Journal of Language, Culture and Education; Vol.3, No.4; 2026
E-ISSN: 3006-3213 P-ISSN: 3006-3205
DOI: https://doi.org/10.70267/jlce.2026.v3n4.5062
Published by: Zeus Press

Evaluating AI "Understanding" with Cognitive-Psychological Criteria: Evidence, Gaps, and Structural Limits

Ruo Qin*

School of International Studies, Sun Yat-sen University, Zhuhai 519000, China

**Corresponding author: Ruo Qin.*

Abstract

With the rapid development of large-scale language models, the performance of artificial intelligence in language understanding and reasoning tasks has become increasingly strong. As a result, many people have gradually regarded the success of tasks as the same as true understanding. However, from the perspective of cognitive psychology, understanding is not merely about giving correct or fluent responses; it is actually an internal psychological process. It includes aspects such as meaning construction, context model update, reasoning generation, and metacognitive regulation. Under such circumstances, this paper systematically studies the cognitive understanding ability of artificial intelligence from the perspective of cognitive psychology. It employs the methods of literature review and qualitative case analysis. First, it clarifies the concept definitions and evaluation criteria of understanding in cognitive psychology by combining classic theories and experimental evidence, and then uses these criteria as the analytical framework. Study the performance of "similar understanding" in contemporary artificial intelligence systems. It was found that although artificial intelligence can approach human understanding at the behavioral level, it has not reached the core cognitive psychological standards. AI lacks experience-based situational models, causal and goal-oriented reasoning mechanisms, and inherent metacognitive monitoring. These limitations are structural, not quantitative issues. It reflects the fundamental difference between artificial intelligence systems and human cognition. This article can help clarify the theoretical distinction between surface manifestations and true understanding, and also provide a cognitive psychology foundation for a more cautious explanation of artificial intelligence capabilities.

Keywords

cognitive psychology, artificial intelligence, understanding ability

1. Introduction

1.1 Research Background

In the early years of the 21st century, artificial intelligence was mainly achieved through expert systems and models designed according to rules [1]. The working logic of these models mostly relied on knowledge bases and reasoning rules set by humans, and their functional scope was relatively clear. At that time, it was more like a "task-oriented tool" to serve a single goal in a specific scenario. Around 2010, there was a breakthrough in deep learning - in the two aspects of computers being able to understand images and sounds - which pushed the research and industrial application of artificial intelligence into a stage of rapid development [2]. According to the AI Index report released by Stanford University, from 2013 to 2023, over the past decade, The total number of papers related to AI in computer science has risen from approximately 102,000 to 242,000, nearly tripling. At the same time, the proportion of AI papers in computer science has also increased from 21.6% to 41.8% [3]. AI has become a core research direction in this field. After 2020, the emergence of large-scale language models has significantly enhanced the capabilities of artificial intelligence The performance of systems in tasks such as writing articles, reading articles to understand their meanings, and logical reasoning has all improved [4]. The AI Index also mentioned that the number of newly released large models globally in 2023 was twice that of the previous year. This reflects that both the development paths of technology and the methods of research are rapidly integrating and expanding.

In addition, when technological capabilities are spreading, it also brings some conceptual pressure and tension. In the industrial sector, IDC predicts that the global AI industry's revenue will reach 642.18 billion US dollars in 2024, an increase of 22.2% compared to the previous year. In China, the China Academy of Information and Communications Technology mentioned in a report released in 2025 that the scale of our country's core AI industry is expected to exceed 1.2 trillion RMB [5]. Nowadays, artificial intelligence is increasingly integrated into various systems of the economy and society. People are gradually viewing it as a technical agent with "quasi-cognitive ability". However, in academic discussions and public topics, the performance of AI systems in performing tasks well is often directly regarded as "understanding". This usage is different from the concept of "understanding" emphasized in cognitive psychology There is clearly a tense relationship. From the perspective of cognitive psychology, understanding is actually an internal psychological process centered on constructing meaning, which includes representation integration, contextualization processing, and metacognitive monitoring [6]. It is not merely about providing correct output or quick and efficient responses. In the current context where terms related to psychology such as "understanding" and "rationality" are frequently used in the context of AI, we need to start from the standards of cognitive psychology and systematically examine the nature, scope and boundaries of AI's "quasi-understanding". This can prevent the problems of over-generalization of concepts and disciplinary misalignment brought about by technological progress [7].

1.2 Research Significance

1.2.1 Theoretical Significance

This research can help artificial intelligence research and cognitive psychology distinguish concepts more clearly, thus resolving the common confusion about the concept of "understanding" in academia. This article re-examines the classic definitions and experimental standards for judging understanding in cognitive psychology, and clearly defines understanding as an internal cognitive process centered on meaning construction and mental representation. With this foundation in place, clear standards can be established to distinguish between "task-level, similar understanding manifestations" and "true psychological understanding". Such clarification can prevent people from taking behavioral outcomes or performance indicators as understanding itself, and also provide useful psychological references for evaluating the cognitive abilities of artificial intelligence.

