The Information Propagation Hypothesis: Optimizing Information Flow in Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) have shown remarkable success in various tasks, yet their internal mechanisms remain inadequately understood. This paper investigates these mechanisms by analyzing how input query information propagates within task-specific spaces. Specifically, we propose a prompt-pair detection method that constructs a task-specific label space and projects hidden representations onto it to examine information propagation during the understanding, generation, and decision-making stages. Our findings reveal that LLMs compress and decompress query information into hidden representations near the task-specific label space during the understanding and generation stages. In the decision-making stage, labels with distributions similar to the query are predicted, but these labels do not always match the true labels, leading to errors. To address this, we analyze the query distribution and find that queries tend to cluster around semantically similar queries, regardless of proximity to the true label. Based on this, we propose a similarity-based voting method (SiV) that aggregates votes from semantically similar queries to improve prediction accuracy, mitigating errors caused by relying solely on label similarity. Extensive experiments show that SiV enhances both accuracy and speed, while also enabling incremental updates without training.

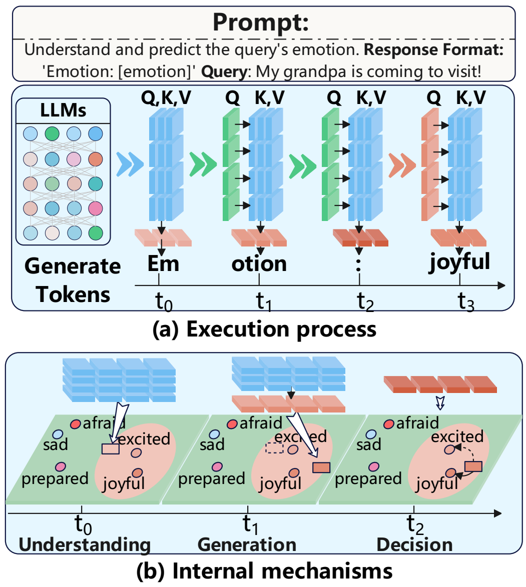

Figure 1: Task completion in LLMs and internal mechanisms under our information propagation hypothesis (IPH). Our IPH: LLMs compress information to the target label space during understanding, decompress and approach the label during generation, and predict results based on similarity during decision-making.

1 Introduction

Large language models (LLMs) have been widely applied to tasks such as text generation (Liu et al., 2024; Long et al., 2024), logical reasoning (Fu et al., 2022; Wang et al., 2022; Yao et al., 2023) and emotion recognition (Yang et al., 2024; Qian et al., 2023), achieving significant results. Existing research mainly focuses on their surface performance (Radford et al., 2019; Luo et al., 2023; Agrawal et al., 2022), neglecting a deeper exploration of task execution, particularly how information flows within the model during the understanding, generation, and decision-making processes and how it impacts final predictions, a question that remains unanswered.

In this paper, we reveal the internal mechanisms of LLMs by observing and analyzing the flow of internal hidden representations in a task-specific space. Inspired by neuroscience research on hidden representations in LLMs (Olah, 2023; Park et al., 2023; Liu et al., 2024), we propose a prompt-pair detection method that constructs the task-specific label space and projects the LLMs' hidden representations into it. By analyzing the mutual information and similarity of the projection during the understanding, generation, and decision-making processes, we identify the following patterns of information transfer: (i) **Understanding Stage**:

LLMs progressively compress the input query information from shallow to deep layers, ultimately mapping it to a target label space. As shown in Figure 1, for the query "my grandpa is coming to visit!" with the emotion "joyful," LLMs compress the information through the layers at time step 0, mapping it to spaces near labels like "joyful" and "excited." (ii) **Generation Stage**: LLMs gradually decompress the compressed information, forming hidden representations closest to the target category at specific time steps. As shown in Figure 1, LLMs gradually decompress the query information to generate tokens, with the hidden representation closest to the target "joyful" emotion formed at the 1st time step. (iii) **Decision Stage**: LLMs compare the decompressed hidden representations to the labels in the label space, with higher similarity increasing the likelihood of the corresponding label being predicted. As shown in Figure 1, at the 2nd time step, the hidden representation is closer to "joyful" than "excited," making the former more likely to be predicted as the result. **Based on these observations, we propose the Information Propagation Hypothesis (IPH)**: LLMs compress query information in the understanding stage, progressively decompress it in the generation stage, and make predictions based on label distribution similarity.

To validate the hypothesis, we manipulate LLMs' hidden representations to block or enhance information transfer in emotion classification (Chatterjee et al., 2019; Rashkin et al., 2019), topic classification (Li and Roth, 2002; Hovy et al., 2001), and question answering (Mallen et al., 2023). The results show that blocking information significantly reduces performance, while enhancing it significantly improves performance. Further analysis reveals the underlying reason: blocking information causes the hidden representations to deviate from the true label distribution, making it harder for LLMs to select the correct label with lower similarity, resulting in poorer performance. Conversely, performance improves when information is enhanced. These findings validate the information propagation hypothesis: LLMs compress and decompress query information and make predictions based on similarity to label distributions.

However, in practice, query information does not always align with the true label distribution, resulting in lower similarity and inaccurate predictions. Thus, relying on labels to predict results is not always effective. To optimize this process, we analyze query distributions and find that semantically similar queries tend to have similar distributions, regardless of their alignment with the true label distribution. Therefore, using semantically similar queries is reliable in predictions. Based on this insight, we propose a **Si**milarity-based **V**oting method (SiV), which retrieves semantically similar queries and uses their corresponding labels to determine the final prediction.

To validate the effectiveness, we conduct experiments using Phi3.5-mini, Llama3.1$_{8b}$, and Mistral-Nemo on question answering (Mallen et al., 2023), emotion classification (Chatterjee et al., 2019), topic classification (Li and Roth, 2002; Hovy et al., 2001), and fine-grained emotion recognition tasks (Rashkin et al., 2019). The results show that SiV improves average accuracy and macro F1 scores by 19% and over 10%, respectively, while achieving a 2.0× speedup. Furthermore, the method's ability to flexibly expand and modify reference queries without retraining allows for incremental iteration and adjustment, offering high flexibility and adaptability.

Overall, our contributions are as follows: (i) We introduce a prompt-pair detection method that enables the construction of a label space, facilitating the exploration of LLMs' internal mechanisms. (ii) Building on the label space, we present the Information Propagation Hypothesis, which posits that LLMs compress information during the understanding stage, progressively decompress it in the generation stage, and make predictions based on distributional similarity in the decision stage. (iii) We propose a simple yet effective similarity-based voting method to enhance the information propagation process in LLMs. (iv) Extensive experiments demonstrate that our approach significantly boosts performance and speed, while enabling incremental updates and iterative improvements without retraining

2 Label Space Construction

To investigate how Large language models (LLMs) perform tasks, we propose a prompt-pair detection method that constructs a task-specific label space, enabling us to analyze changes in internal representations.

2.1 Prompt-Pair Detection Method

Our prompt-pair detection method is based on the linear representation and superposition hypothe-

Infer the dialogue from the perspective of the emotion “**joyful**”. Dialogue Context: <sample s_i>. Response Format: “Emotion: <emotion c_i>”.

