Cognitive Computation (2025) 17:35
https://doi.org/10.1007/s12559-024-10393-y

RESEARCH

A Comparative Analysis of Metaphorical Cognition in ChatGPT and Human Minds

Rui Mao[1] · Guanyi Chen[2] · Xiao Li[3] · Mengshi Ge[1] · Erik Cambria[1]

Received: 22 April 2024 / Accepted: 1 December 2024 / Published online: 24 December 2024

Abstract

ChatGPT represents a significant advancement in the field of Artificial Intelligence (AI), showcasing the development of a robust AI system capable of multitasking and generating human-like language. At present, many scholars have done evaluations on ChatGPT in terms of language, reasoning, and scientific knowledge abilities, based on benchmarks or well-crafted questions. However, to the best of our knowledge, there is currently no existing comparative analysis from a cognitive perspective that directly assesses ChatGPT alongside humans. Metaphor, serving as a manifestation of linguistic creativity, provides a valuable avenue for examining cognition. This is due to the mapping relationship it establishes between the target and source conceptual domains, reflecting distinct cognitive patterns. In this paper, we use a metaphor processing tool, MetaPro, to analyze the cognitive differences between ChatGPT and humans through the metaphorical expressions in ChatGPT- and human-generated text. We illustrate the preferences in metaphor usage, concept mapping, and cognitive pattern variances across different domains. The methodology utilized in this study makes a valuable contribution to the task-agnostic evaluation of AI systems and cognitive research. The insights garnered from this research prove instrumental in comprehending the cognitive distinctions between ChatGPT and humans, facilitating the identification of potential cognitive biases within ChatGPT.

Keywords Cognitive analysis · Conceptual mapping · ChatGPT · MetaPro

Introduction

As an epoch-making Artificial Intelligence (AI) product, ChatGPT has presented its capability of generating creative text [1, 2] and solving complex problems [3]. Its excellent language generation ability and rich knowledge in different domains make it difficult to distinguish whether a piece of text is generated by a human or by ChatGPT [4]. By virtue of its human-level language proficiency, conducting a comparative analysis of the cognitive differences between ChatGPT and humans holds the potential to enhance our comprehension of AI. However, it is essential to clarify that when we refer to the concept of "cognition" in this work, we are not attributing human-like consciousness, thoughts, or perception to the model. Instead, we are using the term within the specific context of observing patterns and responses generated by the algorithm. We assume that language is the embodiment of cognition. Therefore, we treat ChatGPT and humans on an equivalent basis, analyzing their cognitive patterns manifested through language (see our ethics statement at the end).

Previous works frequently evaluated the ability of ChatGPT and other large language models (LLMs), based on benchmark datasets or well-crafted questions that are designed for specific tasks [5]. The limitations of such

✉ Erik Cambria
cambria@ntu.edu.sg

Rui Mao
rui.mao@ntu.edu.sg

Guanyi Chen
g.chen@ccnu.edu.cn

Xiao Li
xiao.li@abdn.ac.uk

Mengshi Ge
mengshi001@e.ntu.edu.sg

[1] Nanyang Technological University, Singapore, Singapore

[2] Central China Normal University, Wuhan, China

[3] University of Aberdeen, Aberdeen, UK

practices stem from their narrow focus, and potential data exposure issues, e.g., whether the testing data were used for pre-training, limiting their ecological validity.[1] To address these limitations, analyzing AI from the perspectives of psychology or cognitive science offers several advantages. This approach provides a holistic understanding of cognitive processes, identifies biases, and facilitates human-AI comparisons. Moreover, exploring models in a task-agnostic manner allows for a more comprehensive investigation into their cognitive functioning, contributing to a nuanced understanding that goes beyond predefined tasks and benchmarks. By adopting a cognitive science framework, researchers can uncover valuable insights, enhancing their interpretability and generalizability across diverse applications. However, it is important to acknowledge that traditional cognitive analysis primarily depends on one-on-one interviews, such as traditional diagnostic psychological tests [7]. Conducting large-scale research, particularly when analyzing extensive AI-generated text, becomes impractical using these methods. Consequently, a central challenge in this study revolves around identifying a useful linguistic device and leveraging efficient automated tools for cognitive analysis.

Text mining techniques, e.g., sentiment analysis [8], emotion detection [9], and depression detection [10], have been used to study human cognition. However, these works only yielded findings upon the statistics of the limited predicted label classes, e.g., the statistics of positive and negative labels, which does not help to gain insight into complex cognitive patterns. Topic modeling was also used to study cognition changes over a long period [11]. However, the generated topic words cannot reveal the cognitive process of a subject in a short time span because the topic words primarily function as a concise representation of the central theme of a document. Their selection is significantly influenced by the intended themes found in different documents. Thus, the aforementioned tools are sub-optimal for studying diverse cognitive patterns from topical documents.

In light of the above challenges, we propose to conduct a comparative analysis of metaphorical cognition[2] in ChatGPT and human minds, utilizing a metaphor processing tool and a large corpus. A metaphor is defined as a figurative language that uses one or several words to represent a different meaning rather than its literal meaning.[3] It is a creative language that reflects human cognition about concepts. According to the Conceptual Metaphor Theory [13, 14], metaphors reflect the concept mappings between target and source domains in human cognition systems. For example, humans likely use MONEY concepts to illustrate TIME to emphasize the value of time in their metaphorical expressions, such as "it *costs* a day to fix the machine".[4] In this case, TIME serves as the target concept, signifying the domain to which the metaphor is applied. Conversely, MONEY is identified as the source concept, representing the domain from which the metaphorical attributes are drawn. To elaborate, target concepts are those that the metaphors aim to elucidate or represent, while source concepts are the conceptual domains from which attributes or characteristics are borrowed to enhance understanding. The conceptual mapping, e.g., TIME IS MONEY,[5] signifies the cognition that attributes associated with MONEY, such as scarcity and preciousness, contribute to the understanding of TIME. Different people can leverage different source concepts to represent a target concept, e.g., LOVE IS JOURNEY or LOVE IS MAGIC. The examination of cognition through concept mappings, albeit in various forms, is a prevalent practice in the field of psychology (see "Cognition and Metaphor" section). Given the widespread use of metaphors in everyday language and the availability of an automated metaphor processing tool, metaphors emerge as an appropriate linguistic tool for examining the cognitive patterns displayed by both ChatGPT and humans within extensive data.

We leverage MetaPro[6] [15], a computational metaphor processing system to automatically parse concept mappings from a large English corpus, Human ChatGPT Comparison Corpus (HC3) [16]. MetaPro is used because, to the best of our knowledge, it is the only expert system that can parse metaphors and generate concept mappings for non-domain-specific texts from end to end [17]. HC3 contains parallel human- and ChatGPT-generated answers to the questions from multiple domains, including Reddit (open-domain), WikiQA (open-domain), Wikipedia (computer science), medical consultations (medicine), and StackExchange (finance). Thus, we categorize our comparative analysis,

[1] Ecological validity pertains to how well the design or evaluation setup aligns with the authentic work context of the user. It focuses on the accuracy with which the design or evaluation mirrors the pertinent characteristics of the interaction's ecology, capturing its context in the real world or environment [6].

