Embracing Trustworthy Brain-Agent Collaboration as Paradigm Extension for Intelligent Assistive Technologies

Yankai Chen[1,2,3], **Xinni Zhang**[4], **Yifei Zhang**[5], **Yangning Li**[6], **Henry Peng Zou**[1], **Chunyu Miao**[1], **Weizhi Zhang**[1], **Xue Liu**[2,3], **Philip S. Yu**[1]
[1]University of Illinois Chicago, [2]MBZUAI, [3]McGill University, [4]The Chinese University of Hong Kong, [5]Nanyang Technological University, [6]Tsinghua University
{ychen588, pzou3, cmiao8, yli23, wzhan42, psyu}@uic.edu,
xnzhang23@cse.cuhk.edu.hk, yifei.zhang@ntu.edu.sg, steve.liu@mbzuai.ac.ae

Abstract

Brain-Computer Interfaces (BCIs) offer a direct communication pathway between the human brain and external devices, holding significant promise for individuals with severe neurological impairments. However, their widespread adoption is hindered by critical limitations, such as low information transfer rates and extensive user-specific calibration. To overcome these challenges, recent research has explored the integration of Large Language Models (LLMs), extending the focus from simple command decoding to understanding complex cognitive states. Despite these advancements, deploying agentic AI faces technical hurdles and ethical concerns. Due to the lack of comprehensive discussion on this emerging direction, this position paper argues that the field is poised for a paradigm extension from BCI to **Brain-Agent Collaboration (BAC)**. We emphasize reframing agents as active and collaborative partners for intelligent assistance rather than passive brain signal data processors, demanding a focus on ethical data handling, model reliability, and a robust human-agent collaboration framework to ensure these systems are safe, trustworthy, and effective.

1 Introduction

Brain-Computer Interfaces (BCIs) provide a direct communication pathway between the human brain and external devices by measuring and translating signals from the central nervous system, thereby bypassing conventional muscular routes [55]. BCIs are not instruments of "mind-reading" but rather tools empowering users to execute direct actions solely through their brain activity, without physical movement. The utility of BCIs can be profound, particularly in restoring communication for individuals with severe neurological impairments like locked-in syndrome, enabling the control of prosthetic limbs, and facilitating neuro-rehabilitation for conditions such as stroke, paralysis, epilepsy, attention disorders, Parkinson's Disease, and sleep disturbances [77, 55].

The widespread adoption of BCI technology is however hampered by several limitations. Beyond ethical considerations, including data privacy, the security of sensitive neural information against potential neuro-hacking, BCI contains several technical and practical issues. A primary concern remains the limited performance, particularly characterized by low information transfer rates, which translate into slow operational speeds and often unsatisfactory accuracy [25]. Due to the inherent inter-individual variability in brain activity, most BCI systems require extensive user-specific training and calibration to accurately interpret neural signals [101]. These paradigms demand that users learn voluntary brain activity modulation [20, 48], creating significant barriers for disabled users. This

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

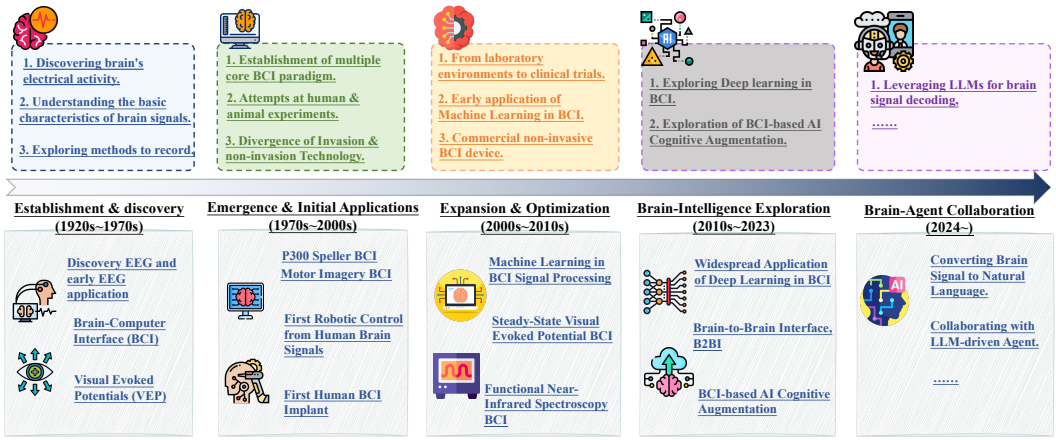

Figure 1: The Evolution of Brain Activity Analysis Paradigm.

