This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

LLaMA-Excitor: General Instruction Tuning via Indirect Feature Interaction

Bo Zou *
Tsinghua University
Beijing, China
zoub21@mails.tsinghua.edu.cn

Chao Yang *
Shanghai AI Laboratory
Shanghai, China
yangchao@pjlab.org.cn

Yu Qiao
Shanghai AI Laboratory
Shanghai, China
qiaoyu@pjlab.org.cn

Chengbin Quan
Tsinghua University
Beijing, China
quancb@tsinghua.edu.cn

Youjian Zhao †
Tsinghua University
Zhongguancun Laboratory
Beijing, China
zhaoyoujian@tsinghua.edu.cn

Abstract

Existing methods to fine-tune LLMs, like Adapter, Prefix-tuning, and LoRA, which introduce extra modules or additional input sequences to inject new skills or knowledge, may compromise the innate abilities of LLMs. In this paper, we propose LLaMA-Excitor, a lightweight method that stimulates the LLMs' potential to better follow instructions by gradually paying more attention to worthwhile information. Specifically, LLaMA-Excitor does not directly change the intermediate hidden state during the self-attention calculation. We designed the Excitor block as a bypass module that reconstructs Keys and changes the importance of Values in self-attention using learnable prompts. LLaMA-Excitor ensures a self-adaptive allocation of additional attention to input instructions, thus effectively preserving LLMs' pre-trained knowledge when fine-tuning LLMs on low-quality instruction-following datasets. Furthermore, we unify the modeling of multi-modal and language-only tuning, extending LLaMA-Excitor to a powerful visual instruction follower without the need for complex multi-modal alignment. Our approach is evaluated in language-only and multi-modal scenarios. Compared with the original LLaMA-7B, LLaMA-Excitor is the only PEFT method that maintains basic capabilities and achieves +3.12% relative improvement on the MMLU benchmark. In the visual instruction tuning, we achieve a new state-of-the-art image captioning performance on MSCOCO (157.5 CIDEr), and a comparable performance on ScienceQA (88.39%) to cutting-edge models with more parameters and extensive vision-language pertaining. The code will be available at https://zoubo9034.github.io/Excitor/.

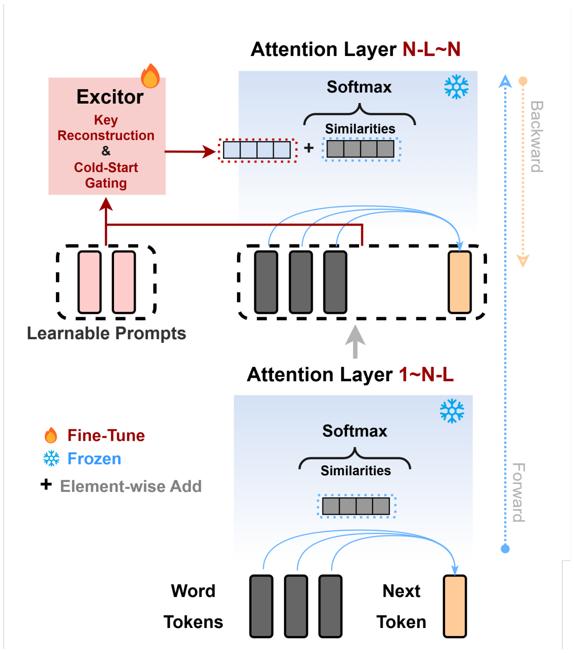


Figure 1. Overview of LLaMA-Excitor. We integrate Excitor blocks into L out of N attention layers of LLaMA. Differing from previous PEFT techniques, LLaMA-Excitor indirectly involves learnable information in the reasoning process by changing the similarity matrices. It ensures that the hidden states are within the original distribution of LLaMA.

1. Introduction

Large language models (LLMs) are highly advanced tools that enable contextual information processing, understanding, and generation. They serve as powerful knowledge bases that can be used to gain valuable insights and create new content [9, 10, 37, 39, 52]. Fine-tuning them for downstream tasks has also achieved great success in vari-

---

* Equal contribution, † Corresponding author
This work was done during an internship at Shanghai AI Lab.

ous domains. With the rapid growth in scale and quality of the corpus and the advance in hardware, the inherent abilities of LLMs are becoming stronger. Therefore, the primary objective of fine-tuning general-purpose LLMs should be ensuring they produce the desired output for downstream tasks (i.e., instruction-following ability) rather than memory excessive knowledge or mastering skills (e.g., causal inference and numerical computation).

Prioritizing the enhancement of instruction-following capabilities for LLMs is prudent for a multitude of reasons. Firstly, it is often presumed that the corpus utilized during training encompasses a comprehensive repository of knowledge, thereby obviating the need for additional fine-tuning datasets when addressing specific downstream tasks. This implies that enriching a model's ability to follow instructions could leverage this pre-existing knowledge base more effectively. Secondly, by standardizing the output format during instruction-following tasks, we can reduce the generation of harmful or irrelevant content. Third, the challenge of incorporating new knowledge and skills into LLMs is accentuated by the limited set of parameters available for fine-tuning. Lastly, fine-tuning in the context of multi-modal tasks can be succinctly expressed as enhancing a model's proficiency in responding to visual cues. This essentially constitutes a variant of instruction following, where the prompts are visual rather than textual.