Table 1: Positive prompt for the emotion recognition.

ses (Olah, 2023; Park et al., 2023), extracting shared label representations of labels across samples to form category-specific representations.

For clarity, we use the label c_i (e.g., “joyful”) as an example. Given this label, we collect N samples $S = [s_1, ..., s_i, ...s_N]$ belonging to the same category and construct positive and negative prompt pairs. The positive prompt is shown in Table 1. The key difference between these prompts is that the positive prompt is labeled with c_i, while the negative prompt uses a random label from the task label set C. Both positive and negative prompts are then input into the LLM, and token generation is performed using a teacher-forcing approach, defined as follows:

$$y^+_{t,s_i} = LLM(P^+_{s_i}, y^+_{<t}) \tag{1}$$

$$y^-_{t,s_i} = LLM(P^-_{s_i}, y^-_{<t}) \tag{2}$$

where $P^+_{s_i}$, $P^-_{s_i}$ are the positive and negative prompts, respectively. y_{t,s_i} and $y_{<t}$ refer to the token at time step t and the tokens before time step t, respectively.

For the generated token at time step t, the hidden representations at l-th layer for the positive and negative representations are denoted as $h^{l,+}_{t,s_i}$ and $h^{l,-}_{t,s_i}$, respectively. By subtracting these representations, we maximize the acquisition of label-specific information (Liu et al., 2024; Turner et al., 2023), yielding the representation h^l_{t,s_i}:

$$h^l_{t,s_i} = h^{l,+}_{t,s_i} - h^{l,-}_{t,s_i} \tag{3}$$

where h^l_{t,s_i}, $h^{l,+}_{t,s_i}$, $h^{l,-}_{t,s_i} \in R^d$, and d is the hidden representation dimension of LLMs. To obtain stable label representations across different contexts, we collect the representations from N samples and apply Principal Component Analysis (PCA) to extract common features. The representation for label c_i is $h^l_{c_i} \in R^d$, as follows:

$$H^l_{c_i} = [h^l_{1,s_1}, ..., h^l_{t,s_i}, ..., h^l_{t,s_N}] \tag{4}$$

$$h^l_{c_i} = PCA(H^l_{c_i}) \tag{5}$$

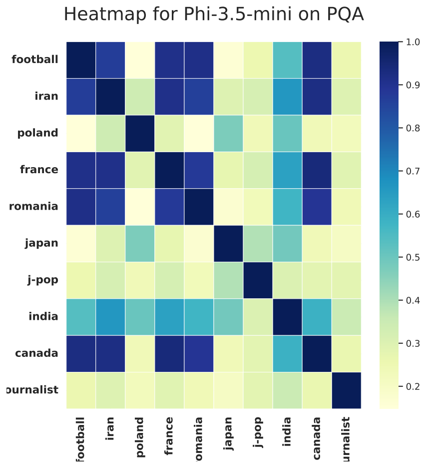

Figure 2: Heatmap for Phi-3.5-mini on PQA dataset.

2.2 Label Space Analysis

After constructing the label representations, we treat the collection of these vectors as the label space. To clearly present the label space, we visualize the label distributions for the models Phi3.5-mini, Llama3.1$_{8b}$, and Mistral-Nemo across emotion recognition (EC) (Chatterjee et al., 2019), topic classification (TREC) (Li and Roth, 2002; Hovy et al., 2001), question answering (PQA) (Mallen et al., 2023), and empathy dialogue datasets (ED) (Rashkin et al., 2019).

Figure 3 shows the label distribution visualization. The results reveal that labels with emotional and semantic meanings are more dispersed. For instance, label distributions in the empathy dialogue dataset (ED) and question answering dataset (PQA) are more scattered, while those in emotion-related datasets (ED and EC) are more concentrated. This suggests that the label representations are generally well-distributed. Figure 2 illustrates label visualizations on the PQA dataset. According to the results, semantically similar categories, like "Iran" and "France," exhibit higher similarity, whereas categories with larger semantic differences, like "Iran" and "J-pop," show lower similarity. Further experiments and analysis in the Appendix A confirm the reasonableness of the label distributions.

3 Information Propagation Hypothesis

LLMs complete specific tasks by understanding the input query and progressively converting the information into the target label. For example, in the emotion recognition task, given the query “my grandpa is coming to visit!”, the LLM understands the emotion and outputs the corresponding target

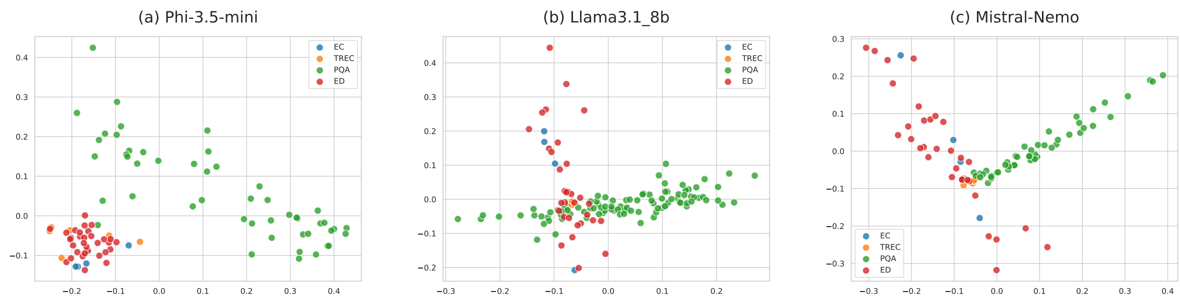

Figure 3: Visualization of label distribution.

emotion "joyful." This process raises an important question: how does the LLM internalize and transfer the information from the input to the target label? From the perspective of the label space, we define this problem as: how do LLMs propagate and convert information during the understanding, generation, and decision-making processes to map it to the target label space?

To address this question, we examine the target labels predicted by LLMs. We find that LLMs often struggle to accurately predict the emotion of a query, such as "my grandpa is coming to visit!"—it may not predict the correct emotion, "joyful," but usually predicts a similar emotion, such as "happy" or "excited." That is, expanding the target label space is necessary. Therefore, we define the correct label c_i and its k_1 most similar neighbors as the target label space h_{ts}, as formulated below:

$$\overline{h}_{c_i} = Mean(h^l_{c_i}); \overline{h}_{c_j} = Mean(h^l_{c_j}) \quad (6)$$

$$h_{ts} = Top_{k_1}(\overline{h}_{c_i}, \overline{h}_{c_j}) \quad (7)$$

where $\overline{h}_{c_i} \in R^d, h^l_{ts} \in R^{k_1 \times d}, c_i, c_j \in C$. c_i and c_j are label categories, while C is the set of labels for the task. $Mean$ refers to the mean pooling function. Top_{k_1} is the selection function, which selects the top k_1 labels with the highest similarity scores. h_{ts} represents the representation of the most similar neighbor labels, including the label itself, while h^l_{ts} refers to the corresponding representation of the l-th layer.

