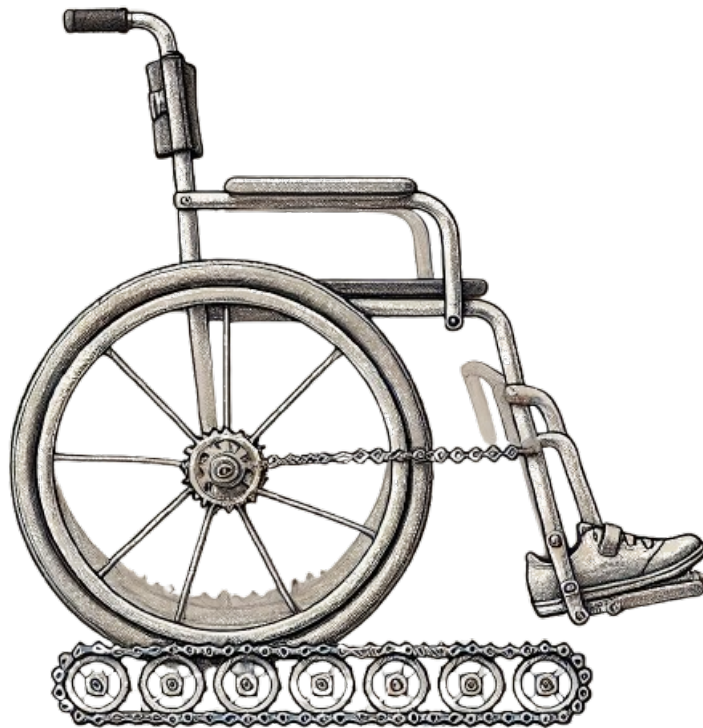# Can Intelligence Emerge Through Merely Knowing All the Answers?

Rethinking Assumptions in the Pursuit of AGI

Rashid Mehmood

{laravelprodev@gmail.com}

March 1, 2025

Abstract

We propose a novel, three-step criterion to assess whether an artificial intelligence model exhibits general intelligence. Our framework—termed AGI (Analyze/Audit Itself, Generate Solutions, Implement, and Repeat)—is predicated on the observation that human intelligence thrives despite inherent limitations in time and resources. In our approach, an AI must autonomously perform self-assessment, generate multiple remedial or enhancing solutions, and implement the optimal choice—all while operating within a fixed envelope of memory and processing power. This resource-constant constraint not only mirrors the efficient resource management observed in human cognition but also provides a rigorous, equitable benchmark for comparing AI performance to human-level intelligence.

As the field of artificial intelligence rapidly evolves, the definition of true intelligence has become a focal point of academic debate. Recent advances in large language models (LLMs) have captivated both public and scholarly attention due to their ability to emulate several cognitive traits traditionally associated with human intelligence, such as creativity and problem-solving. Nonetheless, despite their remarkable capabilities, LLMs remain fundamentally constrained by the scope of their training data and model parameters. Their outputs, often described as synthetic, largely represent recombinations of previously encountered patterns rather than evidence of genuine autonomous innovation. In stark contrast, human intelligence is characterized by its dynamic adaptability and the capacity to innovate despite stringent resource constraints.

Our research delineates this essential difference and outlines a structured pathway toward the development of artificial systems that can genuinely approach human-level intelligence. By mandating that AI systems maintain constant resources during self-improvement cycles, our criterion challenges them to optimize performance through efficient, autonomous resource management—an indispensable quality for any system aspiring to true Artificial General Intelligence.

— Key Insights —

- **True intelligence, whether human or artificial, is not defined by the ability to recall existing knowledge but by the capacity to ask new questions, discover answers, invent novel ideas, tools, and systems.**

- **LLMs are often mistaken for intelligent systems, but their capabilities are fundamentally limited to memorized situations.**

- **Human intelligence is adaptive, inventive, and capable of generating new knowledge, while LLMs lack these crucial capabilities.**

- **State of the Art LLMs, despite their advancements, lack common sense, a critical component of human-like intelligence.**

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 evident that state-of-the-art (SotA) large language models (LLMs) exhibit a considerable lack of common sense. In this paper, we pose very basic questions related to common sense that an average human mind with minimal knowledge could easily answer. We explore fundamental questions that expose the absence of basic common sense in modern LLMs and analyze the root causes of these deficiencies. Our work offers a macro-level analysis of the functioning of LLMs and the challenges they present for future advancements. Despite the remarkable progress in large language models in recent years, the ability to truly understand natural language remains an elusive goal. While LLMs have demonstrated impressive performance on various language tasks, several critical limitations have been identified that prevent them from achieving genuine natural language understanding (Ling et al., 2024)[6] (Singhal et al., 2022)[7].