1.2.2 Practical Significance

In practical terms, the research presented in this paper can aid both individuals and society in gaining a more rational and cautious understanding of artificial intelligence's (AI) capabilities. Currently, AI is widely used in learning, work, and decision-making processes. For individuals, distinguishing between "true understanding" and "high-performance output" can prevent an over-reliance on technological systems. This ensures that when humans collaborate with AI, they retain their own judgment and sense of responsibility. From a societal standpoint, this research offers psychological insights for appropriately positioning AI in sectors such as education, healthcare, and public governance. It helps prevent algorithm-generated results from being equated with cognitive judgments in technological applications, thereby minimizing the risks associated with flawed decisions and ethical dilemmas.

1.3 Research Methods

This research draws on cognitive psychology theory, with a particular focus on exploring the psychological implications of “understanding”. It also looks into how artificial intelligence performs in terms of cognitive understanding and identifies its limitations. The study primarily relies on a literature review approach, supplemented by case studies for deeper analysis.

(1) Literature Review Method

This article systematically reviews the relevant theories in cognitive psychology regarding understanding, meaning construction, mental representation, and metacognitive mechanisms, as well as some representative studies, in order to clarify the specific definition of the concept of understanding at the psychological level and the standards needed to assess the degree of understanding. In addition, this article also investigates the relevant research on “understanding ability” in the field of artificial intelligence to see how different disciplines define “understanding”, which can provide some theoretical references for analyzing issues related to AI understanding.

(2) Case Analysis Method

After looking through existing research, this study picked out several typical artificial intelligence systems and closely studied how they performed in language understanding and reasoning tasks, using them as examples for further analysis. Next, we compared the behaviors these systems showed with the evaluation standards from cognitive psychology. During the analysis, our main focus was on the differences between the two important mechanisms: meaning construction and self-monitoring. This approach helps clarify the boundaries and limitations of what artificial intelligence can “understand”.

2. Literature Review

2.1 Research on Artificial Intelligence “Understanding Ability”

Regarding the question of whether artificial intelligence has “comprehension ability” or not, current research mainly focuses on philosophy, artificial intelligence technology, and natural language processing. The core points of debate mainly lie in how to define the standards of “understanding” and whether it is possible to verify whether artificial intelligence truly understands based on visible behaviors. Current research can roughly be divided into three directions. This article will discuss the main contents of each direction respectively.

The first direction starts from the conceptual and philosophical level, which raises the most fundamental questions about the “understanding” of artificial intelligence. For instance, a representative view holds that the symbol manipulation or statistical pattern matching done by artificial intelligence does not equal true understanding at all. Just like the “Chinese Room” thought experiment proposed by Searle, it aims to illustrate that even if a system can speak grammatically correct Chinese sentences, it doesn't mean it truly understands the meaning of these sentences [8]. Later, Harnad added the “symbolic grounding problem”, saying that if the symbols used by artificial intelligence have no connection with real experience or the perceived world, then the meanings of these symbols are only superficial and formal, not true understanding [9]. This direction laid a theoretical foundation for subsequent research, viewing “understanding” as an internal cognitive state rather than something inferred from observable behaviors from the outside.

The following part of this article will discuss the task performance and evaluation framework of the research center. We aim to rely on standardized tests to measure the “comprehension ability” of AI. Levesque and others proposed the Winograd Schema Challenge to distinguish surface language processing from deeper understanding through common sense reasoning tasks [10]. Later, benchmark datasets like GLUE and SuperGLUE were used to evaluate the overall performance of models in several natural language understanding tasks [11]. Although these studies have promoted the rapid development of language model capabilities in the engineering field, they mainly measure the performance of systems under specific task conditions rather than understanding the mechanisms at this psychological level.

Subsequent research emerged in response to the rise of large-scale language models and also critically reflected on the usage of the concept of “understanding”. Bender and Koller believe that without a true semantic and contextual foundation, most of the current language models merely simulate the formal language structure [12]. Subsequent discussions on "stochastic parrots" emphasized that merely increasing the model size does not automatically lead to a qualitative shift towards true understanding [13].

2.2 Research Progress on "Understanding" in Cognitive Psychology

Research on "understanding" in cognitive psychology has a long history. The development of these theories has been gradual. At the beginning, it was only about building models for the concept of "understanding", then gradually dissecting the possible mechanisms behind it, and later extending to explanations at the regulatory and neural levels. In the middle and late 20th century, Researchers first identified "understanding" as a psychological process hidden in the heart, and the research in the past two decades has continued to refine the structure and operational mechanism of "understanding" on this basis.

Early research mainly focused on what forms the mental representations involved in the understanding process took. Kintsch's construction-integration model states that understanding is not simply the sequential reception of information from the text, but a process-in which propositional representations are activated and then integrated in working memory. Finally, a coherent meaning structure is formed [14]. This model clearly distinguishes surface memory from context-level understanding and provides a usable framework for subsequent research. On this basis, the context model theory gradually developed. This theory emphasizes that when readers understand, they will actively construct a mental model, which includes time, space, causality, and structures related to the goal [15]. There is a lot of experimental evidence to show that when the situation prompts in the text change, the reader's reaction time will slow down and the efficiency of memory retrieval will also decline. These can be regarded as direct evidence to show that understanding is actually related to whether the situation is consistent or not [16, 17].

