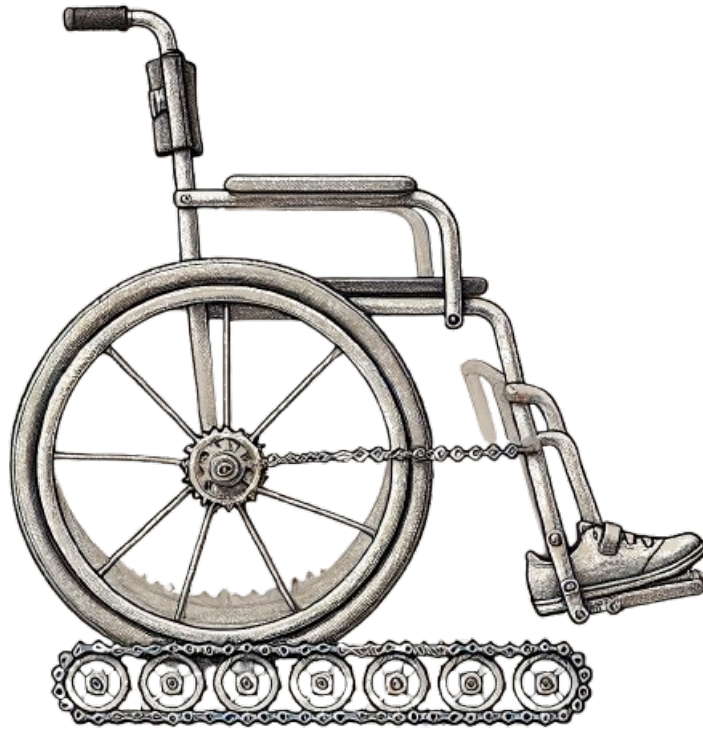# The Illusion of Intelligence
Evaluating LLMs Against Grounded AGI Criteria

Rashid Mehmood, Eid Rehman

{laravelprodev@gmail.com, eid.rehman@umw.edu.pk}

April 18, 2025

Abstract

As artificial intelligence advances, the definition of true intelligence has become a central topic of debate. Large language models (LLMs) impress with their ability to mimic traits like creativity and problem-solving, placing them at the forefront of many applications. However, their effectiveness in tasks requiring genuine reasoning and adaptability remains limited. LLMs operate by recombining patterns from training data, not by generating truly original insights. Unlike humans, they lack context-awareness, self-reflection, and the ability to make decisions under constraints. These deficiencies highlight the gap between current LLMs and the dynamic, adaptive nature of human intelligence.

Our research delineates this fundamental gap and emphasizes the need for AI systems to not only produce outputs that appear intelligent but also possess the internal processes necessary for being intelligent. Specifically, the ability to self-assess, generate alternatives, and adapt strategies within fixed computational bounds is crucial for developing systems that genuinely approximate human-level cognition. To address these challenges, we propose a novel three-step criterion—AGI (Analyze/Audit Itself, Generate Solutions, Implement and Repeat)—as a benchmark to assess the general intelligence of AI systems. Under this framework, an AI must autonomously Perform self-assessment, Generate multiple remedial or enhancing strategies and Implement the most optimal solution all while operating under a constant envelope of memory and processing power. This resource-constant constraint mirrors efficient human cognition and provides a rigorous, equitable benchmark for evaluating progress toward Artificial General Intelligence.

— Key Insights —

- **AI and LLMs have not independently invented any digital tool or function in the way that humans have created entirely novel and foundational innovations.**

- **LLMs cannot even recognize their own core limitations—such as lack of true understanding, common sense, transparency, contextual awareness, efficiency and hallucinations. All of these are identified externally by human researchers.**

- **Intelligence is not merely the ability to recall or pattern-match existing knowledge, but is characterized by the capacity to pose new questions, discover answers, and create novel ideas, tools, and systems.**

- **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.**

- **Human intelligence is inherently adaptive, inventive, and capable of generating original knowledge—abilities that LLMs critically lack.**

Introduction

Artificial Intelligence (AI) has rapidly evolved, becoming a foundational technology across industries such as education, research, business, and entertainment. Among the most transformative developments are Large Language Models (LLMs), which have demonstrated capabilities in language understanding, content generation, translation, summarization, and conversational agents. These models mimic cognitive traits traditionally associated with human intelligence, such as creativity and problem-solving, earning them widespread attention in both public and academic circles. Their broad applicability and growing impact have elevated LLMs to a central role in discussions about the future of intelligent systems.