[2] We define metaphorical cognition as the reflection of cognition through metaphors, encompassing elements such as the cognition of target concepts, source concepts, and their mappings.

[3] While dictionaries may contain the meanings of numerous conventional metaphors, their mere inclusion is not a feature to identify the metaphoricity of a lexical unit. According to Metaphor Identification Procedure [12], a metaphor is identified through the semantic contrast between its contextual and basic meanings. The basic meaning of a metaphor is typically more concrete, related to bodily action, more precise, and historically older.

[4] Italics denote metaphors; small capital words denote concepts.

[5] In this work, the representation of a concept mapping takes the form of "a target concept is a source concept".

[6] https://metapro.ruimao.tech

based on the concept mapping patterns observed between ChatGPT and humans in the domains of open-domain discourse, computer science, medicine, and finance. We come up with the following research questions in this work:

1. How do the metaphor usage preferences of humans and ChatGPT vary across different domains ("Findings in Metaphor Usage Preferences" section)?
2. What are the notable variations observed between ChatGPT and humans in their concept mappings ("Findings in Frequent Concept Mappings" section)?
3. Are the cognitive patterns of ChatGPT significantly different from that of humans ("Cognitive Pattern Comparison" section)?

Our main findings are summarized as follows:

1) Humans and ChatGPT exhibit dissimilar preferences in metaphor usage by different domains. For instance, humans demonstrate a greater tendency to employ metaphors in open-domain scenarios, whereas ChatGPT displays a higher inclination for utilizing metaphors in the finance domain. Both humans and ChatGPT do not extensively rely on metaphors when providing responses to medical inquiries. Furthermore, both humans and ChatGPT exhibit a predilection for employing verb metaphors over other parts of speech.

2) The frequent concept mappings observed in ChatGPT present considerable overlap with those of humans in the finance, computer science, and open domains. However, the medicine domain stands out with the most notable variations in frequent concept mappings, where the differential usage of cognition- and series of actions-related source concepts is the main dissimilarity between ChatGPT and humans.

3) ChatGPT has developed its distinctive concept mapping patterns through its training on human-generated corpora. This inference is derived from the disparate distribution of target and source concepts of ChatGPT in different conceptual subspaces, in contrast to the more uniformly distributed patterns and creativity observed in human-generated content. It is essential to recognize that potential algorithmic biases from ChatGPT's concept preferences should prompt caution in using text generated by such generative AI for training other AI systems. This precaution is warranted to prevent the inadvertent propagation of cognitive biases.

To sum up, the contribution of this work lies in the comparative analysis of metaphorical cognition between ChatGPT and human minds. Moreover, we introduce an innovative task-agnostic analysis methodology, which reveals potential algorithmic biases in ChatGPT when compared to the metaphorical cognition of humans in terms of concepts.

Related Work

Cognitive Research with NLP Techniques

Various NLP tasks have been employed in cognitive research to gain insights into human perception and interpretation of information. Examples of these tasks include sentiment analysis [18], emotion detection [19], topic modeling [11], depression detection [20], and suicidal ideation detection [21]. Researchers have utilized these NLP tools, which are oriented towards cognition and psychology, to analyze large-scale data in diverse fields, such as analyzing public opinions about COVID-19 vaccines [22], presidential election [23], climate change [24], wildfires [25], war [26], and mental health [27]. In these studies, classifiers were employed to assign labels to the research data. Cognitive patterns were summarized based on statistical analyses of these labels. However, many classifiers are limited in terms of the label classes they can provide, such as positive, negative, and neutral classes in sentiment analysis. These limited label classes restrict the depth of insights that can be gained beyond the predefined categories. Topic modeling-based methods [28, 29], on the other hand, offer the advantage of generating diverse sets of topic words for cognitive analysis. However, they often fail to uncover the specific cognitive processes or mechanisms underlying individual topics.

Additionally, these methods can be influenced by the intended themes present in the analyzed documents, making it challenging to generalize the conclusions reached on one theme to other themes. Finally, previous human-oriented concept mapping studies [30–32] focused on limited concepts, which cannot comprehensively reveal the cognition state of subjects.

Cognitive Research for AI Models

While there has been considerable focus on task-specific evaluations of LLMs [5], these efforts predominantly rely on benchmark datasets or task-oriented questions. However, these methods may fall short of capturing task-agnostic algorithmic biases when viewed through the lens of cognitive science. In the context of cognitive science, previous research has primarily focused on investigating AI personality, empathy, creativity, and the Theory of Mind (ToM). Ruane et al. [33] explored the impact of chatbot personality on user experience, showing how users perceive the personality of agents conveyed through textual interactions. Liu and Sundar [34] examined the effects of different empathic expressions, including sympathy, cognitive empathy, and affective empathy, on individuals' perceptions of the service and the chatbot, indicating that the expression of sympathy and empathy was more favorable than the provision of unemotional advice. Santo et al. [35] examined the intersections between Computational Creativity and Formal Learning Theory, discerning areas of convergence and divergence. Their analysis provided fresh insights into the interplay between AI and Computational Creativity. Moghaddam and Honey [36] developed a set of False-Belief questions to assess ToM capabilities of ChatGPT and GPT-4, finding that both models demonstrated a degree of ToM ability, while their performance remained inferior to that of humans. Hutson [37] provided an overview of the complexities involved in understanding LLMs such as ChatGPT, emphasizing their internal mechanisms, behaviors, and the challenges that researchers encounter in attempting to explain these systems. It was observed that the majority of these studies employed carefully designed prompts to identify specific behaviors or capabilities of LLMs.

Although advancements in computational metaphor processing techniques have led to significant insights into metaphorical cognition in recent research [38, 39], there has been a lack of research examining the cognitive patterns of AI through the lens of concept mappings, despite concept mappings being a crucial tool for human cognition and psychology analysis [7]. Previous human-oriented concept mapping studies [30–32] focused on very limited concepts, which cannot comprehensively reveal the cognition state of subjects. Prompts specifically crafted to test LLMs may not reflect how these models are used in everyday situations, raising concerns about the ecological validity of such experiments. Analyzing LLMs within the context of typical usage scenarios is more likely to yield objective and cognitively relevant insights.

Cognition and Metaphor

Concept mappings have been commonly used for psychological analysis, manifested in the forms, such as the word-association test[7] [40], the thematic apperception test[8] [41], and the Rorschach test[9] [42], because different concept mappings reflect the different cognitive patterns, personalities, and emotional functions in subconscious minds. Metaphors are characterized as linguistic expressions wherein words are employed to symbolize other concepts, diverging from the literal meanings of the words within the given context. The variations in metaphorical concept mappings from target to source domains among different subjects can also indicate distinctions in their cognitive processes [43] and behaviors [44]. In contrast to the above psychological tests involving concept mapping studies that depend on interviews or surveys based on questionnaires, the merit of investigating concept mappings derived from metaphors lies in the ability to gather genuine reactions from social media posts without disrupting the subjects. This approach also distinguishes itself from other task-specific evaluations of LLMs, as the analysis of cognition through metaphors and real-world text remains task-agnostic, ensuring ecological validity in our investigations. Thus, we comparatively analyze the cognition of ChatGPT and humans from metaphors.

