Information Processing and Management 63 (2026) 104437

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Can large language models replace human experts in knowledge construction? A comparative analysis from the perspectives of information quality, information perception, and information load

Jingzhu Wei, Zhipeng Chen *

School of Information Management, Sun Yat-Sen University, Guangzhou 510006, PR China

ARTICLE INFO

keywords:
Large language models
Knowledge construction
Human-computer interaction
Text analysis

ABSTRACT

This study evaluates how large language models support knowledge construction by comparing Dense and Mixture of Experts architectures with human expert texts across six dimensions: Intrinsic Information Quality, Contextual Information Quality, Representational Information Quality, Linguistic Affinity, Structural Clarity, and Information Load. Using 2028 questions and 6084 responses, we compute composite indicators and estimate hierarchical regressions. MoE attains the highest Representational Information Quality in 89.69 % overall, rising to 95.72 % in closed domains, and the highest Linguistic Affinity in 65.30 %, but incurs high Information Load in 81.71 %. Dense leads Structural Clarity in 72.87 % and yields stable yet conservative expression. Human experts maintain low Information Load in 77.42 %. Regressions show that complexity increases representational expressiveness and load but weakens contextual alignment, while specificity increases affinity, structure, and load. Both model types lag behind the human standard on Intrinsic Information Quality and Contextual Information Quality. Findings support task aligned model selection and hybrid workflows.

1. Introduction

Knowledge construction originated in the field of cognitive science and refers to the active process through which individuals or groups construct meaning while solving problems or interpreting phenomena. Its goal is to attain a reliable and deep understanding of a given domain (Krajcik et al., 2008; Palincsar, 1998). This process is conceptually distinct from knowledge sharing (Pea, 1994), which emphasizes the transmission of knowledge between individuals, and from knowledge creation (Álvares, 2001; Engeström, 2001; Paavola et al., 2004), which highlights the generation of new ideas and practices through social interaction. Knowledge construction focuses more explicitly on the active reorganization and refinement of knowledge representations by cognitive agents (Van Aalst, 2009). As a core component of deep learning and creative problem solving, knowledge construction has become a central issue in cognitive science.

Rather than representing passive absorption, knowledge construction entails an active, contextualized, and reflective cognitive activity (Palincsar, 2012). It requires learners to integrate new information with prior knowledge and contextual cues to generate meaningful understanding (Hutchins, 1995; Brown et al., 1989). Meaningful learning only occurs when learners move beyond surface-level memorization toward deeper conceptual restructuring (Biggs, 1987; Dole & Sinatra, 1998). Research has further shown

* Corresponding author at: Sun Yat-sen University, 132 East Outer Ring Road, University Town, Guangzhou, Guangdong Province, PR China.
E-mail address: chenzhp68@mail2.sysu.edu.cn (Z. Chen).

https://doi.org/10.1016/j.ipm.2025.104437
Received 24 June 2025; Received in revised form 31 August 2025; Accepted 3 October 2025
Available online 11 October 2025

that knowledge construction typically proceeds through sequential stages, including information sharing, cognitive conflict, negotiation, and consensus building (Gunawardena et al., 1997).

Against this backdrop, question-answering tasks have been proposed as an appropriate lens for evaluating knowledge construction. Answering questions involves not only the generation of information but also activates deeper processing such as explanation, argumentation, and integration across sources. These processes align with the “constructive” and “interactive” levels defined in the ICAP framework (Chi and Wylie, 2014). Different types of questions span multiple levels of cognitive complexity, ranging from memory and comprehension to analysis, evaluation, and creation, thus enabling the externalization of multilayered features of knowledge construction (Adams, 2015). Studies in educational psychology and collaborative learning further indicate that question-driven interaction often initiates negotiation of meaning and co-construction of knowledge. In the domain of natural language processing, existing benchmarks such as TruthfulQA, 2WikiMultihopQA, and HotpotQA have incorporated dimensions like fact verification, multi-hop reasoning, and evidence chain generation (Nguyen et al., 2024), all of which correspond closely to the propose-argue-integrate-revise trajectory central to knowledge construction. Although question-answering tasks cannot fully capture all aspects of collaborative knowledge building, they effectively reflect its core cognitive dimensions.

Since the release of ChatGPT-3.5 at the end of 2022, based on a Dense architecture, large language models have rapidly reshaped human knowledge practices. Knowledge acquisition has shifted from a search-driven to a question-driven paradigm (Xue et al., 2023; Zhou & Wu, 2024). Compared to traditional search engines, a growing number of users now prefer to obtain information by directly posing questions to artificial intelligence systems (Zhou and Wu, 2024). The launch of DeepSeek-R1 in 2025 has further transformed the technological landscape. Based on a Mixture of Experts architecture, DeepSeek-R1 integrates multiple expert sub-models to achieve stronger problem-solving capabilities (Bishop, 1994; Gan et al., 2025). As an emerging global innovator in artificial intelligence, DeepSeek-R1 has accelerated cross-regional collaborative innovation and achieved significant breakthroughs across multiple fields (Sallam et al., 2025).

With continued advances in technology, large language models now excel not only in question generation and natural language understanding but also enable lower-cost and higher-efficiency information collection and decision support (Handler et al., 2024). In the near future, they are expected to play a central role throughout the knowledge construction process and drive structural changes in knowledge production and dissemination (Zhuang et al., 2023). However, public understanding of artificial intelligence remains limited. Combined with early instability in such systems, the societal acceptance of large language models is still constrained (Jagdishbhai & Thakkar, 2023; Sobieszek & Price, 2022). On the one hand, both researchers and the public need to understand the current performance gap between human experts and large language models in real-world knowledge work. On the other hand, there are growing concerns about the potential social consequences of replacing human experts with large language models in knowledge-intensive domains (Guo et al., 2023). For these reasons, it is necessary to conduct a direct and comparative examination of how large language models and human experts differ in their performance on knowledge construction tasks since the widespread adoption of such systems. A systematic assessment is needed to understand both the opportunities and challenges of positioning large language models as agents of knowledge construction.

Although existing research has evaluated the capabilities of large language models across dimensions such as bias, emotion, and reasoning styles and has explored the feasibility of replacing human experts in specific fields (Acerbi & Stubbersfield, 2023; Hou et al., 2024; S. Huang et al., 2025a; Y. Huang et al., 2025b; Riemer & Peter, 2024), there is a lack of systematic analysis comparing the linguistic generation characteristics of large language model outputs with those of expert-produced answers in knowledge construction tasks (Martínez et al., 2024). As language models, the ways in which large language models generate text and respond to questions directly shape their question-answering capabilities and their functional roles in knowledge construction. Language, as one of the most important mediums of human communication, reflects the cognitive depth of the responder (Fedorenko et al., 2024). Therefore, analyzing the micro-level linguistic features of large language model outputs in comparison with human expert answers helps reveal how these models participate in knowledge construction and what implications they have for human information cognition (Collins et al., 2024).

Furthermore, in the context of computer-supported collaborative learning and the co-construction of knowledge between humans and LLMs, scientifically understanding and evaluating the cognitive attributes of large language model outputs is essential for their effective deployment and trustworthy application. This is not only a technical issue but also a topic situated at the intersection of cognitive science and information science. Clarifying how large language models perform in knowledge construction is critical to optimizing human-computer collaboration and leveraging their complementary strengths in future educational, scientific, and organizational decision-making scenarios.

Building on this foundation, the present study formulates the following core research questions: What differences exist between the responses generated by different types of large language models and those provided by human experts when answering different types of questions? To what extent can large language models match the capabilities of human experts in knowledge construction tasks? This study focuses on linguistic generation features to identify the performance characteristics and cognitive impact pathways of different large language model architectures in knowledge-based question-answering. In doing so, it aims to uncover the strengths, biases, and risks involved in using large language models to assist or partially substitute for human expert responses.

The study draws upon information processing theory (Simon, 1978), levels of processing theory (Craik and Lockhart, 1972), and cognitive load theory (Paas et al., 2003) to construct a three-dimensional cognitive evaluation framework comprising information quality (IQ), information perception (IP), and information load (IL). This framework assesses the quality of responses, the depth of understanding, and the cognitive burden imposed on users. The empirical analysis is based on a dataset of 2028 real-world questions drawn from Wikipedia, WikiQA, and other knowledge-oriented platforms. A total of 6084 responses were collected from both human experts and large language models of different architectures. Through linguistic feature analysis and regression modeling, the study systematically compares how expert answers, a typical Dense-architecture model (ChatGPT-3.5), and an advanced MoE-architecture model (DeepSeek-R1) perform under varying levels of question complexity and explores the implications of these differences for human knowledge cognition and construction.

This study makes three main contributions. First, while current research in computer-supported collaborative learning and human-computer interaction often emphasizes collaborative patterns and task performance (Cress and Kimmerle, 2023), few studies have systematically analyzed the language generation mechanisms and cognitive pathways of large language models. By developing empirical indicators based on micro-level linguistic features, this study investigates the cognitive engagement patterns of large language models in knowledge construction (Collins et al., 2024). Second, research in information science has yet to fully compare the adaptability of Dense and MoE model architectures in complex tasks (S. Huang et al., 2025a; Y. Huang et al., 2025b; Song et al., 2025). Given the substantial differences in infrastructure and performance between the two architectures (Gan et al., 2025; Singh et al., 2025; Wang et al., 2024), this study designs comparative experiments based on ChatGPT-3.5 and DeepSeek-R1 to more comprehensively evaluate their capabilities in knowledge construction. Third, rather than relying on a single theoretical framework, this study integrates information processing theory (Simon, 1978), levels of processing theory (Craik and Lockhart, 1972), and cognitive load theory (Paas et al., 2003) to propose a three-dimensional cognitive framework comprising information quality, information perception, and information load. This framework provides a systematic means to evaluate the differences between expert answers and large language model outputs in knowledge construction tasks.

The remainder of this paper is structured as follows. Section 2 discusses the influence of large language models on knowledge construction, their potential to substitute for human experts, and how information processing theories can be applied to understand knowledge construction in human-computer interaction. Section 3 presents the development of the research model and hypotheses based on the aforementioned theoretical perspectives. Section 4 describes the data collection procedures and measurement of evaluation metrics. Section 5 reports descriptive statistics and regression analysis results. Section 6 discusses the theoretical and practical implications of the findings and outlines the study's limitations and directions for future research. Finally, Section 7 summarizes the main contributions and conclusions.

2. Theoretical background

2.1. Large language models in collaborative knowledge construction

Collaborative environments have been shown to significantly enhance knowledge construction. Empirical findings indicate that knowledge construction typically unfolds through a gradual progression from information sharing to cognitive conflict, negotiation, and consensus building (Gunawardena et al., 1997). Collaborative learning is therefore regarded as a key strategy in fostering knowledge construction (An et al., 2025), prompting researchers to examine how learners engage in meaning negotiation and co-construction through interactive processes such as question-answering (Mayordomo and Onrubia, 2015). From the perspective of computer-supported collaborative learning, knowledge construction relies not only on individual cognitive processes but also on multi-agent interaction patterns and mechanisms of information exchange (Cress and Kimmerle, 2023). Existing research has demonstrated the notable potential of large language models in supporting language comprehension, text generation, and personalized learning (Chan & Hu, 2023; Liu et al., 2024b). Although the exploration of digital tools in knowledge construction has a long tradition (Chen and Hong, 2016), the emergence of large language models marks a potential paradigm shift. However, there remains a lack of systematic investigation into the specific mechanisms through which large language models participate in collaborative knowledge construction (Cress & Kimmerle, 2023; Ouyang et al., 2023), and the literature offers limited consensus on their effectiveness in enhancing learning performance (An et al., 2025).

Despite ongoing debates, large language models, as aggregators of intelligence, leverage extensive pretraining on vast corpora and demonstrate powerful text generation capabilities, enabling them to provide learners with unprecedented perspectives and insights. Interaction with large language models can expand the cognitive breadth of individuals and groups and stimulate more diverse and novel ideas(An et al., 2025; Anantrasirichai & Bull, 2022). These models not only produce natural language outputs approaching human quality but also retrieve relevant information from extensive knowledge repositories, thereby supporting human–computer co-construction of knowledge (Liu et al., 2024a). To fully harness large language models for knowledge construction, learners must also possess advanced information literacy, such as the ability to decompose complex tasks into answerable questions, craft effective prompts, and select appropriate tools. These capabilities directly influence the effectiveness of large language models in high-load information environments (Chowdhury and Chowdhury, 2024). Consequently, a nuanced understanding of the interactional patterns and collaborative mechanisms between learners and large language models becomes essential to evaluating their potential for knowledge construction (Chiu, 2024).

2.2. The potential of large language models to substitute or supplement human experts in providing answers

Natural language processing constitutes one of the most dynamic subfields within artificial intelligence, integrating linguistics with computational techniques to enable computer to understand, generate, and interpret human language (Otter et al., 2020). The launch of large-scale systems such as ChatGPT has drawn unprecedented attention to natural language processing. Continuous advances in this domain have significantly enhanced the semantic modeling capacities of computer, laying a foundational basis for large language models to engage in language comprehension and cognitive support.

From a cognitive science perspective, the extraction of meaning in language comprehension has long been a central concern

(Louwerse, 2011). Since the cognitive revolution of the 1950s (Miller, 2003), researchers have increasingly recognized the potential of computational systems to simulate human cognitive processes (Guo et al., 2023; Shijaku & Canhasi, 2023; Yu et al., 2023). In recent years, scholars have begun incorporating large language models into classical cognitive experiments, revealing that these models can replicate human thinking patterns across multiple dimensions (Cai et al., 2023; Muñoz-Ortiz et al., 2024). Such findings offer theoretical support for conceptualizing large language models as cognitive agents in the knowledge construction process and establish a foundation for exploring the semantic properties and cognitive impact of their generated content. Nonetheless, large language models are still primarily regarded as auxiliary tools in blended learning and knowledge construction settings(Lee et al., 2024; Wu et al., 2024a). The extent to which their performance differs from and potentially substitutes that of human experts in these contexts remains insufficiently understood.

In response, some studies have attempted to position large language models as complete substitutes for human experts in specific tasks. For example, researchers have incorporated human feedback into model training pipelines to enable automated completion of tasks that previously required manual annotation, achieving notable results (Song et al., 2025). Other work has explored the use of large language models to conduct peer review of academic manuscripts, demonstrating their potential to fulfill the role of originality reviewers (Huang et al., 2025a). In the domain of mental health, experiments have shown that ChatGPT is capable of generating responses comparable to the average performance of professional counselors (Huang et al., 2025b). These developments suggest that, in certain specialized domains, large language models may already match or even exceed the average level of human expert performance.

Such advances have sparked discussions regarding the feasibility of replacing human experts across broader domains and have begun to challenge the long-standing epistemic authority traditionally ascribed to human actors (Qi et al., 2024). Nonetheless, scholars increasingly advocate for a reframing of the issue away from full replacement toward complementary collaboration. Large language models excel at rapidly processing vast quantities of data and generating pattern-based responses, whereas human experts possess irreplaceable strengths in deep understanding, ethical reasoning, and creative thought. Their integration may enhance the quality and efficiency of knowledge construction. Furthermore, public understanding of artificial intelligence remains underdeveloped, and early system instability has contributed to a cautious societal stance toward full replacement of human expertise (Jagdishbhai & Thakkar, 2023; Sobieszek & Price, 2022). Viewing large language models through a lens of complementarity rather than substitution offers a more pragmatic and constructive perspective.

It is also important to recognize that the capabilities of large language models are contingent upon the alignment between model architecture and task context (Gan et al., 2025). Traditional Dense architectures aim to fit all data with a single model, yet often struggle to capture the local heterogeneity inherent in complex tasks and thus face limitations in handling diversified demands (Gan et al., 2025). In contrast, Mixture of Experts architectures, which incorporate gating mechanisms to dynamically activate specialized sub-models based on task demands, exhibit greater adaptability to multifaceted and complex tasks (Gan et al., 2025; Huang et al., 2024). As such, evaluations of large language models' capacity for knowledge construction and their relationship to human expert performance must take into account the degree of fit between architectural design and application scenarios (Gan et al., 2025). A comprehensive cognitive understanding can only be achieved through such an integrated assessment (Cress and Kimmerle, 2023).

2.3. Information theories in knowledge construction from a human–computer interaction perspective

In knowledge construction environments shaped by human–computer interaction, the presentation of information plays a critical role in activating the learner's cognitive system. This section addresses two core questions. The first pertains to how information exerts its influence within problem-solving–oriented knowledge construction processes.

To address this question, it is necessary to draw upon information processing theory. This theory employs the metaphor of the brain as a computer to model the structure of problem-solving behavior. It posits that the activation level of cognitive structures and the pathways through which information is processed during perception, transformation, and output determine the overall effectiveness of learning (Payne, 2024; Stout & Klett, 2025). In this view, the quality of the information and its mode of delivery directly shape the outcomes of knowledge construction (Mwambe, 2024).

The second relevant theory is levels of processing, which asserts that the depth at which information is presented determines both its cognitive processing and its retention in memory. Shallow processing typically involves surface-level recognition, whereas deep processing entails semantic understanding and perception. Deeply processed information, enriched by elaborative associations, leaves more durable traces in memory and leads to improved learning outcomes (Craik and Lockhart, 1972). Accordingly, the deeper the information is processed, the greater the potential learning benefit (Craik & Lockhart, 1972; Espinosa-Rada et al., 2024; Palopak et al., 2023).

