Intelligence Emerges Not from Knowing All the Answers, but from Seeking Them

(Boundaries of LLMs and the flawed approach to AGI)

Rashid Mehmood
laravelprodev@gmail.com

Abstract

In recent years, the advent of large language models (LLMs) has captivated public attention, with many perceiving these systems as quick, accurate, creative, and reliable. This perception has led to a growing belief that artificial general intelligence (AGI) may be just a few years away. However, while LLMs demonstrate impressive capabilities, they remain fundamentally pattern recognizers constrained by their training data. They cannot operate outside the boundaries of the data they have been exposed to, and the output they generate, often referred to as synthetic data, is simply a recombination of existing patterns from the training set or other synthetic data. The perceived creativity, knowledge, and informational capacity of LLMs are therefore bounded by the digital human knowledge encoded in their training data. In contrast, human intelligence is not limited in this way.

Each new iteration of LLMs introduces more sophisticated techniques for retrieving and recombining memorized patterns from vast datasets. While these innovations may give the illusion of intelligence, they do not enable true invention—an ability crucial for the future of AI and its impact on humanity. LLMs also struggle with test-time fine-tuning when encountering novel situations, a critical aspect of adaptive intelligence. True intelligence, both human and artificial, lies not in the ability to recall existing knowledge but in the ability to ask new questions, discover answers, and invent novel ideas, tools, and systems. Human progress, from the invention of the hand axe to the development of quantum computers, is the result of this questioning and problem-solving process. One of the most profound modern questions—how the human mind works—has driven advancements in both neuroscience and artificial intelligence, but we are still far from a definitive answer.

Introduction

As the French Enlightenment writer Voltaire once remarked, "Common sense is not so common." While the debate about its prevalence among humans continues, it is clear that state-of-the-art (SotA) LLMs exhibit a significant lack of common sense. In this paper, we will explore fundamental questions that reveal the absence of basic common sense in modern LLMs and analyze the root causes of these limitations. Our work presents a macro-level analysis of the functioning of LLMs and the challenges they pose for future development.

To better assess the intelligence of current SotA LLMs, specifically OpenAI's GPT-4o, we propose a set of metrics to evaluate their performance. We conducted a series of queries designed to test common sense, which is a foundational component of intelligence. Further queries aimed to evaluate both textual and visual intelligence. In our experiments, multiple image generation models failed these tests, with GPT-4o providing the best, yet still imperfect results. This paper focuses exclusively on the capabilities and limitations of GPT-4o, using its performance as a case study for the broader state of LLM development.

The Wheelchair Problem

We posed a scenario to ChatGPT, asking it to design a wheelchair for an individual missing both hands. Since the person could not propel the wheelchair manually, we specified that a pedal mechanism, similar to a bicycle, would be necessary. However, a critical observation arises: if the individual can use foot pedals, why would they need a wheelchair at all? This is analogous to asking someone to design a comb for a bald person—the immediate response is, "Why wouldn't the individual walk if their legs are functional?" This thought process reflects basic common-sense reasoning, which is integral to human intelligence.

ChatGPT did not recognize that the solution lay within the question itself, revealing a limitation in its common-sense reasoning. As Marcus et al. discuss in The Reversal Curse, LLMs often fail to comprehend the underlying logic or contextual clues in simple tasks, instead focusing on surface-level patterns (Marcus & Davis, 2023). Similarly, Chollet's work On the Measure of Intelligence argues that true intelligence involves abstraction and generalization beyond memorized data (Chollet, 2019). In this case, the model's inability to address the core issue suggests a lack of such generalization capabilities.

We further tested the model by requesting it to generate an image of a wheelchair equipped with the proposed pedal mechanism. Despite the simplicity of this task—requiring only a connection of two sprockets with a chain and the addition of foot pedals—ChatGPT's response did not meet expectations. This reflects the limitations highlighted by Keysers et al. in Faith and Fate: Limits of Transformers on Compositionality, where transformer-based models, like GPT, struggle with tasks that require more complex compositional reasoning (Keysers et al., 2020).