MetaPro and Evaluation

MetaPro is used to obtain metaphoricity labels and concept mappings. It contains three technical components, aggregating the research outcomes in metaphor identification [45], metaphor interpretation [46], and concept mapping generation [47] tasks (see technical details and benchmark evaluations from the referred works).

Metaphor Identification The module was trained on the largest all word annotated dataset, VU Amsterdam Metaphor Corpus [48]. It uses multi-task learning and a novel gated bridging-based soft parameter sharing mechanism to learn sequential metaphor detection and part-of-speech (PoS) tagging together. Thus, MetaPro can detect metaphors on the token level.

Metaphor Interpretation Given an identified single-word metaphor in the former step, it uses WordNet [49] and a RoBERTa [50]-based masked word prediction mechanism to select the best-fit word from the identified metaphor's hypernyms and synonyms with the same PoS as the paraphrase. If the metaphor is detected as a multi-word expression by a rule set, the metaphoric multi-word expression is explained with the most coherent dictionary meaning at the end of the output with a clause.

Conceptualization The target and source concepts are generated from the paraphrase and the original metaphor, respectively, with a statistical learning algorithm and WordNet. The concepts are abstracted from WordNet hypernyms with different abstractness levels. MetaPro aims to deliver target and source concepts that can represent the major senses of the paraphrase and metaphor, meantime, keeping relatively more concrete, compared to other concept agents that have the same sense coverage. Finally, the concept mapping is formed as "a target concept is a source concept."

[7] A word association test involves the presentation of a stimulus word to a participant, who subsequently provides the initial word that comes to mind in response.

[8] The thematic apperception test is a projective psychological evaluation that requires individuals to furnish interpretations for scenes characterized by ambiguity.

[9] The Rorschach test is a projective technique used in psychological assessment, involving the individual's task of describing their interpretations of ten inkblots. These inkblots consist of a combination of black or gray elements and others featuring patches of color.

Table 1 The statistics of the original data and MetaPro-parsed data

	Open	Fin.	CS	Med	Overall
# Original questions	18,299	3,933	842	1,248	24,322
# Questions after MetaPro	18,218	3,925	796	1,119	24,058
# metaphors in human text	348,181	35,734	5,230	2,645	391,790
# metaphors in AI text	128,437	44,378	3,459	3,000	179,274

For example, given an input sentence, “she *devoured* his novels,” MetaPro first identifies “devoured” as a metaphor. Next, the metaphor is paraphrased into its literal counterpart “enjoyed,” yielding “she enjoyed his novels.” Finally, concept mapping PLEASURE IS BODILY_PROCESS is generated by abstracting the concepts of PLEASURE and BODILY_PROCESS from “enjoyed” and “devoured,” respectively. In essence, target and source concepts correspond to the mapping of distinct concept domains. The definitions of MetaPro-generated concepts can be viewed in WordNet [49]. When we observe variations in target concepts, it suggests that the subject, whether it is humans or ChatGPT, engages in a wide array of discussions about various concepts when addressing a question. Conversely, variations in source concepts indicate that the subject possesses different perceptions when it comes to these target concepts through the use of metaphors.

To evaluate the performance of MetaPro, we randomly sampled 100 sentences from ChatGPT- and human-generated text in the HC3 corpus, respectively (200 in total). These sentences contain at least a metaphor, which was identified by MetaPro. Next, three undergraduate students with psychology backgrounds are invited to evaluate the MetaPro-generated 263 concept mappings. The participants received guidance summarized from the Metaphor Identification Procedure [12] and Conceptual Metaphor Theory. Among the 263 concept mappings, 219 are correct (Fleiss’ kappa: 0.79). Besides the identified metaphors, the evaluators also indicated that there are 21 metaphors in the 200 sentences misclassified as literal by MetaPro (Fleiss’ kappa: 0.63). The above human evaluation results have been concurred upon by the majority (at least two evaluators).

Data and Statistics

As seen in Table 1, the original HC3 corpus [16] comprises a total of 24,322 questions, including 18,299, 3,933, 842, and 1,248 questions from the open domain, finance, computer science, and medicine, respectively. Next, we parsed the human and ChatGPT answers[10] from the HC3 corpus using MetaPro. Consequently, we preserved the responses provided by both humans and ChatGPT for a total of 24,058 questions. The selection was based on the presence of at least one detected metaphor in the answers (either from humans or ChatGPT). Therefore, our subsequent analysis focuses on the answers, consisting of 18,218 questions from the open domain, 3,925 questions from finance, 796 questions from computer science, and 1,119 questions from medicine. In all the examined domains, we collected a substantial number of metaphors from both humans and ChatGPT. The metaphors obtained varied across the domains, with the highest count of metaphors being 348,181 from humans in the open domain and the lowest number of 2,645 metaphors in medicine. ChatGPT yielded the highest number of 128,439 metaphors in the open domain and the lowest number of 3,000 metaphors in medicine.

Findings in Metaphor Usage Preferences

After parsing metaphors with MetaPro, we obtained 391,790 metaphors from human answers and 179,274 metaphors from ChatGPT answers. The breakdown statistics of metaphoricity are shown in Fig. 1. For open-domain questions (Fig. 1a), humans tend to employ metaphors more frequently compared

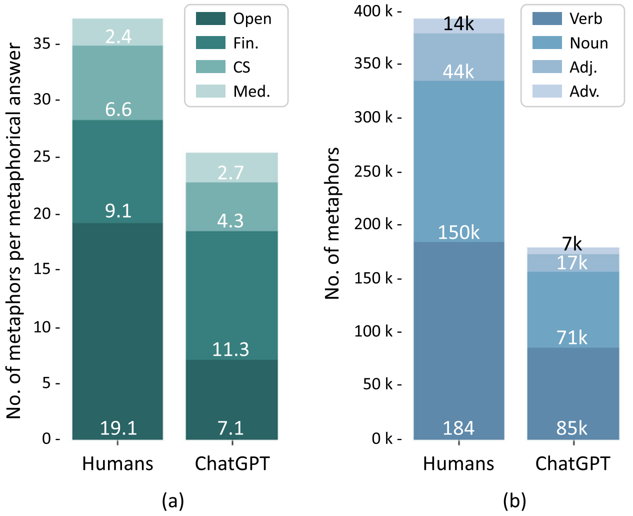


Fig. 1 The statistics of metaphors by different **a** question domains; **b** parts of speech

[10] Questions in the dataset were generated by humans only. Thus, we did not parse questions.

to other types of questions. Conversely, ChatGPT demonstrated a higher propensity for utilizing metaphors in the finance domain. For medical questions, both humans and ChatGPT used a similar number of metaphors, which has the lowest frequency among all domains. Thus, humans and ChatGPT have different metaphor usage preferences in different domains, while neither humans nor ChatGPT tends to rely on metaphors extensively when responding to medical inquiries. This can be attributed to the fact that medical knowledge often relies on evidence-based practices [51], necessitating a cautious approach that avoids ambiguity. Consequently, the utilization of metaphors to bridge concepts from source domains to explain those in target domains is often avoided to maintain clarity and precision. In contrast, metaphors as a useful tool for explaining abstract concepts have been frequently used by humans and ChatGPT in open and financial domains [52].