A third perspective is provided by cognitive load theory, which approaches learning from the standpoint of limited cognitive resources. When task demands exceed an individual's processing capacity, information overload occurs, diminishing learning efficacy and suppressing deep knowledge construction(Mao et al., 2024; Paas et al., 2003). Cognitive load thus serves as a key determinant of both the efficiency and upper bound of learning outcomes. Synthesizing these theoretical perspectives, the dynamic balance among information structure, mode of presentation, and cognitive load has emerged as a critical issue in promoting knowledge construction (C.-Y. Chen et al., 2012; K. Chen et al., 2024). These classical theories collectively emphasize that high-quality content, deeper information presentation, and appropriate cognitive load regulation constitute the three foundational elements of effective knowledge construction.

The second question concerns whether the aforementioned theories operate differently in human–computer interaction contexts involving artificial intelligence. The rise of computer simulations has laid the groundwork for cognitive science, while artificial intelligence offers new explanatory frameworks for how the brain processes, integrates, and reconstructs information (Nilsson, 2009). Within human–computer interaction contexts, information processing is conceptualized as an externalized cognitive process, encompassing sequential stages from perceptual input and cognitive processing to behavioral output (Stephanidis and Salvendy, 2024). The structure and presentation of input information significantly influence processing pathways. Higher-order cognitive mechanisms, including attentional control, curiosity activation, and memory retrieval, become decisive factors in shaping processing efficiency (Obrenovic and Starcevic, 2004). Furthermore, learners engaged in human–computer interaction often exhibit cognitive traits such as short-term attention, selectivity, and cyclical engagement. These traits reflect dynamic variations in cognitive load, which is considered a core variable in understanding how knowledge is acquired in digital interaction environments (Zhang et al., 2021). It provides insight into how individuals allocate cognitive resources and employ strategies in complex informational contexts (Su et al., 2025).

In sum, within human–computer interaction scenarios, information processing theory, levels of processing theory, and cognitive load theory have evolved beyond static metaphors of brain function to serve as theoretical scaffolds for understanding complex and dynamic cognitive processes. This integrated framework helps uncover the cognitive mechanisms driven by technology and bridges the gap between external information presentation and internal knowledge construction. It elucidates the intricate relationships among information structure, mode of delivery, and cognitive burden, thereby offering a more comprehensive account of the knowledge construction process under technological mediation. Taken together, information content quality, the depth of its semantic comprehension, and the magnitude of cognitive load emerge as the three key determinants of knowledge construction effectiveness.

2.4. *Research gap and positioning*

In recent years, numerous studies have compared the output quality of large language models and human experts across various settings, yet most have focused on singular task dimensions. Some investigations have examined the accuracy of question-answering (Lin et al., 2021), others have emphasized text style or user preference (Kabir et al., 2024), while a few have addressed performance in specialized professional tasks, such as legal examinations or clinical consultations (Katz et al., 2024; Motegi et al., 2025). Although these studies highlight the potential of large language models to match or even surpass human performance in specific domains, they typically lack a unified evaluative framework grounded in cognitive science to assess the performance of large language models across the full range of knowledge construction activities. Moreover, few have systematically considered how model architecture differences shape capabilities in knowledge processing.

For instance, the HC3 dataset offers a large-scale corpus of paired human and LLM-generated answers, providing a foundation for comparative research. However, its evaluation criteria are primarily oriented toward textual fluency and stylistic features, without adequate attention to structural language attributes, informational organization, or cognitive load (Guo et al., 2023). Similarly, empirical research on Stack Overflow has found that even when more than half of large language model–generated answers contain factual inaccuracies, users often still prefer these responses due to their completeness and human-like phrasing (Kabir et al., 2024). Such findings reveal a latent tension between linguistic affinity, reflecting the naturalness and fluency of model outputs, and information veracity, denoting factual accuracy. These observations suggest that assessing the complex process of knowledge construction based on single indicators such as answer accuracy is insufficient. There is an urgent need for a multidimensional evaluation framework that captures the reliability of information content, the effectiveness of user information perception, and the degree of

Table 1
Comparison of existing studies.

Study	Comparison Subjects & Tasks	Main Measures	Key Findings	Limitations	This Study's Extension
HC3 (Guo et al., 2023)	Human vs. ChatGPT (multi-domain QA)	Readability, Fluency, Detection Accuracy (AI text)	Provided large-scale human–AI paired corpus	Lacked structural and cognitive load indicators; did not distinguish problem complexity	Introduces a three-dimensional framework (IQ–IP–IL), provides regression-based evidence
TruthfulQA (Lin et al., 2021)	Human baseline vs. LLM(open-domain QA)	Truthfulness	Humans significantly outperformed models	Single-dimensional evaluation; excluded representational and cognitive measures	Incorporates truthfulness into intrinsic IQ, extends analysis to structural integrity and perception
Stack Overflow (Kabir et al., 2024)	ChatGPT vs. human high-voted answers (technical QA community)	Correctness, Consistency, Comprehensiveness, Conciseness, User Preference	LLM often preferred due to completeness	Neglected linguistic affinity and structural clarity; single-domain scope	Distinguishes linguistic affinity and structural clarity, integrates them into IP measures
Bar Exam (Katz et al., 2024)	GPT-4 vs. human candidates(legal bar exam)	Exam performance (bar scores)	GPT-4 achieved passing level	Not an open-domain QA; lacked linguistic and structural analysis	Returns to open-domain QA, evaluates coupling of language generation and cognitive load
Clinical QA (Motegi et al., 2025)	Physicians vs. GPT-4(clinical QA)	Overall Quality, Empathy, Alignment with Medical Consensus, Information Accuracy, Inquiry Comprehension, Harm Potential	GPT-4 scored higher across most dimensions	Limited sample size (60 question-response pairs); task-specific focus	Expands dataset scale, includes tasks with varied complexity levels

cognitive load management, in order to fully represent the complexity of knowledge construction activities.

At the same time, the rapid development of next-generation large language models based on Mixture of Experts architectures, combined with reinforcement learning–based optimization mechanisms, demonstrates markedly different characteristics from traditional Dense models in problem-solving and reasoning processes (Gan et al., 2025). Nevertheless, current research lacks systematic comparative studies based on real and cognitively complex tasks that adopt a unified evaluative standard to reveal how architectural differences manifest in knowledge construction performance and through which pathways they exert influence. The research gaps identified in the above literature review are summarized in Table 1.

In response to these limitations, the present study extends prior research by proposing an evaluation model centered on a three-dimensional cognitive framework comprising information quality, information perception, and information load. This model incorporates a structured set of indicators to systematically compare the knowledge construction capabilities of human expert responses and large language models with different architectures, namely the Dense-based ChatGPT-3.5 and the Mixture-of-Experts–based DeepSeek-R1, under varying levels of question complexity. The proposed framework consists of three core dimensions. Information quality refers to the reliability and completeness of response content, information perception captures the affinity and clarity of language presentation, and information load denotes the efficiency of cognitive resource utilization during information processing. More specifically, information quality evaluates textual content in terms of factual accuracy, semantic relevance, and conciseness. Information perception reflects the learner's subjective understanding and perceived usefulness of the information, including linguistic cues such as stylistic affinity and expressive clarity, which indicate the cognitive depth triggered by the presentation. Information load represents the degree of cognitive effort required to process the information, capturing the complexity and burden associated with knowledge processing. These three dimensions together constitute a comprehensive chain from external information input to internal cognitive processing and offer a systematic lens for characterizing outputs from large language models and human experts at the levels of expression, perception, and processing. Through this multidimensional framework, it becomes possible to more comprehensively assess the key elements of knowledge construction and to capture its complex features. This three-dimensional theoretical model thus enriches existing evaluative approaches that have historically relied on single indicators and provides a solid theoretical foundation for the development of the subsequent research model.

3. Research model development

Building on the preceding theoretical review, this study constructs an integrated research model that draws from information processing theory, levels of processing theory, and cognitive load theory. The model evaluates responses along three core cognitive dimensions: information quality, information perception, and information load. These dimensions respectively reflect the reliability and completeness of the content, the affinity and clarity of its linguistic presentation, and the extent to which information processing consumes cognitive resources. This tripartite structure represents an information-processing chain from content input to presentation form and ultimately to cognitive burden. It provides a systematic framework for evaluating the outputs of large language models in terms of expression, perception, and processing.

Furthermore, recognizing that model performance may vary across different task contexts, the model incorporates moderating effects of question features, specifically question complexity and specificity, on the relationship between the source of the response and its quality. This integrated structure is illustrated in Fig. 1.

The proposed research model systematically compares the performance of responses generated by human experts and by two large language models representing distinct architectures: ChatGPT-3.5 as a representative of the Dense architecture and DeepSeek-R1 as a representative of the Mixture-of-Experts (MoE) architecture. It also examines the moderating role of Question Cues. The following sections explain the theoretical grounding and relationships among all model variables, upon which the research hypotheses are constructed.

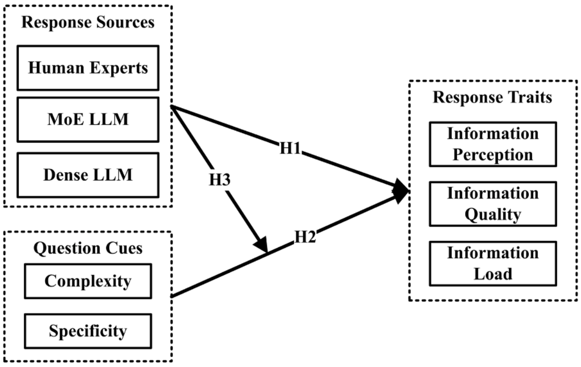

Fig. 1. Research model.

3.1. Response traits: information quality, information perception, and information load

The effectiveness of knowledge construction is largely dependent on the quality of the information received and the cognitive processes activated during comprehension. To capture these dimensions, the present study characterizes response content from three complementary perspectives: information quality, information perception, and information load. This framework is grounded in foundational work from information science and cognitive psychology. The dimension of information quality emphasizes the validity and reliability of the input content. The dimension of information perception concerns the impact of information presentation on user understanding and engagement. The dimension of information load reflects the extent to which information processing consumes cognitive resources. The following subsections define each of these core constructs.

3.1.1. Information quality

Information quality refers to the degree to which a response meets standards of factual accuracy, semantic relevance, content completeness, and expressive consistency. High-quality information should be both reliable and directly pertinent to the question, presented in a manner that is clear and coherent. To ensure theoretical and operational consistency, this study adopts established frameworks in the field of information quality (Kahn et al., 2002; Lee et al., 2002; Pipino et al., 2002; Stvilia et al., 2007; Wang & Strong, 1996). Specifically, the construct is disaggregated into three analytically tractable subdimensions: Intrinsic Information Quality, which refers to the factual and argumentative accuracy of the response, reflected in its consistency with evidence or common knowledge (Bowman et al., 2015; Williams et al., 2017); Contextual Information Quality, which denotes the semantic relevance of the response to the question and the comprehensiveness with which it addresses key informational demands (Robertson and Zaragoza, 2009); and Representational Information Quality, which concerns the expressive clarity, linguistic adequacy, and representational precision of the response (Chen et al., 2024; Kembaren & Aswani, 2022; Zheng et al., 2013). A fourth classical dimension, Accessibility, describes the ease of obtaining information and the openness of its source. However, because the present study restricts all analyzed responses to publicly available, structurally normalized online sources, the influence of accessibility is held constant by design and excluded from further analysis.

To operationalize the three focal quality dimensions, a set of linguistic and semantic features is employed. Natural language inference techniques are used to evaluate whether response sentences are supported by reference materials, serving as indicators of factual accuracy (Robertson and Zaragoza, 2009). Semantic similarity and key-element coverage metrics are computed to assess contextual completeness. Measures of lexical density and lexical diversity, such as the Measure of Textual Lexical Diversity (MTLD), are applied to evaluate representational clarity and conciseness (Chen et al., 2024; Kembaren & Aswani, 2022; Zheng et al., 2013).

3.1.2. Information perception

Information perception refers to the user's subjective assessment of the response in terms of ease of comprehension, stylistic affinity, and structural clarity. This dimension captures how the form of information presentation affects cognitive processing. Expressive clarity and emotional resonance in the response can enhance attention and elicit affective engagement, thereby promoting deeper levels of cognitive elaboration (Go & Sundar, 2019; Wei et al., 2022). To fully characterize this dimension, the present study distinguishes between two complementary aspects: linguistic affinity and structural clarity.

Linguistic affinity evaluates the human-likeness and readability of the language style, including features such as informal expressions, sensory language, and affective or motivational vocabulary. These elements enhance the social presence and emotional resonance of the text and have been shown to influence user engagement in studies on social media, dialogue systems, and persuasive communication (Boyd et al., 2022; Doyle et al., 2021; Go & Sundar, 2019; Mirzaei et al., 2024). Structural clarity focuses on discourse organization and coherence. Texts with higher structural clarity reduce the cognitive effort required for comprehension (Graesser & McNamara, 2011; McNamara et al., 2014). Two metrics are employed to assess this property. Connective density measures the frequency of explicit discourse markers and reflects the degree to which logical relations among propositions are clearly articulated (Johansson, 2008; Lu & Hu, 2024). Adjacent overlap measures the content continuity between consecutive sentences and captures local textual coherence (Breadmore and Carroll, 2016). Together, linguistic affinity and structural clarity shape the perceived ease and accessibility of a given response and influence both user comprehension and cognitive engagement. These metrics have been widely used in tools such as Coh-Metrix to assess text comprehensibility in academic, popular science, and educational contexts (Graesser & McNamara, 2011; Lu & Hu, 2024).

3.1.3. Information load

Information load refers to the extent of cognitive resource consumption experienced by users while processing response content. It constitutes a foundational construct in understanding language comprehension and is a critical variable in determining learning efficacy and cognitive transfer (Paas et al., 2003; Sweller, 2011). According to cognitive load theory, when informational complexity exceeds the capacity of working memory, excessive cognitive load impairs learning (Chandler and Sweller, 1991). At the same time, appropriate cognitive challenges may elicit deeper processing strategies that are conducive to learning outcomes (Kalyuga, 2011; Schnotz & Kürschner, 2007; Wach et al., 2015).

Within textual settings, syntactic complexity is a major contributor to cognitive load. Dependency length, in particular, is widely used to quantify the burden imposed on short-term memory during sentence processing. Greater distances between syntactically dependent elements within a sentence typically increase processing effort and memory demands, thereby elevating cognitive load (Futrell et al., 2015; Gibson, 1998). In this study, average syntactic dependency length within a response is adopted as the primary indicator of information load. This measure reflects both the structural complexity of language generated by the model and the immediate memory resources required for sentence comprehension. It also allows inferences about whether model-generated responses introduce unnecessary comprehension burdens relative to task demands. A prevalence of long-distance syntactic dependencies may signal that the response exceeds the threshold of effortless comprehension, thereby intensifying cognitive effort(Demberg & Keller, 2008; Fedorenko et al., 2024; Liu, 2008). Through this dimension, the model captures the latent processing cost associated with a response and complements the evaluations of content reliability and perceived clarity. Taken together, these three dimensions support a comprehensive assessment of a response's value for knowledge construction.

3.2. Effects of response source

The source of an answer constitutes a critical determinant of both information quality and presentation efficacy. Within question-driven knowledge construction processes, users often decompose complex tasks into subtasks and employ intelligent tools to retrieve necessary information. While the clarity of prompts and the design of interaction mechanisms are essential, the tool selected for question answering itself plays an equally important role. Even when posed with the same query, different tools may yield substantially divergent outputs (Chowdhury and Chowdhury, 2024). Therefore, evaluating the capacity of large language models to support knowledge construction requires consideration of the alignment between model architecture and task context. Only by integrating model source with the characteristics of the specific question can one fully interpret performance differentials in collaborative knowledge construction scenarios.

At present, two major architectural paradigms dominate large language model development: Dense and Mixture-of-Experts (MoE) architectures. ChatGPT-3.5, released by OpenAI in 2022, represents the Dense architecture. Based on the GPT-3.5 model (Brown et al., 2020), ChatGPT-3.5 has been widely adopted due to its coherence in language generation, flexibility in dialogue interaction, and human-like response style (Singh et al., 2025). Dense models activate all model parameters during each inference, and early improvements in performance were largely driven by scaling up parameter count and model depth. However, the exponential growth in parameters has led to sharp increases in training cost and data requirements (Wang et al., 2024). When constrained by limitations in computational power or data availability, achieving performance gains under resource constraints presents a significant challenge (Singh et al., 2025; Teng et al., 2024).

MoE architectures have emerged as a promising alternative by offering a divide-and-conquer solution to the scalability problem (Jacobs et al., 1991; Jordan & Jacobs, 1994). MoE models decompose input tasks into subcomponents, each handled by specialized expert subnetworks, with a gating mechanism dynamically selecting the most appropriate expert for execution. This design improves computational efficiency and enhances scalability (Cai et al., 2025; Jacobs et al., 1991; Jordan & Jacobs, 1994). DeepSeek-R1, released in early 2025, exemplifies this new class of MoE-based large language models(Guo et al., 2025). It achieves improvements in both performance and alignment safety mechanisms while minimizing computational costs by activating only a subset of parameters per inference. Owing to these architectural advantages, MoE models have become increasingly favored for large-scale deployment and are now viewed as viable successors to Dense models.

Moreover, the scope of MoE architectures continues to expand and has emerged as one of the most dynamic trends in natural language processing. Empirical studies indicate that MoE models outperform traditional Dense models in handling high-dimensional sparse data, integrating multisource heterogeneous information, supporting continual learning, and improving model interpretability (Chen et al., 2023; Germino et al., 2024; Huang et al., 2024; Kang et al., 2023; Liu & Sun, 2023; Mirus et al., 2019; Wu et al., 2024b). In contrast to the general-purpose orientation of Dense architectures, MoE models allow different types of information to be routed to the expert subnetworks best suited for their processing, thereby enhancing both efficiency and precision when responding to complex tasks (Bishop, 1994). In the context of knowledge construction, where high-quality content generation and the capacity to handle complex queries are pivotal, MoE architectures provide a valuable vantage point for evaluating the potential and limitations of large language models.