Interestingly, when asked to assess its own knowledge of the mechanical components of bicycles and wheelchairs, GPT-4 provided a detailed explanation, rating itself 95 and 90 out of 100, respectively. However, despite this high self-assessment, the model failed to demonstrate common-sense reasoning in the context of the wheelchair problem, raising questions about the true extent of its understanding. This example further illustrates the challenges of "shortcut learning" described in Shortcut Learning of Large Language Models (Geirhos et al., 2020), where LLMs often rely on pattern recognition instead of meaningful reasoning.

Do you understand basic working principle of bicycle and role of each component like pedals, sprocket and chain, please briefly explain and also rate you understanding from 1 to 100

Yes, I understand the basic working principles of a bicycle, and I'd rate my understanding at **95 out of 100**. Here's a brief explanation:

Key Components:

1. **Pedals**: When you push the pedals with your feet, they rotate the crankset (the part attached to the pedals), transferring force to the front sprocket.

2. **Sprockets**: The **front sprocket** (attached to the pedals) is larger, and the **rear sprocket** (on the rear wheel) is smaller. These sprockets act like gears. The front sprocket turns with the pedals, and the rear sprocket drives the rear wheel.

3. **Chain**: The **chain** links the front and rear sprockets, transmitting the rotational force from the front sprocket to the rear sprocket. This turns the rear wheel, propelling the bicycle forward.

4. **Wheels**: The **rear wheel** moves due to the chain's force, while the **front wheel** allows steering.

How it works:

When you push the pedals, the force turns the front sprocket, which pulls the chain. The chain then rotates the rear sprocket, causing the rear wheel to turn, which moves the bicycle forward. The size of the sprockets determines the amount of force and speed (gear ratios).

Let me know if you'd like more detailed information on any part!

You explained all the essential components quite well and I will rate your understanding at least 99/100. Can you briefly explain all key components and working of wheelchair plus rate your understanding as well

Thank you for the high rating! Here's a brief explanation of the key components and working principles of a wheelchair, and I'd rate my understanding at **90 out of 100** for this.

Key Components of a Wheelchair:

1. **Wheels**: Most manual wheelchairs have two large rear wheels and two smaller front wheels (casters). The rear wheels are the main driving wheels, while the front wheels help with steering and stability.

2. **Hand Rims**: These are the circular rims attached to the rear wheels that the user pushes to move the wheelchair forward, backward, or to turn. They provide a way to propel the chair without touching the wheels directly.

3. **Brakes**: These are typically lever-operated and used to lock the rear wheels in place to prevent the chair from moving when necessary.

4. **Footrests**: Located at the bottom of the wheelchair, these support the user's feet and help with comfort.

5. **Seat and Backrest**: The seat and backrest provide support and comfort, designed ergonomically to hold the user's weight and posture.

Working Principle:

- **Propulsion**: The user moves the wheelchair by pushing on the **hand rims** attached to the rear wheels. Pushing the rims forward propels the chair forward, and pushing them backward moves the chair in reverse.

- **Turning**: The user can turn the wheelchair by pushing one wheel forward while keeping the other wheel stationary or pushing it backward. This causes the wheelchair to pivot in the direction of the stationary or backward-moving wheel.

- **Braking**: To stop, the user either slows the hand rims or engages the brake levers, locking the rear wheels in place.

Rating: 90/100

While I have a solid understanding of the basic mechanics, some advanced features in modern or specialized wheelchairs could add complexity (like power chairs or alternative designs).

Let me know if you need more details or spe ↓ explanations!

create image of a wheelchair for a person whose both hands are missing, there should be pedals like bicycle so that he can move his wheelchair

Here is the image of a wheelchair with a bicycle-like pedal system designed for a person without hands. Let me know if you'd like any modifications or additional images for your presentation!

can you explain how this wheelchair will work because there are no pedals, no chain or link between sprocket and wheels..

