10 Directions and Conclusions

We have presented our initial exploration of GPT-4 across a wide range of tasks and domains, providing supporting evidence to the claim that GPT-4's abilities are comparable to human-level for many of them. This conclusion is consistent with the findings by OpenAI presented in [Ope23]. A primary goal of our experiments is to give a preliminary assessment of GPT-4's *intelligence*, which is an arduous task given the lack of formal definition for this concept, especially for artificial systems. We hope that our exploration provides a useful and necessary first step to appreciate the remarkable capabilities and challenges of GPT-4, and that it opens up new opportunities for developing more formal and comprehensive methods for testing and analyzing future AI systems with such broad intelligence. The capabilities of the model, which have been demonstrated above, both in terms of depth and generality, suggest that the machine learning community needs to move beyond classical benchmarking via structured datasets and tasks, and that the evaluation of the capabilities and cognitive abilities of those new models have become much closer in essence to the task of evaluating those of a human rather than those of a narrow AI model. We hope our investigation stimulates further research on GPT-4 and similar systems, both in terms of exploring new applications and domains, and in terms of understanding the mechanisms and principles that underlie their intelligence.

The central claim of our work is that GPT-4 attains a form of *general* intelligence, indeed showing *sparks of artificial general intelligence*. This is demonstrated by its core mental capabilities (such as reasoning, creativity, and deduction), its range of topics on which it has gained expertise (such as literature, medicine, and coding), and the variety of tasks it is able to perform (e.g., playing games, using tools, explaining itself, ...). A lot remains to be done to create a system that could qualify as a complete AGI. We conclude this paper by discussing several immediate next steps, regarding defining AGI itself, building some of missing components in LLMs for AGI, as well as gaining better understanding into the origin of the intelligence displayed by the recent LLMs.

10.1 Definitions of intelligence, AI, and AGI

In this paper, we have used the 1994 definition of intelligence by a group of psychologists [Got97] as a guiding framework to explore GPT-4's artificial intelligence. This definition captures some important aspects of intelligence, such as reasoning, problem-solving, and abstraction, but it is also vague and incomplete. It does not specify how to measure or compare these abilities. Moreover, it may not reflect the specific challenges and opportunities of artificial systems, which may have different goals and constraints than natural ones. Therefore, we acknowledge that this definition is not the final word on intelligence, but rather a useful starting point for our investigation. There is a rich and ongoing literature that attempts to propose more formal and comprehensive definitions of intelligence, artificial intelligence, and artificial general intelligence [Goe14, Cho19], but none of them is without problems or controversies. For instance, Legg and Hutter [Leg08] propose a goal-oriented definition of artificial general intelligence: Intelligence measures an agent's ability to achieve goals in a wide range of environments. However, this definition does not necessarily capture the full spectrum of intelligence, as it excludes passive or reactive systems that can perform complex tasks or answer questions without any intrinsic motivation or goal. One could imagine as an artificial general intelligence, a brilliant oracle, for example, that has no agency or preferences, but can provide accurate and useful information on any topic or domain. Moreover, the definition around achieving goals in a wide range of environments also implies a certain degree of universality or optimality, which may not be realistic (certainly human intelligence is in no way universal or optimal). The need to recognize the importance of priors (as opposed to *universality*) was emphasized in the definition put forward by Chollet in [Cho19] which centers intelligence around skill-acquisition efficiency, or in other words puts the emphasis on a single component of the 1994 definition: learning from experience (which also happens to be one of the key weaknesses of LLMs). Another candidate definition of artificial general intelligence from Legg and Hutter [LH07] is: a system that can do anything a human can do. However, this definition is also problematic, as it assumes that there is a single standard or measure of human intelligence or ability, which is clearly not the case. Humans have different skills, talents, preferences, and limitations, and there is no human that can do everything that any other human can do. Furthermore, this definition also implies a certain anthropocentric bias, which may not be appropriate or relevant for artificial systems. While we do not adopt any of those definitions in the paper, we recognize that they provide important angles on intelligence. For example, whether intelligence can be achieved without any agency or intrinsic motivation is an important philosophical question. Equipping LLMs with agency and intrinsic motivation is a fascinating and important direction for future work. With

this direction of work, great care would have to be taken on alignment and safety per a system's abilities to take autonomous actions in the world and to perform autonomous self-improvement via cycles of learning. We discuss a few other crucial missing components of LLMs next.

