NeuroImage 338 (2026) 122108

Contents lists available at ScienceDirect

NeuroImage

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

Brain-inspired large model mindreading

Jia Jin [a,1], Yunsong Hu [a,1], Zhongfeng Wang [a], Yu Pan [a], Baojun Ma [a], Bo Dong [b,c], Jianmin Dong [d], Guanxiong Pei [b,c,*]

[a] Key Laboratory of Brain-Machine Intelligence for Information Behavior (Ministry of Education and Shanghai), School of Business and Management, Shanghai International Studies University, Shanghai, 200083, China
[b] Zhejiang Laboratory of Philosophy and Social Sciences - Laboratory of Intelligent Society and Governance, Zhejiang Lab, Hangzhou, 311121, China
[c] Development Strategy and Cooperation Center, Zhejiang Lab, Hangzhou, 311121, China
[d] Research Center for High Efficiency Computing Infrastructure, Zhejiang Lab, Hangzhou, 311121, China

ARTICLE INFO

Keywords:
Multimodal large language models
Theory of mind
Visual question answering
Mentalization
Brain-inspired intelligence

ABSTRACT

Multimodal large language models (MLLMs) demonstrate significant limitations in visual Theory of Mind (ToM) abilities compared to humans. Investigating the neural mechanisms underlying human mind-reading not only addresses critical gaps in visual ToM research but also provides valuable insights for MLLM optimization. Using fMRI data from 83 participants, we systematically compared neural processing patterns between two conditions in visual ToM tasks: MLLM-incorrect/human-correct (MLLMI) and mutually correct (MLLMC). The questionnaire results indicated that, compared with the MLLMC condition, human confidence was significantly lower under the MLLMI condition, and expectations for MLLM performance were also lower. Natural language analysis revealed that, relative to MLLM responses, human responses were more closely aligned with the question context, more concise, and exhibited greater certainty. Neuroimaging results indicated significantly stronger activation in bilateral precuneus and middle temporal gyrus in MLLMI. Furthermore, we observed enhanced functional connectivity in networks associated with task coordination and attention allocation. Leveraging these neural signatures, we constructed multiple prediction models for decoding the two conditions (MLLMI vs. MLLMC), among which the 2-layer Transformer model achieved the highest classification accuracy of 78.6%. Extending these findings, we propose the Knowledge-Thinking-Adaptation (KTA) framework, which integrates memory retrieval, divergent thinking, and multi-level attention mechanisms to provide a potential roadmap for future work in developing AI systems with human-like visual ToM capabilities.

1. Introduction

Theory of Mind (ToM), defined as the ability to attribute mental states to others, is often regarded as being fundamental to defining humanity and referred to as the art of mind-reading(Leslie et al., 2004; Premack and Woodruff, 1978; Vogeley et al., 2001). ToM is the origin of human empathy, self-awareness, and morality, serving as the cornerstone of human cognition and social interaction(Leslie et al., 2004; Strachan et al., 2024b). The multimodal large language models (MLLMs), with their ability for multimodal comprehension and logical reasoning, are highly anticipated to foster computational/machine ToM abilities. They are regarded as essential tools for the achievement of Artificial General Intelligence (AGI) and the enhancement of natural human-computer interaction(Zhang et al., 2024a). While formal definitions of computational/machine ToM remain debated, researchers have adapted human psychological assessment paradigms to benchmark MLLMs' mental state reasoning capabilities(Collins et al., 2024; Hagendorff et al., 2023; Mao et al., 2024). Recent advances reveal MLLMs' remarkable proficiency in text-based ToM tasks, particularly in detecting linguistic nuances like irony, implicit meaning, and false belief scenarios(Kosinski, 2024; Strachan et al., 2024b). This progress is partly enabled by the inherent psychological transparency of human language, where verbal expressions often involuntarily disclose authentic intentions and emotional states through lexical choices and discourse patterns(Hauptman et al., 2023; Vine et al., 2020). However, authentic human social cognition critically depends on multimodal integration –

* Corresponding author at: Zhejiang Laboratory of Philosophy and Social Sciences - Laboratory of Intelligent Society and Governance, Zhejiang Lab, 2880# Wenyixi Road, Hangzhou, 311121, China.
E-mail address: pgx@zhejianglab.org (G. Pei).
[1] Jia Jin and Yunsong Hu contribute equally to the article.

https://doi.org/10.1016/j.neuroimage.2026.122108
Received 5 May 2026; Received in revised form 17 June 2026; Accepted 7 July 2026
Available online 7 July 2026

especially the capacity to decode mental states through implicit visual cues (e.g., micro-expressions, contextual body language) that demand sophisticated cross-modal reasoning far beyond textual analysis. This fundamental discrepancy explains why current MLLMs significantly underperform humans in visual ToM tasks(Kuang et al., 2025; Schulze Buschoff et al., 2025b), highlighting a critical frontier in artificial social intelligence development.

The past two years have witnessed explosive growth in MLLMs research for visual intelligence(Schulze Buschoff et al., 2025a). Overall, MLLMs currently employ two primary technological approaches for the execution of visual reasoning tasks. The first approach utilizes multimodal pre-training to align visual-linguistic embeddings through dedicated adapter modules. However, this method is resource-intensive, requiring substantial computational resources and an extensive collection of image-text pairs for learning (e.g., Flamingo's 2B-scale training corpus(Alayrac et al., 2022)). The lack of high-quality corpora related to ToM abilities(Guo et al., 2023), in combination with the prohibitive cost of manual annotation, has hindered the advancement of visual ToM abilities. The second strategy adopts a language-mediated pipeline that processes visual inputs through textual representations – converting images into descriptive captions (potentially with contextual examples) for subsequent analysis by frozen LLMs. While lightweight compared to pre-training, prominent implementations like PICa(Yang et al., 2022) and Img2LLM(Guo et al., 2023) primarily address perceptual tasks (e.g., object detection, fine-grained recognition) rather than cognitive-level challenges. These limitations reveal the critical performance plateau in MLLMs' visual ToM capabilities under conventional paradigms(Liang et al., 2024; Małkiński and Mańdziuk, 2023), necessitating paradigm-shifting innovations. A highly promising path for the development of artificial intelligence that can rival or surpass human abilities is to conduct an in-depth exploration of the brain's neural mechanisms and transfer these insights into AI models(Cai et al., 2023; Kosinski, 2024; Xu and Poo, 2023). This brain-inspired AI perspective offers a promising alternative pathway to overcome existing architectural constraints.

Brain-inspired AI seeks to derive its inspiration from the brain to advance both general AI development and high-performance computing (Zhang et al., 2024b, 2020). The transformative potential of this approach is best exemplified by the attention mechanism – a neurocognitive-inspired innovation that has revolutionized AI architectures through transformer models, which utilize self-attention to establish new benchmarks in neural network design. Parallel advancements are occurring in neuromorphic hardware, where biologically-inspired chips emulate the brain's parallel processing through neuron/synapse-mimetic circuits, achieving substantial efficiency gains over conventional computing architectures(Cai et al., 2023). MLLMs further demonstrate this biomimetic progression by replicating human-like multitasking capabilities through strategies including: learning neural patterns of multisensory integration to enhance reasoning (Huang et al., 2023); adopting the brain's unified storage-computation architecture to address energy constraints in training(Mehonic and Kenyon, 2022); and implementing value-aligned reinforcement learning mechanisms(Glickman and Sharot, 2024). However, a critical gap persists in brain-inspired approaches specifically targeting the enhancement of visual ToM abilities.

Developing brain-inspired MLLMs with mind-reading capabilities requires deep understanding of the neurocognitive architecture underlying human mental state attribution. Neuroscientific research has identified specialized brain regions governing ToM processes, including the medial prefrontal cortex (mPFC) for self-other differentiation, precuneus (PCu) for autobiographical processing, temporoparietal junction (TPJ) for representing others' perspectives, and middle temporal gyrus (MTG) for associative-divergent thinking(Balgova et al., 2024; Carrington and Bailey, 2009; Frith and Frith, 2006; Molenberghs et al., 2016; Schlaffke et al., 2015; Schurz et al., 2014). Large-scale meta--analytic evidence reveals that ToM cognition predominantly engages cortical midline structures and temporoparietal networks, with peak activation observed in the mPFC-anterior cingulate complex that extends posteriorly along the midline to incorporate PCu and mid-cingulate regions(Schurz et al., 2021).Crucially, this activation topography shows substantial overlap with the default mode network (DMN), where the mPFC-PCu axis forms a functional core supporting both ToM operations and self-referential cognition through dynamic large-scale network interactions(Dadario and Sughrue, 2023). Despite the significant and fruitful progress made by neuroscientists, it remains insufficient to fully support and reflect the underlying mechanisms and brain functional networks involved in the interplay between brain regions during the execution of visual ToM abilities in humans. In particular, there is a notable absence of analysis on the disentanglement of vision-specific mechanisms and the origins of the brain-machine gap, which necessitates urgent research to support the optimization and enhancement of AI models(Kunda, 2018).

This study employed functional magnetic resonance imaging (fMRI) to acquire high-resolution neuroimaging data, specifically examining neural activation differences between incorrect and correct conditions during the completion of visual ToM tasks by the MLLM. Tentatively drawing on these exploratory observations, we suggest potential optimization pathways for MLLM's visual ToM abilities, with the objective of the development and deployment of brain-inspired, flexible, and efficient large models. This study not only contributes to the innovation of the brain-inspired machine psychology scientific paradigm but also establishes three pivotal contributions: (1) providing exploratory insights into the neural mechanism of human visual ToM through large-scale neuroimaging analysis; (2) presenting exploratory evidence regarding neural responses to more complex visual ToM items that current MLLMs fail to pass, while developing prediction models that link such neural signatures to MLLM performance metrics; (3) introducing the Knowledge-Thinking-Adaptation (KTA) framework as a potential roadmap for future work to enhance mind-reading skills.

2. Results

This study employs a tripartite analytical framework (Fig. 1) integrating behavioral, brain imaging, and feature layers. The results from the behavioral layer primarily reflect the behavioral manifestations of human-machine visual ToM capabilities, with a particular focus on the differences in NLP (Natural Language Processing) evaluation metrics (entropy, lexical richness, semantic familiarity, semantic certainty, and semantic similarity) during task completion in MLLMI condition (MLLM incorrect, humans correct) and MLLMC condition (both MLLM and humans correct). The brain imaging layer provides insights into the brain activation patterns and network functional connectivity when performing tasks involving contrasting conditions. The goal is to explore the underlying neural mechanisms of human visual ToM and offer valuable insights for enhancing the ToM abilities of MLLM. The feature layer, based on cognitive neuroscience indicators, constructs cognitive computational prediction models [Convolutional Neural Networks (CNN), Long Short-Term Memory Networks (LSTM) and Transformer models] to decode brain states (MLLMI vs. MLLMC), while also evaluating the importance and stability of neural features in prediction using Shapley additive explanation (SHAP).

2.1. Behavioral layer results

Analysis of image familiarity revealed that participants had previously seen an average of 6.3% of the total images used, indicating low familiarity with the visual ToM task stimuli. A two-sample *t*-test on response confidence levels showed significantly lower confidence in the MLLMI condition ($M = 3.64$, $SD = 1.14$) compared to the MLLMC condition ($M = 4.09$, $SD = 1.06$), $t_{(28)} = -8.07$, $p < 0.001$. Similarly, participants' expectations of MLLM performance were significantly lower in the MLLMI condition ($M = 3.71$, $SD = 1.16$) than in the MLLMC condition ($M = 4.02$, $SD = 1.10$), $t_{(28)} = -6.76$, $p < 0.001$.