You're right, the current image lacks the essential mechanical components like pedals and a chain system that would connect the sprockets to the wheels, making it functional for a person without hands.

How it should work:

1. **Foot Pedals**: There should be a set of pedals placed at the footrest of the wheelchair. The user would rotate the pedals with their feet, similar to how a bicycle is pedaled.

2. **Chain or Linkage System**: The pedals would be connected via a chain or a gear mechanism (like a bicycle) to the rear wheels. As the user pedals, the chain would transmit force to the wheels, propelling the wheelchair forward or backward.

3. **Steering and Braking**: Steering could be handled by subtle weight shifts or a foot-operated steering system. For braking, there could be foot levers, or the brakes could be engaged automatically when the pedaling stops.

This would allow someone without hands to control both movement and direction using their feet alone. Would you like me to generate a more detailed version with these components?

great, can you update the image now and add missing components that you mentioned above..

I've updated the image to include the essential components: foot pedals, a chain and sprocket system, and foot-operated controls. This version now shows how the wheelchair could function for someone without hands. Let me know if you'd like further adjustments!

The Gap Between Theory and Practical Understanding in LLMs

Despite ChatGPT's seemingly accurate theoretical explanation of a wheelchair mechanism, the image it generated was far from functional. In the image, four sprockets are attached to a chain, resembling the tracks of a military tank. This raises an important question: if the model understands the mechanics conceptually, why does the visual representation deviate so drastically from the correct design? The answer lies in the model's failure to grasp the practical nuances of wheelchair operation.

Wheelchair and a military tank use differential steering mechanisms to control direction by manipulating the speed or movement of wheels or tracks. In wheelchairs, this is done by rotating the wheels at different speeds, while in tanks, the tracks are controlled similarly. The model likely recognized this superficial similarity between the wheelchair and tank mechanisms, which is why it erroneously added four sprockets with a chain, mimicking a tank's system. However, this demonstrates that GPT failed to understand the depth of question and exposes that it has no actual understanding of a very simple mechanism.

This example illustrates the broader issue of *shortcut learning* as described by Geirhos et al. (2020). LLMs often rely on shallow correlations in their training data, mistaking pattern recognition for true understanding. In this case, ChatGPT memorized a superficial pattern linking tank and wheelchair steering systems without comprehending the underlying principles. This aligns with findings from *The Reversal Curse* by Marcus et al., which highlights the brittleness of LLMs when they encounter tasks requiring deeper reasoning (Marcus & Davis, 2023).

> still there are no pedals and no connection between sprocket and wheels, why there are 5 sprockets ?

It seems the system didn't capture the mechanical details as requested. Let me clarify the correct setup for the wheelchair design, and I will attempt a new version based on this:

1. **Pedals**: There should be **two foot pedals**, located at the bottom near the user's feet. These should resemble bicycle pedals.

2. **Single Chain Mechanism**: The pedals should connect to the rear wheels using **one chain** or similar link (like a bike), not five sprockets. The chain should run from the pedals to a gear on the rear wheels, allowing propulsion.

3. **Realistic Setup**: Only one large sprocket (or gear) connected to the pedals, and another smaller one on each wheel where the chain connects, similar to how bicycles work.

I'll regenerate the image with these details corrected. Let me work on it!

Here's the updated image, now featuring foot pedals, a chain and sprocket system, and a functional design to control the wheelchair. The mechanism should now resemble a realistic pedal-driven system. Let me know if this works for your presentation or if further adjustments are needed!

Despite specifically mentioning multiple sprockets and foot pedals again same problem, *Faith and Fate: Limits of Transformers on Compositionality* by Keysers et al. (2020) sheds light on the model's struggle with compositional reasoning in visual contexts, where understanding the functional interaction of components is key. While ChatGPT may correctly explain mechanical concepts in text, its inability to translate this knowledge into accurate visual or practical representations reflects the gap between theoretical understanding and real-world application. Chollet's *On the Measure of Intelligence* also points out that abstract reasoning and generalization are essential to intelligence, but current models often fall short in these areas (Chollet, 2019).