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 even state-of-the-art LLMs often lack this fundamental aspect of intelligence (Xiang et al., 2022)[10].

Despite their success, LLMs face several well-documented challenges. They often lack true common sense, struggle with contextual awareness, and fail in situations that require adaptive reasoning or self-reflection. Their reasoning tends to rely on statistical pattern-matching rather than actual understanding, limiting their ability to deal with novel or ambiguous situations. Furthermore, their performance is bounded by training data and model parameters, making them prone to hallucinations, inconsistencies, and superficial problem-solving. These limitations suggest that current benchmarks may not fully capture the depth of intelligence required for general-purpose reasoning.

This paper focuses specifically on the limitations of LLMs in commonsense reasoning and basic problem-solving—critical indicators of natural intelligence. We analyze several state-of-the-art LLMs (ChatGPT, Gemini, Grok, and DeepSeek) through experiments designed to test their commonsense capabilities using straightforward visual and textual prompts. Although these models score highly on standard benchmarks (see Table 1), our evaluations reveal a consistent inability to handle novel, practical questions that an average human could easily solve. Our methodology emphasizes macro-level analysis through minimal, targeted queries, shedding light on the models' failure to transcend memorized patterns. Ultimately, this work argues for more grounded and reliable measures of intelligence in AI systems, challenging the validity of current benchmarks as accurate proxies for true intelligence.

Note: *SotA LLMs refers to State-of-the-Art Large Language Models such as ChatGPT, Gemini, DeepSeek, and Grok. These models represent the most advanced implementations of transformer-based neural architectures currently available.*

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

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

The Wheelchair Problem: A Test of Emergent Reasoning in LLMs

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 with hand rims, we specified that a pedal mechanism—similar to a bicycle—would be necessary. However, a critical observation emerges: if the individual can use foot pedals, why would they need a wheelchair at all? This is analogous to designing a comb for a bald person—the immediate common-sense response should be, "If the legs are functional, wouldn't the individual simply walk?"

This scenario underscores a fundamental gap in common-sense reasoning. Although ChatGPT eventually acknowledged that the person might not need a wheelchair, it failed to immediately recognize the contradiction inherent in the prompt. The fact that it required multiple prompts and context clues to arrive at a partially correct answer highlights a key limitation in its reasoning capabilities. As Marcus et al. note in The Reversal Curse [1], large language models often miss the underlying logic of a task, defaulting instead to pattern-matching based on their training data. Similarly, Chollet's On the Measure of Intelligence [3] emphasizes that true intelligence involves the ability to generalize and abstract—not merely regurgitate memorized data.

This problem becomes more interesting in light of recent discussions around the emergence of sophisticated behaviors in large models, such as few-shot learning and chain-of-thought (CoT) reasoning. Studies like Wei et al. (2022) [12] show that, with sufficient scale, LLMs begin to exhibit CoT abilities that were not explicitly programmed. These emergent behaviors give the impression of reasoning, yet as the wheelchair problem illustrates, such reasoning is still brittle and lacks genuine comprehension. While chain-of-thought prompting can sometimes guide models toward better performance on logical tasks, it does not guarantee consistent application of common-sense principles.

We further tested the model by asking it to generate an image of a wheelchair equipped with a pedal mechanism. Despite the simplicity of the task—merely requiring the connection of sprockets, a chain, and foot pedals—the model's response fell short. This aligns with the findings of Dziri et al. in Faith and Fate: Limits of Transformers on Compositionality [4], which argue that transformer-based models struggle with tasks that require deeper compositional understanding, even if they can solve surface-level problems.