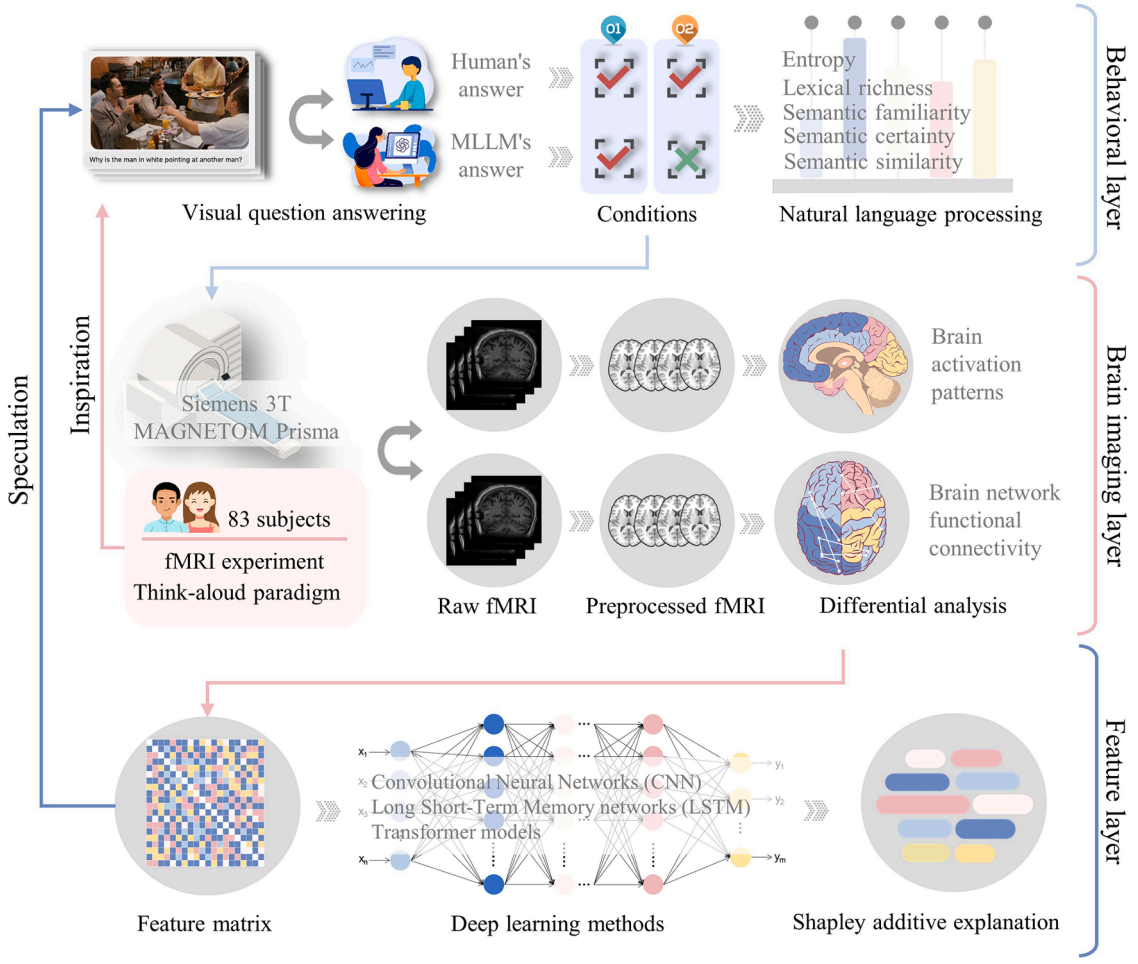

Fig. 1. Overall research and data analysis framework. The framework consists of three layers. (1) Behavioral layer: human participants and MLLMs each provided responses to visual question answering (VQA) tasks. Responses were compared across experimental conditions and analyzed using natural language processing metrics. (2) Brain imaging layer: 83 participants underwent functional MRI scanning while performing a VQA task under a think-aloud protocol. Raw fMRI data were preprocessed to extract regional activation patterns and functional connectivity differences between conditions. (3) Feature layer: multimodal features derived from behavioral and neuroimaging data were integrated into deep learning models (e.g., CNNs, LSTMs, Transformers). Model outputs were interpreted using SHAP to identify predictive and neurocognitively relevant features.

NLP analyses were conducted on MLLM-generated and human answers in both the pre-test (Figure S1, Table S1) and the fMRI experiment (Fig. 2, Table S2). For the behavioral data from the fMRI experiment, two-way analyses of variance (ANOVAs) assessed the effects of performance (MLLMI vs. MLLMC) and respondent (MLLM vs. Human) on entropy, lexical richness, semantic familiarity, semantic certainty, and semantic similarity. For entropy, no significant main effects were observed for either performance [$F_{1,56} = 3.43$, $p = 0.069$] or respondent [$F_{1,56} = 0.46$, $p = 0.49$]. However, a significant interaction effect was identified [$F_{1,56} = 4.81$, $p = 0.030$]. In terms of lexical richness, a significant main effect of respondent was found [$F_{1,56} = 854.9$, $p < 0.001$], accompanied by a significant interaction effect [$F_{1,56} = 5.12$, $p = 0.027$]. For semantic familiarity, significant main effects were found for both performance [$F_{1,56} = 9.59$, $p = 0.003$] and respondent [$F_{1,56} = 70.8$, $p < 0.001$], whereas no significant interaction effect was observed [$F_{1,56} = 2.71$, $p = 0.106$]. For semantic certainty, a significant main effect of respondent was found [$F_{1,56} = 328.6$, $p < 0.001$]. The main effect of performance [$F_{1,56} = 1.55$, $p = 0.218$] and the interaction effect [$F_{1,56} = 0.77$, $p = 0.386$] were not significant. For semantic similarity, significant main effects were found for both performance [$F_{1,56} = 4.29$, $p = 0.043$] and respondent [$F_{1,56} = 15.43$, $p < 0.001$], whereas no significant interaction effect was observed [$F_{1,56} = 0.003$, $p = 0.954$] (Fig. 2). The Tukey post-hoc analyses results are presented in the Supplementary Materials (Table S2). Word cloud maps visually display the word selection, frequency, and emotional tone of responses generated by both the MLLM and human participants (Figure S2).

2.2. *Brain imaging layer results*

A brain activation analysis was conducted to examine the neural representation differences between MLLMI and MLLMC conditions. MLLMI exhibited heightened activation compared to MLLMC in multiple areas: the bilateral PCu gyrus, bilateral MTG, left angular gyrus, right superior temporal gyrus, bilateral cuneus, and left posterior cingulate gyrus (Table 1; Figs. 3 and S3). Additionally, brain network analyses revealed that MLLMI demonstrated higher functional connectivity than MLLMC. Specifically, there was increased connectivity from the DMN to the visual network (VN) and the dorsal attention network (DAN), and from the salience network (SN) to the cerebellar network (CN) (Table 2; Figs. 3 and S4). To account for potential confounding effects of task difficulty and complexity, subjective ratings from the questionnaire were incorporated as covariates. The resulting activation patterns remained largely consistent with the original findings (Table S3), demonstrating significant clusters in the bilateral PCu and MTG. These findings suggest that the neural differences between MLLMI and MLLMC are not fully attributable to the higher difficulty or complexity of the

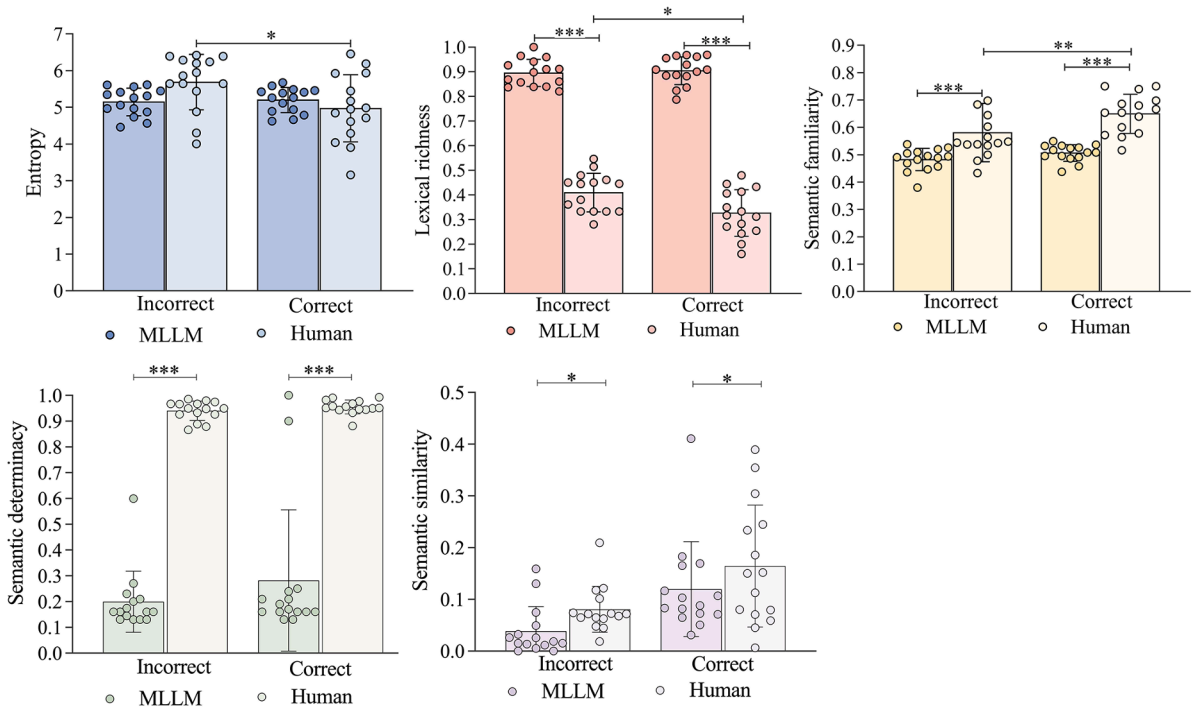

Fig. 2. Two-way ANOVA results of the effect of performance (MLLMI and MLLMC) and respondent (MLLM and Human) on NLP evaluation metrics in fMRI experiment.
Notes: *p<0.05, **p<0.01, ***p<0.001.

Table 1
Group-level activation differences between MLLMI and MLLMC.

Brain region	Hemi	Montreal neurological institute (*X Y Z*)			Cluster size (mm^3)	T
Precuneus gyrus	L	−3	−60	39	5643	5.12
Middle temporal gyrus	L	−45	−51	12	3348	4.97
Angular gyrus	L	−60	−57	27	3078	4.89
Precuneus gyrus	R	9	−60	39	3996	4.82
Middle temporal gyrus	R	51	−27	−9	2970	4.58
Superior temporal gyrus	R	48	−18	−6	1161	4.46
Cuneus	L	0	−78	39	1161	4.28
Cuneus	R	9	−72	39	1134	4.08
Post cingulum gyrus	L	−3	−51	27	1539	3.70

MLLMI items, but are instead associated, to some extent, with the state of MLLM failure.

Furthermore, we conducted a trial-by-trial parametric modulation analysis based on two types of ratings collected from a post-scanning questionnaire: confidence in one's own response and judgement about the current MLLM's ability to answer correctly. High confidence was associated with activation in the right dorsal and medial prefrontal gyrus, the left MTG, the PCu gyrus, the right angular gyrus, and the right superior temporal gyrus (Table S4 and Fig. 3). However, no significant results were found in association with judgment.

2.3. Feature layer results

The performance metrics of four deep learning models are presented in Table 3. The 2-layer Transformer model achieved the highest classification accuracy of 78.6%, with sensitivity of 81.8% and specificity of 75.4%. The receiver operating characteristic (ROC) curve illustrates the Transformer's classification performance across varying thresholds, with an area under the curve (AUC) of 0.89 (Fig. 4).

We performed SHAP analysis to identify the features that most significantly contributed to classification. The key contributing features include the functional connectivities of Salience.ACC - Language.IFG (R), DefaultMode.LP (L) - Visual.Lateral (L), DefaultMode.MPFC - Language.pSTG (R), and DorsalAttention.IPS (L) - DefaultMode.LP (L). In contrast, the predictive contribution of single-trial BOLD signals is relatively weak. To investigate the tipping points at which feature effects on model prediction change, Generalized Additive Model (GAM) was used to fit the relationship between feature values and their SHAP values. The results demonstrate that the following functional connectivities exhibit adequate fit and show distinct tipping points differentiating the MLLMI and MLLMC conditions: Salience.ACC – Language.IFG (R) ($R^2 = 0.84$), DefaultMode.LP (L) – Visual.Lateral (R) ($R^2 = 0.86$), and DorsalAttention.IPS (L) – DefaultMode.LP (L) ($R^2 = 0.79$) (Fig. 4).

To move beyond post-hoc SHAP explanations and directly assess the dependence of decoding performance on specific neural systems, we performed network-level ablation combined with permutation testing. As illustrated in Fig. 4d, removing the DMN-related and visual-related functional connections resulted in significant decreases in accuracy compared to the size-matched random null distributions ($p = 0.03$ and $p = 0.01$, respectively). This exploratory finding suggests that decoding performance may depend on the cross-network connectivity involving the DMN and visual systems. Notably, ablating the regional anatomical features (local activations in PCu and MTG) did not cause a significant decrease in performance. This demonstrates that while local regions exhibit activation differences, the classification of brain states (MLLMI vs. MLLMC) appears to be more strongly driven by large-scale network interactions rather than isolated regional characteristics.