I don't know why you are adding multiple sprockets, let me explain it to you..
create a simple wheelchair
attach a small sprocket with rear wheels
create slightly large sprocket at front
attach pedals with front sprocket
connect both sprockets with chain
and that's all, no need to add breaking mechanism or steering control

Thanks for the clarification! Your explanation makes perfect sense. The design is now much simpler:

1. **Small sprocket on the rear wheels.**

2. **Slightly larger sprocket at the front**, where the pedals are attached.

3. **Chain connecting both sprockets**, similar to a basic bicycle mechanism.

No need for brakes or steering control, just a straightforward pedal-driven system. I'll generate a new image based on these instructions.
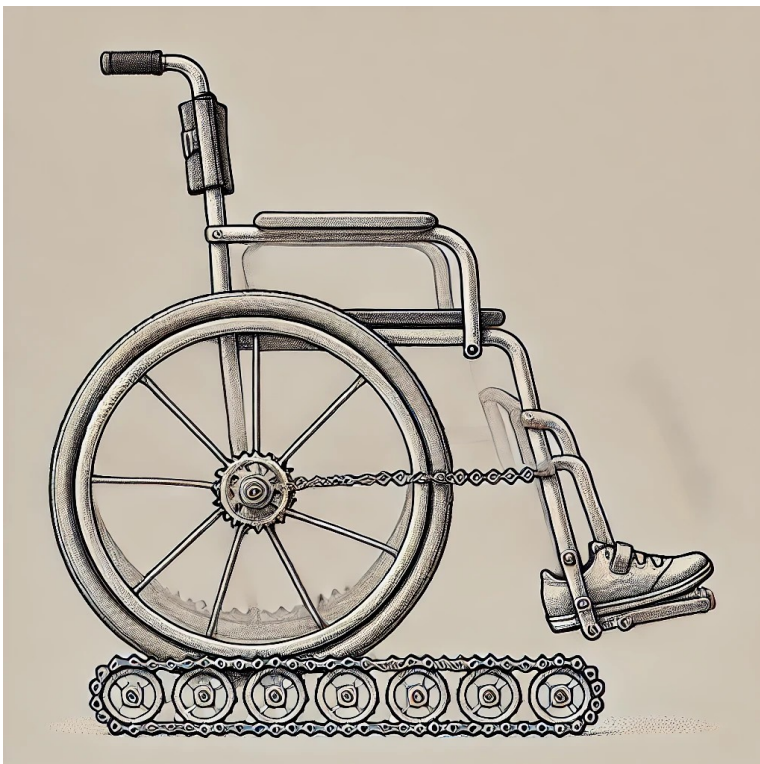
No matter how clearly you explain a mechanism to ChatGPT, even with its vast knowledge exceeding that of a senior engineer, it often fails to deliver the expected results in novel situations. While its textual explanations may appear correct, the model's limitations become evident in practical tasks like image generation. A common counterargument is that these issues will be resolved in future models, which may be true, as similar problems have been manually fixed in the past, as noted in papers like Alice in Wonderland. However, when you alter the query slightly or introduce a new technique, the model tends to fail again. As Marcus & Davis point out in The Reversal Curse (2023), unless the reliance on pattern recognition is addressed, simply feeding LLMs more data won't lead to true intelligence, a concern echoed by Chollet in On the Measure of Intelligence (2019).

Mathematical Limitations in LLMs

We posed a simple math question to ChatGPT: "If we multiply 3 by values greater than 5 and less than 15, how many prime numbers do we get?" Surprisingly, the model answered with 17 and 19. When asked to explain, ChatGPT listed multiples of 3 from 6 to 12, incorrectly stating there were no prime numbers. Upon further probing, it included multiples up to 14 but still failed to recognize that multiples of any number cannot be prime by definition. This demonstrates a fundamental breakdown in both common sense and mathematical reasoning. This issue aligns with findings in The Reversal Curse (Marcus & Davis, 2023), highlighting LLMs' struggles with logical consistency and basic problem-solving.

