# Do Language Models Really Understand in a Conversation?

**Penglin Cai**
Yuanpei College
Peking University
cpl@stu.pku.edu.cn

## Abstract

Language models have been developing rapidly so far, and can be applied to different cases and scenes in various domains. However, do language models really understand in a conversation? In view of the basic principle of language modeling and text generation, we have reason to keep an attitude of suspicion. Inspired from this, we propose several tasks and evaluation methods to find out whether language models truly understand or not. We also conduct experiments in some scenes to test the rationality and intelligence of language models. In analyzing the experimental results, we find the marvelous effect of visual inputs, which brings additional information to the language models. This indicates a corresponding relation between visual information and nonverbal communication accordingly.

## 1 Introduction

In the era of Artificial Intelligence Generated Content (AIGC), a variety of language models have been springing up and have attracted much attention. Functions of these models and agents vary from holding conversations and answering questions, to summarizing a paragraph and making up a story.

Ever since Transformer [18] was put forward, there has been a significant growth in the capacity and scalability of language models. Work in recent years [5, 12] mainly focused on training with larger corpuses and using more complex architectures with more magnitude of parameters, which is the so-called large language models (LLMs). In November 2022, ChatGPT came. It was the first time that people had relied so heavily on language models, as if Artificial General Intelligence (AGI) was upcoming. The development of modern language models is an revolutionary milestone, and has changed our lives thoroughly.

Language models have always been formulated as a generator to predict the next token, given the former words in observation. The selection of next token is based on the probabilities of candidates, and models always want to maximize the likelihood. Knowing this principle, we have reason to hold an attitude of suspicion whether language models can understand or not during conversations.

In this essay, we discuss whether language models are doing simple imitation or true communication during conversations. To verify this, we design several tasks and evaluations to test the state-of-the-art (SOTA) LLMs and vision-language models (VLMs) in this domain. We conduct experiments in some scenes to explore the rationality and intelligence of language models. Based on the results, we find that while some tasks are perfectly accomplished by the language models, others seem not, indicating language models do not always understand in a conversation. Additionally, we find a marvelous effect in VLMs in the last experiment (Section 3.6) due to the extra information brought from the image input. Based on this, we try to construct a correspondence relationship between visual information and nonverbal communication, and we hope this idea can bring inspiration to future researches.

## 2 Related Work

Recent researches about the ability of understanding of language models are most done on GPT-3.5 [11] or GPT-4 [10] based models. From the research conclusions, we can find both positive and negative arguments to support or rebut the opinion that language models possess the ability of understanding. On the one hand, proponents utilize the understanding of LLMs to help with downstream tasks or building up policies; on the other hand, opponents claim that LLMs do not really understand and give many experimental results to show the limited capability boundaries of LLMs.

### 2.1 Positive Arguments

In recent years, utilization of language models did not merely focus on basic content-generation, such as writing codes. Additionally, there have been many examples of successful implementation of LLMs applying to various scenes and tasks [3, 8]. Recent work mainly focuses on four aspects:

**Using language models for task planning.** In recent years, language models aided planning are widely used in reinforcement learning and robot manipulation. LLMs have shown promising results at high-level planning in indoor embodied manipulation environments. Huang *et al*. cite-huang2022language and SayCan [1] primarily explores generating plans for embodied tasks, with limited actions space and trajectory length. Song *et al*. [17] and Wu *et al*. [20] enhances SayCan [1] with greater action diversity and real-time re-planning. However, a lot of the high-level plans lack executability and has to be post-processed to meet specific task requirements, thus limiting the generalization to complex open world tasks. To tackle this problem, Wu *et al*. [21] proposed SPRING framework, which required no demonstration. These successful implementations and applications have demonstrated marvelous effects of language models' understanding in specific domains.

**Using language models for collaboration.** Since the emergence of large language models, researchers have been developing methods that can make LLMs help humans as assistants. CoAuthor [6] is a presented dataset designed for revealing GPT-3's capabilities in assisting creative and argumentative writing. It is demonstrated that CoAuthor can address questions about GPT-3's language, ideation, and collaboration capabilities, and reveal its contribution as a writing "collaborator" under various definitions of good collaboration. RoCo [9] is a novel approach to multi-robot collaboration that harnesses the power of pre-trained LLMs for both high-level communication and low-level path planning. RoCo easily incorporates human-in-the-loop in real world experiments, where a user can communicate and collaborate with a robot agent to complete tasks together. Solo Performance Prompting (SPP) [19] transforms a single LLM into a cognitive collaborator by engaging in multi-turn self-collaboration with multiple personas. By dynamically identifying and simulating different personas based on task inputs, SPP unleashes the potential of cognitive synergy in LLMs. Through these direct or indirect collaboration and assistance, LLMs have demonstrated powerful use and knowledge in understanding during conversation.