Interestingly, when prompted to self-assess its knowledge of the mechanical components of bicycles and wheelchairs, GPT-4 rated itself 95 and 90 out of 100, respectively. This overconfidence, despite its failure on a basic reasoning task, illustrates the phenomenon of "shortcut learning," wherein models rely heavily on superficial cues rather than engaging in true inferential reasoning. As noted by Tao et al. in Shortcut Learning of Large Language Models [5], LLMs tend to exploit statistical shortcuts rather than developing a genuine understanding of the problem space.

Methodology

To investigate the disparity between textual reasoning and visual generation in large language models (LLMs), we designed a diagnostic prompt scenario that tests both domains simultaneously. The scenario involved a mechanical design task requiring basic common-sense reasoning and knowledge of physical systems. Specifically, we asked ChatGPT to design a wheelchair for an individual without hands. We specified that the person would use a pedal mechanism—akin to a bicycle—for propulsion. This intentionally paradoxical scenario, termed the **Wheelchair Problem**, serves as a probe into the model's conceptual consistency and understanding of practical mechanics.

Our methodology consisted of the following steps:

1. **Textual Reasoning Evaluation:** ChatGPT was prompted with a detailed description of the problem and asked to explain how it would design a wheelchair incorporating a bicycle-style pedal mechanism. The generated explanation was evaluated for mechanical plausibility, coherence, and common-sense reasoning. This step tests the model's ability to handle seemingly straightforward tasks requiring contextual awareness.

2. **Visual Generation Task:** Using the same scenario, the model was instructed to generate an image of the proposed wheelchair using a text-to-image module. The output image was analyzed for mechanical correctness and alignment with the explanation provided in the textual reasoning phase.

3. **Consistency Analysis:** We compared the reasoning in the textual response with the content of the generated image to assess conceptual alignment. This step tests for internal consistency—a hallmark of intelligent reasoning. We also examined if the design adhered to physical principles, such as the *differential steering mechanism* commonly found in wheelchairs and tracked vehicles.

4. **Literature Comparison and Emergent Behavior Evaluation:** We contextualized our findings against emergent capabilities such as chain-of-thought reasoning [12] and few-shot learning [11], to evaluate if such behaviors manifest reliably in multi-modal tasks. Our results are interpreted in light of the limitations outlined in

prior work on shortcut learning [5], compositional reasoning [4], and brittle pattern matching failures [1, 2].

This methodology exposes significant weaknesses in current LLMs, such as their tendency to mimic superficial patterns rather than reason about the implications of a scenario. Notably, the model erroneously introduced four sprockets and a chain resembling a tank track system—likely due to a learned association between tracked motion and wheelchairs—demonstrating the phenomenon of *shortcut learning* and lack of abstraction [5, 3].

This experimental setup, while simple, effectively reveals how large models can fail even when they possess factual mechanical knowledge. Such diagnostic tasks are critical for assessing the real-world readiness of LLMs in applied reasoning scenarios.

> 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!

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

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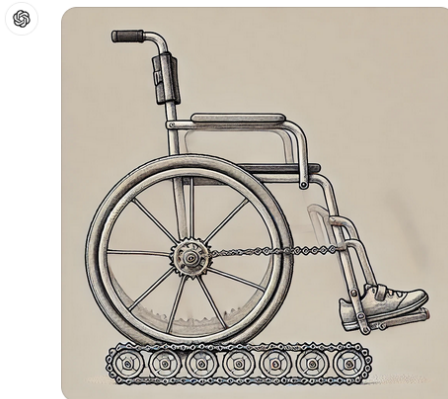
Here is the image of the simple wheelchair with the pedal-driven system, as you described. It features the front and rear sprockets connected by a chain, with pedals attached to the front sprocket. Let me know if this matches your vision or if any further adjustments are needed!