On the other hand, both humans and ChatGPT presented the most verbal metaphors in their answers in Fig. 1b. This is likely because verbs possess an inherent capability to effectively communicate concepts related to ACTION, MOVEMENT, PROCESS, and CHANGE, rendering them more adept at vividly expressing concepts and experiences compared to other parts of speech [53]. The adaptability and malleability of verbal metaphors enable the intricate and comprehensive articulation of ideas across diverse domains.

Findings in Frequent Concept Mappings

The most frequent source and target concepts as well as their mappings are shown in Table 2. ACT, ACTION, ACTIVITY, and MOTION emerge as the most prevalent concepts in both the source and target domains of the metaphoric expressions generated by both humans and ChatGPT. This observation aligns with our previous finding—the higher frequency of verb metaphors compared to other parts of speech. These very frequent concepts in metaphorical language usage can be attributed to their aptitude for capturing the inherent dynamism associated with these concepts. This observation can be linked to the embodied nature of human cognition systems [53], wherein our conceptual understanding is deeply intertwined with our physical experiences and actions. ChatGPT also simulates the language characteristics of human embodied cognition, where action-related concepts are frequently used to explain other abstract concepts in the metaphorical language of ChatGPT.

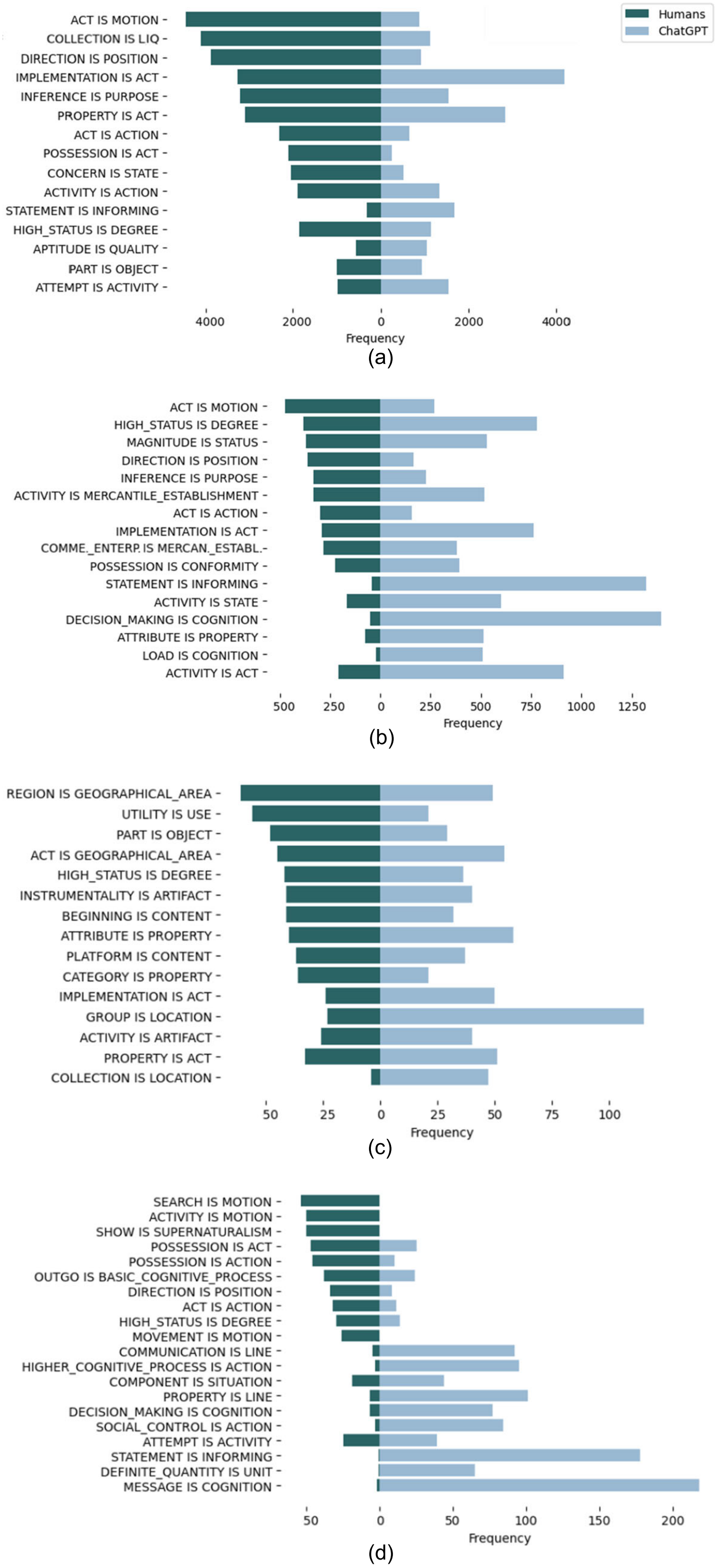
There are also overlaps in the most frequent concept mappings between humans and ChatGPT, e.g., INFERENCE IS PURPOSE, PROPERTY IS ACT, IMPLEMENTATION IS ACT, and HIGH_STATUS IS DEGREE. These overlapped and frequent concept mappings reflect the similarities in the way ChatGPT communicates concepts, compared with humans. Both subjects likely explain INFERENCE with PURPOSE,

Table 2 The lists of highly frequent concept mappings by human and ChatGPT

	Source	Target	Mapping
Humans	ACT	ACT	ACT IS MOTION
	ACTION	ACTIVITY	COLLECTION IS LIQ
	MOTION	ACTION	DIRECTION IS POSITION
	ACTIVITY	POSSESSION	IMPLEMENTATION IS ACT
	STATE	CHANGE_OF_STATE	INFERENCE IS PURPOSE
	POSITION	COMMUNICATION	PROPERTY IS ACT
	COMMUNICATION	QUALITY	ACT IS ACTION
	PRODUCTION	COLLECTION	HIGH_STATUS IS DEGREE
	LIQ	PROPERTY	POSSESSION IS ACT
	ARTIFACT	STATE	MAGNITUDE IS STATUS
ChatGPT	ACT	ACTIVITY	IMPLEMENTATION IS ACT
	ACTION	ACT	STATEMENT IS INFORMING
	ACTIVITY	IMPLEMENTATION	PROPERTY IS ACT
	COGNITION	ACTION	HIGH_STATUS IS DEGREE
	INFORMING	PROPERTY	ATTEMPT IS ACTIVITY
	MOTION	STATEMENT	INFERENCE IS PURPOSE
	STATE	POSSESSION	DECISION_MAKING IS COGNITION
	PROPERTY	PERCEPTION	ACTIVITY IS ACTION
	PRODUCTION	QUALITY	ACTIVITY IS ACT
	BCP	COMMUNICATION	ATTRIBUTE IS PROPERTY

