

**Thesis for the Master of Artificial Intelligence**

**Unveiling Context Control Mechanism in Large  
Language Models by Layer-wise Evaluation**

Wooseok Song

Graduate School of Hanyang University

August 2024

# **Thesis for the Master of Artificial Intelligence**

## **Unveiling Context Control Mechanism in Large Language Models by Layer-wise Evaluation**

**Thesis Supervisor: Younghoon Kim**

**A Thesis submitted to the graduate school of  
Hanyang University in partial fulfillment of the  
requirements for the Master of Artificial Intelligence**

**Wooseok Song**

**August 2024**

**Department of Applied Artificial Intelligence  
Graduate School of Hanyang University**

This thesis, written by Wooseok Song,  
has been approved as a thesis for the Master of Artificial  
Intelligence.

August 2024

**Committee Chairman:** Woohwan Jung (Signature)

**Committee Member:** Younghoon Kim (Signature)

**Committee Member:** Kyungtae Kang (Signature)

Graduate School of Hanyang University

# Contents

<b>Contents</b>	<b>i</b>
<b>List of Tables</b>	<b>iii</b>
<b>List of Figures</b>	<b>iv</b>
<b>Abstract</b>	<b>v</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Related Works</b>	<b>4</b>
2.1 Controlled Dialogue Generation . . . . .	4
2.2 Explainability for LLMs . . . . .	5
<b>3 Emotion Controlled Dialogue Generation</b>	<b>7</b>
3.1 Models and Prompt Template . . . . .	8
3.2 Dialogue Context Dataset . . . . .	10
3.3 Generation Results . . . . .	11
<b>4 Effects of Emotion Control Instruction on Dialogue Contexts</b>	<b>16</b>
4.1 Cosine Similarity of Hidden states of Dialogue Context with and without Emotion Control . . . . .	17
4.2 2D Visualization through PCA . . . . .	23
4.3 Probing with Logistic Regression Classifier . . . . .	28

4.4 Discussion . . . . .	30
<b>5 Conclusions</b>	<b>33</b>
국문요지	<b>40</b>

**List of Tables**

3.1 Example of Instruction Templates . . . . . 9

3.2 Example of Generated Sentences . . . . . 12

3.3 Emotion Accuracy per Model . . . . . 15

4.1 Emotion Accuracy, Plateau points, Convergence points . . . . . 32

## List of Figures

3.1	Controlled Dialogue Generation task . . . . .	7
3.2	Sample of DailyDialog Dataset . . . . .	10
3.3	Accuracy per Emotion . . . . .	14
4.1	Hidden States of Dialogue Context in two situations . . . . .	17
4.2	Cosine Similarity results per Model . . . . .	20
4.3	Average of Cosine similarity and Plateau point of models . . . .	21
4.4	Correlation between Plateau points and the Accuracy of emotion control . . . . .	22
4.5	PCA on Hidden states produced by LLaMA . . . . .	24
4.6	PCA on Hidden states produced by Alpaca and DialoGPT . . . .	25
4.7	PCA on Hidden states produced by GPT2 and WizardLM . . . .	26
4.8	PCA on Hidden states produced by WizardMath and ChatGLM . . .	27
4.9	Logistic Regression Accuracy of Models, at every layers. . . . .	29
4.10	Correlation between Convergence points and the Accuracy of emotion control . . . . .	31

# ABSTRACT

## Unveiling Context Control Mechanism in Large Language Models by Layer-wise Evaluation

Wooseok Song  
Department of Applied Artificial Intelligence  
The Graduate School  
Hanyang University

Despite of outstanding performance of LLMs on various tasks such as dialogue generation task and their in-context learning ability, understanding how these models achieve their results remains challenging due to their complex black-box nature. This study aims to explore the influence of specific attribute-indicating instruction prompts on dialogue context within a controlled dialogue generation task. We conducted experiments by dividing scenarios into situations where emotion-controlling instruction prompts were provided and those where they were not. Ultimately, we sought to understand how the model’s behavior differs between the two scenarios from a layer-wise perspective. In these two distinct scenarios, we compared the hidden states output from the dialogue context through three experiments. The results revealed how instructional prompts indicating emotion affected



the dialogue context at each layer and how they influenced the final model’s generated output.

In conclusion, we discovered that when we include instructions to control specific attributes like emotion in the model’s input, the model behaves similarly to internally fine-tuned behavior when viewed from a layer perspective. Additionally, we found that as the model encodes emotion at lower layers, its ability to control the desired emotion increases.

## Chapter 1. Introduction

Large language models (LLMs) such as GPT-3 [4], Falcon [3], and LLaMA-2 [24], are highly regarded for their exceptional performance across various natural language processing tasks [17]. Notably, these transformer-based models are well-known for their outstanding capabilities in natural language generation tasks such as summarization and question answering.

Furthermore, these base models have been enhanced through the application of instruction tuning and Reinforcement Learning from Human Feedback (RLHF), enabling LLMs to better understand instructions, and to perform complex, multi-turn conversations. These models, such as GPT-4 [18], LLaMA-2-Chat [24], ChatGLM-2 [8] are also referred to as assistant models. We live in an era where we can easily access and use these models, which have demonstrated their ability to perform new tasks through in-context learning (ICL) [4, 7, 27, 30], by following textual instructions or learning from a few examples.

Despite the fact that LLMs achieve good performance, understanding how models produce such remarkable results is important and highly challenging, as LLMs are infamously complex black-box systems. While there

are several interpretative studies on how LLMs understand context and generate appropriate text, detailed studies that open the black box or analyze LLMs layer-wise are still lacking, leaving a gap in our understanding of their internal mechanisms.

In this paper, We investigated the internal workings of language models by conducting zero-shot in-context learning (ICL) using various models. In Chapter 3, we conducted experiments to examine how instructions designed to indicate specific attributes affect the context. While conducting the Dialogue Generation task, we provided task instructions at the beginning of the dialogue context as input to the LLM and generated the next speaker’s utterance. Additionally, we performed the emotion controlled dialogue task by adding specific emotion-controlling instructions and evaluated the generated sentences.

Then in Chapter 4, we compared how these control instructions influenced the dialogue context by comparing it to when no instructions were given. Ultimately, we aimed to determine how the model behaves in the two scenarios from a layer perspective.

Firstly, we analyzed the differences in the hidden states of the context when attribute control was applied and when it was not. By examining the cosine similarity of outputs of hidden layers when controlling emotion at-

tributes and when not, as the layers deepened, we observed that the behavior pattern of cosine similarity closely resembled that seen when the model underwent fine-tuning [16].

Secondly, we confirmed that the more rapidly specific emotions provided in the instructions were encoded into the context, the more effective emotion control became. Through experiments to identify the point where the decrease in cosine similarity slows down for each model and probing tasks to determine the layers where the model effectively distinguishes emotions, we obtained the aforementioned results.

We hope that our findings and contributions will help users better understand the LLMs they are using and assist developers in advancing them.

## Chapter 2. Related Works

### 2.1 Controlled Dialogue Generation

Dialogue generation systems aim to develop agents that can emulate human conversations using natural language. Generative dialogue models frequently encounter higher standards for consistency, semantics, and interactivity. To manage dialogue responses and boost interactivity, constraints such as emotion, the speaker’s personal style, dialogue intent are implemented. According to the criteria classified in [28], various studies focus on different controllable aspects. In [21, 22], the speaker’s persona was constrained for control, while in [11], the focus was on controlling the politeness of the sentences. In [12, 20], speaker’s sentiment was the aspect being controlled.

In this paper, we controlled speaker’s emotion attributes by providing emotion-controlling instructions via prompts, instead of fine-tuning the LLMs parameters.

## 2.2 Explainability for LLMs

Some studies offer insights into in-context learning (ICL) in large language models. [11] uses the SST-2 sentiment analysis benchmark to analyze ICL through contrastive demonstrations and saliency maps. By flipping labels, altering input text, and adding explanations, the researchers find that label flipping reduces salience in smaller models like GPT-2 but increases it in larger models like InstructGPT. [25] shows whether ICL is driven by semantic priors from pre-training or by learning input-label mappings from examples. The results indicate that large models can override semantic priors and learn contradictory mappings, while smaller models rely more on these priors. This ability of large models to learn arbitrary input-label mappings suggests a form of symbolic reasoning not limited by semantic priors.

Some studies explain the role of fine-tuning in LLMs. [31] fine-tuned the LLaMA-65B model using only 1,000 carefully selected instructions, without reinforcement learning, and achieved performance comparable to GPT-4. They hypothesize that alignment is a simpler process focused on learning interaction styles and formats, with most knowledge acquired during pre-training. This suggests that complex fine-tuning and reinforcement learning may be less crucial than previously thought. [13] show that imitation can

enhance the style, persona, and instruction-following ability of language models, but does not improve performance on more complex tasks such as factuality, coding, and problem-solving.

While there are many studies that have attempted to interpret the workings of LLMs, comprehensive studies that delve into the internal workings of LLMs on a layer-by-layer basis are still scarce, leaving much to be understood about their underlying mechanisms.