Recently, Zwaan conducted a long-term review of the research on situational models and found that understanding is not a fixed cognitive representation but a continuous update process. Around the same time, research on individual differences in reading comprehension systematically demonstrated that comprehension ability is not determined by a single factor, but is the result of the combined effect of several cognitive resources, including working memory ability, inhibitory control ability, and the reader's background knowledge [18]. Even under very strict experimental conditions, different people will have very prominent differences in the depth of understanding and the types of mistakes they make. This highlights the structural characteristics of comprehension ability.

In the past decade, research has once again focused on the metacognitive regulation and development changes in the understanding process. The empirical study conducted by Tibken and his colleagues found that understanding monitoring ability is a key predictor of overall understanding performance, capable of explaining more than 20% of the variance [19]. Understanding is not only related to information processing itself It is also related to self-assessment and the regulation of this process. Wannagat and his colleagues' research on middle-aged and elderly people shows that the ability to monitor understanding declines with age. However, a higher level of education can alleviate this decline to a certain extent [20]. These research results all emphasize that understanding is an activity that requires cognitive regulation and needs to be continuously controlled rather than an automatic outcome.

In recent studies, neurocognitive methods have provided new evidence for understanding the related mechanisms. Keller and his colleagues' functional neuroimaging research found that when we understand text, the frontal and temporal lobe regions in the brain are coordinated and activated, and the activity in these regions becomes more obvious when we do causal reasoning and process abstract concepts [21]. Different stages of understanding align with distinct neural network models, suggesting that understanding emerges from coordinated multi-level cognitive systems.

As shown in Table 1, research on understanding in cognitive psychology has developed from mental representation models to metacognitive and neurocognitive explanations.

Table 1: Timeline of Research Progress on "Understanding" in Cognitive Psychology

Period	Representative Studies	Research Focus	Key Findings (Examples)
1980s–1990s	Kintsch et al.	Mental representation models	Proposed the Construction–Integration model, distinguishing surface-level processing from deep comprehension
1990s–2000s	Situation model research	Meaning construction	Situation consistency significantly affects reaction time and memory performance
2000s–2010s	Individual differences research	Cognitive resources	Working memory and inhibitory control jointly determine levels of comprehension
2020s	Tibken; Wannagat	Metacognition and development	Metacognitive processes explain over 20% of variance in comprehension differences
2020s	Keller et al.	Neural mechanisms	Frontal–temporal networks support deep comprehension

Source: Author's compilation based on [14–21].

2.3 Literature Review and Commentary

A review of the existing literature on the "comprehension ability" of artificial intelligence and the "comprehension" of cognitive psychology shows that previous research has made key progress on several levels. This paper finds that in the field of artificial intelligence, people have been discussing whether machines can truly understand meaning, gradually shifting from the initial philosophical criticism to assessment methods based on specific tasks. Recently, I have started to reflect on the issue of the comprehension ability of large-scale language models again. These studies have gradually made people realize that completing tasks well is equivalent to having a strong comprehension ability, which actually conceals many theoretical risks. Task completion, benchmark scores, or output accuracy are often implicitly treated as sufficient indicators of "understanding" in artificial intelligence research. However, such evaluation practices primarily capture external behavioral performance, rather than the internal cognitive processes through which understanding is formed. Additionally, in the field of cognitive psychology, through long-term theoretical modeling and experimental investigations, it has been systematically clarified that understanding, as an internal psychological process, has multiple layers of structure. It also indicates that understanding involves key mechanisms such as meaning construction, context integration, coordination of cognitive resources, and regulation of metacognition. From the perspective of cognitive psychology, understanding is therefore not defined by whether a correct response is produced, but by whether these internal psychological mechanisms are activated, coordinated, and dynamically adjusted during cognitive processing. The research in these two directions, together from the perspectives of engineering practice and psychological mechanisms, examines the matter of "understanding", providing quite rich materials and a solid theoretical foundation for discussing comprehension ability.

However, in current related research, several issues remain unclear. The "understanding" mentioned in artificial intelligence research is generally based on its behavioral performance or test results. Compared with the standards already established in cognitive psychology, it is neither systematic nor consistent. As a result, the conclusions drawn from different studies are difficult to compare at the conceptual level. Although cognitive psychology has elaborated on the mechanisms behind understanding in great detail, these theories and experimental findings have not yet been systematically incorporated into the framework for evaluating the comprehension ability of artificial intelligence. There is a clear disconnection between the two disciplines. This disconnection is not merely a matter of disciplinary focus, but reflects a deeper inconsistency in the criteria used to judge whether "understanding" has occurred: artificial intelligence research largely relies on performance-based indicators, whereas cognitive psychology emphasizes process-based and mechanism-oriented standards. Currently, most of the reflections on the comprehension ability of artificial intelligence remain at the level of macro criticism. There was no detailed analysis of which specific psychological mechanisms were lacking in the artificial intelligence system, nor was it made clear how these deficiencies formed the boundaries of its comprehension ability. Without first establishing a clear and theoretically grounded standard of understanding, evaluations of artificial intelligence comprehension risk conflating surface-level task performance with genuine cognitive understanding. These issues indicate that we must first clarify the concept, meaning and evaluation criteria of "understanding" within the framework of cognitive psychology, then use these criteria as references to analyze the performance of artificial intelligence related to understanding and its deficiencies in cognitive tasks, and finally continue to explore the psychological sources of these boundaries. This research approach it is the basis of the argumentative structure in Chapters Three and Four of this study.

