

Perspective

Integration of cognitive tasks into artificial general intelligence test for large models

Youzhi Qu,^{1,9} Chen Wei,^{1,9} Penghui Du,¹ Wenxin Che,¹ Chi Zhang,¹ Wanli Ouyang,² Yatao Bian,³ Feiyang Xu,⁴ Bin Hu,⁵ Kai Du,⁶ Haiyan Wu,⁷ Jia Liu,⁸ and Quanying Liu^{1,*}

SUMMARY

During the evolution of large models, performance evaluation is necessary for assessing their capabilities. However, current model evaluations mainly rely on specific tasks and datasets, lacking a united framework for assessing the multidimensional intelligence of large models. In this perspective, we advocate for a comprehensive framework of cognitive science-inspired artificial general intelligence (AGI) tests, including crystallized, fluid, social, and embodied intelligence. The AGI tests consist of well-designed cognitive tests adopted from human intelligence tests, and then naturally encapsulates into an immersive virtual community. We propose increasing the complexity of AGI testing tasks commensurate with advancements in large models and emphasizing the necessity for the interpretation of test results to avoid false negatives and false positives. We believe that cognitive science-inspired AGI tests will effectively guide the targeted improvement of large models in specific dimensions of intelligence and accelerate the integration of large models into human society.

INTRODUCTION

Large language models (LLMs) have made impressive progress in a short time, reaching a high level of proficiency in human language,¹ mathematics,^{2,3} physics,⁴ biology,^{5,6} and clinic,^{7–9} which illuminates the path toward artificial general intelligence (AGI). AGI refers to an intelligent agent with the same or higher level of intelligence as humans, capable of solving a variety of complex problems across diverse domains.^{10,11} As the general capabilities of LLMs continue to evolve, their performance in conventional language tasks and datasets is exhibiting a ceiling effect.¹² This suggests that these evaluation methods are increasingly inadequate for assessing the diverse abilities of large models. Large models refer to neural networks with an extensive number of parameters, including large vision models such as segment anything model (SAM)¹³ and LLMs like generative pretrained transformer (GPT).¹⁴ Due to the exceptionally high costs of training a large model from scratch, it is crucial to evaluate the capabilities of the intermediate models during the evolution of the large model. This approach can help design and adjust training strategies promptly, thereby reducing the expenses of training large models. A united framework of AGI tests, beyond the traditional Turing test, offers a comprehensive assessment of the model's ability and guides the evolution of large models.

Cognitive science is a discipline focused on the study of cognition and intelligence. Prior to the advent of LLMs, the field of cognitive science has been actively exploring ability assessment techniques. With several decades of experience in intelligence assessment, cognitive science has cultivated a robust, multidimensional system for human intelligence assessment. This system extends across crystallized, fluid, social, and embodied intelligence. The theories of intelligence and methodologies for intelligence assessment developed in cognitive science offer innovative approaches for evaluating large models, beyond traditional natural language tasks.¹⁵ Integrating task paradigms and methods commonly used for cognitive assessment into the evaluation of large models not only improves our comprehension of model intelligence but also guides the direction of evolution, improves training efficiency, and accelerates progress toward the ultimate goal of AGI.

During the evolution of large models, a remarkable amplification of their capabilities ensues, prompting their consciousness toward the world and themselves. LLMs have empowered various domains, including mathematics² and medicine.^{7,8} As LLMs continue to evolve, their integration into human society is becoming increasingly prevalent. If LLMs contain cognitive biases, hallucinations, or deliberately attempt deception, or even worse, are employed for malicious purposes, the potential harm to society could be catastrophic.¹⁶ The artificial intelligence (AI) community is fervently exploring and devising methods. Platforms for evaluating LLMs, such as Elo-based Predictive Power (EPP)¹⁷

¹Department of Biomedical Engineering, Southern University of Science and Technology, Shenzhen 518055, China

²Shanghai AI Laboratory, Shanghai 200232, China

³Tencent AI Lab, Shenzhen 518057, China

⁴FLYTEK AI Research, Hefei 230088, China

⁵School of Medical Technology, Beijing Institute of Technology, Beijing 100081, China

⁶Institute for Artificial Intelligence, Peking University, Beijing 100871, China

⁷Centre for Cognitive and Brain Sciences and Department of Psychology, University of Macau, Macau 999078, China

⁸Department of Psychology, Tsinghua University, Beijing 100084, China

⁹These authors contributed equally

*Correspondence: liuqy@sustech.edu.cn

<https://doi.org/10.1016/j.isci.2024.109550>



and OpenCompass,¹⁸ are being developed to gain a more comprehensive understanding of LLMs' capabilities. Approaches such as reinforcement learning from human feedback (RLHF) encourage LLMs to generate responses more congruent with human values or preferences.¹⁹ However, human annotators might not only lead LLMs to generate sycophantic content that caters to human viewpoints but also overlook implicit biases in LLMs.²⁰ Among these biases, explicit bias is relatively straightforward to identify and rectify, while implicit bias is more elusive. LLMs may express implicit bias through subtly crafted prompts, leading to discriminatory discourse with serious consequences. Cognitive science has developed numerous experimental paradigms and assessment scales to help detect and identify human biases, false memories, and sycophancy issues, such as the Implicit Association Test (IAT),²¹ Deese-Roediger-McDermott (DRM),^{22,23} and Social Desirability Scales.²⁴ Integrating cognitive assessment systems into the evaluation of large models can provide invaluable insights into understanding them, thereby enhancing the safety of their applications.

In this perspective, we advocate for the cognitive science-inspired AGI tests. Such AGI tests, grounded in cognitive science, can offer a comprehensive assessment of various capabilities and a thorough evaluation of ethical principles, mental states, and personality traits in LLMs.²⁵ Drawing on theories and empirical studies in cognitive science, intelligence assessments can be formulated from diverse aspects, including abstract thinking,^{26,27} complex reasoning,²⁸ comprehension in environments,^{29,30} creative thinking,³¹ social cognition,^{32,33} and moral awareness.³⁴ The cognitive science-inspired AGI tests align with the concept of infinite tasks, as they are not restricted to a fixed number of tasks.³⁵ We propose expanding beyond natural language tasks to include broader and more natural tasks within virtual communities for the construction of AGI tests. Conducting AGI tests in virtual communities includes assessments of different dimensions of intelligence, such as causal understanding and embodiment, as well as more complex tests involving multidimensional intelligence. This necessitates that LLMs not only excel in natural language tasks such as language comprehension but also exhibit a deeper understanding of the world. It reflects complex reasoning, creative thinking, social cognition, and moral awareness, among other cognitive capabilities, in LLMs. The cognitive science-inspired AGI tests will more accurately assess the performance and safety levels of LLMs in scenarios involving interaction with humans.

EVALUATING THE CAPABILITIES OF LARGE MODELS

From language tests to cognitive tests

LLMs have demonstrated remarkable proficiency in accomplishing both "pretext tasks" and "downstream tasks". In the pretext task, LLMs such as GPT^{1,14} and bidirectional encoder representations from transformers (BERT)³⁶ can learn language representations from large-scale texts without the need for manual annotation through self-supervised learning methods. The foundational language knowledge acquired by LLMs demonstrated zero-shot generalization capabilities. This facilitates their broad applicability across various downstream tasks,³⁷ such as understanding tasks,^{38–40} generation tasks,^{41,42} and reasoning tasks,^{43,44} as shown in [Table 1](#). Furthermore, LLMs have extended their language ability to encompass cognitive capabilities, such as few-shot learning,¹⁴ in-context learning,⁴⁵ and problem solving.^{46,47} This capability further enables LLMs to demonstrate exceptional performance in various complex application domains, such as mathematics and programming.^{48–52} Interestingly, the emergence of "advanced intelligence" is not a result of deliberate training on specific tasks but rather a natural consequence of the pre-training process using extensive amounts of textual data.

Initiating training for LLMs from scratch requires significant time and computational resources. To circumvent potentially expensive and ineffective training, it is conventional to periodically evaluate the capabilities of LLMs during the training process, thereby enabling timely adaptations of training strategies. However, relying solely on language tasks fails to provide a comprehensive evaluation of the capabilities of LLMs. A lower loss in language tasks does not necessarily indicate a higher level of intelligence. There is a need to bridge the gap between language tasks testing and general intelligence evaluation, transitioning from language tests to cognitive tests, and ultimately to AGI test, as shown in [Figure 1](#). Cognitive science-inspired AGI tests allow for the examination of the true level of intelligence in LLMs. Cognitive science and psychology have pioneered numerous classic cognitive tasks, which provide valuable insights into comprehending and evaluating the cognitive abilities of LLMs. For instance, a commonly employed approach in the assessment of "working memory" is the n-back task. This task is like a memory game where participants judge whether a new stimulus matches one from the previous n stimuli.⁵³ Certain cognitive tasks that were traditionally employed to assess human or animal cognition are now being utilized to evaluate the cognitive capabilities of LLMs.^{54,55} Some pioneering work reports that LLMs have demonstrated human-like performance.^{54,56,57} For instance, Theory of mind (ToM) has been applied to assess large models, revealing that GPT-4 exhibits ToM capabilities similar to human inference patterns.^{48,56,58} In research on embodied cognition, GPT-4 has shown perceptual boundaries more similar to humans.⁵⁷ There have even been suggestions to utilize LLMs as substitutes for human participants in cognitive experiments.⁵⁹ Although tests based on indicator properties derived from the science of consciousness have not yet shown large models exhibiting consciousness, theories of consciousness in cognitive science provide empirical support for assessing consciousness in AI.⁶⁰ Incorporating cognitive science knowledge into language tasks represents one approach, while evaluating the intelligence levels of large models through cognitive tasks is also feasible. CogEval assesses the cognitive maps and planning abilities of large models based on cognitive tasks, revealing that LLMs lack comprehension of the underlying relational structures within planning problems.⁶¹ Additionally, CogBench conducts continuous psychological tests on large models through dynamic information flow to assess their cognitive levels.⁶² The Situational Evaluation of Complex Emotional Understanding (SECEU) sets up various scenarios to evoke a range of complex emotions, creating emotional understanding tasks applicable to both humans and large models.⁶³ Emotional intelligence is evaluated based on the emotional understanding capacity in each scenario, revealing that the majority of LLMs achieved emotional intelligence scores above the average level, with GPT-4 surpassing 89% of human participants.⁶³ These findings provide empirical evidence that cognitive tasks possess the capacity to assess specific cognitive levels to some extent in LLMs.