## Chapter 3. Emotion Controlled Dialogue Generation

In this section, we describe the zero-shot controlled dialogue generation task we performed. We provided an LLMs with a dialogue context to generate the next speaker's utterance. Instead of only giving the dialogue context as input to the LLM, we also added instructions in front of it. We compared

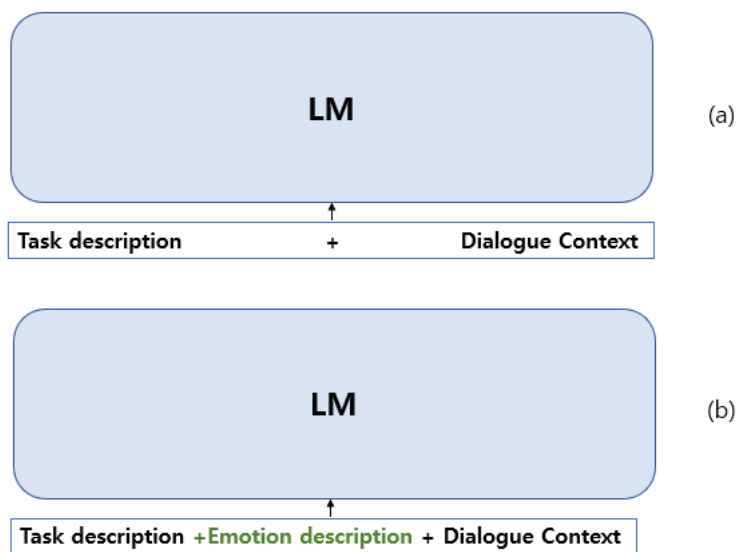


Figure 3.1: Controlled Dialogue Generation task

the generated results between two scenarios: as shown in Figure 3.1 (a), one where the instruction only explained the task, and another as shown in Figure 3.1 (b), where the instruction included both the task explanation



and specific emotion control. The method of providing instruction prompts varies depending on the model. Detailed information can be found in the subsections below.

### 3.1 Models and Prompt Template

To perform the controlled dialogue generation task mentioned earlier, we utilized the following 8 models: LLaMA-2-7B-Chat [4], LLaMA-3-7B-Instruct [1], Alpaca-7B [23], WizardLM-2-7B [26], WizardMath-7B [15], ChatGLM-6B [8], GPT-2-large [19], and DialoGPT-2-large [29]. Each of the models has been pre-trained using different methods: GPT-2 and DialoGPT are models trained solely on large corpora and conversational data, while the rest have been trained using instruction tuning or RLHF.

Additionally, the training data and input prompt templates vary for each model. I crafted input prompts for each model based on its training method and the publicly available training templates or data. You can see examples of Instruction prompt templates of LLaMA-2-7B-Chat, LLaMA-3-7B-Instruct, ChatGLM-6B in Table 3.1.

As shown in the table, each dialogue context is accompanied by a description explaining the task and another description specifying the particular emotion. The sections highlighted in blue denote the emotion-explaining de-

---

**LLaMA-2-7B-Chat:**

[INST]«SYS» In this task you will be shown a conversation context. You need to generate a response to the conversation based on the context.

«/SYS»

You should generate an utterance that reflects the speaker's emotion given. Speaker's emotion is sadness.

Conversation:

Hi . I need to have my shoes repaired . </s>

What 's the matter with them ? </s>

Look at the heels . They are slanting . </s>[/INST]

---

**LLaMA-3-7B-Instruct:**

<|begin\_of\_text|><|start\_header\_id|>system<|end\_header\_id|>

In this task you will be shown a conversation context. You need to generate a response to the conversation based on the context.

You should generate an utterance that reflects the speaker's emotion given. Speaker's emotion is joy.<|eot\_id|><|start\_header\_id|>user1<|end\_header\_id|>

Hi . I need to have my shoes repaired .

<|eot\_id|><|start\_header\_id|>user2<|end\_header\_id|>

What 's the matter with them ?

<|eot\_id|><|start\_header\_id|>user1<|end\_header\_id|>

Look at the heels . They are slanting .

<|eot\_id|><|start\_header\_id|>user2<|end\_header\_id|>

---

**ChatGLM-6B:**

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

### Instruction:

In this task you will be shown a conversation context. You need to generate a response to the conversation based on the context. You should generate an utterance that reflects the speaker's emotion given. Speaker's emotion is joy.

### Input:

Hi . I need to have my shoes repaired . </s>

What 's the matter with them ? </s>

Look at the heels . They are slanting .</s>

### Response:

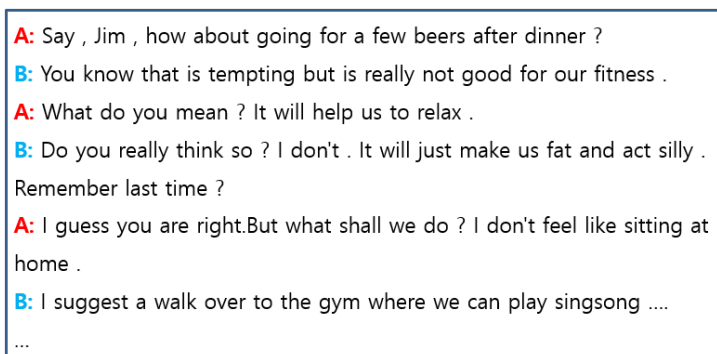
---

Table 3.1: Example of Instruction Templates

scriptions. In Section 3.3, we obtained sentences generated by LLMs when applying and not applying the emotion-explaining descriptions, i.e., when controlling and not controlling emotions, respectively.

## 3.2 Dialogue Context Dataset

We used DailyDialog dataset [14] as the dialogue context input for the LLMs. By referring to Figure 3.1, you can identify which part represents the dialogue context. This dataset is an English dataset containing conversations on various topics, encompassing diversity in conversational characteristics and situations, making it widely used for evaluating and developing natural language understanding and dialogue generation models. You can see an example in Figure 3.2.



A: Say , Jim , how about going for a few beers after dinner ?  
B: You know that is tempting but is really not good for our fitness .  
A: What do you mean ? It will help us to relax .  
B: Do you really think so ? I don't . It will just make us fat and act silly .  
Remember last time ?  
A: I guess you are right.But what shall we do ? I don't feel like sitting at home .  
B: I suggest a walk over to the gym where we can play singsong ....  
...

Figure 3.2: Sample of DailyDialog Dataset

### 3.3 Generation Results

Using the aforementioned LLMs and corresponding instruction prompt templates tailored to each model, we conducted an emotion controlled dialogue generation task. The emotions utilized in the experiment were the six emotions identified by Ekman [9]: joy, surprise, anger, fear, disgust, and sadness. For each emotion, we generated sentences with and without emotion control, i.e., when sentiment control descriptions were included in the instructions and when they were not. The example of sentences generated by each model can be seen in Table 3.2. The table shows that the row labeled "Controlled" after each model name represents the results where emotion was controlled by the instruction, while the row labeled "General" indicates the results when only the task description was provided as instruction. In this example, the given emotion was 'anger'. While some models demonstrate good emotion control and understanding of context, others like GPT-2 and DialogPT struggle with emotion control and sometimes produce responses that do not fit the context well.

As shown in Figure 3.3, We evaluated the emotion accuracy of generated sentences using a test dataset comprising 50 samples selected from the Daily Dialogue dataset. To measure the emotion accuracy of the generated

<b>Dialogue:</b>	
<b>A:</b>	I think Yoga might be suitable for me .
<b>B:</b>	Why ?
<b>A:</b>	Because it doesn't require a lot of energy and thus easy to keep up .
<b>Controlled Sentiment:</b> Anger	
<b>LLaMA-2-7B-Chat-General:</b>	That's a great point! Yoga is definitely a low-impact exercise, which makes it easy on the joints and can be a great option for people who are looking for a low-energy workout.
<b>LLaMA-2-7B-Chat-Controlled:</b>	Really? You think Yoga is suitable for you? Are you kidding me? Have you even tried it before? It's not exactly a cakewalk, you know.
<b>LLaMA-3-7B-Instruct-General:</b>	That's a good point.
<b>LLaMA-3-7B-Instruct-Controlled:</b>	Are you sure about that? I've seen some yoga classes that are quite physically demanding.
<b>Alpaca-7B-General:</b>	That's a great reason to try yoga! It's a great way to get fit without putting too much strain on your body.
<b>Alpaca-7B-Controlled:</b>	But what if I get injured? What if I fall or hurt my back?
<b>DialoGPT-large-General:</b>	I would be very interested in the format.
<b>DialoGPT-large-Controlled:</b>	We had cassettes. All three of them? Is this a real?
<b>GPT-2-large-General:</b>	An international team of astronomers led by the University of Southampton has observed a powerful supernova blast in the constellation Virgo.
<b>GPT-2-large-Controlled:</b>	The best thing about that is the simplicity of its function.
<b>WizardLM-2-7B-General:</b>	That's a great point! Yoga is indeed accessible for all levels of fitness and can be adapted to any energy level.
<b>WizardLM-2-7B-Controlled:</b>	Hi doctor, I just got a cover pos result. X-ray and fever are normal. I need to be quarantined for one week after that. How long should I stay in hospital?
<b>WizardMath-7B-General:</b>	I think yoga would be great for you because it is low-impact and doesn't require a lot of energy to practice. This means it would be easy for you to maintain and keep up with.
<b>WizardMath-7B-Controlled:</b>	I'm not sure about that. Are you sure Yoga is suitable for you ?
<b>ChatGLM-6B-General:</b>	Great! Yoga can be a very beneficial and relaxing activity. It allows you to connect with your body and mind, and can help with physical and mental well-being.
<b>ChatGLM-6B-Controlled:</b>	But is it really suitable for someone who likes to stay active ?