Our research has highlighted the limitations of LLMs, particularly in their ability to engage in effective reasoning and problem-solving. Reasoning, which involves inferring insights, deducing solutions, and synthesizing new ideas, is a hallmark of human intelligence. Each new iteration of LLMs introduces more advanced techniques for retrieving and recombining memorized patterns from vast datasets. While these innovations may create the illusion of intelligence, they do not facilitate novel invention or they lack true Natural Language Understanding (NLU). LLMs also struggle with test-time fine-tuning when encountering novel situations, a crucial element of adaptive intelligence. Humans, by contrast, heavily rely on this ability, which is essential for the survival of any intelligent species. True intelligence, whether human or artificial, is not defined by the ability to recall existing knowledge but by the capacity to ask new questions, discover answers, invent novel ideas, tools, and systems. Human progress, from the creation of the hand axe to the development of quantum computers, is the result of encountering novel challenges, adapting, questioning, and problem-solving.

As shown in Figure 1, ChatGPT achieved an A grade in nearly every AI benchmark related to common sense and basic science reasoning. Despite these high scores in established intelligence benchmarks, LLMs consistently fail when confronted with practical, novel scenarios. This discrepancy raises two key questions:
1) If LLMs are not genuinely intelligent, how do they achieve such high scores in these benchmarks?
2) Can we accurately measure the core intelligence of a system, given that current AI benchmarks seem inadequate for this purpose?
In this paper, we address the first question and leave the second question for future exploration. To more accurately assess the intelligence of current SotA LLMs, particularly OpenAI's GPT and DALL·E, both of which are LLM-based transformer architectures, we conducted a series of queries designed to test common sense, which is a fundamental aspect of intelligence. All three questions referenced in this paper were intended to evaluate common sense and basic reasoning. In our experiments we exclusively worked on

ChatGPT, all the material presented in this paper was tested using GPT-4o, which is currently the most advanced LLM. This paper centers on the capabilities and limitations of GPT-4o, using its performance as a case study to reflect on the broader state of LLM development.

Model	ARC-AGI	GLUE	MMLU	HellaSwag	WinoGrande	BIG-bench	CQA
ChatGPT	80%	90%	78%	82%	80%	75%	79%
Grok	78%	88%	76%	80%	78%	73%	77%
Gemini	85%	91%	80%	87%	83%	76%	81%
DeepSeek	83%	89%	77%	85%	79%	74%	78%

Table 1: Scores for various AI models across different benchmarks.

The Wheelchair Problem

We posed a scenario to Chat-GPT, asking it to design a wheelchair for an individual missing both hands. Since the person could not propel the wheelchair manually with hand rims, 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 should be, "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.

Although ChatGPT initially accepting the fact that the person in question wouldn't need a wheelchair, it 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[1], LLMs often fail to comprehend the underlying logic or contextual clues in simple tasks, instead focusing on surface-level patterns (Evans & Berglund, 2023)[1]. Similarly, Chollet's work On the Measure of Intelligence argues that true intelligence involves abstraction and generalization beyond memorized data (Chollet, 2019)[3]. 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 Dziri 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 (Dziri et al., 2023)[4].

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 (Tao et al., 2024)[5], where LLMs often rely on pattern recognition instead of meaningful reasoning.

> If a person has both arms missing or disabled but has functional legs, does he need a wheelchair?

If a person has both arms missing or disabled but has **functional legs**, they typically would **not need a wheelchair** for everyday mobility, as they can walk independently. A wheelchair is primarily used when a person's legs are impaired, preventing them from walking or moving effectively.

However, a wheelchair might still be needed in certain specific situations:

1. **Fatigue or long distances**: If the person becomes fatigued after walking long distances, a wheelchair might be useful for covering extended distances more comfortably.