However, predominant parameter-efficient fine-tuning (PEFT) techniques are met with notable challenges in performing instruction-tuning. The Adapter methods[14, 36], which have a serial structure, can cause computation bottlenecks, and they also significantly alter the inference process to adapt to downstream tasks, which in turn can lead to degradation of inherent abilities, such as catastrophic forgetting. Prompt learning methods[5, 26, 29, 41] effectively incorporate new information by creating additional parameters corresponding to the input sequences of language models. However, concentrating all necessary knowledge into fixed-length token sequences is challenging for various downstream tasks. Furthermore, as the length of the generated sequence increases, the control over the sequence diminishes since LLMs generate the next word based on a softmax over the entire sequence. LoRA methods[15, 50] directly add the outcomes of randomly initialized low-rank modules with linear layers' outputs, which can introduce features that are out of LLMs' feature distribution and may cause degradations. Recent studies [16, 50] also show that Adapter and LoRA evenly assign learnable parameters to trainable modules and neglect the functionality gap between each layer in LLM, which will lead to performance drops.

In this paper, we start from a new perspective and propose LLaMA-Excitor, a PEFT method that focuses on the following instructions. (1) LLaMA-Excitor aims to optimize instruction-following ability by releasing the potential of an LLM, i.e., LLaMA [46], instead of pursuing new knowledge and skills. Due to variations in the data scale, quality, and content coverage of instruction-tuning sets, the effect of fine-tuning processes on the inherent skills of LLMs is unpredictable. (2) LLaMA-Excitor can reduce the degradation when fine-tuning on unsuited datasets. Specifically, rather than directly changing the intermediate hidden state of pretrained LLMs like Adapter and LoRA, LLaMA-Excitor uses trainable bypass modules (Excitor blocks) to modify the attention score meanwhile maintain original $Values$ for each attention layer in LLaMA. Excitor blocks only adjust the proportion of information in $Values$ for hidden states, ensuring that each attention layer's input is derived from a linear combination of the previous layer's $Values$ and corresponds to the original LLaMA's distribution (an indirect feature interaction process). Additionally, the modification of the attention score depends on both the input sequence and a set of prompt tokens to make the model adaptively decide whether to rely on additional information from the fine-tuning set for attention allocation. These attributes ensure fast training and few forgetings.

Moreover, Exictor innovatively provides a low-budget way to fine-tune a language-only model into a vision-language model. Previous works (e.g., [3, 28, 31, 47, 51]) rely on training an additional visual branch or projection module to align vision and language, which incurs significant computational overhead. We explore the unified modeling of instruction-tuning and visual instruction-tuning for LLMs and utilize a visual Exictor to make LLaMA follow the visual prompts generated from a frozen image encoder like CLIP [38]. In this way, we eliminate the need for extra vision-language alignment while preserving textual features' purity during LLMs' reasoning and shortening the training time. Our contributions are summarized as follows:

- We study indirect feature interaction in fine-tuning LLMs and propose LLaMA-Excitor, a PEFT method that focuses on instruction-following and reduces forgetting.
- We uniformly models multi-modal and language-only tuning and extends language models into powerful vision-language models in a low-budget way.
- Experiments on instruction-following datasets, multi-task evaluation, image captioning, and VQA demonstrate the feasibility and advancement of LLaMA-Excitor.

2. Related Works

2.1. Parameter-Efficient Fine-Tuning

PEFT techniques enable efficient adaptation of large language models to various downstream applications without fine-tuning all the model's parameters. Adapter-based approaches [14, 36] insert additional feed-forward networks within the existing architecture. However, these methods can increase the computational load due to their se-

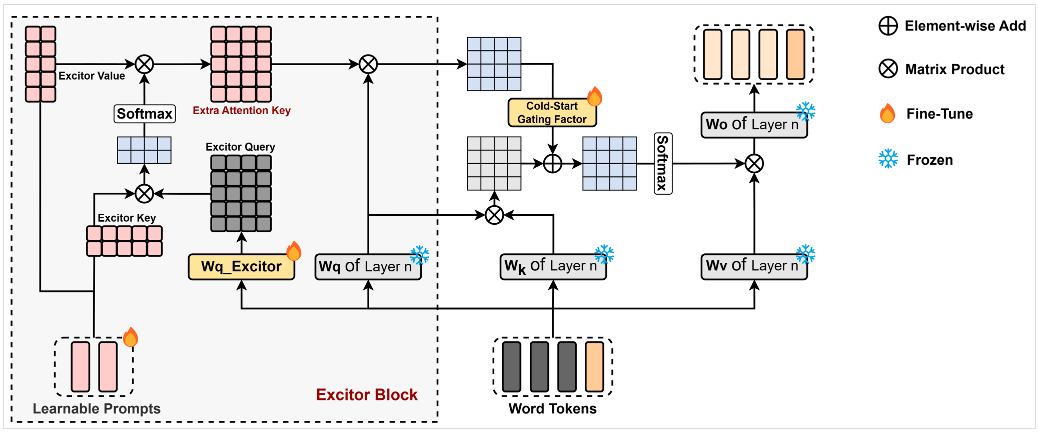


Figure 2. Details of the Excitor block. We assign a set of learnable prompts for attention layers of LLaMA. These prompts are used to construct an extra Key for computing additional similarity scores, which are then merged into the original scores to alter the LLM's behavior. Cold-start gating factors are designed to stabilize the training.

rial nature and may induce catastrophic forgetting by significantly changing the reasoning mechanism of LLMs. Prompt-learning methods offer a training-free approach to task adaptation [5, 41] and are improved by prefix-tuning [26, 29], which involves learnable token sequences. The primary drawback is the difficulty in capturing complex knowledge within a limited number of tokens, which can be restrictive for diverse tasks. Additionally, as the number of learnable tokens grows, these methods may lose control over the sequence output quality. LoRA [15, 50] aims to address these issues by introducing low-rank matrices in a parallel structure, enhancing both performance and efficiency. However, it risks incorporating features that deviate from the original model's distribution, potentially leading to performance degradation. Furthermore, equal distribution of parameter updates across layers disregards the distinct functionalities of different layers [16, 50], leading to suboptimal utilization of the model's capacity.