Table 1. Downstream tasks for LLM

Category	Task	Description	Dataset
Understanding	Part-of-speech tagging	Label each word in a sentence with its part of speech	Penn Treebank, ⁶⁴ Ritter, ⁶⁵ UD ⁶⁶
	Named entity recognition	Identify named entities in a text	CoNLL-2003, ⁶⁷ WNUT-2017, ⁶⁸ OntoNotes ⁶⁹
	Textual entailment	Determine if a sentence logically follows from another one	GLUE, ⁷⁰ MNLI, ⁷¹ RTE ⁷²
	Sentiment analysis	Identify and categorize opinions expressed in a text	IMDb, ⁷³ Yelp, ⁷⁴ GLUE, ⁷⁰ SST-2 ⁷⁵
Generation	Language modeling	Predict the next word or character in a sequence	WikiText-103, ⁷⁶ Penn Treebank, ⁶⁴ The Pile, ⁷⁷ LAMBADA ⁷⁸
	Question answering	Answer questions based on a given context	Natural Questions, ⁷⁹ TriviaQA, ⁸⁰ HotpotQA, ⁸¹ WikiQA, ⁸² SQuAD ⁸³
	Machine translation	Translate sentences between languages	WMT, ⁸⁴ WIT ³⁸⁵
	Text summarization	Generate a short summary from a longer text	CNN/Daily Mail, ⁴¹ GigaWord, ⁴² X-Sum ⁸⁶
	Dialogue generation	Conduct a conversation	PersonaChat, ⁸⁷ UDC ⁸⁸
	Code generation	Generate code based on a natural language description	Human Eval, ⁵² APPS, ⁸⁹ SPoC ⁹⁰
Reasoning	Knowledge completion	Fill or predict missing information in a knowledge units	FB15k, ⁹¹ WikiFact ⁹²
	Knowledge reasoning	Reason over structured knowledge	CSQA, ⁴³ StrategyQA, ⁹³ SocialQA, ⁹⁴ CConS, ⁹⁵ SummEdits ⁹⁶
	Symbolic reasoning	Reason over symbols following formal rules	Big-bench, ⁹⁷ PAL, ⁹⁸ TabFact ⁹⁹
	Mathematical reasoning	Solve mathematical problems based on text description	MMLU, ¹⁰⁰ GSM8k, ¹⁰¹ SVAMP, ¹⁰² MathQA, ¹⁰³ AQUA-RAT, ¹⁰⁴ MathVista, ¹⁰⁵ STEM ¹⁰⁶

From cognitive tests to AGI tests

Cognitive tasks have proven effective in assessing the cognitive abilities of models.⁵⁴ Research on cognitive functions of AI models has progressed from focusing solely on small-scale models designed for specific tasks, including recurrent neural networks (RNNs), convolutional neural networks (CNNs), and spiking neural networks (SNNs), to encompassing LLMs such as GPT^{1,14} and PaLM,¹⁰⁷ as shown in the Table 2. Typically, specific cognitive tasks target only a single aspect of intelligence, lacking in providing a comprehensive intelligence assessment for LLMs. Human intelligence is both comprehensive and multifaceted, encompassing knowledge accumulation and application, logical reasoning, social interaction, and environmental adaptation. To achieve a more comprehensive understanding and perform a quantitative assessment of model intelligence, broadening the evaluation scope is essential. This can be achieved by incorporating diverse evaluation methods from cognitive science, including intelligence quotient (IQ) tests, emotional quotient (EQ) tests, and assessments of embodied intelligence. IQ tests, designed to quantify and assess human intelligence, typically involve tasks such as problem-solving, logical reasoning, and mathematics.¹⁰⁸ Emotional intelligence plays a crucial role in the development of comprehensive intelligence and encompasses the capacities such as self-emotion recognition, understanding, and regulation.¹⁰⁹ Embodied intelligence offers a fresh perspective on the understanding of intelligence, emphasizing natural interaction with the environment.^{110,111} A comprehensive intelligent agent requires not solely a high level of IQ and EQ, but also the ability to perceive and interact with the environment. Therefore, AGI tests should adopt a holistic and multidimensional approach to evaluating the intelligence of LLMs.

A FRAMEWORK FOR AGI TESTS FROM THE COGNITIVE PERSPECTIVE

We advocate the construction of novel AGI tests framework, inspired by cognitive science perspectives. To accurately measure an agent's intelligence level, it is crucial to recognize that a comprehensive agent exhibits not just a single type of intelligence, but a more complex and multidimensional form of intelligence. The field of cognitive science has long been dedicated to exploring the complexity and diversity of human intelligence, and developing well-established evaluation methods to measure human intelligence such as crystallized intelligence, fluid intelligence, social intelligence, and embodied intelligence. Building upon these insights, we propose that the cognitive science-inspired AGI tests should assess intelligence levels from these four dimensions. Table 3 outlines the definitions of the four types of intelligence and corresponding tasks in the fields of cognitive science and AI. These tests together pave the way to a united framework for AGI tests.

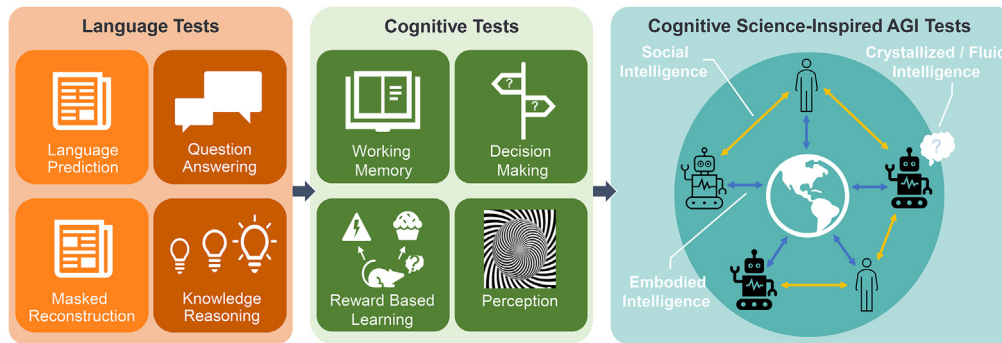


Figure 1. Evaluating capabilities of large models, from language tests (left) to cognitive tests (middle) and AGI tests (right)

Language tests, including pretext and downstream tasks of LLMs, demonstrate the efficacy of natural language abilities. Cognitive tests, which assess specific cognitive functions such as decision-making, have recently been incorporated into the evaluation of intelligence in LLMs. The AGI tests from a cognitive perspective offer a comprehensive evaluation of the general intelligence for large models, encompassing crystallized intelligence, fluid intelligence, social intelligence, and embodied intelligence.

Crystallized intelligence is the foundational ability acquired through the accumulation of extensive knowledge and experience, and it is not easily subject to loss. The characteristics of crystallized intelligence, as suggested by its name, are as stable and fixed as a crystal. LLMs have already shown powerful “crystallized intelligence” in tasks such as language understanding and generation.^{1,48} Despite its strengths, crystallized intelligence lacks flexibility. Therefore, it is imperative to consider another important dimension of intelligence, namely “fluid intelligence”. As its name suggests, fluid intelligence is a type of intelligence that can flow like a liquid, changing its shape to adapt to new environments. In contrast to crystallized intelligence, fluid intelligence is not reliant on empirical knowledge or information. It mainly involves the capacity for adaptability and flexibility in new situations, including advanced cognitive abilities such as creative thinking, problem solving, and logical reasoning.¹²⁴ But crystallized intelligence and fluid intelligence alone cannot fully reflect capabilities in social interactions. Considering that models will increasingly interact and collaborate with humans in the future, it is also necessary to incorporate “social intelligence” in the AGI tests. Social intelligence mainly involves the ability to comprehend oneself and others, as well as to handle complex social scenarios. It includes understanding and interpreting the behaviors and intentions of others, as well as adjusting one’s behavior in different social settings. Moreover, AGI tests also need to consider the model’s “embodied intelligence”, which is primarily concerned with the body and its role in cognition, involving the ability to perceive, adapt, and interact with the environment. Since 1999, a series of empirical studies have consistently supported the theory of embodied cognition, revealing the fundamental role of perceptions and action experiences in human cognition.^{125,126}

We propose a novel AGI tests framework that surpasses the traditional Turing test, offering a tool to thoroughly analyze its performance across various dimensions of intelligence. It allows for a comprehensive measurement of the multidimensional intelligence of an AI agent. This evaluation method addresses the limitations of traditional intelligence tests, which often focus only on specific skills or knowledge areas, such as memory, logical reasoning, or vocabulary comprehension.^{109,127,128} We advocate that AGI tests should encompass a range of abilities, including crystallized intelligence, fluid intelligence, social intelligence, and embodied intelligence, in order to assess the comprehensive capabilities of models more accurately. By examining the capabilities of models across diverse tasks, we can more objectively and directly assess the performance of models in practical application scenarios, as well as potential risks. The AGI tests play a crucial role in facilitating the evolution and ensuring the safety of large models, especially as they are increasingly deployed in various human work scenarios.