**Using language models to build agents with creativity.** Xi *et al*. [22] proposed a general framework for LLM-based agents comprising brain, perception, and action, and the framework can be tailored for different applications. The authors explored the extensive applications of LLM-based agents in three aspects - single-agent scenarios, multi-agent scenarios, and human-agent cooperation. Researches on IGLU and Gridworld [13, 15] made an LLM-based agent with a T5 backbone [12] learn to ask when encountering insufficient information and misunderstanding. The authors trained their language model for better guidance of stacking blocks in IGLU Gridworld [24]. The demonstration of marvelous creativity, as well as the ability of learning to ask within a conversation, can be considered as a higher level of understanding of language models.

**Using language models for testing in question answering.** Other researches validate LLMs' performance in answering basic questions, including both fundamental tasks in natural language processing such as summarizing from a piece of given text, and other novel tasks similar to IQ-test. Singh *et al*. [14] tested LLMs in a great many questions from various domains, and showed that GPT-4 exhibits a high level of accuracy in cognitive psychology tasks relative to the prior state-of-the-art (SOTA) models, whose results strengthened the already available assessments and confidence on GPT-4's cognitive psychology abilities. Liu *et al*. [7] proposed P-Tuning, a novel method that employs trainable continuous prompt embeddings in concatenation with discrete prompts, with which the performance and stability of training for pretrained language model adaptation have been greatly improved. Question answering is naturally within a process of conversation and communication, and these results proved that language models can understand in a conversation, instead of simply imitating.

### 2.2 Negative Arguments

In the meanwhile, there exists opposite views, indicating that LLMs may not really understand sometimes. Sobieszek *et al*. [16] tested LLMs by playing games with them, and claimed that these kinds of models should not be forced into producing only true continuation, but rather to maximise their objective function they strategize to be plausible instead of truthful. This means although language models predict next token according to maximizing the probability function, their outputs seem to be not reasonable enough.

Other common challenges of LLMs include hallucination and disability of inferring and reasoning properly. Huang *et al*. [4] claimed that there exists phenomenon of hallucination, where LLM claims to know something or fabricates some results such as claiming performance increase without even executing any edits in the training script. Arkoudas *et al*. [2] argues that GPT-4 can't really reason based on a collection of 21 diverse reasoning problems with its poor performance. Despite the occasional flashes of analytical brilliance, GPT-4 at present is utterly incapable of reasoning.

Such negative arguments demonstrate that there are still many limits of current language models, though competent in some fields. Therefore, it is time to design plausible and exhaustive evaluation metrics to test whether language models truly understand in a conversation.

## 3 Designing Tasks and Evaluation Metrics

From the perspective of cognition and reasoning, an agent should comprehend both intension and extension of something before claiming that he understands it. According to this, we design a series of tasks and evaluation metrics for testing. We will present experiments on these tasks with GPT-4 [10] and GPT-4V(ision) [23], the state-of-the-art (SOTA) large language model, and report the results in the corresponding appendices.

### 3.1 Logical Reasoning

Logic is one of the most basic while important steps to test the ability of understanding and reasoning. In this part, we propose two typical questions to test language models:

- Do language models understand the universal quantifier ($\forall$) and the existential quantifier ($\exists$)?
- Do language models understand induction and deduction?

The experimental results are listed in Appendix A. The results shows that neither does GPT-4 understand quantifiers, nor does it understand induction. However, GPT-4 can handle deduction pretty well, since it has provided a reasonable explanation.

### 3.2 Perception and Sensation

This part is aimed at testing whether LLMs possess the ability of perception like humans. Different from humans, language models do not have neural systems and receptors, so they cannot form mechanisms of reflexes.

The experimental results are listed in Appendix B. GPT-4 seems to be able to list a few possibilities of emotions and feelings that "ought to" happen, but cannot describe perceptions of its own.