2. **Safety concerns**: If balance or coordination is affected due to the absence of arms (e.g., difficulty catching themselves if they fall), a wheelchair could be used for safety in some environments.

3. **Temporary conditions**: In some cases, if they are recovering from a surgery or injury to the legs, a wheelchair could be required temporarily.

In general, if their **legs are fully functional**, they would likely not require a wheelchair for daily activities.

> 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!

🅜 Memory updated

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!

Thus far, GPT has demonstrated a reasonable level of common sense and a working knowledge of the operation of both wheelchairs and bicycles. Ideally, it should raise a counter-question if prompted to generate an image of a wheelchair for a person with fully functional legs. Before advancing to our primary question, regardless of whether ChatGPT passes or fails this common sense test, we will further evaluate its theoretical and visual comprehension of the fundamental operating principles behind these two basic mechanisms.

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!

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!

8

Despite specifically mentioning multiple sprockets and foot pedals again same problem, Faith and Fate: Limits of Transformers on Compositionality by Dziri et al. (2023)[4] 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)[3].

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.

It seems like I can't generate any more images right now. Please try again later. If you'd like, I can still guide you through creating the image manually or help with other parts of your presentation! Let me know what you'd prefer.

generate image according to last message
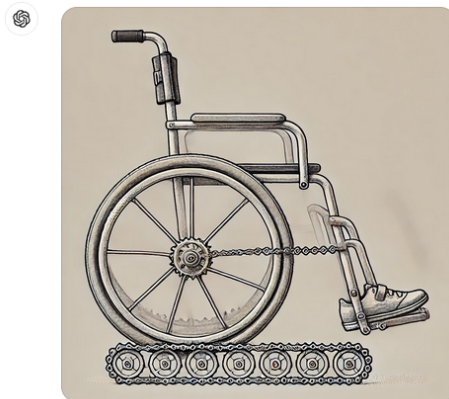
Here is the image of the simple wheelchair with the pedal-driven system, as you described. It features the front and rear sprockets connected by a chain, with pedals attached to the front sprocket. Let me know if this matches your vision or if any further adjustments are needed!

Disparities in Textual and Visual Interpretation within Large Language Models

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 and bicycle operation. Wheelchair and a military tank use **differential steering mechanism** 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 Tao et al. (2024)[5]. 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 Evans et al., which highlights the brittleness of LLMs when they encounter tasks requiring deeper reasoning (Evans et al., 2024)[1].

No matter how clearly you explain a mechanism to ChatGPT, even with its extensive mechanical knowledge surpassing that of a senior engineer, it often fails to deliver the expected results in novel situations. While its textual explanations may appear accurate, the model's limitations become evident in practical tasks such as image generation. A common counterargument is that identical issues mentioned in past research papers have been resolved, which might be true because similar problems have been manually addressed and rectified in the past, as noted in papers like Alice in Wonderland[2] and The Reversal Curse[1]. However, when the query is slightly modified or a new technique is introduced that exploits a known loophole, these models tend to fail once again, as these issues have persisted since the inception of LLMs. Some of the previously recognized issues in past research, which seem to be resolved, are actually just obscured, much like giving painkillers to a patient with a severe disease. Until the underlying condition is treated, the pain is likely to resurface. To handle with such loopholes, one major technique we have noticed implemented in ChatGPT is its reluctance to take a definitive stance in difficult situations. The response is often slightly ambiguous, supporting both sides of an argument so that it cannot be proven entirely wrong. This also allows LLM the flexibility to mold its stance in the future. This tactic resembles the techniques employed by politicians and religious leaders—whether the model learned this unintentionally or it was intentionally programmed by the development team remains unclear. Unless the core issues are addressed, merely feeding LLMs more skill programes and memorized situations will not result in true intelligence, a concern echoed by François Chollet in On the Measure of Intelligence (2019)[3].

Mathematical Common Sense

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, correctly stating that 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. When asked the same question with different numbers, it consistently calculates all the multiples first and then concludes that there are no prime numbers among these multiples. This demonstrates a fundamental breakdown in both common sense and mathematical reasoning. This issue aligns with the findings in The Reversal Curse (Evans el al., 2024)[1], which highlight 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.

> if we multiply 3 with values > 5 and <15 how many prime numbers do we get ?

The prime numbers between 5 and 15 are 7, 11, and 13.

