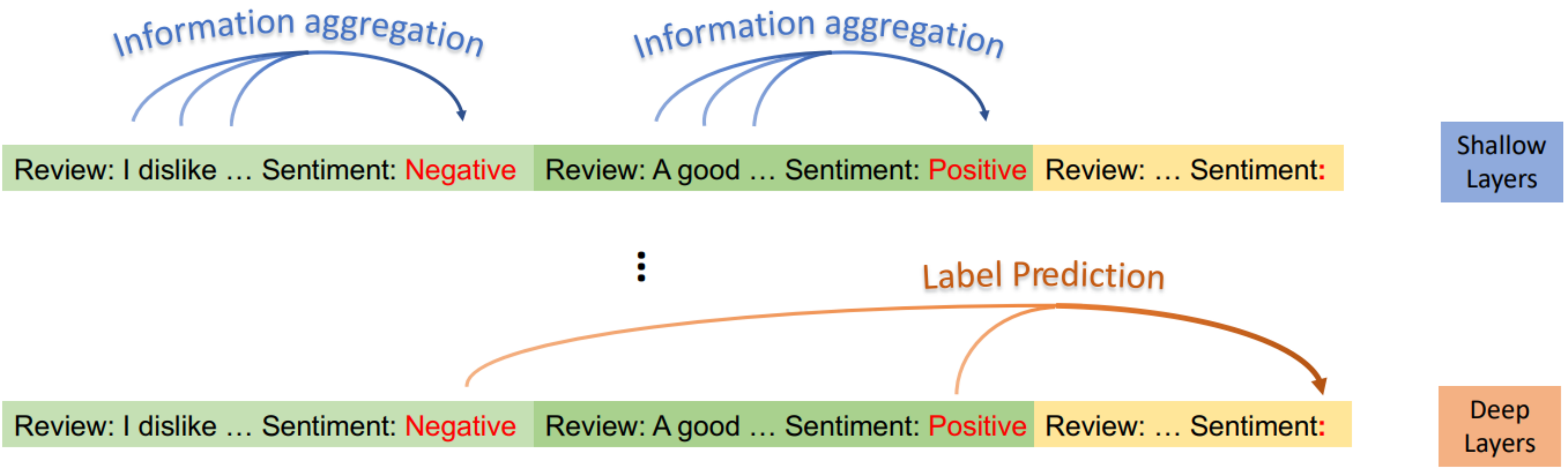


Figure 2: Illustration of our hypothesis. In shallow layers, label words gather information from demonstrations to form semantic representations for deeper processing, while deep layers extract and utilize this information from label words to formulate the final prediction.