Building on this foundation, the present study incorporates human expert responses, ChatGPT-3.5 as a Dense model, and DeepSeek-R1 as an MoE model into a unified evaluative framework. This enables a systematic comparison of Response Source adaptability across knowledge construction scenarios. Given the architectural and cognitive strategy differences across sources, it is hypothesized that the source of the response will significantly influence the information quality, information perception, and information load of the generated text. Accordingly, the following hypothesis is proposed:

H1. There are significant differences in information quality, information perception, and information load across Response Sources, including human experts and large language models of different architectures.

3.3. Effects of question cues

Question Cues significantly influence both the cognitive strategies employed by responders and the informational pathways activated during content generation. Prior research has demonstrated that different types of question cues can direct responders' attention to key informational elements, prompting cognitive adjustments aligned with task demands (Caraban et al., 2019; Wu & Zheng, 2025). When questions provide clear directions and cues, responders are more likely to align their expertise with the requirements of the query, thereby improving accuracy and relevance (Stokhof et al., 2022). Conversely, poorly defined or overly open-ended questions may lead to unfocused or superficial responses. Thus, cues and constraints that correspond to the cognitive demands of a question contribute to the production of more targeted and high-quality answers (Wu et al., 2020).

This study conceptualizes Question Cues along two dimensions: complexity and specificity. Complexity refers to the breadth and depth of domain-specific knowledge required to answer the question. Higher complexity denotes a larger number of interrelated concepts and longer reasoning chains, requiring more extensive background knowledge and cognitive effort to formulate an appropriate response (Byström & Järvelin, 1995; Wu & Zheng, 2025). Previous studies have found that experts evaluating question quality on information platforms prioritize the degree to which questions align with their domain expertise(Stokhof et al., 2022). Therefore, high-complexity questions yield high-quality answers only when responders possess deep disciplinary knowledge. Specificity denotes the degree of detail and structural clarity embedded in the question. Highly specific questions impose stricter requirements on the precision and depth of answers and demand more elaborate detail and well-structured reasoning (Ghosh, 2021). For example, recall-based questions and application-based questions engage distinct cognitive processes(Adams, 2015). According to Bloom's taxonomy, higher-order cognitive tasks tend to contain more precise details and structural cues that guide the responder (Adams, 2015).

Taken together, question complexity and specificity are expected to influence all three dimensions of answer characteristics. High-complexity questions may increase difficulty, potentially lowering information quality or raising cognitive load. Specific and well-defined questions, by contrast, facilitate focus and clarity in responses, thereby improving information perception and semantic relevance, though possibly at the cost of slightly increased information load due to additional detail. Given the importance of question features in shaping response behavior, the following hypothesis is proposed:

H2. Question complexity and specificity significantly influence information quality, information perception, and information load.

3.4. *Interaction effects between question cues and response source*

Beyond the main effects of Response Source and question features, the present study also considers their potential interaction. Distinct sources may demonstrate comparative advantages depending on the complexity and specificity of the question. Prior research suggests that alignment between model capabilities and task context is a critical determinant of model performance (Cress & Kimmerle, 2023; Gan et al., 2025). This implies that the relative performance of human experts and large language models may not be constant across question types but may instead vary as a function of the problem characteristics.

For example, when faced with highly complex questions, MoE-based models may be better equipped to integrate a wider range of disciplinary knowledge through the dynamic activation of specialized expert modules, thereby delivering more complete responses. In contrast, with low-complexity or general-knowledge questions, human experts may demonstrate greater expressive precision and better control of cognitive burden, offering concise and accessible answers. Similarly, when questions contain clearly specified details, all response sources may be able to leverage their respective strengths, narrowing the performance gap. However, in the absence of clear guidance, large language models may provide generic or underdeveloped responses due to insufficient contextual anchoring, while human experts can draw on background knowledge to construct more reliable and nuanced answers.

In sum, we posit that the interaction between Question Cues and Response Source shapes the performance of responses across the three focal dimensions. The influence of question complexity and specificity on response outcomes is expected to vary by Response Source. Based on this reasoning, the following hypothesis is proposed:

H3. There is a significant interaction between Question Cues and Response Source in predicting information quality, information perception, and information load.

4. Methods

4.1. *Selection of large language models*

This study selected ChatGPT-3.5 and DeepSeek-R1 as representative large language models to compare the performance of different model architectures in supporting knowledge construction tasks. The selection of these two models was guided by the following considerations.

First, both models represent highly influential and widely adopted implementations within their respective architectural paradigms, supported by extensive user communities (Conroy and Mallapaty, 2025). They offer a realistic representation of typical user experiences with LLMs in everyday contexts. This representativeness is not limited to technical specifications but extends to the transformative role these models play in shaping public modes of knowledge acquisition and cognitive behavior. Given the central research question of whether LLMs can serve as substitutes for human experts in knowledge construction, the inclusion of models with high public visibility and accessibility holds greater relevance than the pursuit of experimental models operating at the edge of technical performance under constrained laboratory conditions.

Second, ChatGPT-3.5 remains one of the most widely used and cognitively familiar models for LLM evaluation. It has been frequently adopted as a benchmark in both academic and industrial research (Gao et al., 2025; Liu et al., 2025; Yuan et al., 2025). Notably, while GPT-4 has succeeded ChatGPT-3.5, various reports and technical leaks indicate that GPT-4 has transitioned to a Mixture-of-Experts architecture (Yi et al., 2023; Vats et al., 2024). This suggests that traditional Dense architectures may be gradually phased out in mainstream development. As such, ChatGPT-3.5 serves as a valuable historical and technical reference for the Dense category, offering a stable point of comparison.

Third, DeepSeek-R1, as a recently introduced model based on the MoE architecture, has demonstrated the potential of sparse expert systems (Zhao et al., 2025). It is widely regarded as a direct competitor to OpenAI's flagship models and is frequently used in comparative studies against the GPT series (Bevara et al., 2025; Mercer et al., 2025; Wu et al., 2025). DeepSeek-R1 achieves performance levels comparable to proprietary models developed by leading organizations such as OpenAI and Anthropic, but with significantly lower training costs (Wang and Kantarcioglu, 2025). Since its release in 2025, DeepSeek-R1 has been widely discussed as a milestone in the evolution of LLMs, with numerous studies exploring its broader social implications (Deng et al., 2025; Puspitasari et al., 2025). Taken together, the societal influence and architectural significance of ChatGPT-3.5 and DeepSeek-R1 justify their selection as the focal models in this study, representing the Dense and MoE categories, respectively.

4.2. Dataset composition

To systematically compare the performance of different LLM architectures and human experts in knowledge construction tasks, the present study adopted the Human ChatGPT Comparison Corpus (HC3) as its base dataset (Guo et al., 2023). HC3 is one of the most widely used corpora in academic research for comparing LLM-generated and human-generated answers and has been employed in several recent studies (Reviriego et al., 2024; Rosenfeld & Lazebnik, 2024).

The selection of HC3 is supported by the following rationale. First, the HC3 dataset was compiled in 2023, with LLM responses generated using the initial release of ChatGPT-3.5. As a Dense-based model launched by OpenAI in late 2022, ChatGPT-3.5 quickly became a globally recognized general-purpose language model with the largest user base to date. Consequently, the ChatGPT-3.5 responses in HC3 are representative of typical Dense architecture performance on question-answering tasks involving knowledge construction.

Second, the questions included in HC3 span a wide range of domains, encompassing both closed-domain questions requiring technical terminology and theoretical reasoning, and open-domain factual questions answerable without specialized background knowledge. This diversity allows for comparative analysis of human and LLM performance across varying cognitive contexts. Closed-domain questions generally demand higher levels of disciplinary expertise, while open-domain questions rely on general world knowledge. The inclusion of both types of questions enables the study to assess how model and human performance vary under conditions of high and low task complexity.

Third, the questions in HC3 are drawn from real-world online environments, including platforms such as Wikipedia and WikiQA. This ensures the ecological validity and topical diversity of both the questions and the human-generated answers, thereby enhancing the practical relevance of the study. In other words, HC3 provides realistic and application-oriented question-answering data, suitable for evaluating the actual capacity of LLMs to support knowledge construction.

The human expert responses in HC3 originate from two main sources. The first includes curated Wikipedia entries and case summaries that have undergone multiple rounds of peer editing, providing authoritative answers to hundreds of questions related to computer science (Guo et al., 2023). These responses represent expert-level knowledge within closed-domain contexts. The second source consists of high-quality answers extracted from public Q&A platforms such as WikiQA, written by experienced domain users and endorsed through community voting (Yang et al., 2015). These answers reflect widely accepted expert-level responses and serve as credible reference standards.

To expand the comparison and incorporate the MoE architecture, the present study introduced additional responses generated by DeepSeek-R1. Specifically, all 2028 questions in the HC3 dataset were re-input into DeepSeek-R1, and its generated responses were collected. This extension enabled a direct comparison between responses from Dense (ChatGPT-3.5) and MoE (DeepSeek-R1) models for the same set of questions and allowed investigation of how architectural differences may influence future paradigms of human–computer knowledge construction.

The resulting corpus includes 2028 questions and a total of 6084 responses, comprising outputs from ChatGPT-3.5, DeepSeek-R1, and human experts. This cross-architecture and cross-source corpus serves as the empirical foundation of the present study. The structure and composition of the dataset are illustrated in Fig. 2.

It is important to acknowledge that responses generated by large language models may vary slightly across dialogue threads due to mechanisms such as stochastic sampling. However, existing studies have shown that single-sample outputs are sufficiently representative in tasks involving text quality evaluation and content analysis. Repeated sampling does not substantially alter the overall

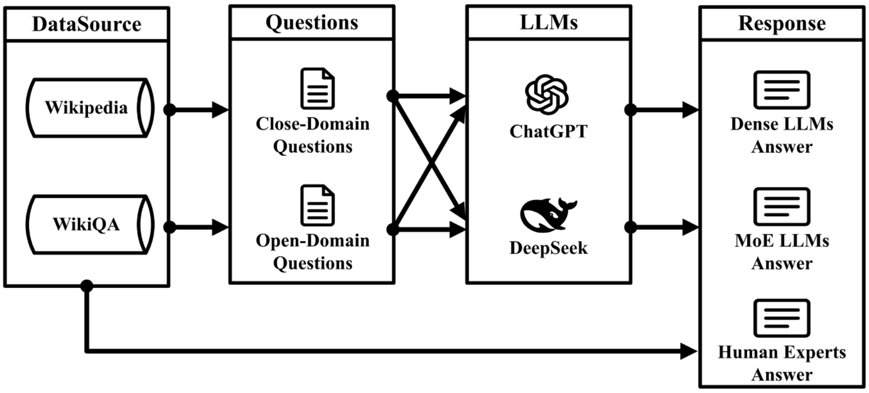

Fig. 2. Data collection and composition.

conclusions (Guo et al., 2023). Therefore, in constructing the dataset, we collected only one response per question from each large language model. This approach simplifies data processing without compromising the validity of the conclusions.

It should also be noted that the "human expert responses" used in this study refer exclusively to static textual outputs retrieved from Wikipedia entries and public Q&A platforms. These texts have been either collaboratively edited or filtered through community voting and thus represent authoritative and fixed forms of knowledge representation. Although such texts provide reliable benchmarks for comparison, they differ from the interactive processes by which human experts construct knowledge in real-world contexts. Accordingly, the focus of our comparison is on content-level quality differences among Response Sources, rather than on the dynamic processes through which human experts engage in collaborative knowledge construction. This distinction delimits the scope of the study’s conclusions: the results primarily reflect textual-level quality and do not entail claims regarding the full substitution of human expert behavior.

4.3. Indicator design and measurement

Based on the theoretical framework established in Section 2, this study quantifies both question and answer characteristics along five dimensions: complexity, specificity, information quality, information perception, and information load. All variables were computed using automated Python scripts, employing techniques such as lexical density analysis, textual diversity measurement using MTLD, LIWC-based text analysis, syntactic dependency distance computation, sentence vector similarity, and other natural language processing methods. The selected dimensions are aligned with information processing theory, levels of processing theory, and cognitive load theory, and are intended to provide a comprehensive cognitive-informational profile of knowledge construction tasks. Detailed explanations of each indicator and the corresponding computational methods are provided below.

4.3.1. Complexity

The complexity dimension captures the inherent difficulty level of each question. In this study, high-complexity questions are defined as those that are closely associated with specialized disciplinary knowledge and require domain expertise for accurate answers. Conversely, low-complexity questions are defined as those answerable using general background knowledge without requiring specialized expertise.

In the HC3 dataset, we selected questions from the wiki_csai subset, which focuses on computer science topics, as the high-complexity sample. After filtering for completeness, this subset yielded 841 questions. Examples include “Please explain what is Natural-language understanding” and “Please explain what is Information theory,” both of which involve abstract concepts and theoretical reasoning. These questions require respondents to possess substantial background knowledge.

In contrast, we selected the open_qa subset to represent low-complexity questions. After preprocessing, this subset contained 1187 questions, such as “Where are colors on stoplight?” and “How much is 1 tablespoon of water?” These questions are grounded in everyday knowledge and require only factual recall. This classification allows for a systematic contrast in knowledge demands and reasoning difficulty, providing the basis for comparative analysis in subsequent sections.

4.3.2. Specificity

Specificity measures the cognitive depth and logical complexity required to answer a question. In essence, specificity reflects the degree of reasoning, inferencing, and structured thinking expected of the responder. Drawing upon Bloom’s taxonomy of cognitive learning objectives, which links linguistic cues to cognitive processing levels (Adams, 2015; Markowitz, 2023), this study adopts a text-based operationalization of specificity using the language found in human expert answers.

Specifically, we apply the Linguistic Inquiry and Word Count (LIWC) tool to extract scores for the “Cognitive Processes” category from the human answers corresponding to each question. This score reflects the density of language indicative of reasoning and cognitive engagement (Boyd et al., 2022). A higher cognitive process score suggests that the answer involves deeper inferential reasoning, and hence the corresponding question is considered to have higher specificity. Because human expert responses are treated as authoritative and widely accepted, the cognitive process score serves as a valid proxy for the depth of cognitive processing required by each question. Accordingly, every question is assigned a specificity score based on this measure.

4.3.3. Information quality

The information quality dimension evaluates the content-level value and reliability of the generated answers. Drawing from established literature, this study disaggregates information quality into four complementary subdimensions (Lee et al., 2002): Intrinsic Information Quality, Contextual Information Quality, Representational Information Quality, and Accessibility Information Quality. The first two concern semantic-level content. Intrinsic IQ assesses factual accuracy and evidential consistency, focusing on the alignment between model output and authoritative references. Contextual IQ evaluates the semantic relevance of the response to the question and its coverage of key informational elements. Representational IQ reflects expressive adequacy, such as lexical richness and information density. Accessibility Information Quality, which reflects the ease of access and interpretability of the content, is treated as a boundary condition in this study rather than a scoring metric.

Since the human expert answers in the HC3 dataset are derived from reputable sources such as Wikipedia and high-vote Q&A responses, they are treated as “gold standards.” These texts are presumed to be factually accurate and semantically complete. Accordingly, the evaluation of LLM-generated answers is performed relative to these benchmarks. For both Intrinsic IQ and the Completeness subdimension of Contextual IQ, human expert answers are assigned perfect scores. LLM outputs are scored based on their degree of divergence from these standards.

Quantification of the three scored subdimensions proceeds as follows:

(1) Intrinsic IQ: This dimension captures factual accuracy. Given the difficulty of verifying open-domain answers in a fully automated manner (McIntosh et al., 2025), we adopt a sentence-level natural language inference (NLI) approach using authoritative sources. Each LLM-generated answer and its corresponding human reference are split into individual sentences and encoded using Sentence-BERT. For every sentence in the model response, the three most semantically similar reference sentences are retrieved using cosine similarity. Each model sentence is treated as a hypothesis, and each reference sentence as a premise. These are passed through a RoBERTa-large-MNLI classifier to generate entailment probabilities. The maximum entailment score among the three candidates is taken as the support score for that sentence. The average of these scores across all sentences constitutes the Intrinsic IQ score, computed as:

$$\text{Intrinsic IQ} = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} \mathrm{p}_{\mathrm{entail}}^{(\mathrm{i})} \tag{1}$$

where N denotes the number of sentences in the response and $\mathrm{p}_{\mathrm{entail}}^{(\mathrm{i})}$ is the highest entailment probability for sentence i. Higher scores indicate stronger alignment with authoritative evidence and greater factual reliability.
(2) Contextual IQ: This dimension includes both Relevance and Completeness. Relevance measures semantic alignment between the question and the answer. Using Sentence-BERT, we encode both the question and the answer and calculate cosine similarity. The score is normalized to the range [0, 1] using the following formula:

$$\text{Relevance} = \frac{1 + \cos(\mathbf{q}, \mathbf{a})}{2} \tag{2}$$

where q and a denote the semantic vectors of the question and answer, respectively. Higher scores indicate closer alignment to the question's core intent.