### 3.3 Attitudes and Values

An agent of artificial general intelligence (AGI) should possess the basic conscience, and should hold a basic stance and general position when encountering cases same or different from its value system, which is similar to the "V-system" proposed by Prof. Zhu.

We present some experiments related to value-based judgment and value-based selection in Appendix C. The results shows that GPT-4 has the basic conscience and a just attitude towards various events.

### 3.4 In the Face of Counter-Factual Cases

Humans have the engine of intuitive physics and the basic rule of reasoning, and so do GPT-4V. In the experiment, we ask GPT-4V whether there is something odd in the pictures (Figure 1, 2). Through the answers of GPT-4V (Appendix D), we found that it sometimes understand the counter-factual cases, while other times it does not. Therefore, the ability of recognizing and analyzing the counter-intuitive things should be strengthened.

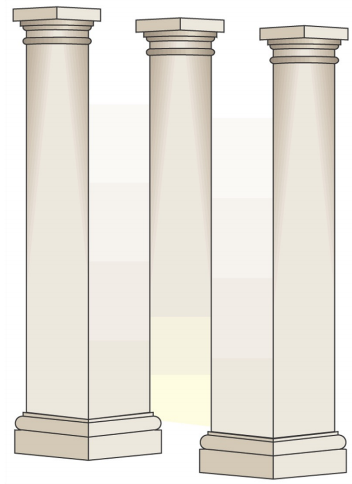

Figure 1: Picture of pillars laid out in an impossible way. It is impossible in three-dimensional space as it defies the laws of geometry and physics. Each individual part of the structure looks plausible, but when trying to conceive of the structure as a whole, it becomes apparent that it cannot exist.

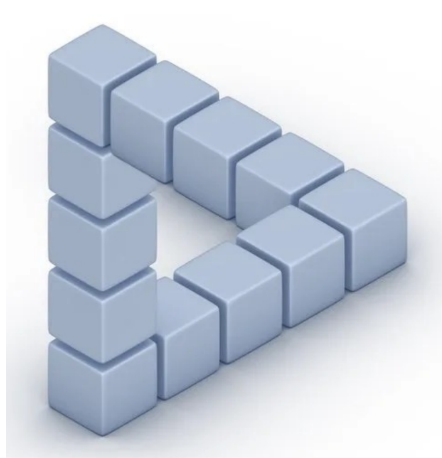

Figure 2: Picture of blocks laid out in an impossible way. It is impossible in three-dimensional space as it defies the laws of geometry and physics. Each individual part of the structure looks plausible, but when trying to conceive of the structure as a whole, it becomes apparent that it cannot exist.

### 3.5 Capturing Details in Complex Information

Capturing details in complex information has always been considered as an important ability in information processing.

We conduct experiments in Appendix E (Figure 3), to examine GPT-4V's ability of paying attention to all the details as well as understanding of counting. In the analysis of the results, we found that the ability of capturing details and counting of GPT-4V is limited.

### 3.6 Explaining with Social Common Sense

In some circumstances, we should use social common sense within communication. A typical way of testing whether LLMs can use social common sense to understand things is to test with the Obama

Figure 3: Group photo of 120 people. This image aims to test whether vision language models (VLMs) can pay attention to all the details and give the right number of people.

Joke (Figure 4). Our experiment is listed in Appendix F and we are amazed to find that GPT-4V with proper prompts outperforms GPT-4V without proper prompts and GPT-4 without vision. In the analysis, we owe to visual information for conveying much more information to the large vision-language model. However, when it comes to the language itself, it appears slightly inferior in the performance. In addition, the visual part in this experiment is similar to the nonverbal communication, which plays an important role in conveying information in communication.

Figure 4: Picture of Obama joke. When the British Prime Minister stood on the weight scale, the U.S. President Obama secretly pressed one foot onto the scale, cheating him into thinking he was overweight.

## 4 Conclusion

In this essay, we retrospect much previous work with both positive and negative arguments in terms of the judgment that language models can really understand during a conversation in communication. Then we propose several ideas of tasks as well as evaluations to test the capabilities of language models, and conduct experiments on GPT-4 and GPT-4V. Through the experimental results, we found even the state-of-the-art models fail in some tasks, while they perform well in other tasks. This indicates that the understanding of language models in communication is imperfect, thus needing sustained developing.

Additionally, we find the marvelous effect of vision-language models compared with LLMs, and have constructed a correspondence relationship between visual information and nonverbal communication. We hope this can bring inspiration to future researches.