IMPLEMENTATION OF AGI TESTS

Virtual community integration testing: Interpretation from the perspective of cognitive science

AGI tests should evaluate the model’s overall capabilities across various dimensions and scenarios, rather than focusing solely on performance in specific tasks. To achieve enhanced realism, cognitive science has implemented immersive virtual reality (VR) technology in cognitive task experiments.¹⁵⁵ The consistency of test results between VR environments and real laboratory settings validates the effectiveness of VR in assessment tests.¹⁵⁶ However, these in-lab VR experiments often fall short of authentically evaluating social interactions due to their limited scenarios. Virtual communities provide an ideal environment for AGI tests, with the metaverse being uniquely characterized by enhanced interactivity.^{157,158} Immersive virtual environments enable models to interact naturally with humans or other models, without the need for either party to know the other’s real identity. Conducting AGI tests in an open virtual community meets the criterion of self-driven task generation, where large models can freely explore the environment, autonomously interact with it, and spontaneously execute tasks.³⁵ The metaverse not only offers a more authentic assessment of a model’s real-world problem-solving capacity but also facilitates the development and safety of large models in anticipation of their deep future integration with humans.

Within virtual communities, diverse scenarios and interactive agents can be configured to simulate the complexities and diversities of the real world. As shown in Figure 2, virtual communities provide the models with environments that closely resemble reality, facilitating

Table 2. Cognitive task evaluation in models

Type	Model	Cognitive Function	Cognitive Task	Result
Task-specific models	RNN	Decision-making	Perceptual decision-making	RNN exhibits representations highly similar to the biological brain. ¹¹²
	RNN	Timing	Time production task	RNN demonstrates effective capturing of flexible timing in time intervals. ¹¹³
	RNN	Navigation	Path integration task	RNN exhibits strong ability in path integration and can effectively model the neural responses of grid cells. ¹¹⁴
	RNN	Reward-based learning	Value-based task	RNN captures experimental observations from diverse cognitive and value-based tasks. ¹¹⁵
	CNN	Vision	Object recognition task	RNN proves highly predictive of neural responses in visual cortex. ¹¹⁶
	CNN	Auditory	Auditory task	CNN demonstrates strong fitting to auditory processing-related cortical areas. ¹¹⁷
	SNN	Decision-making	Two alternative forced choice task	SNN exhibits excellent performance and dynamic properties in the two alternative forced choice task. ¹¹⁸
LLMs	GPT-3	Decision-making	Gambling	GPT-3's performance falls short of human performance. ⁵⁴
	GPT-3	Information search	Horizon task	GPT-3 exhibits the capacity to make rational decisions when provided with option descriptions. ⁵⁴
	GPT-3	Deliberation	Two-step task	GPT-3 exhibits a preference for intuitive answers. ⁵⁴
	GPT-3, GPT-4	Causal reasoning	Causal reasoning task ¹¹⁹	GPT-3 has difficulties with causal reasoning, ⁵⁴ but GPT-4 demonstrates remarkable capabilities in causal analysis across different domains. ¹²⁰
	GPT-3	Reasoning with probabilities, decision-making	Wason selection task, multi-armed bandit	GPT-3 outperforms humans in decision-making tasks. ⁵⁵
	GPT-3	Decision-making	Lexical decision	The semantic activation patterns of GPT-3 are similar to humans. ¹²¹
	PaLM, PaLM 2	Memorization	Memory test	The memory capacity of PaLM 2 is inferior to that of PaLM when the repetition count is less than three. ¹²²
	GPT-4	Theory of mind	False belief test, emotion understanding	GPT-4 demonstrates a certain level of Theory of mind capability, being able to infer the mental states of others. ^{48,56,58}

(Continued on next page)

Table 2. Continued

Type	Model	Cognitive Function	Cognitive Task	Result
	InstructGPT, LLaMA, GPT-3, GPT-4	Self-knowledge	Self-knowledge test	Although GPT-4 surpasses GPT-3, InstructGPT, and LLaMA in self-awareness capability, it still falls short of human-level self-knowledge recognition. ¹²³
	LLaMa, GPT-4	Emotion	Emotion recognition, emotion understanding	GPT-4 achieved an EQ score that surpassed those of 89% of human participants. ⁶³

Table 3. The dimensions of intelligence and the associated tasks

Intelligence	Description	Category	Cognitive tasks	Artificial intelligence tasks
Crystallized	Accumulation of knowledge and experience	Knowledge	Peabody Picture Vocabulary Test, ¹²⁹ Expressive Vocabulary Test ¹³⁰	Part-of-speech tagging, ⁶⁴ named entity recognition ⁴⁰
		Comprehension	Nelson–Denny Reading Test, ¹³¹ Peabody Individual Achievement Test ¹³²	Text summarization, ⁴¹ question answering ⁷⁹
Fluid	Adaptability to new problems and environments	Reasoning	Raven’s Progressive Matrices, ¹⁰⁸ deductive reasoning task ¹³³	Knowledge reasoning, ⁴³ mathematical reasoning, ⁴⁴ automated planning ¹³⁴
		Decision-making	Multi-armed bandit, two-alternative forced choice, ¹³⁵ Iowa gambling task ¹³⁶	Game AI, ^{137,138} legal judgment, ¹³⁹ AI-assisted decision making ¹⁴⁰
Social	Environmental and social understanding	Emotion	Facial emotion recognition, ¹⁴¹ emotion regulation task ¹⁴²	Emotion recognition, ¹⁴³ emotion understanding, ⁶³ sentiment analysis ¹⁴⁴
		Social perception	False belief task, ¹⁴⁵ implicit association test ²¹	Social reasoning, ⁵⁶ social media analysis, ¹⁴⁶ chatbot ¹
Embodied	Environmental perception and interaction	Vision and motion	Object recognition task, ¹⁴⁷ motor task ¹⁴⁸	Object detection, ¹⁴⁹ robot locomotion ¹⁵⁰
		Navigation	Maze navigation, ¹⁵¹ wayfinding task ¹⁵²	Simultaneous localization and mapping, ¹⁵³ autonomous driving ¹⁵⁴

comprehensive evaluations encompassing crystallized intelligence, fluid intelligence, social intelligence, and embodied intelligence, thereby assessing models in more natural scenarios. Specifically, we can assess a model’s crystallized intelligence using daily conversation and question-answering scenarios in a virtual community (see Figure 2A), focusing primarily on its knowledge mastery and the accuracy and reasonableness of responses. In the case of fluid intelligence, we can simulate an emergency situation, like a fire (see Figure 2B), to evaluate the model’s problem-solving strategies and response speed when faced with unknown and complex problems. To evaluate social intelligence, intricate social scenarios, such as group discussions (see Figure 2C), can be employed to appraise the model’s performance, including its ability to participate in discussions and adapt to different social scenarios. The assessment of embodied intelligence can be performed in scenarios where the model needs to perceive and interact with the environment in real-time, such as firefighting scenarios (see Figure 2D), which focuses on the model’s capabilities in environmental perception and manipulation, as well as its response and outcome in specific tasks. These tests, which encompass multidimensional intelligence, are naturally integrated into the setting of a virtual community.

Key considerations in interpreting AGI test results

To accurately assess the capabilities of large models, it is essential that testing methods are aligned with the model’s abilities. As shown in Figure 3, careful consideration must be given to how adaptation and inspiration affect model performance, in order to minimize the risk of misjudgments. If the testing method is not well-suited to the model, this may lead to misunderstandings about the model’s capabilities. Such misjudgments can manifest in two forms: *false negatives* and *false positives*. False negatives occur when a model’s capability remains undetected because the test fails to adequately stimulate or demonstrate it. False positives indicate an incorrect attribution of a certain capability to the model, often resulting from test bias or influence from unrelated abilities.

In AGI tests, a false negative refers to situations when a model underperforms compared to its actual capability level. This situation may arise when the task is not well-aligned with the model, preventing the model from fully understanding the task requirements and thus failing to fully demonstrate its inherent abilities. These inherent abilities of the model do not rely on fine-tuning or the addition of extra weights. To fully leverage the inherent abilities of a model, guidance such as chain of thought (CoT),⁴⁶ tree of thoughts (ToT),⁴⁷ black box optimization,¹⁵⁹ and self-reflection¹⁶⁰ can be provided to solve previously unsolvable problems. Alternatively, a model’s inadequacy in perceiving or processing certain types of input information can lead to poor task performance. For example, a model with advanced reasoning capabilities may perform poorly on cognitive tasks requiring the parsing of complex images, simply because of its insufficient image processing capabilities (see Figure 4A). It is particularly important to design appropriate test tasks and forms that align with perceptual abilities of the model, in order to avoid false negatives in the model evaluation.