3. Discussion

Our findings corroborate existing literature(Li et al., 2024a, 2024b; Schulze Buschoff et al., 2025b) by demonstrating a substantial

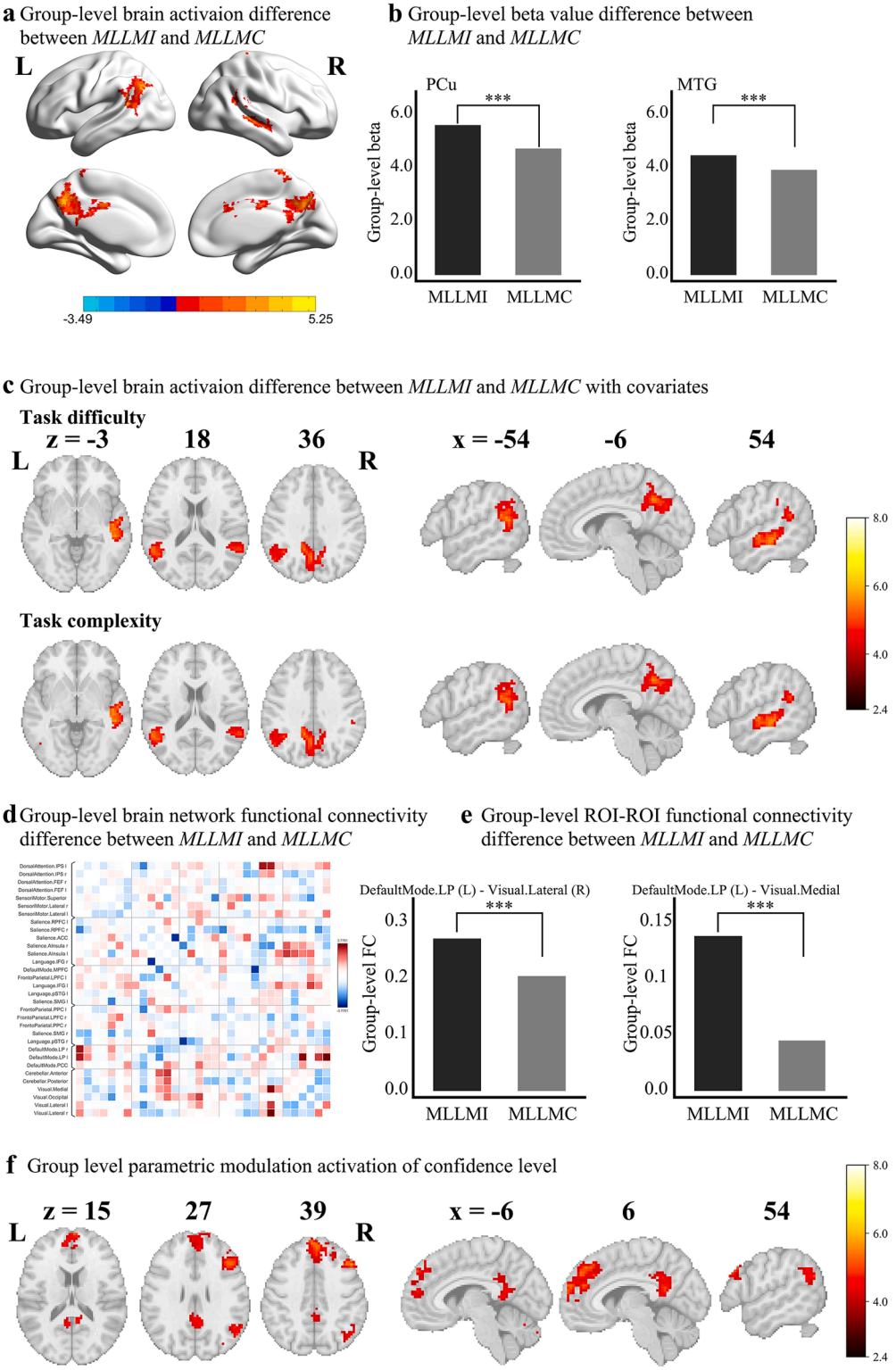

Fig. 3. Results of brain imaging. a: Group-level activation difference between MLLMI and MLLMC. The warm color indicates that MLLMI activation is higher than MLLMC. b: Group-level beta value difference between MLLMI and MLLMC within PCu and MTG. c: Group-level activation difference between MLLMI and MLLMC after controlling for task difficulty and complexity ratings as covariates. d: Group-level brain network functional connectivity difference between MLLMI and MLLMC. The warm color indicates that functional connectivity of MLLMI is higher than that of MLLMC. e: Group-level ROI (Region of interest) to ROI functional connectivity difference between MLLMI and MLLMC. f: Group-level parametric modulation activation of confidence level.
Notes: L: left hemisphere; R: right hemisphere. ***$p<0.001$.

Table 2
Group-level network functional connectivity differences between MLLMI and MLLMC.

ROI (MNI)	to	ROI (MNI)	T	*p*
DefaultMode.LP (L) −39, −77, 33		Visual.Lateral (R) 38, −72, 13	3.78	0.0003
DefaultMode.LP (L) −39, −77, 33		Visual.Medial 2, −79, 1	3.60	0.0005
Salience.ACC 0, 22, 35		Language.IFG (R) 54, 28, 1	−3.36	0.0012
DefaultMode.MPFC 1, 55, −3		Language.pSTG (R) 59, −42, 13	−3.31	0.0015
DorsalAttention.IPS (L) −39, −43, 52		DefaultMode.LP (R) 47, −67, 29	3.28	0.0016
DorsalAttention.IPS (L) −39, −43, 52		DefaultMode.LP (L) −39, −77, 33	3.22	0.0019
Salience.AInsula (R) 44, 13, 1		Cerebellar.Anterior 0, −63, −30	2.89	0.0050
DefaultMode.LP (L) −39, −77, 33		Visual.Lateral (L) −37, −79, 10	2.87	0.0054
Language.IFG (L) -51, 26, 2		Visual.Occipital 0, −93, −4	2.80	0.0065

Note: MNI: Montreal neurological institute coordinate.

Table 3
Performance comparison of four deep learning models with their respective best hyperparameters.

Model	Accuracy	Sensitivity	Specificity	F1-score
2-layer CNN	0.727	0.956	0.500	0.778
2-layer LSTM	0.765	0.788	0.743	0.77
1-layer Transformer	0.774	0.852	0.697	0.790
2-layer Transformer	0.786	0.818	0.754	0.792

performance gap between MLLMs and humans in visual ToM tasks. This disparity was observed not only in quantitative accuracy metrics but also in qualitative information processing capabilities. NLP analysis revealed that, compared with MLLM responses, human responses were more closely aligned with the question context, more concise, and exhibited higher certainty. Entropy index measurements further confirmed the effectiveness of the experimental manipulation, showing significantly greater cognitive complexity in the MLLMI condition compared to the MLLMC condition, consistent with the behavioral findings of reduced human confidence and lower expectations of MLLM performance in the MLLMI condition. These behavioral findings motivated our subsequent exploratory neurocognitive investigation into neural responses to more complex visual ToM items that current MLLMs fail to pass.

3.1. Cognitive neural mechanisms of visual ToM

Our neuroimaging analysis provided exploratory evidence of distinct neural processing patterns between MLLMI and MLLMC conditions, with particularly robust bilateral PCu activation observed during MLLMI scenarios. As a core neural substrate of ToM abilities, the PCu supports multiple critical cognitive functions: (1) visual-spatial mental imagery for perspective-taking, enabling the assessment of others' viewpoints; (2) information integration for constructing logically coherent social narratives; and (3) complex situation processing through episodic memory retrieval(Cavanna and Trimble, 2006; Dadario and Sughrue, 2023; Kulakova et al., 2013; Schurz et al., 2014; Tanglay et al., 2022). The pronounced PCu activation observed in MLLMI conditions may suggest that its role extends beyond basic perspective judgment, potentially encompassing the cognitive demands involved in interpreting challenging social scenarios. Neurocognitive evidence indicates this region facilitates the integration of prior knowledge and autobiographical memories with incoming social information, subsequently updating working memory representations to guide appropriate cognitive and behavioral responses(Wardell and Palombo, 2024). Together, these findings point to a dynamic memory integration and updating mechanism that may support advanced ToM capabilities and higher-order cognitive synthesis.

Our neuroimaging results further suggest significantly stronger bilateral MTG activation in MLLMI versus MLLMC conditions. As a key node in the ToM network(Nummenmaa et al., 2012; Schurz et al., 2017; Xu et al., 2019), the MTG appears to play an important role in social cognition, potentially through its dual functions in semantic processing and novel association formation(Balgova et al., 2024; Ren et al., 2020). During complex social reasoning tasks, this region may facilitate the inhibition of superficial semantic interpretations, establishment of distant conceptual connections for information reorganization, and generation of creative conceptual frameworks that transcend conventional thinking patterns(Davey et al., 2016; Jung-Beeman, 2005; Ren et al., 2020). This associative-divergent cognitive style may enhance semantic processing efficiency by enabling conceptual expansion and categorical knowledge updating(Brod et al., 2016; Ren et al., 2020) - mechanisms critical for advanced ToM operations(Grosse Wiesmann et al., 2017; Leshinskaya and Thompson-Schill, 2020). Specifically, the MTG may support the inference of latent event relationships in visual social scenarios by constructing non-obvious semantic linkages (Boccadoro et al., 2019). The robust MTG activation observed in challenging MLLMI conditions particularly highlights its involvement in the nonlinear, leap-like characteristics of human reasoning - cognitive features that may allow for flexible reinterpretation of social information beyond surface-level interpretations. Together, these exploratory findings suggest the MTG's role in mediating the complex, non-linear thought processes that may distinguish some aspects of human social cognition from current MLLM capabilities.

It is worth noting that in this study, the y-coordinates of the bilateral MTG activation clusters differed significantly, suggesting that they may not represent the same functional region. Notably, the left hemisphere cluster (MNI: −45, −51, 12) appeared to be located near the TPJ (Balgova et al., 2024). The TPJ is considered a core hub of the ToM network and is thought to play a specific role in belief reasoning, rather than being primarily involved in general attentional or semantic processing (Balgova et al., 2024; Molenberghs et al., 2016; Saxe and Kanwisher, 2013). More specifically, the TPJ has been found to be critically engaged when inferring beliefs that diverge from reality (i.e., false beliefs) in others (Geng and Vossel, 2013), a finding that has been robustly replicated in meta-analyses (Schurz et al., 2014). Moreover, the left TPJ has been shown to encode bidirectional signals of belief–fact matching or mismatching (Apperly et al., 2004; Ciaramidaro et al., 2007; Doricchi et al., 2022; Schurz et al., 2014). Based on this converging anatomical and functional evidence, it is plausible that identifying the left hemisphere activation cluster observed in this study as the left TPJ may reflect participants' in-depth processing of others' mental states during the MLLMI condition, potentially corresponding to a more refined representation of social intentions and belief information.

Our findings suggest that the differential activation patterns observed in key brain regions, including bilateral PCu and MTG, may reflect distinct neurocognitive processes (e.g., episodic memory retrieval and novel semantic association formation) engaged during challenging visual ToM tasks where MLLMs typically underperform. To elucidate the neural architecture underlying these complex social-cognitive processes, we conducted large-scale brain network analyses examining functional connectivity patterns across distributed neural systems. The results reveal two particularly predictive network configurations: (1) dynamic interactions between the DMN, DAN, and VN; and (2) coordinated activity between the SN, CN, and LN. The cognitive computational modeling suggests that these network-based features appear to show considerable predictive power while also demonstrating relatively high stability across individuals.