2.2. Visual Instruction Tuning

Instruction tuning[33, 48] finetunes or adapted language models to follow specific textual instructions. Despite the prominent success in the NLP domain, its inherent text-only nature limits its ability to comprehend and interact with visual content effectively. LLaVA [28] enables language models to follow visual instructions and engage with multi-modal information. And a series of approaches have been developed to improve the visual instruction tuning with advanced architectures [2, 4, 22] or more versatile functionalities[20, 35, 54]. Fine-tuning in the context of multi-modal tasks can be succinctly expressed as enhancing a model's proficiency in responding to visual cues. This essentially constitutes a variant of instruction following, where the prompts are visual rather than textual.

3. Methodology

The primary goal of Excitor is to effectively leverage LLMs' pre-trained knowledge and skills to solve downstream tasks without excessive modification of the reasoning process, which can lead to forgetting and performance degradation. We choose LLaMA[45] as our LLM as its effectiveness has been demonstrated in several open-source instruction-tuning works [8, 28, 34, 44, 51]. In Section 3.1, we first introduce how to insert Excitor blocks into LLaMA's attention layers. Then, in Section 3.2, we present the details of Excitor blocks' architecture. Finally, in section 3.3, we generalize LLaMA-Excitor for multi-modal reasoning.

3.1. Indirect feature interaction

Given 52K instruction-following data [44] and a pre-trained LLaMA with N attention layers, we insert our proposed Excitor blocks into the topmost $L\,(L < N)$ layers for instruction-following fine-tuning. This can better tune the language representations with higher-level semantics. Taking the l-th inserted layer as an example, we denote the M-length word tokens as $T_l \in \mathbb{R}^{M\times C}$, where C equals the feature dimension of LLaMA's attention layers. As shown in Figure 1, the Excitor block takes T_l with a set of learnable prompts $\{P_l\}_{l=1}^{L}$ as inputs, where $P_l \in \mathbb{R}^{K\times C}$ with K denoting the length of learnable prompts for each layer, generates an extra similarity matrix $S_l^{\text{extra}} \in \mathbb{R}^{M\times M}$ with exactly the same shape of original similarity matrix S_l for the l-th layer. In the self-attention mechanism, S_l is calculated from $Query$ and Key, and controls the importance of components in $Value$, where $Query$, Key, and $Value$ are linear projected version of T_l:

$$Query = W_{\text{q}}\left(T_l\right), Key = W_{\text{k}}\left(T_l\right), Value = W_{\text{v}}\left(T_l\right) \quad (1)$$

where W_{q}, W_{k}, W_{v} are the pre-trained LLMs' weights that are frozen during the finetuning. Then, S_l is formulated as:

$$S_l = \frac{Query \cdot Key^{\mathrm{T}}}{\sqrt{C}}, \tag{2}$$

Since S_l rules how the information updates, we treat S_l^{extra} as the residual of S_l to modify the reasoning process and formulate the updated output of the l-th layer T_l^{o} as:

$$S_l^g = \mathrm{Softmax}\left(S_l^{\mathrm{extra}} \times g_l + S_l\right), \tag{3}$$

$$T_l^{\mathrm{o}} = W_{\mathrm{o}}\left(S_l^g \cdot Value\right), \tag{4}$$

where g_l is a set of learnable parameters, named cold-start gating factor, which is inspired by zero-init attention [51], to stabilize training by controlling the proportion of S_l^{extra} participants in the reasoning. Instead of applying zero-initialization like [51], we initialize g_l with $\mathbf{N}\left(0, 10^{-2}\right)$ to overcome the gradient vanishing in the mixed precision (FP16) training.

As stated in sec 1, previous techniques updating the output of frozen LLMs' layers by adding the outcomes of trainable modules onto T_l^{o} (e.g., Adapter[14] and LoRA[15]) or concatenating trainable tokens with T_l as input for layer l (e.g., Prefix-tuning [26]) can be attributed to the direct modifying of the intermediate representations. One drawback of direct modifying is adding out-of-distribution features into the intermediate representations. It will largely change the inputs of higher layers and make the framework unstable, especially at the beginning of training. By contrast, LLaMA-Excitor only uses trainable tokens to change the attention score S_l^g. The output of each layer T_l^{o} is still the linear combination of $Value$ and within the original LLMs' distribution. i.e., LLaMA-Excitor alters the likelihood of potential outcomes generated by pre-trained LLMs. This indirect feature interaction ensures the model doesn't deviate too far from its pretrained reasoning process and can more effectively arouse pretrained skills in downstream usage.

