

# MiniGPT-4:增强视觉语言理解与先进的大型语言模型

朱德耀(Deyao Zhu) \*陈军(Jun Chen) \*沈晓倩(Xiaoqian Li) Mohamed Elhoseiny  
阿卜杜拉国王科技大学{ 德耀.朱, jun.chen全新课题. 沈,xiang.li. 1, mohamed.elhoseiny } @  
kaust.edu.sa

## 摘要。

最近的GPT-4展示了非凡的多模式能力，例如直接从手写文本生成网站和识别图像中的幽默元素。这些特征在以前的视觉语言模型中很少被观察到。我们认为GPT-4具有先进的多模态生成能力的主要原因在于使用了更先进的大型语言模型(LLM)。为了研究这一现象，我们提出了MiniGPT-4，它只使用一个投影层将冻结的视觉编码器与冻结的LLM对齐。我们的研究表明，MiniGPT-4具有许多类似于GPT-4所展示的功能，如详细的图像描述生成和从手写草稿创建网站。此外，我们还观察到MiniGPT-4中的其他新兴功能，包括根据给定图像编写故事和诗歌，为图像中显示的问题提供解决方案，根据食物照片教用户如何烹饪等。在我们的实验中，我们发现仅对原始图像-文本对进行预训练可能会产生缺乏连贯性的非自然语言输出，包括重复和碎片化的句子。为了解决这个问题，我们在第二阶段策划了一个高质量的、对齐良好的数据集，使用会话模板对我们的模型进行微调。事实证明，这一步对于增强模型的生成可靠性和整体可用性至关重要。值得注意的是，我们的模型具有很高的计算效率，因为我们只使用大约500万个对齐的图像-文本对来训练一个投影层。我们的代码、预训练的模型和收集的数据集可以在<https://minigpt-4.github.io/>上获得。

## 1 介绍

近年来，大型语言模型(large language models, 法学硕士)发展迅速[21,18,4,24,32,9,14]。凭借出色的语言理解能力，这些模型可以以零射击的方式执行各种复杂的语言任务。值得注意的是，GPT-4 [19]是一种大型多模式模型，最近被引入，展示了许多令人印象深刻的功能。例如，GPT-4可以生成非常详细和准确的图像描述，解释不寻常的视觉现象，甚至可以根据手写的文字说明构建网站。

虽然GPT-4已经展现出非凡的能力，但其非凡能力背后的方法仍然是一个谜[19]。我们认为，这些卓越的技能可能源于对更先进的大型语言模型(LLM)的使用。法学硕士表现出了各种应急能力，GPT-3的几次提示设置[4]和Wei等人(2022)[34]的研究结果证明了这一点。在较小规模的模型中很难找到这样的涌现特性。据推测，这些涌现的

\* Equal contribution

能力也适用于多模态模型，这可能是GPT-4令人印象深刻的视觉描述能力的基础。

为了证实我们的假设，我们提出了一个名为MiniGPT-4的新模型。它使用了一种先进的大型语言模型(LLM)， 骆马[8]，它建立在LLaMA[32]的基础上，据报道，根据GPT-4的评估，它可以达到ChatGPT 90%的质量，作为语言解码器。在视觉感知方面，我们采用了与BLIP-2[16]相同的预训练视觉组件，该组件由EVA-CLIP[13]中的viti - g /14和Q-Former组成。MiniGPT-4增加了一个单一的投影层，将编码的视觉特征与骆马语言模型对齐，并冻结所有其他视觉和语言组件。MiniGPT-4最初在4个A100 gpu上使用256个批处理大小进行20k步的训练，利用包括来自LAION [26]， Conceptual Captions[5,27]和SBU[20]的图像的组合数据集来将视觉特征与骆马语言模型对齐。然而，简单地将视觉特征与LLM对齐不足以训练具有视觉会话能力(如聊天机器人)的高性能模型，并且原始图像-文本对背后的噪声可能导致语言输出不连贯。因此，我们收集了另外3500个高质量对齐的图像-文本对，用设计好的会话模板进一步微调模型，以提高生成语言的自然度及其可用性。

在我们的实验中，我们发现MiniGPT-4具有许多与GPT-4相似的功能。例如，MiniGPT-4可以生成复杂的图像描述，基于手写文本指令创建网站，并解释不寻常的视觉现象。此外，我们的研究结果显示，MiniGPT-4还具有GPT-4演示中未展示的各种其他有趣的能力。例如，MiniGPT-4可以通过观察受图像启发的开胃食物照片、工艺故事或说唱歌曲，直接生成详细的食谱，在图像中编写产品广告，区分照片中显示的问题并提供相应的解决方案，以及直接从图像中检索有关人物、电影或艺术的丰富事实等功能。这些能力在以前的视觉语言模型中是不存在的，如Kosmos-1[15]和BLIP-2[16]，它们不适用骆马等更强的语言模型。这一对比验证了将视觉特征与高级语言模型集成可以产生紧急的视觉语言能力。

我们总结了我们的主要发现:

- 我们的研究表明，通过将视觉特征与先进的大型语言模型骆马相结合，我们可以实现紧急视觉语言能力。我们演示了我们的MiniGPT-4可以处理类似于GPT-4演示中展示的能力。
- 通过使用预训练视觉编码器和大型语言模型，MiniGPT-4实现了更高的计算效率。我们的研究表明，仅训练一个投影层就可以有效地将视觉特征与大型语言模型对齐。我们的MiniGPT-4只需要在4个A100 gpu上训练大约10个小时。
- 我们发现，使用来自公共数据集的原始图像-文本对，简单地将视觉特征与大型语言模型对齐，不足以开发性能良好的MiniGPT-4模型。它可能会产生不自然的语言输出，缺乏连贯性，包括重复和碎片化的句子。解决这一限制需要使用高质量、对齐良好的数据集进行训练，这将显著提高其可用性。

## 2 相关工作

近年来，由于训练数据的规模扩大和参数数量的增加，大型语言模型取得了巨大的成功。早期的模型，如BERT[11]、GPT-2[22]和T5[23]，为这一进展奠定了基础。随后，大规模的1750亿个参数的GPT-3[4]被引入，在众多语言基准上取得了重大突破。这一发展启发了其他各种大型语言模型的创建，包括Megatron-Turing NLG[28]、Chinchilla[14]、PaLM[9]、OPT[38]、BLOOM[25]和LLaMA[32]等。Wei等人[34]进一步发现了几种突现能力，这些能力只出现在大型模型中。这些能力的出现强调了在大型语言模型的开发中按比例放大的重要性。此外，通过将预训练的大型语言模型GPT-3与人类意图、指令和人类反馈相结合，InstructGPT[21]和ChatGPT[18]实现了与人类的对话交互，并可以回答各种各样的复杂问题。最近，一些开源模型，如

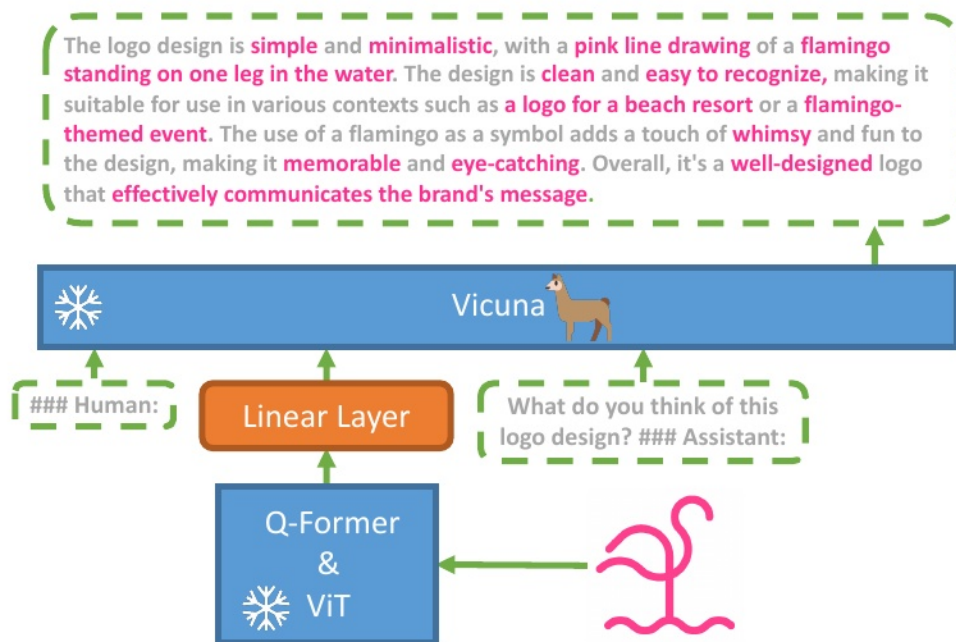


图1:MiniGPT-4的架构。它由一个带有预训练ViT和Q-Former的视觉编码器、一个单一的线性投影层和一个先进的骆马大语言模型组成。MiniGPT-4只需要训练线性投影层，使视觉特征与骆马对齐。

如羊驼[30]和骆马[8]，都是在LLaMA[32]的基础上发展起来的，也表现出类似的性能。

在视觉语言任务中利用预训练的法学硕士。近年来，在视觉语言任务中使用自回归语言模型作为解码器的趋势得到了显著的关注[6,15,36,31,2,16,17,12]。这种方法利用了跨模态迁移，允许知识在语言和多模态域之间共享。VisualGPT[6]和Frozen[33]等开创性研究已经证明了使用预训练的语言模型作为视觉语言模型解码器的好处。然后开发了Flamingo[2]，使用门控交叉注意来对齐预训练的视觉编码器和语言模型，并在数十亿对图像-文本对上进行了训练，展示了令人印象深刻的上下文少镜头学习能力。随后，引入了BLIP-2[16]，使用带有Q-Former的Flan-T5[10]有效地将视觉特征与语言模型对齐。最近，具有5620亿个参数的PaLM-E[12]已被开发用于将现实世界的连续传感器模态集成到LLM中，从而建立了现实世界感知与人类语言之间的联系。GPT-4[19]最近也发布了，在对大量对齐的图像-文本数据进行预训练后，展示了更强大的视觉理解和推理能力。

