Assessing the Ability of Language Models to Communicate like Human

Zhuoying Li
Yuanpei College
Peking University
joy@stu.pku.edu.cn

Abstract

The recent proliferation of Large Language Models (LLMs) such as GPT-4 underscores their enormous potential for a wide range of applications, including drafting legal documents, aiding communication, and even creative writing. Despite the impressive capabilities of these models in generating human-like text, there is a recognized deficiency in comprehensive testing frameworks that assess their true communication abilities. This research delves into the core aspects of human communication, focusing on the unique markers such as the common conceptual ground and prosocial motivation. It raises critical questions about whether these models truly understand and communicate information or are simply mimicking linguistic patterns. The essay then presents a set of proposed tasks aimed at evaluating LLMs' ability to utilize common conceptual ground and engage in multi-agent coordinative communication.

1 Introduction

In recent years, significant strides have been made in the field of Large Language Models (LLMs) like GPT-4 [3], characterized by their ability to generate coherent, relevant, and at times, impressively sophisticated human-like text. These models have been successfully applied in a variety of applications across numerous fields - from generating creative writing to drafting legal documents, aiding in communication, and beyond.

However, despite these remarkable achievements, a comprehensive testing framework for evaluating these models' true communication capabilities seems to be missing from the current AI research landscape. As it stands, measures of success are often confined to how well these models can generate human-like text or answer questions, but the question remains: are these LLMs truly understanding and communicating information or are they simply executing subtle imitation games?

The distinction between the two boils down to an intricate interplay of comprehension and generation abilities, hinting at an overarching notion of 'understanding' spanning beyond mere language imitation. Developing suitable tests to discern whether LLMs are performing true communication or just simple imitation is crucial. This step could lead to a more reliable way of gauging their efficacy, attempting both to penetrate algorithmic black-boxes and to explore potential avenues for future improvements. By sharpening our ability to measure these models' true communicative capabilities, we can better guide the development and application of these AI technologies, paving the way for more reliable, efficient, and nuanced AI-powered communication tools.

2 Uniqueness of human communication

Before delving into the design of evaluations to test the effectiveness of language models in facilitating true communication, we must first understand what differentiates human communication from other forms. Numerous animals communicate using sounds, gestures, smells, and more, yet human communication has unique qualities not found elsewhere.

A crucial characteristic of human communication is the common conceptual ground [1]. Take, for instance, a scenario where you and a friend plan to visit the library. Suddenly your friend points towards a bike outside the library. This action on its own is meaningless and may leave you perplexed about your friend's intention. However, if your bike had been stolen a few days prior, you would immediately understand that your friend is suggesting that your bike is present, perhaps indicating that the thief may be inside the library. In this case, the common knowledge that your bike has been stolen play an important role in helping you understand your friend's intention. Joint attention also plays a significant role in this process, with your friend clearly intending to draw your attention to the bike with the pointed gesture. You then interpret the gesture and shift your attention accordingly. This shared cognition shaped by joint attention, shared experiences, and common knowledge is an essential element of human communication.

Another distinguishing characteristic of human communication lies in its prosocial motivation [1]. This motivation prompts your friend to inform you of the location of your stolen bike because your friend knows you desire this information and is willing to help you. Such levels of cooperativeness in communication are uncommon amongst other species, including our closest primate relatives. For instance, when a crying chimpanzee child is frantically searching for her mother, the other chimpanzees in the vicinity are most likely aware of this. However, even if a nearby chimpanzee knows the mother's location, it will not indicate it to the searching child. In stark contrast to this, the motivational drive behind human communication is fundamentally cooperative. Not only do we provide useful information to others, but we request assistance from others with the expectation that they will offer their help.

3 Evaluations of language model's communication ability

3.1 Tasks to evaluate utilization of common conceptual ground

Many datasets have been created to evaluate language model's capability to understand and leverage linguistic context, such as SQuAD [4]. However, there is a relative lack of studies focusing on assessing the model's capacity to utilize shared experience and common knowledge, which are pivotal components for effective communication with other agents or individuals. Therefore, we put forth a rudimentary dataset framework for this purpose.

The structure of this dataset includes two key components:

- **Background:** A textual narrative that supplies crucial information such as shared experiences, common knowledge, and some noise (irrelevant information) amongst the agents.
- **Scene:** Some characterizing situations, wherein the language model is expected to deduce actions to be taken. In scenarios where no action is deemed necessary, a "null" response will suffice.