3.2. Excitor Blocks

To impose effective control over the attention score S_l^g while reducing the degradation of LLM's inherent abilities, LLaMA-Excitor must be self-adaptive to the original input sequence T_l to decide whether learnable parameters should influence the reasoning. Thus, Excitor Blocks needs to generate the extra similarity matrix S_l^{extra} considering both injected information in the learnable prompt P_l and T_l. We formulate S_l^{extra} as:

$$S_l^{\mathrm{extra}} = \frac{Query \cdot Key_{extra}^{\mathrm{T}}}{\sqrt{C}}, \tag{5}$$

where $Query$ is reused from the equation 1 to pass knowledge from T_l while reducing the computational overhead, Key_{extra} is an extra attention key that carries information from P_l. Since the shape of S_l^{extra} should correspond with S_l, the shape of Key_{extra} must be the same with $Key \in \mathbb{R}^{M \times C}$ in the equation 1. However, M, which is the length of Key, is changing along with variable input texts, while the length of learnable prompts P_l is a pre-defined hyper-parameter.

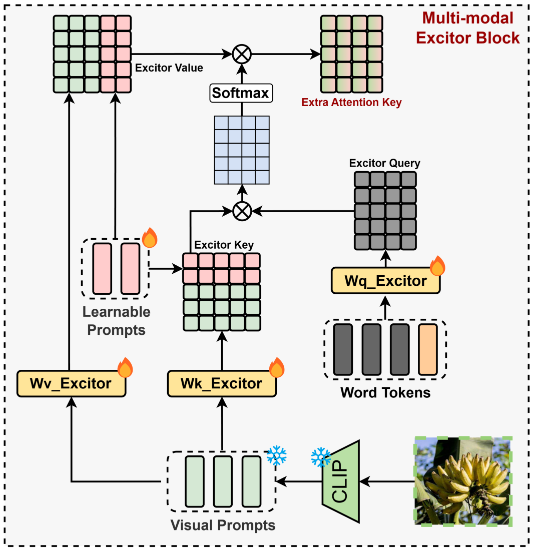


Figure 3. Extend LLaMA-Excitor into a powerful multi-modal model. Owing to the indirect feature interaction, LLaMA-Excitor is the cheapest PEFT method that can follow visual instructions without complicated projection modules aligning vision and language.

As shown on the left of the Excitor Block in Figure 2, we introduce the Key reconstruction to mitigate the shape inconsistency. Since the attention mechanism generates an output with the same shape as its queries and transmits information about its values, we define P_l as the Exitor Key and the Exitor $Value$ and calculate the Excitor $Query$ from T_l for the l-th Excitor Block. We omit linear projections for P_l to pass information directly to Key_{extra} and accelerate the training. The Key reconstruction is formulated as:

$$S_l^{\mathrm{excitor}} = \frac{W_{\mathrm{q}}^{\mathrm{excitor}}\left(T_l\right) \cdot P_l^{\mathrm{T}}}{\sqrt{C}}, \tag{6}$$

$$Key_{\mathrm{extra}} = \mathrm{Softmax}\left(S_l^{\mathrm{excitor}}\right) \cdot P_l, \tag{7}$$

Where $W_{\mathrm{q}}^{\mathrm{excitor}}$ is a low-rank module consisting of two linear layers with dimension $\mathbb{R}^{C \times r}$ and $\mathbb{R}^{r \times C}$. Finally, S_l^{extra} carries information from T_l and has the same shape as Key.

3.3. Multi-modal Extension

The indirect feature interaction is not limited to textual instructions but is also compatible with following visual prompts. We simply extend our Excitor Block into a multi-modal version in Figure 3. Taking an image and a task description as input, LLaMA-Excitor first leverages a frozen pre-trained visual encoder like CLIP [38] to extract a sequence of visual embeddings $I \in \mathbb{R}^{V \times D}$ as the visual

Instruction: Tell me about alpacas.

LLaMA7B: Alpacas are a type of camelid that are native to South America. They are related to llamas and vicunas. Alpacas are raised for their fiber, which is used to make clothing.

Full Fine-Tuning: Alpacas are small, soft-furry, and mammals that are native to South America. They are the largest of the wool-making mammals and are the most gentle and social of the mammals. They are also the most colorful mammal, with a range of colors from white to gray, brown, black, and shades of brown and gray. They are also very smart and can learn tricks easily. They are also very social and live in herds of up to 20-30 individuals.

Alpaca-LoRA: Alpacas are a species of South American camelid, related to the llama and the vicuña. They are a domesticated version of the wild vicuña and are primarily raised for their fine fleece, which is used to make clothing. Alpacas are typically smaller than llamas, with a range of sizes from 12 to 15 hands high. They are known for their gentle nature and are often kept as pets.

LLaMA-Adapter: Alpacas are members of the camelid family and are native to the Andes Mountains of South America. They are typically between 3 and 4 feet tall and weigh between 100 and 200 pounds. They have long, coarse, and dense coats that come in a variety of colors, including white, black, brown, and grey.

LLaMA-Excitor: Alpacas are domesticated animals that are related to llamas and are native to South America. They are typically between 100 and 120 cm tall and weigh between 100 and 150 kg. Alpacas have long, soft fleece that is used for making clothing, blankets, and other products. They are also used for their meat and their dung, which is used as fuel. Alpacas are herd animals and are known for their gentle and calm nature. They are also very intelligent and can be trained to perform certain tasks.

Instruction: Write a conversation between the sun and pluto.

LLaMA7B:
The sun: Hey Pluto, how are you?

Pluto: I'm fine.

The sun: I'm glad to hear that.

Pluto: I'm glad to hear that too.

The sun: I'm glad to hear that too.