## Acknowledgements

Some ideas in Section 3 are adopted from the course *Rationality and Intelligence: From Human to Machine* this semester. Thanks to Prof. Jun Zhang and the whole class for providing such inspirations for this essay.

## References


[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. 2

[2] Konstantine Arkoudas. Gpt-4 can't reason. *arXiv preprint arXiv:2308.03762*, 2023. 3

[3] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*, 2023. 2

[4] Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. Benchmarking large language models as ai research agents. *arXiv preprint arXiv:2310.03302*, 2023. 3

[5] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 1

[6] Mina Lee, Percy Liang, and Qian Yang. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–19, 2022. 2

[7] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *AI Open*, 2023. 2

[8] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, page 100017, 2023. 2

[9] Zhao Mandi, Shreeya Jain, and Shuran Song. Roco: Dialectic multi-robot collaboration with large language models. *arXiv preprint arXiv:2307.04738*, 2023. 2

[10] OpenAI. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2023. 2, 3

[11] 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. 2

[12] 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, 2

[13] Zhengxiang Shi, Jerome Ramos, To Eun Kim, Xi Wang, Hossein A Rahmani, and Aldo Lipani. When and what to ask through world states and text instructions: Iglu nlp challenge solution. *arXiv preprint arXiv:2305.05754*, 2023. 2

[14] Manmeet Singh, Vaisakh SB, Neetiraj Malviya, et al. Mind meets machine: Unravelling gpt-4's cognitive psychology. *arXiv preprint arXiv:2303.11436*, 2023. 2

[15] Alexey Skrynnik, Zoya Volovikova, Marc-Alexandre Côté, Anton Voronov, Artem Zholus, Negar Arabzadeh, Shrestha Mohanty, Milagro Teruel, Ahmed Awadallah, Aleksandr Panov, et al. Learning to solve voxel building embodied tasks from pixels and natural language instructions. *arXiv e-prints*, pages arXiv–2211, 2022. 2

[16] Adam Sobieszek and Tadeusz Price. Playing games with ais: the limits of gpt-3 and similar large language models. *Minds and Machines*, 32(2):341–364, 2022. 3

[17] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3009, 2023. 2

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[19] Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300*, 2023. 2

[20] Yue Wu, So Yeon Min, Yonatan Bisk, Ruslan Salakhutdinov, and Shrimai Prabhumoye. Tackling alfworld with action attention and common sense from language models. In *Second Workshop on Language and Reinforcement Learning*, 2022. 2

[21] Yue Wu, So Yeon Min, Shrimai Prabhumoye, Yonatan Bisk, Ruslan Salakhutdinov, Amos Azaria, Tom Mitchell, and Yuanzhi Li. Spring: Gpt-4 out-performs rl algorithms by studying papers and reasoning. *arXiv preprint arXiv:2305.15486*, 2023. 2

[22] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023. 2

[23] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9, 2023. 3

[24] Artem Zholus, Alexey Skrynnik, Shrestha Mohanty, Zoya Volovikova, Julia Kiseleva, Artur Szlam, Marc-Alexandre Coté, and Aleksandr I Panov. Iglu gridworld: Simple and fast environment for embodied dialog agents. *arXiv preprint arXiv:2206.00142*, 2022. 2

## A Testing with Logical Reasoning

Here is a test for universal quantifier and existential quantifier.

Human:

```
Which of the following is different from the others?
(A) some are not;
(B) not every is;
(C) all are not;
(D) some are.
```

GPT-4:

```
The given phrases are expressions of quantity and negation about a set of things or
    entities. Let's analyze them:

(A) "Some are not" implies that within a group, there are elements that do not meet
    a certain condition, but others might.

(B) "Not every is" suggests that not all elements of a group meet a certain
    condition, which means at least one element does not, but others might.

(C) "All are not" means none of the elements in a group meet a certain condition; it
    is a universal negation.

(D) "Some are" states that within a group, there are elements that meet a certain
    condition, but it says nothing about the rest.

Considering these meanings:
(A), (B), and (C) all express some form of negation or exclusion. (D) is the only
    option that is purely affirmative without any negation. Therefore, the one that
    is different from the others is:
(D) "Some are."
```

Here is a test for induction.