10.2 On the path to more general artificial intelligence

Some of the areas where GPT-4 (and LLMs more generally) should be improved to achieve more general intelligence include (note that many of them are interconnected):

- **Confidence calibration:** The model has trouble knowing when it should be confident and when it is just guessing. It both makes up facts that have not appeared in its training data, and also exhibits inconsistencies between the generated content and the prompt, which we referred to as *open-domain* and *closed-domain* hallucination in Figure 1.8. These hallucinations can be stated in a confident and persuasive manner that can be difficult to detect. Thus, such generations can lead to errors, and also to confusion and mistrust. While hallucination is a good thing when generating creative content, reliance on factual claims made by a model with hallucinations can be costly, especially for uses in high-stakes domains such as healthcare. There are several complementary ways to attempt to address hallucinations. One way is to improve the calibration of the model (either via prompting or fine-tuning) so that it either abstains from answering when it is unlikely to be correct or provides some other indicator of confidence that can be used downstream. Another approach, that is suitable for mitigating open-domain hallucination, is to insert information that the model lacks into the prompt, for example by allowing the model to make calls to external sources of information, such as a search engine as in Section 5.1. For closed-domain hallucination the use of additional model computation through post-hoc checks is also promising, see Figure 1.8 for an example. Finally, building the user experience of an application with the possibility of hallucinations in mind can also be part of an effective mitigation strategy.
- **Long-term memory:** The model's context is very limited (currently 8000 tokens, but not scalable in terms of computation), it operates in a "stateless" fashion and there is no obvious way to teach the model new facts. In fact, it is not even clear whether the model is able to perform tasks which require an evolving memory and context, such as reading a book, with the task of following the plot and understanding references to prior chapters over the course of reading.
- **Continual learning:** The model lacks the ability to update itself or adapt to a changing environment. The model is fixed once it is trained, and there is no mechanism for incorporating new information or feedback from the user or the world. One can fine-tune the model on new data, but this can cause degradation of performance or overfitting. Given the potential lag between cycles of training, the system will often be out of date when it comes to events, information, and knowledge that came into being after the latest cycle of training.
- **Personalization:** Some of the applications require the model to be tailored to a specific organization or end user. The system may need to acquire knowledge about the workings of an organization or the preferences of an individual. And in many cases, the system would need to adapt in a personalized manner over periods of time with specific changes linked to the dynamics of people and organizations. For example, in an educational setting, there would be an expectation of the need for the system to understand particular learning styles as well as to adapt over time to a student's progress with comprehension and prowess. The model does not have any way to incorporate such personalized information into its responses, except by using meta-prompts, which are both limited and inefficient.
- **Planning and conceptual leaps:** As suggested by the examples in Section 8, the model exhibits difficulties in performing tasks that require planning ahead or that require a "Eureka idea" constituting a discontinuous conceptual leap in the progress towards completing a task. In other words, the model does not perform well on tasks that require the sort of conceptual leaps of the form that often typifies human genius.
- **Transparency, interpretability and consistency:** Not only does the model hallucinate, make up facts and produce inconsistent content, but it seems that the model has no way of verifying whether or not the content that it produces is consistent with the training data, or whether it's self-consistent. While the model is often able to provide high-quality post-hoc explanations for its decisions (as demonstrated in Section 6.2), using explanations to verify the process that led to a certain decision or conclusion only works when that process is accurately modeled and a sufficiently powerful explanation process is also accurately modeled (Section 6.2). Both of these conditions are hard to verify, and when they fail there

are inconsistencies between the model's decisions and its explanations. Since the model does not have a clear sense of its own limitations it makes it hard to establish trust or collaboration with the user without extensive experimentation in a narrow domain.

- **Cognitive fallacies and irrationality:** The model seems to exhibit some of some of the limitations of human knowledge and reasoning, such as cognitive biases and irrationality (such as biases of confirmation, anchoring, and base-rate neglect) and statistical fallacies. The model may inherit some of the biases, prejudices, or errors that are present in its training data, which may reflect the distribution of opinions or perspectives linked to subsets of the population or larger common views and assessments.
- **Challenges with sensitivity to inputs:** The model's responses can be very sensitive to details of the framing or wording of prompts and their sequencing in a session. Such non-robustness suggests that significant effort and experimentation is often required with engineering prompts and their sequencing and that uses in the absence of such investments of time and effort by people can lead to suboptimal and non-aligned inferences and results.