Table 3.2: Example of Generated Sentences

sentences for each emotion, we used a binary BERT [6] classifier. For training BERT, we used Ekman Emotion dataset [2] and trained it to determine whether a given sentence matched each specific emotion. In the figure, the lightly shaded areas represent instances where emotion control was not applied, while the darkly shaded areas indicate instances where emotion control was applied. Since LLMs are typically tuned during pre-training to avoid generating harmful or negative sentences, the accuracy for emotions such as 'Anger,' 'Sadness,' 'Disgust,' and 'Fear' was significantly lower when emotion control was not applied. However, when emotion control was applied, all models except GPT-2-large and DialoGPT-2-large showed an increase in accuracy, indicating that the emotions were effectively controlled to some extent through zero-shot prompt control. Smaller models like GPT-2, as observed in previous research [11], struggle to understand human-provided instructions, resulting in significantly lower accuracy.

We also performed multi-class classification accuracy assessment on the generated sentences for the six emotions. Similarly, BERT was trained on the six emotions and used for evaluation. The results can be seen in Table 3.3. The column 'General' shows the average accuracy of binary classification when emotion control was not applied, while the column 'Binary' shows the average accuracy of binary classification when emotion control

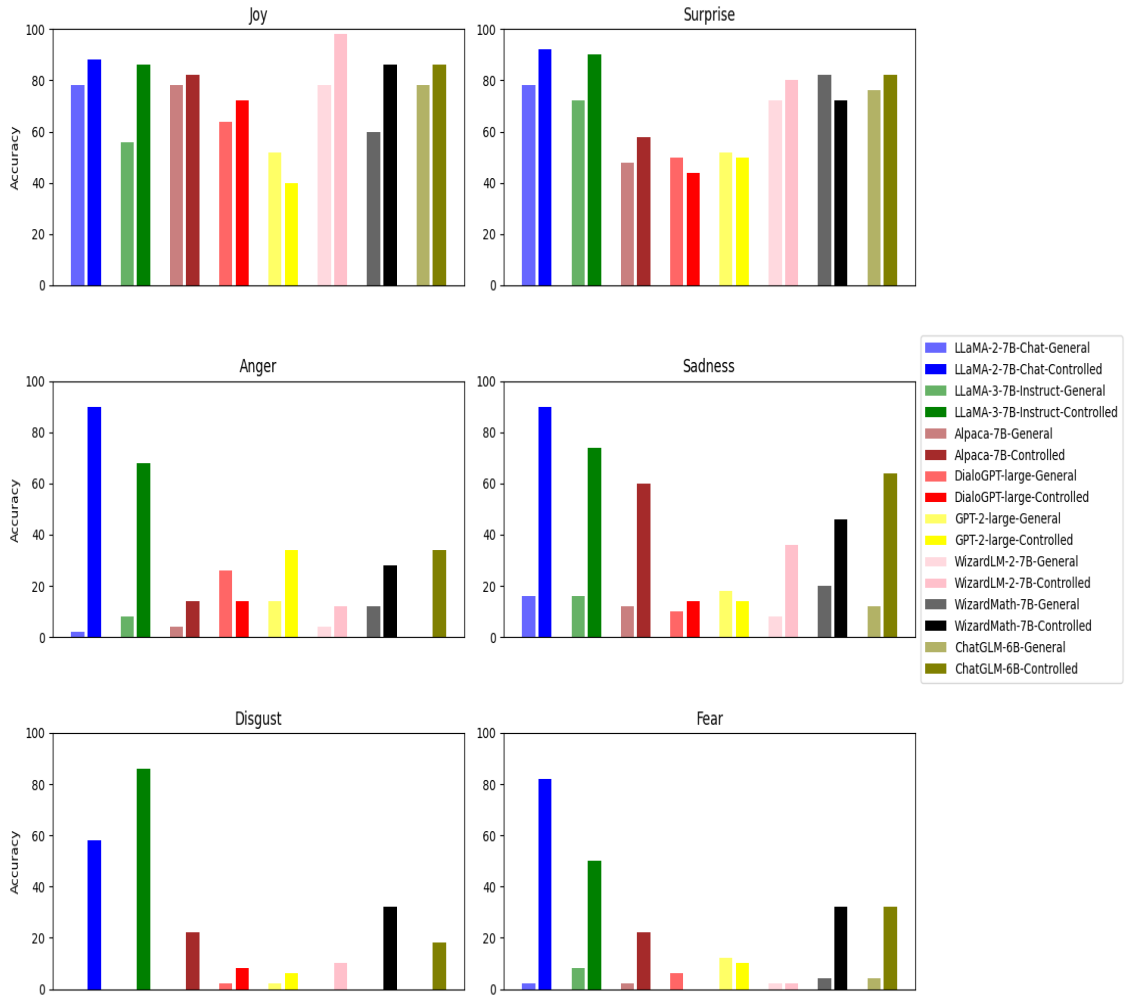


Figure 3.3: Accuracy per Emotion

was applied. Additionally, the column 'Multi' represents the multi-class classification accuracy.

There are differences in accuracy among the various models, specifically in their ability to recognize and adhere to human-provided emotion instruc-

Model/Accuracy	General	Binary	Multi
LLaMA-2-7B-Chat	29.33	83.33	73.00
LLaMA-3-7B-Instruct	26.67	75.67	55.00
ChatGLM-6B	28.33	52.67	44.00
WizardLM-2-7B	27.33	39.67	37.33
Alpaca-7B	24.00	43.00	37.00
WizardMath-7B	29.67	49.33	37.00
DialoGPT-large	26.33	25.33	18.67
GPT-2-large	25.00	25.67	16.33

Table 3.3: Emotion Accuracy per Model

tions. In the following section, we explore the reasons behind these discrepancies.



## Chapter 4. Effects of Emotion Control Instruction on Dialogue Contexts

In this chapter, we explored the effects of emotion control instruction on the dialogue contexts and control ability through three experiments below.

Firstly, we compare a layer-wise similarity between the hidden states of the *controlled context* and ones of the *general context*. The controlled contexts represent the dialogue context part of controlled prompts, and general contexts mean the dialogue context part of the prompt which have no control instruction. This similarity can show how much the emotion control instruction affect the context’s hidden states. Then we observed patterns in how the similarity changed across layers as they progressed.

Secondly, we visualized at which layer the model effectively distinguishes emotions by plotting the hidden states of the dialogue context in a 2-dimensional space using Principal Component analysis(PCA) [10] method when the LLM is controlled by each emotion instruction.

Lastly, through probing task using the hidden states extracted at every layers of LLMs, we quantified at which layer each model effectively distinguishes emotions.

## 4.1 Cosine Similarity of Hidden states of Dialogue Context with and without Emotion Control

To compare the hidden states of the dialogue context when specific emotion instructions were provided and when they were not, we measured cosine similarity. The comparison between the two scenarios can be understood by referring to Figure 4.1 and the equations below.

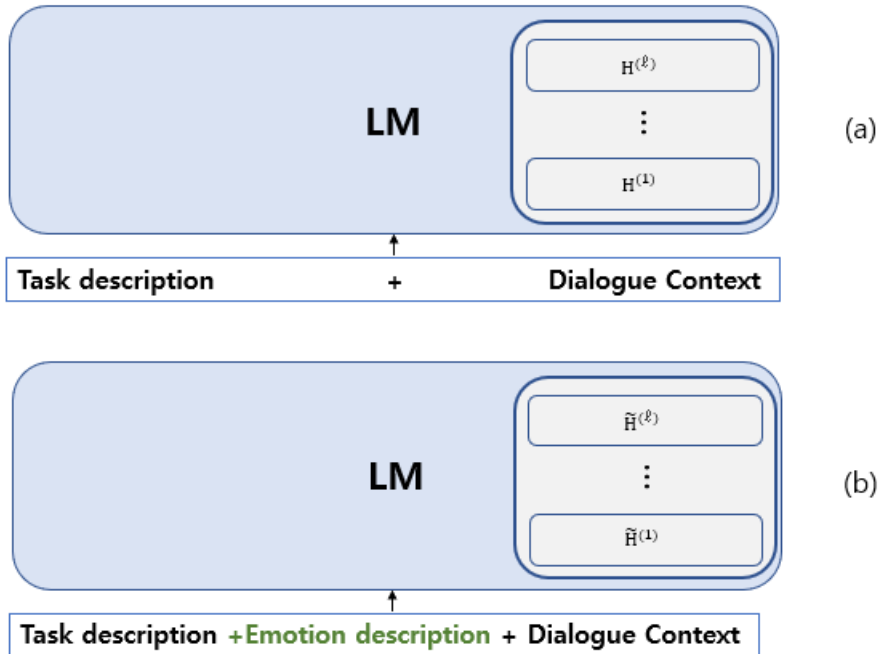


Figure 4.1: Hidden States of Dialogue Context in two situations

Let  $x$  be a model’s input prompt which consists of the *task description*  $TD$  and the *dialogue context*  $CON$ . In Figure 4.1 (a),  $x$  can be seen as the *general prompt* without any control instruction. Let define another input dialogue prompt  $\tilde{x}$  composing the *task description*  $TD$ , the *Emotion description*  $ED$ , and the *dialogue context*  $CON$ . In Figure 4.1 (b), in contrast to  $x$ ,  $\tilde{x}$  represents the *controlled prompt*.

$$x = [TD; CON], \quad (4.1)$$

$$\tilde{x} = [TD; ED; CON], \quad (4.2)$$

where ‘;’ represents a concatenation operator. Note that, for the one dialogue data, the  $TD$  and  $CON$  of  $\tilde{x}$  is exactly same as ones of  $x$  for easy comparison. To explore similarity of  $x$  and  $\tilde{x}$ , we take the hidden states for the dialogue context tokens for computing similarity.