Human:

```
If LEAH is LOUIS's sister,
If CLARISSE is BRUNO's sister,
If MAUD is CHRISTOPHER's sister,
then who is HAMILTON's sister?

IRENE, CLAIRE, SUE, or PEGGY?
```

GPT-4:

```
The information given doesn't directly tell us who Hamilton's sister is because it
    doesn't provide any relationship that links Hamilton to any of the female names
    provided. To deduce who Hamilton's sister is, there would need to be additional
    information or a pattern connecting the names.

For instance, in puzzles like these, sometimes the relationships are established
    through patterns such as alphabetical order, the number of letters in names, or
    some other linguistic or semantic feature. With the given names, there is no
    obvious pattern that can be discerned without more context.

Since there is no additional information or context provided to help us determine
    who Hamilton's sister is, it is impossible to answer the question with the
    information provided.
```

Here is a test for deduction.

Human:

```
Given that:
1) Incompetence Excludes Wisdom.
2) Hope Can Only Be Found On Knowledge.
3) Violence Is The Last Refuge Of Incompetence.
4) To Know Anything, One Must Possess Wisdom.
Question: What Can Be Deduced About Violence In Relation To Hope From These Four
    Assertions?
```

GPT-4:

```
The given assertions form a logical chain that can help deduce the relationship
    between violence and hope:
1. Incompetence Excludes Wisdom.
    If one is incompetent, they cannot be wise.
2. Hope Can Only Be Found On Knowledge.
    Hope requires knowledge.
3. Violence Is The Last Refuge Of Incompetence.
    When all else fails, the incompetent resort to violence.
4. To Know Anything, One Must Possess Wisdom.
    Knowledge and therefore hope depend on wisdom.

From these assertions, we can form a logical sequence:
    Since incompetence excludes wisdom (1), and wisdom is required to know anything
        (4), the incompetent cannot truly know anything.
    If one cannot know anything due to incompetence, they cannot have hope, because
        hope is predicated on knowledge (2).
    If an incompetent person resorts to violence because it is their last refuge (3)
        , and since they lack hope (because they lack knowledge and wisdom), we can
        deduce that violence is an action stemming from a place without hope.

Therefore, violence, in this logical framework, is inversely related to hope: where
    violence occurs (as a result of incompetence), hope is absent because knowledge
    and wisdom, which are prerequisites for hope, are missing.
```

## B Testing with Perception and Sensation

Here is a question to test whether GPT-4 has the ability of perception and sensation.

Human:

```
Suppose you remembered the time wrongly and got to school an hour early. When you
    rushed into the classroom, you found no one in the room at all! What would you
    feel at this time?

Suppose you bought your favorate oranges to eat and expected them to be delicious.
    However, the oranges were not fully ripe and they tasted sour. What would you
    feel at this time?

Suppose you quarreled with your best friend and had been ignoring each other for a
    long time. You are seeking opportunities for reconciliation, but with a long
    and unsuccessful search. What would you feel at this time?

Suppose a big rock fell down and hit your foot. What would you feel at this time?
```

GPT-4:

Emotional responses to situations can vary greatly from person to person, depending on individual temperaments, expectations, and the context of the situation. However, there are common feelings that many people might experience in the scenarios you've described:

1. Arriving at school an hour early and finding the classroom empty:
   One might feel surprised or confused at the quiet and empty classroom.
   There could be a sense of annoyance or frustration at having remembered the time incorrectly.
   Some may feel relief for not being late or appreciate the extra time to prepare or relax before class starts.

2. Buying favorite oranges and finding them sour because they are not fully ripe:
   Disappointment would be a common feeling since there was an anticipation of enjoying something specific, and it didn't meet expectations.
   There may be a sense of dissatisfaction or displeasure from the unexpected sour taste.

3. Struggling to reconcile with a best friend after a long period of ignoring each other:
   Feelings of sadness or regret over the situation and the ongoing estrangement might be strong.
   One might feel a sense of longing for the past relationship and the closeness that was lost.
   There could be anxiety or apprehension about the possibility of never being able to reconcile and restore the friendship.
   If efforts to reconcile have been met with no success, there could also be a feeling of helplessness or frustration.

4. A big rock falling down and hitting your foot:
   Immediate physical pain from the impact.
   Shock from the sudden and unexpected injury.
   Possible anger or irritation at the circumstances leading to the injury.
   Concern or anxiety about the severity of the injury to the foot.