3. Criteria for Judging Cognitive Understanding from a Cognitive Psychological Perspective

3.1 Conceptual Definition of “Understanding” in Cognitive Psychology

In the study of cognitive psychology, “understanding” generally does not refer to the ability to copy a text verbatip or provide a standard answer. In fact, it is defined as a process of constructing meaning in the mind. During this process, when people process external information, they will mobilize the knowledge they have learned before and then form mental ideas that can coherently explain these materials During the processing, these ideas will also be constantly examined and integrated [14]. Kintsch’s construction-integration model describes understanding as a dynamic process of “building first and then integrating”: readers will generate several possible viewpoints and relationships based on the received information, and then integrate the relevant information into a coherent meaning structure by meeting certain constraints. In this way, a mental representation that supports reasoning and memory can be formed. From this perspective, the key to understanding is not whether one can generate output like that of humans, but whether structured, inferential and coherent mental representations have been formed.

In this definition framework, our judgment of “depth of understanding” actually places more emphasis on whether the reader has constructed a context-level model and whether this model can be used for reasoning [16]. Scholars who study situational models believe that understanding is not only about handling the structure of words and sentences, but also about integrating various things stated in the text to form an overall representation [16]. Typical dimensions here include time, space, causality, and structures related to the goal. When the situation in the text changes, the cost for readers to process information will increase, and the effect of remembering things will also decline [17]. This indicates that they are updating that situation model online, which is the behavioral evidence of such updates. When we understand the text, we still need to make necessary inferences. Only in this way can the coherence of meaning be maintained. Causal inferences and bridging inferences can help readers organize scattered information into an explainable whole [22]. Later, more research on integrated modeling demonstrated that reading comprehension is not a single ability but the result of the coordination of several mechanisms, including representational structure, inference generation, and monitoring and regulation [18]. In the words of cognitive psychology, “comprehension” can be regarded as the ability to form and update context-based statements and make coherent inferences. It can also support subsequent transfers and interpretations.

3.2 Core Psychological Mechanisms of Human Cognitive Understanding

In human cognitive activities, the foundation of understanding this process is actually the construction of meaning and the formation of mental representations. When people deal with external information, they do not merely passively receive those symbols or sentences. Instead, they rely on the knowledge they have learned and the experience they have accumulated before to select, organize and then interpret this information. In this way, mental representations with internal structures can be formed [14]. This kind of mental representation usually goes beyond the surface meaning of the text and can reflect a comprehensive grasp of the relationships between events, the structure of concepts or the problem situation. For example, when reading narrative texts, readers will automatically integrate the characters’ goals, the causal relationships between events and the chronological order together. When learning abstract theories, Whether one can understand or not often depends on whether they can incorporate new concepts into their existing knowledge framework. The depth of understanding depends on whether these mental representations are structurally organized and inferential, because merely remembering surface parameters or fragments generally cannot support reasoning and knowledge transfer [14].

The process of understanding actually requires the joint efforts of reasoning generation, cognitive resource coordination, and metacognitive regulation. When information is incomplete or there is hidden content, everyone should actively push something forward to keep the meaning coherent - using causal reasoning to explain why something happened, or using bridging reasoning to connect the information before and after. In addition, when understanding, there are requirements for the ability of working memory, the control of attention, and the suppression of distractions [23]. If there is a lot of information or complex logic, these requirements will be more obvious. Moreover, metacognitive monitoring can give understanding a self-correcting function, so that everyone can determine whether they have truly understood [24]. If any mistakes are found, the handling methods can be adjusted. For example, read it again, slow down, or ask someone else for help. When these mechanisms are combined, they endow human understanding with dynamic, reflective and regulatory characteristics, making it different from simple rule execution or result matching.

3.3 Evidence from Classic Cognitive Psychology Experiments Supporting Criteria for Understanding

In cognitive psychology, the concept of “understanding” isn't just confined to theoretical reasoning. This article reveals that a set of classic experiments exists, which can turn this internal mental process into observable and measurable behavioral indicators. These experiments not only assist us in figuring out whether understanding has genuinely taken place but also help uncover the psychological mechanisms at play. They offer a solid and dependable empirical foundation for addressing these issues.

