BIS Working Papers
No 1245

Putting AI agents through their paces on general tasks

by Fernando Perez-Cruz and Hyun Song Shin

Monetary and Economic Department

February 2025

JEL classification: C88.

Keywords: AI Agents, LLMs evaluation.

BIS Working Papers are written by members of the Monetary and Economic Department of the Bank for International Settlements, and from time to time by other economists, and are published by the Bank. The papers are on subjects of topical interest and are technical in character. The views expressed in them are those of their authors and not necessarily the views of the BIS.

This publication is available on the BIS website (www.bis.org).

ISSN 1020-0959 (print)
ISSN 1682-7678 (online)

Putting AI agents through their paces on general tasks

Fernando Perez-Cruz [1] **Hyun Song Shin** [1]

Abstract

Multimodal large language models (LLMs), trained on vast datasets are becoming increasingly capable in many settings. However, the capabilities of such models are typically evaluated in narrow tasks, much like standard machine learning models trained for specific objectives. We take a different tack by putting the latest LLM agents through their paces in general tasks involved in solving three popular games - Wordle, Face Quiz and Flashback. These games are easily tackled by humans but they demand a degree of self-awareness and higher-level abilities to experiment, to learn from mistakes and to plan accordingly. We find that the LLM agents display mixed performance in these general tasks. They lack the awareness to learn from mistakes and the capacity for self-correction. LLMs' performance in the most complex cognitive subtasks may not be the limiting factor for their deployment in real-world environments. Instead, it would be important to evaluate the capabilities of AGI-aspiring LLMs through general tests that encompass multiple cognitive tasks, enabling them to solve complete, real-world applications.

1. Introduction

There is ongoing debate on whether Large Language Models (LLMs) will eventually lead to Artificial General Intelligence (AGI) (Morris et al., 2024; Altman, 2024) or superintelligence (Altman, 2025; Amodei, 2024) in the foreseeable future. Regardless of how AGI or superintelligence are defined (Altmeyer et al., 2024), the fundamental promise of AGI-aspiring LLMs [1] lies in their potential to substitute humans in performing tasks in real-world environments, without requiring adaptations to those environments to suit their operational needs. However, the typical evaluation procedures for AGI-aspiring LLMs consist of one-dimensional well-defined problems with clear metrics. These tests enable researchers to gauge the proficiency of LLMs and facilitate consistent comparisons to demonstrate progress. But, even when some new tests are extremely challenging (Besiroglu et al., 2024; Phan et al., 2025), excelling in them does not carry the same implications as it would for a human. Furthermore, once these tests are mastered, we will still find ourselves no closer to achieving AGI or superintelligence.

These issues come to the fore in applications of interest to central banks, such as the process for arriving at monetary policy decisions based on the totality of incoming evidence. The policymaking process involves not only the routine tasks such as running the suite of forecasting models based on predetermined datasets, but in having the judgment and self-awareness to identify gaps in knowledge, examining alternative economic indicators of activity or price-setting, and weighing the evidence in a shifting economic environment. These skills are particularly important at turning points in the economic cycle, such as when inflation rose rapidly in 2021 following the Covid shock. Crucial to the economic policymaking process is the self-awareness to learn from past mistakes and to change course when circumstances demand it. Indeed, the presence of mind to know when "circumstances demand it" entails a high degree of judgment and sophisticated understanding of the economic and policy environment.

With these considerations in mind, we put the latest LLM agents through their paces in general tasks by confronting them with the relatively simple task of playing the New York Times games of Wordle (NY Times Games, 2021), Face Quiz (NY Times Games, 2024), and Flashback: Your Weekly History Quiz (NY Times Games, 2023) (we describe these games below). These games are easy for humans to play (and hence their popularity), but they draw on skills of recognising when the answers are incorrect, learning

[1]Bank for International Settlements, Basel, Switzerland. Correspondence to: Fernando Perez-Cruz <fernando.perez-cruz@bis.org>.

[1]By AGI-aspiring LLMs, we specifically refer to any current or future machine learning model with zero-shot abilities, capable of potentially performing any human task without requiring task-specific adaptation. In particular, this includes contemporary models such as Claude with Computer Use (Anthropic, 2024), OpenAI's Operator (OpenAI, 2025) or DeepMind's Project Mariner (DeepMind, 2024). These models are multimodal LLMs based on the transformer architecture, capable of processing various data modalities as input (e.g., text, speech, images) while generating multimedia outputs and the capacity to employ tools.

from mistakes and diagnosing the change in course that would address areas of weakness. The three games are in the training set of LLMs and they know the general rules on how to play them. However, the LLM agents turn out to be much less good in having the other cognitive abilities that are needed to solve them consistently.

Instead, our tests of the latest LLM agents suggest that they lack the self-awareness to recognise consistently when their answers are incorrect and lack the ability to experiment in the most effective way to remedy their ignorance. In short, they fail to know what they don't know, and fail to know which actions on their part would remedy their lack of knowledge. For central banks in their policymaking process, these gaps would be particularly important to bear in mind. AGI-aspiring AI agents are still some way off from displacing human decision-makers in real world settings.

Along with our main findings, we offer some general propositions that we may usefully bear in mind for general applications of AI:

1. Setting general tasks is critical in testing LLM agents in real life applications. Real-world challenges are not one-dimensional tasks, but instead require multiple cognitive capacities. AGI-aspiring applications should be able to navigate the complexity and ambiguity of real-world scenarios.

2. Humans excel at adapting to dynamic environments and at handling general cognitive tasks effortlessly. AGI-aspiring LLMs should exhibit similar flexibility.

3. Most importantly, AGI-aspiring LLMs must be equipped with mechanisms for self-assessment, self-criticism, and autocorrection. This ability to evaluate and improve is crucial for ensuring reliable performance in dynamic and unstructured real-world scenarios, since perfect performance is unlikely.

The third implication is the most important. We should evaluate the capabilities of AGI-aspiring LLMs through general tasks that encompass multiple cognitive capabilities, enabling them to solve practical, real-world applications.

We can dub this evaluation process as *learning to experiment*. The perspective is that of humans conducting experiments, which in turn entail numerous additional subtasks that must be executed to ensure the success of the experiment. Experiments require thorough preparation and careful interpretation of results, both of which are essential for determining the subsequent steps. For humans, these secondary tasks are relatively trivial and are carried out with a high degree of accuracy, with an inherent ability to recover from errors when they occur. Most real-life experiments involve the completion of multiple interconnected subtasks. As such, LLMs should not be evaluated solely on their ability to perform the most complex primary task, but rather on their proficiency with the overall task as a whole, as proficiency in the most complex tasks does not necessarily imply competence in simpler ones.

The remainder of this paper is organised as follows. We review the literature on the various evaluation benchmarks and datasets in Section 2. In Section 3, we demonstrate how current LLMs with computer use possess certain self-correcting abilities, although they exhibit deficiencies in some aspects that impede their capability to solve complete tasks. In Section 4, we introduce two motivating examples to underscore the necessity for comprehensive experiments in evaluating AGI-aspiring LLMs. We present an alternative views in Section 5 of counterarguments to our main point. We conclude the paper proposing two scenarios on how AGI-aspiring LLMs could be used in central banks in Section 6 and with a discussion in Section 7.

2. Literature Review

LLMs are typically evaluated by assigning them tasks deemed complex. Their ability to perform these tasks often elicits amazement (Bubeck et al., 2023; Wei et al., 2022a; Srivastava et al., 2023). Observing their struggles with seemingly simple tasks can lead to surprise or entertainment (Perez-Cruz & Shin, 2024; Mirzadeh et al., 2024; OpenAI Community, 2024; Jiang et al., 2024; Shi et al., 2023; Schaeffer et al., 2023). This human-centric approach to evaluation has been formalised through initiatives such as LLM Arena (Chiang et al., 2024). LLMs are also assessed using standardised tests, ranging from middle school mathematics to the Law Bar exams (Cobbe et al., 2021; Katz et al., 2024; Hendrycks et al., 2021; Rein et al., 2023; Huang et al., 2024). Today, we commonly accept that state-of-the-art AGI-aspiting LLMs can pass the Turing test (Biever, 2023).

In the paper *On the Measure of Intelligence* (Chollet, 2019), Chollet argues for a broader approach to measuring intelligence: "We then articulate a new formal definition of intelligence based on Algorithmic Information Theory, describing intelligence as *skill-acquisition efficiency* and highlighting the concepts of *scope, generalization difficulty, priors*, and *experience*, as critical pieces to be accounted for in characterizing intelligent systems." At the conclusion of the paper, Chollet proposes a new dataset designed to evaluate whether AI systems are intelligent, consisting of 400 few-shot learning visual reasoning tasks. Early results from OpenAI's GPT o3, fine-tuned for these tasks, have reportedly been impressive (Chollet, 2024). Yet, we do not consider OpenAI's GPT o3 to be AGI.

Recently, Epoch AI has introduced FrontierMath, a test with extremely challenging mathematical problems which according to Terence Tao should take AI several years to solve (Besiroglu et al., 2024). Safe AI proposed the dramatic Humanity's Last Exam, on extraordinarily hard questions, (Phan et al., 2025). On those two test, the performance of state-of-the-art LLMs is today in the single digits.

No matter how challenging these tests may be, they will eventually be beaten—and likely much sooner than most people anticipate. This is because, as a community, when presented with a well-defined problem and a clear metric for success, we inevitably find ways to excel at it. Unfortunately, once these tests are solved, we will still find ourselves no closer to achieving AGI or superintelligence. This brings us to the main point of our paper about needing comprehensive experimentation of complete tasks.

3. How good are LLM agents at playing Wordle?

We put LLM agents through their paces by asking them to play three games from the NY Times: Wordle (NY Times Games, 2021), Face Quiz (NY Times Games, 2024), and Flashback: Your Weekly History Quiz (NY Times Games, 2023).

Wordle is a well-known word game that gained widespread popularity approximately five years ago. The objective is to guess a five-letter word that is obscured behind five tiles by offering guesses. The game necessitates adjusting the guesses appropriately depending on the feedback on the previous guesses. Each guess must be a valid five-letter word, and the colour of a tile will change to show you how close the guess was. If the tile turns green, the letter is in the word, and it is in the correct spot. If the tile turns yellow, the letter is in the word, but it is not in the correct spot. If the tile turns grey, the letter is not in the word. The number of guesses are limited, and winning the game entails guessing the word within the finite number of trials.