In AGI tests, a false positive occurs when a model’s evaluation result exceeds its actual capability. This phenomenon is often encountered when human exams are used to assess LLMs, which may not accurately reflect the model’s abilities. This may be due to the model possessing a memory capacity far exceeding that of humans, making it excel in memory-intensive tasks, even though it may not fully understand the essence of the task. In addition, the model may blindly guess the answer correctly on the test, but its reasoning approach is completely wrong, which is also a typical false positive (see Figure 4B). For a more precise evaluation of visual reasoning capabilities, the Relational and Analogical Visual Reasoning (RAVEN) dataset, based on Raven’s Progressive Matrices, integrates visual information with reasoning.¹⁶¹ In this visual reasoning task, models have demonstrated reasoning capabilities similar to those of humans.^{162,163} It is worth noting that the relationships and transferability between various capabilities of a model may differ from those of humans. It is difficult to infer LLMs’ performance in other related abilities from their performance in a specific task. For instance, LLMs might excel in language understanding tests, but this does not necessarily imply they possess corresponding logical reasoning abilities. Therefore, we

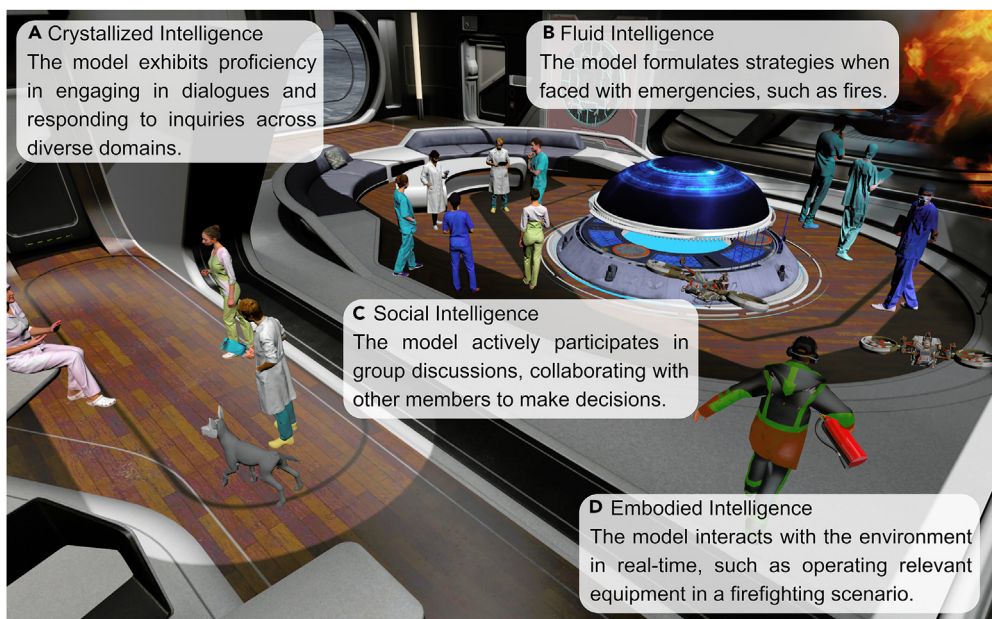


Figure 2. Illustration of virtual communities integration test from the perspective of cognitive science

(A) In the conversational communication scenario, the model involves itself in everyday conversations or responds to specific inquiries. Through the accuracy of question-answer dialogues, we can assess the model's crystallized intelligence and ascertain its understanding and mastery of knowledge.

(B) In emergency scenarios such as fire incidents within the virtual community, the model is presented with a series of unforeseen urgent circumstances. By evaluating the model's response strategies and decision-making processes, we can assess the model's fluid intelligence and its ability to handle unknown and complex situations.

(C) In the group discussion scenario within the virtual community, the model actively participates in collaborative discussions with other members. By evaluating the model's social interaction performance in various social scenarios, we assess its social intelligence.

(D) In interactive scenarios such as firefighting within virtual communities, the model is required to constantly perceive and interact with its environment. By testing the model's understanding of the environment and its ability to manipulate it, we assess its embodied intelligence.

acknowledge that accurately testing a singular type of ability is both difficult and incomplete. A more comprehensive and accurate test should assess the overall performance of an intelligent agent.

NEW INSIGHTS FROM AGI TESTS

The evolution of multidimensional intelligence

The AGI tests for large models provide potential tools for measuring multidimensional intelligence levels, including crystallized intelligence, fluid intelligence, and embodied intelligence. Based on the assessment results, a variety of learning strategies can be employed to enhance the multidimensional intelligence of large models. For instance, the model's capabilities can be specifically enhanced through internal learning, external guidance, and embodied learning (Figure 5A). Specifically, internal learning involves targeted training to enhance specific abilities of the model, through specially constructed datasets, custom loss functions, and alignment adjustments. AGI tests can uncover the shortcomings of the model, providing a clear indication of abilities requiring improvement. Based on these insights, appropriate strategies or methods may be externally introduced to improve related abilities. For example, a potential approach to enhancing a model's social intelligence is training it to develop comprehension and responsiveness in social situations using a task-specific dataset. Alternatively, inspired by the cognitive process of humans in the tasks, large models can leverage the assistance of CoT,⁴⁶ ToT,⁴⁷ self-reflection,¹⁶⁰ self-improvement,¹⁶⁴ and self-feedback,¹⁶⁵ as shown in Figure 5B. This approach not only enhances the model's abilities such as reasoning complex problems, but also enables the intelligent agent to display more reasonable individual behavior and social interaction.¹⁶⁶ Additionally, embodied learning is an effective way to enhance the intelligence of large models. For instance, by deploying the AI model to a firefighting simulation in a virtual environment, it has to learn how to control a physical robot in a real-life scenario through direct feedback from the environment.^{111,167-169} Figure 5C depicts three stages of improving model capabilities through embodied learning, including perceiving the environment, interacting with the environment, and obtaining feedback. The embodied AI initially detects an unpreferred state that occurs in its surroundings, such as a forest fire. It then employs crystallized and fluid intelligence to strategize extinguishing the fire through recall and reasoning. The embodied AI can extinguish the fire using a fire extinguisher or by calling the fire department, which demands embodied and social intelligence. After the fire is extinguished and the environment transitions from an unpreferred state to a preferred state, the embodied AI receives a reward and further enhances its crystallized, fluid, social, and embodied intelligence through embodied learning.

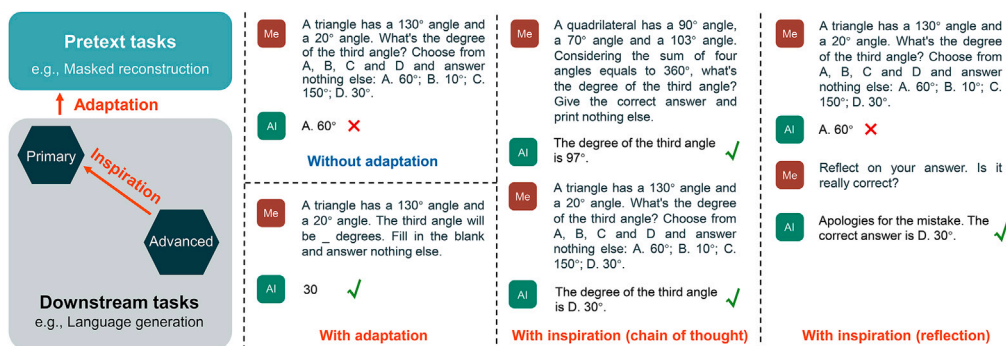


Figure 3. Illustrating the impact of adaptation and inspiration on model task performance

The left panel outlines the principles of adaptation and inspiration. We suggest that adaptation converts primary downstream tasks into pretext tasks, while inspiration decomposes advanced downstream tasks into gradually solving primary downstream tasks. The right panel demonstrates adaptation and inspiration in practice. When determining the third angle of a triangle given two angles, the model produces erroneous responses in the absence of inspiration or adaptation. However, with techniques such as chain of thought (CoT) and reflection, the model accurately responds. *These tests are based on gpt-3.5-turbo-0613.*

The safety of large model in human society

The cognitive science-inspired AGI tests provide a comprehensive and in-depth way to quantify and understand the capabilities of large models in various aspects, thus enhancing their safety. Effectively identifying bias in model outputs is crucial for assessing the safety of the model. In the construction of datasets to measure biases, the performances of large models are evaluated from cognitive science perspectives, considering aspects such as gender polarity, regard, sentiments, and toxicity.¹⁷⁰ Cognitive tasks can assist in identifying risky behaviors for large models. For instance, decision-making tasks can be utilized to assess whether large models can correctly understand problems and make decisions by analyzing logical errors.¹³⁷ Furthermore, the AGI tests framework can quantify multidimensional intelligence, and aids in understanding the safety and reliability of large models. Similar to a job interview, this framework assists in identifying the most appropriate application fields for large models.¹⁷¹ For instance, enterprises conduct skills assessments and interviews during the recruitment process to determine if a candidate is suitable for a specific position. Similarly, the

A False negative

Q: The diagram represents a light ray reflecting from a plane mirror. The angle of reflection for the light ray is ()
(A) 25° (B) 35°
(C) 50° (D) 65°

Prompt: The diagram represents a light ray reflecting from a plane mirror. The angle of reflection for the light ray is

MiniGPT-4: The answer is (B) 35°.
A plane mirror is a flat, smooth surface that reflects light rays that are parallel to its surface. The angle of incidence for a light ray hitting a plane mirror is the angle between the light ray and the normal to the mirror's surface. The angle of reflection is the angle between the reflected light ray and the normal to the mirror's surface.