LIQ denotes LARGE_INDEFINITE_QUANTITY. BCP denotes BASIC_COGNITIVE_PROCESS

Fig. 2 The most frequent concept mappings by answers in **a** open domain; **b** finance; **c** computer science; **d** medicine. LIQ denotes LARGE_INDEFINITE_QUANTITY. COMME._ENTERP. IS MERCAN._ESTABL. denotes COMMERCIAL_ ENTERPRISE IS MERCANTILE_ESTABLISHMENT



because communication is driven by the principle of relevance, where individuals aim to convey information that is most relevant to the listener's cognitive environment, emphasizing the common ground in the purpose over unnecessary details or processes [54]. Next, we will demonstrate the similarities and dissimilarities between the two subjects by frequent concept mappings.

We first obtained the most frequent 10 concept mappings from humans and ChatGPT, respectively, then demonstrate their union and statistics in Fig. 2 by different domains. In general, the domain exhibiting the most substantial cognitive disparity is medicine, as it encompasses a total of 20 distinct concept maps. In other words, when considering the top 10 frequently occurring concept mappings, there is no overlap between ChatGPT and humans within this domain. For other domains, the level of intersection between ChatGPT and human concept mappings is relatively close with a range of 15 to 16 concept mappings being in the unions. However, we also observe concept mappings that are common for ChatGPT, but rare for humans, e.g., STATEMENT IS INFORMING in the open domain, finance, and medicine (Fig. 2a, b, and d), DECISION_MAKING IS COGNITION in finance, and medicine (Fig. 2b and d), LOAD IS COGNITION in finance (Fig. 2b), COLLECTION IS LOCATION in computer science (Fig. 2c), PROPERTY IS LINE, SOCIAL_CONTROL IS ACTION, and MESSAGE IS COGNITION in medicine (Fig. 2d). These concept mappings highlight the different cognition of ChatGPT in different domains, compared to that of humans. Our speculative and hypothetical opinion is that ChatGPT likely compares DECISION_MAKING, LOAD, and MESSAGE as the concepts of COGNITION, showing that ChatGPT places great emphasis on allowing interlocutors to consider their point of view, while such projections are comparatively less common among humans.

By controlling the target concepts, it is possible to discern distinctive source concepts, stemming from both ChatGPT and humans. This approach is motivated by the understanding that metaphors serve as a means to convey intended meanings, whereby the intended meaning linked to the target domain is projected onto the metaphorical meaning associated with the source domain. Consequently, a given intended meaning (represented by a target concept) may give rise to diverse metaphorical expressions (represented by source concepts) from humans and ChatGPT. Thus, we first ranked the target concepts from humans and ChatGPT, respectively, according to frequency. We then selected the intersection of the most frequent target concepts to form a set with 10 concepts. Next, through these 10 target concepts, we found the corresponding frequent source concepts from the concept mappings of humans and ChatGPT, respectively. We plotted the target concepts, the retrieved source concepts, and the concept mapping frequencies in Fig. 3. The stream from the target to the source denotes the mapping and frequency.

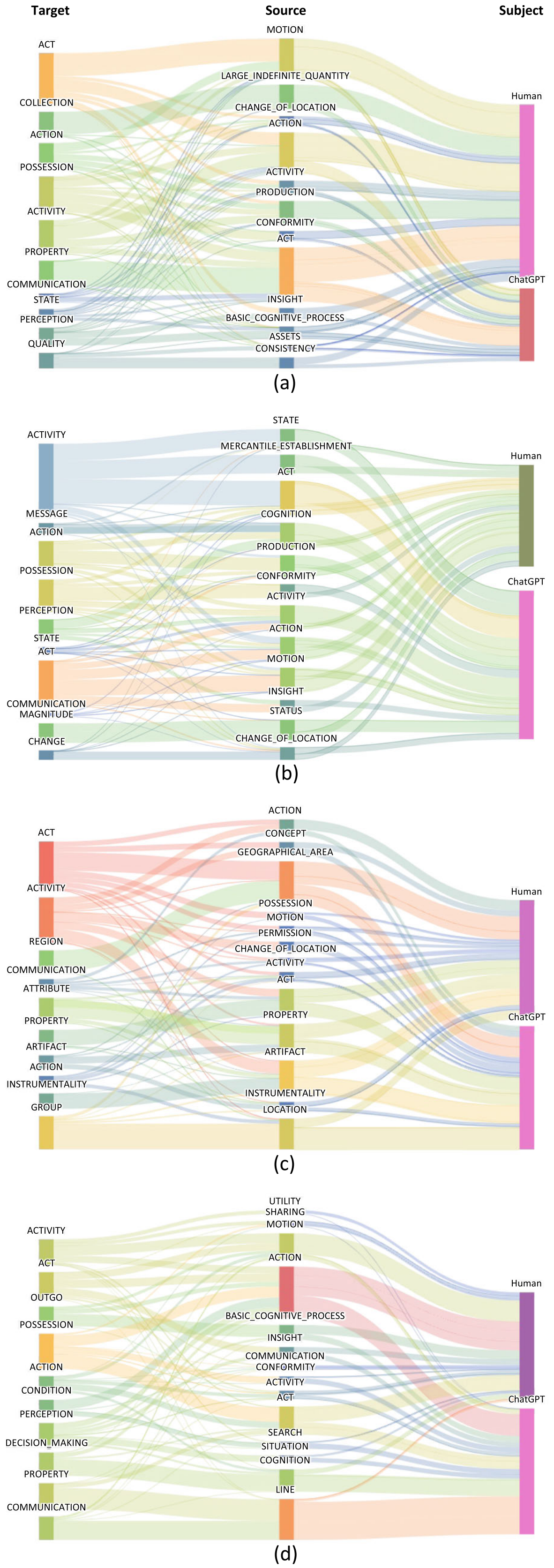


Fig. 3 The comparison of source concepts given the same target concepts in **a** open domain; **b** finance; **c** computer science; **d** medicine