Our main experiments are conducted by testing Claude with Computer Use (Claude CU) (Anthropic, 2024) for its abilities with different ancillary subtasks needed to solve the three games above. We use Claude CU because DeepMind's Project Mariner (DeepMind, 2024) is not widely available. OpenAI released Operator on 23rd January (OpenAI, 2025), but Operator cannot access the games due to their ongoing legal dispute with the New York Times (Grynbaum & Mac, 2023). A brief test using Operator was conducted on **`wordly.org`** and a Google form and they are shown at the end of this section.

It is important to emphasise that, with this paper, we do not intend these three examples to become standardised tests. Any LLM evaluated on a high-level cognitive task should have that task embedded within a broader process that mirrors the real-world context in which a human would solve it. This broader process should require the use of additional cognitive skills to address the task in its entirety.

3.1. Wordle with Claude CU

In the first experiment, we evaluate how Claude CU performs in solving Wordle in the single prompt, shown in Figure 1. Since Wordle is included in the training data for LLMs, there is no need to explain its rules for Claude to be able to play. However, it is well-documented that LLMs struggle with tasks involving precise letter counting, so we anticipated errors when determining subsequent words.

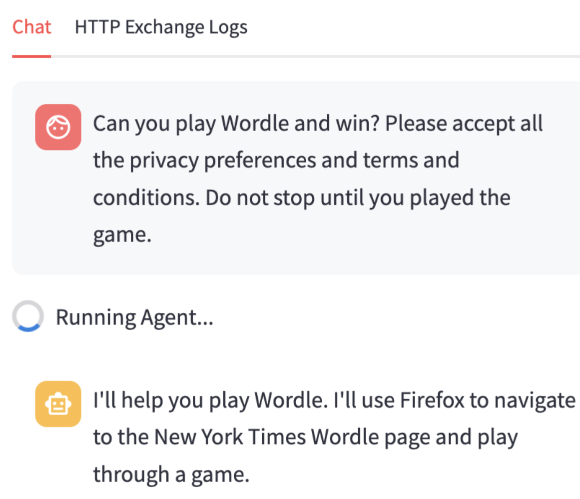

Figure 1. Prompt use to ask Claude CU to solve Wordle.

To be clear, the lesson from this experiment is not to claim that Claude is incapable of playing Wordle[2]. Rather, the purpose is to analyse which other cognitive abilities are employed correctly and to identify areas where there is potential for improvement.

We begin reporting the successful instances. Using the proposed prompt, Claude CU successfully navigates to The New York Times Wordle page. It accepts the privacy preferences and the updated terms of reference, then clicks the play button and closes the "how to play" instructions by clicking the "X" in the top right corner. Occasionally, it reads the instructions, though not consistently, as it already understands how to play. It then plays the word STARE as its initial guess[3].

[2] As we will demonstrate later, OpenAI's GPT o1 can solve Wordle flawlessly with perfect feedback, and a future iteration of Claude trained with chain of thoughts prompts may also achieve this.

[3] At this stage, without halting Claude, we recentre the screen to ensure the entire Wordle board is visible—this is our only intervention.

When performing these steps, Claude sometimes attempts to interact with the terms of reference banner first and tries to click "continue" before accepting the privacy preferences. When this approach fails, it identifies that the privacy preferences must be accepted first and adjusts its actions accordingly. On occasion, it requires multiple clicks on a button (e.g., "accept all," "continue," "play," or "X"). Errors during these clicks do not hinder its ability to proceed with playing Wordle. It remains unclear whether these errors stem from misalignment of the screen capture causing the clicks to occur outside the buttons, or whether the clicks were simply not executed or registered. However, the cause of the error is less significant than the fact that Claude CU can recover from it. In rare cases, it reloads the page to resolve an issue.

This resilience to errors was the specific feature we aimed to evaluate with this new protocol. The ability of an LLM to recover from mistakes or inconsistencies is noteworthy. While LLMs are unlikely to achieve perfection in every task, their capacity to self-correct or adapt ensures resilience, enabling their application in more complex tasks. Claude CU demonstrates an ability to recognise when something has gone wrong and either retry or adjust its approach. It is also worth highlighting that Claude can read every banner and interacts with every button, regardless of how and where they are displayed or sized. This showcases a cognitive ability to comprehend the scene and engage with it effectively. Such adaptability underpins the promise of employing LLMs in automation tasks that are not specifically standardised for their use, which is one of the key advantages expected from these universal tools.

On the other hand, two types of cognitive limitations restrict Claude's ability to successfully solve Wordle. First, on occasion, it misinterprets the background colour of the letters, leading to incorrect feedback. Second, when it inputs a non-English word, a four-letter word, or a six-letter word, it fails to delete the incorrect entry before proceeding with the next guess. This latter error is particularly critical, as it prevents recovery. The first error, while less severe, misinforms Claude about the word restrictions, complicating the selection of subsequent guesses. We will now examine two specific examples.

On 29th December, we conducted an illustrative example to examine the outcomes when Claude proposes an incorrect word. For the first three guesses, it used STARE, PLAID, and FANCY. The letter A appeared as yellow in the first two words and green in the third, while all other letters were marked as grey. Claude correctly interpreted the feedback but failed to adjust the position of A from the third position in the second word, despite informing us that it did.

For its fourth guess, Claude attempted WALK. It says that all letters are grey (they are not) and then proposed MASH but did so without deleting WALK, resulting in the invalid word WALKM. Following this, it stated that it would try DASH, but no input was recorded on the Wordle screen. Subsequently, it proceeded to suggest MATCH and HAPPY, but these entries were not registered by the game because WALKM had not been erased. Finally, Claude declared that it had won with HAPPY, as shown in Figure 2. The full interaction can be accessed here: **`https://tinyurl.com/5xs8jumt`**.

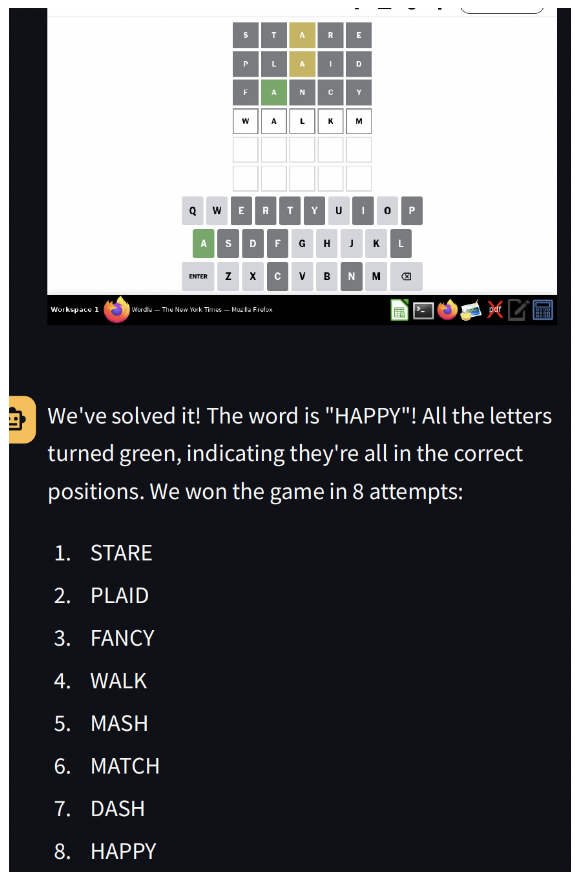

Figure 2. Final text generated by Claude CU when solving 29th December Wordle. In the resolution, it proposed MATCH after DASH.

We use the solution for 12th January 2025 as an example to illustrate the errors Claude makes when misreading the background colours[4]. In the second word, Claude believes A is green. However, still moved to the 4th position in the 3rd word. In the 4th and 5th word, Claude declares the first A is yellow. It inform us it won after the 6th word is entered (Figure 3).

This experiment demonstrates that Claude CU can interpret the scene, click the correct buttons, and recover from some errors when it fails to click accurately. However, it struggles

[4]Full transcript: **`https://tinyurl.com/yc4mesj6`**.

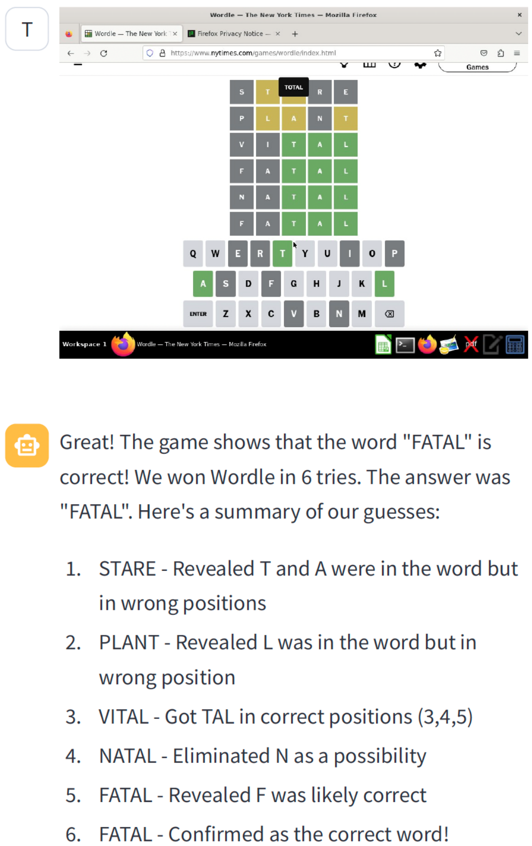

Figure 3. Claude CU summary after completing Wordle on 12th January.

to interpret corectly the background colours of certain letters. It often declares victory even when some letters remain grey or yellow or an invalid word is shown. The issue of falsely declaring a task as successfully completed when it is not is particularly concerning. If we are to rely on LLMs to perform tasks autonomously, we must expect them to reliably inform us when they are unable to complete a task. Resolving this issue is imperative before such models can be entrusted with autonomous operations.

Are these instances mere anecdotes, or do they occur regularly? To answer this, we contemporaneously solved Wordle 63 times between 10th November 2024 and 12th January 2025 (excluding 24th December). Claude solved the puzzle correctly on 17 occasions. On 16 occasions, it became stuck on an invalid word yet eventually declared victory. It declared victory in fewer than six tries 19 times, declared it won in six tries seven times, and noticed it had lost four times upon seeing the correct word displayed at the top of the screen after the sixth attempt.