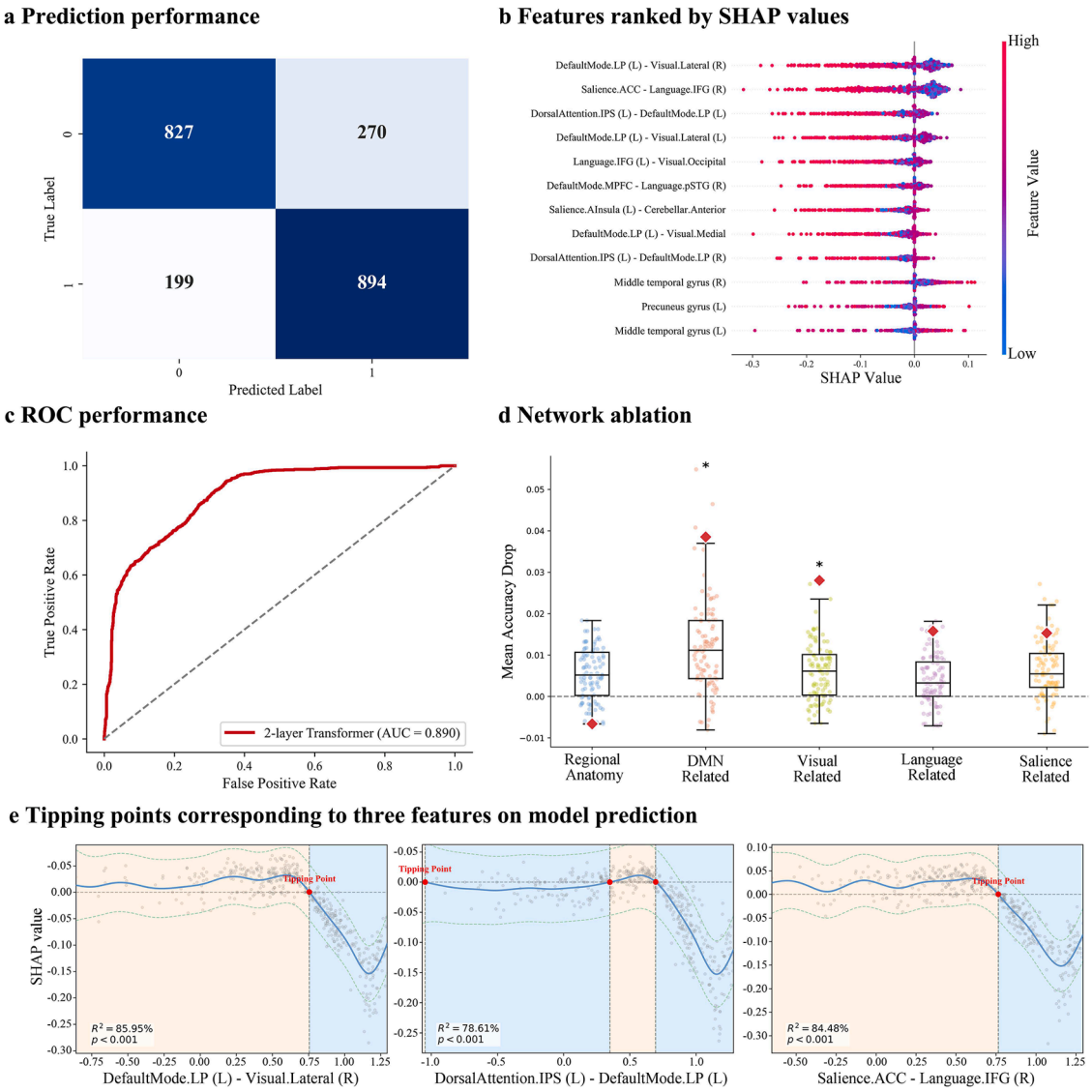

Fig. 4. SHAP analysis results. a and c: Confusion matrix of the Transformer prediction results and AUC of the ROC curve. b: Features ranked by SHAP values. Each dot in the plot corresponds to a sample. For a given feature, the horizontal axis shows the sample's SHAP value, and the color indicates the sample's feature value. A positive SHAP value corresponds to a classification of MLLMC, while a negative value corresponds to MLLMI. d: Red diamonds denote the true mean accuracy drop when all features belonging to a specific functional network are systematically removed from the prediction model. To control for biases introduced by the varying number of features within each network, size-matched random null distributions were generated through 100 permutations. Boxplots overlaid with colored jittered points illustrate the spread of these random null distributions. The horizontal dashed line at zero indicates no change in classification accuracy relative to the baseline model (* indicates $p < 0.05$). e: Tipping points for three features in model prediction. The vertical axis represents SHAP values, reflecting the magnitude and direction of each feature's contribution to the model output. Positive values (upper region) signify a positive impact, while negative values (lower region) denote a negative influence. The horizontal axis shows the range of the feature's values, progressing from low to high, with the corresponding SHAP value trends highlighting nonlinear effects. The blue solid line illustrates the trend of SHAP values as the feature changes, and the green dashed line represents the confidence interval, indicating the reliability of the model's trend predictions. Red markers denote critical points where the marginal contribution of the feature transitions between positive and negative, indicating key shifts in its influence on the model output.

Specifically, compared to MLLMC, the DMN, DAN, and VN appear to exhibit heightened functional connectivity in the condition of MLLMI. While traditional neurocognitive frameworks often conceptualize the DMN as a task-negative network that exhibits antagonistic interactions with task-positive networks like the DAN(Chang et al., 2024; Kucyi et al., 2020), our data challenge this dichotomous perspective. Specifically, our findings suggest that the dorsal medial subsystem of the DMN - crucial for mentalizing and memory processes - can effectively coordinate with the perceptual systems of the DAN and VN when serving the unified cognitive goal of visual ToM tasks. This synergistic interaction may imply that successful ToM reasoning depends not on the suppression of any particular network, but rather on the flexible coordination of multiple distributed systems working in concert.

Our findings suggest robust functional coupling and temporal synchronization among the SN, CN, and LN. The SN appears to serve as a critical attentional modulator(Chen et al., 2016), dynamically regulating the interplay between CN-mediated visuospatial processing and LN-supported linguistic encoding. As a multifunctional hub, the CN appears to not only govern sensory-motor integration (Fernandez-Iriondo et al., 2021; Wagner and Luo, 2020) but also orchestrate higher-order cognitive operations including visuospatial cognition, emotional regulation, and executive control(Brissenden et al.,

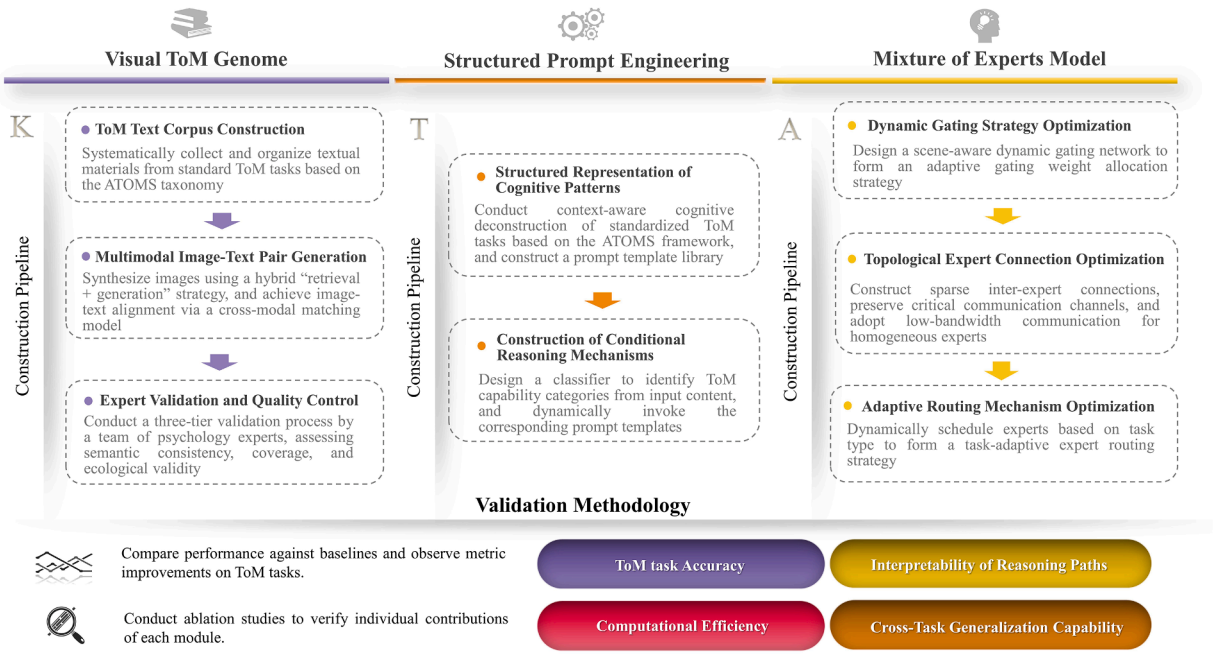

Fig. 5. The KTA framework. The framework integrates three synergistic components: (1) knowledge representation optimization through constructing a structured Visual ToM Genome that enables post-training acquisition of ToM knowledge, significantly improving accuracy on frequent social cognition tasks; (2) thinking process optimization via neurally-inspired prompt engineering that transforms ATOMS-based ToM priors into computational cognitive templates, enabling MLLMs to emulate human-like associative and divergent thinking for discovering implicit relationships in complex visual ToM scenarios; and (3) adaptation mechanism optimization employing a dynamic mixture of experts model architecture that implements neuroplasticity principles through on-demand module activation, autonomous reasoning path adjustment, and resource-efficient allocation.

2018). This network may enable context-appropriate allocation of visuospatial attention while facilitating the translation of visual information into symbolic representations and narrative structures. The LN, conversely, appears to specialize in linguistic organization - processing textual information and generating coherent narrative descriptions of visual stimuli(Yablonski et al., 2024). Notably, comparative analysis between MLLMI and MLLMC conditions suggests task-dependent modulation of these network interactions: while SN—CN connectivity strengthens to meet task demands, SN-LN coupling shows relative suppression. This adaptive reconfiguration may reflect the brain's capacity for optimal distribution of limited attentional resources, potentially influencing the speed and efficiency of cognitive processing during complex ToM tasks. Such dynamic network adjustments point to the sophisticated neural mechanisms underlying human social cognition that current MLLMs have yet to replicate.

3.2. *Brain-inspired roadmap for future work*

Inspired by the neural mechanisms of the human brain, this study proposes an innovative Knowledge-Thinking-Adaptation (KTA) framework (Fig. 5). It is worth noting that there are multiple possible functional interpretations of the neural basis. Reverse inference, as an important research tool, offers valuable insights for the design of artificial intelligence models but does not yield definitive conclusions (Poldrack, 2006, 2011; Young and Saxe, 2009). Therefore, the KTA framework is an exploratory framework grounded in neuroscientific findings, whose core value lies in providing a brain-inspired roadmap for subsequent model construction and engineering validation.

Knowledge representation optimization: The neural mechanisms underlying ToM autobiographical memory provide critical insights for constructing MLLMs knowledge systems, thereby enhancing their cognitive capabilities. The pronounced activation of the PCu and the coordinated engagement of the DMN-DAN-VN suggest that humans may leverage goal-directed mechanisms to retrieve prior knowledge from autobiographical memory when performing complex ToM tasks, facilitating novel task representation. To develop MLLMs with equivalent cognitive capacities, this study proposes a systematic framework based on the "Abilities in Theory of Mind Space" (ATOMS)(Conway and Bird, 2018) to develop a Visual ToM Genome. The framework includes three standardized processes: (1) ToM text corpus construction: Using the systematic categorization provided by the ATOMS (which covers 7 core mental states and 39 dimensions of ToM sub-abilities), textual materials can be collected from canonical ToM tasks, including but not limited to the hinting task(Corcoran, 2003), the false belief task (Bernstein et al., 2011; Wimmer and Perner, 1983), the recognition of faux pas(Baron-Cohen et al., 1999), the strange stories(Happé, 1994; White et al., 2009), and the unexpected transfer (Sally-Anne) task (Baron-Cohen et al., 1985). This phase aims to ensure comprehensive coverage of all ATOMS-defined ability dimensions. (2) Multimodal image-text pair generation: A hybrid data acquisition strategy is employed, which involves semantic search to retrieve candidate images from existing visual databases and using generative models to synthesize contextually relevant images. Pretrained cross-modal matching models (e.g., Contrastive Language-Image Pre-training(Pan et al., 2022), CLIP) can be used to automate preliminary image-text alignment. (3) Expert validation and quality control: A review panel consisting of at least three psychological experts is formed to implement a three-tier verification protocol. This includes reviewing the semantic consistency between images and text, evaluating the coverage of ToM ability dimensions, and verifying the ecological validity of cognitive tasks. This process aims to remove representation biases caused by automated generation, ensuring the construct validity of the dataset. The Visual ToM Genome can provide structured cognitive priors for post-training the model, significantly enhancing robustness and domain generalization in novel ToM tasks while improving interpretability.