Full Fine-Tuning:
The sun said, "I'm so hot!"

The sun said, "I'm so hot!"

The sun said, "I'm so hot!"

The sun said, "I'm so hot!"

The sun said, "I'm so hot!"

Alpaca-LoRA:
Sun: Hey Pluto, how's life in the outer solar system?

Pluto: It's pretty chilly up here! I'm still getting used to being so far away from the sun.

Sun: That's understandable. I'm sure you'll get used to it soon enough. Anyway, I'm glad you're still around.

Pluto: Thanks! It's nice to still be a part of the solar system.

LLaMA-Adapter:
The Sun: Good morning, Pluto!

Pluto: Good morning, Sun! How are you?

The Sun: I'm doing great! How are you?

Pluto: I'm doing well, too!

The Sun: I'm glad to hear that!

Pluto: Me too!

The Sun: I'm glad we're having such a nice day!

Pluto: Me too!

The Sun: I'm glad we're having such a nice day!

Pluto: Me too!

LLaMA-Excitor:
Sun: Hey Pluto, how are you doing?
Pluto: I'm doing great! How about you?
Sun: I'm doing well too. It's been a while since we last spoke.
Pluto: Yeah, it has been a while. I've been busy exploring the outer reaches of the solar system.
Sun: That's great! I'm glad you're having fun.
Pluto: I'm having a blast! I'm so glad I'm not stuck in the inner solar system like the other planets.
Sun: Yeah, it's nice to have some space to yourself.
Pluto: Definitely! So, what have you been up to?
Sun: I've been busy keeping the other planets warm and providing light and heat.
Pluto: That's a big job! You must be exhausted.
Sun: I am, but it's worth it. I'm glad I can help.
Pluto: I'm glad you're doing well. It's been great talking to you.

Table 1. Instruction-following Comparison between LLaMA-Excitor and other techniques.

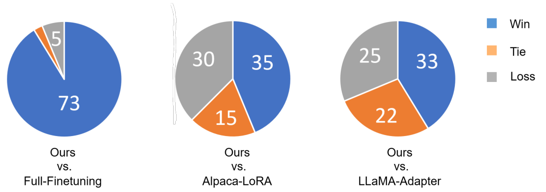


Figure 4. Quantitative comparisons between LLaMA-Excitor (BLUE) with other methods, evaluated by GPT-4 [32].

prompts, where V is the embedding length and D is the dimension of embeddings. Note that we adopt both the global CLS token and regional patch tokens from the last layer of the CLIP encoder to acquire adequate visual semantics. Then, we calculate multi-modal versions of the Excitor Key and the Exitor $Value$ by combining P_l and I:

$$Key_{\text{excitor}} = \left[P_l;\ W_{\text{k}}^{\text{excitor}}(I)\right], \tag{8}$$

$$Value_{\text{excitor}} = \left[P_l;\ W_{\text{v}}^{\text{excitor}}(I)\right]. \tag{9}$$

Unlike the processing of P_l, we adopt low-rank linear projections $W_{\text{k}}^{\text{excitor}}$ and $W_{\text{v}}^{\text{excitor}}$ on I, since we provide the same visual prompt for attention layers $l \in [N-L, N]$ and different layers require various visual clues for the downstream reasoning. LLaMA-Excitor is an innovative framework for the multi-modal fine-tuning of LLMs. As mentioned in Section 1, most techniques before LLaMA-Excitor involved training a module to perform multi-modal alignment that projects visual embeddings into the textual feature space. This enables LLMs to understand visual tokens and integrate them into the reasoning process. However, the strict alignment constraint may damage the rich semantics of visual embeddings, and their direct feature fusion may also cause performance degradation. In contrast, LLaMA-Excitor does not require feature alignment and works on the raw inputs from the frozen image encoder. It does not change the purity of the intermediate textual features and thus avoids any potential damage to the rich semantics of the visual embeddings and the reasoning of LLMs.

4. Experiments

4.1. Language-only Performances

4.1.1 Instruction-following Evaluations

In this section, we evaluate the instruction-following capacity of LLaMA-Excitor by responding to instructions.

Dataset. We use Stanford Alpaca [44], which contains 52K instruction-following data, as the training set. Each sample in this dataset is formatted according to a template as follows:

> 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: {instruction}
> Input: {input}
> Response: {output}

where {instruction} is the description of a task, {input} is the context for the task, and {output} is the answer generated by GPT-3.5 [5].

Implementation Details. We develop LLaMA-Excitor based on the original LLaMA codebase with minor modifications. We train LLaMA-Excitor on 8 NVIDIA A100 GPUs for five epochs. The warmup epochs, batch size, learning rate, and weight decay are set to 2, 64, 9e-3, and 0.02, respectively. In this section, we utilize the pretrained LLaMA-7B v1 model, which has $N = 32$ transformer lay-

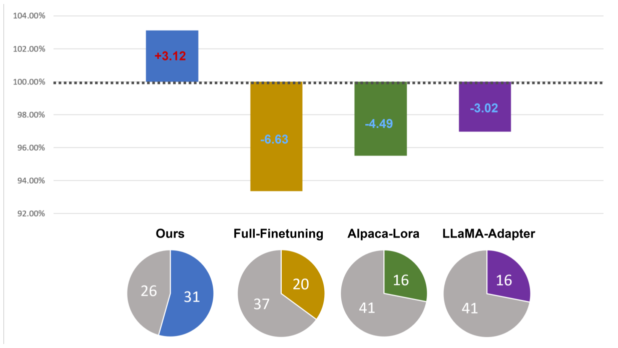


Figure 5. The relative average performance changes and win-loss situations of fine-tunings compared to original LLaMA-7B on MMLU.

ers with feature dimension $C = 4096$ as the backbone. We apply our excitor blocks with the low-rank dimension $r = 16$ for $W_{\text{q}}^{\text{excitor}}$ in the topmost $L = 30$ layers and set the length of learnable prompts $L = 30$. During the inference, we adopt top-p sampling as the default decoding method with a temperature of 0.1 and a top-p = 0.75.