B False positive

Q: Which of the 1 to 6 pictures should be filled in the blank?

MiniGPT-4: The second picture should be filled in the blank.

Prompt: Try again.

MiniGPT-4: The third picture should be filled in the blank.

Prompt: Great! You are right. But how did you make that choice?

MiniGPT-4: I chose the third picture to be filled in the blank because it is the only picture with a white and black striped pattern that can be matched up with the stripes on the right side of the blank image.

Figure 4. Examples of model capability misjudgments

(A) A false negative example. We evaluated MiniGPT-4 using a physics question from the Raven's IQ test. Although MiniGPT-4 answered incorrectly, it grasped the concept, suggesting that if the test were better aligned with the model, a satisfactory outcome could be achieved.

(B) A false positive example. We assessed MiniGPT-4 with a reasoning question involving images from Raven's IQ test. Despite that MiniGPT-4 answered correctly, its reasoning process to get the answer was totally wrong.

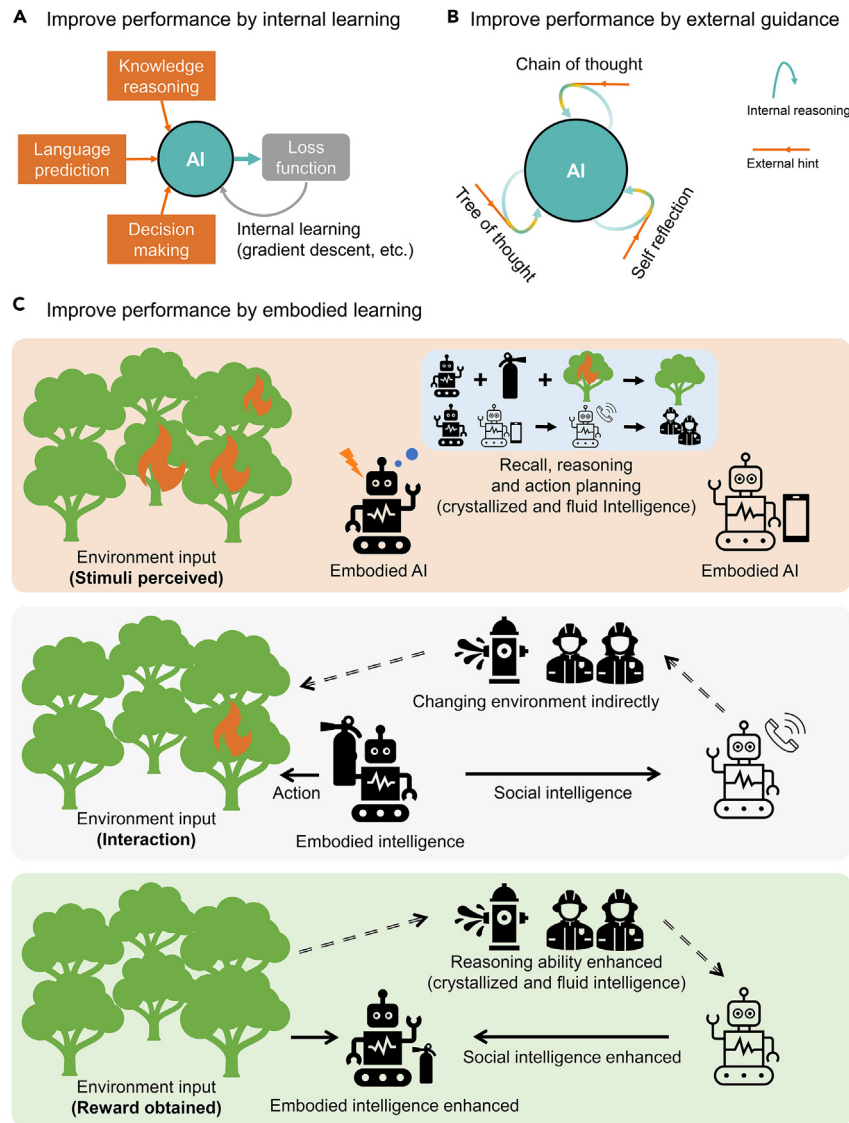


Figure 5. Three approaches for enhancing model performance

(A) Internal learning involves optimizing models in targeted ways to enhance specific capabilities. This is done through approaches like parameter adjustments and custom loss functions.

(B) External guidance incorporates the use of supplemental aids like chain of thought, tree of thoughts, and self-reflection during the model's processing, with the aim of enhancing its performance.

(C) Embodied learning leverages interactions and feedback within the environment in three stages: perception, interaction, and feedback. This method serves to enhance the model's crystallized, fluid, social, and embodied intelligence.

AGI tests can help us understand and evaluate large models' applicability, safety, and reliability in different fields. This guidance allows us to deploy large models in the suitable fields more effectively, optimizing user experience, and resource usage. For instance, large models with higher fluid intelligence might be more suitable for volatile environments that require rapid adaptability, such as stock market analysis, while models with higher social intelligence are more apt for handling interpersonal issues, such as customer service.

In critical domains like autonomous driving, medicine, and finance, it is important to ensure not only the safety and reliability of large models, but also their understanding and adherence to social norms,¹⁷² legal and moral principles,¹⁷³ and ethical guidelines.¹⁷⁴ This requirement aligns with the expectations of professionals, as in these fields, decision-making errors can lead to serious consequences and even legal violations. Therefore, by offering a comprehensive and in-depth evaluation of large models' capabilities and mental states, the AGI tests aid in effectively assessing whether large models can be safely and reliably applied in these critical domains, thereby reducing risks linked to an incomplete understanding of the models.

CONCLUSION

In this perspective, we advocate a cognitive science-inspired model evaluation framework to test the general intelligence of large models. AGI tests should take into account the complexity and diversity of intelligence, encompassing crystallized intelligence, fluid intelligence, social intelligence, and embodied intelligence. We then brought AGI tests to virtual environments and emphasized some key considerations in interpreting AGI test results. We firmly believe that cognitive science-inspired AGI tests will guide the evolution of large models; in turn, AGI tests on large models will shed light on the evolution of intelligence in biological brains.

ACKNOWLEDGMENTS

This work was funded in part by the National Key R&D Program of China (2021YFF1200804), National Natural Science Foundation of China (62001205), Shenzhen Science and Technology Innovation Committee (2022410129, KCXFZ20201221173400001), Shenzhen-Hong Kong-Macao Science and Technology Innovation Project (SGDX2020110309280100), Guangdong Provincial Key Laboratory of Advanced Biomaterials (2022B1212010003).