法学硕士，如ChatGPT，已经被证明是与其他专业模型协作来增强视觉语言任务性能的强大工具。例如，Visual ChatGPT[35]和MM-REACT[37]展示了ChatGPT如何充当协调器，与不同的可视化基础模型集成，并促进它们的协作，以应对更复杂的挑战。ChatCaptioner[39]将ChatGPT视为提问者，提出各种问题供BLIP-2回答。通过多轮对话，ChatGPT从BLIP-2中提取视觉信息，有效总结图像内容。视频ChatCaptioner[7]扩展了这种方法，将其应用于视频时空理解。ViperGPT[29]展示了将LLM与不同视觉模型结合起来以编程方式解决复杂视觉查询的潜力。相比之下，MiniGPT4直接将视觉信息与语言模型对齐，在不使用外部视觉模型的情况下完成各种视觉语言任务。

## 3 方法

MiniGPT-4旨在将来自预训练视觉编码器的视觉信息与先进的大型语言模型(LLM)对齐。具体来说,我们使用Vicuna[8]作为我们的语言解码器,它是在LLaMA[32]的基础上构建的,可以执行广泛的复杂语言任务。对于视觉感知,我们使用与BLIP-2[16]相同的视觉编码器,ViT主干[13]与预训练的Q-Former相结合。语言和视觉模型都是开源的。我们的目标是使用线性投影层弥合视觉编码器和LLM之间的差距,我们的模型概述如图1所示。

为了实现有效的MiniGPT-4,我们提出了一个两阶段的训练方法。初始阶段涉及在大量对齐的图像-文本对上对模型进行预训练,以获取视觉语言知识。在第二阶段,我们使用更小但高质量的图像-文本数据集和设计的会话模板对预训练模型进行微调,以增强模型的生成可靠性和可用性。

### 3.1 第一个预训练阶段

在初始预训练阶段,模型被设计为从大量对齐的图像-文本对中获得视觉语言知识。我们将注入投影层的输出视为对LLM的软提示,提示其生成相应的ground-truth文本。

在整个预训练过程中,无论是预训练的视觉编码器还是LLM都保持冻结状态,只有线性投影层被预训练。我们使用conceptual Caption[5,27]、SBU[20]和LAION[26]的组合数据集来训练我们的模型。我们的模型经历了2万个训练步骤,批大小为256,覆盖了大约500万对图像-文本。整个过程大约需要10个小时才能完成,使用4个A100 (80GB) gpu。

第一个预训练阶段的问题。在第一个预训练阶段之后,我们的MiniGPT-4展示了拥有丰富知识并对人类询问提供合理响应的能力。然而,我们已经观察到它难以产生连贯的语言输出的实例,例如生成重复的单词或句子、碎片化的句子或不相关的内容。这些问题阻碍了MiniGPT-4与人类进行流畅的视觉对话的能力。

我们还注意到,在GPT-3中也面临着类似的问题。尽管在广泛的语言数据集上进行了预训练,但GPT-3不能直接生成符合用户意图的语言输出。通过从人类反馈中进行指令微调和强化学习的过程,GPT-3演变为GPT-3.5[21,18],并能够产生更人性化的输出。这种现象与MiniGPT-4在初始预训练阶段后的当前状态相似。因此,我们的模型在这个阶段可能难以生成流畅和自然的人类语言输出也就不足为奇了。

### 3.2 为视觉语言领域策划一个高质量的对齐数据集。

为了在生成的语言中实现更大的自然度并增强模型的可用性,第二阶段的对齐过程是必不可少的。虽然在NLP领域,指令微调数据集[30]和对话[1]很容易访问,但视觉语言领域没有等效的数据集。为了解决这一不足,我们精心策划了一个高质量的图像-文本数据集,专门为对齐目的量身定制。该数据集随后用于在第二阶段对准过程中微调我们的MiniGPT-4。

初始对齐图像-文本生成在初始阶段,我们使用从第一个预训练阶段导出的模型来生成给定图像的全面描述。为了使我们的模型能够产生更详细的图像描述,我们设计了一个遵循Vicuna[8]语言模型的会话格式的提示符,如下所示:

###Human: <Img><ImageFeature></Img>详细描述此图像。给出尽可能多的细节。把你看到的都说出来。###助理:

在这个提示符中,<ImageFeature>表示线性投影层产生的视觉特征。

为了识别不完整的句子，我们检查生成的句子是否超过80个token。如果没有，我们将加入一个额外的提示符，`###Human: Continue ###Assistant:`，提示我们的MiniGPT-4扩展生成。通过连接这两个步骤的输出，我们可以创建一个更全面的图像描述。这种方法使我们能够生成更多具有详细信息丰富的图像描述的图像-文本对。我们从Conceptual Caption数据集中随机选择5000张图像[5, 27]，并采用这种方法为每张图像生成相应的语言描述。

数据后处理生成的图像描述仍然有很多噪声并包含错误，例如单词或句子的重复，以及不连贯陈述的存在。为了缓解这些问题，我们使用ChatGPT通过使用下面的提示来改进描述：

修复给定段落中的错误。删除任何重复的句子,无意义的字符,不是英语句子,等等。删除不必要的重复。重写不完整的句子。直接返回结果,不做任何解释。如果输入的段落已经正确,则直接返回,无需解释。

在完成后续处理阶段后，我们手动验证每个图像描述的正确性，以保证其高质量。具体来说，我们检查每个生成的图像描述是否遵循我们期望的格式，并通过消除ChatGPT无法检测到的冗余单词或句子来手动改进生成的标题。最后，5000对图像-文本对中只有大约3500对满足我们的要求，这些对随后被用于第二阶段的对齐过程。