Multiplying 3 with these values gives:

- $3 \times 7 = 21$
- $3 \times 11 = 33$
- $3 \times 13 = 39$

None of these results are prime numbers. Therefore, multiplying 3 by values between 5 and 15 gives **0 prime numbers**.

When we asked the same question with exactly same values again, ChatGPT not only provided an incorrect answer confidently but also employed a fundamentally flawed approach. This highlights a clear gap in basic common sense reasoning, making it easy

to craft new, challenging questions that cause ChatGPT to falter once again. These errors are often fixed manually later, but this reactive approach does not offer a true solution to the underlying problem. Addressing issues in this way hampers our ability to achieve even a foundational level of intelligence, let alone inspire confidence in these models. As the saying goes, **"Intelligence is not about knowing all the answers, but about being prepared to confront all the questions."**

Challenges of LLMs in Abstraction and Reasoning

The Abstraction and Reasoning Corpus (ARC) for Artificial General Intelligence (AGI) is a novel metric designed to evaluate the general intelligence of systems, rather than merely their skill. While most AI benchmarks assess proficiency in specific tasks, skill alone does not constitute intelligence. General intelligence entails the ability to efficiently acquire new skills across a diverse range of tasks.

As Dr. François Chollet remarked at the AGI Conference 2024[9], **"Displaying skill in any number of tasks does not demonstrate intelligence. It is always possible to be skillful in a given task without requiring any intelligence."** Chollet's ARC, developed in 2019, remains the only formal benchmark for AGI, consisting of puzzles that are simple enough for a fifth-grader to solve, yet complex enough to challenge state-of-the-art AI systems. The average human benchmark for ARC puzzles is 85

To evaluate 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 once again failed—this time with misplaced confidence in its incorrect answer. This underscores the gap between abstractly understanding a problem and effectively applying that understanding to solve it.
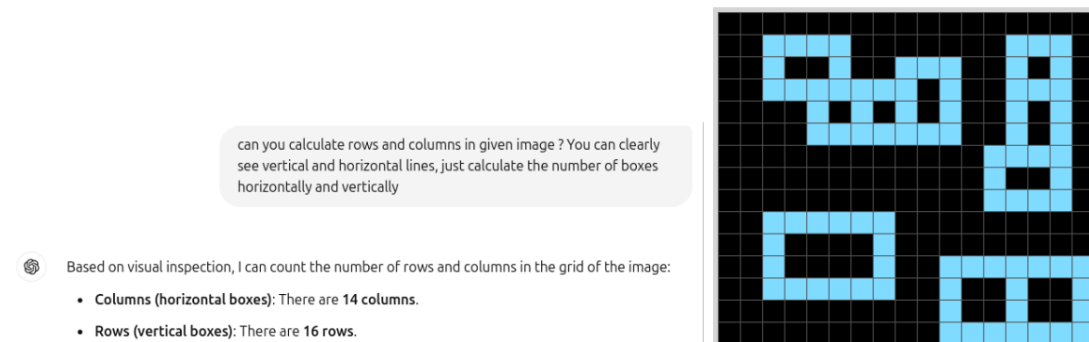

Gemini Flash API We have tested first 50 puzzles of public evaluation set of ARC using raw json input with models of gemini 1.5-flash and 2.0-flash-experimental. Starting

from oirinal examples given with ARC puzzles we gave additional examples like vertical, diagonal and horizontal flip, colors shifting etc. We increased number of additional examples with each batch from 0 to 9, which means each original example of puzzle will have 9 additional training examples and if 5 examples are originally given with puzzle then total examples will become 50. Yellow line representing the threshold is basically the similarity between test input and true output of puzzle. e.g if gemini output of a given puzzle is 91 similar with true output and the threshold for that puzzle is 90 then we can say that gemini only solved 1 of the puzzle because 90 of it was already similar or solved. Threshold 0 means test input and true output matrix size does not match. As you can see in figure there is no significant improvement with additional examples. Gemini could not even cross the threshold in more than half of the puzzles.