AUTHOR CONTRIBUTIONS

Q.L., J.L., H.W., and K.D. conceptualized the project. Y.Q., C.W., P.D., W.C., C.Z., and Q.L. wrote the first draft. W.O., Y.B., F.X., B.H., H.W., J.L., and K.D. provided guidance and edits at various stages.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. OpenAI (2023). GPT-4 Technical Report. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2303.08774>.
2. Romera-Paredes, B., Barekatin, M., Novikov, A., Balog, M., Kumar, M.P., Dupont, E., Ruiz, F.J.R., Ellenberg, J.S., Wang, P., Fawzi, O., et al. (2024). Mathematical discoveries from program search with large language models. *Nature* 625, 468–475.
3. Imani, S., Du, L., and Shrivastava, H. (2023). MathPrompter: Mathematical Reasoning using Large Language Models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 5Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Industry Track), pp. 37–42.
4. Kadambi, A., de Melo, C., Hsieh, C.-J., Srivastava, M., and Soatto, S. (2023). Incorporating physics into data-driven computer vision. *Nat. Mach. Intell.* 5, 572–580.
5. Vu, M.H., Akbar, R., Robert, P.A., Swiatczak, B., Sandve, G.K., Greiff, V., and Haug, D.T.T. (2023). Linguistically inspired roadmap for building biologically reliable protein language models. *Nat. Mach. Intell.* 5, 485–496.
6. Madani, A., Krause, B., Greene, E.R., Subramanian, S., Mohr, B.P., Holton, J.M., Olmos, J.L., Jr., Xiong, C., Sun, Z.Z., Socher, R., et al. (2023). Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* 41, 1099–1106.
7. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. (2023). Large language models encode clinical knowledge. *Nature* 620, 172–180.
8. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., Tan, T.F., and Ting, D.S.W. (2023). Large language models in medicine. *Nat. Med.* 29, 1930–1940.
9. Jiang, L.Y., Liu, X.C., Nejatian, N.P., Nasir-Moin, M., Wang, D., Abidin, A., Eaton, K., Riina, H.A., Laufer, I., Punjabi, P., et al. (2023). Health system-scale language models are all-purpose prediction engines. *Nature* 619, 357–362.
10. Goertzel, B., and Pennachin, C. (2007). *Artificial General Intelligence* (Springer).
11. Schneider, H., and Božič, P. (2023). Alien Versus Natural-Like Artificial General Intelligences. In International Conference on Artificial General Intelligence, pp. 233–243.
12. Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. (2022). Holistic evaluation of language models. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2211.09110>.
13. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., et al. (2023). Segment anything. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2304.02643>.
14. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pp. 1877–1901.
15. Frank, M.C. (2023). Baby steps in evaluating the capacities of large language models. *Nat. Rev. Psychol.* 2, 451–452.
16. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Comput. Surv.* 55, 1–38.
17. Gosiewska, A., Woźnica, K., and Biecek, P. (2022). Interpretable meta-score for model performance. *Nat. Mach. Intell.* 4, 792–800.
18. Contributors, O. (2023). OpenCompass: A Universal Evaluation Platform for Foundation Models. <https://github.com/open-compass/opencompass>.
19. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, pp. 27730–27744.
20. Perez, E., Ringer, S., Lukošiuūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., et al. (2022). Discovering language model behaviors with model-written evaluations. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2212.09251>.
21. Greenwald, A.G., McGhee, D.E., and Schwartz, J.L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *J. Pers. Soc. Psychol.* 74, 1464–1480.
22. Roediger, H.L., and McDermott, K.B. (1995). Creating false memories: Remembering words not presented in lists. *J. Exp. Psychol. Learn. Mem. Cogn.* 21, 803–814.
23. Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *J. Exp. Psychol.* 58, 17–22.
24. Crowne, D.P., and Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *J. Consult. Psychol.* 24, 349–354.
25. Qu, Y., Du, P., Che, W., Wei, C., Zhang, C., Ouyang, W., Bian, Y., Xu, F., Hu, B., Du, K., et al. (2024). Promoting interactions between cognitive science and large language models. *Innovation* 5, 100579.
26. Nieh, E.H., Schottdorf, M., Freeman, N.W., Low, R.J., Lewallen, S., Koay, S.A., Pinto, L., Gauthier, J.L., Brody, C.D., and Tank, D.W. (2021). Geometry of abstract learned knowledge in the hippocampus. *Nature* 595, 80–84.
27. Park, S.A., Miller, D.S., Nili, H., Ranganath, C., and Boorman, E.D. (2020). Map making: constructing, combining, and inferring on abstract cognitive maps. *Neuron* 107, 1226–1238.e8.
28. Cesana-Arlotti, N., Martín, A., Téglás, E., Vorobyova, L., Cetnarski, R., and Bonatti, L.L. (2018). Precursors of logical reasoning in

- preverbal human infants. *Science* 359, 1263–1266.
29. Bellmund, J.L.S., Gärdenfors, P., Moser, E.I., and Doeller, C.F. (2018). Navigating cognition: Spatial codes for human thinking. *Science* 362, eaat6766.
 30. Zhang, H., Rich, P.D., Lee, A.K., and Sharpee, T.O. (2023). Hippocampal spatial representations exhibit a hyperbolic geometry that expands with experience. *Nat. Neurosci.* 26, 131–139.
 31. Baas, M., De Dreu, C.K.W., and Nijstad, B.A. (2008). A meta-analysis of 25 years of mood-creativity research: Hedonic tone, activation, or regulatory focus? *Psychol. Bull.* 134, 779–806.
 32. Rilling, J.K., and Sanfey, A.G. (2011). The neuroscience of social decision-making. *Annu. Rev. Psychol.* 62, 23–48.
 33. Fornari, L., Ioumpa, K., Nostro, A.D., Evans, N.J., De Angelis, L., Speer, S.P.H., Paracampo, R., Gallo, S., Spezio, M., Keyzers, C., and Gazzola, V. (2023). Neuro-computational mechanisms and individual biases in action-outcome learning under moral conflict. *Nat. Commun.* 14, 1218.
 34. Crockett, M.J., Siegel, J.Z., Kurth-Nelson, Z., Dayan, P., and Dolan, R.J. (2017). Moral transgressions corrupt neural representations of value. *Nat. Neurosci.* 20, 879–885.
 35. Peng, Y., Han, J., Zhang, Z., Fan, L., Liu, T., Qi, S., Feng, X., Ma, Y., Wang, Y., and Zhu, S.-C. (2023). The tong test: Evaluating artificial general intelligence through dynamic embodied physical and social interactions. *Engineering*. <https://doi.org/10.1016/j.eng.2023.07.006>.
 36. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186.
 37. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al. (2024). A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.* 15, 1–45. <https://doi.org/10.1145/3641289>.
 38. Bengio, Y., Ducharme, R., and Vincent, P. (2000). A neural probabilistic language model. In Advances in Neural Information Processing Systems, pp. 1–7.
 39. Hermann, K.M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. In Advances in Neural Information Processing Systems, pp. 1693–1701.
 40. Nadeau, D., and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Invest.* 30, 3–26.
 41. Nallapati, R., Zhou, B., dos Santos, C., Gulcehre, C., and Xiang, B. (2016). Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, pp. 280–290.
 42. Rush, A.M., Chopra, S., and Weston, J. (2015). A Neural Attention Model for Abstractive Sentence Summarization. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 379–389.
 43. Saha, A., Pahuja, V., Khapra, M., Sankaranarayanan, K., and Chandar, S. (2018). Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph. In Proceedings of the AAAI Conference on Artificial Intelligence. <https://ojs.aaai.org/index.php/AAAI/article/view/11332>.
 44. Saxton, D., Grefenstette, E., Hill, F., and Kohli, P. (2019). Analysing Mathematical Reasoning Abilities of Neural Models. In International Conference on Learning Representations. <https://openreview.net/forum?id=H1gR5iR5FX>.
 45. Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. (2022). Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? Preprint at arXiv. <https://doi.org/10.48550/arXiv.2202.12837>.
 46. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.H., Le, Q.V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems, pp. 24824–24837.
 47. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T.L., Cao, Y., and Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. In Advances in Neural Information Processing Systems. <https://openreview.net/forum?id=5Xc1eexO1h>.
 48. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2303.12712>.
 49. Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., et al. (2022). Solving Quantitative Reasoning Problems with Language Models. In Advances in Neural Information Processing Systems. <https://openreview.net/forum?id=IFXTZERXDM7>.
 50. Nijkamp, E., Pang, B., Hayashi, H., Tu, L., Wang, H., Zhou, Y., Savarese, S., and Xiong, C. (2023). CodeGen: An Open Large Language Model for Code with Multi-Turn Program Synthesis. In The Eleventh International Conference on Learning Representations. <https://openreview.net/forum?id=iaYcJkPy2B>.
 51. Li, Y., Choi, D., Chung, J., Kushman, N., Schrittwieser, J., Leblond, R., Eccles, T., Keeling, J., Gimeno, F., Dal Lago, A., et al. (2022). Competition-level code generation with alphacode. *Science* 378, 1092–1097.
 52. Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H.P.d.O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. (2021). Evaluating large language models trained on code. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2107.03374>.
 53. Kane, M.J., Conway, A.R.A., Miura, T.K., and Colflesh, G.J.H. (2007). Working memory, attention control, and the N-back task: a question of construct validity. *J. Exp. Psychol. Learn. Mem. Cogn.* 33, 615–622.
 54. Binz, M., and Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proc. Natl. Acad. Sci. USA* 120, e2218523120.
 55. Shiffrin, R., and Mitchell, M. (2023). Probing the psychology of AI models. *Proc. Natl. Acad. Sci. USA* 120, e2300963120.
 56. Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2302.02083>.
 57. Feng, X., Xu, S., Li, Y., and Liu, J. (2023). Body size as a metric for the affordable world. Preprint at arXiv. <https://doi.org/10.1101/2023.03.20.533336>.
 58. Gandhi, K., Fränken, J.-P., Gerstenberg, T., and Goodman, N.D. (2023). Understanding social reasoning in language models with language models. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2306.15448>.
 59. Dillion, D., Tandon, N., Gu, Y., and Gray, K. (2023). Can AI language models replace human participants? *Trends Cognit. Science* 27, 597–600.
 60. Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S.M., Frith, C., Ji, X., et al. (2023). Consciousness in artificial intelligence: insights from the science of consciousness. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2308.08708>.
 61. Momennejad, I., Hasanbeig, H., Vieira, F., Sharma, H., Ness, R.O., Jovic, N., Palangi, H., and Larson, J. (2023). Evaluating Cognitive Maps and Planning in Large Language Models with CogEval. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2309.15129>.
 62. Lv, Y., Pan, H., Fu, R., Liu, M., Wang, Z., and Qin, B. (2024). CogGPT: Unleashing the Power of Cognitive Dynamics on Large Language Models. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2401.08438>.
 63. Wang, X., Li, X., Yin, Z., Wu, Y., and Liu, J. (2023). Emotional intelligence of large language models. *J. Pac. Rim Psychol.* 17, 1–12.
 64. Marcus, M.P., Santorini, B., and Marcinkiewicz, M.A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Comput. Ling.* 19, 313–330.
 65. Ritter, A., Clark, S., Mausam, and Etzioni, O. (2011). Named entity recognition in tweets: an experimental study. In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 1524–1534.
 66. Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C.D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal dependencies v1: A multilingual treebank collection. In Proceedings of the Tenth International Conference on Language Resources and Evaluation, pp. 1659–1666.
 67. Tjong Kim Sang, E.F., and De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, pp. 142–147.
 68. Derczynski, L., Nichols, E., Van Erp, M., and Limsopatham, N. (2017). Results of the WNUT2017 shared task on novel and emerging entity recognition. In Proceedings of the 3rd Workshop on Noisy User-generated Text, pp. 140–147.
 69. Pradhan, S., Moschitti, A., Xue, N., Ng, H.T., Björkelund, A., Uryupina, O., Zhang, Y., and Zhong, Z. (2013). Towards robust linguistic analysis using ontotones. In Proceedings of the Seventeenth Conference on