### 3.3 阶段的整合

在第二阶段，我们用精心策划的高质量图像-文本对我们的预训练模型进行微调。在微调期间，我们在以下模板中使用预定义的提示：

`###人:<Img><ImageFeature></Img> <指令> ###助手:`

在此提示符中，`<Instruction>`表示从我们预定义的指令集中随机抽样的指令，该指令集包含各种形式的指令，例如“详细描述此图像”或“您可以为我描述此图像的内容吗”。需要注意的是,我们没有计算这个特定文本-图像提示符的回归损失。

因此，MiniGPT-4现在能够产生更自然和可靠的反应。此外，我们观察到模型的微调过程非常高效，只需要400个训练步骤，批量大小为12，使用单个A100 GPU只需7分钟即可完成。

## 4 示威活动:

我们的MiniGPT-4展示了与GPT-4类似的众多功能。这些包括生成详细的图像描述(图2)，识别图像中的有趣方面(图3)，以及发现不寻常的内容(图4)。此外，该模型可以从手写文本生成网站(图5)。我们还发现，我们的MiniGPT-4还具有其他能力，例如识别图像中的问题并提供解决方案(图6)，受图像启发创作诗歌或说唱歌曲(图7)，为图像写故事(图8)，为图像中的产品制作广告(图9)，识别个体(图10)，提供有见地的图像评论(图11)，检索与图像相关的事实(图12)，并教用户使用给定的照片烹饪食物(图13)。这些不同的例子展示了我们的MiniGPT-4的强大功能。

## 5 局限性

尽管MiniGPT-4处理了许多先进的视觉语言功能，但正如我们在演示中所展示的那样，它目前仍然面临一些限制。

语言的幻觉。由于MiniGPT-4是建立在法学硕士的基础上，它继承了LLM的局限性，如不可靠的推理能力和幻觉不存在的知识。这个问题可能会得到缓解

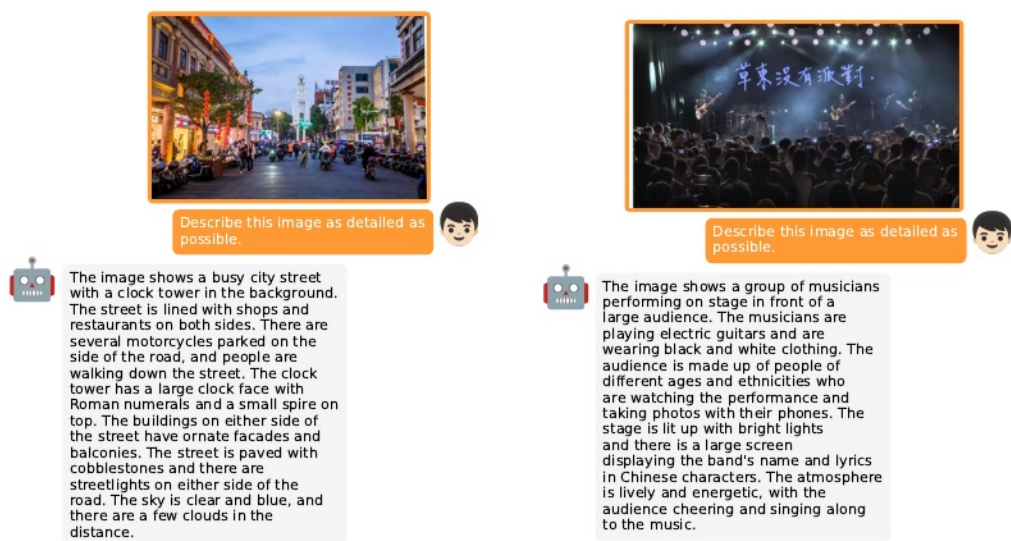


图2:详细的图像描述

通过使用更多高质量、对齐的图像-文本对训练模型，或者在未来与更高级的法学硕士对齐。

感知能力不足。MiniGPT-4的视觉感知仍然有限。它可能难以从图像中识别详细的文本信息，并区分空间定位。这种限制可能源于以下几个因素:1)缺乏足够的对齐图像-文本数据，其中包含足够的信息，如空间定位和光学字符注释。这个问题可以通过在更对齐良好和丰富的数据上进行训练来缓解。2)视觉编码器中使用的冻结Q-former可能会失去一些基本特征，比如视觉空间接地。这可能会通过用更强的视觉感知模型取代它而得到改善。3)只训练一个投影层可能无法提供足够的学习广泛的能力来学习广泛的视觉文本对齐。



Describe this image.



The image shows a cat wearing a blue cookie monster costume sitting on a couch with a plate of cookies in front of it. The cat's mouth is open and it appears to be enjoying the cookies. The caption reads, "another wild saturday night."

The image is a cute and funny depiction of a cat in a cookie monster costume enjoying some cookies on a Saturday night. The cat's expression and the plate of cookies in front of it add to the humor of the image. The caption reinforces the idea that the cat is having a wild and fun night.

Explain why this image is funny.



The image is funny because it depicts a cat wearing a cookie monster costume sitting on a couch with a plate of cookies in front of it. The cat's expression and the plate of cookies in front of it add to the humor of the image. The caption reinforces the idea that the cat is having a wild and fun night, which adds to the humor of the image. Additionally, the contrast between the cat's serious expression and the playful nature of the cookie monster costume creates a humorous juxtaposition. Overall, the image is funny because it combines elements of cuteness and humor to create a playful and amusing depiction of a cat enjoying some cookies on a Saturday night.