For the query's hidden representation, we project it onto the target label space,called projection, and represent it as:

$$h^l_p = \sum_{k=1}^{k_1} \frac{h^l \cdot h^l_{ts}}{|h^l_{ts} \cdot h^l_{ts}|} h^l_{ts} \quad (8)$$

where $h^l_p \in R^d$ is the projection of the hidden representation.

To further observe the internal information changes within the task, we compute the Mutual Information (MI) at each stage of the task execution. Mutual Information measures the dependency between two random variables X and Y. It quantifies the reduction in uncertainty of one variable given the other. In practice, exact computation of mutual information is infeasible as the true probability distributions are unknown. Therefore, we employ K-Nearest Neighbors (KNN)-based methods to approximate the densities. Using these approximations, the MI can be expressed as:

$$I(X;Y) \approx \frac{1}{M} \sum_{i=1}^{M} \log \frac{\hat{p}(x_i, y_i)}{\hat{p}(x_i)\hat{p}(y_i)} \quad (9)$$

where $\hat{p}(x_i, y_i)$, $\hat{p}(x_i)$, and $\hat{p}(y_i)$ are the estimated joint and marginal probabilities for the samples x_i and y_i. M represents the number of samples.

3.1 Definition of Task Completion Stages

In the following sections, we divide the task execution process of LLMs into three stages: understanding, generation, and decision-making. Taking the emotion recognition task as an example, LLMs are required to understand the query and generate a response in the form of "Emotion: [emotion]."

Understanding Stage. At the time step 0, LLMs learn the query as a hidden representation, encoding its information; we call this the understanding stage. Note that we use decoder-based LLMs, which, although lacking a distinct encoder, still encode information at step 0 through self-attention.

Generation Stage. From time step 0 to t_k, LLMs generate the prompt-specified content, "Emotion:" We refer to this as the generation stage.

Decision-making Stage. At the k-th key time step, LLMs generate the token ":". Based on previous research (Wang et al., 2023), this step consolidates the most important information for prediction. We denote this time step as t_k, called the decision point, and refer to the process of converting the hidden representation into the result at this step as the decision process.

3.2 Understanding Stage: Information Compression

Hypothesis. The first step in completing the task is to interpret the query as task-relevant information. In this process, we hypothesize that LLMs continuously compress the query information towards the target label space.

Experiment. According to information bottleneck theory (Saxe et al., 2019; Slonim, 2002; Tishby et al., 2000; Tishby and Zaslavsky, 2015), for the hidden representations z_j from shallow to deep layers, if the mutual information $I(x, z_j)$ between the input representation x and the hidden representation z_j decreases, while the mutual information $I(z_j, y)$ between the hidden representation and the target representation y increases, it indicates that the model is compressing information towards the target. To verify this, we calculate the mutual information between the input representation at 0-th layer, x=h_p^0, the intermediate hidden representation z_j=$h_p^{l_j}$, and the target representation y=h_{ts}^l.

$$I_{x,z} = I(h_p^0, h_p^{l_j}); I_{z,y} = I(h_p^{l_j}, h_{ts}^l) \qquad (10)$$

where h_p^0, $h_p^{l_j}$, and h_{ts}^l represent the query's input projection, the intermediate projection, and the target representation at the l_j-th layer. The target representation is the label space representation of the query's correct category (see Eq. 7).

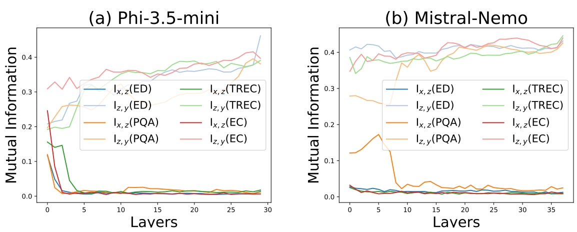

Figure 4: Variation of mutual information in LLMs during the understanding stage.

Results and Analysis. Figure 4 shows the results of Phi-3.5-mini, Llama3.1$_{8b}$, and Mistral-Nemo on multiple datasets. The results indicate that as the depth of layers increases, the mutual information between the intermediate projection and the input projection gradually decreases until it stabilizes, while the mutual information between the intermediate projection and the target representation steadily increases. Notably, due to differences in training data and methods, the degree and efficiency of compression vary. Nevertheless, the results still demonstrate that LLMs compress information towards the target label space, confirming our hypothesis.

3.3 Generation Stage: Information Decompression

The second step in task completion is to generate the corresponding tokens based on the understanding. To track the information changes in this process, we calculate the mutual information between the projection at step t_i of the generation stage and the projection at the understanding stage (t_i = 0) or the target label. Since the key time step t_k consolidates the most crucial information for prediction (Wang et al., 2023), we also observe the mutual information changes at each layer of LLMs at this time step.

3.3.1 Mutual Information in Time Steps

Hypothesis. Since the goal of LLMs in the generation stage is to output the target category, we hypothesize that during this stage, they continue to accumulate information towards the target label space until the key time step reaches its peak.

Experiment. To verify this hypothesis, we calculate the mutual information between the projection at the time step t_i and the projection at the time step t_i=0 or the target label space, as follows:

$$\overline{I}_{x,z}^{t_i} = \frac{1}{L}\sum_{l=1}^{L} I(h_p^{l,t_0}, h_p^{l,t_i}) \qquad (11)$$

$$\overline{I}_{z,y}^{t_i} = \frac{1}{L}\sum_{l=1}^{L} I(h_p^{l,t_i}, h_{ts}^{l}) \qquad (12)$$

where h_p^{l,t_0} is the projection at the l-th layer in the understanding stage (t_i=0), h_p^{l,t_i} is the projection at l-th layer in the generation stage.

Results and Analysis. As shown in Figure 5, the mutual information between the projection in the generation process and the projection in the understanding stage remains stable, indicating that LLMs retain some query information during generation. Meanwhile, the mutual information between the projection in the generation process and the target label space increases before the key time step and decreases afterward. This suggests that LLMs extract information towards the target space until the key time step. In summary, these results show that LLMs continuously extract and decompress information into the target space during the generation process, until the key time step. Additional experiments in Appendix B confirm the same conclusion.

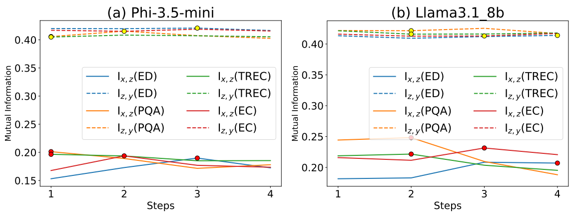

Figure 5: Mutual Information at Generation Time Steps.

3.3.2 Mutual Information at Key Time Step

Hypothesis. Previous studies show that at the key time step, deep layers in LLMs aggregate information essential for prediction (Wang et al., 2023). Therefore, we hypothesize that at this point, the deep-layer projection is more closely aligned with the target space.