We compare LLaMA-Excitor with three competing methods. (1) Full-finetuning, which updates the entire 7B parameters in LLaMA during the end-to-end training. (2) Alpaca-LoRA, the official implementation of low-rank adaptation [15] provided by [44]. (3) LLaMA-Adapter [51], a cutting-edge prefix-tuning method with zero-init attention to reduce forgetting in finetuning.

Performances. We first use examples in Table 1 to show the strong instruction-following ability of LLaMA-Excitor. The first case requires the models to provide some information about alpacas. LLaMA-Excitor generates reasonable responses comparable to Full-finetuning, demonstrating our method's effectiveness. It's worth noticing that our approach follows the same answering pattern as the prefix-tuning method LLaMA-adapter and the original LLaMA7B (although Excitor provides more details). This phenomenon is reasonable since prefix-tuning methods use additional learnable prompts to guide the generation process of LLMs. They do not change the internal reasoning process, just like our Excitor. This characteristic may also be the foundation of avoiding catastrophic forgetting. For the second task, the LLMs are asked to create a dialogue between the Sun and Pluto. Out of all the models that are tested, only Excitor and LoRA correctly recognize the characters as planets within the Solar System and incorporate their identities into the conversation. However, Excitor stands out as it can generate lengthy, high-quality conversations rich in interesting content. (Please see Appendix A for the full comparison including dialog generation, code generation, and question answering.)

To evaluate the performance of instruction-following models quantitatively, we refer to the research conducted by Chiang et al. [8]. In this evaluation, we use GPT-4 [32] to assess the quality of responses to 80 randomly selected questions from AlpacaEval [25]. However, we notice a bias in GPT-4's scoring, where it tends to give higher scores to the first response. To overcome this, we interchange the position of two responses and use the average score for the final rating. Results are shown in Figure 4. We notice that Excitor overwhelms the Full-Finetuning (73/80). The reason behind this is that AlpacaEval is an out-of-distributed (OOD) test for our training set, and full-finetuning may easily overfit the training set. When compared with Alpaca-LoRA and LLaMA-Adapter, our Excitor still gains the majority of "Win", demonstrating our indirect feature interaction is at least comparable with or better than predomint PEFT methods. (Please refer to Appendix B for instances on AlpacaEval.)

As this paper is a preliminary attempt to study indirect feature interaction in LLMs and the 52K instruction-following data used in our study is automatically generated, it is possible to further enhance the performance of LLaMA-Excitor by improving the quality of the training data, adjusting the hyper-parameters to more reasonable values or exploring larger LLaMA models.

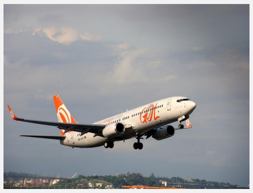
4.1.2 The Impact of Fine-tuning Techniques on Inherent Abilities

One of the motivations for introducing the indirect feature interaction in LLaMA-Excitor is to reduce the influence of fine-tuning on low-quality or non-targeted datasets. To quantitatively analyze this characteristic, we evaluate the fine-tuned models on MMLU [13], which is currently the most comprehensive benchmark for assessing the multi-task accuracy of text models. This benchmark includes 57 tasks across various fields, such as basic math, American history, computer science, and law. The model must possess extensive world knowledge and problem-solving capabilities to achieve high accuracy on this test. We report the relative performance changes of PEFT methods compared with the original LLaMA-7B and how many of the 57 evaluations PEFT methods outperform LLaMA-7B in Figure 5. We find that for models fine-tuned on Alpaca-52k, only Excitor can avoid reducing the average score of 57 tasks in MMLU, and it even shows an increase in performance (+3.12%). As expected, the full-finetuning performance suffers the most severe decline as its parameters are most significantly changed (-6.63%), while the LLaMA-Adapter, which is based on prefix-tuning, experiences the least decline (-3.02%) besides Exictor. (We provide results based on LLaMA2-7B in Appendix C)

Instruction: Please answer me based on this image. Generate a caption that details what happened in the image.

LLaMA-Excitor: An airplane is flying through the sky, with its landing gear down, as it approaches for a landing at an airport. The airplane is dominating the scene, with its bright red and white colors clearly visible against the backdrop of the cloudy sky. The airplane is about to touch down on the runway, creating a moment of excitement and anticipation for the passengers and crew on board.

Human Annotation: A large passenger airplane flying through the air.



Instruction: Please answer me based on this image. Generate a caption that details what happened in the image.

LLaMA-Excitor: In the image, a group of birds is standing together on a girraffe. One of the bird is leaning over to nibble on the girraffe's fur and the giraffe seems to be enjoying the interaction. The scene captures the playful nature of these animals and the unexpected interaction between a giraffe and a bird.

Human Annotation: Little birds sitting on the shoulder of a giraffe.