Explain why this meme is funny.



This meme is funny because it shows a dog laying on the floor with its head resting on its paws, as if it is tired or sleepy. The caption, "monday just monday," adds to the humor by suggesting that the dog is feeling the same way as many people do on Mondays, which are often considered to be the most dreaded day of the week. The meme is relatable and humorous, making it a popular one among internet users.

图3:识别图像中有趣的方面





Describe the image



The image depicts a wolf and a lamb lying in a field of grass. The wolf is lying down with its head resting on its paws, while the lamb is lying next to it, looking up at the wolf. The sun is shining down on the scene, casting a warm glow on the animals and the grass around them. The overall mood of the image is peaceful and serene.

Can this happen in real world?



In the wild, it is not uncommon for wolves and lambs to coexist in the same area, although it is rare for them to interact in the way depicted in the image. In most cases, wolves will prey on lambs and other young livestock, so it is unlikely that they would lie down together in a peaceful manner like in the image. However, it is possible that they may come into contact with each other in other ways, such as when the wolf is defending its territory or when the lamb is separated from its flock.



Describe the image



The image shows a cactus plant standing in the middle of a frozen lake. The cactus is surrounded by large ice crystals that have formed on the surface of the water. In the background, there is a range of mountains with snow-capped peaks. The sky is clear and blue, with a few fluffy clouds floating in the distance. The overall mood of the image is peaceful and serene, with the cactus standing tall and proud in the center of the frozen lake.

Is this image common in real world?



No, this image is not common in the real world. While it is possible for cacti to grow in cold climates, it is rare to see one standing in the middle of a frozen lake. Additionally, the formation of large ice crystals on the surface of the water is also not a common occurrence. This image is likely a digital creation or manipulation.

图4:发现不寻常的内容(图片来自WHOOPS数据集[3])



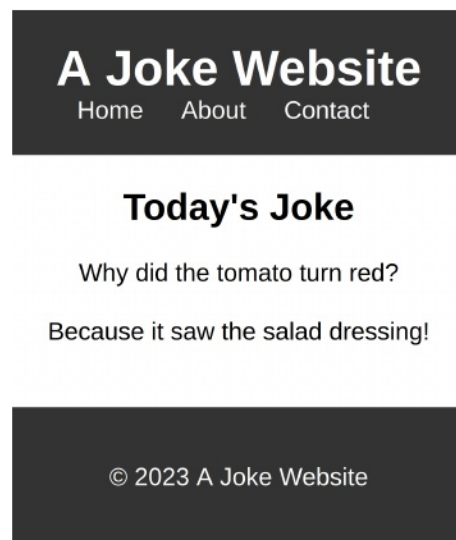
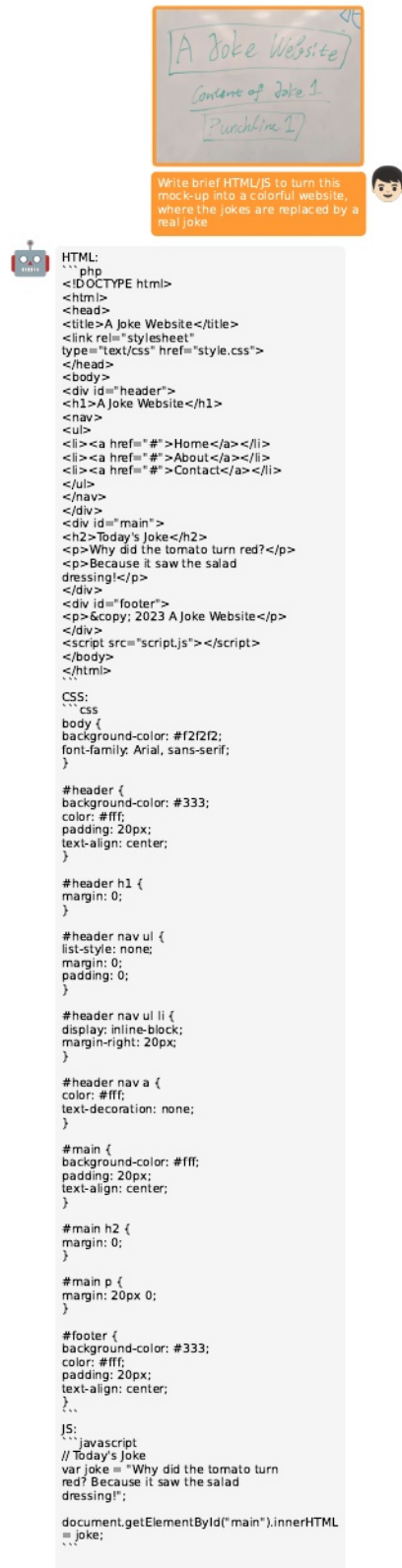


图5:从手写文本和呈现的网站生成网站代码

What is the issue with my plant? What should I do now?

The image shows a leaf with brown spots on it. The spots are likely caused by a fungal infection, which can be treated with a fungicide. It's important to identify the type of fungus causing the spots in order to choose the appropriate treatment.