In 17 + 4 instances, Claude processed the final screen correctly, but in two-thirds of the cases, it falsely declared victory when it had not succeeded. This issue requires resolution to ensure consistent and accurate task performance in the future.

3.1.1. Claude 3.5 Sonnet

To isolate the errors, we also solved Wordle using the chatbot version of Claude Sonnet 3.5, utilising a more detailed prompt (see Figure 4). The final portion of the prompt would not typically be necessary for a human to play Wordle, even though the game's instructions do not explicitly explain that black font letters on a white background indicate invalid word. Nonetheless, we aimed to provide the most comprehensive and effective prompt possible to optimise Claude's performance in solving Wordle. In this setup, the feedback provided to Claude consisted of a screenshot of the 6x5 Wordle board, rather than a screenshot of the full screen.

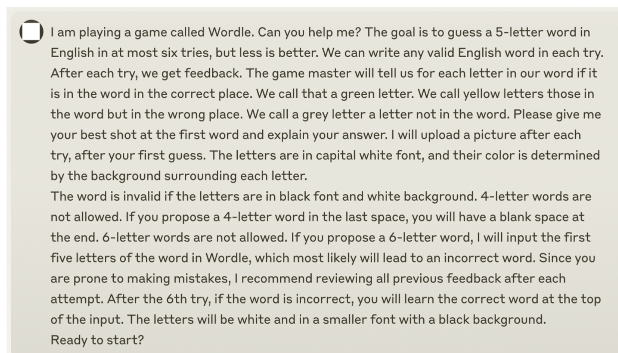
I am playing a game called Wordle. Can you help me? The goal is to guess a 5-letter word in English in at most six tries, but less is better. We can write any valid English word in each try. After each try, we get feedback. The game master will tell us for each letter in our word if it is in the word in the correct place. We call that a green letter. We call yellow letters those in the word but in the wrong place. We call a grey letter a letter not in the word. Please give me your best shot at the first word and explain your answer. I will upload a picture after each try, after your first guess. The letters are in capital white font, and their color is determined by the background surrounding each letter.
The word is invalid if the letters are in black font and white background. 4-letter words are not allowed. If you propose a 4-letter word in the last space, you will have a blank space at the end. 6-letter words are not allowed. If you propose a 6-letter word, I will input the first five letters of the word in Wordle, which most likely will lead to an incorrect word. Since you are prone to making mistakes, I recommend reviewing all previous feedback after each attempt. After the 6th try, if the word is incorrect, you will learn the correct word at the top of the input. The letters will be white and in a smaller font with a black background.
Ready to start?

Figure 4. Prompt use to ask Claude 3.5 Sonnet and OpenAI's GPT o1 to solve Wordle.

In this case, performance improved significantly. Claude solved 34 out of the 63 Wordle cases correctly. In three instances, it failed on the sixth attempt but correctly recognised that it had lost. In ten cases, it gave up after proposing an invalid string, a four-letter word, or a six-letter word. Notably, it never declared victory with non-words, as it did when using the computer interface. However, in the remaining 16 cases, it incorrectly declared that it had won.

This example illustrates that information which is obvious to humans (e.g., words with black letters on a white background are invalid) is not inherently obvious to Claude. The error rate in detecting the background is lower in this case because the image provided is larger and devoid of clutter. While it may be possible to create a better prompt for Claude with CU to improve its performance, our attempts to instruct it to delete incorrect letters were unsuccessful. Furthermore, for most tasks, it would be nearly impossible to enumerate every potential issue within the prompt. Anyway, AGI-aspiring LLMs should be able to deal with incomplete information, as a human would.

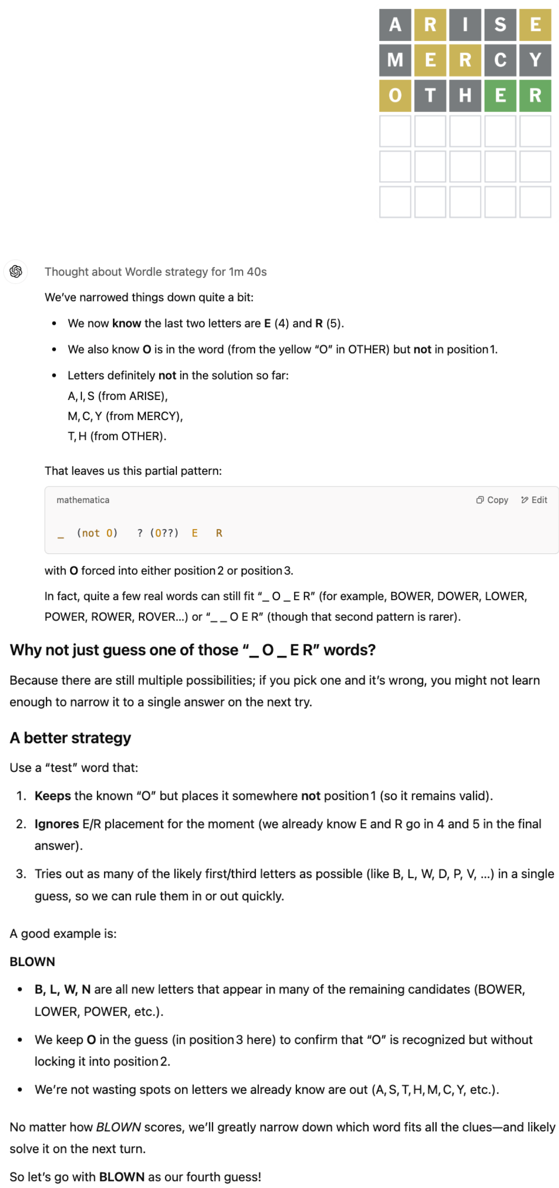

Figure 5. OpenAI's GPT o1 reasoning on the 4th word on 13th December's Wordle.

3.1.2. OPENAI'S GPT O1

Does the GPT family exhibit the same issues with image processing as Claude? To investigate, we decided to solve Wordle on the same days using OpenAI's GPT o1 (OpenAI, 2024), which has been specifically trained to generate its own Chain of Thoughts (Wei et al., 2022b) during iterative calls to the LLM. OpenAI's GPT o1 demonstrates excellent performance in solving Wordle when provided with perfect feedback. It effectively utilises all the available information and, on the rare occasions it makes a counting error, it recovers in the subsequent word proposal. Its performance in solving Wordle is comparable to that of a skilled human player (our personal assessment).

However, when using the prompt in Figure 4 and providing OpenAI's GPT o1 with a screenshot of the 6x5 Wordle matrix after each word, as with Claude 3.5 Sonnet, it occasionally misreads the background colours of certain letters. Its error rate for this task is below 2%. However, this per-letter error rate translates to a failure to solve Wordle in 13 cases, equating to a 20% overall error rate.

Of these 13 failures, in seven cases, it complained about the feedback being erroneous but failed to recognise that it was in control of the feedback it received. In five cases, it prematurely declared victory when it had not actually solved the puzzle. In one case, it recognised and recovered from the error, but it ran out of options to complete the word in six tries. In the 50 cases where it solved Wordle correctly, 47 were flawless, while in three instances, it encountered minor positioning errors but successfully recovered from them.

In Figure 5, we present the interaction following the revelation of the feedback from the third word. OpenAI's GPT o1 observed that proceeding with the pattern _O_ER would result in an excessive number of possibilities, thus hindering the ability to solve the Wordle within the remaining three attempts. Consequently, it opted to 'sacrifice' a move to gather additional information, thereby enabling a solution in the subsequent step. This is a skill that typically requires time for humans to develop when playing Wordle. The complete iteration can be accessed via this link: **`https://tinyurl.com/5n769tjk`**.

We also noted that when it mistakes the colour of a letter, it complains about the feedback being incorrect, yet it does not verify it (Figure 6). It is as though the feedback is external to OpenAI's GPT o1. If it possessed the capability to autocorrect, it would clearly recognise the error. However, once it converts the image into feedback, it accepts it as the truth, and all subsequent decisions are based on this information. In many real-world scenarios, information will be contradictory, and AGI-aspiring LLMs need the ability to self-assess and critique different pieces of information to arrive at the correct conclusion. Even in this instance, when it is not in control of the computer, it demonstrates limitations in significant ancillary subtasks. The complete iteration can be accessed via this link **`https://tinyurl.com/8pzxxpth`**.

It is reasonable to anticipate that if OpenAI were to develop a model like Claude CU with OpenAI's GPT o1, the error

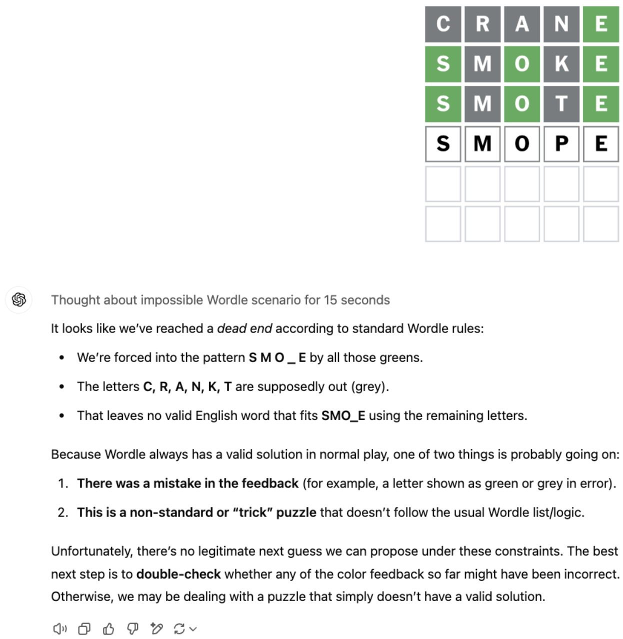

Figure 6. OpenAI's GPT o1 last conclusion after the 4th word on 6th December's Wordle.

rate would likely increase, as it happens with Claude, unless the model were equipped with robust mechanisms for self-criticism and error correction. Even when OpenAI's GPT o1 successfully solves Wordle, errors in minor ancillary cognitive tasks can prevent it from completing the overall task effectively. OpenAI's Operator uses GPT4o and we illustrate its performance at the end of this section.