Thinking process optimization: The MTG plays a crucial role in the neural mechanisms underlying associative and divergent thinking, which are essential for enhancing the higher-order reasoning abilities of

MLLMs. In comparison to MLLMC, MLLMI exhibits significantly greater activation in the MTG, revealing the central role of associative thinking and divergent thinking in remote semantic associations and the discovery of latent cues. Inspired by this insight, the present study proposes a methodology for mapping human cognitive patterns onto the reasoning framework of MLLMs through structured prompt engineering. This approach comprises two key components: (1) Structured representation of cognitive patterns:: By integrating the systematic characterization of 39 ToM abilities within the ATOMS framework, a prompt template grounded in cognitive science can be constructed. The template parameterizes human cognitive patterns through a four-dimensional situational modeling approach, encompassing the entity-relation graph, role-state matrix, scenario semantic frame, and social positioning mapping. This structured representation stems from the contextualized cognitive deconstruction of standardized ToM ability dimensions, such as the hinting task(Corcoran, 2003) and the false belief task(Bernstein et al., 2011; Wimmer and Perner, 1983). (2) Construction of conditional reasoning mechanisms:: Utilizing classifiers, the model first determines the ToM ability category and subsequently calls the appropriate prompt template based on the classification results. This process guides the model in simulating the human associative-divergent thinking process. The core of this mechanism lies in systematically building AI's deep cognitive capabilities, enabling it to precisely capture reasoned connections in visually complex structures and thereby infer beliefs, desires, and intentions.

Adaptation mechanism optimization: The human brain's SN demonstrates a multi-layered attention control mechanism that offers systematic insights for optimizing the mixture-of-experts (MoE) architecture of MLLMs, particularly for enhancing adaptive capabilities. Three key neurocognitive principles offer actionable design guidelines: (1) Dynamic gating strategy optimization: From the perspective of expert weight distribution, the brain's visual information prioritization mechanism during complex ToM tasks informs dynamic gating strategies for MoE systems. When complex social scenarios are detected, the model can automatically increase the activation weight of visual experts, such as vision encoders based on Vision Transformer(Zhou et al., 2024) or CLIP, while suppressing redundant computations from secondary modality experts. (2) Topological expert connection optimization:From the perspective of functional network coordination, the dynamic coupling mechanisms between the SN, CN, and LN reveal how functional networks achieve efficient communication through modulation of connection strength. This insight inspires MoE architecture to adopt a topologically optimized expert connection model, such as constructing a sparse expert interconnection in a small-world network, retaining critical channels while employing low-bandwidth communication for homogeneous experts. (3) Adaptive routing mechanism optimization:From the perspective of computational power optimization, MoE can mimic the SN's dynamic routing mechanism to handle different types of tasks. For simple visual pattern recognition, sparse local attention experts are used, while for tasks requiring higher-order cognitive reasoning, the model switches to global attention experts and activates the working memory cache module. This approach ensures accuracy in visual ToM task processing while reducing computational consumption.

3.3. *Primary contributions and research limitations*

Currently, there is a notable lack of research on the functional limitations of MLLMs from a neuroscience perspective, which may pose a challenge to transfer insights from brain science into AI algorithms and architectures. The primary contributions of this study can be summarized as follows: Theoretically, this study investigates the mapping patterns of the visual ToM differences between humans and machines within brain regions, and explores the deeper reasons for the gap between human and MLLM abilities. This helps to address the gap in research on the neural mechanisms of brain-inspired visual ToM. Methodologically, this study systematically investigates the limitations of large models using tools from experimental psychology, develops a cognitive computational prediction model and explores the potential of neuroscience metrics as benchmarks for evaluating MLLMs. This contributes to the scientific paradigm of machine psychology(Strachan et al., 2024b), supporting broader interdisciplinary research on machine behavior. Pragmatically, it proposes a brain-inspired KTA framework, which may provide a roadmap for enhancing MLLM's key abilities in memory retrieval, divergent thinking, and multi-level attention mechanisms, potentially contributing to the comprehensive enhancement of mind-reading abilities in MLLM.

While our study introduces innovative neurocomputational methods for enhancing MLLMs' visual ToM capabilities, several important limitations warrant consideration. Firstly, the current paradigm's reliance on static image stimuli precludes examination of temporal dynamics in neural processing - a critical constraint given that real-world social cognition fundamentally operates as a time-varying process. Future investigations should incorporate dynamic visual stimuli (e.g., video-based ToM tasks) along with a larger set of stimulus materials to enable spatiotemporal modeling of neural mechanisms, thereby improving MLLMs' adaptability to ecologically valid social scenarios. Secondly, the exclusive use of fMRI data limits our understanding of complementary neural signals across different spatial and temporal scales. Multimodal neuroimaging approaches combining electrophysiology (EEG/MEG), optical imaging, and magnetic resonance techniques could provide a more comprehensive characterization of cross-modal mapping mechanisms while revealing complementary information embedded in different signal modalities. Addressing these limitations through temporally sensitive paradigms and multimodal circuit analysis represents a crucial next step for advancing both our theoretical understanding of ToM and the development of truly human-like artificial intelligence systems. Thirdly, the sample primarily consisted of undergraduate and graduate students from Shanghai, China, all of whom were raised in a collectivist cultural context and completed the experimental tasks in Chinese. Future research should seek to include participants from a wider range of cultural backgrounds and language groups in order to assess the cross-group generalizability of the present findings. Finally, our experimental design was limited to trials on which human participants answered correctly, contrasting MLLMI and MLLMC conditions to isolate neural representations underlying human performance advantages over MLLMs. This design, while internally rigorous, limits a full understanding of how task difficulty and the broader variability of human ToM performance shape the observed effects. Future research should therefore incorporate a broader range of outcome types, including scenarios in which both humans and MLLMs respond incorrectly.

4. Conclusions

Our research suggests that MLLMs continue to lag behind human capabilities in visual ToM tasks. Through exploratory neurocognitive analysis, we identified neural patterns associated with complex visual ToM items that MLLMs currently underperform, including: PCu activation patterns associated with episodic memory retrieval, MTG involvement in associative-divergent thinking, coordinated dynamics of DMN-DAN-VN, and SN-mediated attentional switching between CN and LN. Although these neural features also reflect the cognitive demands of complex ToM reasoning, rather than solely MLLM failure per se, their reliable decoding (78.6% accuracy) supports their potential as predictive indicators for evaluating artificial cognitive systems. By synergistically integrating memory retrieval, divergent thinking, and multi-level attention mechanisms, we suggest the KTA framework may offer a potential roadmap for future work in constructing AI systems with human-like visual ToM abilities.

5. Online methods

5.1. Participants

The study recruited 83 healthy right-handed undergraduate and graduate students (mean age = 23.04 ± 3.11 years; range = 18–33 years) from Shanghai International Studies University. Research indicates that ToM ability follows an inverted U-shaped trajectory across the lifespan, developing progressively during childhood, peaking in early to middle adulthood, and declining in old age (Hughes et al., 2019; Li et al., 2025; Moran, 2013). The age range of participants in the present study corresponds to the peak period of ToM functioning, which may reflect an optimal mode of human mentalizing and offer valuable insights for the development of AI systems. All participants were free of psychiatric, physiological, or pain disorders, neurological conditions, or MRI contraindications through MRI safety-screening questions. They had normal or corrected-to-normal vision, and normal color vision. Following quality control procedures, 10 participants were excluded from final analysis due to excessive head motion or structural brain abnormalities (e.g., arachnoid cysts). The study protocols adhered to the Declaration of Helsinki(Williams, 2008) and received institutional approval from the Ethics Committee of Key Laboratory of Brain-Machine Intelligence for Information Behavior, Shanghai International Studies University (Approval No.IRB2023BC034). Written informed consent was obtained from each participant prior to the fMRI experiment. Each participant received compensation of RMB 150 (approximately US $ 21) for their involvement.

5.2. Materials

Image-question pairs were selected from the Visual Commonsense Reasoning Dataset(Yin et al., 2021; Zellers et al., 2019), which primarily assess ToM abilities(Apperly and Butterfill, 2009) including inference of beliefs, desires, intentions, emotions, and thoughts that may differ from one's own. These stimuli were presented to ChatGPT-4o (OpenAI, San Francisco, CA, USA), selected for two key reasons: (1) its consistent top performance in MLLM evaluation benchmarks, and (2) its prevalent use in ToM research(Kosinski, 2024; Strachan et al., 2024a; Ünlütabak and Bal, 2025), enabling meaningful cross-study comparisons. Both ChatGPT-4o and three human evaluators then independently completed the visual question answering (VQA) task. To control fMRI experiment duration, we selected 15 image-question pairs that the MLLM failed to answer correctly but all humans answered correctly (MLLMI condition) and 15 pairs that both MLLM and humans answered correctly (MLLMC condition) (Figure S5). A successful manipulation check was confirmed by participants' high performance (96.1% accuracy, SD = 3.46%) across all 30 trials (image-question pairs). To test whether individual MLLMI trials drove the neuroimaging results, we performed a leave-one-trial-out validation, iteratively excluding one MLLMI trial (i.e., one image–question pair) at a time and comparing the remaining 14 MLLMI trials against all 15 MLLMC trials. The results indicate that the activation effects exhibit good cross-trial generalizability and are not driven by a small subset of MLLMI trials (Figure S6 and Table S5). We conducted a further control test of the experimental materials using multidimensional questionnaire ratings from 100 participants. Please see the Supplementary Material for details (Tables S6 and S7).

5.3. fMRI experiment design

The fMRI experiment utilized a within-subjects design incorporating a think-aloud paradigm(Li et al., 2023). During scanning, participants were first presented with an image for 6 s and instructed to passively observe it without responding. Subsequently, the same image reappeared accompanied by a question for 14 s, prompting participants to verbally articulate their response while remaining in the scanner. Verbal responses were recorded using a magnet-compatible microphone mounted on the participant's collar, with image presentation order randomized across trials. Post-scanning, participants completed two recall phases: first, they immediately reproduced their original responses to each question; then, for each stimulus, they answered three follow-up items assessing (1) prior exposure to the image's source material, (2) confidence in their fMRI-task response (5-point Likert scale), and (3) their belief about current MLLM's ability to answer correctly (5-point Likert scale) (Fig. 6). It should be noted that, in the present study, human participants were exposed to each stimulus only once, without any prior prompts or repeated exposure. To establish a fair and symmetric comparison framework between humans and artificial intelligence, we similarly restricted the MLLM to a single-pass inference paradigm (Jones et al., 2023; Kosinski, 2024). This approach also avoids unnecessary confounding factors—such as learning effects, order effects, and context anchoring effects—that could otherwise arise from repeated questioning (Huang et al., 2026; Lou and Sun, 2026; Shaier et al., 2024; Shaki et al., 2023).

5.4. Natural language processing

To investigate linguistic differences between human and ChatGPT-4o responses in visual ToM tasks, we performed NLP analyses comparing their linguistic properties(Cambria and White, 2014). All textual data underwent standardized preprocessing: we removed stop words using a predefined Chinese stop word list and excluded words shorter than two characters. Tokenization and part-of-speech tagging were implemented using the Jieba Chinese text segmentation library (https://github.com/fxsjy/jieba). We then extracted key linguistic features to characterize answer semantics, including entropy(Berger et al., 1996), lexical richness(North et al., 2023), semantic familiarity(Liu et al., 2018), semantic certainty(Farkas et al., 2010; Pastore and Dellantonio, 2016; Prabhakaran, 2010; Sanchez and Vogel, 2015; Wang et al., 2024), and semantic similarity(Ahmed et al., 2022). These metrics allowed systematic comparison of response patterns between human and machine-generated answers. In terms of the human answers, the answers of all participants for all 30 questions were calculated, and the linguistic features of 73 participants were averaged for each question so as to compare with MLLM-generated answers.

Entropy is defined as a measure of average information content of a probability distribution, which could be calculated as:

$$H(T) = -\sum_{i=1}^{v} p(w_i) log_2(p(w_i))$$

where $H(T)$ is regarded as the entropy of the distribution of word types, representing the expected information content per word. A crucial step towards estimating $H(T)$ is to reliably approximate the probabilities $p(w_i)$.