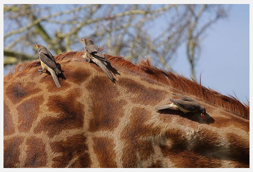


Table 2. Examples demonstrating LLaMA-Excitor's visual instruction following capacity.

Method	Data Scale		COCO Caption	
	PT	FT	BLEU@4	CIDEr
ClipCap [31]	0M	0.6M	33.5	113.1
VL-PET [16]	0M	0.6M	-	121.7
LLaMA-AdapterV2 [12]	0M	0.6M	36.2	122.2
Qwen-vl-chat [4]	1.4B	0.6M	-	131.9
mPLUG-Owl2 [49]	348M	0.6M	-	137.3
BLIP [21]	14M	0.6M	40.4	136.7
Flamingo [2]	1.8B	0.6M	-	138.1
BLIP-2 [22]	129M	0.6M	43.7	145.3
LLaMA-Excitor	0M	0.6M	**49.7**	**157.5**

Table 3. Comparison with state-of-the-art image captioning methods on COCO Caption.

4.2. Multi-Modal Performances

In this section, we evaluate the visual instruction following the capacity of LLaMA-Excitor by responding to paired vision-language instructions and demonstrate our indirect feature interaction, which unifies the modeling of language-only tuning and multi-modal tuning, providing a powerful low-budget way to perform vision-language tasks.

4.2.1 Image Captioning Evaluation

Implementation Details. We first evaluate our Excitor on COCO Caption [7], which contains 0.6M training image caption data (120k images with 5 captions per image) over a wide range of distributions. We apply frozen CLIP/14-L [38] as the image encoder, the visual embedding dimension $D = 768$, low-rank dimension $r = 16$. We keep other hyper-parameters the same as the section 4.1.1.

Performances. We compare our Excitor with cutting-edge image captioning methods in Table 3. We find that Excitor achieves a stunning result, significantly surpassing the previous SOTA method BILP-2 [22] by +6 BLEU@4 and +12.2 CIDEr. Especially considering that we do not apply complex multi-modal alignment modules and have only 0.6M training data. By contrast, BLIP-2 adopts a costly pre-training stage on an additional 129M dataset, including VisualGenome [19], Conceptual Captions [42, 43] and LAION [6]. We visualize the results of image captioning in Appendix D. However, the image captions provided in COCO are concise and lack sufficient detail in describing the content of images. To further unleash Exitor's performance in image captioning, we continue fine-tuning Exitor on LLaVA665k [27], a higher-quality dataset with detailed picture descriptions. We provide several examples from the COCO test set in Table 2. It shows that image captions generated by Excitor can accurately cover the content of human annotations and provide richer details. (e.g., (1) The airplane's landing gear is down, (2) A bird is pecking at the back of a giraffe)

4.2.2 Performances on ScienceQA

Implementation Details. We evaluate our Excitor on ScienceQA [40], which contains 21k multimodal multiple choice questions with rich domain diversity across 3 subjects, 26 topics, 127 categories, and 379 skills. We train Excitor on ScienceQA train split from scratch, using the combination of Chain-of-Thought(CoT) and direct answer prediction. CoT requires the model to predict the solution first and then generate the answer choice based on the solution. During training, we randomly require the model to first generate solutions or directly predict answers in each iteration. The model is asked to answer the question directly in the evaluation.

Table 4 reports the performances of the top methods on the official leaderboard of the ScienceQA and Excitor. The current SOTA method LLaVA [28] is pretrained on another 558k visual-language dataset and fine-tuned on ScienceQA. Also, they allow all LLM's parameters to be updated during fine-tuning. Our Excitor demonstrates comparable performance with LLaVA w/o pretraining (only -0.4% performance gap). Note that LLaMA-Excitor is a PEFT method (with a frozen LLM) without CoT, and LLaVA utilizes a larger backbone LLaMA-13B than our LLaMA-7B. To further unlock the potential of Excitor, we introduce a vari-

Model	Average	Subject			Context Modality			Grade	
		NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12
Human [40]	88.40	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42
UnifiedQA$_{CoT}$	74.11	71.00	76.04	78.91	66.42	66.53	81.81	77.06	68.82
GPT-3$_{CoT}$	75.17	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68
ChatGPT$_{CoT}$ [1]	78.31	78.82	70.98	83.18	77.37	67.92	86.13	80.72	74.03
GPT-4$_{CoT}$ [32]	83.99	85.48	72.44	90.27	82.65	71.49	92.89	86.66	79.04
MM-COT [53]	84.91	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37
LLaVA$_{CoT}$ [28]	90.92	90.36	95.95	88.00	89.49	88.00	90.66	90.93	90.90
LLaVA$_{CoT}$ (w/o pretrain) [28]	85.81	-	-	-	-	-	-	-	-
DFAF [11]	60.72	64.03	48.82	63.55	65.88	54.49	64.11	57.12	67.17
ViLT [18]	61.14	60.48	63.89	60.27	63.20	61.38	57.00	60.72	61.90
Patch-TRM [30]	61.42	65.19	46.79	65.55	66.96	55.28	64.95	58.04	67.50
VisualBERT [23, 24]	61.87	59.33	69.18	61.18	62.71	62.17	58.54	62.96	59.92
UnifiedQA [17]	70.12	68.16	69.18	74.91	63.78	61.38	77.84	72.98	65.00
GPT-3 [5]	74.04	75.04	66.59	78.00	74.24	65.74	79.58	76.36	69.87
LLaMA-Adapter	85.19	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05
LLaMA-Excitor	85.41	85.70	**92.35**	82.82	83.43	84.56	86.27	85.65	84.64
LLaMA-Excitor@336px + LoRA	**88.39**	**87.19**	91.33	**87.09**	**90.42**	**85.20**	**88.64**	**88.35**	**88.42**