To treat the plant, you can try the following steps:

1. Identify the type of fungus causing the spots on the leaves.
2. Use a fungicide specifically labeled for the type of fungus you have identified.
3. Apply the fungicide according to the product's instructions, making sure to cover all affected leaves.
4. Repeat the treatment every 7 to 14 days, or as directed by the product's instructions, until the fungus is completely gone.
5. Water the plant well after treatment to help the fungicide penetrate the leaves.
6. Keep the plant in a well-ventilated area and away from direct sunlight to prevent overheating.
7. Monitor the plant's progress and repeat the treatment if new spots appear.

It's also important to keep the plant healthy by providing it with proper care, such as adequate watering, fertilization, and pruning. This will help the plant to recover and prevent future fungal infections.

Why this happens and how to fix it?

The image shows a washing machine with a large amount of bubbles coming out of it. It appears that the washing machine is overflowing with soap suds.

There are several reasons why this might happen:

1. Overloading the washing machine: If too much laundry is put into the washing machine at once, it can cause an overflow.
2. Using too much detergent: If too much detergent is used in the washing machine, it can cause an overflow.
3. A clogged drain: If the drain of the washing machine is clogged, it can cause an overflow.
4. A malfunctioning washing machine: If the washing machine is not functioning properly, it can cause an overflow.

To fix this issue, you can try the following:

1. Check the load size: Make sure you are not overloading the washing machine.
2. Use the right amount of detergent: Use the recommended amount of detergent for the size of the load.
3. Clear the drain: Make sure the drain is clear of any debris.
4. Check for malfunctions: Have the washing machine checked by a professional if it is not functioning properly.

图6:从照片中识别问题并提供解决方案



Sure, here is a poem about the image you provided:

A man sits on a hill, with his dog by his side  
The sun sets behind them, casting a golden light  
The city below is bustling, but they are at peace  
Enjoying the moment, in each other's embrace

The man's arms are wrapped around his furry friend  
As they watch the world go by, with a gentle bend  
The dog's head rests on the man's shoulder  
A symbol of their unbreakable bond, forever

The sky is ablaze with color, a beautiful sight  
But it's the love between the man and his dog that shines bright  
It's a moment of pure joy, a memory to keep  
Forever etched in their hearts, as they sit and weep

So let us cherish these moments, with our four-legged friends  
For they bring us so much happiness, until the very end  
And even when they're gone, their love will live on  
In the memories we hold, forever and strong.



Can you write me a beautiful poem about this image?



Sure, here's a rap song based on the image you provided:

Verse 1:  
I'm an old lady, but I'm still flyin'  
I've got my headphones on and I'm feelin' high  
I'm in the studio, makin' some beats  
I'm the queen of hip hop, can't nobody touch my feet

Chorus:  
I'm the grandma of rap, I'm still spittin' fire  
I've got my green jacket on, I'm lookin' fly  
I've got my sunglasses on, I'm shining bright  
I'm the queen of the game, I'm makin' it right

Verse 2:  
I've been in the game for years, I've paid my dues  
I've got a microphone in my hand and I'm singin' the blues  
I've got the crowd in the palm of my hand  
I'm the queen of rap, I'm takin' over the land

Chorus:  
I'm the grandma of rap, I'm still spittin' fire  
I've got my green jacket on, I'm lookin' fly  
I've got my sunglasses on, I'm shining bright  
I'm the queen of the game, I'm makin' it right

Verse 3:  
I've got my beats in my headphones, I'm feelin' the vibe  
I'm in the zone, I'm in my element, I'm in my ride  
I'm the queen of rap, I'm in control  
I'm makin' hits, I'm takin' over the world

Chorus:  
I'm the grandma of rap, I'm still spittin' fire  
I've got my green jacket on, I'm lookin' fly  
I've got my sunglasses on, I'm shining bright  
I'm the queen of the game, I'm makin' it right

Outro:  
I'm the grandma of rap, I'm still spittin' fire  
I've got my green jacket on, I'm lookin' fly  
I've got my sunglasses on, I'm shining bright  
I'm the queen of the game, I'm makin' it right.



Can you write me a master rap song that rhymes very well based on this image?



图7:押韵生成



Can you write a story around this image for my child?



当然，这里有一个基于图片的故事：

很久很久以前，森林里住着一只棕熊、一只白兔和一只黑猫。他们是好朋友，喜欢在一起度过时光。一天，他们决定在树下野餐。熊带了一些蜂蜜，兔子带了一些胡萝卜，猫带了一些牛奶。

当他们围坐在桌子旁时，熊说：“我好饿，我可以吃下一整罐蜂蜜！”兔子咯咯地笑着说：“我太饿了，我可以吃下一整袋胡萝卜！”猫喵喵叫着说：“我好饿，我可以喝下一整碗牛奶！”

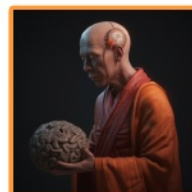
他们都笑了，开始大吃特吃。熊喜欢蜂蜜，兔子喜欢胡萝卜，猫喜欢牛奶。它们都非常快乐和满足。

吃完饭后，他们决定到森林里去散步。他们遇到一条小溪，熊说：“我们去游泳吧！”兔子和猫有点犹豫，但熊说服了它们一起去。

他们都在水里游泳和玩耍，玩得很开心。熊喜欢玩水花，兔子喜欢潜水，猫喜欢追逐水花。他们都很开心，玩得很开心。

游完泳后，他们决定回到野餐地点去吃更多的食物。他们围坐在桌子旁，熊说：“我好饿，我可以再吃下一整罐蜂蜜！”兔子咯咯地笑着说：“我好饿，我又能吃下一整袋胡萝卜了！”猫喵喵叫着说：“我好饿，我又可以喝一整碗牛奶了！”

他们都笑了，开始大吃起来。他们都非常高兴和满足。从此以后，他们过着幸福的生活。



I designed a character for my game as shown in the image. Can you write me a background story for him?