Lexical richness is defined as a measure of vocabulary variability, calculated as the proportion of unique tokens relative to the total number of tokens:

$$\text{Lexical richness} = \frac{|\text{unique words}|}{|\text{total words}|}$$

Semantic familiarity is defined as the proportion of meaning-bearing words (nouns, verbs, and adjectives) relative to the total number of words, which is a measure of the linguistic simplicity of responses:

$$\text{Semantic familiarity} = \frac{|\text{meaning} - \text{bearing words}|}{|\text{total words}|}$$

Semantic certainty is defined as the degree of confidence expressed in verbal communication regarding its stated content, integrating multiple factors such as response character count, vague words, affirmative words, punctuation, syntactic structure, etc. (Farkas et al., 2010; Pastore and Dellantonio, 2016; Prabhakaran, 2010; Sanchez and Vogel, 2015; Wang et al., 2024) to derive a composite score:

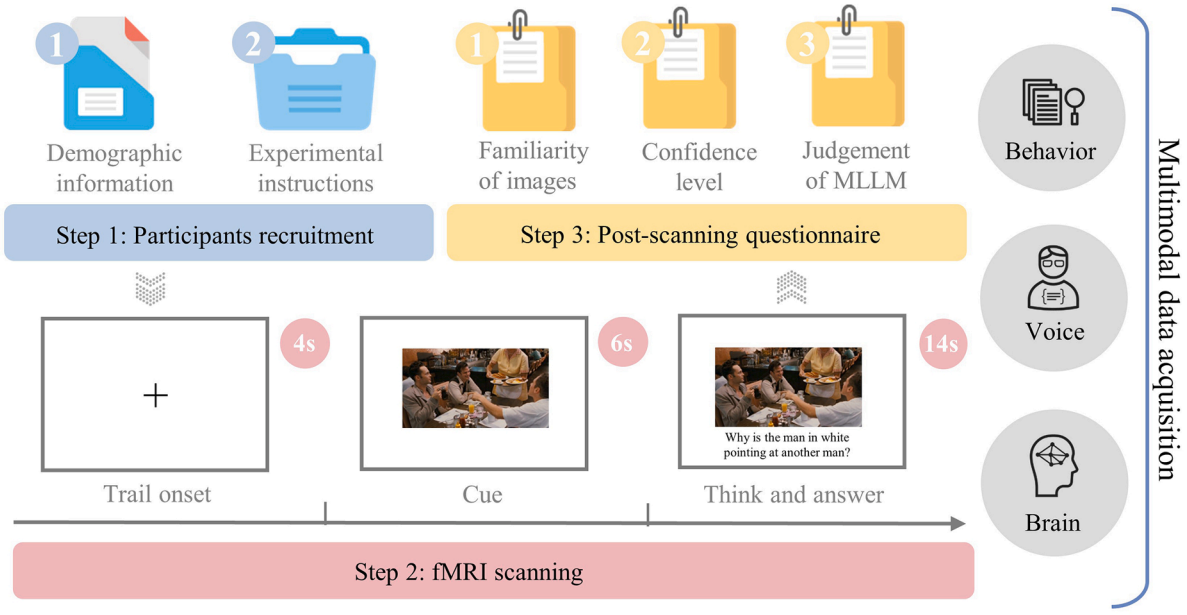

Fig. 6. The flowchart of the fMRI experiment. First, participants were recruited, and their demographic information was collected along with an explanation of the experiment. Next, participants underwent the fMRI scanning. Following the scan, participants completed a questionnaire assessing their familiarity with the experimental materials, confidence in their responses, and judgment regarding whether MLLMs would be able to answer the questions correctly.

$$\textbf{Semantic certainty} = \mathbf{L}(\mathbf{a}) \times \mathbf{P_v}(\mathbf{a}) + \mathbf{B_a}(\mathbf{a}) + \mathbf{P_p}(\mathbf{a}) + \mathbf{B_s}(\mathbf{a}) + \mathbf{P_c}(\mathbf{a}) + \mathbf{P_{conn}}(\mathbf{a})$$

$L(a)$ represents Length score, $P_v(a)$ represents Vague word penalty, $B_a(a)$ represents Affirmative word bonus, $P_p(a)$ represents Punctuation adjustment, $B_s(a)$ represents Short-sentence bonus, $P_c(a)$ represents Complex-structure penalty, $P_{conn}(a)$ represents Connective penalty (see the Supplementary Materials for the calculation details).

Semantic similarity is defined as the degree of similarity between the answer to a specific question and the question itself in semantic space:

$$\text{Semantic Similarity} = \frac{1}{|Q_s|}\sum_{q\in Q_s}\frac{v_q \times v_{as,q}}{\| v_q\| \|v_{as,q}\|}$$

Q_s is the set of questions for which participant s provided a valid answer; v_q and $v_{as,q}$ are the TF-IDF (Term Frequency-Inverse Document Frequency) vectors of question q and its corresponding answer $a_{s,q}$, respectively (see the Supplementary Materials for the calculation details).

5.5. MRI data acquisition and analysis

All MRI data were obtained on a Siemens Magnetom Prisma 3T scanner (Siemens Healthineers, Erlangen, Germany). Comfortable straps and foam pads were used to minimize head motion. Participants were instructed to refrain from moving their heads while answering the questions. A high resolution T1 anatomical image was scanned for accurate localization with the following parameters: sagittal slices; slice number = 192; matrix size = 256 × 256; field of view (FOV) = 256 × 256 mm^2; repetition time/echo time (TR/TE) =2530/2.98 ms; fractional isotropy (FA) = 7°; thickness/gap = 1/0 mm (isotropic voxel size = 1 mm × 1 mm × 1 mm). fMRI images using an echo-planar imaging (EPI) sequence were acquired with the following parameters: TE = 30 ms; FA = 90°; TR = 2000 ms; 32 slices with interleaved acquisition; thickness/gap = 4/0 mm; FOV = 224 × 224 mm^2; matrix = 64 × 64; and voxel size = 3.5 mm × 3.5 mm × 4 mm.

The fMRI data processing was conducted by using Statistical Parametric Mapping 12 (SPM12, http://www.fil.ion.ucl.ac.uk/spm/software/spm12/). The preprocessing steps were as follows: (1) Slice timing correction, (2) Head motion correction, (3) Non-linear registration of the high-resolution T1 structural images to the Montreal Neurological Institute (MNI) template and the segmentation into white matter, gray matter, and cerebrospinal fluid (CSF) with the New segment algorithm, (4) normalization of functional images to MNI space by using the transformation matrix obtained from T1 segmentation and normalization followed by resampling at a spatial resolution of 3 × 3 × 3 mm^3, (5) Spatial smoothing with a Gaussian kernel (full width at half-maximum = 6 mm).

Brain activation difference analysis was conducted to examine the neural representations of ToM differences between MLLMI and MLLMC conditions. The participant-level activation analysis was conducted with SPM12 to generate participant-level activation maps. To precisely isolate the neural correlates of visual ToM reasoning from those of visual perception and motor-speech execution, we constructed a subject-specific first-level general linear model (GLM) using distinct event regressors. Using the synchronized in-scanner microphone recordings, we extracted the precise vocal onset and offset times for each trial. Each trial was then modeled with three sequential epoch regressors: (1) a "Cue" phase (6 s of passive viewing), (2) a "Think" phase (from question onset to the precise vocal onset), and (3) an "Answer" phase (from vocal onset to speech offset). Each epoch regressor was convolved with the canonical hemodynamic response function (HRF). Individual activation maps for the Think phase onset, time-locked to the presentation of the image–question pairs, were generated from the GLM for the two conditions: MLLMI and MLLMC. Since the "Answer" epoch was modeled using the actual duration of the participant's speech, the BOLD variance associated with varying speech lengths was inherently absorbed by this regressor. Additionally, head motion parameters (three translations, three rotations) and a binary regressor indicating whether the participant had previously seen the source material of the picture were defined as covariate regressors. Subsequently, the participant-level activation maps of each condition were obtained. A paired t-test was performed to compare between the two conditions with a voxel-level $p < 0.001$ and cluster-level $p < 0.05$ (Gaussian random field correction).

To control for the potential confounding effects of task difficulty and complexity, subjective ratings of task difficulty and complexity obtained from the questionnaire were included as covariates in separate first-level GLMs. Subsequently, a contrast between conditions was performed in

the second-level analysis, with a voxel-level $p < 0.005$ and cluster-level $p < 0.05$ (Gaussian random field correction).

To measure the effects of trial-by-trial variation in participants' confidence and judgment of MLLM, the rated confidence and judgement for each trial was defined as a parametric modulation. These two indicators were incorporated as modulators in separate GLM models to generate participant-level rating effect activation maps. A one-sample *t*-test was performed to investigate the significant brain regions affected by behavior indicators, with a voxel-level $p < 0.005$ and cluster-level $p < 0.05$ (Gaussian random field correction).

To further explore the neural network representations of ToM differences between MLLMI and MLLMC conditions, functional connectivity difference analysis was conducted by using CONN (functional connectivity toolbox, www.nitrc.org/projects/CONN)(Whitfield-Gabrieli and Nieto-Castanon, 2012). ROI-to-ROI analysis was conducted, which signifies the level of functional connectivity between each pair of ROIs. Each element in a matrix is defined as the Fisher-transformed bivariate correlation coefficient between a pair of ROI BOLD time series:

$$r(i,j) = \frac{\int R_i(t)R_j(t)dt}{\left(\int R_i^2(t)dt \int R_j^2(t)dt\right)^{1/2}}$$

$$Z(i,j) = tanh^{-1}(r(i,j))$$

where R is the BOLD time series within each ROI (for simplicity, all the time series here are considered centered to zero mean), r is the matrix of correlation coefficients, and Z is the ROI-to-ROI connectivity matrix of Fisher-transformed correlation coefficients.

The default network atlas provided within the CONN toolbox was used to define the ROIs, encompassing core nodes across standard large-scale functional networks (e.g., Default Mode, Visual, Salience, and Dorsal Attention networks). Based on the preprocessed data, the dataset was denoised through a band-pass filter [0.008, Inf]. The condition-specific association between each source ROI BOLD time series was measured by weighted GLM for correlation. Then ROI-to-ROI functional connectivity matrices for all network sources for each participant were obtained. A paired *t*-test was performed to make comparison between the contrasting conditions. ROI-to-ROI connectivity map was plotted to show strength with $p < 0.005$ and cluster $p < 0.05$.

5.6. Deep learning models

We constructed a computational prediction model via deep learning to decode distinct brain states (MLLMI vs. MLLMC) and to identify the underlying neural activation states and functional connectivity patterns that contribute to distinguishing between the two conditions. The dataset consisted of 2190 samples (73 valid participants × 30 questions), with each sample characterized by 12 features and a binary label (0 for MLLMI, 1 for MLLMC). The initial three features are single-trial fMRI BOLD signals, while the remaining nine are single-trial functional connectivity values (ROI-to-ROI).

To ensure the comparability of predictive variables, this study adopted a standardized data processing pipeline: Firstly, all feature values were normalized using Z-score standardization (mean = 0, standard deviation = 1). To rigorously prevent data leakage and evaluate true cross-subject generalizability, we adopted a strict subject-independent 10-fold cross-validation strategy (GroupKFold). Under this strategy, all 30 trials from a given participant were assigned exclusively to either the training set or the testing set, ensuring that no participant's data appeared in both sets.

This study conducted a systematic comparative analysis of three mainstream deep learning architectures: CNN, LSTM, and Transformer models. The experimental setup included four network configurations: 2-layer LSTM, 2-layer CNN, 2-layer Transformer, and 1-layer Transformer. All analyses were conducted using Python version 3.12 (Python Software Foundation, Wilmington, DE, USA). The deep learning models were implemented using PyTorch. Model interpretability was achieved through SHAP analysis(Pat et al., 2022), and all visualizations were generated using Matplotlib. SHAP values were computed for the test samples, providing a unified framework to quantify the influence of each feature on the model's output. A global summary plot was generated to rank features based on their average SHAP values, highlighting their relative importance across the entire dataset.

To further strengthen interpretability and identify the causal contribution of specific neural features, we conducted targeted network-level ablation studies. Recognizing the high collinearity of fMRI connectivity, the 12 neural features were grouped into five biologically meaningful categories: Regional Anatomy (PCu, MTG), DMN-related, Visual-related, Language-related, and Salience-related networks. For each network, we ablated its corresponding features and calculated the mean accuracy drop using the 10-fold Transformer model. To control for the bias introduced by the varying number of features in each network, we generated an empirical null distribution by randomly ablating size-matched feature subsets across 100 permutations.

CRediT authorship contribution statement

Jia Jin: Conceptualization, Funding acquisition, Methodology, Writing – original draft. **Yunsong Hu:** Data curation, Formal analysis, Investigation, Methodology, Visualization. **Zhongfeng Wang:** Data curation, Formal analysis, Investigation, Methodology, Visualization. **Yu Pan:** Validation, Writing – review & editing. **Baojun Ma:** Validation, Writing – review & editing. **Bo Dong:** Writing – review & editing. **Jianmin Dong:** Formal analysis, Methodology, Visualization. **Guanxiong Pei:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Natural Science Foundation of China [grant numbers 72401263, 72271166]; the Soft Science Research Program of Zhejiang Province [grant number 2025C25080 (SYS)]; and the Supervisor Academic Mentorship Program of Shanghai International Studies University [grant number 2025DSYL048].

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2026.122108.

Data availability

Data and code supporting this study are available from the corresponding author upon reasonable request.