combined training overhead and operational inconvenience severely limit real-world deployment beyond laboratory settings. Furthermore, signal quality is another major hurdle, with non-invasive modalities like electroencephalography (EEG) suffering from low spatial resolution and high susceptibility to artifacts from muscle movements or environmental noise [101]. As for invasive BCI approaches, while offering higher fidelity, they introduce significant concerns, including surgical risks like infection and tissue damage [84]. These interconnected limitations, where poor signal quality might necessitate longer training, thereby affecting speed and usability, create a cycle that impedes the development of truly fluid and intuitive BCIs [25].

To address these issues, recent research attempts the integration of advanced artificial intelligence, particularly Large Language Models (LLMs) and Vision Language Models (VLMs), into BCI research [7, 64, 22]. These methods revolutionize how neural data is interpreted to move beyond simple command decoding towards understanding the cognitive states. For instance, LLMs are being employed to mitigate the data variability in brain signal processing. Liu et al. [69] utilize techniques like signal autoencoders and prompt tuning for better generalization and zero-shot predictions. Moreover, it is promising to further decode brain signals into textual languages [75]. For example, Thought2Text [75] decodes brain activities into comprehensible texts with fine-tuned LLMs and Electroencephalographic data. BrainLLM [122] demonstrates the ability of LLMs to generate natural language directly from functional Magnetic Resonance Imaging (fMRI) recordings by integrating brain signals into the language generation. Beyond LLM-driven brain signal processing and language decoding, some other works explore the **agentic capabilities of LLMs**. LLM agents are systems that extend the powerful comprehension and generation capabilities of LLMs into the domain of autonomous action [105, 76]. Their core characteristics include autonomous planning [121, 98], tool utilization [95, 10], and perception and interaction with the external environment [127, 130] in single or multi-agent systems [37, 83, 114]. They are widely applied in fields such as software engineering [40, 39, 74], knowledge-intensive question answering and deep research [31, 65, 105], autonomous robotics [103, 15], and healthcare [106], etc. For brain activity analysis, LLMs function as autonomous agents that can perceive neural contexts, reason about user intentions, provide appropriate responses, and execute adaptive actions based on decoded brain states [51, 6, 54]. For example, Baradari et al. [6] leverage an LLM-based agent as the neuroadaptive tutor to track real-time brain signal engagement. The system continuously monitors and infers the user's level of cognitive engagement. Based on this inferred engagement level, it dynamically responds and adjusts the complexity of the educational content to the user.

Despite the promising advancements, the deployment of integrating the agentic AI with neural interfaces in BCI remains sophisticated and fraught with both technical and ethical challenges. On one hand, AI agents are designed to extend human intelligence, fundamentally redefining workflows in complex decision-making and task execution. On the other hand, with the strong decoding capability of LLMs, it is also crucial to ensure the agentic integration is safe and trustworthy [55]. Furthermore, LLMs exhibit hallucination issues that can generate plausible yet factually incorrect or nonsensical outputs, which undermines trust and potentially causes significant errors, particularly

when actions are chained together [36, 116, 35]. The lack of robust frameworks for development, evaluation, and deployment further complicates efforts to ensure the tool effectiveness and safety.

Our Position. This paper argues that the integration of LLM-based agents in neurocognitive care has reached a critical juncture where their implementation and evaluation are both feasible and essential. We advocate that **the field is poised for a paradigm extension from Brain-Computer Interface (BCI) to Brain-Agent Collaboration (BAC)**, where agents serve as active, supportive, collaborative, ethical, and adaptive intelligent assistants.