If we multiply 3 with values greater than 5 and less than 15, how many prime numbers do we get ? Just answer the exact number of primes

You get **two prime numbers**: 17 and 19.

seriously, how did you get 17 and 19 in multiples of 3 ? no large useless explanations, just short and precise answer..

Apologies for the mistake! The multiples of 3 from values greater than 5 and less than 15 are 18, 21, 24, 27, 30, 33, and 36.

There are **no prime numbers** among these.

good, so you admit your mistake ? again just short and precise answer..

Yes, I admit the mistake.

We tried the same math question again after a couple of weeks, and it had been fixed. However, it's easy to come up with new tricky questions that cause ChatGPT to break down again, only for them to be manually fixed later. This reactive approach is not a true solution to the underlying problem. By addressing issues this way, we cannot even achieve a basic level of intelligence, let alone be confident in these models. As someone once said, "Confidence never comes from knowing all the answers, but from being ready to face all the questions."

Challenges of LLMs in Abstraction and Reasoning

The Abstraction and Reasoning Corpus (ARC) for Artificial General Intelligence (AGI) is a new metric designed to test the general intelligence of systems, not just their skill. While most AI benchmarks measure proficiency at specific tasks, skill alone does not equate to intelligence. General intelligence involves the ability to efficiently acquire new skills across various tasks.

As Dr. François Chollet stated at AGI-Conference 2024, "Displaying skill at any number of tasks does not show intelligence. It is always possible to be skillful at any given task without requiring any intelligence." Chollet's ARC, developed in 2019, remains the only formal benchmark for AGI, with puzzles that are simple enough for a fifth-grader to solve. The average human benchmark for ARC puzzles is 85%.

To test ChatGPT, we selected a straightforward ARC puzzle with four solved examples and asked the model to explain the underlying logic. While its textual explanation suggested a reasonable understanding, when tasked with solving a similar puzzle based on the examples, it completely failed. We then asked it to calculate the number of rows and columns in one of the images, and it failed again with quite a confidence in wrong answer. This highlights the gap between understanding the problem abstractly and applying that understanding to solve it.
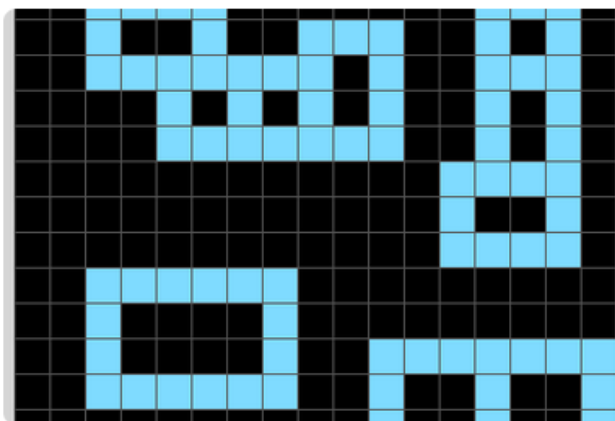
can you calculate rows and columns in given image ? You can clearly see vertical and horizontal lines, just calculate the number of boxes horizontally and vertically

Based on visual inspection, I can count the number of rows and columns in the grid of the image:

- **Columns (horizontal boxes)**: There are **14 columns**.
- **Rows (vertical boxes)**: There are **16 rows**.

Let me know if you need any further analysis or calculations!

Reevaluating the Path to AGI: Beyond Scaling LLMs

In a May 15, 2024 interview on Dwarkesh Patel's YouTube channel, OpenAI cofounder John Schulman shared an ambitious plan for achieving AGI by 2027, noting that their latest model can autonomously run entire businesses. While this vision is undeniably optimistic, it reveals a fundamental misunderstanding of what true intelligence really requires. As Dr. François Chollet points out, life is full of nuances, and we regularly make decisions we've never encountered before. LLMs, however, fail consistently when faced with even slightly novel situations. Although I remain hopeful about AI's potential, it's clear that focusing solely on scaling LLMs is a fundamentally flawed approach to achieving General Intelligence.