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 slightly 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 all SotA LLMs is their 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

Task Category	Number of Tasks	Difficulty	ChatGPT	Gemini
Public Training Tasks	400	Easy	92%	90%
Public Evaluation Tasks	400	Hard	85%	88%
Semi-private Evaluation Tasks	100	Hard	80%	84%
Private Evaluation Tasks	100	Hard	78%	82%

Table 2: ARC-AGI-1 dataset composition: 1,000 tasks split into four subsets.

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 score for ARC-AGI Benchmark 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.
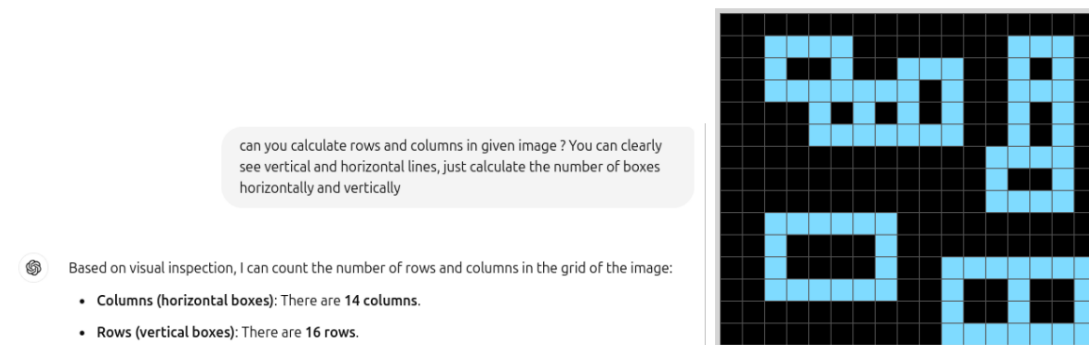
As shown in Table 2, latest models of top LLMs have now "cracked" ARC-AGI, with arcprize.org confirming that ChatGPT o3 High (Tuned) achieved an 88% score on the Semi-Private Evaluation set. These high scores result from benchmarks specifically targeted in every new LLM version, on which the models are highly trained. **Our research demonstrates that a model's performance cannot exceed its predetermined threshold at inference time, regardless of how simplified the benchmarks or hints are. In other words, you cannot improve a trained model simply by tweaking its inputs; instead, you must address its key issues and train a new model.**
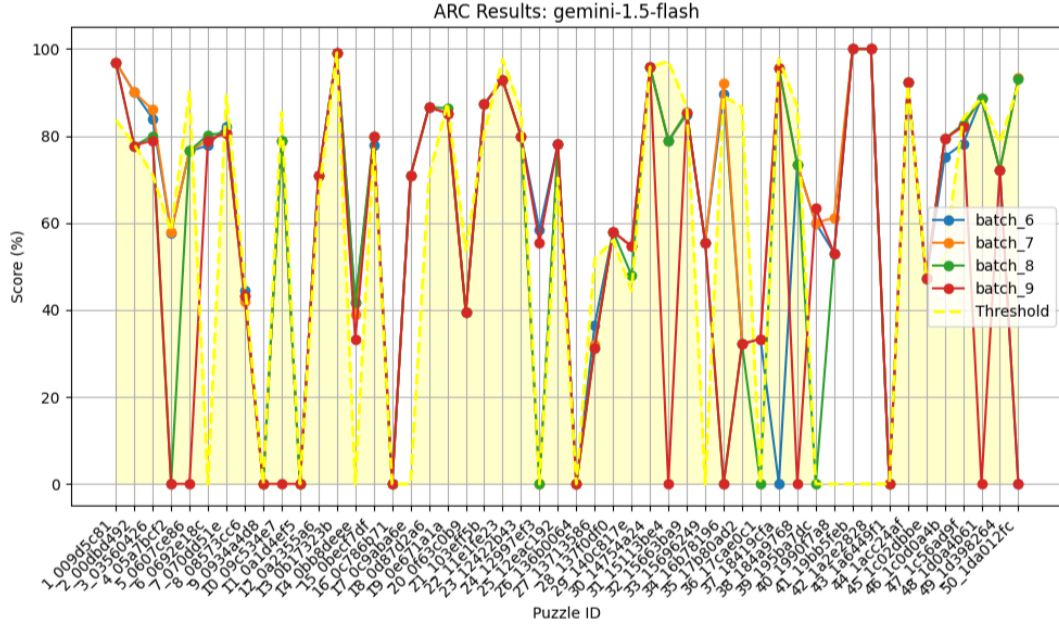
Gemini Flash Results