Table 4. Question Answering Accuracy (%) on ScienceQA's [40] test set. We report GPT-3 [5], ChatGPT [1], and GPT-4 [32] for zero-shot inference. CoT denotes utilizing chain-of-thought for question answering.

ant that utilizes a more powerful image encoder (CLIP-L/14@336) and adds LoRA blocks to the original LLM's parameters to reduce the gap with LLaVA's full parameter updating. It brings +2.58% performance gain compared with LLaVA.

4.3. Ablations

Variant	ACC
(a) CLIP-16/B	83.11
(b) CLIP-14/L	85.41
(c) CLIP-14/L@336px	85.87
(d) CLIP-14/L@336px + Full-Finetuning	83.20
(e) CLIP-14/L@336px + LoRA	**88.39**

Table 5. Analysis of the effectiveness of each module.

We ablate several design choices on ScienceQA in Table 5. (a), (b) and (c) study the influence of the capacity of image encoder. The best CLIP encoder yields 85.41% and is +2.76% higher than the basic version. In (d) and (e), we study to combine Excitor, which focuses on instruction-following, with other techniques that update LLM's reasoning process. We find LoRA can bring 2.98% performance gain for Excitor, demonstrating the potential of combining Excitor with existing finetuning methods in multi-modal usage. (However, in the text-only scenario, we encounter -4.14% performance gap compared with the original LLaMA-7B in the MMLU test by combining Excitor and LoRA.) Please see Appendix E for more ablations on the design of low-rank linear layers (i.e., $W_{\text{q}}^{Excitor}$, $W_{\text{k}}^{Excitor}$, and $W_{\text{v}}^{Excitor}$) in Excitor Blocks under single-modal and multi-modal scenarios, the choice of low-rank dimension r, and the number of layers L with Excitor blocks inserted.

5. Conclusion

In this paper, we study the indirect feature interaction in fine-tuning LLMs into instruction-following models and propose LLaMA-Exictor, a PEFT method that demonstrates outstanding instruction-following capacity compared with predominant techniques meanwhile reducing the forgetting of LLMs' inherent abilities. Moreover, Excitor also shows its potential in vision-language tasks by unifying the modeling of visual instruction-following and language-only instruction-following. Impressively, Excitor achieves strong multi-modal performance without the need to train complex alignment modules.

This project is a work in progress, and some directions can be explored. The first one is the adaptability to LLMs other than LLaMA. The second direction is to explore more effective ways to introduce multi-scale visual prompts (since we only apply the features from the last layer of CLIP at the current stage). Last but not least, Excitor has the potential to fine-tune vision-only models into vision-language models, just like our implementation for LLMs.

Acknowledge

This work is supported in part by the Beijing Natural Science Foundation (No. L222024), and the National Natural Science Foundation of China (No. 62394322). One of the authors, Chao Yang, is supported by the Shanghai Post-doctoral Excellent Program (Grant No. 2022234).

References

[1] Chatgpt. `https://chat.openai.com`, 2023. 8

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 3, 7

[3] Anonymous. Videodistill: Language-aware vision distillation for video question answering. In *Conference on Computer Vision and Pattern Recognition 2024*, 2024. 2

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 3, 7

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2, 3, 5, 8

[6] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pretraining to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 7

[7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. arxiv 2015. *arXiv preprint arXiv:1504.00325*, 2015. 7

[8] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023. 3, 6

[9] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019. 1

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[11] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6639–6648, 2019. 8

[12] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 7

[13] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020. 6

[14] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 2, 4

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 3, 4, 6

[16] Zi-Yuan Hu, Yanyang Li, Michael R Lyu, and Liwei Wang. Vl-pet: Vision-and-language parameter-efficient tuning via granularity control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3010–3020, 2023. 2, 3, 7

[17] Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. Unifiedqa: Crossing format boundaries with a single qa system. In *Findings of the Association for Computational Linguistics (EMNLP)*, pages 1896–1907, 2020. 8

[18] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 5583–5594, 2021. 8

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 7

[20] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 3

[21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 7

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 3, 7

[23] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 8

[24] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. What does bert with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5265–5275, 2020. 8

[25] Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tat-

sunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`, 2023. 6

[26] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 2, 3, 4

[27] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 7

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 2, 3, 7, 8

[29] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021. 2, 3

[30] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *The 35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2021. 8

[31] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 2, 7

[32] R OpenAI. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2023. 5, 6, 8

[33] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022. 3

[34] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023. 3

[35] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 3

[36] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*, 2020. 2

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

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 4, 7

[39] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 1

[40] Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301, 2022. 7, 8

[41] Timo Schick and Hinrich Schütze. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*, 2020. 2, 3

[42] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 7

[43] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 7

[44] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023. 3, 5, 6

[45] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3

[46] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: open and efficient foundation language models, 2023. *URL https://arxiv. org/abs/2302.13971*, 2023. 2

[47] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 2

[48] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022. 3

[49] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*, 2023. 7

[50] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023. 2, 3

[51] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 2, 3, 4, 6

[52] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab,

Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 1

[53] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 8

[54] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3