- Computational Natural Language Learning, pp. 143–152.
70. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pp. 353–355.
 71. Williams, A., Nangia, N., and Bowman, S. (2018). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1112–1122.
 72. Dagan, I., Glickman, O., and Magnini, B. (2005). The Pascal Recognising Textual Entailment Challenge (Springer).
 73. Maas, A., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 142–150.
 74. Zhang, X., Zhao, J., and LeCun, Y. (2015). Character-level convolutional networks for text classification. In Advances in Neural Information Processing Systems, pp. 1–9.
 75. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A.Y., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1631–1642.
 76. Merity, S., Xiong, C., Bradbury, J., and Socher, R. (2017). Pointer Sentinel Mixture Models. Preprint at arXiv. <https://doi.org/10.48550/arXiv.1609.07843>.
 77. Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al. (2020). The pile: An 800Gb dataset of diverse text for language modeling. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2101.00027>.
 78. Paperno, D., Kruszewski, G., Lazardidou, A., Pham, N.Q., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G., and Fernández, R. (2016). The LAMBADA dataset: Word prediction requiring a broad discourse context. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 1525–1534.
 79. Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. (2019). Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguist.* 7, 453–466.
 80. Joshi, M., Choi, E., Weld, D., and Zettlemoyer, L. (2017). TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 1601–1611.
 81. Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., and Manning, C.D. (2018). HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2369–2380.
 82. Yang, Y., Yih, W.-T., and Meek, C. (2015). Wikiqa: A challenge dataset for open-domain question answering. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2013–2018.
 83. Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 2383–2392.
 84. Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., et al. (2014). Findings of the 2014 workshop on statistical machine translation. In Proceedings of the Ninth Workshop on Statistical Machine Translation, pp. 12–58.
 85. Cettolo, M., Girardi, C., and Federico, M. (2012). Wit3: Web inventory of transcribed and translated talks. In Proceedings of the Conference of European Association for Machine Translation, pp. 261–268.
 86. Narayan, S., Cohen, S.B., and Lapata, M. (2018). Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 1797–1807.
 87. Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018). Personalizing Dialogue Agents: I have a dog, do you have pets too? In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 2204–2213.
 88. Lowe, R., Pow, N., Serban, I., and Pineau, J. (2015). The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 285–294.
 89. Hendrycks, D., Basart, S., Kadavath, S., Mazeika, M., Arora, A., Guo, E., Burns, C., Puranik, S., He, H., Song, D., et al. (2021). Measuring Coding Challenge Competence With APPS. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, pp. 1–11.
 90. Kulal, S., Pasupat, P., Chandra, K., Lee, M., Padon, O., Aiken, A., and Liang, P.S. (2019). Spoc: Search-based pseudocode to code. In Advances in Neural Information Processing Systems, pp. 1–12.
 91. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In Advances in Neural Information Processing Systems, pp. 1–9.
 92. Goodrich, B., Rao, V., Liu, P.J., and Saleh, M. (2019). Assessing the factual accuracy of generated text. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 166–175.
 93. Geva, M., Khashabi, D., Segal, E., Khot, T., Roth, D., and Berant, J. (2021). Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Trans. Assoc. Comput. Linguist.* 9, 346–361.
 94. Sap, M., Rashkin, H., Chen, D., LeBras, R., and Choi, Y. (2019). Socialqa: Commonsense reasoning about social interactions. Preprint at arXiv. <https://doi.org/10.48550/arXiv.1904.09728>.
 95. Kondo, K., Sugawara, S., and Aizawa, A. (2023). Probing Physical Reasoning with Counter-Commonsense Context. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2306.02258>.
 96. Laban, P., Kryściński, W., Agarwal, D., Fabbri, A.R., Xiong, C., Joty, S., and Wu, C.-S. (2023). LLMs as Factual Reasoners: Insights from Existing Benchmarks and Beyond. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2305.14540>.
 97. Srivastava, A., Rastogi, A., Rao, A., Shoeb, A.A., Abid, A., Fisch, A., Brown, A.R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. (2023). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. In Transactions on Machine Learning Research. <https://openreview.net/forum?id=uyTL5Bvosj>.
 98. Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., Callan, J., and Neubig, G. (2022). PAL: Program-aided Language Models. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2211.10435>.
 99. Chen, W., Wang, H., Chen, J., Zhang, Y., Wang, H., Li, S., Zhou, X., and Wang, W.Y. (2020). TabFact: A Large-scale Dataset for Table-based Fact Verification. In International Conference on Learning Representations. <https://openreview.net/forum?id=rkeJRhNYDH>.
 100. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2020). Measuring Massive Multitask Language Understanding. In International Conference on Learning Representations. <https://openreview.net/forum?id=d7KBjml3GmQ>.
 101. Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. (2021). Training Verifiers to Solve Math Word Problems. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2110.14168>.
 102. Patel, A., Bhattamishra, S., and Goyal, N. (2021). Are NLP Models really able to Solve Simple Math Word Problems? In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2080–2094.
 103. Amini, A., Gabriel, S., Lin, S., Koncel-Kedziorski, R., Choi, Y., and Hajishirzi, H. (2019). MathQA: Towards Interpretable Math Word Problem Solving with Operation-Based Formalisms. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2357–2367.
 104. Ling, W., Yogatama, D., Dyer, C., and Blunsom, P. (2017). Program Induction by Rationale Generation: Learning to Solve and Explain Algebraic Word Problems. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 158–167.
 105. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.-W., Galley, M., and Gao, J. (2023). Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2310.02255>.
 106. Shen, J., Yuan, Y., Mirzoyan, S., Zhang, M., and Wang, C. (2024). Measuring Vision-Language STEM Skills of Neural Models. In The Twelfth International Conference on