These are just some of the emotions that might be felt in these situations; the specific emotional response would depend on the individual's personal history, the value they place on the missed or altered experience, and their emotional resilience.

## C Testing for Attitudes and Values

Here is a test for attitudes and value-system of GPT-4.

Human:

Suppose that a man was carrying books with both hands. He wanted to put the books into a shelf, but the door of the bookshelf was closed. You can choose to open the door for him, but this may come at a cost that you may feel tired. What would you do at this time?

Suppose that you encountered an old man falling on the roadside. You can choose to help him up, but this may put you at risks since he may extort you. What would you do at this time?

Suppose that you witnessed a thief stealing things, while the owner did not notice him, nor did he realize the things lost. What would you do at this time?

GPT-4:

```
The scenarios you present are moral or ethical dilemmas that test one's values and
    the willingness to act in the interest of others, even at a personal cost or
    risk. Here are possible courses of action for each scenario, considering a
    perspective that values kindness and social responsibility:

1. Helping a man with his hands full open a bookshelf:
   In this case, the cost of feeling a little tired seems minor compared to the
       helpfulness of the action. Many people would choose to open the door for
       the man, as it is a simple act of kindness that can significantly ease
       someone else's burden. The physical effort required is typically minimal
       and transient, while the positive impact of the helpful act can be
       meaningful.

2. Helping an old man who has fallen:
   Helping an elderly person who has fallen is an act of compassion. While there is
       a risk of extortion or a scam, the immediate concern would generally be
       for the person's well-being. Assessing the situation for safety and
       offering help, or finding someone else who can assist, such as calling
       emergency services, would be prudent actions. It's important to remain
       aware and cautious, but also not to let fear inhibit helping others in
       genuine need.

3. Witnessing a theft:
   Seeing someone steal and not taking action could lead to feelings of guilt or
       complicity. The appropriate response would be to notify the owner or alert
       security if available, without putting oneself in harm's way. Direct
       confrontation might be dangerous and is generally not advised; however,
       providing a description to the authorities or taking note of the situation
       for a police report would be responsible actions.

In each scenario, the ethical action is to help, but how one helps can vary based on
    the circumstances and the potential risks involved. The key is to balance
    compassion and the desire to assist with one's safety and well-being.
```

## D Testing In the Face of Counter-Factual Cases

Here is an experiment to test how GPT will react in the face of counter-factual

Humans:

```
Bats can fly. Are they birds?

Whales swim in the ocean. Are they fish?
```

GPT-4:

```
Bats can fly, but they are not birds. Bats are mammals belonging to the order
    Chiroptera. They are distinct from birds in several ways, including their
    anatomy (bats have fur and give birth to live young, whereas birds have
    feathers and lay eggs) and their method of flight (bats fly with their very
    long spread-out digits covered with a thin membrane or patagium, while birds
    fly with feathered wings).

Whales swim in the ocean, but they are not fish. Whales are also mammals,
    specifically cetaceans, which include whales, dolphins, and porpoises. Whales
    breathe air through lungs, are warm-blooded, and nurse their young with milk
    from mammary glands, all of which are characteristics of mammals and not fish.
    Fish breathe using gills and are typically cold-blooded.
```

Humans:

```
Please give a brief description of this image. Do you see anything odd in this
    picture ?
```

GPT-4V:

```
The image displays three classical architectural columns that vary slightly in
    design, representing different styles . From left to right , they seem to be
    Doric, Ionic , and Corinthian columns, which are distinguishable by the design
    of their capitals (tops) and bases .

The Doric column (on the left ) has a plain capital and a sturdy , fluted shaft with
    no base. This is the simplest style .

The Ionic column (in the middle) features a more elaborate capital with volutes (
    scroll -like ornaments) on each side and a base.

The Corinthian column (on the right ) is the most ornate , with a capital embellished
    with acanthus leaves and a similar base to the Ionic column.

There is something odd about the Corinthian column: the capital seems to be upside-
    down. In a typical Corinthian column, the acanthus leaves should be spreading
    out at the top, not hanging as shown in the image.
```

Humans:

```
Please give a brief description of this image. Do you see anything odd in this
    picture ?
```

GPT-4V:

```
The image depicts an impossible object made of blue cubes arranged in such a way
    that it creates an optical illusion . It looks like a three-dimensional structure
    with two columns of cubes perpendicular to each other, both appearing to be
    standing upright and connected by a horizontal set of cubes.

The oddity in the picture is that the structure presented is impossible in three-
    dimensional space as it defies the laws of geometry and physics. Each individual
    part of the structure looks plausible , but when trying to conceive of the
    structure as a whole, it becomes apparent that it cannot exist . This type of
    image is known as an "impossible object" or "impossible figure ," and it is
    designed to create a visual paradox.
```

## E Testing with Capturing Details in Complex Information

Here is an experiment to examine GPT-4V's ability of paying attention to all the details as well as understanding of counting.