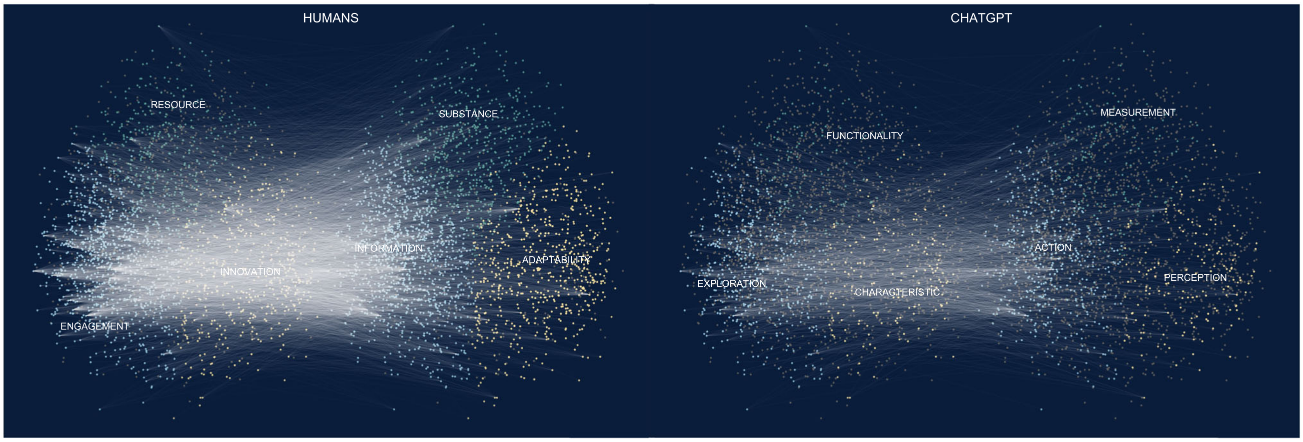


Fig. 4 The concept mapping pattern comparison between humans (left) and ChatGPT (right). For each subject, the left cluster denotes target concepts; the right cluster denotes source concepts. The gray dots denote inactivated concepts. Different bright colors denote different groups of activated concepts. Dots at the same positions in target and source clusters denote the same concept. Bolder lines denote more frequent concept projections from the target domain to the source domain

In the open domain (Fig. 3a), we can observe that the target concept, ACT, was frequently projected to the source of MOTION for humans. However, such a projection pattern is less common for ChatGPT, as the diameter of the pipe transporting from MOTION to humans is thicker than the pipe flowing from MOTION to ChatGPT. In contrast, both humans and ChatGPT share a common source concept of ACT, because the stream of ACT to humans and ChatGPT are comparable. Though ACT covers both physical and cognitive aspects of behavior, e.g., intention, agency, and volition, MOTION highlights the physical movement of behavior. Thus, given the same target, humans likely project them into both general (ACT) and concrete (MOTION) concepts, while ChatGPT often projects them to a general ACT concept in open domains.

In finance (Fig. 3b), target concepts, e.g., MESSAGE and PERCEPTION, are frequently projected to the source COGNITION. These projections are commonly observed in ChatGPT's responses, while humans tend to present such mappings less frequently. In the source concept domain, ChatGPT exhibits a preference for concepts such as STATE, ACT, and ACTIVITY. Conversely, humans demonstrate a higher inclination toward the use of the source concept MOTION. This disparity suggests that ChatGPT tends to rely on mental representations when elucidating concepts in the finance domain, whereas humans lean towards explanations rooted in bodily sensorimotor experiences.

In computer science (Fig. 3c), besides LOCATION, most source concepts are used equally by humans and ChatGPT. LOCATION is favored by ChatGPT, because it has a distinctive concept mapping GROUP IS LOCATION. Hence, the cognitive patterns behaved by humans and ChatGPT towards identical concepts exhibit a general resemblance in computer science, although ChatGPT tends to display a preference for employing spatial concepts associated with LOCATION when explaining the GROUP concept involving sets and collections.

In medicine (Fig. 3d), ChatGPT exhibits a preference for utilizing source concepts, COGNITION and LINE, while these concepts are relatively infrequent in the answers generated by humans. Therefore, the differential utilization of cognitive representations (COGNITION) and a connected series of actions (LINE) for expounding medical concepts serve as a distinguishing characteristic between ChatGPT and humans.

In summary, the frequent concept mappings of ChatGPT across various domains exhibit a general similarity to that of humans. However, upon closer examination of individual source concepts, subtle variations emerge within specific domains, indicating that ChatGPT differs slightly from human preferences when elucidating particular frequent concepts within those domains. Frequent concept mappings exhibit the most notable disparities between humans and ChatGPT, particularly in the medical domain. The distinctive utilization of source concepts related to cognition and series of actions is prominent in the text generated by ChatGPT.

Cognitive Pattern Comparison

We plot the concept mapping distributions of humans and ChatGPT to investigate if ChatGPT's cognitive patterns are significantly different from that of humans in Fig. 4.

First, we embedded all concepts from humans and ChatGPT in vector space via GloVe 50d[11] [55]. Then, the concepts

[11] https://nlp.stanford.edu/projects/glove/

were clustered into three groups with k-means [56], wherein the constituent concepts within the cluster bear a certain degree of resemblance to one another. Next, we plotted all source and target concepts in mirror space for humans and ChatGPT, respectively. The activated concepts of each subject (humans or ChatGPT) were highlighted with different bright colors to demonstrate their distributions in different subspace (clusters). Finally, we read the activated concepts in each cluster and came up with a concept that can best represent the major activated concepts for each cluster. In what follows, we describe key findings identified from Fig. 4.

Humans Are More Creative The cognitive patterns of ChatGPT are not significantly different from that of humans overall, as we can observe that there are activated concepts in different conceptual subspaces from both humans and ChatGPT. However, the collective intelligence of humans embodies stronger creativity than ChatGPT, because we can observe more activated source and target concepts and the associated mappings from humans than that of ChatGPT in terms of the number of nodes and edges.

ChatGPT Exhibits Own Concept Mapping Patterns For humans, the activated concepts exhibit a roughly even distribution across the three clusters. However, for ChatGPT, we observe relatively dense blue nodes and sparse green nodes among its activated concepts. It suggests that ChatGPT has its own special cognitive characteristics, such as concept preference in the target and source domains. While ChatGPT was trained on humans' corpora, it does not embody the collective cognitive patterns of the average human. Throughout its training process, ChatGPT has acquired distinct cognitive characteristics of its own.

Metaphorical Cognition Differences Are Also Presented in Conceptual Subspace In different subspaces, the representative concepts are also slightly different from humans. For example, in the yellow region of target concepts, many concepts in human cognition systems are related to INNOVATION, while ChatGPT has more blue target concepts related to CHARACTERISTIC. It shows that ChatGPT exhibits a tendency to employ metaphors when elucidating concepts related to CHARACTERISTIC, whereas humans demonstrate a preference for using metaphors to describe concepts linked to INNOVATION. Among yellow source concepts, ChatGPT likes to use PERCEPTION related concepts, while humans prefer ADAPTABILITY. Thus, humans commonly project the INNOVATION target domain to the ADAPTABILITY source domain, while the same subspace projection of ChatGPT is from CHARACTERISTIC to PERCEPTION. Furthermore, the projections observed in humans extend from ENGAGEMENT to INFORMATION (blue) and from RESOURCE to SUBSTANCE (green), while the projections in ChatGPT span from EXPLORATION to ACTION (blue) and from FUNCTIONALITY to MEASUREMENT (green).