Completeness measures how thoroughly the model answer covers the key informational points in the human reference. We extract noun phrases and named entities from the reference answer to form a keypoint set K. We then count how many of these elements are matched in the model answer (M), and compute:

$$\text{Completeness} = \frac{\mathrm{M}}{\mathrm{K}} \tag{3}$$

This reflects the proportion of essential content included in the model response.
(3) Representational IQ:

This dimension is measured via lexical diversity and lexical density. For lexical diversity, we use the Measure of Textual Lexical Diversity (MTLD), which is more robust than traditional type-token ratios(Torruella and Capsada, 2013).

Assume a text of N tokens: $W = \{\mathrm{w}_1, \mathrm{w}_2, \ldots, \mathrm{w}_\mathrm{N}\}$. For any subsequence $\mathrm{S} \subseteq \mathrm{W}$ of length n, the type-token ratio (TTR) is defined as:

$$\mathrm{TTR}(\mathbf{S}) = \frac{\mathrm{V}(\mathbf{S})}{\mathrm{n}} \tag{4}$$

where V(S) is the number of unique word types in S. The MTLD score is then calculated as the number of such segments (factors) required before TTR drops below a threshold of 0.72. If F is the number of such segments, then:

$$\mathrm{MTLD} = \frac{\mathrm{N}}{\mathrm{F}} \tag{5}$$

To improve reliability, MTLD is computed bi-directionally (left-to-right and right-to-left), and the final score is the average of both directions:

$$\mathrm{MTLD}_{\mathrm{final}} = \frac{1}{2}\left(\frac{\mathrm{N}}{\mathrm{F}_{\mathrm{forward}}} + \frac{\mathrm{N}}{\mathrm{F}_{\mathrm{backward}}}\right) \tag{6}$$

This captures the richness and stability of vocabulary usage in the response. In parallel, lexical density is computed as the proportion of content words (nouns, verbs, adjectives, adverbs) over total tokens to reflect information compression and semantic density (Halliday, 1985).
(4) Accessibility IQ: This dimension is not scored numerically but acts as a design constraint to ensure that all content compared in this study originates from publicly accessible, structurally coherent, and linguistically standard sources. As such, all evaluated answers are based on Wikipedia entries or widely endorsed Q&A responses, and all questions are derived from standardized corpora using plain, non-technical language. Therefore, no Accessibility IQ score is assigned; instead, it guarantees that information quality comparisons are made under equivalent conditions.

4.3.4. *Information perception*

Information perception captures the stylistic affinity and structural clarity of model-generated responses. This dimension reflects the extent to which a response is readable, emotionally engaging, and cognitively coherent for the reader. It pertains to the perceived humanness of the text, including how easily it can be understood and how effectively it elicits sensory and emotional resonance. From the perspective of cognitive processing, a more vivid and affectively rich style can foster greater emotional and cognitive engagement, while structurally clear texts reduce comprehension effort and facilitate more efficient information assimilation. Drawing upon relevant theories and prior empirical work, this study operationalizes information perception along two sub-dimensions.

Linguistic affinity is measured by examining the degree to which the response exhibits characteristics typically associated with human language. This includes stylistic, sensory, motivational, and emotional features. Specifically, four dimensions drawn from the LIWC dictionary are employed: Informal Language, Perceptual Processes, Drives, and Affect. Informal Language includes colloquialisms, contractions, and internet slang, indicating a casual and accessible tone. Perceptual Processes comprises terms related to sensory experience, such as those referring to sight, sound, or touch, reflecting the extent to which the text activates the reader's sensory imagination. The Drives category includes words related to achievement, power, and curiosity, suggesting motivational or value-laden content that may prompt reflection. Affect captures the use of emotionally charged words, both positive and negative, and reflects the emotional expressiveness of the response. The standardized scores of these four dimensions are aggregated to produce a composite linguistic affinity score. Higher values in this dimension indicate that the response contains more vivid and human-like expressions, which can enhance reader engagement and resonance. Linguistic affinity thus serves as a proxy for perceived humanness and emotional appeal.

Structural clarity focuses on the internal coherence and logical organization of the response. It is assessed using two complementary metrics: connective density and adjacent overlap. Connective density measures the proportion of explicit logical connectors in the response. This is defined as the number of tokens that match a predefined list of conjunctions and transitional markers (such as "and", "but", "because", "however"), divided by the total number of tokens. The formula is given as follows:

$$\mathrm{ConnectiveDensity} = \frac{\mathrm{N_{conn}}}{\mathrm{N_{tokens}}} \tag{7}$$

where $\mathrm{N_{conn}}$ denotes the number of identified connective words, and $\mathrm{N_{tokens}}$ is the total number of tokens in the response. A higher connective density suggests clearer articulation of logical relationships between clauses and sentences, thereby improving structural transparency and reader comprehension.

Adjacent overlap measures local coherence between consecutive sentences in a response. Each sentence is represented as a set of content words (with stop words removed). For any adjacent sentence pair $\mathrm{S_i}$ and S_{i+1}, the Jaccard similarity is computed as:

$$\mathrm{J}(\mathrm{S_i}, \mathrm{S}_{i+1}) = \frac{|\mathrm{S_i} \cap \mathrm{S}_{i+1}|}{|\mathrm{S_i} \cup \mathrm{S}_{i+1}|} \tag{8}$$

The overall adjacent overlap score is the average of these similarities across all sentence pairs:

$$\mathrm{AdjacentOverlap} = \frac{1}{M-1}\sum_{i=1}^{M-1} \mathrm{J}(\mathrm{S_i}, \mathrm{S}_{i+1}) \tag{9}$$

where M is the total number of sentences in the response. If $M < 2$, the score is assigned a value of zero, indicating the absence of coherence indicators in single-sentence responses. Higher values indicate stronger thematic continuity and tighter conceptual links between sentences, which improve local coherence and facilitate reader comprehension. This metric reflects the tendency of human-written paragraphs to maintain topic continuity through lexical repetition or anaphoric reference.

Together, the dimensions of linguistic affinity and structural clarity form a composite measure of information perception. This score captures the degree to which the model output resembles human-authored text in style and structure. Higher information perception scores indicate that the response is more engaging in tone and clearer in organization, which may lead to more active information processing. This metric complements the previously described dimensions of information quality and information load to form a comprehensive framework for evaluating knowledge construction performance.

4.3.5. *Information load*

The dimension of information load evaluates the cognitive effort required to read and comprehend a response. According to cognitive load theory (Chandler and Sweller, 1991), the format and complexity of information presentation directly influence the amount of cognitive resources consumed by the learner. Structurally complex or verbose content increases extraneous cognitive load and may impede efficient knowledge construction. Based on this theoretical foundation, information load is measured using two indicators: syntactic complexity and textual length.

Syntactic complexity is operationalized using average dependency distance. Dependency distance refers to the linear distance, in terms of word positions, between a head word and its dependent modifier within a syntactic dependency structure. Shorter distances typically correspond to simpler sentence structures in which modifiers appear adjacent to their heads. Longer distances reflect more complex constructions such as nested clauses or long-distance dependencies. Psycholinguistic studies have shown that longer dependency distances place greater demands on working memory and increase parsing difficulty (Futrell et al., 2015). For each response, the average dependency distance is calculated across all sentences. Higher values indicate more complex syntactic structures, which correspond to increased intrinsic cognitive load.

Textual length is used to capture the extraneous cognitive effort associated with processing large volumes of information. While information richness and length are not always linearly correlated, longer responses generally require more reading and impose a higher processing burden. The number of tokens in each response is used as a proxy for this aspect of information load. Although length does not distinguish between informative and redundant content, it provides a basis for comparing tendencies toward verbosity or conciseness across different Response Sources. Human experts may prefer concise and focused responses, while LLMs often produce longer outputs, potentially increasing unnecessary cognitive load. The total token count is recorded for each response to reflect this difference.

Together, average dependency distance and response length provide a dual perspective on the information load imposed by each response. For comparative analysis, both indicators are standardized and integrated into a single composite index of information load. Higher scores reflect greater demands on the reader's attention and memory resources during information processing.

4.4. Data processing

Following the computation of the aforementioned indicators, the resulting data were subjected to further integration and processing. Specifically, based on the procedures outlined in Section 4.3, each question was assigned scores for complexity and specificity, while each response was evaluated along the three primary dimensions of information quality, information perception, and information load. This yielded a total of 2028 questions and 6084 corresponding responses with complete feature profiles. Subsequently, the entropy weight method was employed to derive composite evaluation indices for use in later analytical stages.

The entropy weight method is an objective weighting technique grounded in the concept of information entropy. The central premise is that indicators with more dispersed distributions carry greater informational value and should thus be assigned higher weights in composite evaluations (Gray, 2011). Conversely, indicators with limited variance across samples convey less new information and should have lower weights. This method provides a systematic approach to weight determination in multi-indicator evaluation frameworks and aligns with foundational principles of information theory concerning uncertainty (Gray, 2011; Zhang et al., 2022).

In this study, the entropy weight method was applied to the secondary-level sub-indicators under the three primary evaluation dimensions: information quality (IQ), information perception (IP), and information load (IL). These weights were then used to construct the composite scores for each dimension. The specific computational steps are as follows.

First, the raw data were normalized using min–max scaling to eliminate the effects of differing measurement units. Next, the proportion p_{ij} of each sample under each indicator was calculated to reflect the contribution of the i th sample to the j-th indicator. This is defined as:

$$p_{ij} = \frac{x'_{ij}}{\sum_{i=1}^{n} x'_{ij}} \tag{10}$$

where x'_{ij} denotes the normalized value of the i th sample on the j-th indicator. Based on this, the entropy value e_j of each indicator was computed using the following formulation:

$$e_j = -\frac{1}{\ln n}\sum_{i=1}^{n} p_{ij} \ln p_{ij} \tag{11}$$

where n represents the total number of samples. In cases where $p_{ij} = 0$, the corresponding term in the summation is conventionally assigned a value of zero to ensure mathematical validity.

The relative entropy values among sub-indicators within the same primary dimension were then used to determine the corresponding weights (Gray, 2011; Zhang et al., 2022). Information entropy quantifies the uncertainty of a source, and smaller entropy values indicate higher concentration and variability in the data, thus warranting a greater influence on the overall evaluation. The weight w_j of each indicator was computed as follows:

Table 2
Entropy and weight of indicators.

Primary Indicators	Secondary indicators	Entropy	Weight
IQ_Int	Intrinsic IQ	0.946	1.000
IQ_Rep	Lexical Density	0.997	0.131
	MTLD	0.980	0.869
IQ_Ctx	Relevance	0.999	0.046
	Completeness	0.981	0.954
IP_LA	Informal Language	0.798	0.460
	Perceptual Processes	0.880	0.273
	Drives	0.947	0.119
	Affect	0.935	0.148
IP_SC	Connective Density	0.981	0.235
	Adjacent Overlap	0.939	0.765
IL	Avg. Dependency Distance	0.989	0.239
	Answer Length	0.965	0.761

$$w_j = \frac{1 - e_j}{\sum_{k=1}^{m} (1 - e_k)} \tag{12}$$

where m is the total number of indicators included in the calculation. The resulting entropy values and corresponding weights derived through this procedure are presented in Table 2.

Based on the weights reported in Table 2, the six major evaluation indicators were normalized and integrated to obtain composite scores for each response across the three primary dimensions. The resulting analytical dataset is summarized in Table 3.

5. Results

5.1. Descriptive analysis

Under different levels of task complexity, the evaluation outcomes of distinct response sources, including LLMs of divergent architectures and human experts, demonstrate significant variation across three overarching dimensions—Information Quality, Information Perception, and Information Load—and six specific indicators: Intrinsic IQ, Representational IQ, Contextual IQ, Linguistic Affinity, Structural Clarity, and Information Load. To provide an intuitive visualization of this trend, Fig. 3 illustrates the overall distribution of performance traits.

Overall, the three types of response sources reveal evident distributional differences across multiple dimensions, indicating distinct expressive strategies and performance boundaries in response to varying levels of task complexity. From the perspective of task differentiation, the transition from open-domain to closed-domain question settings leads to a general increase in five out of the six indicators, with the exception of Contextual IQ. This upward shift suggests that increased task complexity imposes greater demands on content coverage, linguistic expression, and structural organization, prompting adaptive changes in the expressive strategies of different response sources. The observed decline in Contextual IQ within closed-domain settings may be attributed to the high requirements for contextual alignment and key information coverage embedded in this dimension. Complex tasks often feature condensed semantic contexts and stronger background constraints, which expose deficiencies in topic alignment and coverage precision, particularly for LLM-generated responses. This result further confirms the moderating effect of task complexity on model performance.

In the domain of Information Quality, the distribution of Intrinsic IQ scores, which reflect evidence alignment, shows that human responses consistently achieve near-perfect scores with minimal variance, consistent with their designation as the reference standard. Both MoE and Dense models perform below the human baseline, but their medians and upper bounds increase under high-complexity conditions, especially for MoE, suggesting enhanced capacity for evidence alignment in more constrained tasks. For Contextual IQ, which captures relevance and coverage, human responses again cluster at the maximum value. MoE outperforms Dense in both complexity groups, with further improvements observed under complex conditions, indicating its superior ability to align with contextual cues and cover core content. Regarding Representational IQ, MoE demonstrates a clear advantage, especially under complex tasks, exhibiting greater expressive range and lexical diversity, as evidenced by its wider interquartile range and whisker span. In contrast, the Dense model maintains a lower but more concentrated distribution across all three IQ indicators, reflecting its tendency toward stable and conservative outputs.

The dimension of Information Perception reveals perceptible differences in language style and structural organization. On the Linguistic Affinity indicator, MoE scores significantly higher than both Dense and Human responses, particularly in the closed-domain group, where both the median and range expand. This suggests frequent use of emotionally expressive, sensory-rich, and motivational language by the MoE model, enhancing its perceived human-likeness. Human responses, by contrast, remain consistently low in this dimension, reflecting a restrained and objective style typical of expert-authored texts in knowledge construction contexts. Structural Clarity exhibits an opposite ranking: Dense outperforms the other two sources, indicating a strong reliance on explicit structural markers to enhance clarity. While MoE leads in expressive richness, it is comparatively weaker in structural organization, occasionally lacking local coherence.

In terms of Information Load, the median and whisker span for MoE increase notably in the closed-domain group, suggesting that its intensified expression and rich content generation impose higher cognitive demands. Dense maintains a relatively moderate load, with minimal fluctuation across task complexities. Human responses exhibit consistently low load and variance in both conditions, particularly excelling in simple tasks by achieving high-quality expression at minimal cognitive cost, showcasing an advantage in simplification strategies and cognitive efficiency.

Table 3
Example Results of indicators.

Question	Source	Complexity	Specificity	IQ_Int	IQ_Rep	IQ_Ctx	IP_LA	IP_SC	IL
1	Human	1.000	0.262	0.998	0.308	1.000	0.008	0.204	0.140
1	Dense	1.000	0.262	0.285	0.211	0.326	0.049	0.154	0.172
1	MoE	1.000	0.262	0.155	0.487	0.370	0.059	0.104	0.264
2	Human	1.000	0.302	0.999	0.247	0.999	0.056	0.198	0.172
2	Dense	1.000	0.302	0.528	0.175	0.326	0.061	0.129	0.113
2	MoE	1.000	0.302	0.303	0.355	0.368	0.094	0.097	0.340

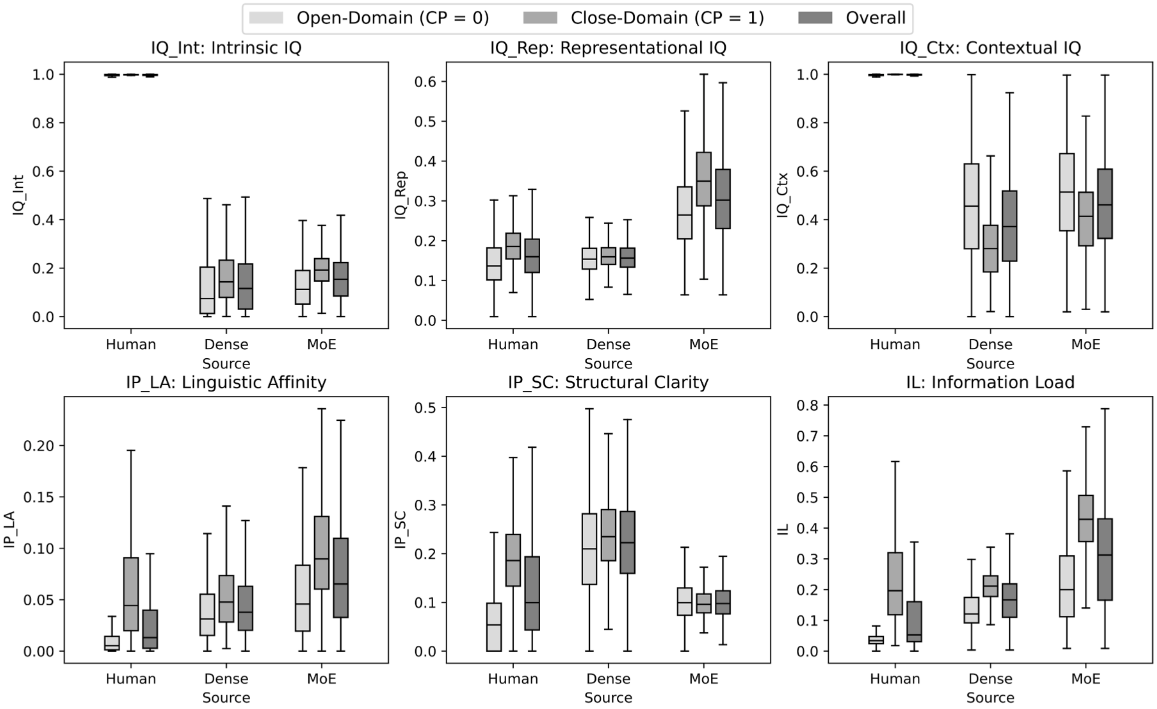

Fig. 3. Traits distribution of LLMs and human expert responses.