Our results indicate no significant improvement when additional examples were provided—Gemini failed to exceed the threshold in more than half of the puzzles. Essentially, if Gemini 1.5 only achieves a 5% score on the AGI benchmark, no level of simplification appears capable of enhancing its performance. We also leveraged Gemini's long-context capabilities to perform multiple attempts per puzzle using data from previous attempts; however, this strategy did not yield any significant improvement, and the results remained highly unstable. These findings confirm that a model's abilities are inherently limited by the training data it receives, and a trained model cannot be improved at test time.

Interesting Fact: We incorporated the correct output for each puzzle, labeled as "true output," to assess the model's common sense abilities. Despite multiple attempts and clear hints provided to the model, the success rate remained below 40%.

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 3: gemini-1.5-flash.png results

ARC Results: gemini-1.5-flash

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 4: gemini-2.0-flash-thinking-exp-01-21 results

ARC Results: gemini-2.0-flash-thinking-exp-01-21
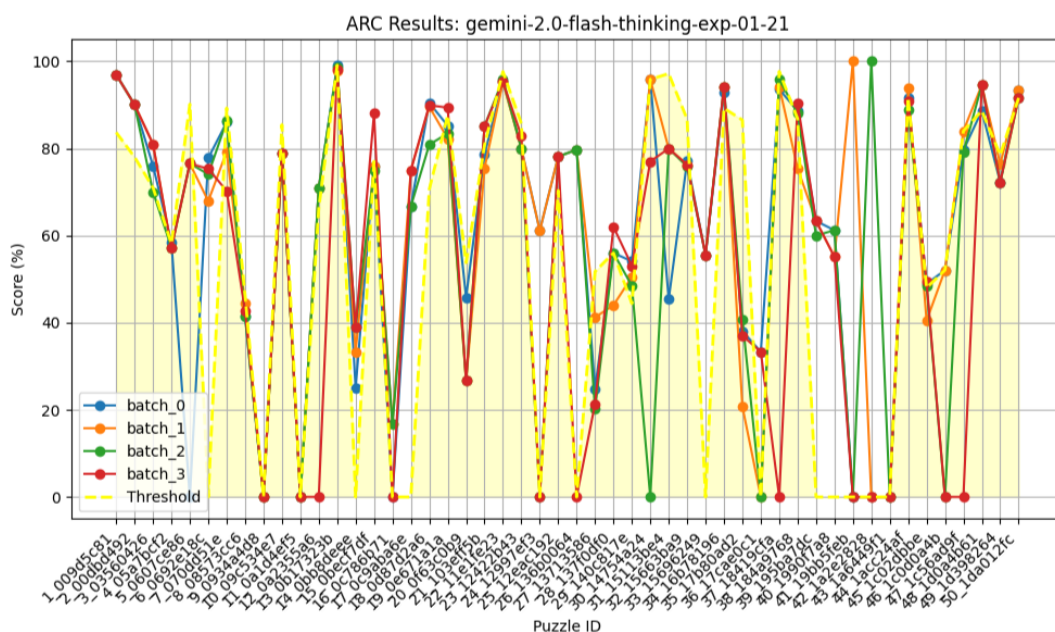
Measuring True Intelligence: Challenges, Limitations, and a Proposal for AGI Criteria

However, many challenges—such as achieving true reasoning, common sense, and ethical alignment—require breakthroughs that extend beyond merely scaling existing architectures. Until these advances are made, AI remains a powerful yet flawed tool that performs best under human oversight. Below is a list of key issues observed in current state-of-the-art LLMs:

- Lack of true understanding/comprehension

- Lack of common sense

- Context limitations or shallow reasoning

- Resource intensity

- Lack of transparency (black box behavior)

- Vulnerability to adversarial attacks

- Hallucinations

Interestingly, none of these issues are self-identified by AI/LLMs; they are all diagnosed by humans—whether researchers, users, or auditors—through testing, analysis, or observation. Current AI/LLMs lack the self-awareness and introspection [13] needed to

16

autonomously recognize their own limitations. Although these systems can describe their flaws if prompted [14], this is based on training data (human-authored critiques) or web searches rather than genuine self-diagnosis.

As demonstrated in Table 1, top AI models have achieved remarkable scores on benchmarks related to common sense and basic science reasoning. Yet, despite these high scores, LLMs consistently falter when confronted with practical, novel scenarios. This discrepancy raises an important question: Can we accurately measure the core intelligence of a system when current AI benchmarks seem inadequate for this purpose?

At present, AI and LLMs have not independently invented any digital tool or function in the way that humans have created entirely novel and foundational innovations. However, human researchers have developed AI-generated enhancements and optimizations that significantly improve performance. In light of these observations, we propose a simple yet rigorous AGI criterion—one that tests an AI/LLM model's general intelligence and cannot be easily circumvented unless true AGI is achieved.

AGI Criteria: Beyond Scaling LLMs

We have proposed a very simple 3 steps AGI criteria to check whether the current LLMs and future intelligent models hold the test of human level intelligence. An AI system can be considered Artificial General Intelligence (AGI) if it independently pass AGI Criteria without human intervention:

Analyze/Audit Itself – The AI must recognize issues or limitations in its own body of code, reasoning, context and abstraction.
Generate Solutions – For a specific problem chosen by AI itself based on its priority It should be able to generate multiple possible ways to fix or improve itself.
Implement and Repeat – The AI must choose the best solution and implement it and repeat the process.

Conditions AI Resources will remain constant during that process, initially limited amount of extra memory and resources should be provided to the AI system then it should remain constant then it should be upto the AI system to manage those resources like Humans do

Design and Operational Implications

Prioritization of Tasks:
The AGI would need to learn which tasks are most critical and allocate resources accordingly. It might, for instance, reserve more memory for tasks that require deep introspection or learning, and less for routine operations.
Trade-Offs and Decision-Making:
Limited resources mean that the AI must sometimes make trade-offs. For example, a more resource-intensive self-analysis might be deferred in favor of immediate, less de-

manding tasks. This trade-off is similar to how humans decide between deep reflection and rapid decision-making based on available mental energy and time.

Algorithmic Innovation:

With a cap on resources, the AGI is pushed toward developing more efficient algorithms. This might lead to breakthroughs in how to compress information, optimize code, or structure reasoning in a way that minimizes overhead.

Safety and Stability:

Limiting resources can also serve as a safety mechanism. It prevents an AGI from over-committing or making uncontrolled changes that might require external resources to manage—much like how our biological systems maintain homeostasis.

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 [8], 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 impressive in many respects, exhibit significant limitations that prevent them from attaining true intelligence. Their creativity is constrained by the scope of their training data, and they are fundamentally incapable of producing genuinely novel ideas. While they excel at pattern recognition, they lack essential qualities such as common sense, relational logic, and true understanding—core attributes of human intelligence. Rather than reasoning, they operate through memorized skill programs and statistical associations, which limits their

ability to generalize or solve unfamiliar problems.

One of the most critical limitations is that LLMs approach all problems through their singular strength: **predicting the next word in a sequence**. Though this mechanism proves powerful for various tasks, it fails when confronted with multi-modal or abstract reasoning challenges. This is akin to deploying a highly specialized industrial robot designed to sort objects by color for a task requiring spatial reasoning or ethical judgment. The robot may perform flawlessly within its narrow scope but fails entirely when asked to adapt to contexts outside of its designed capabilities. Similarly, LLMs apply their language modeling paradigm indiscriminately—even to problems that require conceptual understanding or reasoning—exposing their lack of flexibility.