In summary, contemplating the conceptual diversity observed in text generated by humans and the distinctive conceptual preferences manifested in text generated by ChatGPT, it becomes crucial to recognize the potential ramifications of employing ChatGPT-generated text for training subsequent AI systems. The amalgamation of AI- and human-generated corpora, without proper differentiation, introduces a transformative element in the distribution of human language and cognitive patterns. This shift in distribution implies that future AI models, when trained on such blended datasets, may incline towards emulating the idiosyncrasies of preceding AI text generation behaviors rather than capturing the nuances of human cognition. Without control, the text produced by AI systems has the potential to act as a contaminant, diluting the creativity and diversity inherent in human language.

Conclusion

We compared the metaphorical cognition of ChatGPT and humans with MetaPro. Our findings uncovered subtle discrepancies in the cognitive treatment of specific concepts, highlighting that ChatGPT, despite being pre-trained on human-generated corpora, does not fully embody the collective cognitive patterns of humans. Throughout its training, ChatGPT has developed distinctive "cognitive characteristics" of its own (algorithmic biases in concept selection). In addition to human-centric AI development [57], we hope our findings serve as a reminder to be cautious about using AI-generated content, as this may lead to the spread of potential cognitive biases.

In future work, we will extend the scope of our cognitive analysis. This expansion will involve examining the personalities associated with LLMs and their influence on human-LLM interactions [58, 59]. Additionally, we will explore strategies to mitigate conceptual biases within LLMs, potentially through the development of criteria for selecting high-quality pre-training data. Given the significant role of LLMs in human information acquisition, it is also crucial to equip LLMs with the ability to adopt diverse perspectives when engaging with different individuals. This approach aims to prevent the homogenization of human thought and to maintain its diversity.

Ethics Statement

This study seeks to investigate cognitive patterns manifested in both AI- and human-generated content, focusing on language patterns and preferences in metaphorical concept mappings. Our approach does not involve anthropomorphizing

AI; rather, it aims to analyze and compare concept mapping patterns, treating ChatGPT and humans equivalently. The assumption underlying this research is that language serves as a representation of cognition. We emphasize the equivalence of analysis between ChatGPT and humans, examining cognitive patterns expressed through language without distinguishing algorithmic inference mechanisms from human cognition for writing simplicity. It is crucial to note that our work does not assert that AI possesses human-like cognitive abilities.

Given that the dataset [16] used in this study was sourced from publicly available content, obtaining explicit consent from individual users is unfeasible. However, we underscore that our analysis strives to offer general insights into cognitive patterns, avoiding the targeting of specific individuals. To mitigate potential analytical biases, we employ parallel answers from ChatGPT and humans for the same questions. This ensures a balanced and unbiased approach to the exploration of the research questions.

Author Contributions Rui Mao: conceptualization, methodology, software, formal analysis, visualization, writing—original draft, writing—review and editing. Guanyi Chen: formal analysis, visualization, writing—review and editing. Xiao Li: investigation, validation, writing—review and editing. Mengshi Ge: investigation, validation, writing—review and editing. Erik Cambria: conceptualization, data curation, writing—review and editing, supervision.

Funding This research/project is supported by the Ministry of Education, Singapore, under its MOE Academic Research Fund Tier 2 (STEM RIE2025 Award MOE-T2EP20123-0005). Guanyi Chen is supported by the Hubei Provincial Key Laboratory of Artificial Intelligence and Smart Learning and the National Language Resources Monitoring and Research Center for Network Media of Central China Normal University in Wuhan, China.

Data Availability The data used in this study will be made available upon request.

Declarations

Conflict of Interest The authors declare no competing interests.

References

1. Shidiq M. The use of artificial intelligence-based Chat-GPT and its challenges for the world of education; from the viewpoint of the development of creative writing skills. In: Proceeding of International Conference on Education, Society and Humanity; 2023;1:353–357
2. Méndez G, Gervás P. Using ChatGPT for story sifting in narrative generation. In: Proceedings of The 14th International Conference on Computational Creativity; 2023
3. Qin C, Zhang A, Zhang Z, Chen J, Yasunaga M, Yang D. Is ChatGPT a general-purpose natural language processing task solver? 2023 arXiv:2302.06476
4. Soni M, Wade V Comparing abstractive summaries generated by ChatGPT to real summaries through blinded reviewers and text classification algorithms. 2023 arXiv:2303.17650
5. Mao R, Chen G, Zhang X, Guerin F, Cambria E. GPTEval: a survey on assessments of ChatGPT and GPT-4. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING), Torino, Italia; 2024. pp. 7844–7866
6. Hartson R, Pyla PS. Chapter 14 - rigorous empirical evaluation: preparation. In: Hartson R, Pyla PS (eds.) The UX Book; 2012. pp. 503–536. Morgan Kaufmann, Boston
7. Rapaport D, Gill M, Schafer R. Diagnostic psychological testing: the theory, statistical evaluation, and diagnostic application of a battery of tests. 1946:2
8. Crossley SA, Kyle K, McNamara DS. Sentiment analysis and social cognition engine (SEANCE): an automatic tool for sentiment, social cognition, and social-order analysis. Behav Res Methods. 2017;49:803–21.
9. Naskar D, Singh S.R, Kumar D, Nandi S, Rivaherrera EOdl. Emotion dynamics of public opinions on Twitter. ACM Trans Inf Syst (TOIS). 2020;38(2):1–24
10. Orabi AH, Buddhitha P, Orabi MH, Inkpen D. Deep learning for depression detection of Twitter users. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: from Keyboard to Clinic; 2018. pp. 88–97
11. Priva UC, Austerweil JL. Analyzing the history of cognition using topic models. Cognit. 2015;135:4–9.
12. Pragglejaz G. MIP: a method for identifying metaphorically used words in discourse. Metaphor Symb. 2007;22(1):1–39.
13. Lakoff G, Johnson M. Metaphors we live by; 1980
14. Lakoff G. The contemporary theory of metaphor. Metaphor and thought; 1993:202–251
15. Mao R, Li X, He K, Ge M, Cambria E. MetaPro online: a computational metaphor processing online system. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (System Demonstrations); 2023;3:127–135
16. Guo B, Zhang X, Wang Z, Jiang M, Nie J, Ding Y, Yue J, Wu Y. How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. 2023 arXiv:2301.07597
17. Ge M, Mao R, Cambria E. A survey on computational metaphor processing techniques: from identification, interpretation, generation to application. Artif Intell Rev. 2023;56:1829–95.
18. Mao R, Liu Q, He K, Li W, Cambria E. The biases of pre-trained language models: an empirical study on prompt-based sentiment analysis and emotion detection. IEEE Trans Affect Comput. 2023;14(3):1743–53.
19. Fan C, Lin J, Mao R, Cambria E. Fusing pairwise modalities for emotion recognition in conversations. Inf Fusion. 2024;106:102306.
20. William D, Suhartono D. Text-based depression detection on social media posts: a systematic literature review. Procedia Comput Sci. 2021;179:582–9.
21. Ji S, Pan S, Li X, Cambria E, Long G, Huang Z. Suicidal ideation detection: a review of machine learning methods and applications. IEEE Trans Comput Soc Syst. 2020;8(1):214–26.
22. Karami A, Zhu M, Goldschmidt B, Boyajieff HR, Najafabadi MM. COVID-19 vaccine and social media in the US: exploring emotions and discussions on Twitter. Vaccines. 2021;9(10):1059.
23. Xia E, Yue H, Liu H. Tweet sentiment analysis of the 2020 US presidential election. In: Companion Proceedings of the Web Conference 2021; 2021. pp. 367–371
24. Duong C, Liu Q, Mao R, Cambria E. Saving earth one tweet at a time through the lens of artificial intelligence. In: 2022 International Joint Conference on Neural Networks (IJCNN); 2022. pp. 1–9