Experiment. To validate this hypothesis, we compute the mutual information at the key time step t_k between the projection in the generation stage and the projection in the understanding stage or the target label space, as follows:

$$I_{x,z} = I(h_p^{l,t_0}, h_p^{l,t_k}); I_{z,y} = I(h_p^{l,t_k}, h_{ts}^{l}) \quad (13)$$

where h_p^{l,t_k} represents the projection at the l-th layer at time step t_k.

Results and Analysis. As shown in Figure 6, the results reveal that at the key time step, the mutual information between the projection in the generation stage and projection in the understanding stage fluctuates while maintaining a high level of information. This suggests that LLMs adjust and retain connections with the understood information. Additionally, the mutual information between the projection in the generation stage and the target label space fluctuates from shallow to mid-layers, reaching or approaching its maximum in deeper layers. This indicates that LLMs aggregate target-related information in the deeper layers. Overall, LLMs retain understood information while extracting and decompressing target-related information in the deeper layers, making it crucial for prediction. Additional experiments in Appendix C support this conclusion.

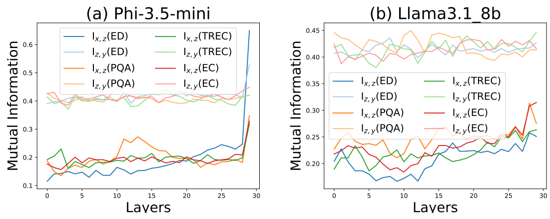

Figure 6: Mutual information at key time step t_k.

3.4 Decision Stage: Similarity-Based Prediction

The final step in task completion is to decide the output category based on the understanding and generation content.

Hypothesis. Since the hidden representation at the key time step heavily influences the output (Wang et al., 2023), we focus on how the projection at this moment affects the decision. We hypothesize that the stronger the correlation between the projection and a specific category, the more likely LLMs are to predict that category.

Experiment. To validate this, we compute the dot product between the hidden representation and label space representations at the key time step, as shown below:

$$\overline{h}_{c_j} = Mean(h_{c_j}^{l}); \overline{h}_{t_k} = Mean(h_{t_k}^{l}) \quad (14)$$

$$o_{i,j} = \overline{h}_{c_j} \cdot \overline{h}_{t_k} \quad (15)$$

where $\overline{h}_{c_j}, \overline{h}_{t_k} \in R^d$, are the category representations in the label space and the hidden representation at time step t_k, respectively. $o_{i,j}$ is the dot product score.

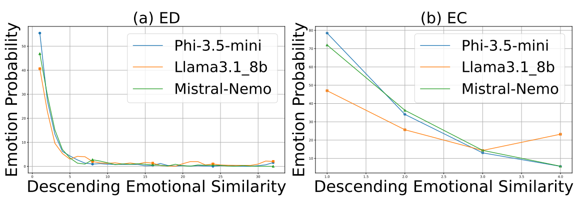

Figure 7: Category probabilities based on descending similarity.

Results and Analysis. Figure 7 shows the results. The x-axis represents the accuracy of categories sorted by descending dot product scores, and the

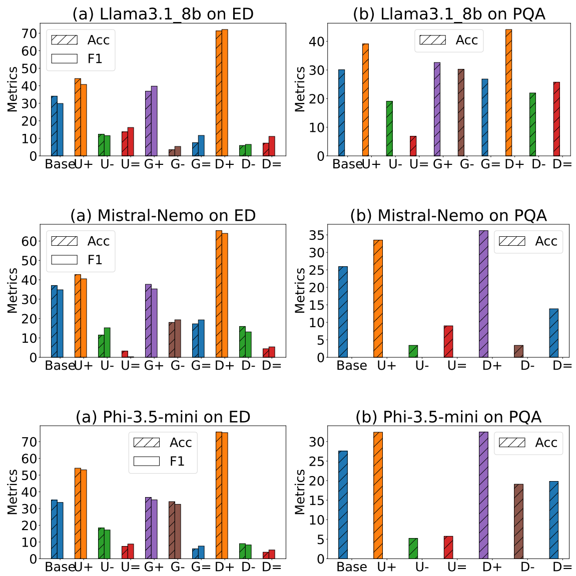

Figure 8: Validation results of LLMs on the ED and PQA datasets.

y-axis shows the predicted probability for each category. The results indicate that categories with higher dot product scores are more likely to be predicted as the target. This suggests that LLMs make decisions based on similarity during the decision-making process. Further experiments in Appendix D further confirms this conclusion.

4 Validation of Hypotheses on Tasks

In this section, we propose the information propagation hypothesis based on our experiments and analysis, and validate it across multiple datasets.

Information Propagation Hypothesis. LLMs compress information towards the target label space in the understanding stage, decompress and extract information in the generation stage, and make decisions based on similarity in the decision stage.

Blocking and Enhancement Experiments. To validate the hypothesis on real tasks, we manipulate the information propagation in LLMs by enhancing or disrupting it at different stages. For a query, enhancing information involves adding the ground truth label representation (in Eq. 5) at the corresponding stage, while disrupting information involves subtracting or replacing it, as below:

$$h_{t_i}^{l,s} = h_{t_i}^{l} - \alpha h_{c_i}^{l}; h_{t_i}^{l,r} = \alpha h_{c_i}^{l} \tag{16}$$

$$h_{t_i}^{l,a} = h_{t_i}^{l} + \alpha h_{c_i}^{l} \tag{17}$$

where $h_{t_i}^{l,a}$, $h_{t_i}^{l,s}$, $h_{t_i}^{l,r} \in R^d$, and α is a hyperparameter.

The enhancement experiment brings the hidden representation closer to the correct category's label space. In contrast, subtraction moves the hidden representation further away. This experiment helps verify the impact of similarity-based decision-making. The replacement experiment retains only category-related information. If performance improves, it suggests that LLMs can make accurate predictions relying solely on label information; otherwise, it indicates that other semantic information in the hidden representation is also crucial. We conduct experiments during the understanding stage, early generation steps (e.g., generating "Emotion"), and the key time step (e.g., generating ":"). For simplicity, we refer to these stages as the understanding stage, generation stage, and decision point.

Results and Analysis. The results are shown in Figure 8. From the results, we find that: (i) Enhancing at the decision point greatly improves accuracy, while enhancing during the understanding stage provides a moderate boost. Little change occurs during the generation process. This suggests that the decision point is the most critical, followed by the understanding stage. This also suggests that bringing the hidden representation closer to the correct label space improves decision accuracy. (ii) Disrupting the understanding stage and decision point significantly reduces performance, while the effect during the generation process varies by model. This emphasizes the importance of information at the understanding stage and decision point, with varying results in the generation stage due to model characteristics. This also suggests that disrupting representations reduces accuracy by distancing them from the correct label space. (iii) Replacing information at the understanding stage and decision point significantly lowers performance, with varying effects during the generation process. This suggests that LLMs' decisions depend not only on label information but also on other semantic factors. (iv) The above experiments show that the similarity between the query's hidden representation at the decision point and the correct label space significantly impacts the prediction result. From this perspective, we find that LLMs' inability to consistently propagate the query's hidden representation to the correct label space leads to poor performance.