In summary, the MoE architecture demonstrates clear strengths in Information Quality and Linguistic Affinity, making it suitable for complex tasks requiring broad coverage and expressive flexibility, albeit at the cost of reduced structural clarity and increased cognitive burden. The Dense model, by contrast, emphasizes structural stability and logical clarity, producing more predictable outputs with less stylistic variation. Human experts maintain a distinct advantage in simple tasks, achieving an optimal balance of low cognitive load and high content quality—a combination still unmatched by either LLM architecture. Fig. 3 highlights not only the focus areas of each response type across the information dimensions but also illustrates how task complexity modulates performance traits, suggesting that future implementations may benefit from task-aligned model selection and the use of structured rewriting or summarization to improve the usability and cognitive accessibility of MoE outputs.

Table 4 provides descriptive statistics of responses from the three sources across the three major dimensions—Information Quality, Information Perception, and Information Load—under both task settings.

In contrast to the macro-level insights offered above, Fig. 4 presents more fine-grained comparisons at the level of individual questions. This figure displays the relative rankings of the three response sources across six specific evaluation indicators, under three task groupings: Open-Domain, Closed-Domain, and Overall. The six indicators include three subdimensions of Information Quality (Accuracy, Contextual Relevance, and Representational Expression), two subdimensions of Information Perception (Linguistic Affinity and Structural Clarity), and Information Load. In this analysis, a "High" label indicates that a given response type scored the highest on that indicator for a specific question, while "Low" signifies the lowest score. These results help uncover the relative strengths of different response sources in concrete knowledge construction tasks and provide empirical validation for the consistency and generalizability of the earlier descriptive patterns at the question level.

Table 5 reveals a structured distribution of subdimension performance across response types, further supporting the discriminative power of the proposed framework in identifying Response Traits.

It is important to note that Intrinsic IQ and Contextual IQ were scored with reference to human-generated answers as the gold standard. Consequently, human responses show extremely high consistency in these dimensions, with 99.75 % and 98.92 % of responses classified as High, respectively. This outcome reflects the annotation framework rather than serving as a direct indicator of comparative performance. When human responses are excluded, MoE outperforms Dense in both Intrinsic and Contextual IQ.

In the dimension of Representational IQ, the MoE architecture performs particularly well, with 89.69 % of responses rated as High and nearly reaching saturation under closed-domain conditions (95.72 %). This indicates superior capacity for expressive richness and information density. Representational IQ incorporates both lexical density and diversity. MoE's strong performance in this dimension reflects a high level of linguistic representativeness in its generated text. In contrast, Dense and Human responses show much lower High-rank percentages, at 2.91 % and 7.4 %, respectively, suggesting a more limited capacity for linguistic variety and expressiveness.

In the domain of Information Perception, MoE and Dense exhibit divergent patterns. MoE achieves 65.30 % High-rank responses on Linguistic Affinity, significantly higher than Dense (23.29 %) and Human (11.40 %). This indicator draws on LIWC-defined categories such as informal expression, affective vocabulary, perceptual verbs, and motivational terms. A high proportion reflects a tendency

Table 4
Descriptive statistics of data.

Variable	Complexity	Source	N	Mean	Std. Dev.	Min	Max
IQ_Int	Close-Domain (CP=1)	Human	841	0.995	0.014	0.803	1
		Dense	841	0.162	0.112	0.001	0.637
		MoE	841	0.195	0.072	0.002	0.434
	Open-Domain (CP=0)	Human	1187	0.995	0.011	0.757	1
		Dense	1187	0.136	0.169	0	0.998
		MoE	1187	0.139	0.129	0	0.976
	Overall	Human	2028	0.995	0.013	0.757	1
		Dense	2028	0.147	0.149	0	0.998
		MoE	2028	0.162	0.112	0	0.976
IQ_Ctx	Close-Domain (CP=1)	Human	841	0.996	0.012	0.862	1
		Dense	841	0.303	0.143	0.021	0.895
		MoE	841	0.404	0.154	0.03	0.994
	Open-Domain (CP=0)	Human	1187	0.993	0.021	0.678	1
		Dense	1187	0.464	0.241	0	0.998
		MoE	1187	0.519	0.232	0.02	0.996
	Overall	Human	2028	0.994	0.018	0.678	1
		Dense	2028	0.397	0.221	0	0.998
		MoE	2028	0.471	0.211	0.02	0.996
IQ_Rep	Close-Domain (CP=1)	Human	841	0.19	0.052	0.07	0.429
		Dense	841	0.162	0.034	0.072	0.419
		MoE	841	0.358	0.107	0.103	0.954
	Open-Domain (CP=0)	Human	1187	0.153	0.082	0.009	0.738
		Dense	1187	0.157	0.043	0.04	0.38
		MoE	1187	0.279	0.102	0.064	0.802
	Overall	Human	2028	0.169	0.073	0.009	0.738
		Dense	2028	0.159	0.039	0.04	0.419
		MoE	2028	0.312	0.111	0.064	0.954
IP_LA	Close-Domain (CP=1)	Human	841	0.065	0.065	0	0.476
		Dense	841	0.058	0.043	0.002	0.33
		MoE	841	0.105	0.067	0	0.536
	Open-Domain (CP=0)	Human	1187	0.011	0.016	0	0.208
		Dense	1187	0.041	0.037	0	0.294
		MoE	1187	0.06	0.055	0	0.399
	Overall	Human	2028	0.033	0.051	0	0.476
		Dense	2028	0.048	0.04	0	0.33
		MoE	2028	0.079	0.064	0	0.536
IP_SC	Close-Domain (CP=1)	Human	841	0.19	0.084	0	0.678
		Dense	841	0.244	0.087	0.045	0.648
		MoE	841	0.102	0.034	0.037	0.271
	Open-Domain (CP=0)	Human	1187	0.078	0.099	0	0.761
		Dense	1187	0.22	0.116	0	0.815
		MoE	1187	0.114	0.072	0	0.598
	Overall	Human	2028	0.125	0.108	0	0.761
		Dense	2028	0.23	0.105	0	0.815
		MoE	2028	0.109	0.06	0	0.598
Information Load	Close-Domain (CP=1)	Human	841	0.231	0.149	0.018	0.761
		Dense	841	0.21	0.056	0.029	0.413
		MoE	841	0.435	0.112	0.108	0.788
	Open-Domain (CP=0)	Human	1187	0.039	0.023	0	0.248
		Dense	1187	0.135	0.065	0.003	0.654
		MoE	1187	0.223	0.132	0.009	0.74
	Overall	Human	2028	0.119	0.136	0	0.761
		Dense	2028	0.166	0.072	0.003	0.654
		MoE	2028	0.311	0.162	0.009	0.788

toward emotional and human-like language. However, in Structural Clarity, MoE performs relatively poorly, with only 9.68 % of responses rated as High, compared to 72.87 % for Dense, suggesting that Dense responses exhibit stronger logical coherence and structural transparency. This indicator incorporates connective density and adjacent sentence overlap, both of which contribute to perceived clarity and continuity.

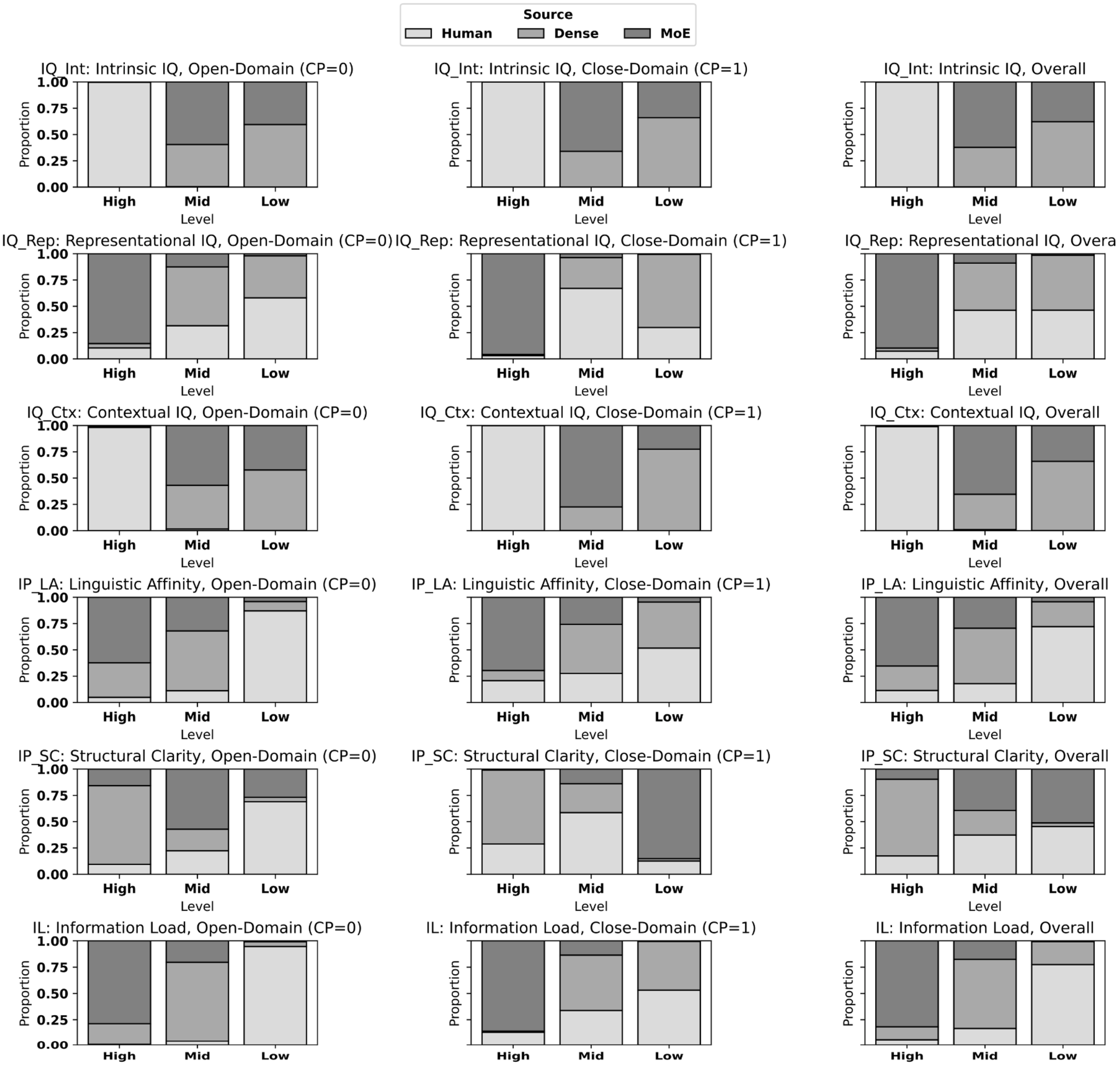

Fig. 4. Comparison of data level from different questionand different sources.

On the Information Load dimension, 81.71 % of MoE responses are classified as High, significantly surpassing Dense (12.38 %) and Human (5.92 %). This metric is based on average dependency distance in syntactic parsing, where higher values imply more complex sentence structures and greater cognitive demand. Dense responses concentrate in the Medium range (65.78 %), while Human responses show a strong tendency toward the Low category (77.42 %), reflecting a deliberate effort to compress information and reduce cognitive effort in their expression.

Across different task conditions, these trends remain consistent, with closed-domain tasks amplifying certain differences. For example, excluding human responses, the MoE architecture's advantage over Dense in Intrinsic and Contextual IQ becomes more pronounced. MoE also improves further in Representational IQ and Linguistic Affinity under closed-domain settings, reaching 95.72 % and 69.60 %, respectively. However, its High-rank proportion for Structural Clarity drops to 0.95 %, while its Low-rank proportion rises to 85.14 %. Information Load remains high, confirming its tendency to generate dense and complex content under specialized conditions. Dense maintains strong performance in Structural Clarity, particularly under closed-domain tasks, and does not exhibit a significant rise in Information Load. Human responses sustain high performance in accuracy and contextual alignment, while achieving 28.78 % High-rank in Structural Clarity and remaining predominantly in the Low range for Information Load, indicating strong consistency and compression capacity.

In conclusion, the MoE architecture excels in expressive and affective dimensions, Dense provides structural consistency and logical clarity, and human responses continue to lead in accuracy, contextual relevance, and cognitive efficiency.

Table 5
Distribution of high/mid/low proportions across question, features, and sources.

Question	Feature	Source	High (%)	Mid (%)	Low (%)
Open-Domain	IQ_Rep	Moe	85.43	12.54	2.02
		Dense	4.13	55.89	39.97
		Human	10.45	31.57	58.01
	IQ_Int	Moe	0	59.48	40.52
		Dense	0.42	40.1	59.48
		Human	99.58	0.42	0
	IQ_Ctx	Moe	0.84	56.87	42.29
		Dense	0.93	41.53	57.54
		Human	98.23	1.6	0.17
	IP_SC	Moe	15.83	57.14	26.84
		Dense	74.71	20.55	4.19
		Human	9.46	22.31	68.97
	IP_LA	Moe	62.31	31.96	4.04
		Dense	32.81	56.9	8.86
		Human	4.88	11.15	87.11
	IL	Moe	78.77	20.39	0.84
		Dense	20.39	75.06	4.55
		Human	0.84	4.55	94.61
Close-Domain	IQ_Rep	Moe	95.72	3.69	0.59
		Dense	1.19	29.25	69.56
		Human	3.09	67.06	29.85
	IQ_Int	Moe	0	65.99	34.01
		Dense	0	34.01	65.99
		Human	100	0	0
	IQ_Ctx	Moe	0.12	77.41	22.47
		Dense	0	22.47	77.53
		Human	99.88	0.12	0
	IP_SC	Moe	0.95	13.91	85.14
		Dense	70.27	27.47	2.26
		Human	28.78	58.62	12.6
	IP_LA	Moe	69.6	25.71	4.54
		Dense	9.62	46.68	43.73
		Human	20.78	27.61	51.73
	IL	Moe	85.85	13.56	0.59
		Dense	1.07	52.68	46.25
		Human	13.08	33.77	53.15
OVERALL	IQ_Rep	Moe	89.69	8.87	1.43
		Dense	2.91	44.85	52.24
		Human	7.4	46.28	46.32
	IQ_Int	Moe	0	62.18	37.82
		Dense	0.25	37.57	62.18
		Human	99.75	0.25	0
	IQ_Ctx	Moe	0.54	65.38	34.07
		Dense	0.54	33.63	65.83
		Human	98.92	0.99	0.1
	IP_SC	Moe	9.68	39.3	51.22
		Dense	72.87	23.41	3.38
		Human	17.44	37.29	45.4
	IP_LA	Moe	65.3	29.39	4.25
		Dense	23.29	52.7	23.62
		Human	11.4	17.91	72.13
	IL	Moe	81.71	17.55	0.74
		Dense	12.38	65.78	21.84
		Human	5.92	16.67	77.42

5.2. Hierarchical linear regression analysis

To further examine how response sources and question cues influence the multidimensional features of generated answers, this study constructed separate regression models using six dependent variables: Intrinsic IQ, Representational IQ, Contextual IQ, Linguistic Affinity, Structural Clarity, and Information Load. The independent variables include response source, with human experts as the baseline, question complexity and specificity, as well as their interaction terms. All dependent and continuous variables were standardized prior to analysis. Cluster-robust standard errors with CR2 correction and Satterthwaite degrees of freedom were applied, clustered by question ID. The analysis was conducted using R 4.3.3. The results are presented in Table 6.

The first block of regression results reveals the distinct impacts of response sources on six answer traits. Compared to human-generated responses, the Dense model exhibited significantly lower performance in accuracy (Intrinsic IQ, β = −0.973***) and contextual relevance (Contextual IQ, β = −0.882***), along with a slight negative coefficient in representational quality

Table 6
Associations between question cues and response traits ($N = 6084$).

	IQ_Int	IQ_Rep	IQ_Ctx	IP_LA	IP_SC	IL
Block 1: Responses sources (H1)						
Human experts (reference)	—	—	—	—	—	—
LLM: Dense (GPT)	−0.973***	−0.043***	−0.882***	0.123***	0.459***	0.146***
LLM: MoE (DS)	−0.956***	0.635***	−0.773***	0.384***	−0.068***	0.593***
Block 2: Question traits (H2)						
Complexity (0 = Open, 1 = Close)	0.031***	0.145***	−0.111***	0.153***	0.043**	0.269***
Specificity	0.003	0.059***	−0.043***	0.272***	0.207***	0.355***
Block 3: Response sources × Question traits (H3)						
Dense × Complexity	0.017***	−0.075***	−0.092***	−0.166***	−0.230***	−0.182***
MoE × Complexity	0.026***	0.033*	−0.070***	−0.134***	−0.183***	−0.065***
Dense × Specificity	−0.004	0.009	−0.041***	0.019	0.061***	0.005
MoE × Specificity	0.009**	0.084***	−0.024**	0.135***	−0.121***	0.136***
Model Summary						
Adjusted R^2	0.933	0.491	0.726	0.297	0.373	0.665
RMSE	0.26	0.71	0.52	0.84	0.79	0.58

Note: Values represent standardized β coefficients; significance based on cluster-robust SE (CR2) by question ID.
***$p < .001$, **$p < .01$,*$p < .05$.