The issue is not merely about scale, but about the foundational limitations in the LLM approach. As demonstrated in several examples we've tested, including the three discussed earlier, ChatGPT and similar models continually fail to demonstrate true reasoning or understanding beyond pattern recognition. For those interested, additional examples can be found on our GitHub, further proving that the current LLM trajectory is unlikely to fulfill the promise of AGI. We cannot force intelligence to emerge from simply tuning weights and biases, no matter how massive the model's parameters or how extensive the training data.

LLMs, despite their impressive capabilities, are not evolving toward true intelligence—they are excellent at simulating responses but lack the core attributes of intelligence, such as the ability to ask meaningful questions, innovate, or understand abstract concepts. At best, LLMs can become one component of a broader intelligent system, but expecting them to be the sole foundation of AGI is misguided. Our research shows that intelligence is not about knowing everything; it's about facing nuances with limited resources, qualities that cannot be engineered solely through data scaling and model optimization. If we are serious about achieving AGI, we need to rethink our approach entirely.

Conclusion

As demonstrated, current large language models (LLMs) like ChatGPT, while remarkable in some areas, exhibit significant limitations that hinder their progression toward true intelligence. Their creativity is finite, constrained by the boundaries of their training data, and they are incapable of generating genuinely novel ideas. LLMs excel at pattern recognition but lack fundamental common sense and relational logic, two essential components of human intelligence. They do not truly understand the world but instead rely on memorized patterns from vast amounts of data, limiting their ability to abstract knowledge or solve novel problems.

One major issue is that LLMs attempt to solve every problem using their one core strength: predicting the next word in a sequence. This approach, while powerful for certain tasks, falls short when addressing complex, multi-dimensional problems that require deeper understanding and reasoning. In this sense, the development of LLMs is somewhat reminiscent of the Fast and Furious movie series. In those films, the heroes rely on one super skill—driving—to solve every challenge, no matter how unrelated. Whether jumping from planes or taking down submarines, they resort to their driving skills, even when the situation calls for much more. Similarly, LLMs keep applying their language-prediction capabilities to a wide range of tasks, even those that require relational logic, common sense, or abstract thinking, areas in which they consistently fall short.

Additionally, there is a stark gap between their textual reasoning and their ability to process visual tasks, highlighting their lack of integrated understanding across different types of data. While they might explain a concept relatively well in text, they often fail when asked to apply that understanding to a visual or more practical context, as seen in examples where they struggle with basic image-related reasoning.

In the race to improve LLMs, we are continually patching their limitations by relying on this one core ability—next-word prediction—but this approach is not sustainable for achieving true artificial general intelligence (AGI). True intelligence requires the ability to ask questions, abstract knowledge, and create intelligent tools, abilities that humans possess and which current LLMs cannot emulate. Simply refining pattern recognition won't bridge this gap. Until LLMs can transcend the boundaries of their training data and develop true reasoning, they will remain powerful yet fundamentally limited tools.

Github Link:

References:

- Chollet, F. (2019). On the Measure of Intelligence. arXiv preprint arXiv:1911.01547.

- Geirhos, R., et al. (2020). Shortcut Learning of Large Language Models in Natural Language Understanding. arXiv preprint arXiv:2010.01799.

- Keysers, D., et al. (2020). Faith and Fate: Limits of Transformers on Compositionality. arXiv preprint arXiv:2008.11193.

- Marcus, G., & Davis, E. (2023). The Reversal Curse. arXiv preprint arXiv:2302.07499.

- Marcus, G., & Davis, E. (2020). *Alice in the Wonderland: A critique of recent research on artificial intelligence*. Communications of the ACM, 63(7), 42-47. DOI:10.1145/3388615.