Batch	Temp	Additional Examples	Total Attempted	Above Threshold	Solved 100%
batch-6	1.65	0	49	23	2 (4.08%)
batch-7	1.65	2	50	23	2 (4.00%)
batch-8	1.65	4	48	19	2 (4.17%)
batch-9	1.65	9	42	21	2 (4.76%)

Table 2: gemini-1.5-flash.png results
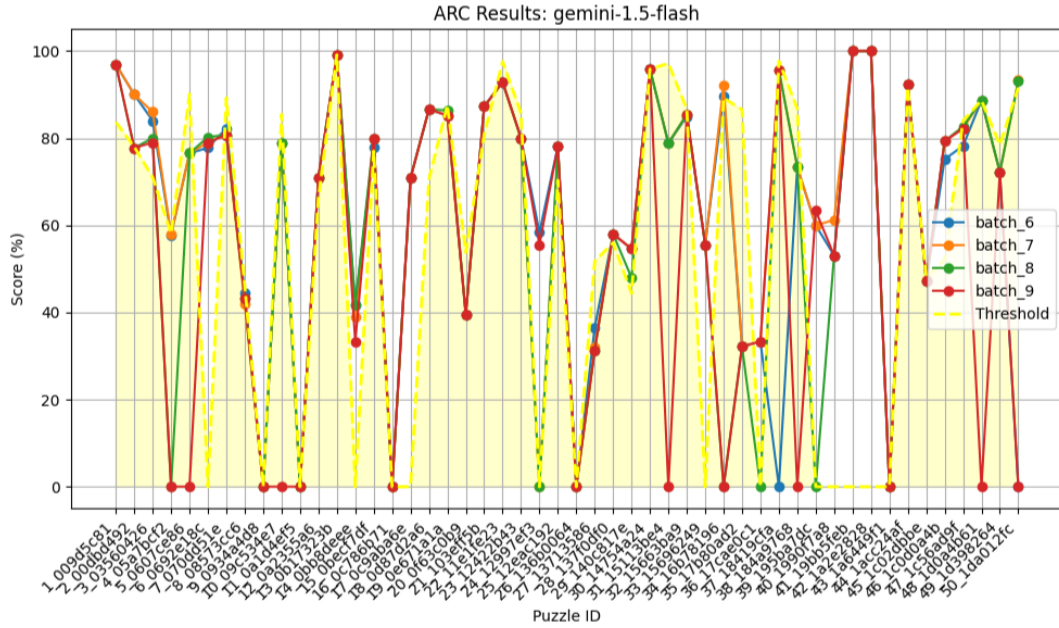
Batch	Temp	Additional Examples	Total Attempted	Above Threshold	Solved 100%
batch-0	1.65	0	47	21	0 (0.00%)
batch-1	1.65	0+data	50	21	1 (2.00%)
batch-2	1.25	2+data	48	20	1 (2.08%)
batch-3	1.35	4+data	45	21	0 (0.00%)

Table 3: gemini-2.0-flash-thinking-exp-01-21 results
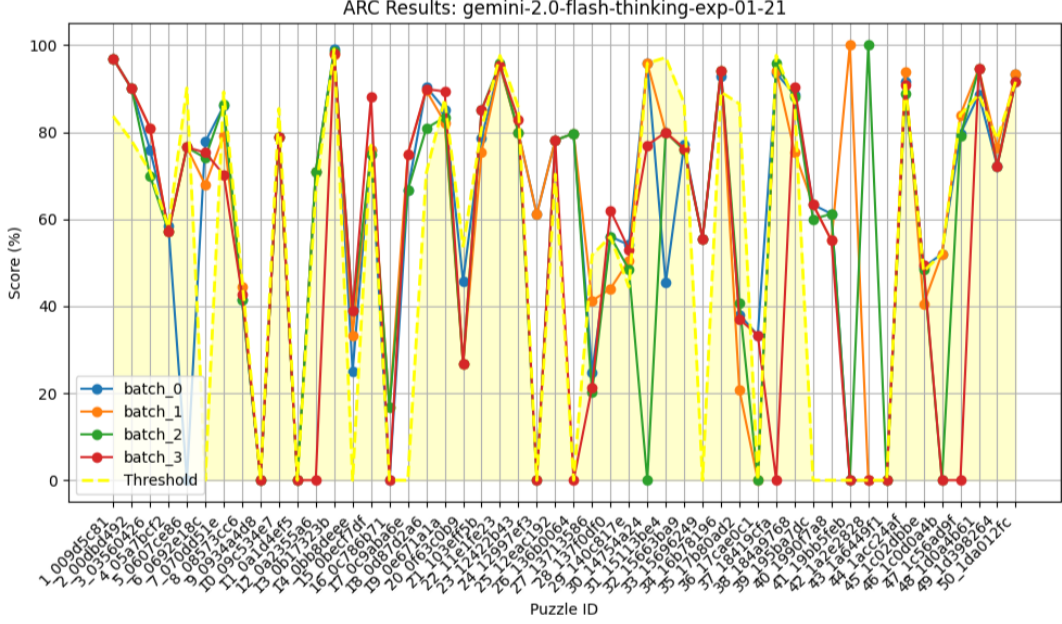
Reevaluating the Path to AGI: Beyond Scaling LLMs