5 Methodology

The above experiments show that the distributional mismatch between the query's hidden representation and the correct label space leads to inaccurate predictions. To address this, we analyze the query's distribution at the decision point and propose a similarity-based voting method based on the characteristics of the query representation.

Sample Distribution Observation. Experiment and analysis details are in Appendix D. The results show that, regardless of whether the query matches the correct label distribution, it is often close to semantically similar queries.

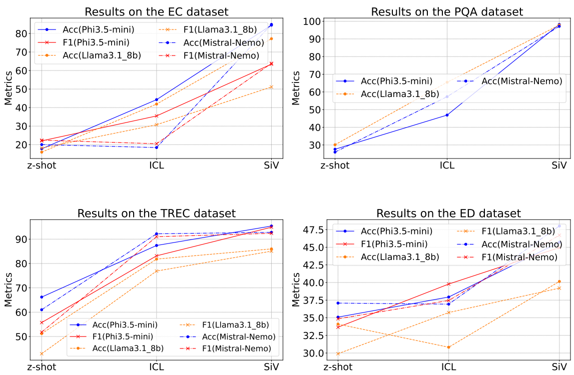

Figure 9: Main results on the datasets.

Similarity-Based Voting. We propose a Similarity-based Voting (SiV) method to address the issue of LLMs relying on a similar label space. Given a query, we retrieve the top k_2 most similar queries from the training set, then use their labels to vote, selecting the most frequent label as the result. The process is formalized as follows:

$$h_{t_k}^{q_i} = Mean(h_{t_k,q_i}^{l}); h_{t_k}^{q_j} = Mean(h_{t_k,q_j}^{l}) \quad (18)$$

$$c_{q_i} = Vote(Top_{k_2}(h_{t_k}^{q_i} \cdot h_{t_k}^{q_j})) \quad (19)$$

where $h_t^{q_i}$, $h_t^{q_j} \in R^d$. Top_{k_2} selects the top k_2 queries with the highest scores, and $Vote$ assigns the label with the most votes as the prediction.

To validate SiV, we conduct experiments with multiple LLMs across EC (Chatterjee et al., 2019), TREC (Li and Roth, 2002; Hovy et al., 2001), ED (Rashkin et al., 2019), PQA (Mallen et al., 2023) datasets. Experimental setup details are in Appendix E.

Main Results. As shown in Figure 9 (See Table 4 in Appendix F for details), the results show that, compared to zero-shot learning, in-context learning (ICL) performs better. This is mainly because ICL incorporates task-relevant information through samples, leading to more accurate decisions. Our proposed SiV, however, significantly outperforms ICL, demonstrating that our method enhances the decision-making process of LLMs.

Time Consumption. We evaluate SiV's experimental efficiency on the Empathetic Dialogue (ED) dataset. As shown in Table 5 (see Appendix F), our method doubles the inference speed compared to ICL. This is because it generates fewer tokens to make predictions, reducing time consumption. Additionally, generating fewer tokens reduces cache usage during generation, lowering computational resource demands. Moreover, since the training set queries can be adjusted and expanded anytime, our method offers high scalability and flexibility.

6 Related Work

Recent studies have explored information flow and reasoning in LLMs, including multimodal interactions (Zhang et al., 2025), zero-shot Chain-of-Thought mechanisms (Yuan et al., 2024), and knowledge conflicts (Jin et al., 2024), while others focus on interpretability through gradient projection (Katz et al., 2024) and methods for controlling information flow and ensuring security (Men et al., 2024; Tiwari et al., 2024; Siddiqui et al., 2024). These methods overlook how information flows and impacts outcomes in LLM task execution. To address this gap, this paper explores the information propagation process.

7 Conclusion

This paper investigated the internal mechanisms of large language models (LLMs) by analyzing how query information propagates within task-specific spaces. We proposed the information propagation hypothesis, explaining how LLMs compress and decompress query information during the understanding, generation, and decision-making stages through compression-decompression and similarity-based decision propagation. Based on this, we introduced the similarity-based voting method (SiV) to optimize internal information flow and enhance LLM performance. Extensive experiments show that SiV improved accuracy, speed, and flexibility. In the future, we will explore deeper mechanisms within LLMs.

Limitations

Our paper has the following limitations: (i) The label space representation constructed by the prompt-based detection method contains some noise, which affects certain experimental results. We plan to address this issue in the future. (ii) The proposed method has been tested only on classification and question-answering tasks, both of which specify the response format. We have not validated it on natural text generation tasks, but we aim to explore this further in the future.

Ethical Considerations

Regarding the potential ethical impacts of our work: (i) The data we use is open-source and does not pose any ethical risks. (ii) The baseline models and methods we use are also open-source and do not involve ethical concerns. (iii) Moreover, the components employed in our approach are either open-source or innovative, and do not present potential ethical risks.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation. *arXiv preprint arXiv:2212.02437*.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*.

Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran. 2001. Toward semantics-based answer pinpointing. In *Proceedings of the First International Conference on Human Language Technology Research*.

Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. *arXiv preprint arXiv:2402.18154*.

Shahar Katz, Yonatan Belinkov, Mor Geva, and Lior Wolf. 2024. Backward lens: Projecting language model gradients into the vocabulary space. *arXiv preprint arXiv:2402.12865*.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Huanshuo Liu, Hao Zhang, Zhijiang Guo, Jing Wang, Kuicai Dong, Xiangyang Li, Yi Lee, Cong Zhang, and Yong Liu. 2024. Ctrla: Adaptive retrieval-augmented generation via inherent control. *arXiv*.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126*.

Man Luo, Xin Xu, Zhuyun Dai, Panupong Pasupat, Mehran Kazemi, Chitta Baral, Vaiva Imbrasaite, and Vincent Y Zhao. 2023. Dr. icl: Demonstration-retrieved in-context learning. *arXiv preprint arXiv:2305.14128*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Tianyi Men, Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Unlocking the future: Exploring look-ahead planning mechanistic interpretability in large language models. *arXiv preprint arXiv:2406.16033*.

Chris Olah. 2023. Distributed representations: Composition & superposition. *Transformer Circuits Thread*, 24.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*.

Yushan Qian, Wei-Nan Zhang, and Ting Liu. 2023. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. *arXiv preprint arXiv:2310.05140*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *ACL*, page 5370–5381.

Andrew M Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D Tracey, and David D Cox. 2019. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124020.

Shoaib Ahmed Siddiqui, Radhika Gaonkar, Boris Köpf, David Krueger, Andrew Paverd, Ahmed Salem, Shruti Tople, Lukas Wutschitz, Menglin Xia, and Santiago Zanella-Béguelin. 2024. Permissive information-flow analysis for large language models. *arXiv preprint arXiv:2410.03055*.

Noam Slonim. 2002. *The information bottleneck: Theory and applications*. Ph.D. thesis, Citeseer.

Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. IEEE.

Trishita Tiwari, Suchin Gururangan, Chuan Guo, Weizhe Hua, Sanjay Kariyappa, Udit Gupta, Wenjie Xiong, Kiwan Maeng, Hsien-Hsin S Lee, and G Edward Suh. 2024. Information flow control in machine learning through modular model architecture. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 6921–6938.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv e-prints*, pages arXiv–2308.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Label words are anchors: An information flow perspective for understanding in-context learning. *arXiv preprint arXiv:2305.14160*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *ArXiv*, abs/2203.11171.