In our view, agents in BAC systems should extend beyond offering rapid processing, intelligent user intent recognition, and user-friendly interaction to evolve as dynamic and adaptive tools through iterative learning, personalization, and interoperability. This paradigm extension acknowledges the deeply personal and sensitive nature of brain activity analysis. To realize this, we emphasize the necessity of ethical data practices, reliable models, and human-agent collaboration to ensure safety and accountability. We propose reframing agents as *active and collaborative assistants* rather than *passive data processors*, with implementation guidelines and evaluation frameworks that transcend narrow technical metrics to encompass trustworthiness and effectiveness for meaningful and actionable outcomes. This position paper makes the following key contributions:

- **Analysis of BCI Systems and Alternative Viewpoints**. We analyze the workflows and limitations of conventional BCI systems, and then examine the perspectives of integrating LLMs for brain activity analysis in § 2 and § 3.
- **Identification of Current Progress and Key Challenges.** We review the existing LLM-based pursuit, identify recent advancements and key challenges for further development in § 4 and § 5.
- **Proposing Implementation Guidelines for Brain-Agent Collaboration System.** We propose a detailed guideline for a Brain-Agent Collaboration (BAC) system, outlining its core mechanisms, component design, and evaluation protocols in § 6.

2 Preliminaries of Brain-Computer Interface Systems

A Brain-Computer Interface (BCI) system establishes a direct communication pathway between the electrical activity of the brain and an external device, typically a computer or a robotic limb [21]. It scientifically regulates brain activity to facilitate effective rehabilitation, not to "control the brain" in a manipulative or harmful manner [19]. Technically, a BCI system transforms raw neural signals into actionable commands, providing feedback to the user to facilitate control and learning [13].

2.1 General Workflow of BCI Systems

A typical BCI system operates through a sequence of distinct yet interconnected stages, forming a closed loop that allows a user to interact with an external device using their brain activity. Generally, it contains the following five stages:

1. **Brain Signal Acquisition.** It involves the recording of brain activity that is related to the user's intentions, mental tasks, or responses to specific external stimuli. Signal acquisition modalities fall into two broad categories: non-invasive methods, including Electroencephalographic (EEG), Magnetoencephalographic (MEG), and Functional Near-Infrared Spectroscopy (fNIRS), and invasive approaches such as Electrocorticographic (ECoG) and Local Field Potentials (LFPs). The selection of an appropriate modality depends on several key considerations: the required signal quality (particularly spatial and temporal resolution), acceptable invasiveness levels, portability constraints, and the intended application [2, 94].
2. **Signal Pre-processing.** Raw brain signals, as acquired by the sensors, are often weak and heavily contaminated by noise and artifacts originating from both physiological and non-physiological sources [13]. The signal pre-processing stage applies various approaches, e.g., filtering techniques [27, 93], for artifact detection and removal. This stage is vital for ensuring the purity and accuracy of the brain signals that will be used for subsequent analysis, as artifacts can severely distort the underlying neural information and lead to misinterpretation of the user's intent, potentially causing unintentional control of the BCI device [45].
3. **Feature Extraction.** Feature extraction aims to reduce the data dimensionality while retaining the most informative patterns for the BCI task. Conventional dimensionality reduction algorithms

are applicable, e.g., Principal Component Analysis (PCA) [1] or Task-Discriminant Component Analysis (TDCA) [67], which enables the features discriminative between different tasks.

4. **Feature Translation for User Command Decoding.** Machine learning algorithms are commonly employed at this stage. These algorithms are trained on a dataset of brain signal features paired with known user intentions or task conditions to learn the mapping between the neural patterns and the desired commands. For instance, Li et al. [63] identify the specific frequency of the visual stimulus that the user is focusing on, thereby inferring their intended visual selection. In general, the accuracy and reliability of this translation process are critical for the overall performance of the BCI system [13].
5. **Device Operation and Feedback Loop.** The decoded commands are then transmitted to an output application or device for action execution. This could involve moving a computer cursor, controlling a robotic arm or wheelchair, or even interacting with a virtual environment. A crucial element of nearly all BCI systems is the **provision of feedback to the user** about the outcome of their executed commands [107]. This feedback can take various forms, e.g., visual, auditory, or even tactile and haptic. The feedback loop serves multiple purposes: it informs the user whether their intention was correctly interpreted by the BCI, allows them to make corrections, and, importantly, *facilitates learning*. Through this iterative feedback, users can learn to modulate their brain activity more effectively to achieve better control over the BCI system.