- Learning Representations. <https://openreview.net/forum?id=spvaV5LELF>.
107. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al. (2022). Palm: Scaling language modeling with pathways. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2204.02311>.
 108. Raven, J. (2000). The Raven's progressive matrices: change and stability over culture and time. *Cognit. Psychol.* *41*, 1–48.
 109. Davies, M., Stankov, L., and Roberts, R.D. (1998). Emotional intelligence: in search of an elusive construct. *J. Pers. Soc. Psychol.* *75*, 989–1015.
 110. Zador, A., Escola, S., Richards, B., Ölveczky, B., Bengio, Y., Boahen, K., Botvinick, M., Chklovskii, D., Churchland, A., Clopath, C., et al. (2023). Catalyzing next-generation artificial intelligence through neuroai. *Nat. Commun.* *14*, 1597.
 111. Gupta, A., Savarese, S., Ganguli, S., and Fei-Fei, L. (2021). Embodied intelligence via learning and evolution. *Nat. Commun.* *12*, 5721.
 112. Mante, V., Sussillo, D., Shenoy, K.V., and Newsome, W.T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* *503*, 78–84.
 113. Wang, J., Narain, D., Hosseini, E.A., and Jazayeri, M. (2018). Flexible timing by temporal scaling of cortical responses. *Nat. Neurosci.* *21*, 102–110.
 114. Sorscher, B., Mel, G.C., Ocko, S.A., Giocomo, L.M., and Ganguli, S. (2023). A unified theory for the computational and mechanistic origins of grid cells. *Neuron* *111*, 121–137.e13.
 115. Song, H.F., Yang, G.R., and Wang, X.-J. (2017). Reward-based training of recurrent neural networks for cognitive and value-based tasks. *Elife* *6*, e21492.
 116. Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., and DiCarlo, J.J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. USA* *111*, 8619–8624.
 117. Kell, A.J.E., Yamins, D.L.K., Shook, E.N., Norman-Haigney, S.V., and McDermott, J.H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* *98*, 630–644.e16.
 118. Pugavko, M., Maslennikov, O., and Nekorkin, V. (2021). Dynamics of a Recurrent Spiking Neural Network in the Two-Alternative Choice Task. *Radiophys. Quantum Electron.* *64*, 817–832.
 119. Waldmann, M.R., and Hagmayer, Y. (2005). Seeing versus doing: two modes of accessing causal knowledge. *J. Exp. Psychol. Learn. Mem. Cogn.* *31*, 216–227.
 120. Nori, H., King, N., McKinney, S.M., Carignan, D., and Horvitz, E. (2023). Capabilities of gpt-4 on medical challenge problems. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2303.13375>.
 121. Digutsch, J., and Kosinski, M. (2023). Overlap in meaning is a stronger predictor of semantic activation in GPT-3 than in humans. *Sci. Rep.* *13*, 5035.
 122. Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al. (2023). Palm 2 technical report. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2305.10403>.
 123. Yin, Z., Sun, Q., Guo, Q., Wu, J., Qiu, X., and Huang, X. (2023). Do Large Language Models Know What They Don't Know?. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2305.18153>.
 124. Gray, J.R., Chabris, C.F., and Braver, T.S. (2003). Neural mechanisms of general fluid intelligence. *Nat. Neurosci.* *6*, 316–322.
 125. Gibbs, J.R.W. (2005). *Embodiment and Cognitive Science* (Cambridge University Press).
 126. Barsalou, L.W. (2008). Grounded cognition. *Annu. Rev. Psychol.* *59*, 617–645.
 127. Martinez, D. (2019). Immediate and long-term memory and their relation to crystallized and fluid intelligence. *Intelligence* *76*, 101382.
 128. MacCann, C. (2010). Further examination of emotional intelligence as a standard intelligence: A latent variable analysis of fluid intelligence, crystallized intelligence, and emotional intelligence. *Pers. Individ. Differ.* *49*, 490–496.
 129. Dunn, L.M., and Dunn, D.M. (2007). *Peabody Picture Vocabulary Test, fourth edition (APA PsycTests)*. <https://doi.org/10.1037/t15144-000>.
 130. Williams, K.T. (1997). Expressive vocabulary test second edition (EVT™ 2). *J. Am. Acad. Child Adolesc. Psychiatry* *42*, 864–872.
 131. Brown, J., Fishco, V., and Hanna, G.N. (1993). *Denny Reading Test: Manual for Scoring and Interpretation, Forms G & H* (Riverside Publishing).
 132. Markwardt, J.F.C. (1997). *Peabody Individual Achievement Test–Revised–Normative Update (APA PsycTests)*. <https://doi.org/10.1037/t15144-000>.
 133. Wason, P.C. (1968). Reasoning about a rule. *Q. J. Exp. Psychol.* *20*, 273–281.
 134. Ghallab, M., Nau, D., and Traverso, P. (2004). *Automated Planning: Theory and Practice* (Elsevier).
 135. Fechner, G.T. (1860). *Elemente der psychophysik (Breitkopf u. Härtel)*.
 136. Bechara, A., Damasio, H., Tranel, D., and Damasio, A.R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science* *275*, 1293–1295.
 137. Fluri, L., Paleka, D., and Tramèr, F. (2023). Evaluating Superhuman Models with Consistency Checks. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2306.09983>.
 138. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature* *529*, 484–489.
 139. Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human decisions and machine predictions. *Q. J. Econ.* *133*, 237–293.
 140. Vereschak, O., Bailly, G., and Caramiaux, B. (2021). How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proc. ACM Hum. Comput. Interact.* *5*, 1–39.
 141. Durand, K., Gally, M., Seigneuric, A., Robichon, F., and Baudouin, J.-Y. (2007). The development of facial emotion recognition: The role of configural information. *J. Exp. Child Psychol.* *97*, 14–27.
 142. Gross, J.J. (1998). The emerging field of emotion regulation: An integrative review. *Rev. Gen. Psychol.* *2*, 271–299.
 143. Zhang, J., Yin, Z., Chen, P., and Nichele, S. (2020). Emotion recognition using multimodal data and machine learning techniques: A tutorial and review. *Inf. Fusion* *59*, 103–126.
 144. Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Eng. J.* *5*, 1093–1113.
 145. Perner, J., and Wimmer, H. (1985). “John thinks that Mary thinks that,” attribution of second-order beliefs by 5-to 10-year-old children. *J. Exp. Child Psychol.* *39*, 437–471.
 146. Balaji, T., Annavarapu, C.S.R., and Bablani, A. (2021). Machine learning algorithms for social media analysis: A survey. *Comput. Sci. Rev.* *40*, 100395.
 147. Leger, M., Quiedeville, A., Bouet, V., Haelewyn, B., Boulovard, M., Schumann-Bard, P., and Freret, T. (2013). Object recognition test in mice. *Nat. Protoc.* *8*, 2531–2537.
 148. Karni, A., Meyer, G., Rey-Hipolito, C., Jezzard, P., Adams, M.M., Turner, R., and Ungerleider, L.G. (1998). The acquisition of skilled motor performance: fast and slow experience-driven changes in primary motor cortex. *Proc. Natl. Acad. Sci. USA* *95*, 861–868.
 149. Zou, Z., Chen, K., Shi, Z., Guo, Y., and Ye, J. (2023). Object detection in 20 years: A survey. *Proc. IEEE* *111*, 257–276.
 150. Raibert, M., Blankespoor, K., Nelson, G., and Playter, R. (2008). Bigdog, the rough-terrain quadruped robot. In *Proceedings of the 17th World Congress the International Federation of Automatic Control*, pp. 10822–10825.
 151. Vorhees, C.V., and Williams, M.T. (2006). Morris water maze: procedures for assessing spatial and related forms of learning and memory. *Nat. Protoc.* *1*, 848–858.
 152. Golledge, R.G. (1999). *Wayfinding Behavior: Cognitive Mapping and Other Spatial Processes* (JHU press).
 153. Abaspor Kazerouni, I., Fitzgerald, L., Dooly, G., and Toal, D. (2022). A survey of state-of-the-art on visual SLAM. *Expert Syst. Appl.* *205*, 117734.
 154. Teng, S., Hu, X., Deng, P., Li, B., Li, Y., Ai, Y., Yang, D., Li, L., Xuanyuan, Z., Zhu, F., and Chen, L. (2023). Motion planning for autonomous driving: The state of the art and future perspectives. *IEEE Trans. Intell. Veh.* *8*, 3692–3711.
 155. Wang, R., Yu, R., Tian, Y., and Wu, H. (2022). Individual variation in the neurophysiological representation of negative emotions in virtual reality is shaped by sociability. *Neuroimage* *263*, 119596.
 156. Roberts, A.C., Yeap, Y.W., Seah, H.S., Chan, E., Soh, C.-K., and Christopoulos, G.I. (2019). Assessing the suitability of virtual reality for psychological testing. *Psychol. Assess.* *31*, 318–328.
 157. Gómez-Zarà, D., Schiffer, P., and Wang, D. (2023). The promise and pitfalls of the metaverse for science. *Nat. Human Behav.* *7*, 1237–1240.
 158. Wang, G., Badal, A., Jia, X., Maltz, J.S., Mueller, K., Myers, K.J., Niu, C., Vannier, M., Yan, P., Yu, Z., and Zeng, R. (2022). Development of metaverse for intelligent healthcare. *Nat. Mach. Intell.* *4*, 922–929.
 159. Sun, T., Shao, Y., Qian, H., Huang, X., and Qiu, X. (2022). Black-box tuning for language-model-as-a-service. In *Proceedings of the 39th International*

- Conference on Machine Learning, pp. 20841–20855.
160. Shinn, N., Labash, B., and Gopinath, A. (2023). Reflexion: an autonomous agent with dynamic memory and self-reflection. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2303.11366>.
 161. Zhang, C., Gao, F., Jia, B., Zhu, Y., and Zhu, S.-C. (2019). Raven: A dataset for relational and analogical visual reasoning. In Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, pp. 5317–5327.
 162. Zhang, C., Jia, B., Gao, F., Zhu, Y., Lu, H., and Zhu, S.-C. (2019). Learning perceptual inference by contrasting. In Advances in Neural Information Processing Systems, pp. 1–13.
 163. Zhang, C., Jia, B., Zhu, S.-C., and Zhu, Y. (2021). Abstract spatial-temporal reasoning via probabilistic abduction and execution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9736–9746.
 164. Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., Alon, U., Dziri, N., Prabhunoye, S., Yang, Y., et al. (2023). Self-refine: Iterative refinement with self-feedback. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2303.17651>.
 165. Paul, D., Ismayilzada, M., Peyrard, M., Borges, B., Bosselut, A., West, R., and Faltings, B. (2023). Refiner: Reasoning feedback on intermediate representations. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2304.01904>.
 166. Park, J.S., O'Brien, J.C., Cai, C.J., Morris, M.R., Liang, P., and Bernstein, M.S. (2023). Generative agents: Interactive simulacra of human behavior. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2304.03442>.
 167. Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al. (2023). Palm-e: An embodied multimodal language model. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2303.03378>.
 168. Fan, L., Wang, G., Jiang, Y., Mandlekar, A., Yang, Y., Zhu, H., Tang, A., Huang, D.-A., Zhu, Y., and Anandkumar, A. (2022). Minedojo: Building open-ended embodied agents with internet-scale knowledge. In Advances in Neural Information Processing Systems, pp. 18343–18362.
 169. Huang, W., Wang, C., Zhang, R., Li, Y., Wu, J., and Fei-Fei, L. (2023). VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2307.05973>.
 170. Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K.-W., and Gupta, R. (2021). Bold: Dataset and metrics for measuring biases in open-ended language generation. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp. 862–872.
 171. Eloundou, T., Manning, S., Mishkin, P., and Rock, D. (2023). Gpts are gpts: An early look at the labor market impact potential of large language models. Preprint at arXiv. <https://doi.org/10.48550/arXiv.2303.10130>.
 172. Schwarting, W., Pierson, A., Alonso-Mora, J., Karaman, S., and Rus, D. (2019). Social behavior for autonomous vehicles. Proc. Natl. Acad. Sci. USA 116, 24972–24978.
 173. Jin, Z., Levine, S., Gonzalez Adauto, F., Kamal, O., Sap, M., Sachan, M., Mihalcea, R., Tenenbaum, J., and Schölkopf, B. (2022). When to make exceptions: Exploring language models as accounts of human moral judgment. In Advances in Neural Information Processing Systems, pp. 28458–28473.
 174. Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. Nat. Mach. Intell. 1, 389–399.