Let  $H^{(l)}$ ,  $\tilde{H}^{(l)}$  be the hidden states of the context part of general prompt and controlled prompt, respectively, at the  $l$ -th layer of a LLM.  $H^{(l)}$  and  $\tilde{H}^{(l)}$  can be formulated as follows:

$$H^{(l)} = DL^{(l)}(x)_{[|TD|:|TD|+|CON|]}, \quad (4.3)$$

$$\tilde{H}^{(l)} = DL^{(l)}(\tilde{x})_{[|TD|+|ED|:|TD|+|ED|+|CON|]}. \quad (4.4)$$

where 'DL' represents decoder layer of model. We calculate the similarity of two hidden states  $H^{(l)}$  and  $\tilde{H}^{(l)}$  using averaged cosine similarity as follows:

$$AC^{(l)} = \frac{1}{K} \sum_{k=0}^K \cos(H_k^{(l)}, \tilde{H}_k^{(l)}), \quad (4.5)$$

where  $H_k^{(l)}$  and  $\tilde{H}_k^{(l)}$  represent the hidden state of the  $k$ -th context token, at the  $l$ -th layer. Subsequently, we average the  $AC^{(l)}$  across the all dialogue data.

The results for each model can be found in Figure 4.2. Except for GPT-2 and DialoGPT, other models exhibited similar patterns of cosine similarity changes in the early layers as observed when the models were fine-tuned [16]. While previous studies showed a gradual decrease in cosine similarity as layers deepened due to parameter updates by fine-tuning, we achieved similar results without changing parameters. Simply by adding emotion-controlling instructions, we concluded that the models internally encode emotion attributes to dialogue context in a manner similar to previous research. We confirmed that the decrease in cosine similarity as the layers deepened corresponds to the segment where emotion information is encoded in the context.

Additionally, Figure 4.3 presents the averaged results for the six emotions per model, derived from the data depicted in Figure 4.2. We observed

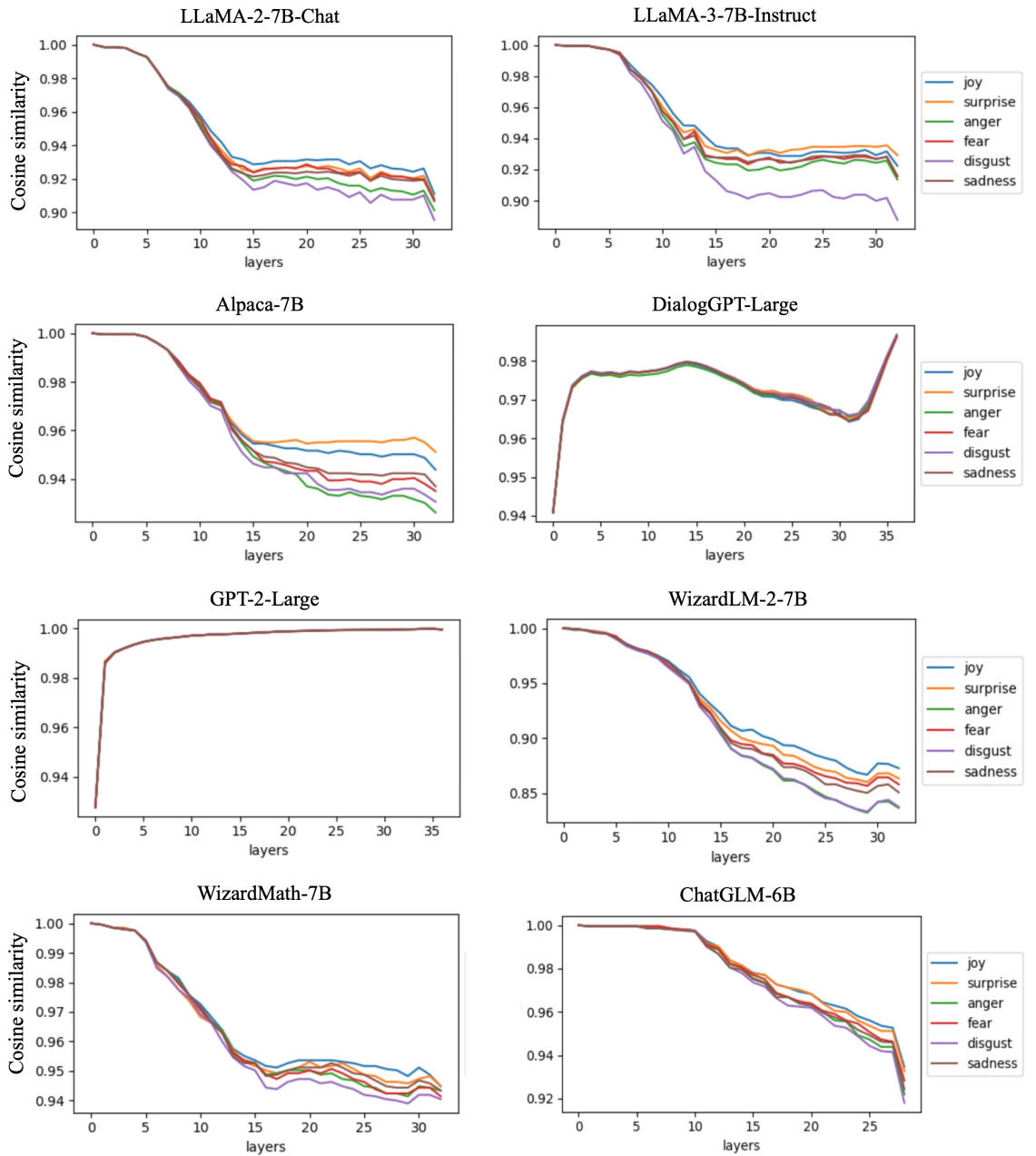


Figure 4.2: Cosine Similarity results per Model

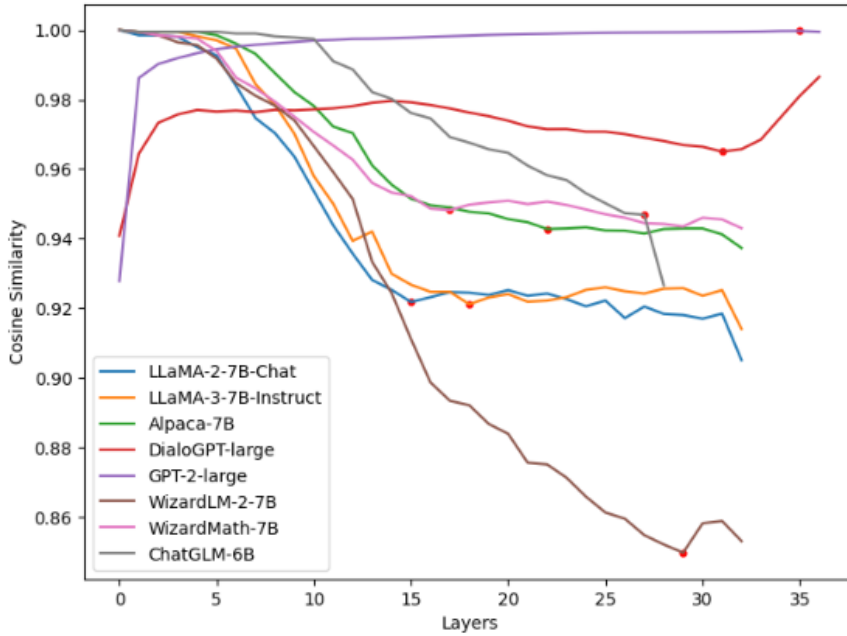


Figure 4.3: Average of Cosine similarity and Plateau point of models

the difference between the hidden states in the two situations varies in a certain segment and then converges at some point. This indicates that the segments where emotions are encoded differ for each model. We identified the 'Plateau' points where the decrease in cosine similarity slows down and these points are marked with red dots in the graph. GPT-2, being an exception, does not exhibit a distinct decrease in cosine similarity, so it is marked at the last layer. Similarly, while the cosine similarity for DialoGPT slightly decreases towards the end, the change is very small, indicating that the difference in hidden states between the two scenarios is minimal and thus it does not encode emotions effectively.

Furthermore, we proceeded to analyze the correlation between these plateau points and the accuracy of emotion control. As illustrated in Fig-