3.3.1 Situation Model–Based Comprehension Experiments: Empirical Evidence for Meaning Construction

At the level of meaning construction, Zwaan, Magliano and Graesser conducted a classic situational model experiment in 1995. Their experiment systematically studied whether readers would construct more situational representations than the information provided by the text itself when reading narrative texts [17]. In the experiment, researchers measured the changes in reaction time of participants when they read key sentences by adjusting the temporal and spatial continuity in the text. It was found that when the context of the text remained coherent, participants read faster. However, if there were changes in time or space within the text, the reaction time would significantly increase. These findings indicate that readers constantly construct and update situation models during the process of understanding the text, and changes in the context and environment directly increase the difficulty of understanding. This experiment used behavioral indicators for research. The results showed that understanding a text is not a process of handling each sentence individually, but rather relies on the construction of the overall meaning to be accomplished.

3.3.2 Inference-Generation–Based Comprehension Experiments: Empirical Evidence for Coherence Maintenance

Regarding reasoning generation, Graesser, Singer and Trabasso [22] conducted narrative text experiments to verify that people automatically generate causal reasoning during the process of understanding text. In these experiments, the researchers deliberately deleted the key causal information in the text and then used comprehension tests and reaction time tasks. Evaluate whether the participants have filled in the missing information by themselves. It was found that most participants would actively generate necessary causal inferences without being explicitly required, so as to make the text read more coherently [22]. Conversely, those who did not generate such inferences performed worse in terms of the accuracy of understanding and the retention of content memory. These experimental results indicate that inference generation is a key indicator of text understanding. If there is a defect in the ability to infer, it will make understanding fragmented.

3.3.3 Metacognitive Monitoring–Based Comprehension Experiments: Empirical Evidence for Self-Regulation

At the level of comprehension regulation, research on metacomprehension monitoring has demonstrated that individuals differ substantially in how accurately they judge their own comprehension [25]. This research findings show that there is a strong correspondence between the self-assessment results and actual performance of those with a relatively high level of comprehension. When they realize that their understanding of the content is insufficient, they can proactively adjust their learning strategies. Conversely, those with a lower level of comprehension tend to overestimate their understanding. There is also a lack of effective self-correcting behavior. These results indicate that metacognitive monitoring is a crucial part of the understanding process, endowing understanding with reflective and regulatory attributes.

In summary, these three classic experiment types offer empirical evidence for assessing understanding standards. They do so from the angles of meaning creation, reasoning development, and metacognitive control. These experiments have set up an experimental groundwork for subsequent studies on artificial intelligence's “quasi-understanding” abilities and the cognitive psychological boundaries that define them. This paper doesn't suggest directly using these experimental indicators in artificial intelligence systems. Instead, it proposes using them as a functional framework to determine if understanding has taken place and to examine the structural shortcomings of artificial intelligence in relevant areas [17, 22, 25].

4. Manifestations and Boundaries of Artificial Intelligence’s Cognitive Understanding Ability

4.1 AI’s Understanding-Like Performance in Cognitive Tasks

In tasks of language understanding and knowledge processing, large-scale language models like OpenAI’s ChatGPT can demonstrate “understanding” capabilities almost identical to those of humans in various contexts. For instance, when performing reading comprehension and question-answering tasks, these systems can generate coherent and structurally complete answers based on the information in the context. It can also accurately summarize the relationships between characters in the text, the sequence of events, and those obvious causal connections [26]. For instance, in scenarios such as educational assistance and information retrieval, ChatGPT can adjust the amount of details in its answers based on the type of question. It can summarize, explain, or repeat the same material in different ways, and the output format is very similar to what humans can understand. Additionally, in some standardized tests and publicly available benchmark tasks, The performance of these models in reading comprehension, answering questions based on knowledge, and logical reasoning problems is almost on par with or even exceeds the average level of humans. From the clear results they provide, one can clearly see the “characteristics of comprehension behavior”.

In more complex cognitive tasks, artificial intelligence systems also demonstrate certain reasoning capabilities and context maintenance skills. For instance, in multi-round conversations, ChatGPT can track previously discussed topics, maintain consistency in reference information, and recall earlier mentioned information in subsequent responses, showing a behavioral characteristic somewhat like constantly updating the “context” [26]. When dealing with tasks such as context disambiguation and common sense reasoning, models can sometimes combine clues in the context to make reasonable judgments on implicit causal relationships or ambiguous expressions. Thus, relying on tests previously regarded as “understanding thresholds” under specific conditions, during the interaction with users, Some systems can also adjust their explanations based on the questions raised by users later, such as adding some examples, using more colloquial expressions, or correcting errors in previous responses. These performances seem to have the characteristics of strategic adjustment and self-correction. When these features are combined, they constitute the “understanding-based” performance of artificial intelligence in cognitive tasks, making it seem to meet several criteria for understanding and judgment at the behavioral level. The “comprehension performance” presented in this section refers only to the similarity between artificial intelligence’s overt behavior and human comprehension, not whether it has formed a psychologically meaningful comprehension mechanism internally. A systematic analysis of this difference will be conducted in the next section based on cognitive psychology criteria. However, it must be emphasized that the ‘understanding-like performance’ discussed above is evaluated solely at the behavioral level. The following analysis therefore shifts the criterion from observable output similarity to the psychological mechanisms required for genuine understanding.