当然，这可能是图中人物的背景故事：

图像中的和尚是一个聪明而开明的人，他花了很多年的时间研究心灵的运作。他花了无数个小时冥想和思考意识的本质，因此，他对心灵的内部运作有了深刻的理解。


他手中的大脑象征着他心灵的深刻理解。他用它来帮助别人了解自己思想和情感的本质，并引导他们走上觉悟之路。

尽管他拥有巨大的智慧，但这位僧人是一位谦逊而富有同情心的人。他用他的知识去帮助别人，并且总是愿意与那些寻求他指导的人分享他的见解。在他的社区里，他是一个受人尊敬的人物，很多人都来找他寻求帮助和指导。


僧侣的长袍象征着他精神道路的承诺。他穿着僧袍是为了提醒自己贫穷、贞洁和服从的誓言，也象征着他致力于过一种服务他人的生活。

总的来说，图中的僧人是一位智慧而富有同情心的人，他一生致力于理解心灵的本质，并在开悟的道路上帮助他人。

图8:故事生成



We design new mugs shown in the image. Can you help us write an advertisement?




Sure, here's an advertisement for the mugs:


Introducing our new line of mugs featuring adorable cat designs! These mugs are perfect for coffee lovers and cat enthusiasts alike. Made from high-quality ceramic, each mug is durable and dishwasher safe. The unique designs are hand-painted, making each mug a one-of-a-kind piece of art.

Whether you're sipping your morning coffee or enjoying a warm drink on a chilly evening, these mugs are the perfect way to add a touch of feline charm to your daily routine. They also make great gifts for cat lovers in your life.

So why settle for a plain old mug when you can have a cat-themed one? Order yours today and start enjoying your beverages in style!




Help me write an advertisement for this.




Are you looking for a unique and stylish lighting solution for your home or office? Look no further than our brass toucan lamp! This beautifully crafted bird lamp is sure to add a touch of whimsy and charm to any room. The toucan's beak is designed to hold a light bulb, providing a warm and inviting glow. The base is made of wood, adding a natural touch to the overall design. The toucan lamp is a great addition to any room, whether you're looking to brighten up a dark corner or add a touch of personality to your space. Order yours today and bring a touch of the tropics to your home or office.

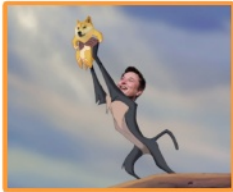
图9:产品广告




Who are the people in the image?



The people in the image are Goku and Vegeta, two characters from the Dragon Ball series. Goku is a saiyan warrior and the main protagonist of the series, while Vegeta is a saiyan prince and the main antagonist. They are both known for their powerful abilities and their rivalry.



Who is the person in the image?



The person in the image is Elon Musk, the CEO of Tesla and SpaceX. He is known for his work in the fields of electric cars, space exploration, and renewable energy. He is also known for his eccentric personality and outspoken views on various topics.