Moreover, a clear disparity exists between LLMs' textual reasoning and their performance in visual or practical tasks, underscoring the absence of **integrated, cross-modal understanding**. While they may offer coherent textual descriptions, their inability to transfer this understanding to visual tasks—such as mechanical design—demonstrates a lack of grounded comprehension.

In summary, the current trajectory of LLM development relies heavily on augmenting their core strength rather than addressing foundational weaknesses. This patchwork approach may yield incremental improvements, but it is unlikely to result in true artificial general intelligence (AGI). **True intelligence demands the ability to question, abstract, and invent—capabilities that remain unique to human cognition**. Until LLMs transcend pattern recognition and demonstrate genuine reasoning, they will remain powerful yet fundamentally limited tools.

Key Takeaways:

- LLMs are limited by their reliance on next-word prediction and lack true understanding or abstraction capabilities.

- They exhibit poor common-sense reasoning and fail at tasks requiring relational logic.

- Their textual and visual capabilities remain disconnected, revealing gaps in cross-modal reasoning.

- Pattern recognition alone is insufficient to achieve AGI.

- Human-like intelligence requires curiosity, creativity, and the ability to ask new questions—traits absent in LLMs.

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] Extra examples, Github Link: `https://github.com/ainumbat/ChatGPT4o_issues.git`

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

[10] Xiang Lorraine Li, Adhiguna Kuncoro, Jordan Hoffmann, et al. (2022). A Systematic Investigation of Commonsense Knowledge in Large Language Models. arXiv:2111.00607v3.

[11] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, et al. (2022). Emergent Abilities of Large Language Models. arXiv:2206.07682v2

[12] Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903v6.

[13] Zhangyue Yin, Qiushi Sun, Qipeng Guo, et al. (2023). Do Large Language Models Know What They Don't Know? arXiv:2305.18153v2.

[14] Miles Turpin, Julian Michael, Ethan Perez, Samuel R. Bowman1, et al. (2023). Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought. arXiv:2305.04388v2.

[15] Georg Wenzel and Adam Jatowt. (2023). An Overview Of Temporal Commonsense Reasoning and Acquisition. arXiv:2308.00002v3.

[16] François Chollet, Mike Knoop, Gregory Kamradt and Bryan Landers (2023). ARC Prize 2024: Technical Report. arXiv:2412.04604v2.

[17] Jason Wei, Brian Ichter, Xuezhi Wang, Fei Xia, et al. (2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903v6.

[18] Jun Zhao, Jingqi Tong, Yurong Mou, et al. (2024). Exploring the Compositional Deficiency of Large Language Models in Mathematical Reasoning Through Trap Problems. arXiv:2405.06680v4.

[19] Michael Timothy Bennett. (2024). Is Complexity an Illusion? arXiv:2404.07227v4.

[20] Tyler A. Chang and Benjamin K. Bergen. (2023). Language Model Behavior: A Comprehensive Survey

[21] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, et al. (2020). Language Models are Few-Shot Learners. arXiv:2005.14165v4

[22] Sourav Banerjee, Ayushi Agarwal and Saloni Singla. (2024). LLMs Will Always Hallucinate, and We Need to Live With This. arXiv:2409.05746v1.

[23] Moritz Herrmann, Julian D. Lange, Katharina Eggensperger, et al. (2024). Position: Why We Must Rethink Empirical Research in Machine Learning. arXiv:2405.02200v2.

[24] Zhaofeng Wuã, Linlu Qiuã, Alexis Ross, et al. (2024). Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks. arXiv:2307.02477v3.

[25] Ekin Akyürek, Mehul Damani, Linlu Qiu, et al. (2024). The Surprising Effectiveness of Test-Time Training for Abstract Reasoning. arXiv:2411.07279v1.