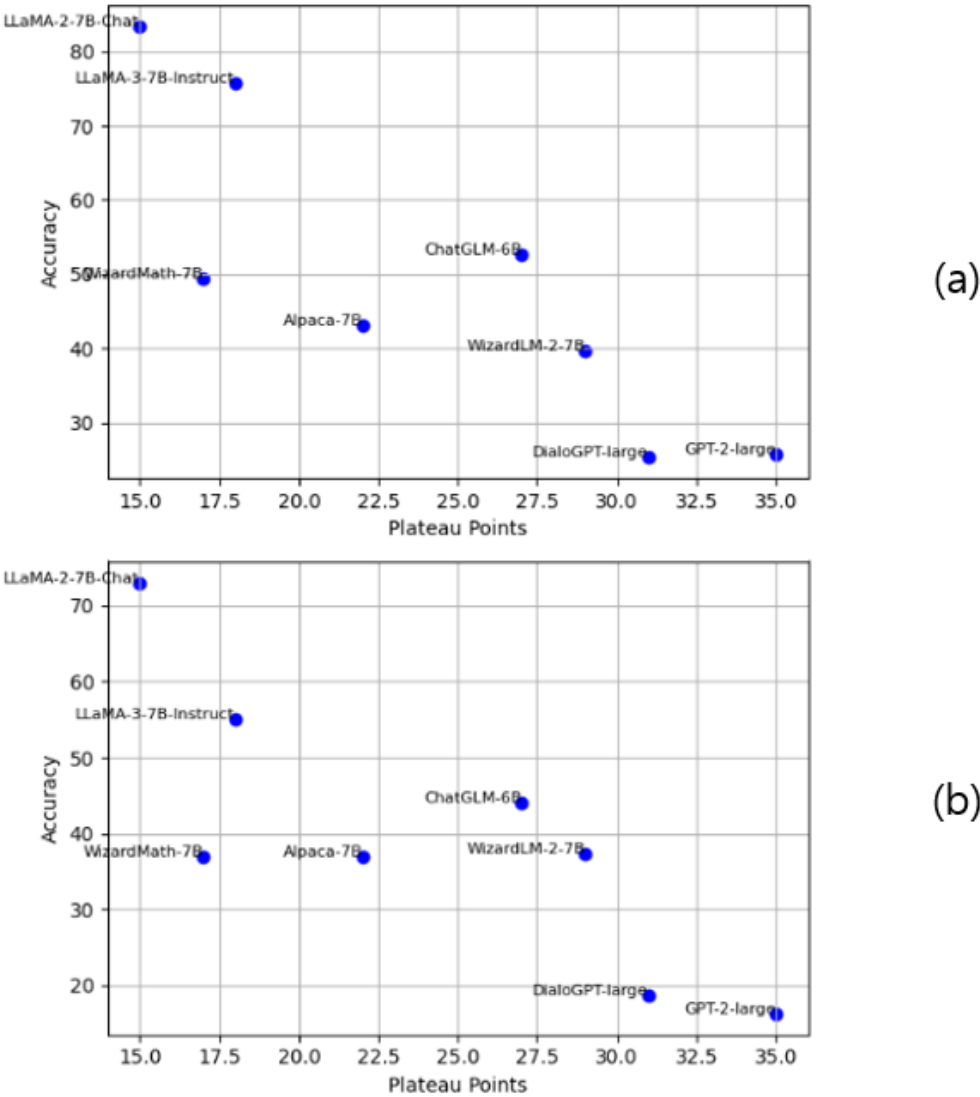


Figure 4.4: Correlation between Plateau points and the Accuracy of emotion control

ure 4.4, we observed a trend where a faster plateau point, indicating a quicker encoding of emotional information into the dialogue context, correlates with higher accuracy in controlling the LLMs. Figure 4.4 (a) represents the results of binary classification and Figure 4.4 (b) represents the results of multi classification that we measured in chapter 3.

## 4.2 2D Visualization through PCA

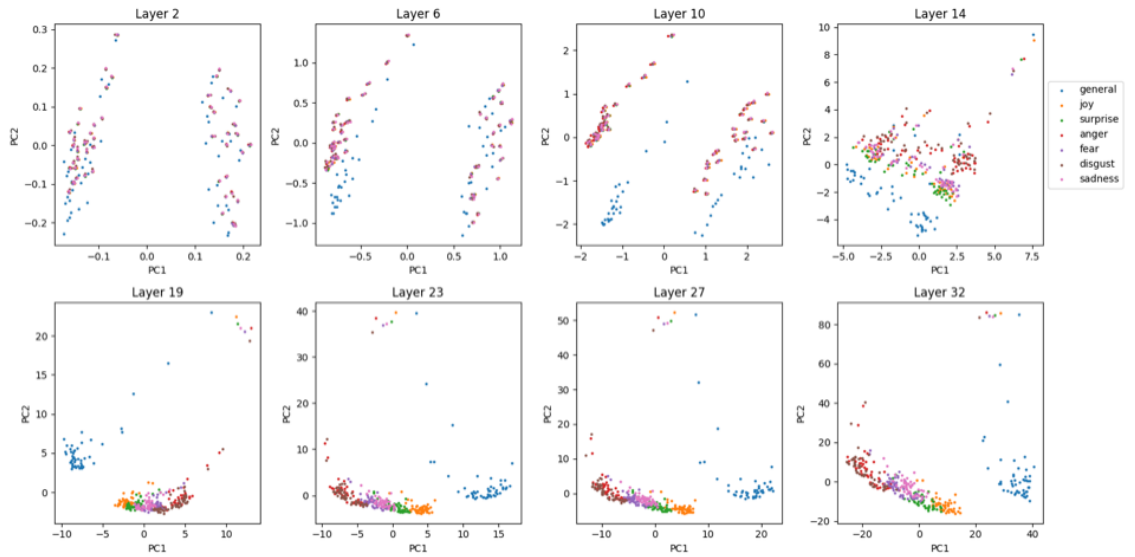
Until now, we demonstrate our experiments for exploring the difference caused by the control instruction, utilizing cosine similarity. However, only seeing cosine similarity cannot provide enough information the difference is really due to the control instruction. We complement this by visualizing  $H^{(l)}$  and  $\tilde{H}^{(l)}$  at the low dimension (2-dimension) using PCA method. Figure 4.5, Figure 4.6, Figure 4.7, Figure 4.8 shows each results of models.

In the initial layers, emotion differentiation was not well pronounced, but it improved gradually as the layers deepened. From the plateau point onward, which signifies the moment when emotion encoding is completed, emotion differentiation appeared to be well maintained up to the final layers. There seemed to be a correlation between the plateau point and the point at which emotions began to be visually distinguished. GPT-2 and DialoGPT appeared to have difficulty in distinguishing emotions even as the



model's layers deepened, compared to other models.

### LLaMA-2-7B-Chat



### LLaMA-3-7B-Instruct

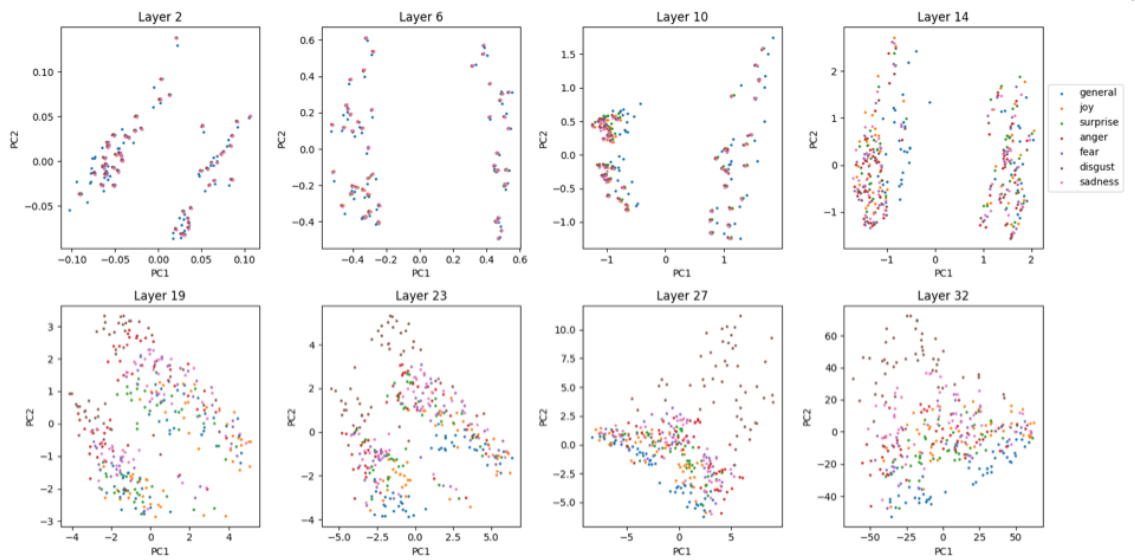
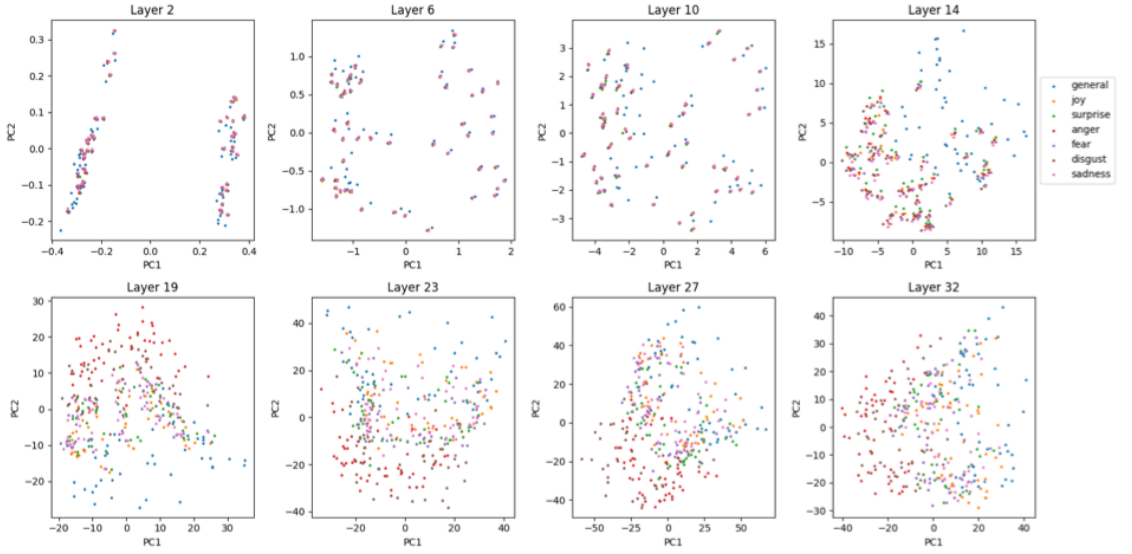