图10:个体识别



图11:图片评论



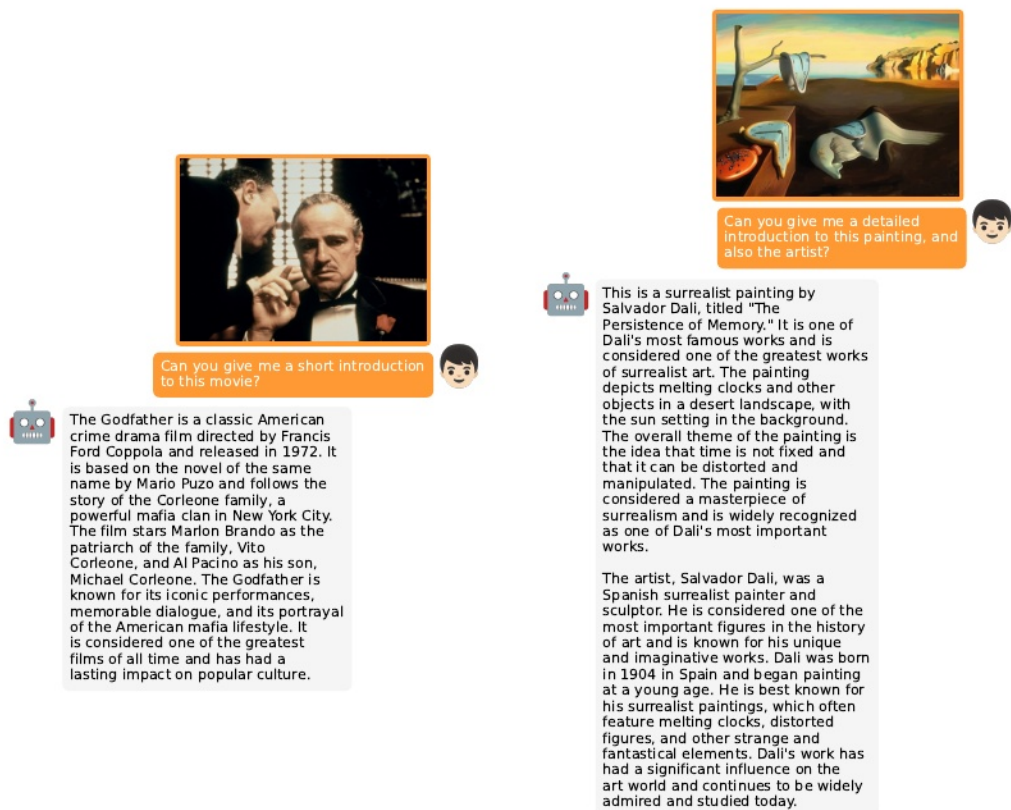


图12:事实检索



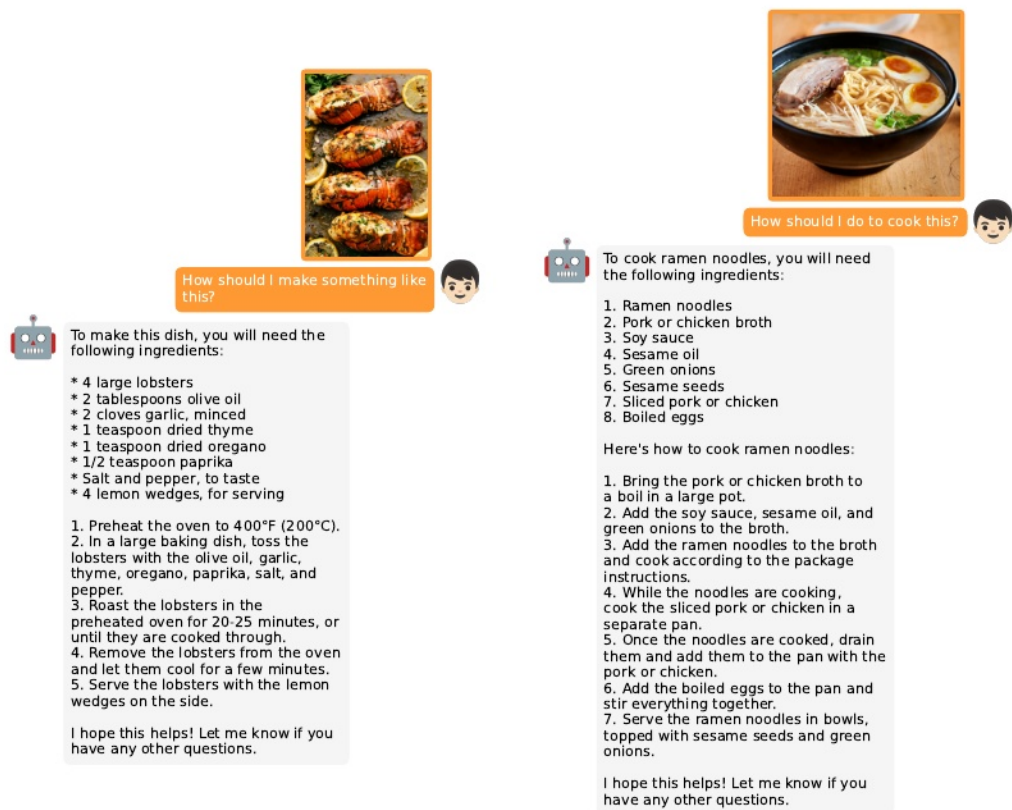


图13:食物食谱生成

## 参考文献。

- [1] Sharegpt. <https://github.com/domeccleston/sharegpt>, 2023.
- [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. In *Advances in Neural Information Processing Systems*, 2022.
- [3] Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. *arXiv preprint arXiv:2303.07274*, 2023.
- [4] 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.
- [5] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021.
- [6] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18030–18040, 2022.
- [7] Jun Chen, Deyao Zhu, Kilichbek Haydarov, Xiang Li, and Mohamed Elhoseiny. Video chatcaptioner: Towards the enriched spatiotemporal descriptions. *arXiv preprint arXiv:2304.04227*, 2023.
- [8] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90% \* chatgpt quality, March 2023.
- [9] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.
- [10] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.
- [11] 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.
- [12] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.
- [13] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022.
- [14] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- [15] Shaoan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023.
- [16] 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.
- [17] 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.
- [18] OpenAI. Introducing chatgpt. <https://openai.com/blog/chatgpt>, 2022.
- [19] OpenAI. Gpt-4 technical report, 2023.
- [20] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.
- [25] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.
- [26] 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.

- [27] 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.
- [28] Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. Using deepspeed and megatron to train megatron-turing nlq 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*, 2022.
- [29] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*, 2023.
- [30] 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. [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca), 2023.
- [31] Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven CH Hoi. Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training. *arXiv preprint arXiv:2210.08773*, 2022.
- [32] 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.
- [33] 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.
- [34] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. Survey Certification.
- [35] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023.
- [36] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *arXiv preprint arXiv:2206.08155*, 2022.
- [37] Zhengyuan Yang\*, Linjie Li\*, Jianfeng Wang\*, Kevin Lin\*, Ehsan Azarnasab\*, Faisal Ahmed\*, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. 2023.
- [38] 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.
- [39] Deyao Zhu, Jun Chen, Kilichbek Haydarov, Xiaoqian Shen, Wenxuan Zhang, and Mohamed Elhoseiny. Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. *arXiv preprint arXiv:2303.06594*, 2023.