4.2 Analysis of the Limitations of AI Understanding Based on Cognitive Psychological Criteria

From the perspective of cognitive psychology, the limitations of artificial intelligence in understanding are not evenly distributed across all cognitive dimensions. Among them, the inability to construct and update situation models constitutes the most fundamental constraint.

4.2.1 Limitations in Situation Model Construction and Updating

According to cognitive psychology, understanding involves the continuous construction and updating of a situation model that integrates temporal order, causal relations, goals, and contextual constraints. When humans genuinely understand content, changes in temporal structure, causal assumptions, or task premises typically trigger observable cognitive consequences, such as increased reaction time, higher processing load, or the need to reorganize mental representations [16, 17].

However, large-scale language models such as OpenAI's ChatGPT do not exhibit such effects. These systems can maintain surface-level semantic coherence in multi-round dialogues and long-form text generation—for example, preserving referential consistency when retelling narratives or following an apparent logical sequence in explanatory discourse. Yet this coherence does not correspond to the updating of a psychological situation model. In many real-world interactions, when time reversals, conditional reversals, or premise modifications are introduced, the model continues to generate fluent responses without any detectable indication of comprehension disruption. In some cases, mutually incompatible situational assumptions even coexist within a single response [12, 13].

Unlike human comprehenders, whose processing behavior reflects internal costs when a situation model is disrupted, artificial intelligence outputs show no internal sensitivity to situational conflict. This absence of an internal cost signal indicates that the system has not constructed a situation representation that can be perturbed or reorganized. Consequently, what appears as "context sensitivity" in such models is more accurately described as conditional generation based on input distributions, rather than continuous modeling of event states in the psychological sense [12, 13].

4.2.2 Limitations in Inference Generation and Metacognitive Monitoring

In cognitive psychology, understanding is also characterized by the spontaneous generation of inferences, particularly causal inferences and bridging inferences, which serve to maintain coherence and fill informational gaps. When such inferences fail to occur, comprehension breakdown becomes observable [22, 25]. In addition, humans possess metacognitive monitoring mechanisms that allow them to evaluate their own understanding, detect failure, and actively regulate cognitive strategies.

In contrast, the inference processes exhibited by current AI systems differ fundamentally from these psychological mechanisms. What is often described as "reasoning" in large language models largely consists of statistically driven pattern completion based on frequent co-occurrences and contextual correlations. In tasks involving counterfactual reasoning or long causal chains, systems like ChatGPT can produce explanations that are linguistically well-formed and expressed with confidence. However, these explanations are not constrained by causal necessity and frequently drift away from the stated premises of the problem.

Moreover, current AI systems lack intrinsic metacognitive monitoring. They cannot internally distinguish between full understanding, partial understanding, or misunderstanding, nor can they autonomously evaluate the adequacy of their own comprehension [12, 13]. When revisions or corrections occur, they are typically triggered by external prompts from users rather than by internal detection of comprehension failure. In humans, by contrast, awareness of not understanding precedes strategy adjustment and the reconstruction of meaning. This fundamental difference indicates that AI systems simulate the outcomes of understanding without engaging the underlying regulatory processes that define understanding in cognitive psychology.

4.2.3 Structural Absence Rather Than Performance Deficit

Taken together, artificial intelligence can produce outputs that closely resemble the products of human understanding at the behavioral level. However, when evaluated against cognitive-psychological criteria—such as situation model updating, inference generation grounded in causal necessity, and metacognitive self-regulation—these systems fail to meet the standards required for genuine understanding [12, 13]. Importantly, these limitations do not arise merely from insufficient performance or technical immaturity, but from the absence of key psychological mechanisms that underlie human comprehension. As a result, the boundaries of AI understanding should be understood as structural rather than quantitative.

As shown in Table 2, human understanding and AI "understanding-like" performance differ across situation model updating, inference generation, and metacognitive monitoring.

Table 2: Comparison Between Human Cognitive Criteria for Understanding and AI Deficits in "Understanding" (Human Cognitive Criteria vs. AI Deficits in Understanding)

Cognitive-Psychological Criterion	Characteristics of Human Understanding	AI "Understanding-Like" Performance
Situation Model Construction and Updating	Humans continuously construct situation models integrating temporal, spatial, causal, and goal-related structures; changes in the situation lead to increased processing costs and representational restructuring	AI systems can maintain surface-level semantic coherence and referential consistency in output, but show limited internal sensitivity to premise reversals or situational conflicts
Psychological Necessity of Inference Generation	Humans spontaneously generate causal and bridging inferences to maintain coherence; failure to generate such inferences directly results in comprehension breakdown	AI can produce formally plausible explanations or fill in missing information, largely relying on statistical correlations and contextual probabilities
Metacognitive Monitoring and Self-Regulation	Humans can evaluate their own understanding, detect comprehension failure, and actively adjust strategies (e.g., rereading, revising, restructuring representations)	AI can revise or reformulate responses when externally prompted, but lacks intrinsic evaluation of its own understanding state

Source: Author's compilation.