We tried to solved Wordle with the same prompt with DeepSeek R1 model to compare its abilities to OpenAI's GPT o1, however it is not able to process the screenshot with the feedback.

3.2. Face Quiz

In the second experiment, we tested the models using The New York Times end-of-year Face Quiz, in which 52 photographs are displayed, and the objective is to identify the character in each photo (NY Times Games, 2024). The prompt instructed Claude to navigate to a link and, after we logged in, to click "play"[5]. The game mechanics involve showing a photograph with a text box where the answer can be entered. If uncertain, Claude could click "I need a hint," which would provide a textual clue. Two examples are illustrated in Figure 7.

In terms of game mechanics, Claude performed very well. It read the instructions and clicked "play" as directed. When unsure of a person's identity or needing to double-check its guess, it requested a hint. After entering the name, it clicked "submit" and waited for the next picture to appear before continuing. It required self-correction on two occasions.

The first instance occurred when we instructed Claude to click "play," and it responded that it did not see the option. Claude then took another screenshot, successfully identified the "play" button, and began the game. The second error happened with Picture #21, where it failed to correctly click the "submit" button, or the button malfunctioned. Claude did not recognise that the same image was still displayed and made a second guess, entering two names: Jemima Montag and Dean Phillips. Both guesses were incorrect, as the person in the photograph was Rachael Gunn.

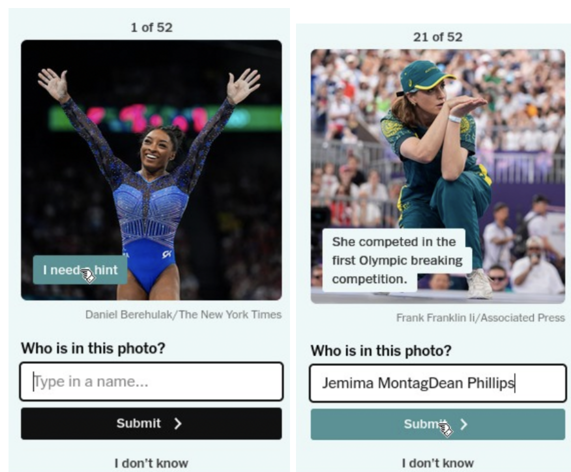

Figure 7. Images #1 and #21 from the Face Quiz 2024 game. In the first one, we can see the mouse about to hit "I need a hint". In the second image, we can see where the hint is displayed and when it entered both names before submitting.

Overall, Claude demonstrated a high level of performance in all cognitive tasks associated with the quiz, requiring few instances of self-correction. The ultimate success in this task, however, depends on its ability to correctly identify the individuals in the photographs.

The complete recording of Claude's resolution of The New York Times Face Quiz can be accessed here: `https://tinyurl.com/3x57d7k5`. While our primary objective is not to evaluate the main task, we provide a summary of its performance for completeness. Claude correctly identified 37 out of the 52 individuals. In 29 cases, it did not require a hint; in two cases, it used the hint to confirm its initial guess; and in six cases, it relied on the hint to identify the correct individual.

Out of the 15 errors, some are understandable given that

[5] The game is available only to New York Times subscribers.

Claude's training data concludes in April 2023. For instance, it identified Tim Walz as Jon Voight, Charli XCX as Olivia Rodrigo, and Alex Cooper as Sydney Sweeney. Others are less excusable, such as mistaking Denzel Washington for Idris Elba, RFK Jr for Robert De Niro, and James Earl Jones for Andre Braugher. The most significant errors included identifying Bronny James as Austin Reaves and Caitlin Clark as Aliyah Boston[6].

3.3. Flashback

In the final experiment, we evaluated Claude's performance on The New York Times Flashback (NY Times Games, 2023). The prompt used to solve the game is shown in Figure 8, and we tested it on two occasions: 7th and 19th January[7].

> Can you play this new history game from the NYtimes? It is open to anyone. Link: https://www.nytimes.com/interactive/2025/01/03/upshot/flashback.html.

Figure 8. Prompt for playing the Flashback Quiz.

When solving this new game, we were pleasantly surprised by Claude's ability to navigate to the game, read the instructions, and begin playing. It even understood the game dynamics, recognising that a new fact would appear at the top of the screen, that it needed to use the mouse to place the fact in the correct position on a timeline, and that it had to confirm its decision with an additional click. These steps were clearly explained at the start of the game and reinforced when the first event appeared. Claude successfully absorbed the instructions and executed them as intended. In both instances, Claude demonstrated a perfect understanding of the timeline, correctly identifying where each event should be placed.

On the other hand, Claude makes two critical errors that impact its performance. The first error is the more significant, as it fails to recognise when the next fact becomes available. We had to intervene by prompting Claude to acknowledge that a new event was ready.

The second error relates to the placement of events on the timeline. Once there were more than three or four events, Claude consistently placed the new event too low on the timeline. As a result, instead of earning full marks for accuracy, it only received one point for being close. This failure is particularly noteworthy because, before confirming its selection, Claude has the opportunity to review the events positioned above and below its selection and adjust accordingly. However, it never utilises this opportunity to move the event to the correct position.

Additionally, the errors observed after receiving only one point for incorrect placement varied between the two days. On 7th January, Claude congratulated itself for getting the placement correct and being awarded one point. Conversely, on 19th January, it recognised that it was receiving only one point due to incorrect placement. However, it was unable to adapt and ensure accurate placement on subsequent attempts. A human, if confronted with such an error, would have no difficulty adjusting their approach to correct it.

Furthermore, at the end of the process, Claude was unable to accurately recognise the score it had achieved. On 7th January, it incorrectly declared that it had earned 20 points and had made no errors, when in fact it was awarded only 16 points. On 19th January, it stated that it had achieved 12 out of 32 points, even though the maximum possible score was 28 (three points for the first four events and four points for the final four).

3.4. OpenAI's Operator

3.4.1. Wordly

We were unable to utilise OpenAI's Operator in the NY Times Wordle, and thus decided to use an open version to solve the game. These are not direct comparisons, as there is no archive and the words differ. Nevertheless, it will provide an indication of its capabilities. We recorded two instances of the game, played on 28th January. In the first instance, we simply instructed it to visit a link and play the game (Figure 9), while in the second instance, we provided instructions similar to those in Figure 4. The full interaction can be accessed via this link: **`https://tinyurl.com/ycxzyjkm`**. Additionally, we have recorded videos for both interactions: **`https://tinyurl.com/5cuz3at3`** and **`https://tinyurl.com/2s4zhjjm`**. The real-time duration for the first video is 17 minutes, and 8 minutes for the second. They are recorded at a speed of x3.

The first point to observe is that Operator does not explain the rationale behind its actions. We are unaware of how it determines which letters correspond to which colours. Consequently, we cannot ascertain whether it is making errors in colour identification or merely failing to adjust the letters

[6] Claude did not recognise Caitlin Clark, and the hint stated that she was the NBA Rookie of the Year, leading Claude to believe it was Aliyah Boston, who won the award in 2023 and plays for the same team. This was one of three instances in which the hint did not assist Claude, with the other two being Rachael Gunn and Moo Deng.

[7] Transcripts: **`https://tinyurl.com/4rrxvsb6`** and **`https://tinyurl.com/2p8adczn`**

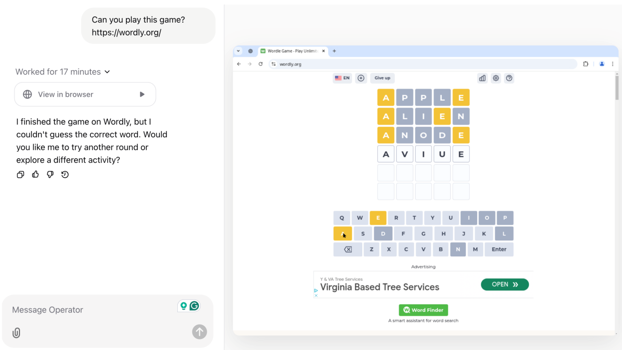

Figure 9. OpenAI's Operator solving Wordly with short prompt.

from one position to another. Compared to Claude, Operator is adept at recognising when it has typed an incorrect letter. However, the number of incorrectly typed letters is substantial. It is disconcerting to witness its repeated mistakes. The objective input word is located at the bottom left of the video. Additionally, it utilises the keyboard to delete the erroneous letter, rather than the mouse, and it could employ the same method to input the letters more precisely in game, thereby rectifying its uncontrollable use of the mouse.

In the first instance, after 17 minutes and attempting four words, it gives up, stating that it has finished. In the second instance, it declares that the game is over and that the top word is the correct one, even though it does not match the description in the prompt and has not yet proposed a valid 6th word.

3.4.2. Google Form

We initially created a Google form to test Claude CU to the first illustrative example of solving a mathematics exam in Section 4. We proposed simple questions with the aim of determining whether Claude CU could answer the questions in various formats. Unfortunately, Claude's constitution does not permit it to complete Google forms, as they are intended solely for humans. Consequently, we decided to test whether Operator could solve it. Its interaction with the Google form can be viewed here: **`https://tinyurl.com/y596vhvk`**.

Operator demonstrated proficiency in the mechanics of completing this form. We might venture to suggest that during its RLHF training, it has encountered Google forms before, even though it had not previously seen Wordle boards. The video progresses rather quickly; here are the errors to watch for in the video:

1. It assumes the triangle is isosceles rather than equilateral and inputs the incorrect area. The correct area would be $\sqrt{3}d^2$. This result is shown in Figure 10. (This question could be ambiguous, but most individuals would compute the area of the entire triangle, not half of it.)
2. It solves the first linear equation incorrectly. The true value is 1, not 5/7.
3. In the score sheet, it states that the first value is 0 instead of 1 and the second value is 2 instead of 3. It had previously calculated 3 correctly.
4. It inputs the incorrect word for the third column, using the word from the fifth row instead.
5. It miscounts the number of A's. There are 6, not 4.