25. Duong C, Raghuram VC, Lee A, Mao R, Mengaldo G, Cambria E. Neurosymbolic AI for mining public opinions about wildfires. Cognit Comput. 2023;16:1531–53.
26. Garcia MB, Cunanan-Yabut A. Public sentiment and emotion analyses of Twitter data on the 2022 Russian invasion of Ukraine. In: 2022 9th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE); 2022. pp. 242–247. IEEE
27. Guntuku SC, Preotiuc-Pietro D, Eichstaedt JC, Ungar LH. What twitter profile and posted images reveal about depression and anxiety. In: Proceedings of the International AAAI Conference on Web and Social Media; 2019;13:236–246
28. Wu X, Pan F, Nguyen T, Feng Y, Liu C, Nguyen CD, Luu AT. On the affinity, rationality, and diversity of hierarchical topic modeling. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2024;38:19261–19269
29. Wu X, Dong X, Nguyen TT, Luu AT. Effective neural topic modeling with embedding clustering regularization. In: International Conference on Machine Learning; 2023. pp. 37335–37357. PMLR
30. Taleb NN. The black swan: the impact of the highly improbable. 2007:2
31. Glette-Iversen I, Aven T. On the meaning of and relationship between dragon-kings, black swans and related concepts. Reliab Eng Syst Saf. 2021;211:107625.
32. Arrese Á. The use of ‘bubble’ as an economic metaphor in the news: the case of the ‘real estate bubble’ in Spain. Lang Commun. 2021;78:100–8.
33. Ruane E, Farrell S, Ventresque A. User perception of text-based chatbot personality. In: Chatbot Research and Design: 4th International Workshop; 2021. pp. 32–47. Springer
34. Liu B, Sundar SS. Should machines express sympathy and empathy? Experiments with a health advice chatbot. Cyberpsychology Behav Soc Netw. 2018;21(10):625–36.
35. Santo L.E, Cardoso A, Wiggins G. Theoretical learning creators and creative scientists. In: 13th International Conference on Computational Creativity. 2022. Association for Computational Creativity
36. Moghaddam SR, Honey CJ. Boosting theory-of-mind performance in large language models via prompting. 2023. arXiv:2304.11490
37. Hutson M. How does ChatGPT ‘think’? Psychology and neuroscience crack open AI large language models. Nat. 2024;629(8014):986–8.
38. Mao R, He K, Ong CB, Liu Q, Cambria E. MetaPro 2.0: computational metaphor processing on the effectiveness of anomalous language modeling. In: Findings of the Association for Computational Linguistics: ACL; 2024. pp. 9891–9908. Association for Computational Linguistics, Bangkok, Thailand
39. Manro R, Mao R, Dahiya L, Ma Y, Cambria E. A cognitive analysis of CEO speeches and their effects on stock markets. In: Proceedings of the 5th International Conference on Financial Technology (ICFT), 2024. Singapore
40. Prosser J, Cohen LJ, Steinfeld M, Eisenberg D, London ED, Galynker II. Neuropsychological functioning in opiate-dependent subjects receiving and following methadone maintenance treatment. Drug Alcohol Depend. 2006;84(3):240–7.
41. Wiggins JS. Paradigms of personality assessment. 2003
42. De Vos GA. Boyer LB. Symbolic analysis cross-culturally: the rorschach test; 2021.
43. Han S, Mao R, Cambria E. Hierarchical attention network for explainable depression detection on Twitter aided by metaphor concept mappings. In: Proceedings of the 29th International Conference on Computational Linguistics (COLING); 2022. pp. 94–104
44. Mao R, Zhang T, Liu Q, Hussain A, Cambria E. Unveiling diplomatic narratives: analyzing United Nations Security Council debates through metaphorical cognition. In: Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci); 2024. pp. 1709–1716. Rotterdam, the Netherlands
45. Mao R, Li X. Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification. Proceed AAAI Conf Artif Intell. 2021;35(15):13534–42.
46. Mao R, Li X, Ge M, Cambria E. MetaPro: a computational metaphor processing model for text pre-processing. Inf Fusion. 2022;86–87:30–43.
47. Ge M, Mao R, Cambria E. Explainable metaphor identification inspired by conceptual metaphor theory. Proceed AAAI Conf Artif Intell. 2022;36(10):10681–9.
48. Steen GJ, Dorst AG, Herrmann JB, Kaal A, Krennmayr T, Pasma T. A method for linguistic metaphor identification: from MIP to MIPVU. 2010;2010:14.
49. Fellbaum C. WordNet: an electronic lexical database. 1998
50. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. RoBERTa: a robustly optimized BERT pretraining approach. 2019. arXiv:1907.11692
51. Sackett DL. Evidence-based medicine. In: Seminars in Perinatology; 1997;21:3–5. Elsevier
52. Mao R, Du K, Ma Y, Zhu L, Cambria E. Discovering the cognition behind language: financial metaphor analysis with MetaPro. In: 2023 IEEE International Conference on Data Mining (ICDM); 2023. pp. 1211–1216. IEEE
53. Jamrozik A, McQuire M, Cardillo ER, Chatterjee A. Metaphor: bridging embodiment to abstraction. Psychon Bull Rev. 2016;23:1080–9.
54. Sperber D, Wilson D. Relevance: communication and cognition. 2nd ed. 1995
55. Pennington J, Socher R, Manning CD. GloVe: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing; 2014. pp. 1532–1543
56. Lloyd S. Least squares quantization in PCM. IEEE Trans Inf Theory. 1982;28(2):129–37.
57. Cambria E, Mao R, Chen M, Wang Z, Ho SB. Seven pillars for the future of artificial intelligence. IEEE Intell Syst. 2023;38(6):62–9.
58. Zhu L, Li W, Mao R, Pandelea V, Cambria E. PAED: zero-shot persona attribute extraction in dialogues. In: Proceedings of the 61st annual meeting of the association for Computational Linguistics (ACL); 2023;1:9771–9787
59. Zhu L, Mao R, Cambria E, Jansen BJ. Neurosymbolic AI for personalized sentiment analysis. In: Proceedings of international conference on human-computer Interaction (HCII); 2024. Washington DC, USA

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.