Zhou Yang, Zhaochun Ren, Wang Yufeng, Shizhong Peng, Haizhou Sun, Xiaofei Zhu, and Xiangwen Liao. 2024. Enhancing empathetic response generation by augmenting llms with small-scale empathetic models. *arXiv preprint arXiv:2402.11801*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *ArXiv*, abs/2305.10601.

Xiaosong Yuan, Chen Shen, Shaotian Yan, Xiaofeng Zhang, Liang Xie, Wenxiao Wang, Renchu Guan, Ying Wang, and Jieping Ye. 2024. Instance-adaptive zero-shot chain-of-thought prompting. *arXiv preprint arXiv:2409.20441*.

Xiaofeng Zhang, Fanshuo Zeng, and Chaochen Gu. 2025. Simignore: Exploring and enhancing multimodal large model complex reasoning via similarity computation. *Neural Networks*, 184:107059.

Appendix

A Label Space Analysis

To clearly present the label space, we visualize the label space of Phi-3.5-mini, Llama3.1$_{8b}$, and Mistral-Nemo across different datasets.

Figure 12a shows the results for Phi-3.5-mini on the TREC dataset. The results indicate a strong similarity between the categories HB (human beings) and ENT (entities). We hypothesize that this similarity arises because “human beings” and “entities” often appear in similar contexts, resulting in higher similarity in their vector representations.

Figure 12b presents the results for Phi-3.5-mini on the EC dataset. The results show a certain degree of similarity between emotional categories, while their similarity with the non-emotional category “others” is lower. Negative emotions like “sad” and “angry” are highly similar, while other emotions have weaker associations.

Figure 10 shows the visualization of labels on the ED dataset. As shown in Figure 10, in the empathy dialogue dataset (ED), categories with similar emotions, such as “annoyed” and “furious,” have higher similarity, while those with larger emotional differences, such as “annoyed” and “hopeful,” have lower similarity.

Additionally, Figures 11 and 13 display the heatmaps for Mistral-Nemo and Llama3.1$_{8b}$, respectively, showing similar patterns. These results suggest that the constructed label space is reasonable.

B Mutual Information in Time Steps

To verify the information changes during the generation process, we calculate the mutual information between the projection in the generation process and the projection in the understanding stage or the target label space using Mistral-Nemo. As shown in Figure 15, the results show that the mutual information between the projection in the generation and understanding stages remains stable, while the mutual information with the target label space increases to a peak and then decreases. This further indicates that LLMs continuously extract and decompress information until the critical time step.

C Mutual Information at Key Time Step

Figure 15 shows the mutual information between the projection at the key time step and the understanding stage or the target label space for LLMs.

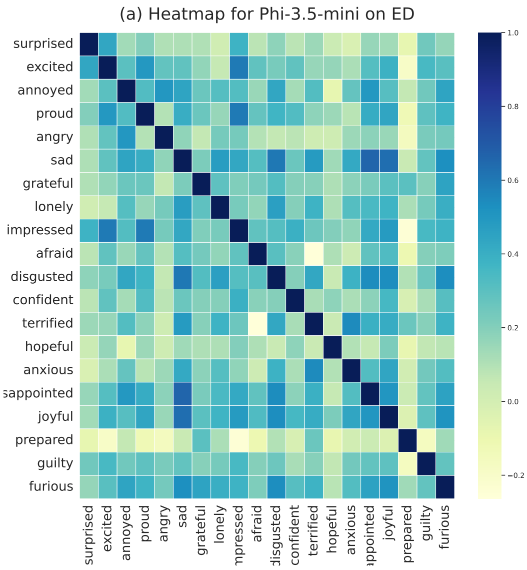

Figure 10: Heatmap for Phi-3.5-mini on the ED datasets.

The results indicate that the mutual information between the projection at the key time step and the understanding stage fluctuates. The mutual information between the projection at the key time step and the target label space fluctuates from shallow to mid-layers, reaching or approaching its maximum at the final layer. Moreover, the change in mutual information with the target label space is much greater than with the understanding stage. This suggests that at this time step, LLMs adjust and extract information from the understanding stage and decompress it towards the target label space in deeper layers.

D Similarity-Based Decision

Figure 16 shows the results on the TREC dataset. The results indicate that on Phi-3.5-mini and Llama3.1$_{8b}$, the category with the highest similarity is more likely to be predicted as the result. In Mistral-Nemo, the second most similar category is also predicted as the result. Overall, these models tend to predict the most or second most similar category, supporting the hypothesis of similarity-based decision making.

We select samples from the ED dataset for experimentation. Given a query q_i, we use Roberta$_{large}$ to generate its representation and label, then select three groups of queries q_j with varying similarity: (i) **High similarity group**: Queries with the same label and cosine similarity greater than $1-\beta$. (ii) **Middle similarity group**: Queries with cosine similarity between $1-2\beta$ and $1-\beta$. (iii) **Low similarity group**: Queries with cosine similarity between $1-3\beta$ and $1-2\beta$. Note that LLMs' emotion recognition accuracy for these samples is below 60%, meaning some query representations are close to the correct label space at the decision point, while others are not.

LLMs	High	Middle	Low
Phi3.5-mini	98.40	95.97	95.50
Llma3.1$_{8b}$	91.52	82.62	75.02
Mistral-Nemo	99.91	99.81	99.74

Table 2: Sample distribution on ED datasets.

We calculate the similarity between the hidden representations of query q_i and q_j at the decision point. As shown in Table 2 and Figure 17, results reveal that the high similarity group scores highest, followed by the middle group, with the low similarity group scoring lowest. This suggests that LLMs often map semantically similar queries to similar hidden representations, regardless of their proximity to the correct label space.

To illustrate the distribution of queries, we provide a visualization. As shown in Figure 17, similar queries tend to cluster together, suggesting that LLMs process semantically similar queries with similar hidden representations at the decision point.

E Implementation

Large Language Models. We conduct experiments using three advanced large language models: Phi-3.5-mini, Llama3.1$_{8b}$, and Mistral-Nemo on the NVIDIA L40. (i) **Phi-3.5-mini**: A lightweight model with approximately 3.8 billion parameters, designed for efficient performance on various tasks. (ii) **Llama3.1$_{8b}$**: A large-scale model with 8 billion parameters, known for its high accuracy in a range of NLP tasks. (iii) **Mistral-Nemo**: A powerful model with 12 billion parameters, offering fine-tuned performance and scalability for complex tasks.

Experimental Setup. During the experiments, we set the value of t_1 based on the characteristics of each model and dataset, as shown in Table 1. For the disruption and enhancement experiments, we set the hyperparameter α to 2 for the ED dataset. For the PQA dataset, we set α to 10 for Mistral-Nemo, 2 for Llama3.1$_{8b}$, and 50 for Phi-3.5-mini.