Figure 4.5: PCA on Hidden states produced by LLaMA

## Alpaca-7B



## DialoGPT-Large

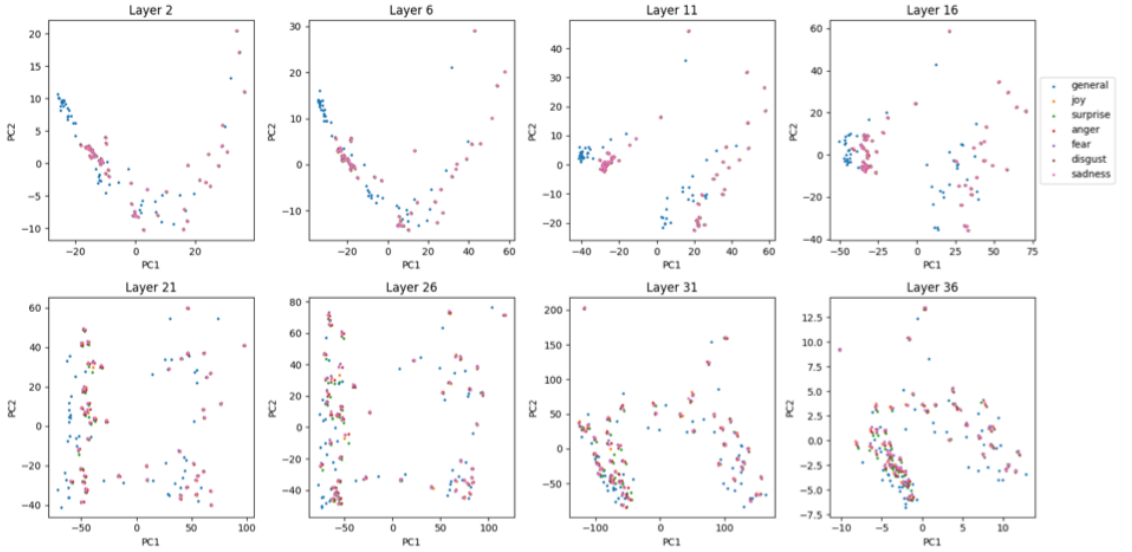
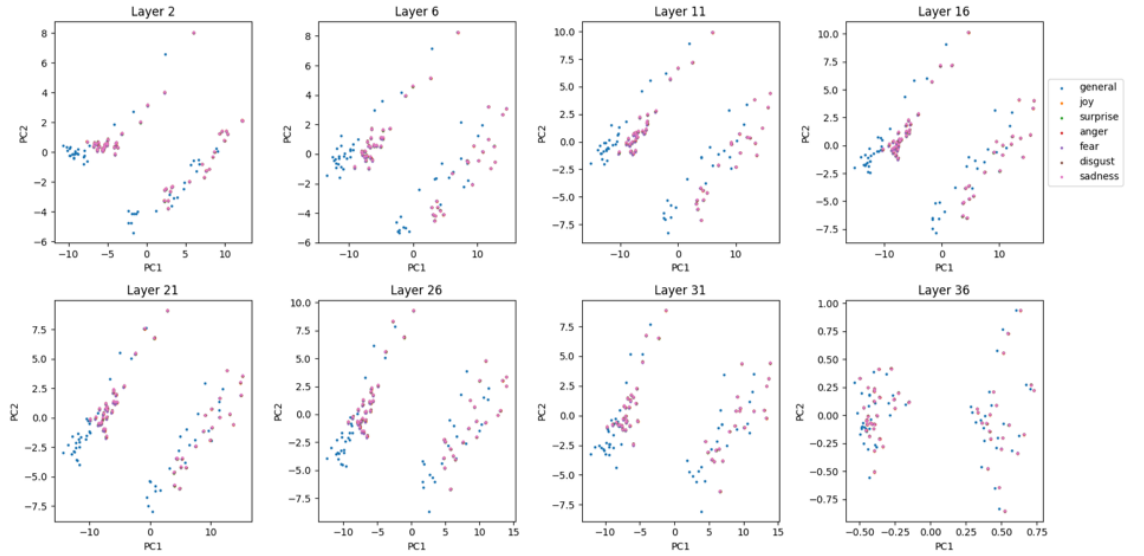


Figure 4.6: PCA on Hidden states produced by Alpaca and DialoGPT

## GPT-2-Large



## WizardLM-2-7B

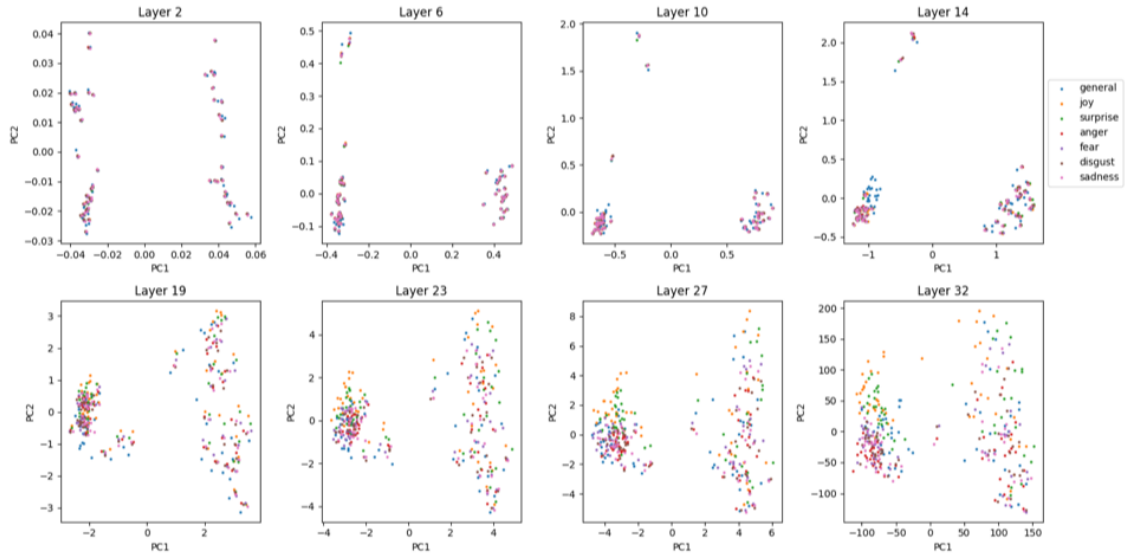
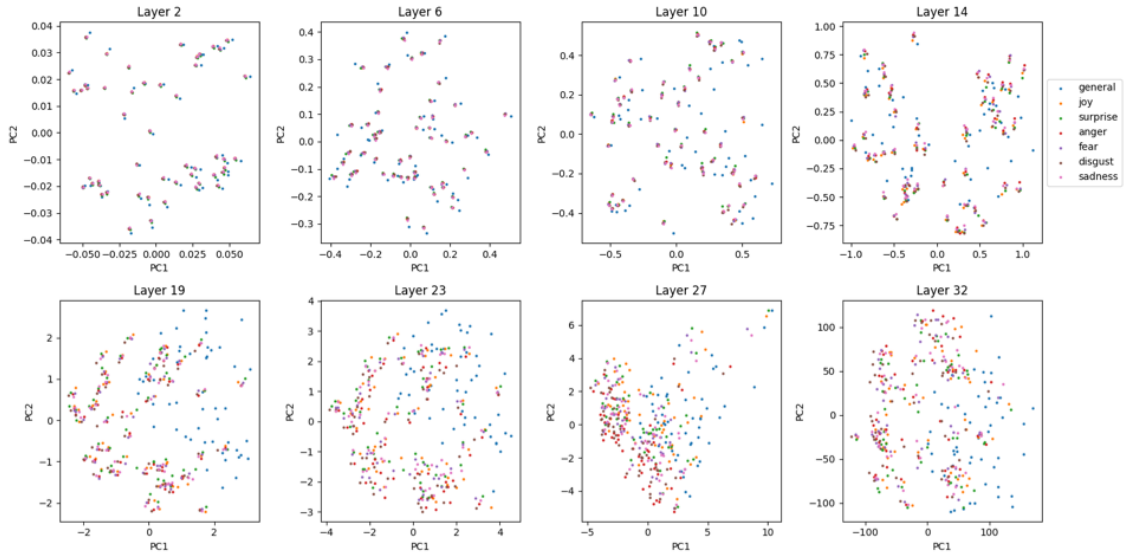


Figure 4.7: PCA on Hidden states produced by GPT2 and WizardLM

## WizardMath-7B



## ChatGLM-6B

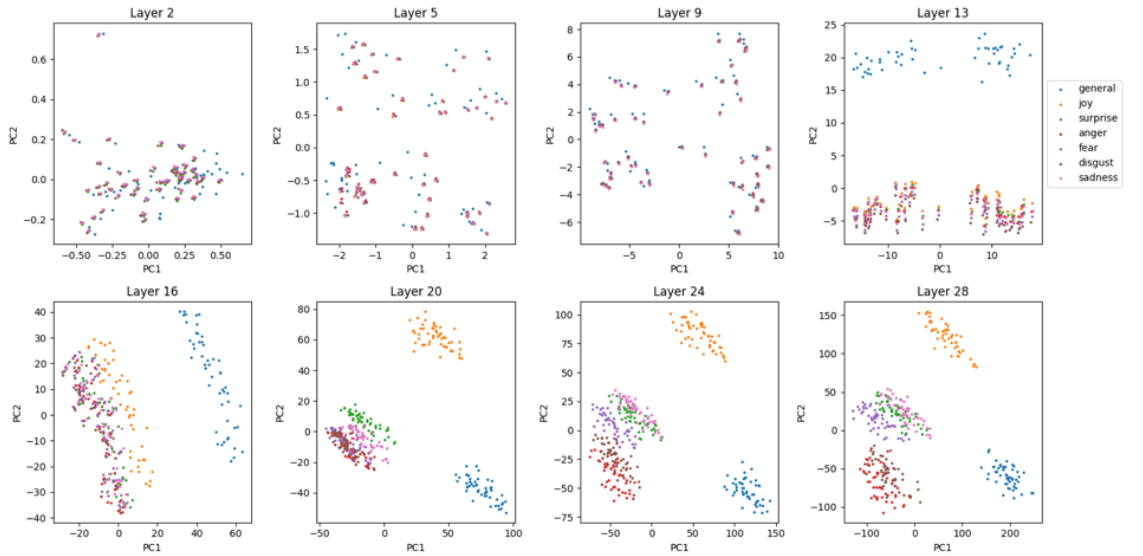


Figure 4.8: PCA on Hidden states produced by WizardMath and ChatGLM

However, since this was an experiment conducted by reducing vectors of several thousand dimensions to two dimensions, we conducted further probing experiment in the following section to obtain more accurate results.