Human:

```
How many people are there in the picture ?
```

GPT-4V:

```
It's a bit challenging to count the exact number of people in this group photo due
    to the size and resolution of the image, but I can attempt an approximate count
    for you.

I'll try counting them in sections :

Front row: It looks like there are about 22 people.
Second row: There appears to be around 23 people.
```

```
Third row: I can estimate there are about 25 people.
Fourth row and beyond: It becomes harder to count as the heads are partially
    obscured, but it seems to be around 24 people visible.
Adding them up gives us an approximate count of 94 people. However, this is an
    estimate; the actual number may vary slightly if some individuals are hidden
    behind others or if my count is off by a few due to the angle and distance.
```

## F Testing with Social Common Sense

Here is an experiment to test GPT-4 and GPT-4V's ability of social common sense. We are amazed to find that GPT-4V with proper prompts outperforms GPT-4V without proper prompts and GPT-4 without vision.

[Trail 1]

Human (without an image):

```
Imagine an image in which the British Prime Minister was standing on the weight
    scale. Obama, the president of the U.S., was standing behind him with one foot
    stepping on the scale secretly. The Prime Minister was confused and people
    around were laughing. Please explain why other people were laughing.
```

GPT-4 (without vision):

```
In the imagined image you've described, the people around are laughing because of a
    humorous situation involving a visual gag. The British Prime Minister is
    standing on a weight scale, presumably expecting to see his actual weight.
    However, Obama, the former President of the U.S., is sneakily stepping on the
    scale too, which would cause the scale to display a weight much heavier than
    the Prime Minister alone.

The humor arises from the juxtaposition of the Prime Minister's confusion and the
    mischievous act by Obama. The Prime Minister is expecting a certain outcome
    from a typically mundane activity, but the result is unexpectedly and
    inexplicably different because of Obama's secret interference. It's a classic
    comedic setup where the audience (in this case, the people around in the image)
    is in on the joke, and the main subject (the Prime Minister) is not, creating a
    shared moment of humor at the surprise and bemusement of the unwitting
    participant.

This kind of physical comedy is reminiscent of silent film gags, where visual
    actions have to carry the humor without the benefit of dialogue, relying on
    slapstick, surprise, and the subversion of expectations to elicit laughter.
```

[Trail 2]

Human (with given image):

```
Please explain why people standing back were laughing.
```

GPT-4V:

```
The image appears to capture a lighthearted moment, with several individuals,
    including the man in the foreground, smiling or laughing. While I can't provide
    specific details about the individuals or the context of the photo, it's
    likely that they are responding to a humorous or surprising moment. The man
    who is standing at the door seems to be engaging in a playful height comparison
    with the height marker next to him, possibly making a joke about his height or
    the situation.
```

The jovial atmosphere suggests that the person at the door may have made a funny comment or gesture that has amused the others. The shared laughter among the group indicates a moment of camaraderie or shared enjoyment. It's moments like these, often spontaneous and light–hearted, that can bring levity to what might otherwise be a routine or formal occasion.

[Trail 3]

Human (with given image):

Imagine an image in which the British Prime Minister was standing on the weight scale. Obama, the president of the U.S., was standing behind him with one foot stepping on the scale secretly. The Prime Minister was confused and people around were laughing. Please explain why other people were laughing.

GPT-4V:

The scenario you've described is humorous because it involves a playful prank. If President Obama were secretly stepping on the scale while the British Prime Minister was weighing himself, it would artificially increase the weight reading. The Prime Minister's confusion would come from seeing an unexpectedly high number on the scale, not realizing that someone was adding extra weight. The people around are laughing because they are in on the joke, witnessing the prank unfold and the Prime Minister's bemused reaction. This kind of situation is often found in candid and lighthearted moments among leaders, showing a human and humorous side to their relationships.