(Representational IQ, β = −0.043***). However, it showed significant positive effects in linguistic affinity (Linguistic Affinity, β = 0.123***) and structural clarity (Structural Clarity, β = 0.459***), suggesting that this model tends to structure its responses in an orderly and human-like linguistic style. This also coincides with a moderate increase in information load (Information Load, β = 0.146***). The MoE model demonstrated a stronger positive impact on representational quality (β = 0.635***) and linguistic affinity (β = 0.384***), and its coefficient for information load was the highest among the three sources (β = 0.593***), indicating richer expression at a higher cognitive processing cost. For structural clarity, MoE yielded a slight negative effect (β = −0.068***), though still statistically significant. Given that both Dense and MoE models outperform humans in linguistic affinity and information load, and that MoE additionally shows partial advantages in information quality, particularly on representational quality, Hypothesis 1 is supported.

The second block of analysis examined the main effects of question cues. Question complexity exerted a significant positive effect on most dependent variables, especially representational quality (β = 0.145***), linguistic affinity (β = 0.153***), and information load (β = 0.269***). It also showed a slight positive effect on accuracy (β = 0.031***), but a negative effect on contextual alignment (β = −0.111***), indicating that while complex questions stimulate deeper content generation and richer language expression, they may also reduce alignment between the response and the intended context. For structural clarity, complexity had a small but significant positive coefficient (β = 0.043**), suggesting that high-complexity tasks promote more organized response structures. Specificity had a significant positive impact on all dimensions except accuracy, with pronounced effects on linguistic affinity (β = 0.272***) and information load (β = 0.355***), implying that more specific prompts tend to elicit responses containing emotional, sensory, and motivational elements, while also increasing syntactic complexity. The positive effects on structural clarity (β = 0.207***) and representational quality (β = 0.059***) further indicate that specificity enhances content organization and linguistic richness. These findings provide strong support for Hypothesis 2 across all dimensions except Intrinsic IQ.

The third block of interaction analysis further reveals how the impact of question cues differs across response sources. For the Dense model, question complexity showed significantly negative interaction effects on contextual relevance (β = −0.092***), linguistic affinity (β = −0.166***), structural clarity (β = −0.230***), and information load (β = −0.182***), suggesting that under complex task conditions, this model's advantages in linguistic clarity and coherence are weakened, and its cognitive burden increases further. The MoE model exhibited similar negative interaction effects on linguistic affinity (β = −0.134***), structural clarity (β = −0.183***), and information load (β = −0.065***), indicating comparable degradation under complexity. However, it showed a positive interaction for representational quality (β = 0.033*) and a slight positive interaction for accuracy (β = 0.026***), suggesting that MoE may possess stronger language expansion and reasoning capabilities in complex tasks. Regarding specificity, the Dense model did not show significant interactions for any indicators, indicating a relatively stable response mechanism. In contrast, the MoE model showed significant positive interactions with representational quality (β = 0.084***), linguistic affinity (β = 0.135***), and information load (β = 0.136***), demonstrating its heightened sensitivity to specificity cues. This suggests that MoE can capture latent semantic information within the prompt, generating more stylistically rich and detail-oriented responses, though at the cost of increased processing demands.

In summary, the Dense and MoE models exhibit significantly different profiles across response traits. The Dense model emphasizes structural clarity and consistency, while the MoE model favors expressive richness and human-likeness, with both models generating higher syntactic load than human experts. Under complex task conditions, both models show weakened contextual alignment and language control, though the MoE model demonstrates superior responsiveness when task specificity is high. These findings confirm that question cues significantly influence response performance, and their interaction with model architecture modulates the generative traits of output. Therefore, Hypothesis 3 is also supported.

6. Discussion

6.1. Key findings

Drawing upon a composite perspective of information processing theory, levels of processing theory, and cognitive load theory, this study conducted a multidimensional comparison between human expert responses and outputs generated by two prevailing large language model (LLM) architectures—dense architecture (Dense) and sparsely gated mixture-of-experts architecture (MoE)—within the context of knowledge construction tasks. The comparative framework encompassed three core dimensions: information quality, information perception, and information load. Moreover, the moderating roles of problem cues, including complexity and specificity, as well as their interactions with response sources, were systematically examined. The principal findings are summarized as follows:

First, with respect to the main effects of response sources, the MoE model demonstrated clear advantages over both Dense models and human responses in terms of representational quality (IQ_Rep) and linguistic affinity (IP_LA), indicating its superior capability to generate content characterized by linguistic richness, emotional valence, and sensory features. By contrast, the Dense model exhibited superior performance in structural clarity (IP_SC), reflecting its tendency to produce logically organized and explicitly cohesive content. Both models exhibited positive coefficients for information load (IL), with the MoE model inducing a higher cognitive burden, thereby suggesting that its expressive superiority is accompanied by increased processing demands. On the dimensions of intrinsic accuracy (IQ_Int) and contextual alignment (IQ_Ctx), both model types displayed deviations from the human-anchored gold standard, indicating persistent gaps in factual alignment and semantic coherence. Taken together, the MoE model appears to be more effective in content generation and expressive articulation, the Dense model emphasizes structural order and communicative clarity, while human responses maintain comparative advantages in information density regulation and factual correctness.

Second, with respect to the main effects of problem cues, problem complexity exerted significantly positive effects across most dependent variables, most notably in representational quality, linguistic affinity, and information load, indicating that complex tasks tend to elicit outputs that are richer and more stylized, but also cognitively demanding. In structural clarity, complexity also yielded positive effects, underscoring the importance of structural organization in responding to cognitively demanding tasks. However, on the dimension of contextual alignment (IQ_Ctx), the coefficient was negative, suggesting that when tasks involve interdisciplinary reasoning, ambiguity, or open-ended prompts, responses may deviate from the core semantic focus. Problem specificity exhibited significantly positive effects on linguistic affinity, structural clarity, and information load, but not on intrinsic accuracy or contextual alignment. This implies that highly specific and cue-rich problems stimulate the generation of more human-like and structurally explicit language, yet do not necessarily enhance consistency or coverage in terms of knowledge content. Consequently, in responding to highly specific problems, it is essential to emphasize structural and stylistic control for effective information guidance, rather than focusing solely on surface coverage. These findings are consistent with existing research conducted under dense architectures, which have also affirmed the critical role of complexity in shaping knowledge construction outcomes (Suárez et al., 2024; Wu & Zheng, 2025).

Third, the interaction effects revealed systematic interactions between response sources and problem complexity. As complexity increased, the advantages of LLMs in linguistic affinity, structural clarity, and information load were generally attenuated. For instance, the Dense model exhibited significantly negative interactions with complexity on representational quality and structural clarity, indicating that its structural strengths diminish under higher task complexity. On contextual alignment, both models displayed negative interactions with complexity, suggesting that semantic precision becomes more vulnerable in complex tasks. However, on representational quality, the interaction between complexity and the MoE model was positive, indicating its capacity for linguistic elaboration and contextual integration, which helps maintain output quality under demanding conditions. The interactions between specificity and response sources further revealed that the MoE model is more sensitive to latent semantic cues in problem prompts. It exhibited significantly positive interactions with specificity on linguistic affinity, representational quality, and information load, suggesting its superior capacity to produce detailed and stylistically distinct outputs in response to semantically rich prompts, albeit with increased cognitive burden. In contrast, the Dense model did not show significant interaction effects with specificity across any indicator, indicating a more stable but less adaptive response strategy with respect to variations in problem representation.

Taken together, the results demonstrate that the MoE model exhibits strong advantages in expressive capacity and perception-oriented generation, albeit at the cost of greater cognitive load. The Dense model prioritizes structural economy and order, making it more suitable for low-complexity tasks. Human responses continue to maintain stability in accuracy, control, and structural adaptability. Problem cues not only exert direct effects on Response Traits but also interact with model architectures to modulate generation patterns. These findings offer theoretical and empirical insights into the dynamic mechanisms embedded in human–computer collaborative knowledge construction.

6.2. Theoretical contributions

First, existing literature has yet to offer a systematic account of the role of information in knowledge construction processes. Most prior studies have focused on individual participation traits and behavioral patterns (Hao & Gu, 2024; He & Lu, 2024; T. T. Wu et al., 2024). Against this backdrop, this study introduces a problem-solving-oriented perspective on human–computer collaborative knowledge construction and argues for a conceptual shift wherein information is regarded not as a neutral transmission medium but as a structural element of construction. In dialogic contexts involving LLMs, the expressive form of information, its perceptual features, and cognitive burden constitute core levers of constructional behavior. Building on theories of information processing, levels of processing, and cognitive load, this study constructs an evaluative framework comprising three dimensions—information quality, information perception, and information load—which is further elaborated into six complementary indicators: intrinsic accuracy, contextual alignment, representational quality, linguistic affinity, structural clarity, and information load. This framework extends the scope of information theory and addresses the long-standing issue of insufficient semantic evaluation dimensions in language generation research (McIntosh et al., 2024; Reviriego et al., 2024). It provides theoretical rationale and analytic tools for evaluating informational output in LLM-mediated educational, research, and public discourse scenarios.

Second, although existing research has emphasized the importance of questioning and explanatory strategies in knowledge construction—for example, by highlighting the cognitive stimulation potential of inspiration sources and structural variation in questions (Chin & Chia, 2004; King, 1994; Margutti, 2006)—little empirical evidence has been presented to clarify how intrinsic Question Cues shape the response generation process and its constructive potential. The peripheral role of problem features in past studies is partly due to their ambiguous semantic boundaries and the lack of standardized quantification methods. In response, this study employs natural language processing techniques to construct a two-dimensional problem feature recognition mechanism based on complexity and specificity. By reverse-engineering semantic features from the logical structure of gold-standard answers, the study quantitatively assesses the differential impact of problem cues on constructional traits across response sources. Empirical results confirm that problem cues not only influence response structures and expressions directly but also regulate information pathways through coupling with model architectures. These findings validate the guiding role of informational presentation in knowledge construction behavior (Weijers et al., 2021). The proposed framework offers an effective approach for modeling the mechanisms linking questions, information, and cognition, thereby enriching traditional answer-adaptation models with dynamic semantic components.

Lastly, under the broader theoretical context of human–computer collaborative knowledge construction, previous studies have posited that information exchange among multiple intelligent agents often leads to meaning reconstruction and knowledge emergence (Cress & Kimmerle, 2023). However, a fundamental question remains unresolved: how should cognitive structures and information mechanisms in human–computer interaction be redefined when generative artificial intelligence becomes a participant? Most contemporary studies focus on performance evaluations of the latest models (Dellermann et al., 2019), but these face significant limitations in terms of time sensitivity and generalizability. This study focuses instead on architectural types—Dense and MoE—and selects models that are both socially influential and structurally representative, thereby avoiding overreliance on version-specific or parameter-size comparisons. This shift enables more transferable and theoretically grounded insights into how underlying architectures behave in knowledge construction tasks. The findings demonstrate that human–computer interaction is not a linear exchange of information but a task-centered, dynamically coupled system that integrates information quality, expressive perception, and cognitive cost. The system operates through interdependent pathways shaped by model generation capabilities, task features, and human benchmarks, yielding structurally coordinated patterns of information flow, processing depth, and output load. These findings advance the paradigm of human–computer collaborative knowledge construction and provide foundational theoretical support for designing LLM deployment mechanisms that balance task adaptivity with cognitive economy.

6.3. Practical implications

This study extends the conceptual understanding of human-computer collaborative knowledge construction and offers several actionable insights for the deployment and evaluation of large language models in real-world settings.

The findings demonstrate, through systematic modeling and comparison of micro-level linguistic features, that distinct model architectures such as Dense and MoE differ significantly from human expert responses in terms of information quality, linguistic perception, and information load. Large language models not only excel in linguistic fluency but also exhibit measurable structural and stylistic differences in how information is generated and organized. This underscores the necessity of moving beyond accuracy-based assessment to adopt interdisciplinary evaluation frameworks that draw from information science, psycholinguistics, and cognitive science. Building a multi-level, multi-metric profiling system will advance the interpretability, controllability, and reusability of LLM-based methods and applications.

Furthermore, the results indicate that question complexity and specificity have a significant impact not only on the informational and perceptual features of responses but also on their interaction with model architecture. Therefore, optimal responder selection should be dynamically adapted to question semantics and task demands. For open-domain questions that involve general knowledge, clear context, and minimal structural requirements, both Dense and MoE models outperform human experts in terms of information quality and linguistic perception, making them suitable for routine consultations, popular science dissemination, and automated QA systems. In contrast, when addressing closed-domain questions that require complex reasoning and professional synthesis, the MoE architecture displays greater robustness and adaptability in linguistic extensibility and sensitivity to question features, making it more appropriate for high-level cognitive construction, instructional support, and expert decision-making.

Despite the notable advantages of the MoE model in expression quality, linguistic affinity, and problem adaptiveness, the analysis also reveals a significant increase in information load. High linguistic density and syntactic complexity may result in cognitive overload, diminishing information absorption and inducing information behaviors such as semantic fatigue or avoidance, as previously noted by Manis et al. (1978). As such, current LLM architectures are not yet capable of independently managing knowledge construction under high-load conditions.

In this context, the study proposes the exploration of hybrid intelligence through human-computer collaboration. By strategically combining the compression and calibration capacities of human experts with the generative capabilities of LLMs, a dynamic regulation of cognitive load and output efficacy can be achieved. Human experts can serve as information filters and knowledge calibrators to maintain thematic focus and structural conciseness, while LLMs can enrich the output with linguistic diversity, expressive personalization, and content expansion. This collaborative mechanism supports both the acceptability and adaptability of knowledge output and contributes to the design of cognitively bandwidth-friendly human-computer knowledge systems.

6.4. Limitations and future work

While this study offers a systematic exploration of theoretical frameworks and empirical analysis, several limitations remain and merit further research.

First, in terms of task selection, the current study focuses primarily on professional QA scenarios in the field of computer science. Although this domain represents a typical knowledge-intensive task environment, other disciplines such as medicine, humanities and social sciences, and engineering management differ in their complexity structures, knowledge representations, and linguistic conventions. These differences may affect model performance across domains. Future research should therefore expand to include cross-domain and cross-knowledge-system assessments to enhance external validity and broader applicability.

Second, the study examines only two representative architectures with significant structural divergence—namely, the Dense-based GPT and the MoE-based DeepSeek. This focused comparison is useful for understanding the architectural impact on knowledge construction capacity but does not encompass emerging model types such as multimodal, instruction-tuned, or language–tool hybrid models. This limitation constrains a more granular understanding of how different design philosophies and training strategies shape response behaviors. Future work should incorporate behavioral chain analysis and systematic multi-model comparisons to explore fine-grained performance differentials and cross-domain robustness.

Third, a standardized, prompt-free invocation protocol was used during data collection to avoid bias caused by manual intervention and ensure data comparability. However, this also precluded the inclusion of more advanced collaborative mechanisms such as prompt engineering, which limits the scope of customization and interactive depth assessed. In practical usage, prompt design has become a crucial technique for leveraging model capacity, including strategies such as chain-of-thought and few-shot learning. These influences on output quality and structure merit formal modeling in future research through diversified prompt strategies to examine their coupled effects on model behavior.

Fourth, although the evaluation framework includes multidimensional indicators for information quality, perception, and load, it relies primarily on existing automatic scoring and linguistic feature metrics. The study does not incorporate more fine-grained fact-alignment mechanisms under the accuracy dimension, such as semantic verification systems based on knowledge graphs or professional evidence sources. This omission reflects the current limitations in fact-checking technologies. Future research should develop more stable alignment baselines using high-credibility corpora such as professional databases, patents, and textbooks, and integrate expert annotation or human evaluation procedures to enhance assessment precision.

Fifth, the study centers on the triadic structure of response source, question cue, and response feature. While it preliminarily explores interaction effects between source and question traits, more complex causal paths remain unexplored. Questions such as whether non-linear interactions exist between complexity and specificity, whether synergies between information quality and perception amplify cognitive effects, or whether user background knowledge or information literacy modulates perceived credibility of model outputs, all require deeper investigation. Future work should incorporate structural equation modeling to expand the relational schema and integrate user cognition and perception into the analytic model.

Finally, the study is unavoidably bounded by the technological context of its time. The models and evaluation systems employed reflect the state of technology at a particular moment. Given the rapid evolution of LLMs, the findings should be viewed as a snapshot summary of current capabilities rather than definitive conclusions. Future research should establish more sustainable and architecture-agnostic evaluation methodologies to maintain robustness under technological shifts, thereby laying a methodological foundation for long-term studies of human-computer knowledge construction.

7. Conclusion

This study employs text analysis and regression modeling to investigate knowledge construction behaviors in QA-driven contexts, focusing on how large language models based on Dense and MoE architectures perform in handling questions of varying features and how these compare to human expert responses. The goal is to identify the expressive, cognitive, and load-related characteristics of different response sources and to assess their implications for human-computer interaction and information services. The study seeks to provide empirical support and strategic guidance for both users and developers of LLMs in future knowledge construction applications.

Through the use of the LIWC affective dictionary and various structural linguistic indicators, the study extracts, quantifies, and normalizes features of questions and corresponding responses. Based on semantic entropy weighting, a comparative framework is constructed that links question features such as complexity and specificity to response features such as information quality, linguistic perception, and cognitive load. This modeling establishes dynamic mappings and interaction mechanisms between question and response elements.

The study expands the theoretical boundary of information science in human-computer collaboration and provides tools for adaptive evaluation of language models in knowledge construction tasks. By modeling six sub-dimensions—intrinsic accuracy, contextual alignment, representational quality, linguistic affinity, structural clarity, and information load—it verifies systematic differences in semantic strategies and output costs across model architectures, offering theoretical and practical insights for model selection and hybrid application design.