A limitation of our exploration is the absence of a clear distinction between drawbacks founded in the way that the reinforcement learning step (RLHF) was carried out, versus drawbacks which are fundamentally inherent in the larger architecture and methodology. For example, it is not clear to what extent the hallucination problem can be addressed via a refined reinforcement learning step or via a focused effort to introduce new forms of calibration about the likelihoods of the veracity of alternative inferences that the system can compute and consider in its generations (see also [Ope23] for more discussion on this). To draw an analogy to humans, cognitive biases and irrational thinking may be based in artifacts of our culture as well as to limitations in our cognitive capabilities. Pursuing better understandings of the sources and potential solutions to challenges of hallucination in GPT-4, will benefit from studies that compare several versions of the RL stage over the same architecture.

A broader question on the identified limitations is: which of the aforementioned drawbacks can be mitigated within the scope of next word prediction? Is it simply the case that a bigger model and more data will fix those issues, or does the architecture need to be modified, extended, or reformulated? Potential extensions to next word prediction include the following:

- External calls by the model to components and tools such as a calculator, a database search or code execution, as suggested in Section 5.1.
- A richer, more complex "slow-thinking" deeper mechanism that oversees the "fast-thinking" mechanism of next word prediction. Such an approach could allow the model to perform long-term planning, exploration, or verification, and to maintain a working memory or a plan of action. The slow-thinking mechanism would use the next word prediction model as a subroutine, but it would also have access to external sources of information or feedback, and it would be able to revise or correct the outputs of the fast-thinking mechanism.
- Integration of long-term memory as an inherent part of the architecture, perhaps in the sense that both the input and output of the model will include, in addition to the tokens representing the text, a vector which represents the context.
- Going beyond single-word prediction: Replacing the sequence of tokens by a hierarchical structure, where higher-level parts of the text such as sentences, paragraphs or ideas are represented in the embedding and where the content is generated in a top-down manner. It is unclear whether richer predictions about the sequencing and interdependency of such higher-level concepts might emerge from large-scale compute and data centered on a next-word–prediction paradigm.

10.3 What is actually happening?

Our study of GPT-4 is entirely phenomenological: We have focused on the surprising things that GPT-4 can do, but we do not address the fundamental questions of why and how it achieves such remarkable intelligence. How does it reason, plan, and create? Why does it exhibit such general and flexible intelligence when it is at its core merely the combination of simple algorithmic components—gradient descent and large-scale transformers with extremely large amounts of data? These questions are part of the mystery and fascination of LLMs, which challenge our understanding of learning and cognition, fuel our curiosity, and motivate deeper research. Key directions include ongoing research on the phenomenon of emergence in LLMs (see

[WTB+22] for a recent survey). Yet, despite intense interest in questions about the capabilities of LLMs, progress to date has been quite limited with only toy models where some phenomenon of emergence is proved [BEG+22, ABC+22, JSL22]. One general hypothesis [OCS+20] is that the large amount of data (especially the diversity of the content) forces neural networks to learn generic and useful "neural circuits", such as the ones discovered in [OEN+22, ZBB+22, LAG+22], while the large size of models provide enough redundancy and diversity for the neural circuits to specialize and fine-tune to specific tasks. Proving these hypotheses for large-scale models remains a challenge, and, moreover, it is all but certain that the conjecture is only part of the answer. On another direction of thinking, the huge size of the model could have several other benefits, such as making gradient descent more effective by connecting different minima [VBB19] or by simply enabling smooth fitting of high-dimensional data [ES16, BS21]. Overall, elucidating the nature and mechanisms of AI systems such as GPT-4 is a formidable challenge that has suddenly become important and urgent.

Acknowledgements. We thank OpenAI for creating such a marvelous tool and giving us early access to experience it. We also thank Miles Brundage at OpenAI, and the numerous people at Microsoft, who have provided thoughtful feedback on this work.