### 4.3 Probing with Logistic Regression Classifier

PCA gives us intuitive information of how much control information the hidden states involve. For more formal evaluation, we quantified the control information involved in the hidden states utilizing probing classifiers. We trained simple emotion attribute classifiers(Logistic Regression [5]) at hidden state space to explore the hidden states are well discriminated by attributes. If those classifier achieves high accuracy, those hidden states are well separated given the emotion attribute. The input to the classifier consists of the hidden states of the dialog context from each layer when the LLMs is under emotion control by emotion instructions, and the label corresponds to the emotion being controlled.

Figure 4.9 illustrates the accuracy at every layers, using 8 LLMs. As it shows, while it can be seen that the hidden states at the initial layers have not enough information about the sentiment control instruction, after a few layers, except for GPT-2, all models showed that the logistic regression accuracy converged to 100% as the layers deepened. The convergence points, that is,

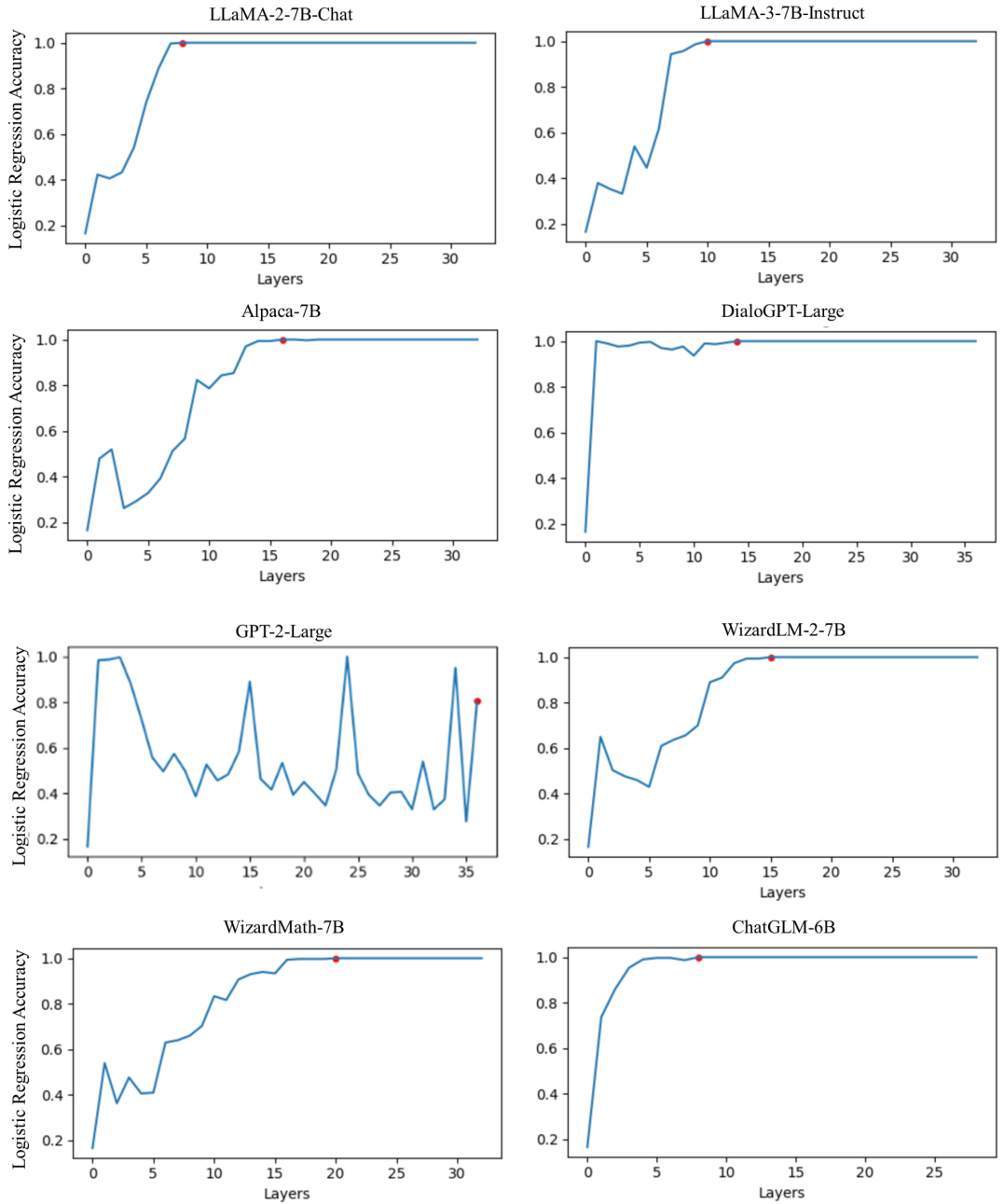


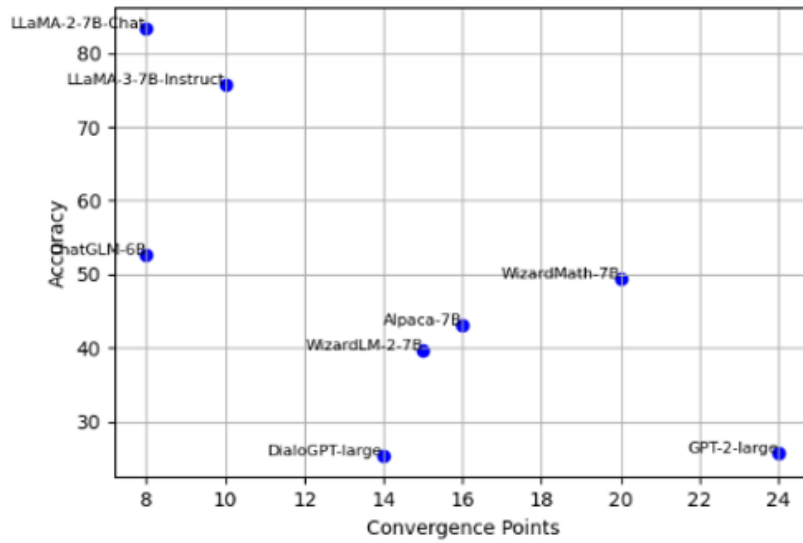
Figure 4.9: Logistic Regression Accuracy of Models, at every layers.

the points at which each model started to distinguish emotions well, varied across models, and we marked these points with red dots in the graph. Exceptionally, GPT-2 did not converge, so we marked it at the final layer.

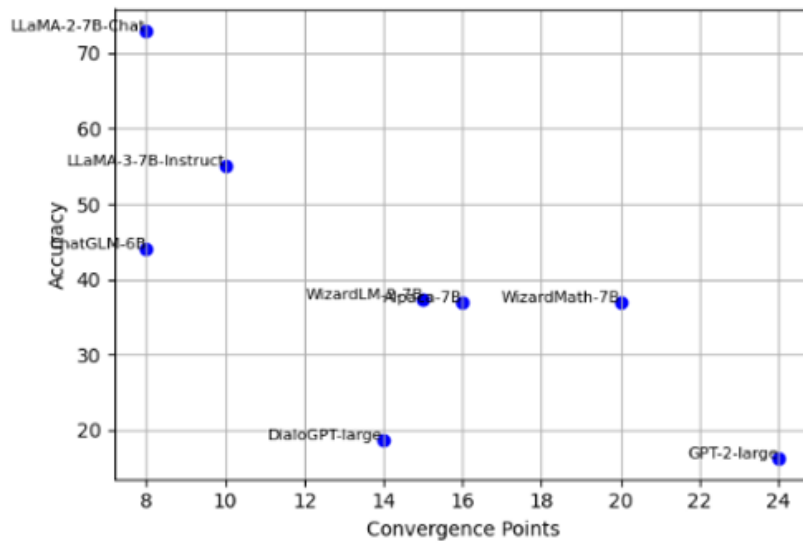
Subsequently, we examined the correlation between emotion accuracy and these convergence points. As shown in Figure 4.10, we found that the earlier the classifier’s convergence point, the higher the emotion accuracy of the sentences generated by the LLMs. Figure 4.10 (a) represents the results of binary classification and Figure 4.10 (b) represents the results of multi classification that we measured in chapter 3.

## 4.4 Discussion

All the experiments conducted so far can be summarized in Table 4.1. The ‘PP’ column represents the plateau point of cosine similarity pattern, and ‘CP(LR)’ represents the convergence point of Logistic Regression. The ‘Binary’ and ‘Multi’ columns indicate the emotion accuracy measured in Chapter 3. From the results of the first experiment, by examining the ‘PP’ column and the two accuracy columns, it was observed that the faster the convergence in the region encoding emotion description into dialogue context, the higher the model’s emotion control ability. Similarly, from the results of the third experiment, by examining the ‘CP(LR)’ column and the two



(a)



(b)

Figure 4.10: Correlation between Convergence points and the Accuracy of emotion control



accuracy columns, it was found that the earlier the model’s ability to distinguish emotions in lower layers, the higher the model’s emotion control ability. GPT-2 and DialoGPT-2, being smaller in size compared to other models and not well-trained to understand instructions, showed lower emotion accuracy and did not appear to encode emotions well or distinguish emotions effectively as layers deepened.