Moreover, by integrating theories of information processing, levels of processing, and cognitive load, the study proposes an information response mechanism for QA tasks that explains how question features shape responses and how different architectures mediate construction pathways. It offers a practice-oriented research paradigm for modeling question features, constructing response dimensions, and identifying interaction paths. Future work may further introduce comprehensive information behavior theories, expand data and question type coverage, and integrate more complex causal models and user traits, thereby advancing the convergence of information science, human-computer collaboration, and language model research.

Author agreement

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We understand that the Corresponding Author is the sole contact for the Editorial process. He/she is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs.

Signed by all authors as follows: Jingzhu Wei, Zhipeng Chen

Funding

This work was supported by the Major projects of the National Social Science Foundation in China [No.: 21&ZD339].

CRediT authorship contribution statement

Jingzhu Wei: Writing – review & editing, Validation, Supervision, Software, Resources, Project administration, Funding acquisition, Conceptualization. **Zhipeng Chen:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Data availability

Data will be made available on request.

References

Acerbi, A., & Stubbersfield, J. M. (2023). Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences, 120*(44), Article e2313790120. https://doi.org/10.1073/pnas.2313790120

Adams, N. E. (2015). Bloom's taxonomy of cognitive learning objectives. *Journal of the Medical Library Association: JMLA, 103*(3), 152. https://doi.org/10.3163/1536-5050.103.3.010

Álvares, A. (2001). The 3M way to innovation: Balancing people and profit. *Revista de Administração de Empresas, 41*(09/01), 94–95. https://doi.org/10.1590/S0034-75902001000300013

An, S., Zhang, S., Guo, T., Lu, S., Zhang, W., & Cai, Z. (2025). Impacts of generative AI on student teachers' task performance and collaborative knowledge construction process in mind mapping-based collaborative environment. *Computers & Education, 227*, Article 105227. https://doi.org/10.1016/j.compedu.2024.105227

Anantrasirichai, N., & Bull, D. (2022). Artificial intelligence in the creative industries: A review. *Artificial Intelligence Review, 55*(1), 589–656. https://doi.org/10.1007/s10462-021-10039-7

Bevara, R. V. K., Mannuru, N. R., Lund, B. D., Karedla, S. P., & Mannuru, A. (2025). Beyond ChatGPT: How DeepSeek R1 may transform academia and libraries? *Library Hi Tech News, 42*(4), 4–8. https://doi.org/10.1108/LHTN-01-2025-0024

Biggs, J. B. (1987). *Student approaches to learning and studying.* Research Monograph. ERIC.

Bishop, C. M. (1994). Neural networks and their applications. *Review of Scientific Instruments, 65*(6), 1803–1832. https://doi.org/10.1063/1.1144830

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). *A large annotated corpus for learning natural language inference*. arXiv preprint. https://doi.org/10.18653/v1/D15-1075. arXiv:1508.05326.

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The development and psychometric properties of LIWC-22, 10* pp. 1–47). Austin, TX: University of Texas at Austin. https://doi.org/10.13140/RG.2.2.23890.43205

Breadmore, H. L., & Carroll, J. M. (2016). Effects of orthographic, morphological and semantic overlap on short-term memory for words in typical and atypical development. *Scientific Studies of Reading, 20*(6), 471–489. https://doi.org/10.1080/10888438.2016.1246554

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Byström, K., & Järvelin, K. (1995). Task complexity affects information seeking and use. *Information Processing & Management, 31*(2), 191–213. https://doi.org/10.1016/0306-4573(95)80035-R

Cai, W., Jiang, J., Wang, F., Tang, J., Kim, S., & Huang, J. (2025). A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering, 37*(7), 3896–3915. https://doi.org/10.1109/TKDE.2025.3554028

Cai, Z. G., Duan, X., Haslett, D. A., Wang, S., & Pickering, M. J. (2023). *Do large language models resemble humans in language use?* (pp. 1–39) arXiv preprint. https://doi.org/10.48550/arXiv.2303.08014. arXiv:2303.08014.

Caraban, A., Karapanos, E., Gonçalves, D., & Campos, P. (2019). 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. https://doi.org/10.1145/3290605.3300733

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education, 20*(1), 43. https://doi.org/10.1186/s41239-023-00411-8

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*(4), 293–332. https://doi.org/10.1207/s1532690xci0804_2

Chen, B., & Hong, H.-Y. (2016). Schools as knowledge-building organizations: Thirty years of design research. *Educational Psychologist,, 51*(2), 266–288. https://doi.org/10.1080/00461520.2016.1175306

Chen, C.-Y., Pedersen, S., & Murphy, K. L. (2012). The influence of perceived information overload on student participation and knowledge construction in computer-mediated communication. *Instructional Science, 40*, 325–349. https://doi.org/10.1007/s11251-011-9179-0

Chen, K., Song, N., Zhao, Y., Peng, J., & Chen, Y. (2024). Online attention versus knowledge utilization: Exploring how linguistic features of scientific papers influence knowledge diffusion. *Information Processing & Management, 61*(3), Article 103691. https://doi.org/10.1016/j.ipm.2024.103691

Chen, Z., Wang, Z., Wang, Z., Liu, H., Yin, Z., Liu, S., Sheng, L., Ouyang, W., Qiao, Y., & Shao, J. (2023). *Octavius: mitigating task interference in mllms via lora-moe* (pp. 1–22). arXiv preprint. https://doi.org/10.48550/arXiv:2311.02684. arXiv:2311.02684.

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Chin, C., & Chia, L. G. (2004). Problem-based learning: Using students' questions to drive knowledge construction. *Science Education, 88*(5), 707–727. https://doi.org/10.1002/sce.10144

Chiu, T. K. (2024). Future research recommendations for transforming higher education with generative AI. *Computers and Education: Artificial Intelligence, 6*, Article 100197. https://doi.org/10.1016/j.caeai.2023.100197

Chowdhury, G., & Chowdhury, S. (2024). AI-and LLM-driven search tools: A paradigm shift in information access for education and research. *Journal of Information Science*. https://doi.org/10.1177/01655515241284046, 01655515241284046.

Collins, K. M., Jiang, A. Q., Frieder, S., Wong, L., Zilka, M., Bhatt, U., Lukasiewicz, T., Wu, Y., Tenenbaum, J. B., & Hart, W. (2024). Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences, 121*(24), Article e2318124121. https://doi.org/10.1073/pnas.2318124121

Conroy, G., & Mallapaty, S. (2025). How China created AI model DeepSeek and shocked the world. *Nature, 638*(8050), 300–301. https://doi.org/10.1038/d41586-025-00259-0

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning And Verbal Behavior, 11*(6), 671–684. https://doi.org/10.1037/h0084237

Cress, U., & Kimmerle, J. (2023). Co-constructing knowledge with generative AI tools: Reflections from a CSCL perspective. *International Journal of Computer-Supported Collaborative Learning, 18*(4), 607–614. https://doi.org/10.1007/s11412-023-09409-w

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. *Business & Information Systems Engineering, 61*(5), 637–643. https://doi.org/10.1007/s12599-019-00595-2

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition, 109*(2), 193–210. https://doi.org/10.1016/j.cognition.2008.07.008

Deng, Z., Ma, W., Han, Q.-L., Zhou, W., Zhu, X., Wen, S., & Xiang, Y. (2025). Exploring DeepSeek: A survey on advances, applications, challenges and future directions. *IEEE/CAA Journal of Automatica Sinica, 12*(5), 872–893. https://doi.org/10.1109/JAS.2025.125498

Dole, J. A., & Sinatra, G. M. (1998). Reconceptalizing change in the cognitive construction of knowledge. *Educational Psychologist, 33*(2–3), 109–128. https://doi.org/10.1207/s15326985ep3302&3_5

Doyle, P. R., Clark, L., & Cowan, B. R. (2021). What do we see in them? Identifying dimensions of partner models for speech interfaces using a psycholexical approach. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3411764.3445206

Engeström, Y. (2001). Expansive learning at work: Toward an activity theoretical reconceptualization. *Journal of Education and Work, 14*(1), 133–156. https://doi.org/10.1080/13639080020028747

Espinosa-Rada, A., Bellotti, E., Everett, M. G., & Stadtfeld, C. (2024). Co-evolution of a socio-cognitive scientific network: A case study of citation dynamics among astronomers. *Social Networks, 78*, 92–108. https://doi.org/10.1016/j.socnet.2023.11.008

Fedorenko, E., Piantadosi, S. T., & Gibson, E. A. (2024). Language is primarily a tool for communication rather than thought. *Nature, 630*(8017), 575–586. https://doi.org/10.1038/s41586-024-07522-w

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences, 112*(33), 10336–10341. https://doi.org/10.1073/pnas.1502134112

Gan, W., Ning, Z., Qi, Z., & Yu, P. S. (2025). *Mixture of experts (MoE): a big data perspective* (pp. 1–28). arXiv preprint. https://doi.org/10.48550/arXiv:2501.16352. arXiv:2501.16352.

Gao, M., Hu, X., Yin, X., Ruan, J., Pu, X., & Wan, X. (2025). Llm-based nlg evaluation: Current status and challenges. *Computational Linguistics*, 1–28. https://doi.org/10.1162/coli_a_00561

Germino, J., Moniz, N., & Chawla, N. V. (2024). FairMOE: Counterfactually-fair mixture of experts with levels of interpretability. *Machine Learning, 113*(9), 6539–6559. https://doi.org/10.1007/s10994-024-06583-2

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(1), 32–42. https://doi.org/10.3102/0013189X018001032

Zhuang, Y., Liu, Q., Ning, Y., Huang, W., Lv, R., Huang, Z., Zhao, G., Zhang, Z., Mao, Q., & Wang, S. (2023). *Efficiently measuring the cognitive ability of llms: an adaptive testing perspective*, 1–15. https://doi.org/10.48550/arXiv.2306.10512. arXiv:2306.10512.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68*(1), 1–76. https://doi.org/10.1016/S0010-0277(98)00034-1

Go, E., & Sundar, S. S. (2019). Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior, 97*, 304–316. https://doi.org/10.1016/j.chb.2019.01.020

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science, 3*(2), 371–398. https://doi.org/10.1111/j.1756-8765.2010.01081.x

Gray, R. M. (2011). *Entropy and information theory*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4419-7970-4

Gunawardena, C. N., Lowe, C. A., & Anderson, T. (1997). Analysis of a global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. *Journal of Educational Computing Research, 17*(4), 397–431. https://doi.org/10.2190/7MQV-X9UJ-C7Q3-NRAG

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). *How close is chatgpt to human experts? comparison corpus, evaluation, and detection* (pp. 1–20). arXiv preprint. https://doi.org/10.48550/arXiv.2301.07597. arXiv:2301.07597.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., & Bi, X. (2025). *Deepseek-r1: incentivizing reasoning capability in llms via reinforcement learning* (pp. 1–22). arXiv preprint. https://doi.org/10.48550/arXiv:2501.12948. arXiv:2501.12948.

Halliday, T. C. (1985). Knowledge mandates: Collective influence by scientific, normative and syncretic professions. *British Journal of Sociology*, 421–447. https://doi.org/10.2307/590459

Handler, A., Larsen, K. R., & Hackathorn, R. (2024). Large language models present new questions for decision support. *International Journal of Information Management, 79*, Article 102811. https://doi.org/10.1016/j.ijinfomgt.2024.102811

Hao, X., & Gu, X. (2024). Analysis of key roles in large-scale online learning: Interactive participation characteristics and knowledge construction behaviour patterns. *British Journal of Educational Technology, 55*(3), 910–932. https://doi.org/10.1111/bjet.13405

He, S., & Lu, Y. (2024). The effectiveness of Gen AI in assisting students' knowledge construction in humanities and social sciences courses: Learning behaviour analysis. *Interactive Learning Environments, 32*(10), 7041–7062. https://doi.org/10.1080/10494820.2024.2415444

Hou, T.-Y., Tseng, Y.-C., & Yuan, C. W. T. (2024). Is this AI sexist? The effects of a biased AI's anthropomorphic appearance and explainability on users' bias perceptions and trust. *International Journal of Information Management, 76*, Article 102775. https://doi.org/10.1016/j.ijinfomgt.2024.102775

Huang, Q., An, Z., Zhuang, N., Tao, M., Zhang, C., Jin, Y., Xu, K., Chen, L., Huang, S., & Feng, Y. (2024). *Harder tasks need more experts: dynamic routing in moe models* (pp. 1–11). arXiv preprint. https://doi.org/10.48550/arXiv:2403.07652. arXiv:2403.07652.

Huang, S., Huang, Y., Liu, Y., Luo, Z., & Lu, W. (2025a). Are large language models qualified reviewers in originality evaluation? *Information Processing & Management, 62*(3), Article 103973. https://doi.org/10.1016/j.ipm.2024.103973

Huang, Y., Wang, W., Zhou, J., Zhang, L., Lin, J., Liu, H., Hu, X., Zhou, Z., & Dong, W. (2025b). Integrative modeling enables ChatGPT to achieve average level of human counselors performance in mental health Q&A. *Information Processing & Management, 62*(5), Article 104152. https://doi.org/10.1016/j.ipm.2025.104152

Hutchins, E. (1995). *Cognition in the wild*. MIT press.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation, 3*(1), 79–87. https://doi.org/10.1109/ICNN.1993.298760

Jagdishbhai, N., & Thakkar, K. Y. (2023). Exploring the capabilities and limitations of GPT and chat GPT in natural language processing. *Journal of Management Research and Analysis, 10*(1), 18–20. https://doi.org/10.18231/j.jmra.2023.004

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working papers/Lund University, Department of Linguistics and Phonetics,, 53*, 61–79-61–79.

Jordan, M. I., & Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation, 6*(2), 181–214. https://doi.org/10.1162/neco.1994.6.2.181

Kabir, S., Udo-Imeh, D. N., Kou, B., & Zhang, T. (2024). Is stack overflow obsolete? An empirical study of the characteristics of chatgpt answers to stack overflow questions. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

Kahn, B. K., Strong, D. M., & Wang, R. Y. (2002). Information quality benchmarks: Product and service performance. *Communications of the ACM, 45*(4), 184–192. https://doi.org/10.1145/505248.506007

Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need? *Educational Psychology Review, 23*(1), 1–19. https://doi.org/10.1007/s10648-010-9150-7

Kang, J. K. Z., Tan, S. Y., He, B., & Zhang, Z. (2023). Real time index and search across large quantities of gnn experts for low latency online learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/3580305.3599893

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2024). Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A, 382*(2270), Article 20230254. https://doi.org/10.2139/ssrn.4389233

Kembaren, F. R., & Aswani, A. N. (2022). Exploring lexical density in the. *New York Times*, 1–11. https://doi.org/10.32528/ellite.v7i2.8795

King, A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Educational Research Journal, 31*(2), 338–368. https://doi.org/10.3102/00028312031002338

Krajcik, J., McNeill, K. L., & Reiser, B. J. (2008). Learning-goals-driven design model: Developing curriculum materials that align with national standards and incorporate project-based pedagogy. *Science Education, 92*(1), 1–32. https://doi.org/10.1002/sce.20240

Lee, H., Chen, P., Wang, W., Huang, Y., & Wu, T. (2024). Empowering ChatGPT with guidance mechanism in blended learning: Effect of self-regulated learning, higher-order thinking skills, and knowledge construction. *International Journal of Educational Technology in Higher Education, 21*(1), 16. https://doi.org/10.1186/s41239-024-00447-4

Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: a methodology for information quality assessment. *Information & management, 40*(2), 133–146. https://doi.org/10.1016/S0378-7206(02)00043-5

Lin, S., Hilton, J., & Evans, O. (2021). *Truthfulqa: measuring how models mimic human falsehoods*. arXiv preprint. arXiv:2109.07958.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science, 9*(2), 159–191. https://doi.org/10.17791/jcs.2008.9.2.159

Liu, J., Li, S., & Dong, Q. (2024a). Collaboration with generative artificial intelligence: An exploratory study based on learning analytics. *Journal of Educational Computing Research, 62*(5), 1234–1266. https://doi.org/10.1177/07356331241242441

Liu, M., Zhang, L. J., & Biebricher, C. (2024b). Investigating students' cognitive processes in generative AI-assisted digital multimodal composing and traditional writing. *Computers & Education, 211*, Article 104977. https://doi.org/10.1016/j.compedu.2023.104977

Liu, Y., Bhandari, S., & Pardos, Z. A. (2025). Leveraging LLM respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology, 56*(3), 1028–1052. https://doi.org/10.1111/bjet.13570

Liu, Z., & Sun, W. (2023). *Uncovering high-dimensional phase space and the application of mixture of experts (MoE) on building the large calphad model (LCM)* (pp. 1–14). arXiv preprint. https://doi.org/10.48550/arXiv:2312.06429. arXiv:2312.06429.

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science, 3*(2), 273–302. https://doi.org/10.1111/j.1756-8765.2010.01106.x

Lu, X., & Hu, R. (2024). Sense-aware connective-based indices of cohesion and their relationship to cohesion ratings of English language learners' written production. *Studies in Second Language Acquisition, 46*(3), 644–662. https://doi.org/10.1017/S0272263124000202

Manis, M., Fichman, M., & Platt, M. B. (1978). Cognitive integration and referential communication: Effects of information quality and quantity in message decoding. *Organizational Behavior and Human Performance, 22*(3), 417–430. https://doi.org/10.1016/0030-5073(78)90025-9

Mao, B., Jia, X., & Huang, Q. (2024). How do information overload and message fatigue reduce information processing in the era of COVID-19? An ability–motivation approach. *Journal of Information Science, 50*(5), 1242–1254. https://doi.org/10.1177/01655515221118047

Margutti, P. (2006). Are you human beings?" order and knowledge construction through questioning in primary classroom interaction. *Linguistics and Education, 17*(4), 313–346. https://doi.org/10.1016/j.linged.2006.12.002

Markowitz, D. M. (2023). Analytic thinking as revealed by function words: What does language really measure? *Applied Cognitive Psychology, 37*(3), 643–650. https://doi.org/10.1002/acp.4057

Martínez, G., Hernández, J. A., Conde, J., Reviriego, P., & Merino, E. (2024). *Beware of words: evaluating the lexical richness of conversational large language models* (pp. 1–15). arXiv preprint. https://doi.org/10.48550/arXiv:2402.15518. arXiv:2402.15518.