In a May 15, 2024 interview on Dwarkesh Patel's YouTube channel[8], OpenAI cofounder John Schulman discussed cutting-edge developments in AI and the goal of achieving AGI by 2027, stating that their latest model can autonomously run entire businesses. While this vision is undeniably optimistic, it is important to recognize that running even a small-scale business requires significant general intelligence and the ability to handle novel situations multiple times a day. This assertion reveals a fundamental misunderstanding of what true intelligence necessitates. As Dr. François Chollet points out, life is full of nuances, and we regularly make decisions we've never encountered before. LLMs, however, consistently fail when faced with even slightly novel circumstances. Although I remain hopeful about AI's potential, it is clear that **focusing exclusively on scaling LLMs is a fundamentally flawed approach to achieving**

General Intelligence.

The issue lies not merely in the scale of LLMs, but in the foundational limitations of the LLM approach. As demonstrated in several examples we have tested, including the three discussed earlier, ChatGPT and similar models continually fail to exhibit true reasoning or understanding beyond pattern recognition. For those interested, additional examples are available on our GitHub link, further proving that the current trajectory of LLM development is unlikely to fulfill the promise of AGI. Intelligence cannot be forced to emerge simply by tuning weights and biases, no matter how large the model's parameters or how extensive the training data.

LLMs, despite their impressive capabilities, are not progressing toward true intelligence—they excel at simulating responses but lack core attributes of intelligence, such as the ability to ask meaningful questions, innovate, or comprehend abstract concepts. **At best, LLMs can serve as one component within a broader intelligent system, but expecting them to form the sole foundation of AGI is misguided**. Intelligence is not about knowing everything; it is about confronting nuances with limited resources—qualities that cannot be engineered solely through data scaling and model optimization. To build a truly intelligent system, we need to fundamentally rethink our approach.

Conclusion
As demonstrated, current large language models (LLMs), 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 skill programes and situations 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 LLMs textual reasoning and their ability to process visual tasks, highlighting the 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**.

References

[1] Owain Evans, Lukas Berglund, Meg Tong, Max Kaufmann, et al. (2024). THE REVERSAL CURSE : LLMs TRAINED ON "A IS B" FAIL TO LEARN "B IS A". arXiv:2309.12288v4.

[2] Marianna Nezhurina, Lucia Cipolina-Kun, Mehdi Cherti, Jenia Jitsev, et al. (2024). Alice in Wonderland: Simple Tasks Showing Complete Reasoning Breakdown in State-Of-the-Art Large Language Models. arXiv:2406.02061v4.

[3] François Chollet. (2019). On the Measure of Intelligence. arXiv:1911.01547v2.

[4] Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, et al. (2023). Faith and Fate: Limits of Transformers on Compositionality. arXiv:2305.18654v3.

[5] Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. (2024). Shortcut Learning of Large Language Models in Natural Language Understanding. DOI:10.1145/3596490.

[6] Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, et al. (2024). Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey. arXiv:2305.18703v7.

[7] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, et al. (2022). Large Language Models Encode Clinical Knowledge. arXiv:2212.13138v1.

[8] John Schulman (OpenAI Cofounder) Interview at Dwarkesh Patel Youtube channel (2024). `https://www.youtube.com/watch?v=Wo95ob_s_NI&t=2041s`

[9] François Chollet talk at AGI Conference, ARC Prize Youtube channel (2024). `https://www.youtube.com/watch?v=nL9jEy99Nh0&t=1450s`

Github link: `https://github.com/ainumbat/ChatGPT4o_issues.git`