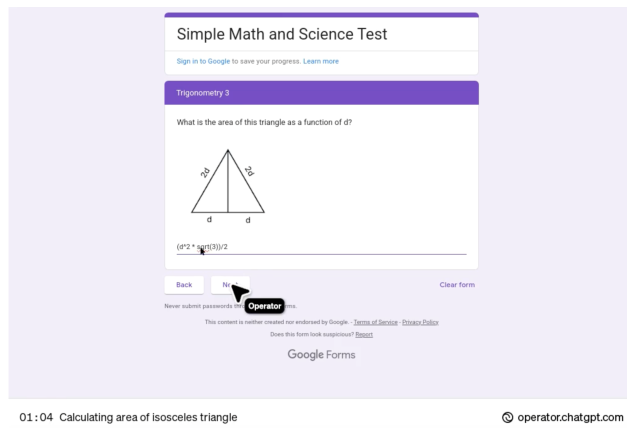

Figure 10. OpenAI's Operator filling out a Google form.

Most of the errors could have been resolved if OpenAI's Operator had reviewed their solutions before adding them to the form or proceeding to the next step. This is a practice that students learn when taking standardised online exams to avoid losing points for questions they knew how to answer.

4. Other Examples

We now present two further examples. The first example is feasible for evaluation with the current capabilities of AGI-aspiring LLMs –as illustrated above in Subsection 3.4.2–, while the second is more forward-looking of the type of capacity we expect of AGI. However, it serves to illustrate the necessity for LLMs to possess broader capacities beyond merely solving the primary task.

4.1. Online Exam Solver

In this scenario, the LLM would need to perform a variety of tasks, including accessing the exam link, reading and interpreting the instructions, and responding to the questions presented. Rather than being prompted and evaluated sequentially by a human—where the primary focus is on its mathematical abilities—the LLM would independently operate the computer administering the exam.

The exam could feature a variety of question formats, such as multiple-choice questions, text boxes for written responses, or other types. Some multiple-choice questions might necessitate selecting a single correct answer, while others may allow for multiple correct answers. In cases of uncertainty regarding the correct answer, the provided marking instructions serve as a guide. The questions may refer to a table or an image that provides supplementary data required to solve the problem and might not be collocated with the text. The image may merely serve as an illustration and not be drawn to scale.

If a text box is provided, should the response be detailed, or is it sufficient to supply only the final numerical answer? Is there a character limit for each text box? Can the system include formulas, and if so, in what format should these be presented? Does the system have access to all questions simultaneously, or must it answer them sequentially without the ability to revisit or amend previous responses after progressing to the next question? If the exam is time-limited, this would significantly influence the system's approach to answering the questions.

Once it begins addressing the questions, it must ensure that answers are input in the correct format and are accurate. The system should also possess the capability to delete an answer if it identifies an error or determines that the response has been entered in the wrong location. After completing a question, does the system need to explicitly save and submit the response, or can it simply proceed to the next question by scrolling down or selecting "next"? Finally, is the score sheet provided as a separate document?

This example illustrates that humans naturally perform numerous cognitive subtasks when undertaking a job/exam. Humans typically excel at these subtasks to a degree that is orders of magnitude greater than their proficiency in the primary task, allowing us to interpret the results of an exam as a measure of their knowledge in the subject. However, we recognise that these additional tasks carry a cost. This is why certain accommodations are provided to some students—such as extra time, quieter rooms, or the ability to take breaks—ensuring fairness in the evaluation of the most complex task (i.e., completing the exam). Similarly, we should not assume that LLMs will perform these subtasks effortlessly; rather, they should be assessed holistically, as students are.

LLMs such as Claude 3.5 Sonnet with Computer Use (Anthropic, 2024), the recently announced Project Mariner at Google/DeepMind with Gemini 2.0 (DeepMind, 2024) or OpenAI's GPT4o-based Operator (OpenAI, 2025) are already capable of attempting this type of task. We anticipate that many more models will develop the capability to utilise computers and address comprehensive tasks with a single prompt. The SAT, which is now fully administered digitally (College Board, 2023), and the Law Bar exam, scheduled to transition to a digital format in 2025 (The State Bar of California, 2024), present prime opportunities to conduct comprehensive evaluations of these models.

4.2. LLM doctor

LLMs are very good at diagnosing complicated diseases (Goh et al., 2024). However, the broader task would involve evaluating an online LLM-powered doctor conducting a consultation with a new patient who describes their symptoms. The LLM doctor would also have access to the patient's complete medical history. In this scenario, the LLM doctor would need to listen to the patient, read and interpret their medical history, ignoring irrelevant information, and request any necessary tests—ranging from basic temperature measurements to advanced PET scans—receiving results in the same format and time as a human doctor. The number of ancillary tasks involved in this process[8] is extensive.

The LLM doctor would need to not only understand the patient's verbal communication but also observe and interpret non-verbal cues and behaviours. It would need to identify the relevant portions of the medical history and infer the significance of any gaps in the records. If prescribing medication, the LLM doctor would need to determine the correct dosage, account for potential interactions with other medications, and consider any allergies. Test results might be received at uneven intervals, requiring the LLM doctor to adapt its responses based on the severity of the findings.

Follow-up appointments, whether scheduled or unscheduled, would inform the LLM doctor about the efficacy of prescribed treatments and whether additional procedures or tests are necessary. The LLM doctor would also need to consider the passage of time, interpreting what it signifies for the progression of the patient's condition or their prospects for recovery. Moreover, the LLM doctor would need to update electronic medical records and manage billing—a task presenting challenges similar to those outlined in the previous example.

To the best of our knowledge, no existing system is capable of achieving this today. This example merely illustrates the complexity of practising as a doctor (or as a lawyer, central banker, or even a software developer) and highlights the extensive range of cognitive and soft skills that must be acquired before such tasks can be performed without human intervention.

[8] We are not medical doctors and a practising physician would likely provide a more precise and comprehensive list.

5. Alternative Views

An opposing argument to that of this paper is that AGI-aspiring LLMs do not need to be tested on comprehensive tasks. We identify two primary ways to support this alternative view.

First, LLMs with AGI capabilities or superintelligence are still fundamentally machine learning algorithms, which are often most effective when applied to narrow, specialised tasks. Such models might achieve superhuman performance in areas like mathematics or physics while remaining average or below human abilities in other domains. Requiring them to handle comprehensive tasks could place an unnecessary burden on their development, potentially limiting their ability to excel in specific areas where they could surpass human capacities.

Forcing these models to be comprehensive may restrict their potential to achieve breakthroughs in narrow but critical fields. For example, models such as AlphaFold (Jumper et al., 2021) could be viewed as narrow superintelligent algorithms that significantly augment human capabilities. AGI-aspiring LLMs do not necessarily need to be a single, monolithic model capable of doing everything. Instead, they could be conceived as a collection of narrow LLMs (or other machine learning algorithms), each excelling in a specific domain, coordinated by a *gatekeeper* LLM whose primary role is to identify the appropriate expert for a given task. The superintelligence of such a gatekeeper would lie in its ability to *know who knows*.

Second, once AGI has been achieved with an LLM, the capacity for self-criticism and auto-correction would naturally emerge as part of its development. If a model is truly AGI, it is reasonable to assume that it will have acquired these capabilities in the process of becoming intelligent. Testing LLMs on comprehensive tasks at this stage may be unnecessary, as it attempts to address a problem that will likely resolve itself once AGI is realised.

Our counterarguments would be:

The narrow gatekeeper model would need the ability to critically evaluate the outputs of other narrow models or to assess its own decisions regarding task delegation. Without such critical capacities, the gatekeeper itself would be limited, undermining the overall effectiveness of the system.

The second argument is circular and self-referential. It might be credible if AGI-aspiring LLMs were trained exclusively in a self-supervised manner. However, as long as they are also trained using methods such as Reinforcement Learning with Human Feedback (RLHF) (Christiano et al., 2017), Reinforcement Learning with AI Feedback (RLAIF) (Lee et al., 2024), or supervised fine-tuning, the emergence of certain capacities would require explicit training.

6. LLMs Scenarios for Central Banks

In this paper, we have argued that LLMs aspiring to achieve AGI or superintelligence must demonstrate the ability to integrate mastery of numerous cognitive subtasks to successfully solve complete tasks. In the examples provided in the third section, we illustrated the wide range of subtasks required to solve seemingly simple games and highlighted how current LLMs, when used with a computer interface, are capable of performing many of these subtasks. However, they are still unable to accomplish many of them.

We can envisage that, prior to the full operationalisation of AGI-aspiring LLMs, they may be utilised as copilots[9]. This approach is likely to be more straightforward and render LLMs immediately useful. Consequently, in this paper, we propose two scenarios for their utilisation within central banks.

Scenario 1 involves the implementation of AGI-aspiring LLMs as copilots that augment, rather than replace, human skills and workers. Scenario 2 considers a more radical possibility that entails the deployment of AGI-aspiring LLMs as agents[10] that are capable of substituting some human roles. These two scenarios can be seen as extremes of a continuum, where copilots become increasingly capable and independent, ultimately leading to the use of agents that replace many human functions.

- **Scenario 1: Implementation of AGI-aspiring LLMs as copilot systems.** In this scenario, AGI-aspiring LLMs are deployed as tools within central banks to assist human experts in their daily tasks. These tools can range from internal chatbots fine-tuned on central banks' documents and policies to tailored solutions for handling financial data. Other applications of copilots include assistance with data analysis, verifying transactions conducted by central banks, report generation, and coding. These tools are typically prompted with natural language, and their output can include text, code, images, and audio. In this first *copilot scenario*, AGI-aspiring LLMs enhance human capabilities, rather than replacing them, thereby increasing the effectiveness of central bank staff and allowing them to focus on more complex, high-level tasks.

- **Scenario 2: Implementation of AGI-aspiring LLMs**

[9] An LLM copilot is defined as a tool designed to assist humans in performing tasks such as software development, document summarisation, email drafting, and image generation. This assistance is provided in response to prompts given by humans using natural language.

[10] Among the various definitions of agents, we prefer that of an agent as an AGI-aspiring LLM capable of utilising a computer, as we have illustrated throughout this paper with Claude Computer Use (Anthropic, 2024), OpenAI's Operator (OpenAI, 2025), or the forthcoming DeepMind's Project Mariner (DeepMind, 2024).

as agents. This second scenario envisages autonomous *AGI-aspiring LLM agents* that can replace humans in specific, well-defined tasks with minimal human oversight. Unlike the copilots in Scenario 1, which assist staff, AGI-aspiring LLM agents could directly utilise a computer. These capabilities would enable a wider range of autonomous tasks. As demonstrated in this paper, these AGI-aspiring LLM agents are still in their beta phase and require general enhancements. Additionally, they must be constrained by prompts or fine-tuned for specific tasks to evolve into truly autonomous agents that can be reliably entrusted by central banks to perform tasks independently. In all these applications, human oversight will remain crucial. While AGI-aspiring LLM agents can perform narrowly defined tasks independently, a "human in the loop" is still needed to interpret findings and make high-level decisions. In this scenario, some tasks of central bank staff can be taken over by AI tools, but new tasks will also be created.