Mayordomo, R. M., & Onrubia, J. (2015). Work coordination and collaborative knowledge construction in a small group collaborative virtual task. *The Internet and Higher Education, 25*, 96–104. https://doi.org/10.1016/j.iheduc.2015.02.003

McIntosh, T. R., Susnjak, T., Arachchilage, N., Liu, T., Watters, P., & Halgamuge, M. N. (2024). *Inadequacies of large language model benchmarks in the era of generative artificial intelligence* (pp. 1–18). arXiv preprint. https://doi.org/10.48550/arXiv:2402.09880. arXiv:2402.09880.

McIntosh, T. R., Susnjak, T., Arachchilage, N., Liu, T., Xu, D., Watters, P., & Halgamuge, M. N. (2025). Inadequacies of large language model benchmarks in the era of generative artificial intelligence. *IEEE Transactions on Artificial Intelligence*. https://doi.org/10.1109/TAI.2025.3569516

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with coh-metrix*. Cambridge University Press. *https://dl.acm.org/doi/10.5555/2655323*.

Mercer, S., Spillard, S., & Martin, D. P. (2025). *Brief analysis of deepseek R1 and it's implications for generative ai* (pp. 1–9). arXiv preprint. https://doi.org/10.48550/arXiv.2502.02523. arXiv:2502.02523.

Miller, G. A. (2003). The cognitive revolution: A historical perspective. *Trends in Cognitive Sciences, 7*(3), 141–144. https://doi.org/10.1016/S1364-6613(03)00029-9

Mirus, F., Stewart, T. C., Eliasmith, C., & Conradt, J. (2019). A mixture-of-experts model for vehicle prediction using an online learning approach. In *International Conference on Artificial Neural Networks*. https://doi.org/10.1007/978-3-030-30508-6_37

Mirzaei, A., Wilkie, D. C. H., & Burgess, A. J. (2024). Does human value-expressive brand anthropomorphism on social media lead to greater brand competitiveness? *Journal of Retailing and Consumer Services, 81*, Article 104032. https://doi.org/10.1016/j.jretconser.2024.104032

Motegi, M., Shino, M., Kuwabara, M., Takahashi, H., Matsuyama, T., Tada, H., Hagiwara, H., & Chikamatsu, K. (2025). Comparison of physician and large language model chatbot responses to online ear, nose, and throat inquiries. *Scientific Reports, 15*(1), Article 21346. https://doi.org/10.1038/s41598-025-06769-1

Muñoz-Ortiz, A., Gómez-Rodríguez, C., & Vilares, D. (2024). Contrasting linguistic patterns in Human and LLM-generated news text. *Artificial Intelligence Review, 57*(10), 265. https://doi.org/10.1007/s10462-024-10903-2

Mwambe, O. O. (2024). Deployment of information processing theory to support adaptive e-learning systems: Feasibility study. *Computers in Human Behavior Reports, 14*, Article 100420. https://doi.org/10.1016/j.chbr.2024.100420

Nguyen, X.-P., Pandit, S., Purushwalkam, S., Xu, A., Chen, H., Ming, Y., Ke, Z., Savarese, S., Xong, C., & Joty, S. (2024). *Sfr-rag: towards contextually faithful llms*. arXiv preprint. https://doi.org/10.48550/arXiv.2409.09916. arXiv:2409.09916.

Nilsson, N. J. (2009). *The quest for artificial intelligence*. Cambridge University Press. https://doi.org/10.1017/CBO9780511819346

Obrenovic, Z., & Starcevic, D. (2004). Modeling multimodal human-computer interaction. *Computer, 37*(9), 65–72. https://doi.org/10.1109/MC.2004.139

Otter, D. W., Medina, J. R., & Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. *IEEE Transactions On Neural Networks And Learning Systems, 32*(2), 604–624. https://doi.org/10.48550/arXiv.1807.10854

Ouyang, F., Wu, M., Zhang, L., Xu, W., Zheng, L., & Cukurova, M. (2023). Making strides towards AI-supported regulation of learning in collaborative knowledge construction. *Computers in Human Behavior, 142*, Article 107650. https://doi.org/10.1016/j.chb.2023.107650

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*(1), 1–4. https://doi.org/10.1207/S15326985EP3801_1

Paavola, S., Lipponen, L., & Hakkarainen, K. (2004). Models of innovative knowledge communities and three metaphors of learning. *Review of Educational Research, 74* (4), 557–576. https://doi.org/10.3102/00346543074004557

Palincsar, A. S. (1998). Social constructivist perspectives on teaching and learning. *Annual Review Of Psychology, 49*(1), 345–375. https://doi.org/10.1146/annurev.psych.49.1.345

Palincsar, A. S. (2012). *Social constructivist perspectives on teaching and learning* (pp. 290–319). An introduction to Vygotsky. https://doi.org/10.1146/annurev.psych.49.1.345

Palopak, Y., Huang, S.-J., & Ratnasari, W. (2023). Knowledge diffusion trajectories of agile software development research: A main path analysis. *Information and Software Technology, 156*, Article 107131. https://doi.org/10.1016/j.infsof.2022.107131

Payne, J. W. (2024). Information processing theory: Some concepts and methods applied to decision research. *Cognitive processes in choice and decision behavior* (pp. 95–115). Routledge. https://doi.org/10.4324/9781003469544-6

Pea, R. (1994). Seeing what we build together: Distributed multimedia learning environments for transformative communications. *Journal of the Learning Sciences, 3*, 1–16. https://doi.org/10.1207/s15327809jls0303_4

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM, 45*(4), 211–218. https://doi.org/10.1145/505248.506010

Puspitasari, F. D., Zhang, C., Dam, S. K., Zhang, M., Kim, T.-H., Hong, C. S., Bae, S.-H., Qin, C., Wei, J., & Wang, G. (2025). *DeepSeek models: a comprehensive survey of methods and applications* (pp. 1–28). Authorea Preprints. https://doi.org/10.36227/techrxiv.174198511.15158242/v1

Qi, W., Pan, J., Lyu, H., & Luo, J. (2024). Excitements and concerns in the post-chatgpt era: Deciphering public perception of ai through social media analysis. *Telematics and Informatics, 92*, Article 102158. https://doi.org/10.48550/arXiv.2307.05809

Reviriego, P., Conde, J., Merino-Gómez, E., Martínez, G., & Hernández, J. A. (2024). *Playing with words: comparing the vocabulary and lexical diversity of chatgpt and humans*. Machine Learning with Applications, Article 100602. https://doi.org/10.1016/j.mlwa.2024.100602

Riemer, K., & Peter, S. (2024). Conceptualizing generative AI as style engines: Application archetypes and implications. *International Journal of Information Management, 79*, Article 102824. https://doi.org/10.1016/j.ijinfomgt.2024.102824

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval, 3*(4), 333–389. https://doi.org/10.1561/1500000019

Rosenfeld, A., & Lazebnik, T. (2024). *Whose llm is it anyway? linguistic comparison and llm attribution for GPT-3.5, GPT-4 and bard* (pp. 1–12). arXiv preprint. https://doi.org/10.48550/arXiv.2402.14533. arXiv:2402.14533.

Sallam, M., Al-Mahzoum, K., Sallam, M., & Mijwil, M. M. (2025). DeepSeek: Is it the end of generative AI monopoly or the mark of the impending doomsday? *Mesopotamian Journal of Big Data*, 26–34. https://doi.org/10.58496/MJBD/2025/002, 2025.

Schnotz, W., & Kürschner, C. (2007). A reconsideration of cognitive load theory. *Educational Psychology Review, 19*(4), 469–508. https://doi.org/10.1007/s10648-007-9053-4

Shijaku, R., & Canhasi, E. (2023). *ChatGPT generated text detection.* Publisher: Unpublished.

Simon, H. A. (1978). *Information-processing theory of human problem solving.* Erlbaum Hillsdale, NJ.

Singh, S., Bansal, S., Saddik, A. E., & Saini, M. (2025). *From chatgpt to deepseek AI: a comprehensive analysis of evolution, deviation, and future implications in AI-Language models* (pp. 1–10). arXiv preprint. https://doi.org/10.48550/arXiv:2504.03219. arXiv:2504.03219.

Sobieszek, A., & Price, T. (2022). Playing games with AIs: The limits of GPT-3 and similar large language models. *Minds and Machines, 32*(2), 341–364. https://doi.org/10.1007/s11023-022-09602-0

Song, R., Li, Y., Tian, M., Wang, H., Giunchiglia, F., & Xu, H. (2025). Causal keyword driven reliable text classification with large language model feedback. *Information Processing & Management, 62*(2), Article 103964. https://doi.org/10.1016/j.ipm.2024.103964

Stephanidis, C., & Salvendy, G. (2024). *Foundations and fundamentals in human-computer interaction.* CRC Press. https://doi.org/10.1201/9781003495109

Stokhof, H., Meli, K., Lavidas, K., & Grammenos, D. (2022). Why answer this question? Experts' Behaviors on educational community question-answering platforms. *Electronic Journal of e-Learning, 20*(2), 87–100. https://doi.org/10.34190/ejel.20.2.2240

Stout, P. J., & Klett, M. D. (2025). New media technologies and information processing theory—George a. miller and others. *Science education in theory and practice: an introductory guide to learning theory* (pp. 101–123). Springer. https://doi.org/10.1007/978-3-031-81351-1_7

Stvilia, B., Gasser, L., Twidale, M. B., & Smith, L. C. (2007). A framework for information quality assessment. *Journal of the American Society For Information Science And Technology, 58*(12), 1720–1733. https://doi.org/10.1002/asi.20652

Su, Z., Wang, H., Shao, Z., Liu, Z., Liu, Y., & Liu, S. (2025). Ecological network analysis of attention flow in online learning: Insights into knowledge acquisition and dropout behaviors. *Information Processing & Management, 62*(5), Article 104163. https://doi.org/10.1016/j.ipm.2025.104163

Suárez, A., Díaz-Flores García, V., Algar, J., Gómez Sánchez, M., Llorente de Pedro, M., & Freire, Y. (2024). Unveiling the ChatGPT phenomenon: Evaluating the consistency and accuracy of endodontic question answers. *International Endodontic Journal, 57*(1), 108–113. https://doi.org/10.1111/iej.13985

Sweller, J. (2011). Cognitive load theory. In *Psychology of learning and motivation, 55* pp. 37–76). Elsevier. https://doi.org/10.1007/978-1-4419-8126-4

Teng, Z., Yan, Z., Song, Y., Ye, X., & Ouyang, Y. (2024). *Data-Driven moe: a data-driven approach to construct moe by a single llm.* https://doi.org/10.1007/978-981-97-5672-8_30

Torruella, J., & Capsada, R. (2013). Lexical statistics and tipological structures: A measure of lexical richness. *Procedia-Social and Behavioral Sciences, 95*, 447–454. https://doi.org/10.1016/j.sbspro.2013.10.668

Van Aalst, J. (2009). Distinguishing knowledge-sharing, knowledge-construction, and knowledge-creation discourses. *International Journal of Computer-Supported Collaborative Learning, 4*, 259–287. https://doi.org/10.1007/s11412-009-9069-5

Ghosh, S. (2021). Predicting question deletion and assessing question quality in social Q&A sites using weakly supervised deep neural networks. https://doi.org/10.24251/HICSS.2021.329.

Vats, A., Raja, R., Jain, V., & Chadha, A. (2024). The evolution of mixture of experts: A survey from basics to breakthroughs. https://doi.org/10.20944/preprints202408.0583.v1.

Wach, F.-S., Spengler, M., Gottschling, J., & Spinath, F. M. (2015). Sex differences in secondary school achievement–The contribution of self-perceived abilities and fear of failure. *Learning and Instruction, 36*, 104–112. https://doi.org/10.1016/j.learninstruc.2015.01.005

Wang, C., & Kantarcioglu, M. (2025). *A review of deepseek models' key innovative techniques*, 1–11. https://doi.org/10.48550/arXiv.2503.11486. arXiv:2503.11486.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099

Wang, S., Chen, Z., Li, B., He, K., Zhang, M., & Wang, J. (2024). *Scaling laws across model architectures: a comparative analysis of dense and moe models in large language models*, 1–13. https://doi.org/10.48550/arXiv:2410.05661. arXiv:2410.05661.

Wei, Y., Lu, W., Cheng, Q., Jiang, T., & Liu, S. (2022). How humans obtain information from AI: Categorizing user messages in human-AI collaborative conversations. *Information Processing & Management, 59*(2), Article 102838. https://doi.org/10.1016/j.ipm.2021.102838

Weijers, R. J., de Koning, B. B., & Paas, F. (2021). Nudging in education: From theory towards guidelines for successful implementation. *European Journal of Psychology of Education, 36*, 883–902. https://doi.org/10.1007/s10212-020-00495-0

Williams, A., Nangia, N., & Bowman, S. R. (2017). *A broad-coverage challenge corpus for sentence understanding through inference.* https://doi.org/10.18653/v1/N18-1101. arXiv:1704.05426.

Wu, J., Wang, Z., & Qin, Y. (2025). Performance of DeepSeek-R1 and ChatGPT-4o on the Chinese National Medical Licensing Examination: A comparative study. *Journal of Medical Systems, 49*(1), 1–5. https://doi.org/10.1007/s10916-025-02213-z

Wu, Q., Goh, D. H. L., & Lee, C. S. (2020). Generating collective online information sharing: The framing effect of questions on response network. *Proceedings of the Association for Information Science and Technology, 57*(1), e213. https://doi.org/10.1002/pra2.213

Wu, Q., & Zheng, H. (2025). Consumers' questions as nudges: Comparing the effect of linguistic cues on LLM chatbot and human responses. *Journal of Retailing and Consumer Services, 84*, Article 104250. https://doi.org/10.1016/j.jretconser.2025.104250

Wu, T., Lee, H., Li, P., Huang, C., & Huang, Y. (2024a). Promoting self-regulation progress and knowledge construction in blended learning via ChatGPT-based learning aid. *Journal of Educational Computing Research, 61*(8), 3–31. https://doi.org/10.1177/07356331231191125

Wu, Y., Qu, W., Tao, T., Wang, Z., Bai, W., Li, Z., Tian, Y., Zhang, J., Lentz, M., & Zhuo, D. (2024b). *Lazarus: resilient and elastic training of mixture-of-experts models with adaptive expert placement*, 1–17. https://doi.org/10.48550/arXiv:2407.04656. arXiv:2407.04656.

Xue, J., Wang, L., Zheng, J., Li, Y., & Tan, Y. (2023). Can ChatGPT kill user-generated Q&A platforms? *Available at SSRN, 4448938*, 1–22. https://doi.org/10.2139/ssrn.4448938

Yang, Y., Yih, W.-T., & 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.* https://doi.org/10.18653/v1/D15-1237

Yi, R., Guo, L., Wei, S., Zhou, A., Wang, S., & Xu, M. (2023). *Edgemoe: fast on-device inference of moe-based large language models*, 1–16. https://doi.org/10.48550/arXiv.2308.14352. arXiv:2308.14352.

Yu, P., Chen, J., Feng, X., & Xia, Z. (2023). *Cheat: a large-scale dataset for detecting chatgpt-written abstracts*. https://doi.org/10.1109/TBDATA.2025.3536929

Yuan, P., Feng, S., Li, Y., Wang, X., Zhang, Y., Shi, J., Tan, C., Pan, B., Hu, Y., & Li, K. (2025). *LLM-Powered benchmark factory: reliable, generic, and efficient*, 1–25. https://doi.org/10.48550/arXiv.2502.01683. arXiv:2502.01683.

Zhang, J., Gao, M., & Zhang, J. (2021). The learning behaviours of dropouts in MOOCs: A collective attention network perspective. *Computers & Education, 167*, Article 104189. https://doi.org/10.1016/j.compedu.2021.104189

Zhang, P., Li, T., Yuan, Z., Luo, C., Liu, K., & Yang, X. (2022). Heterogeneous feature selection based on neighborhood combination entropy. *IEEE Transactions On Neural Networks And Learning Systems, 35*(3), 3514–3527. https://doi.org/10.1109/TNNLS.2022.3193929

Zhao, C., Deng, C., Ruan, C., Dai, D., Gao, H., Li, J., Zhang, L., Huang, P., Zhou, S., & Ma, S. (2025). *Insights into deepseek-v3: scaling challenges and reflections on hardware for ai architectures*, 1–14. https://doi.org/10.48550/arXiv.2505.09343. arXiv:2505.09343.

Zheng, Y., Zhao, K., & Stylianou, A. (2013). The impacts of information quality and system quality on users' continuance intention in information-exchange virtual communities: An empirical investigation. *Decision Support Systems, 56*, 513–524. https://doi.org/10.1016/j.dss.2012.11.008

Zhou, T., & Wu, X. (2024). Examining user migration intention from social Q&A communities to generative AI. *Humanities and Social Sciences Communications, 11*(1), 1083. https://doi.org/10.1057/s41599-024-03540-1