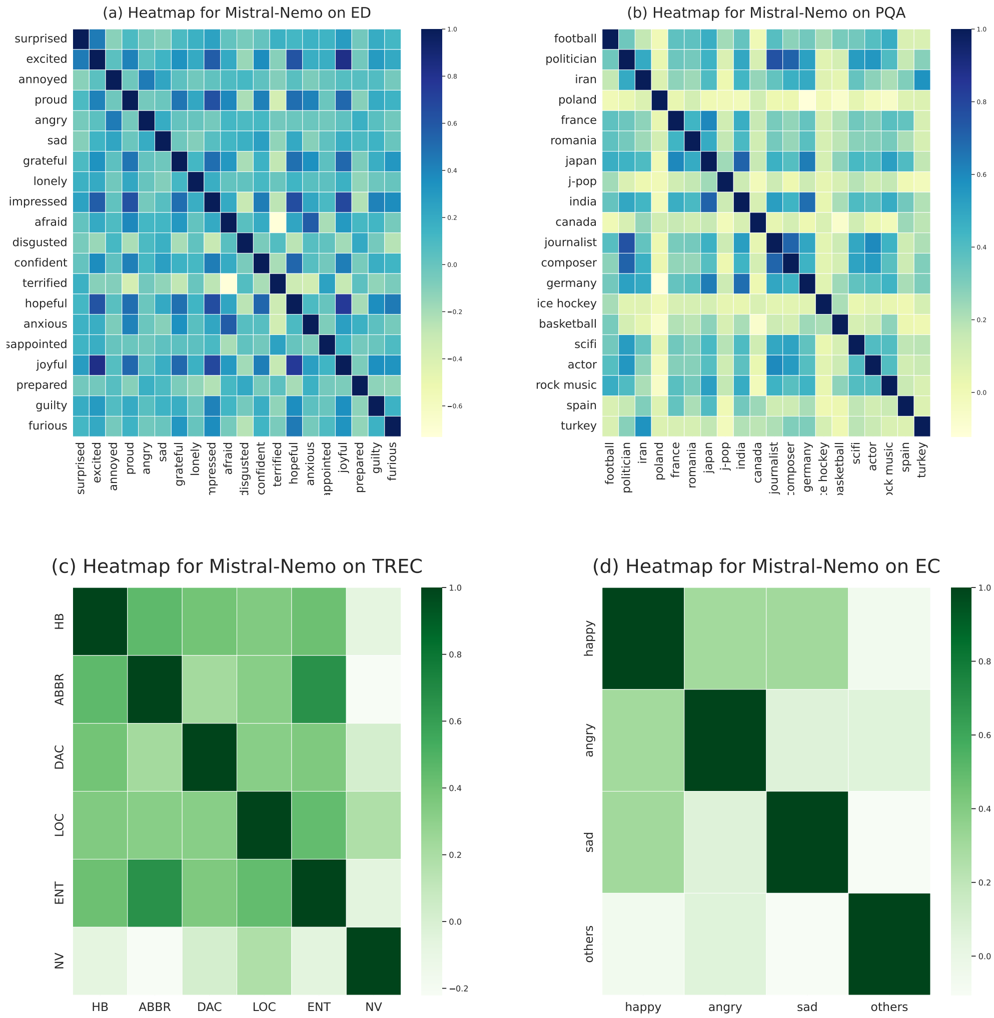

Figure 11: Heatmap for Mistral-Nemo on the ED, PQA, TREC, and EC datasets, with TREC categories including human beings (HB), abbreviations (ABBR), description and abstract concepts (DAC), locations (LOC), entities (ENT), and numeric values (NV).

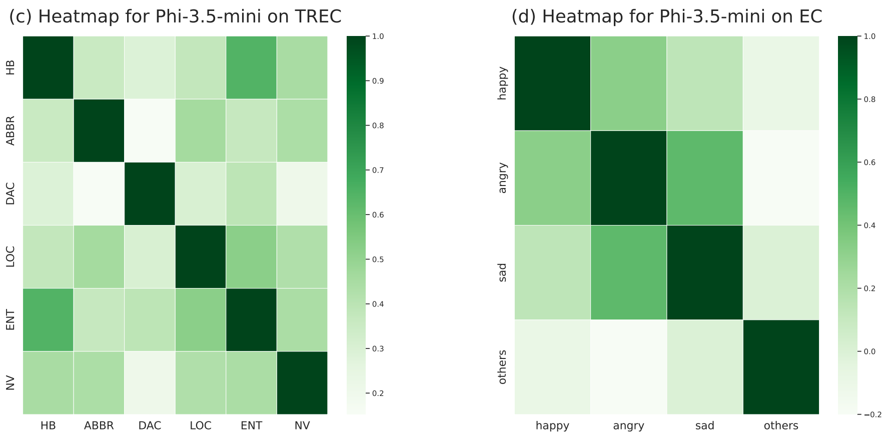

Figure 12: Heatmap for Phi-3.5-mini on the TREC and EC datasets, with TREC categories including human beings (HB), abbreviations (ABBR), description and abstract concepts (DAC), locations (LOC), entities (ENT), and numeric values (NV).

LLMs	EC	TREC	ED	PQA
Phi3.5-mini	1	1	5	3
Llma3.1$_{8b}$	1	2	2	5
Mistral-Nemo	1	1	3	3

Table 3: Hyperparameter k_1 Settings.

In the sample distribution experiment, we set the threshold to $\beta = 0.0005$.

Datasets. For the datasets, we use the EmoContext (EC) dataset with 6 emotion categories, the Text Retrieval Conference Question Classification (TREC) dataset, the Empathetic Dialogues (ED) dataset with 32 emotion categories, and the short-text question answering dataset PopQA (PQA).

Evaluation Metrics. we use Accuracy (Acc) and Macro-F1 (F1) as evaluation metrics. Accuracy measures the overall proportion of correctly classified instances, while Macro-F1 calculates the average F1 score across all classes, treating them equally.

F Experimental Results

To validate the effectiveness of SiV, we conduct experiments. Tables 4 and 5 present the results for performance and speed, respectively.