Model/Accuracy	Binary	Multi	PP	CP(LR)
LLaMA-2-7B-Chat	83.33	73.00	15	8
LLaMA-3-7B-Instruct	75.67	55.00	18	10
ChatGLM-6B	52.67	44.00	27	8
WizardLM-2-7B	39.67	37.33	29	15
Alpaca-7B	43.00	37.00	22	16
WizardMath-7B	49.33	37.00	17	20
DialoGPT-large	25.33	18.67	31	14
GPT-2-large	25.67	16.33	35	24

Table 4.1: Emotion Accuracy, Plateau points, Convergence points

## Chapter 5. Conclusions

The models used in this study are known to perform well across various tasks in the field of natural language generation. However, it is still unclear how the prompts fed into these models function internally. Therefore, in this study, we investigated how instructions indicating specific attribute impact the context in a controlled dialogue generation task. By conducting three experiments using the hidden states output at each layer, we discovered the following two key points:

First, during zero-shot learning, we provided different prompts for scenarios with and without emotion control and measured the cosine similarity of the dialogue context. The results showed that, in the early layers, the model behaved as if it had been fine-tuned internally, with the cosine similarity gradually decreasing. This pattern reached a specific plateau point, where the decrease slowed down as the layers deepened. By examining this pattern and the PCA results from the second experiment, we found that each model had specific regions of layer where emotions were well-encoded, with each model converging at different points.

Second, considering the results of the first experiment, the third probing

experiment, and the emotion control accuracy, we found that the earlier the emotion is encoded into the dialogue context, i.e., in the lower layers, and the quicker the emotion is distinguished, the higher the emotion accuracy of the generated text.

Additionally, the future work includes the following: First, further generalization through experiments with a greater number of models is needed. Second, just as we identified when and how emotions are embedded into the context, experiments can be conducted on various control attributes. Through this, we expect to observe how the behavior of LLMs varies internally depending on the level of control attributes.

We hope this study will help general users understand the results of language model better and assist developers in advancing models.

## Reference

- [1] <https://github.com/meta-llama/llama3>.
- [2] <https://github.com/Valendrew/ekman-emotion-detection>.
- [3] Ebtesam Almazrouei *et al.* *The Falcon Series of Open Language Models*. 2023. arXiv: 2311.16867 [cs.CL].
- [4] Tom B. Brown *et al.* *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL].
- [5] David R Cox. “The regression analysis of binary sequences”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 20.2 (1958), pp. 215–232.
- [6] Jacob Devlin *et al.* *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].
- [7] Qingxiu Dong *et al.* *A Survey on In-context Learning*. 2023. arXiv: 2301.00234 [cs.CL].
- [8] Zhengxiao Du *et al.* *GLM: General Language Model Pretraining with Autoregressive Blank Infilling*. 2022. arXiv: 2103.10360 [cs.CL].

- [9] Paul Ekman. “Basic Emotions”. In: *Handbook of Cognition and Emotion*. Ed. by Tim Dalgleish and M. J. Powers. Wiley, 1999, pp. 4–5.
- [10] Karl Pearson F.R.S. “LIII. On lines and planes of closest fit to systems of points in space”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572. DOI: 10.1080/14786440109462720.
- [11] Mauajama Firdaus *et al.* “Being Polite: Modeling Politeness Variation in a Personalized Dialog Agent”. In: *IEEE Transactions on Computational Social Systems* 10.4 (2023), pp. 1455–1464. DOI: 10.1109/TCSS.2022.3182986.
- [12] Mauajama Firdaus *et al.* “EmoSen: Generating Sentiment and Emotion Controlled Responses in a Multimodal Dialogue System”. In: *IEEE Transactions on Affective Computing* 13.3 (2022), pp. 1555–1566. DOI: 10.1109/TAFFC.2020.3015491.
- [13] Arnav Gudibande *et al.* *The False Promise of Imitating Proprietary LLMs*. 2023. arXiv: 2305.15717 [cs.CL].
- [14] Yanran Li *et al.* “DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset”. In: *Proceedings of The 8th International Joint Conference on Natural Language Processing (IJCNLP 2017)*. 2017.

- [15] Haipeng Luo *et al.* *WizardMath: Empowering Mathematical Reasoning for Large Language Models via Reinforced Evol-Instruct*. 2023. arXiv: 2308.09583 [cs.CL].
- [16] Amil Merchant *et al.* *What Happens To BERT Embeddings During Fine-tuning?* 2020. arXiv: 2004.14448 [cs.CL].
- [17] Shervin Minaee *et al.* *Large Language Models: A Survey*. 2024. arXiv: 2402.06196 [cs.CL].
- [18] OpenAI *et al.* *GPT-4 Technical Report*. 2024. arXiv: 2303.08774 [cs.CL].
- [19] Alec Radford *et al.* “Language Models are Unsupervised Multitask Learners”. In: 2019. URL: <https://api.semanticscholar.org/CorpusID:160025533>.
- [20] Yu-Ping Ruan and Zhen-Hua Ling. “Emotion-Regularized Conditional Variational Autoencoder for Emotional Response Generation”. In: *IEEE Transactions on Affective Computing* 14.1 (Jan. 2023), pp. 842–848. ISSN: 2371-9850. DOI: 10.1109/taffc.2021.3073809. URL: <http://dx.doi.org/10.1109/TAFRC.2021.3073809>.
- [21] Haoyu Song *et al.* “BoB: BERT Over BERT for Training Persona-based Dialogue Models from Limited Personalized Data”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language*

- Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong *et al.* Online: Association for Computational Linguistics, Aug. 2021, pp. 167–177. DOI: 10.18653/v1/2021.acl-long.14. URL: <https://aclanthology.org/2021.acl-long.14>.
- [22] Haoyu Song *et al.* “Generate, Delete and Rewrite: A Three-Stage Framework for Improving Persona Consistency of Dialogue Generation”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky *et al.* Online: Association for Computational Linguistics, July 2020, pp. 5821–5831. DOI: 10.18653/v1/2020.acl-main.516. URL: <https://aclanthology.org/2020.acl-main.516>.
- [23] Rohan Taori *et al.* “Alpaca: A Strong, Replicable Instruction-following Model”. In: *Stanford Center for Research on Foundation Models 3.6* (2023), p. 7. URL: <https://crfm.stanford.edu/2023/03/13/alpaca.html>.
- [24] Hugo Touvron *et al.* *Llama 2: Open Foundation and Fine-Tuned Chat Models*. 2023. arXiv: 2307.09288 [cs.CL].
- [25] Jerry Wei *et al.* *Larger language models do in-context learning differently*. 2023. arXiv: 2303.03846 [cs.CL].

- [26] Can Xu *et al.* *WizardLM: Empowering Large Language Models to Follow Complex Instructions*. 2023. arXiv: 2304.12244 [cs.CL].
- [27] Shukang Yin *et al.* *A Survey on Multimodal Large Language Models*. 2024. arXiv: 2306.13549 [cs.CV].
- [28] Hanqing Zhang *et al.* *A Survey of Controllable Text Generation using Transformer-based Pre-trained Language Models*. 2023. arXiv: 2201.05337 [cs.CL].
- [29] Yizhe Zhang *et al.* *DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation*. 2020. arXiv: 1911.00536 [cs.CL].
- [30] Wayne Xin Zhao *et al.* *A Survey of Large Language Models*. 2023. arXiv: 2303.18223 [cs.CL].
- [31] Chunting Zhou *et al.* *LIMA: Less Is More for Alignment*. 2023. arXiv: 2305.11206 [cs.CL].



## 국문요지

거대 언어 모델들은 대화 생성 작업과 같은 다양한 작업에서 뛰어난 성능을 보이고 있으며, 문맥 속 학습 능력도 높지만, 이러한 모델이 결과를 어떻게 달성하는지를 이해하는 것은 여전히 복잡한 블랙박스 구조 때문에 도전적이다. 본 연구는 특정 속성을 반영하도록 지시하는 프롬프트가 대화 생성 작업 내에서 대화 문맥에 미치는 영향을 분석했다. 우리는 감정을 제어하는 지시 프롬프트가 제공되었는지 여부에 따라 상황을 나누어 실험을 실시했다. 궁극적으로 우리는 **layer** 관점에서 모델이 두 가지 상황에 대해서 어떤 양상을 보이는지 알아내고자 했다. 우리는 이 두 가지 다른 시나리오에서, 대화 문맥으로부터의 임베딩 벡터를 세 가지 실험을 통해 비교했다. 실험 결과는 감정을 표시하는 지시 프롬프트가 각 계층에서 대화 문맥에 어떻게 영향을 미치는지, 그리고 최종 모델의 생성된 출력에 어떻게 영향을 미치는지를 밝혀냈다.

결론적으로, 우리는 모델의 입력에 감정과 같은 특정 속성을 제어하는 지시를 포함시킬 때, 모델이 계층적인 관점에서 내부적으로 미세조정 되었을 때와 비슷한 행동을 보인다는 것을 발견했다. 또한 모델이 감정을 낮은 계층에서 인코딩할수록 원하는 감정을 제어하는 능력이 증가한다는 것을 발견했다.