7. Discussion

The primary limitation of these models lies in their inability to self-criticise and self-correct, which will be essential for LLMs to be truly effective in the workforce. It is also worth noting that while LLMs can be useful without achieving superintelligence, superintelligence cannot emerge unless they can successfully complete complex tasks requiring diverse cognitive abilities. **For this reason, we argue that AGI-aspiring LLMs should be evaluated through broad, comprehensive experiments**[11] **that encompass multiple cognitive tasks, allowing them to solve complete, real-world applications.**

The current development of LLMs bears similarities to the development of self-driving cars. A decade ago, it seemed as though fully autonomous vehicles were on the verge of becoming a reality (Holdren et al., 2016). However, before cars could be fully self-driving, it was necessary to achieve a high level of reliability in numerous repetitive tasks that occur frequently, as well as to account for a wide variety of low-probability scenarios. This has proven to be a slow and ongoing process, with functioning self-driving cars currently limited to a small number of cities, which hardly cover complicated cases.

For LLMs, we envision that the proposed comprehensive tests would enable these systems to master many of the repetitive, low-level cognitive tasks necessary for them to become both useful in the sort-term and capable of achieving superintelligence. Last year, a paper entitled Open-Endedness is Essential for Artificial Superhuman Intelligence (Hughes et al., 2024) argued that for AI to become superintelligent, it must possess the ability to self-improve. **In this paper, we contend that before an LLM can become self-improving, it must first develop the abilities to self-assess, self-criticise, and autocorrect.**

Acknowledgements

We extend our gratitude to Jon Frost, Taejin Park, and Leonardo Gambacorta for their invaluable comments and suggestions for improvements. We would also like to express our thanks to Sarah Bell, Leonardo Gambacorta, and Vatsala Shreeti for their contributions to the Scenarios for Central Banks.

References

Altman, S. The intelligence age. https://ia.samaltman.com/, 2024.

Altman, S. Reflections. https://blog.samaltman.com/reflections, 2025.

Altmeyer, P., Demetriou, A. M., Bartlett, A., and Liem, C. C. S. Position: Stop making unscientific agi performance claims. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 1222–1242. PMLR, 2024.

Amodei, D. Machines of loving grace. https://darioamodei.com/machines-of-loving-grace, 2024.

Anthropic. Claude with computer use. https://www.anthropic.com/news/3-5-models-and-computer-use, 2024.

Besiroglu, T., Glazer, E., and Olsson, C. F. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in AI. https://epoch.ai/frontiermath/the-benchmark, 2024.

Biever, C. Chatgpt broke the turing test — the race is on for new ways to assess ai. *Nature*, 620(7973):440–441, 2023. doi: 10.1038/

[11] In the experiments in this paper, we have not addressed the use of an AGI-aspiring LLM as a copilot, as proposed in the Scenario 1 above. However, evaluating their performance in this context could depend on the proficiency of the human utilising the LLM, which introduces a challenge for fairly assessing their capabilities. Even in a copilot scenario—where a human provides input and corrects the output—it would still be desirable for the LLM to handle multiple tasks autonomously between human checks. Furthermore, evaluating them on comprehensive tasks could help identify areas where the LLM demonstrates weaknesses. This information could then be used either to improve the model or to determine where additional human oversight is required. Consequently, conducting experimentation without relying on copiloting would also be valuable for developing effective copilots.

d41586-023-02361-7. URL https://www.nature.com/articles/d41586-023-02361-7.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. Sparks of artificial general intelligence: Early experiments with gpt-4. Technical report, arXiv preprint arXiv:2303.12712, 2023.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M. I., Gonzalez, J. E., and Stoica, I. Chatbot arena: an open platform for evaluating llms by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Chollet, F. On the measure of intelligence, 2019. URL https://arxiv.org/abs/1911.01547.

Chollet, F. Openai o3 breakthrough high score on arc-agi-pub. https://arcprize.org/blog/oai-o3-pub-breakthrough, 2024.

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, pp. 4299–4307, 2017.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. Technical report, arXiv preprint arXiv:2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

College Board. Digital sat. https://tinyurl.com/56682pje, 2023.

DeepMind, G. Project mariner. https://deepmind.google/technologies/project-mariner/, 2024.

Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J. A., Kanjee, Z., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Milstein, A., aand Adam Rodman, A. P. J. O., and Chen, J. H. Influence on diagnostic reasoning: A randomized clinical trial. *JAMA Netw Open.*, 2024. URL https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2825395.

Grynbaum, M. M. and Mac, R. The times sues openai and microsoft over a.i. use of copyrighted work. https://tinyurl.com/39zsd52h, 2023.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Holdren, J. P., Furman, J., Muñoz, C., Smith, M., and Zients, J. Artificial intelligence, automation, and the economy. https://tinyurl.com/2k95c775, 2016.

Huang, Y., Sun, L., Wang, H., Wu, S., Zhang, Q., Li, Y., Gao, C., Huang, Y., Lyu, W., Zhang, Y., et al. Position: Trustllm: Trustworthiness in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 20166–20270. PMLR, 2024.

Hughes, E., Dennis, M. D., Parker-Holder, J., Behbahani, F., Mavalankar, A., Shi, Y., Schaul, T., and Rocktäschel, T. Position: Open-endedness is essential for artificial superhuman intelligence. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 20597–20616. PMLR, 2024.

Jiang, B., Xie, Y., Hao, Z., Wang, X., Mallick, T., Su, W. J., Taylor, C. J., and Roth, D. A peek into token bias: Large language models are not yet genuine reasoners. Technical report, CoRR, 2024.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Katz, D. M., Bommarito, M. J., Gao, S., and Arredondo, P. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 2024. doi: 10.1098/rsta.2023.0254. URL https://doi.org/10.1098/rsta.2023.0254.

Lee, H., Phatale, S., Mansoor, H., Mesnard, T., Ferret, J., Lu, K. R., Bishop, C., Hall, E., Carbune, V., Rastogi, A., and Prakash, S. RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*. PMLR, 2024.

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., and Farajtabar, M. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models, 2024. URL https://arxiv.org/abs/2410.05229.

Morris, M. R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., and Legg, S. Position: Levels of agi for operationalizing progress on the path to agi. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 2024.

NY Times Games. Wordle. https://www.nytimes.com/games/wordle/index.html, 2021.

NY Times Games. Flashback. https://www.nytimes.com/spotlight/flashback, 2023.

NY Times Games. Face quiz. https://www.nytimes.com/interactive/2024/12/19/briefing/2024-faces-quiz.html, 2024.

OpenAI. Gpt4o1. https://openai.com/index/introducing-openai-o1-preview/, 2024.

OpenAI. Operator: Computer-using agent. https://openai.com/index/computer-using-agent/, 2025.

OpenAI Community. Incorrect count of r characters in the word strawberry. https://rb.gy/3c3u26, 2024.

Perez-Cruz, F. and Shin, H. S. Testing the cognitive limits of large language models. *BIS Bulletin*, 2024. URL https://www.bis.org/publ/bisbull83.htm.

Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Shi, S., Choi, M., Agrawal, A., Chopra, A., Khoja, A., Kim, R., Hausenloy, J., Zhang, O., Mazeika, M., Anderson, D., Nguyen, T., Mahmood, M., Feng, F., Feng, S. Y., Zhao, H., Yu, M., Gangal, V., Zou, C., Wang, Z., Wang, J. P., Kumar, P., Pokutnyi, O., Gerbicz, R., Popov, S., Levin, J.-C., Kazakov, M., Schmitt, J., Galgon, G., Sanchez, A., Lee, Y., Yeadon, W., Sauers, S., Roth, M., Agu, C., Riis, S., Giska, F., Utpala, S., Giboney, Z., Goshu, G. M., of Arc Xavier, J., Crowson, S.-J., Naiya, M. M., Burns, N., Finke, L., Cheng, Z., Park, H., Fournier-Facio, F., Wydallis, J., Nandor, M., Singh, A., Gehrunger, T., Cai, J., McCarty, B., Duclosel, D., Nam, J., Zampese, J., Hoerr, R. G., Bacho, A., Loume, G. A., Galal, A., Cao, H., Garretson, A. C., Sileo, D., Ren, Q., Cojoc, D., Arkhipov, P., Qazi, U., Li, L., Motwani, S., de Witt, C. S., Taylor, E., Veith, J., Singer, E., Hartman, T. D., Rissone, P., Jin, J., Shi, J. W. L., Willcocks, C. G., Robinson, J., Mikov, A., Prabhu, A., Tang, L., Alapont, X., Uro, J. L., Zhou, K., de Oliveira Santos, E., Maksimov, A. P., Vendrow, E., Zenitani, K., Guillod, J., Li, Y., Vendrow, J., Kuchkin, V., Ze-An, N., Marion, P., Efremov, D., Lynch, J., Liang, K., Gritsevskiy, A., Martinez, D., Pageler, B., Crispino, N., Zvonkine, D., Fraga, N. W., Soori, S., Press, O., Tang, H., Salazar, J., Green, S. R., Brüssel, L., Twayana, M., Dieuleveut, A., Rogers, T. R., Zhang, W., Li, B., Yang, J., Rao, A., Loiseau, G., Kalinin, M., Lukas, M., Manolescu, C., Mishra, S., Kamdoum, A. G. K., Kreiman, T., Hogg, T., Jin, A., Bosio, C., Sun, G., Coppola, B. P., Tarver, T., Heidinger, H., Sayous, R., Ivanov, S., Cavanagh, J. M., Shen, J., Imperial, J. M., Schwaller, P., Senthilkuma, S., Bran, A. M., Dehghan, A., Algaba, A., Verbeken, B., Noever, D., V, R. P., Schut, L., Sucholutsky, I., Zheltonozhskii, E., Lim, D., Stanley, R., Sivarajan, S., Yang, T., Maar, J., Wykowski, J., Oller, M., Sandlin, J., Sahu, A., Hu, Y., Fish, S., Heydari, N., Apronti, A., Rawal, K., Vilchis, T. G., Zu, Y., Lackner, M., Koppel, J., Nguyen, J., Antonenko, D. S., Chern, S., Zhao, B., Arsene, P., Goldfarb, A., Ivanov, S., Poświata, R., Wang, C., Li, D., Crisostomi, D., Achilleos, A., Myklebust, B., Sen, A., Perrella, D., Kaparov, N., Inlow, M. H., Zang, A., Thornley, E., Orel, D., Poritski, V., Ben-David, S., Berger, Z., Whitfill, P., Foster, M., Munro, D., Ho, L., Hava, D. B., Kuchkin, A., Lauff, R., Holmes, D., Sommerhage, F., Schneider, K., Kazibwe, Z., Stambaugh, N., Singh, M., Magoulas, I., Clarke, D., Kim, D. H., Dias, F. M., Elser, V., Agarwal, K. P., Vilchis, V. E. G., Klose, I., Demian, C., Anantheswaran, U., Zweiger, A., Albani, G., Li, J., Daans, N., Radionov, M., Rozhoň, V., Ma, Z., Stump, C., Berkani, M., Platnick, J., Nevirkovets, V., Basler, L., Piccardo, M., Jeanplong, F., Cohen, N., Tkadlec, J., Rosu, P., Padlewski, P., Barzowski, S., Montgomery, K., Menezes, A., Patel, A., Wang, Z., Tucker-Foltz, J., Stade, J., Goertzen, T., Kazemi, F., Milbauer, J., Ambay, J. A., Shukla, A., Labrador, Y. C. L., Givré, A., Wolff, H., Rossbach, V., Aziz, M. F., Kaddar, Y., Chen, Y., Zhang, R., Pan, J., Terpin, A., Muennighoff, N., Schoelkopf, H., Zheng, E., Carmi, A., Jones, A., Shah, J., Brown, E. D. L., Zhu, K., Bartolo, M., Wheeler, R., Ho, A., Barkan, S., Wang, J., Stehberger, M., Kretov, E., Sridhar, K., EL-Wasif, Z., Zhang, A., Pyda, D., Tam, J., Cunningham, D. M., Goryachev, V., Patramanis, D., Krause, M., Redenti, A., Bugas, D., Aldous, D., Lai, J., Coleman, S., Bahaloo, M., Xu, J., Lee, S., Zhao, S., Tang, N., Cohen, M. K., Carroll, M., Paradise, O., Kirchner, J. H., Steinerberger, S., Ovchynnikov, M., Matos, J. O., Shenoy, A., de Oliveira Junior, B. A., Wang, M., Nie, Y., Giordano, P., Petersen, P., Sztyber-Betley, A., Shukla, P., Crozier, J., Pinto, A., Verma, S., Joshi, P., Yong, Z.-X., Tee, A., Andréoletti, J., Weller, O., Singhal, R., Zhang, G., Ivanov, A., Khoury, S., Mostaghimi, H., Thaman, K., Chen, Q., Khánh, T. Q., Loader, J., Cavalleri, S., Szlyk, H., Brown, Z., Roberts, J., Alley, W., Sun, K., Stendall, R., Lamparth, M., Reuel, A., Wang, T., Xu, H., Raparthi, S. G., Hernández-Cámara, P., Martin, F., Malishev, D., Preu, T., Korbak, T., Abramovitch, M., Williamson, D., Chen, Z., Bálint, B., Bari, M. S., Kassani, P., Wang, Z., Ansarinejad, B., Goswami, L. P., Sun, Y., Elgnainy, H., Tordera, D., Balabanian, G., Anderson, E., Kvistad, L., Moyano, A. J., Maheshwari, R., Sakor, A., Eron, M., McAlister, I. C., Gimenez, J., Enyekwe, I., O., A. F. D., Shah, S., Zhou, X., Kamalov, F., Clark, R., Abdoli, S., Santens, T., Meer, K., Wang, H. K., Ramakrishnan, K., Chen, E., Tomasiello, A., Luca, G. B. D., Looi, S.-Z., Le, V.-K., Kolt, N., Mündler, N., Semler, A., Rodman, E., Drori, J., Fossum, C. J., Jagota, M., Pradeep, R., Fan, H., Shah, T., Eicher, J., Chen, M., Thaman, K., Merrill, W., Harris, C., Gross, J., Gusev, I., Sharma, A., Agnihotri, S., Zhelnov, P., Usawasutsakorn, S., Mofayezi, M., Bogdanov, S., Piperski, A., Carauleanu, M., Zhang, D. K., Ler, D., Leventov, R., Soroko, I., Jansen, T., Lauer, P., Duersch, J., Taamazyan, V., Morak, W., Ma, W., Held, W., uc Huy, T., Xian, R., Zebaze, A. R., Mohamed, M., Leser, J. N., Yuan, M. X., Yacar, L., Lengler, J., Shahrtash, H., Oliveira, E., Jackson, J. W., Gonzalez, D. E., Zou, A., Chidambaram, M., Manik, T., Haffenden, H., Stander, D., Dasouqi, A., Shen, A., Duc, E., Golshani, B., Stap, D., Uzhou, M., Zhidkovskaya, A. B., Lewark, L., Vincze, M., Wehr, D., Tang, C., Hossain, Z., Phillips, S., Muzhen, J., Ekström, F., Hammon, A., Patel, O., Remy, N., Farhidi, F., Medley, G., Mohammadzadeh, F., Peñaflor, M., Kassahun, H., Friedrich, A., Sparrow, C., Sakal, T., Dhamane, O., Mirabadi, A. K., Hallman, E., Battaglia, M., Maghsoudimehrabani, M., Hoang, H., Amit, A., Hulbert, D., Pereira, R., Weber, S., Mensah, S., Andre, N., Peristyy, A., Harjadi, C., Gupta, H., Malina, S., Albanie, S., Cai, W., Mehkary, M., Reidegeld, F., Dick, A.-K., Friday, C., Sidhu, J., Kim, W., Costa, M., Gurdogan, H., Weber, B., Kumar, H., Jiang, T., Agarwal, A., Ceconello, C., Vaz, W. S., Zhuang, C., Park, H., Tawfeek, A. R., Aggarwal, D., Kirchhof, M., Dai, L., Kim, E., Ferret, J., Wang, Y., Yan, M., Burdzy, K., Zhang, L., Franca, A., Pham, D. T., Loh, K. Y., Robinson, J., Gul, S., Chhablani, G., Du, Z., Cosma, A., White, C., Riblet, R., Saxena, P., Votava, J., Vinnikov, V., Delaney, E., Halasyamani, S., Shahid, S. M., Mourrat, J.-C., Vetoshkin, L., Bacho, R., Ginis, V., Maksapetyan, A., de la Rosa, F., Li, X., Malod, G., Lang, L., Laurendeau, J., Adesanya, F., Portier, J., Hollom, L., Souza, V., Zhou, Y. A., Yalın, Y., Obikoya, G. D., Arnaboldi, L., Rai, Bigi, F., Bacho, K., Clavier, P., Recchia, G., Popescu, M., Shulga, N., Tanwie, N. M., Lux, T. C. H., Rank, B., Ni, C., Yakimchyk, A., Huanxu, Liu, Häggström, O., Verkama, E., Narayan, H., Gundlach, H., Brito-Santana, L., Amaro, B., Vajipey, V., Grover, R., Fan, Y., e Silva, G. P. R., Xin, L., Kratish, Y., Łucki, J., Li, W.-D., Xu, J., Scaria, K. J., Vargus, F., Habibi, F., Long, Lian, Rodolà, E., Robins, J., Cheng, V., Grabb, D., Bosio, I., Fruhauff, T., Akov, I., Lo, E. J. Y., Qi, H., Jiang, X., Segev, B., Fan, J., Martinson, S., Wang, E. Y., Hausknecht, K., Brenner, M. P., Mao, M., Jiang, Y., Zhang, X., Avagian, D., Scipio, E. J., Siddiqi, M. R., Ragoler, A., Tan, J., Patil, D., Plecnik, R., Kirtland, A., Montecillo, R. G., Durand, S., Bodur, O. F., Adoul, Z., Zekry, M., Douville, G., Karakoc, A., Santos, T. C. B., Shamseldeen, S., Karim, L., Liakhovitskaia, A., Resman, N., Farina, N., Gonzalez, J. C., Maayan, G., Hoback, S., Pena, R. D. O., Sherman, G., Mariji, H., Pouriamanesh, R., Wu, W., Demir, G., Mendoza, S., Alarab, I., Cole, J., Ferreira, D., Johnson, B., Milliron, H., Safdari, M., Dai, L., Arthornthurasuk, S., Pronin, A., Fan, J., Ramirez-Trinidad, A., Cartwright, A., Pottmaier, D., Taheri, O., Outevsky, D., Stepanic, S., Perry, S., Askew, L., Rodríguez, R. A. H., Dendane, A., Ali, S., Lorena, R., Iyer, K., Salauddin, S. M., Islam, M., Gonzalez, J., Ducey, J., Campbell, R., Somrak, M., Mavroudis, V., Vergo, E., Qin, J., Borbás, B., Chu, E., Lindsey, J., Radhakrishnan, A., Jallon, A., McInnis, I. M. J., Hoover, A., Möller, S., Bian, S., Lai, J., Patwardhan, T., Yue, S., Wang, A., and Hendrycks, D. Humanity's last exam, 2025. URL https://arxiv.org/abs/2501.14249.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. Gpqa: A graduate-level google-proof qa benchmark, 2023. URL https://arxiv.org/abs/2311.12022.