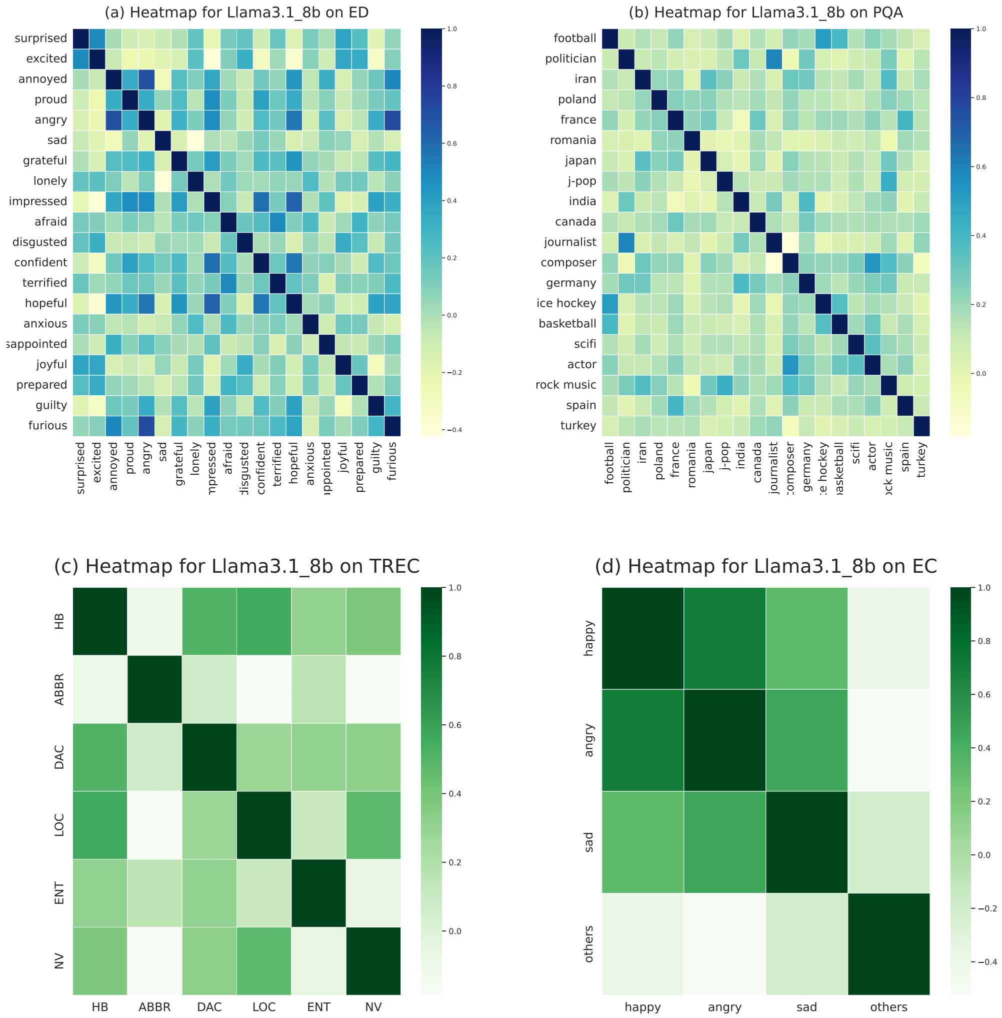

Figure 13: Heatmap for Llama3.1$_{8b}$ on the ED, PQA, TREC, and EC datasets, with TREC categories including human beings (HB), abbreviations (ABBR), description and abstract concepts (DAC), locations (LOC), entities (ENT), and numeric values (NV).

LLM	Models	EC		TREC		ED		PQA	Average	
		Acc	F1	Acc	F1	Acc	F1	Acc	Acc	F1
Phi3.5-mini	z-shot	17.76	22.04	66.2	55.73	35.09	33.66	27.56	36.65	37.14
	ICL	44.28	35.49	87.4	83.15	37.94	39.78	46.9	54.13	52.80
	SiV	**84.49**	**63.44**	**95.39**	**94.9**	**45.7**	**44.95**	**98.2**	**80.94**	**67.76**
Llma3.1$_{8b}$	z-shot	15.96	18.12	51.2	42.93	34.1	29.9	30.09	32.83	30.31
	ICL	41.89	30.75	81.8	76.89	30.82	35.74	65.47	54.99	47.79
	SiV	**77.17**	**51.13**	**86**	**85.06**	**40.15**	**39.2**	**97.84**	**75.29**	**58.46**
Mistral-Nemo	z-shot	20.02	22.4	61	51.8	37.08	34.85	25.94	36.01	36.35
	ICL	18.43	20.5	92.2	91.01	36.93	37.45	57.32	51.22	49.65
	SiV	**84.88**	**63.89**	**92.8**	**92.39**	**48.01**	**46.71**	**97.12**	**80.70**	**67.66**

Table 4: Results for LLMs on the datasets.

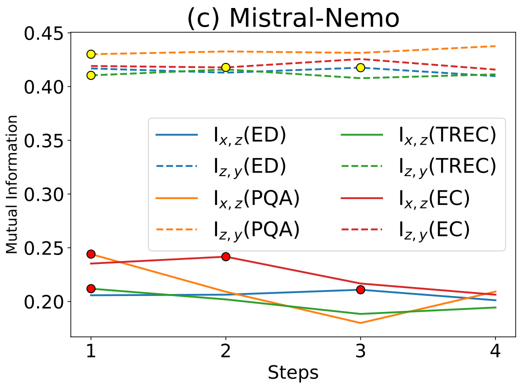

Figure 14: Mutual Information at Generation Time Steps.

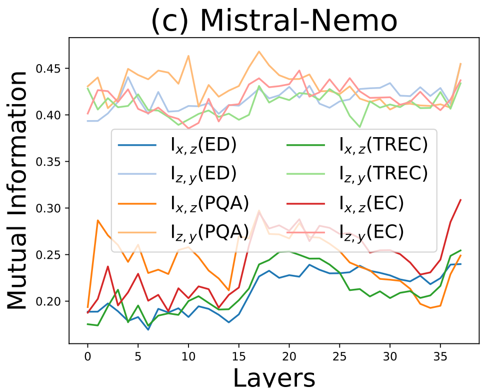

Figure 15: Mutual Information at the Key Time Steps.

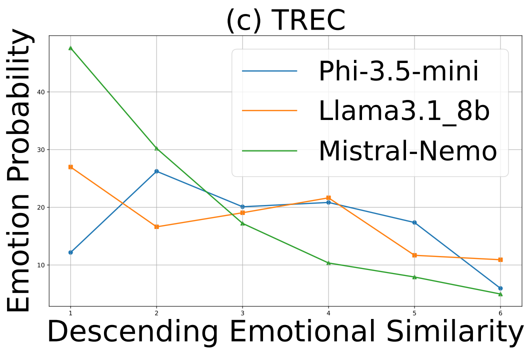

Figure 16: Category probabilities based on descending similarity.

LLMs	ICL	SiV	Speed-up
Phi3.5-mini	13m 33s	6m 31s	2.07×
Llma3.1$_{8b}$	12m 0s	5m 39s	2.12×
Mistral-Nemo	17m 7s	8m 14s	2.07×

Table 5: Comparison of time consumption on the ED dataset.

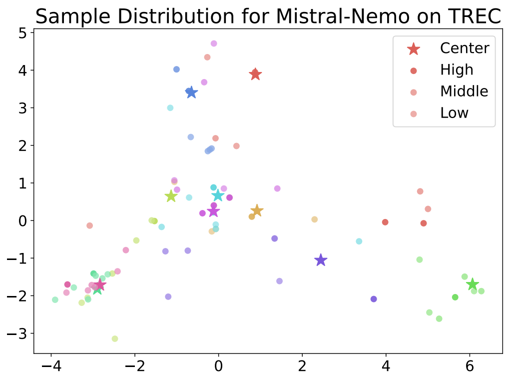

Figure 17: Query distribution for Mistral-Nemo on the TREC dataset.