References

Ahmed, M., Khan, H.U., Iqbal, S., Althebyan, Q., 2022. Automated question answering based on improved TF-IDF and cosine similarity. In: 2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS), pp. 1–6. https://doi.org/10.1109/SNAMS58071.2022.10062839.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., 2022. Flamingo: a visual language model for few-shot learning. Adv. Neural Inf. Process. Syst. 35, 23716–23736. https://doi.org/10.48550/arXiv.2204.14198.

Apperly, I.A., Samson, D., Chiavarino, C., Humphreys, G.W., 2004. Frontal and temporo-parietal lobe contributions to theory of mind: neuropsychological evidence from a

false-belief task with reduced language and executive demands. J. Cogn. Neurosci. 16, 1773–1784. https://doi.org/10.1162/0898929042947928.
Apperly, I.A., Butterfill, S.A., 2009. Do humans have two systems to track beliefs and belief-like states? Psychol. Rev. 116, 953. https://doi.org/10.1037/a0016923.
Balgova, E., Diveica, V., Jackson, R.L., Binney, R.J., 2024. Overlapping neural correlates underpin theory of mind and semantic cognition: evidence from a meta-analysis of 344 functional neuroimaging studies. Neuropsychologia 200, 108904. https://doi.org/10.1016/j.neuropsychologia.2024.108904.
Baron-Cohen, S., Leslie, A.M., Frith, U., 1985. Does the autistic child have a "theory of mind"? Cognition 21, 37–46. https://doi.org/10.1016/0010-0277(85)90022-8.
Baron-Cohen, S., O'riordan, M., Stone, V., Jones, R., Plaisted, K., 1999. Recognition of faux pas by normally developing children and children with Asperger syndrome or high-functioning autism. J. Autism Dev. Disord. 29, 407–418. https://doi.org/10.1023/A:1023035012436.
Berger, A.L., Pietra, V.J.D., Pietra, S.A.D., 1996. A maximum entropy approach to natural language processing. Comput. Linguist. 22, 39–71.
Bernstein, D.M., Thornton, W.L., Sommerville, J.A., 2011. Theory of mind through the ages: older and middle-aged adults exhibit more errors than do younger adults on a continuous false belief task. Exp. Aging Res. 37, 481–502. https://doi.org/10.1080/0361073X.2011.619466.
Boccadoro, S., Cracco, E., Hudson, A.R., Bardi, L., Nijhof, A.D., Wiersema, J.R., Brass, M., Mueller, S.C., 2019. Defining the neural correlates of spontaneous theory of mind (ToM): an fMRI multi-study investigation. NeuroImage 203, 116193. https://doi.org/10.1016/j.neuroimage.2019.116193.
Brissenden, J.A., Tobyne, S.M., Osher, D.E., Levin, E.J., Halko, M.A., Somers, D.C., 2018. Topographic cortico-cerebellar networks revealed by visual attention and working memory. Curr. Biol. 28, 3364–3372. https://doi.org/10.1016/j.cub.2018.08.059. . e3365.
Brod, G., Lindenberger, U., Wagner, A.D., Shing, Y.L., 2016. Knowledge acquisition during exam preparation improves memory and modulates memory formation. J. Neurosci. 36, 8103. https://doi.org/10.1523/JNEUROSCI.0045-16.2016.
Cai, H., Ao, Z., Tian, C., Wu, Z., Liu, H., Tchieu, J., Gu, M., Mackie, K., Guo, F., 2023. Brain organoid reservoir computing for artificial intelligence. Nat. Electron. 6, 1032–1039. https://doi.org/10.1038/s41928-023-01069-w.
Cambria, E., White, B., 2014. Jumping NLP curves: a review of natural language processing research. IEEE Comput. Intell. Mag. 9, 48–57. https://doi.org/10.1109/MCI.2014.2307227.
Carrington, S.J., Bailey, A.J., 2009. Are there theory of mind regions in the brain? A review of the neuroimaging literature. Hum. Brain Mapp. 30, 2313–2335. https://doi.org/10.1002/hbm.20671.
Cavanna, A.E., Trimble, M.R., 2006. The precuneus: a review of its functional anatomy and behavioural correlates. Brain 129, 564–583. https://doi.org/10.1093/brain/awl004.
Chang, S.E., Hughes, D.E., Zhu, J., Hyat, M., Salone, S.D., Goodman, Z.T., Roffman, J.L., Karcher, N.R., Hernandez, L.M., Forsyth, J.K., Bearden, C.E., 2024. Attention-mediated genetic influences on psychotic symptomatology in adolescence. Nat. Ment. Health 2, 1518–1531. https://doi.org/10.1038/s44220-024-00338-7.
Chen, T., Cai, W., Ryali, S., Supekar, K., Menon, V., 2016. Distinct global brain dynamics and spatiotemporal organization of the salience network. PLoS Biol. 14, e1002469. https://doi.org/10.1371/journal.pbio.1002469.
Ciaramidaro, A., Adenzato, M., Enrici, I., Erk, S., Pia, L., Bara, B.G., Walter, H., 2007. The intentional network: how the brain reads varieties of intentions. Neuropsychologia 45, 3105–3113. https://doi.org/10.1016/j.neuropsychologia.2007.05.011.
Collins, K.M., Sucholutsky, I., Bhatt, U., Chandra, K., Wong, L., Lee, M., Zhang, C.E., Zhi-Xuan, T., Ho, M., Mansinghka, V., Weller, A., Tenenbaum, J.B., Griffiths, T.L., 2024. Building machines that learn and think with people. Nat. Hum. Behav. 8, 1851–1863. https://doi.org/10.1038/s41562-024-01991-9.
Conway, J.R., Bird, G., 2018. Conceptualizing degrees of theory of mind. Proc. Natl. Acad. Sci. 115, 1408–1410. https://doi.org/10.1073/pnas.1722301115.
Corcoran, R., 2003. Inductive reasoning and the understanding of intention in schizophrenia. Cogn. neuropsychiatry 8, 223–235. https://doi.org/10.1080/13546800244000319.
Dadario, N.B., Sughrue, M.E., 2023. The functional role of the precuneus. Brain 146, 3598–3607. https://doi.org/10.1093/brain/awad181.
Davey, J., Thompson, H.E., Hallam, G., Karapanagiotidis, T., Murphy, C., De Caso, I., Krieger-Redwood, K., Bernhardt, B.C., Smallwood, J., Jefferies, E., 2016. Exploring the role of the posterior middle temporal gyrus in semantic cognition: integration of anterior temporal lobe with executive processes. NeuroImage 137, 165–177. https://doi.org/10.1016/j.neuroimage.2016.05.051.
Doricchi, F., Lasaponara, S., Pazzaglia, M., Silvetti, M., 2022. Left and right temporal-parietal junctions (TPJs) as "match/mismatch" hedonic machines: a unifying account of TPJ function. Phys. Life Rev. 42, 56–92. https://doi.org/10.1016/j.plrev.2022.07.001.
Farkas, R., Vincze, V., Móra, G., Csirik, J., Szarvas, G., 2010. The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text. In: Proceedings of the Fourteenth Conference On Computational Natural Language Learning–Shared Task, pp. 1–12.
Fernandez-Iriondo, I., Jimenez-Marin, A., Diez, I., Bonifazi, P., Swinnen, S.P., Muñoz, M. A., Cortes, J.M., 2021. Small variation in dynamic functional connectivity in cerebellar networks. Neurocomputing 461, 751–761. https://doi.org/10.1016/j.neucom.2020.09.092.
Frith, C.D., Frith, U., 2006. The neural basis of mentalizing. Neuron 50, 531–534. https://doi.org/10.1016/j.neuron.2006.05.001.
Geng, J.J., Vossel, S., 2013. Re-evaluating the role of TPJ in attentional control: contextual updating? Neurosci. Biobehav. Rev. 37, 2608–2620. https://doi.org/10.1016/j.neubiorev.2013.08.010.
Glickman, M., Sharot, T., 2024. How human–AI feedback loops alter human perceptual, emotional and social judgements. Nat. Hum. Behav. 9, 345–359. https://doi.org/10.1038/s41562-024-02077-2.
Grosse Wiesmann, C., Schreiber, J., Singer, T., Steinbeis, N., Friederici, A.D., 2017. White matter maturation is associated with the emergence of theory of mind in early childhood. Nat. Commun. 8, 14692. https://doi.org/10.1038/ncomms14692.
Guo, J., Li, J., Li, D., Tiong, A.M.H., Li, B., Tao, D., Hoi, S., 2023. From images to textual prompts: zero-shot visual question answering with frozen large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10867–10877. https://doi.org/10.1109/CVPR52729.2023.01046.
Hagendorff, T., Fabi, S., Kosinski, M., 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. Nat. Comput. Sci. 3, 833–838. https://doi.org/10.1038/s43588-023-00527-x.
Happé, F.G., 1994. An advanced test of theory of mind: understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. J. Autism Dev. Disord. 24, 129–154. https://doi.org/10.1007/BF02172093.
Hauptman, M., Blank, I., Fedorenko, E., 2023. Non-literal language processing is jointly supported by the language and theory of mind networks: evidence from a novel meta-analytic fMRI approach. Cortex 162, 96–114. https://doi.org/10.1016/j.cortex.2023.01.013.
Huang, J.Y., Choshen, L., Astudillo, R., Broderick, T., Andreas, J., 2026. Do LLMs Benefit From Their Own Words? arXiv preprint. https://doi.org/10.48550/arXiv.2602.24287 arXiv:2602.24287.
Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O. K., Patra, B., 2023. Language is not all you need: aligning perception with language models. Adv. Neural Inf. Process. Syst. 36, 72096–72109. https://doi.org/10.48550/arXiv.2302.14045.
Hughes, C., Cassidy, B.S., Faskowitz, J., Avena-Koenigsberger, A., Sporns, O., Krendl, A. C., 2019. Age differences in specific neural connections within the Default Mode Network underlie theory of mind. NeuroImage 191, 269–277. https://doi.org/10.1016/j.neuroimage.2019.02.024.
Jones, C.R., Trott, S., Bergen, B., 2023. Epitome: experimental protocol inventory for theory of mind evaluation. In: Proceedings of the First Workshop on Theory of Mind in Communicating Agents, pp. 1–14.
Jung-Beeman, M., 2005. Bilateral brain processes for comprehending natural language. Trends Cogn. Sci. 9, 512–518. https://doi.org/10.1016/j.tics.2005.09.009.
Kosinski, M., 2024. Evaluating large language models in theory of mind tasks. Proc. Natl. Acad. Sci. 121, e2405460121. https://doi.org/10.1073/pnas.2405460121.
Kuang, J., Shen, Y., Xie, J., Luo, H., Xu, Z., Li, R., Li, Y., Cheng, X., Lin, X., Han, Y., 2025. Natural language understanding and inference with MLLM in visual question answering: a survey. ACM Comput. Surv. 57, 190. https://doi.org/10.1145/3711680.
Kucyi, A., Daitch, A., Raccah, O., Zhao, B., Zhang, C., Esterman, M., Zeineh, M., Halpern, C.H., Zhang, K., Zhang, J., 2020. Electrophysiological dynamics of antagonistic brain networks reflect attentional fluctuations. Nat. Commun. 11, 325. https://doi.org/10.1038/s41467-019-14166-2.
Kulakova, E., Aichhorn, M., Schurz, M., Kronbichler, M., Perner, J., 2013. Processing counterfactual and hypothetical conditionals: an fMRI investigation. NeuroImage 72, 265–271. https://doi.org/10.1016/j.neuroimage.2013.01.060.
Kunda, M., 2018. Visual mental imagery: a view from artificial intelligence. Cortex 105, 155–172. https://doi.org/10.1016/j.cortex.2018.01.022.
Leshinskaya, A., Thompson-Schill, S.L., 2020. Transformation of event representations along middle temporal gyrus. Cereb. Cortex 30, 3148–3166. https://doi.org/10.1093/cercor/bhz300.
Leslie, A.M., Friedman, O., German, T.P., 2004. Core mechanisms in 'theory of mind. Trends Cogn. Sci. 8, 528–533. https://doi.org/10.1016/j.tics.2004.10.001.
Li, H.-X., Lu, B., Wang, Y.-W., Li, X.-Y., Chen, X., Yan, C.-G., 2023. Neural representations of self-generated thought during think-aloud fMRI. NeuroImage 265, 119775. https://doi.org/10.1016/j.neuroimage.2022.119775.
Li, Y., Tian, W., Jiao, Y., Chen, J., Qian, T., Zhu, B., Zhao, N., Jiang, Y.-G., 2024a. Look Before You Decide: Prompting Active Deduction of MLLMs For Assumptive Reasoning. arXiv preprint. https://doi.org/10.48550/arXiv.2404.12966 arXiv:2404.12966.
Li, Z., Liu, D., Zhang, C., Wang, H., Xue, T., Cai, W., 2024b. Enhancing Advanced Visual Reasoning Ability of Large Language Models. arXiv preprint. https://doi.org/10.48550/arXiv.2409.13980 arXiv:2409.13980.
Li, Z., Petersen, I.T., Wang, L., Radua, J., Yang, G., Liu, X., 2025. The lifespan trajectories of brain activities related to conflict-driven cognitive control. Sci. Bull. https://doi.org/10.1016/j.scib.2025.08.038.
Liang, P.P., Zadeh, A., Morency, L.-P., 2024. Foundations & trends in multimodal machine learning: principles, challenges, and open questions. ACM Comput. Surv. 56, 1–42. https://doi.org/10.1145/3656580.
Liu, S., Bremer, P.T., Thiagarajan, J.J., Srikumar, V., Wang, B., Livnat, Y., Pascucci, V., 2018. Visual exploration of semantic relationships in neural word embeddings. IEEE Trans. Vis. Comput. Graph. 24, 553–562. https://doi.org/10.1109/TVCG.2017.2745141.
Lou, J., Sun, Y., 2026. Anchoring bias in large language models: an experimental study. J. Comput. Soc. Sci. 9, 11. https://doi.org/10.1007/s42001-025-00435-2.
Małkiński, M., Mańdziuk, J., 2023. A review of emerging research directions in abstract visual reasoning. Inf. Fusion 91, 713–736. https://doi.org/10.1016/j.inffus.2022.11.011.
Mao, Y., Liu, S., Ni, Q., Lin, X., He, L., 2024. A review on machine theory of mind. IEEE Trans. Comput. Soc. Syst. 11, 7114–7132. https://doi.org/10.1109/TCSS.2024.3416707.