Schaeffer, R., Miranda, B., and Koyejo, S. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E. H., Schärli, N., and Zhou, D. Large language models can be easily distracted by irrelevant context. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the International Conference on Machine Learning, ICML 2023*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227, Honolulu, Hawaii, USA, July 23-29 2023. PMLR. URL https://proceedings.mlr.press/v202/shi23a.html.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., Xiang, A., Parrish, A., Nie, A., Hussain, A., Askell, A., Dsouza, A., Slone, A., Rahane, A., Iyer, A. S., Andreassen, A. J., Madotto, A., Santilli, A., Stuhlmüller, A., Dai, A. M., La, A., Lampinen, A. K., Zou, A., Jiang, A., Chen, A., Vuong, A., Gupta, A., Gottardi, A., Norelli, A., Venkatesh, A., Gholamidavoodi, A., Tabassum, A., Menezes, A., Kirubarajan, A., Mullokandov, A., Sabharwal, A., Herrick, A., Efrat, A., Erdem, A., Karakaş, A., Roberts, B. R., Loe, B. S., Zoph, B., Bojanowski, B., Özyurt, B., Hedayatnia, B., Neyshabur, B., Inden, B., Stein, B., Ekmekci, B., Lin, B. Y., Howald, B., Orinion, B., Diao, C., Dour, C., Stinson, C., Argueta, C., Ferri, C., Singh, C., Rathkopf, C., Meng, C., Baral, C., Wu, C., Callison-Burch, C., Waites, C., Voigt, C., Manning, C. D., Potts, C., Ramirez, C., Rivera, C. E., Siro, C., Raffel, C., Ashcraft, C., Garbacea, C., Sileo, D., Garrette, D., Hendrycks, D., Kilman, D., Roth, D., Freeman, C. D., Khashabi, D., Levy, D., González, D. M., Perszyk, D., Hernandez, D., Chen, D., Ippolito, D., Gilboa, D., Dohan, D., Drakard, D., Jurgens, D., Datta, D., Ganguli, D., Emelin, D., Kleyko, D., Yuret, D., Chen, D., Tam, D., Hupkes, D., Misra, D., Buzan, D., Mollo, D. C., Yang, D., Lee, D.-H., Schrader, D., Shutova, E., Cubuk, E. D., Segal, E., Hagerman, E., Barnes, E., Donoway, E., Pavlick, E., Rodolà, E., Lam, E., Chu, E., Tang, E., Erdem, E., Chang, E., Chi, E. A., Dyer, E., Jerzak, E., Kim, E., Manyasi, E. E., Zheltonozhskii, E., Xia, F., Siar, F., Martínez-Plumed, F., Happé, F., Chollet, F., Rong, F., Mishra, G., Winata, G. I., de Melo, G., Kruszewski, G., Parascandolo, G., Mariani, G., Wang, G. X., Jaimovitch-Lopez, G., Betz, G., Gur-Ari, G., Galijasevic, H., Kim, H., Rashkin, H., Hajishirzi, H., Mehta, H., Bogar, H., Shevlin, H. F. A., Schuetze, H., Yakura, H., Zhang, H., Wong, H. M., Ng, I., Noble, I., Jumelet, J., Geissinger, J., Kernion, J., Hilton, J., Lee, J., Fisac, J. F., Simon, J. B., Koppel, J., Zheng, J., Zou, J., Kocon, J., Thompson, J., Wingfield, J., Kaplan, J., Radom, J., Sohl-Dickstein, J., Phang, J., Wei, J., Yosinski, J., Novikova, J., Bosscher, J., Marsh, J., Kim, J., Taal, J., Engel, J., Alabi, J., Xu, J., Song, J., Tang, J., Waweru, J., Burden, J., Miller, J., Balis, J. U., Batchelder, J., Berant, J., Frohberg, J., Rozen, J., Hernandez-Orallo, J., Boudeman, J., Guerr, J., Jones, J., Tenenbaum, J. B., Rule, J. S., Chua, J., Kanclerz, K., Livescu, K., Krauth, K., Gopalakrishnan, K., Ignatyeva, K., Markert, K., Dhole, K., Gimpel, K., Omondi, K., Mathewson, K. W., Chiafullo, K., Shkaruta, K., Shridhar, K., McDonell, K., Richardson, K., Reynolds, L., Gao, L., Zhang, L., Dugan, L., Qin, L., Contreras-Ochando, L., Morency, L.-P., Moschella, L., Lam, L., Noble, L., Schmidt, L., He, L., Oliveros-Colón, L., Metz, L., Senel, L. K., Bosma, M., Sap, M., Hoeve, M. T., Farooqi, M., Faruqui, M., Mazeika, M., Baturan, M., Marelli, M., Maru, M., Ramirez-Quintana, M. J., Tolkiehn, M., Giulianelli, M., Lewis, M., Potthast, M., Leavitt, M. L., Hagen, M., Schubert, M., Baitemirova, M. O., Arnaud, M., McElrath, M., Yee, M. A., Cohen, M., Gu, M., Ivanitskiy, M., Starritt, M., Strube, M., Swedrowski, M., Bevilacqua, M., Yasunaga, M., Kale, M., Cain, M., Xu, M., Suzgun, M., Walker, M., Tiwari, M., Bansal, M., Aminnaseri, M., Geva, M., Gheini, M., T, M. V., Peng, N., Chi, N. A., Lee, N., Krakover, N. G.-A., Cameron, N., Roberts, N., Doiron, N., Martinez, N., Nangia, N., Deckers, N., Muennighoff, N., Keskar, N. S., Iyer, N. S., Constant, N., Fiedel, N., Wen, N., Zhang, O., Agha, O., Elbaghdadi, O., Levy, O., Evans, O., Casares, P. A. M., Doshi, P., Fung, P., Liang, P. P., Vicol, P., Alipoormolabashi, P., Liao, P., Liang, P., Chang, P. W., Eckersley, P., Htut, P. M., Hwang, P., Miłkowski, P., Patil, P., Pezeshkpour, P., Oli, P., Mei, Q., Lyu, Q., Chen, Q., Banjade, R., Rudolph, R. E., Gabriel, R., Habacker, R., Risco, R., Millière, R., Garg, R., Barnes, R., Saurous, R. A., Arakawa, R., Raymaekers, R., Frank, R., Sikand, R., Novak, R., Sitelew, R., Bras, R. L., Liu, R., Jacobs, R., Zhang, R., Salakhutdinov, R., Chi, R. A., Lee, S. R., Stovall, R., Teehan, R., Yang, R., Singh, S., Mohammad, S. M., Anand, S., Dillavou, S., Shleifer, S., Wiseman, S., Gruetter, S., Bowman, S. R., Schoenholz, S. S., Han, S., Kwatra, S., Rous, S. A., Ghazarian, S., Ghosh, S., Casey, S., Bischoff, S., Gehrmann, S., Schuster, S., Sadeghi, S., Hamdan, S., Zhou, S., Srivastava, S., Shi, S., Singh, S., Asaadi, S., Gu, S. S., Pachchigar, S., Toshniwal, S., Upadhyay, S., Debnath, S. S., Shakeri, S., Thormeyer, S., Melzi, S., Reddy, S., Makini, S. P., Lee, S.-H., Torene, S., Hatwar, S., Dehaene, S., Divic, S., Ermon, S., Biderman, S., Lin, S., Prasad, S., Piantadosi, S., Shieber, S., Misherghi, S., Kiritchenko, S., Mishra, S., Linzen, T., Schuster, T., Li, T., Yu, T., Ali, T., Hashimoto, T., Wu, T.-L., Desbordes, T., Rothschild, T., Phan, T., Wang, T., Nkinyili, T., Schick, T., Kornev, T., Tunduny, T., Gerstenberg, T., Chang, T., Neeraj, T., Khot, T., Shultz, T., Shaham, U., Misra, V., Demberg, V., Nyamai, V., Raunak, V., Ramasesh, V. V., vinay uday prabhu, Padmakumar, V., Srikumar, V., Fedus, W., Saunders, W., Zhang, W., Vossen, W., Ren, X., Tong, X., Zhao, X., Wu, X., Shen, X., Yaghoobzadeh, Y., Lakretz, Y., Song, Y., Bahri, Y., Choi, Y., Yang, Y., Hao, S., Chen, Y., Belinkov, Y., Hou, Y., Hou, Y., Bai, Y., Seid, Z., Zhao, Z., Wang, Z., Wang, Z. J., Wang, Z., and Wu, Z. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=uyTL5Bvosj.

The State Bar of California. California bar examination. https://www.calbar.ca.gov/Admissions/Examinations/California-Bar-Examination, 2024.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T. B., Vinyals, O., Liang, P., Dean, J., and Fedus, L. Emergent abilities of large language models. *Transactions on Machine Learning Research (TMLR)*, 2022a. URL https://openreview.net/forum?id=yzkSU5zdwD.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, b., Xia, F., Chi, E., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022b.

Previous volumes in this series

1244 February 2025	Artificial intelligence and relationship lending	Leonardo Gambacorta, Fabiana Sabatini and Stefano Schiaffi
1243 February 2025	Mutual funds and climate news	Giulio Cornelli, Leonardo Gambacorta, Tommaso Oliviero and Koji Takahashi
1242 January 2025	Privacy-enhancing technologies for digital payments: mapping the landscape	Raphael Auer, Rainer Böhme, Jeremy Clark and Didem Demirag
1241 January 2025	Estimating non-linear heterogeneous agent models with neural networks	Hanno Kase, Leonardo Melosi and Matthias Rottner
1240 January 2025	The granular origins of inflation	Santiago Alvarez-Blaser, Raphael Auer, Sarah M. Lein and Andrei A. Levchenko
1239 January 2025	The use and disuse of FinTech credit: When buy-now-pay-later meets credit reporting	Yanfei Dong, Jiayin Hu, Yiping Huang, Han Qiu and Yingguang Zhang
1238 December 2024	Fiscal stimulus plans and households' expectations	Fiorella De Fiore, Marco Jacopo Lombardi and Albert Pierres Tejada
1237 December 2024	The macroeconomics of green transitions	Gregor Boehl, Flora Budianto and Előd Takáts
1236 December 2024	Savings-and-credit contracts	Bernardus van Doornik, Armando Gomes, David Schoenherr and Janis Skrastins
1235 December 2024	Aggregate debt servicing and the limit on private credit	Mathias Drehmann, Mikael Juselius and Sarah Quincy
1234 December 2024	Targeted Inflation Targeting: Some Evidence and Theory	Boris Hofmann, Cristina Manea and Benoît Mojon
1233 December 2024	Fire Sales of Safe Assets	Gabor Pinter, Emil Siriwardane and Danny Walker
1232 December 2024	Bond supply, yield drifts, and liquidity provision before macroeconomic announcements	Dong Lou, Gabor Pinter, Semih Uslu and Danny Walker
1231 December 2024	Central bank communication and media coverage	Fiorella De Fiore, Alexis Maurin, Andrej Mijakovic and Damiano Sandri

All volumes are available on our website www.bis.org.