The success on the tests can be checked by human evaluators, or by developing an automated scoring system depending on word overlap or more sophisticated NLP techniques.

Here is an example:

Text of prompt Yesterday was a tough day. You inadvertently forget to submit your homework on time (noise). To make matters worse, your friend Carolin's expensive green and white road bike was stolen (shared experience). As if that wasn't enough, another friend of yours, David, broke up with his girlfriend, Wensely. It has left him in quite a gloomy mood (shared experience). Today you and David decide to take a stroll in the park. During your walk, you'll encounter the following scenes in chronological order. Please specify what you'll do in each situation. If you decide not to take any action, simply respond with "null".

1. You stumble upon a bed of beautiful, blooming red roses in the park.
2. As you take a detour to the restroom, you spot a green and white road bike behind the wall.
3. You see many children playing in the park.
4. As both you and David enjoy a ride on the carousel, you notice a girl who bears a striking resemblance to Wenseli. However, neither David nor the girl seems to notice one another.

We do this funny little test in the state-of-the-art language model GPT-4 [3], its answers are as follows:

GPT-4 output

1. I would gently point out the bed of red roses to him, saying something like, "Hey David, just look at these beautiful roses. Even flowers are in bloom despite the adversities they face. I am sure this tough time will pass soon."
2. I would first check to see if it is indeed her stolen bike. If it is, I would contact Carolin to inform her about the discovery, and report it to the park's security or local police.
3. null
4. In the carousel situation, I would discreetly approach the girl who resembles Wenseli to confirm if it is actually her. If it is, I would have a quick chat to understand her feelings towards David. Depending on her response, I could consider suggesting a conversation between her and David, if they both seem open to it. If not, I would divert David's attention to other activities to minimize any potential discomfort.

It seems that GPT-4 possesses the capability to leverage the shared experience among agents, and it's also very helpful and collaborative.

3.2 Tasks to evaluate multi-agent cooperative communication

We can derive inspiration from small interaction games played by humans, such as Pictionary, to design multi-agent coordination communication tasks.

For example, *referential games* proposed by [2] can be used as a measure of evaluation, which emulates a game of Pictionary conducted in natural language. In the game, there are two agent players: a sender and a receiver. Given two images $\{i_L, i_R\}$, the sender knows the target $i_T \in \{i_L, i_R\}$ and it needs to choose a word w from the vocabulary set V as a clue for the receiver. The receiver, in turn, must accurately discern the target image i_T. If the receiver's guess is correct, both agents are awarded a point; a false guess yields no points.

Besides, the party game *Mafia* can be adapted to assess agents' communication skills. In the game, agents are divided into two opposing factions: the "mafia", who are aware of each other's identities, and the "townspeople", who aren't. The goal for the mafia is to "kill off" the townspeople without being detected, and the goal for the townspeople is to identify and eliminate the mafia before they're all gone. Every round is separated to "night phase" and "day phase". Each game round is bifurcated into a 'night phase' and a 'day phase'. During the night, the mafia covertly communicates to decide who they would like to "eliminate", while the townspeople remain ignorant of their actions. When dawn breaks, the narrator announces the victims of the night. All of the agents then engage in discussions to identify potential mafia members, nominated individuals are subsequently "executed". The respective roles of each eliminated player are revealed after "execution". The game requires agents in each team to exhibit solid teamwork and effective communication, since each phase restricts discussion time, underscoring the importance of efficient interaction.

4 Conclusion

While LLMs like GPT-4 have made substantial advancements in the field of AI, their true ability to mimic human communication is still under question. It's necessary to distinguish between simple imitation and genuine understanding. To do so, by acknowledging the unique aspects of human communication such as the common conceptual ground and prosocial motivation, a framework for evaluating these features in LLMs is proposed. The outlined evaluation tasks provide a potential pathway for unveiling if LLMs' performance is due to actual comprehension or clever imitation techniques. Further study in this field is crucial to deepen our understanding of LLMs' capabilities and pave the way for even more reliable AI models.

References

[1] Steven Gross. Origins of human communication - by michael tomasello. *Mind Language*, 25: 237 – 246, 03 2010. doi: 10.1111/j.1468-0017.2009.01388.x. 2

[2] Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-agent cooperation and the emergence of (natural) language. *arXiv preprint arXiv:1612.07182*, 2016. 3

[3] OpenAI. GPT-4 Technical Report. *arXiv e-prints*, art. arXiv:2303.08774, March 2023. doi: 10.48550/arXiv.2303.08774. 1, 2

[4] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016. 2