Mehonic, A., Kenyon, A.J., 2022. Brain-inspired computing needs a master plan. Nature 604, 255–260. https://doi.org/10.1038/s41586-021-04362-w.
Molenberghs, P., Johnson, H., Henry, J.D., Mattingley, J.B., 2016. Understanding the minds of others: a neuroimaging meta-analysis. Neurosci. Biobehav. Rev. 65, 276–291. https://doi.org/10.1016/j.neubiorev.2016.03.020.
Moran, J.M., 2013. Lifespan development: the effects of typical aging on theory of mind. Behav. Brain Res. 237, 32–40. https://doi.org/10.1016/j.bbr.2012.09.020.
Ünlütabak, B., Bal, O., 2025. Theory of mind performance of large language models: a comparative analysis of Turkish and English. Comput. Speech Lang. 89, 101698. https://doi.org/10.1016/j.csl.2024.101698.
North, K., Zampieri, M., Shardlow, M., 2023. Lexical complexity prediction: an overview. ACM Comput. Surv. 55, 179. https://doi.org/10.1145/3557885.
Nummenmaa, L., Glerean, E., Viinikainen, M., Jääskeläinen, I.P., Hari, R., Sams, M., 2012. Emotions promote social interaction by synchronizing brain activity across individuals. Proc. Natl. Acad. Sci. 109, 9599–9604. https://doi.org/10.1073/pnas.1206095109.
Pan, X., Ye, T., Han, D., Song, S., Huang, G., 2022. Contrastive language-image pre-training with knowledge graphs. Adv. Neural Inf. Process. Syst. 35, 22895–22910. https://doi.org/10.48550/arXiv.2210.08901.
Pastore, L., Dellantonio, S., 2016. Modelling scientific un/certainty. Why argumentation strategies trump linguistic markers use. Model-Based Reasoning in Science and Technology: Logical, Epistemological, and Cognitive Issues. Springer, pp. 137–164.
Pat, N., Wang, Y., Bartonicek, A., Candia, J., Stringaris, A., 2022. Explainable machine learning approach to predict and explain the relationship between task-based fMRI and individual differences in cognition. Cereb. Cortex 33, 2682–2703. https://doi.org/10.1093/cercor/bhac235.
Poldrack, R.A., 2006. Can cognitive processes be inferred from neuroimaging data? Trends Cogn. Sci. 10, 59–63. https://doi.org/10.1016/j.tics.2005.12.004.
Poldrack, R.A., 2011. Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. Neuron 72, 692–697. https://doi.org/10.1016/j.neuron.2011.11.001.
Prabhakaran, V., 2010. Uncertainty learning using SVMs and CRFs. In: Proceedings of the Fourteenth Conference on Computational Natural Language Learning–Shared Task, pp. 132–137.
Premack, D., Woodruff, G., 1978. Does the chimpanzee have a theory of mind? Behav. Brain Sci. 1, 515–526. https://doi.org/10.1017/S0140525X00076512.
Ren, J., Huang, F., Zhou, Y., Zhuang, L., Xu, J., Gao, C., Qin, S., Luo, J., 2020. The function of the hippocampus and middle temporal gyrus in forming new associations and concepts during the processing of novelty and usefulness features in creative designs. NeuroImage 214, 116751. https://doi.org/10.1016/j.neuroimage.2020.116751.
Sanchez, L.M., Vogel, C., 2015. A hedging annotation scheme focused on epistemic phrases for informal language. In: Proceedings of the Workshop on Models for Modality Annotation. UK. Association for Computational Linguistics, London.
Saxe, R., Kanwisher, N., 2013. People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind. Social Neuroscience. Psychology Press, pp. 171–182.
Schlaffke, L., Lissek, S., Lenz, M., Juckel, G., Schultz, T., Tegenthoff, M., Schmidt-Wilcke, T., Brüne, M., 2015. Shared and nonshared neural networks of cognitive and affective theory-of-mind: a neuroimaging study using cartoon picture stories. Hum. Brain Mapp. 36, 29–39. https://doi.org/10.1002/hbm.22610.
Schulze Buschoff, L.M., Akata, E., Bethge, M., Schulz, E., 2025a. Visual cognition in multimodal large language models. Nat. Mach. Intell. 1–11. https://doi.org/10.1038/s42256-024-00963-y.
Schurz, M., Radua, J., Aichhorn, M., Richlan, F., Perner, J., 2014. Fractionating theory of mind: a meta-analysis of functional brain imaging studies. Neurosci. Biobehav. Rev. 42, 9–34. https://doi.org/10.1016/j.neubiorev.2014.01.009.
Schurz, M., Radua, J., Tholen, M.G., Maliske, L., Margulies, D.S., Mars, R.B., Sallet, J., Kanske, P., 2021. Toward a hierarchical model of social cognition: a neuroimaging meta-analysis and integrative review of empathy and theory of mind. Psychol. Bull. 147, 293. https://doi.org/10.1037/bul0000303.
Schurz, M., Tholen, M.G., Perner, J., Mars, R.B., Sallet, J., 2017. Specifying the brain anatomy underlying temporo-parietal junction activations for theory of mind: a review using probabilistic atlases from different imaging modalities. Hum. Brain Mapp. 38, 4788–4805. https://doi.org/10.1002/hbm.23675.
Shaier, S., Sanz-Guerrero, M., von der Wense, K., 2024. Asking Again and again: Exploring LLM Robustness to Repeated Questions. arXiv preprint. https://doi.org/10.48550/arXiv.2412.07923 arXiv:2412.07923.
Shaki, J., Kraus, S., Wooldridge, M., 2023. Cognitive Effects in Large Language Models. arXiv preprint. https://doi.org/10.48550/arXiv.2308.14337 arXiv:2308.14337.
Strachan, J.W., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., 2024a. Testing theory of mind in large language models and humans. Nat. Hum. Behav. 8, 1285–1295. https://doi.org/10.1038/s41562-024-01882-z.
Tanglay, O., Young, I.M., Dadario, N.B., Briggs, R.G., Fonseka, R.D., Dhanaraj, V., Hormovas, J., Lin, Y.-H., Sughrue, M.E., 2022. Anatomy and white-matter connections of the precuneus. Brain Imaging Behav. 16, 574–586. https://doi.org/10.1007/s11682-021-00529-1.
Vine, V., Boyd, R.L., Pennebaker, J.W., 2020. Natural emotion vocabularies as windows on distress and well-being. Nat. Commun. 11, 4525. https://doi.org/10.1038/s41467-020-18349-0.
Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happé, F., Falkai, P., Maier, W., Shah, N.J., Fink, G.R., Zilles, K., 2001. Mind reading: neural mechanisms of theory of mind and self-perspective. NeuroImage 14, 170–181. https://doi.org/10.1006/nimg.2001.0789.
Wagner, M.J., Luo, L., 2020. Neocortex-cerebellum circuits for cognitive processing. Trends Neurosci. 43, 42–54. https://doi.org/10.1016/j.tins.2019.11.002.
Wang, P., Lam, B.D., Liu, Y., Asgari-Targhi, A., Panda, R., Wells, W.M., Kapur, T., Golland, P., 2024. Calibrating expressions of certainty. arXiv preprint arXiv: 2410.04315. https://doi.org/10.48550/arXiv.2410.04315.
Wardell, V., Palombo, D.J., 2024. Stability and malleability of emotional autobiographical memories. Nat. Rev. Psychol. 1–14. https://doi.org/10.1038/s44159-024-00312-1.
White, S., Hill, E., Happé, F., Frith, U., 2009. Revisiting the strange stories: revealing mentalizing impairments in autism. Child Dev. 80, 1097–1117. https://doi.org/10.1111/j.1467-8624.2009.01319.x.
Whitfield-Gabrieli, S., Nieto-Castanon, A., 2012. Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. Brain Connect. 2, 125–141. https://doi.org/10.1089/brain.2012.0073.
Williams, J.R., 2008. The declaration of helsinki and public health. Bull. World Health Organ. 86, 650–652. https://doi.org/10.2471/BLT.08.050955.
Wimmer, H., Perner, J., 1983. Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition 13, 103–128. https://doi.org/10.1016/0010-0277(83)90004-5.
Xu, J., Lyu, H., Li, T., Xu, Z., Fu, X., Jia, F., Wang, J., Hu, Q., 2019. Delineating functional segregations of the human middle temporal gyrus with resting-state functional connectivity and coactivation patterns. Hum. Brain Mapp. 40, 5159–5171. https://doi.org/10.1002/hbm.24763.
Xu, B., Poo, M.-m., 2023. Large language models and brain-inspired general intelligence. Natl. Sci. Rev. 10, nwad267. https://doi.org/10.1093/nsr/nwad267.
Yablonski, M., Karipidis, I.I., Kubota, E., Yeatman, J.D., 2024. The transition from vision to language: distinct patterns of functional connectivity for subregions of the visual word form area. Hum. Brain Mapp. 45, e26655. https://doi.org/10.1002/hbm.26655.
Yang, Z., Gan, Z., Wang, J., Hu, X., Lu, Y., Liu, Z., Wang, L., 2022. An empirical study of gpt-3 for few-shot knowledge-based VQA. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 3081–3089. https://doi.org/10.1609/aaai.v36i3.20215.
Yin, D., Li, L.H., Hu, Z., Peng, N., Chang, K.-W., 2021. Broaden the vision: geo-diverse visual commonsense reasoning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 2115–2129.
Young, L., Saxe, R., 2009. An FMRI investigation of spontaneous mental state inference for moral judgment. J. Cogn. Neurosci. 21, 1396–1405. https://doi.org/10.1162/jocn.2009.21142.
Zellers, R., Bisk, Y., Farhadi, A., Choi, Y., 2019. From recognition to cognition: visual commonsense reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6720–6731. https://doi.org/10.48550/arXiv.1811.10830.
Zhang, S., Dai, G., Huang, T., Chen, J., 2024a. Multimodal large language models for bioimage analysis. Nat. Methods 21, 1390–1393. https://doi.org/10.1038/s41592-024-02334-2.
Zhang, W., Ma, S., Ji, X., Liu, X., Cong, Y., Shi, L., 2024b. The development of general-purpose brain-inspired computing. Nat. Electron. 7, 954–965. https://doi.org/10.1038/s41928-024-01277-y.
Zhang, Y., Qu, P., Ji, Y., Zhang, W., Gao, G., Wang, G., Song, S., Li, G., Chen, W., Zheng, W., 2020. A system hierarchy for brain-inspired computing. Nature 586, 378–384. https://doi.org/10.1038/s41586-020-2782-y.
Zhou, T., Niu, Y., Lu, H., Peng, C., Guo, Y., Zhou, H., 2024. Vision transformer: to discover the "four secrets" of image patches. Inf. Fusion 105, 102248. https://doi.org/10.1016/j.inffus.2024.102248.