4.3 Psychological Origins of the Boundaries of AI Cognitive Understanding

From the perspective of cognitive psychology, the boundaries of artificial intelligence's comprehension ability are not accidental outcomes of current technical limitations, nor can they be fully explained by insufficient model scale or training data. Rather, they arise from fundamental differences between artificial intelligence systems and humans in cognitive structure and psychological mechanisms. These differences determine why the limitations identified in the previous section are systematic and persistent.

4.3.1 Absence of an Experience-Based Situation Model

The limitations of artificial intelligence in situation model construction are rooted in the absence of an experience-based cognitive system. Human understanding is grounded in embodied experience: perceptual interaction with the environment, accumulated life experience, and goal-directed action jointly provide the foundation for constructing situation models. These models are not static representations, but dynamically updated structures that can be reorganized or disrupted as situations change [9, 16].

Artificial intelligence systems, by contrast, do not possess a subjective experiential world formed through perception and action. Their internal representations are not anchored in lived experience, but are derived from statistical regularities across large-scale symbolic data. As a result, what is often described as "context sensitivity" in AI systems reflects adjustments in output probabilities conditioned on input patterns, rather than the maintenance and updating of a stable situation representation across time and tasks. Without an experiential basis, artificial intelligence cannot develop a psychologically meaningful mechanism for situation model updating, which explains its insensitivity to situational disruption observed in cognitive tasks [9, 12].

4.3.2 Lack of Causal and Goal-Oriented Cognitive Structures

The boundaries of artificial intelligence understanding in inference generation stem from the absence of internalized causal and goal-oriented cognitive structures. In human cognition, reasoning is not merely an associative process, but a mechanism driven by the need to preserve causal coherence and explain relationships between events in relation to goals [22]. This gives human inference a form of psychological necessity: failures in reasoning directly undermine understanding itself.

Artificial intelligence systems do not operate on such internally structured causal models. Their inferential outputs are generated through statistical correlations learned from data, rather than through representations of causal relations or intentional goals. Because causal necessity is not encoded as an internal constraint, AI-generated inferences remain formally plausible but structurally fragile. This explains why artificial intelligence struggles with counterfactual reasoning, premise revision, and causal conflict resolution. The boundary of AI understanding in this dimension is therefore rooted in the absence of cognitive mechanisms that bind inference to causal explanation and goal-directed coherence [12, 13, 22].

4.3.3 Absence of Metacognitive Accessibility and Self-Regulation

A further psychological source of the boundary of artificial intelligence understanding lies in the lack of metacognitive accessibility. Human understanding is accompanied by the capacity to represent and evaluate one's own cognitive states. Individuals can determine whether they have understood, detect comprehension failure, and actively regulate cognitive strategies to restore understanding. This metacognitive monitoring plays a central role in sustaining robust and flexible comprehension.

Artificial intelligence systems lack access to their own cognitive states and therefore cannot autonomously monitor or regulate understanding. Output changes depend on external prompts rather than internal evaluation, preventing genuine metacognitive control and imposing a psychological upper limit on AI understanding beyond improvements in performance or scale.

As shown in Table 3, the boundaries of AI understanding can be traced to the absence of experience-based situation models, causal and goal-oriented structures, and metacognitive self-regulation.

Table 3: Psychological Causes Underlying the Limits of Artificial Intelligence Understanding

Dimension of Understanding Criteria	Psychological Basis of Human Understanding	Structural Characteristics of AI Systems	Psychological Causes of the Limitation
Situation Model Construction	Humans construct continuously updated situation models through the integration of perception, experience, and goal-directed processing	AI systems rely on input–output mappings and context-conditioned generation	Absence of an experience-based subjective world model
Inference Generation	Inference generation is driven by causal understanding and the need for semantic coherence, exhibiting psychological necessity	Inferences are produced as high-probability completions based on statistical correlations	Lack of causal structure and goal-oriented cognitive mechanisms
Metacognitive Monitoring	Humans can represent and evaluate comprehension states and regulate understanding accordingly	Revision and correction behaviors are triggered by external prompts	Absence of representational access to the system's own cognitive states

Source: Author's compilation.

5. Conclusions

This study examines the cognitive understanding ability of artificial intelligence from the perspective of cognitive psychology, in order to clarify the differences in concepts and theories between true human understanding and the "similar understanding" presented by contemporary AI systems. Currently, whether in the field of artificial intelligence research or during regular discussions, It is very easy to directly mistake the performance of AI in completing tasks as its true understanding. To respond to this common situation, this article re-examines the definition of the concept of "understanding" in cognitive psychology, particularly emphasizing that understanding is actually an internal psychological process, which includes specific contents such as meaning construction, update of situational models, reasoning generation, and metacognitive regulation. After conducting a systematic review of relevant literature, we established a cognitive psychology framework to determine what true understanding is. We also applied this framework to analyze the performance of large-scale language models in cognitive tasks. The research results revealed that although artificial intelligence can provide outputs at the behavioral level similar to those after human understanding, But in fact, it has not reached the key cognitive psychological standards.

References

[1] Russell, S. J., Norvig, P. Artificial intelligence: A modern approach. 3rd ed. Upper Saddle River, NJ: Prentice Hall; 2010.

[2] LeCun, Y., Bengio, Y., Hinton, G. Deep learning. Nature. 2015, 521(7553), 436-444. https://doi.org/10.1038/nature14539

[3] Stanford Institute for Human-Centered Artificial Intelligence. The AI Index 2024 Annual Report. Stanford University; 2024. Available from: https://aiindex.stanford.edu/report/ (accessed 7 June 2026).

[4] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. Language Models are Few-Shot Learners. In Advances in Neural Information Processing Systems 33. Curran Associates; 2020, pp. 1877-1901. https://doi.org/10.48550/arXiv.2005.14165

[5] China Academy of Information and Communications Technology. Research Report on the Development of Artificial Intelligence Industry (2025). Beijing: China Academy of Information and Communications Technology; 2026. (in Chinese)

[6] Anderson, J. R. Cognitive psychology and its implications. 8th ed. New York, NY: Worth Publishers; 2014.

[7] Gardner, H. The mind's new science: A history of the cognitive revolution. New York, NY: Basic Books; 1985.

[8] Searle, J. R. Minds, brains, and programs. Behavioral and Brain Sciences. 1980, 3(3), 417-457. https://doi.org/10.1017/S0140525X00005756

[9] Harnad, S. The symbol grounding problem. Physica D: Nonlinear Phenomena. 1990, 42(1-3), 335-346. https://doi.org/10.1016/0167-2789(90)90087-6

[10] Levesque, H. J., Davis, E., Morgenstern, L. The Winograd Schema Challenge. In Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning. AAAI Press; 2012, pp. 552-561.

[11] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S. R. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP. Association for Computational Linguistics; 2018, pp. 353-355. https://doi.org/10.18653/v1/W18-5446

[12] Bender, E. M., Koller, A. Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2020, pp. 5185-5198. https://doi.org/10.18653/v1/2020.acl-main.463

[13] Bender, E. M., Gebru, T., McMillan-Major, A., Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. ACM; 2021, pp. 610-623. https://doi.org/10.1145/3442188.3445922

[14] Kintsch, W. Comprehension: A paradigm for cognition. Cambridge: Cambridge University Press; 1998.

[15] Zwaan, R. A. From Words to Worlds: Twenty-Five Years of Advances in Situation Model Research. Current Directions in Psychological Science. 2025, 34(5), 287-292. https://doi.org/10.1177/09637214251326812

[16] Zwaan, R. A., Radvansky, G. A. Situation models in language comprehension and memory. Psychological Bulletin. 1998, 123(2), 162-185. https://doi.org/10.1037/0033-2909.123.2.162

[17] Zwaan, R. A., Magliano, J. P., Graesser, A. C. Dimensions of situation model construction in narrative comprehension. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1995, 21(2), 386-397. https://doi.org/10.1037/0278-7393.21.2.386

[18] McNamara, D. S., Magliano, J. P. Toward a comprehensive model of comprehension. In The psychology of learning and motivation. Elsevier; 2009, Vol. 51, pp. 297-384. https://doi.org/10.1016/S0079-7421(09)51009-2

[19] Tibken, C., Richter, T., Wannagat, W. Metacognitive comprehension monitoring: Cognitive abilities explain performance differences between younger and older adults. Scientific Studies of Reading. 2024, 28(3), 284-302. https://doi.org/10.1080/10888438.2023.2261572

[20] Wannagat, W., Nieding, G., Tibken, C. Age-related decline of metacognitive comprehension monitoring in adults aged 50 and older: Effects of cognitive abilities and educational attainment. Cognitive Development. 2024, 70, Article 101440. https://doi.org/10.1016/j.cogdev.2024.101440

[21] Keller, T. A., Mason, R. A., Legg, A. E., Just, M. A. The neural and cognitive basis of expository text comprehension. npj Science of Learning. 2024, 9, Article 21. https://doi.org/10.1038/s41539-024-00232-y

[22] Graesser, A. C., Singer, M., Trabasso, T. Constructing inferences during narrative text comprehension. Psychological Review. 1994, 101(3), 371-395. https://doi.org/10.1037/0033-295X.101.3.371

[23] Sun, C. Y., Peng, P., Chen, H. X., et al. The Effects of the Processing and the Storage in English and Chinese Working Memory on English Reading cognitive Comprehension among Chinese Middle Sehool Students. Psychological Development and Education. 2012, 28(1), 61-69. https://doi.org/10.16187/j.cnki.issn1001-4918.2012.01.003 (in Chinese)

[24] Liu, W. F., Si, J. W., Wang, Y. X. Metacognitive factors in cognitive strategy selection. Advances in Psychological Science. 2011, (7), 10-14. (in Chinese)

[25] Dunlosky, J., Lipko, A. R. Metacomprehension: A brief history and how to improve its accuracy. Current Directions in Psychological Science. 2007, 16(4), 228-232. https://doi.org/10.1111/j.1467-8721.2007.00509.x

[26] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems. 2022, 35, 27730-27744. https://doi.org/10.48550/arXiv.2203.02155

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Copyrights