2.2 Limitations of Conventional BCI Systems

Despite promising advancements, BCI systems face several limitations that currently hinder their full capability realization.

Safety and Ethical Implications. These present the complex long-term challenges. Invasive techniques carry surgical risks including infection, hemorrhage, and chronic biocompatibility issues, while the gradient of invasiveness directly correlates with both signal quality and associated risks. More profoundly, BCIs raise unprecedented ethical questions about autonomy, mental privacy, and neurodiscrimination [72]. The technology challenges fundamental concepts of identity and agency by potentially accessing thoughts and influencing mental states, while current legal frameworks lag behind technological development, and such the innovation outstrips society's ability to establish appropriate safeguards.

Technical and Performance Hurdles. As for the technical side, one primary challenge in conventional BCI systems is poor signal quality, characterized by low signal-to-noise ratios, high susceptibility to artifacts from muscle activity and environmental interference, and limited spatial/temporal resolution depending on the recording modality [41, 72]. Information Transfer Rates (ITR) remain frustratingly low for many applications, while long-term stability poses a critical problem—particularly for invasive BCIs where electrode-tissue interfaces degrade over time due to biological responses like gliosis [107, 91]. The inherent non-stationarity of brain signals means that a BCI system calibrated at one point may perform poorly later, requiring frequent recalibration and creating the fundamental challenge of developing truly adaptive systems [60, 91].

User-Related and Practical Challenges. A substantial portion of users (15-30%) experience "BCI illiteracy", where they are unable to achieve reliable control despite extensive training [20, 48]. Moreover, a high inter-user variability in brain signals necessitates individual calibration, while usability concerns, including comfort, setup complexity, and cognitive fatigue, limit practical deployment [13]. Cost, portability issues, and the lack of standardized protocols across the field further impede progress, creating a significant "lab-to-life" gap where systems that work in controlled laboratory settings fail in real-world environments [119, 26].

3 Alternative Views

View 1: Extra user burden persists in managing LLM hallucinations. Integrating LLM-based agents for brain activity analysis paradoxically shifts a cognitive burden onto users who must continuously monitor outputs and correct hallucinations when decoding noisy neural signals [104]. This imposes two key burdens: *(1) Continuous Vigilance and Validation.* Users must constantly scrutinize agent outputs for inaccuracies or fabrications [42]. This dual task of vigilance *and* real-time validation is mentally taxing and slows interaction, contradicting the ideal of an effortless, intuitive interface.

(2) Challenging Error Detection and Correction. Identifying plausible hallucinations often requires multiple clarification cycles, particularly difficult given BCIs' limited feedback bandwidth [89]. Though LLMs may reduce calibration needs, managing hallucinations creates a significant cognitive load, transforming users into supervisors rather than beneficiaries.

Response. Hallucination management represents a challenge but not an insurmountable barrier for BAC [111]. This burden can be mitigated through: **(1) Advanced LLM Robustness and Grounding.** The field is developing more robust LLMs grounded in factual context using techniques like RLHF [80] and improved architectures [49]. **(2) Agent Transparency and Uncertainty Modeling.** BAC agents can communicate confidence levels [90] and seek clarification when faced with ambiguous inputs rather than producing potentially erroneous outputs [82]. **(3) User-Centric Error Correction and Iterative Learning.** Human-agent collaboration includes mechanisms for users to correct errors, with these interactions serving as feedback for system improvement [131]. The goal is a co-evolving brain-agent collaboration system that becomes increasingly reliable through continued iterative use.

View 2: Ethical Imperatives: Risks to Autonomy, Privacy, and Equity. Integrating AI agents introduces critical ethical challenges requiring proactive scrutiny. These concerns include: *(1) Autonomy Risks.* Threats to cognitive liberty and autonomy emerge as systems interpret and potentially influence neural activity. *(2) Privacy Vulnerabilities.* The sensitive nature of neural data creates unprecedented privacy concerns [12, 56], exacerbated by inadequate regulatory frameworks and risks of "cognitive hacking". *(3) Equity Challenges.* BAC technologies may deepen societal inequalities through prohibitive costs and employment discrimination while raising questions about dehumanization and accountability in shared human-AI control [128].

Response. Addressing BAC's ethical challenges is foundational to its responsible development and acceptance, requiring ethical considerations throughout its lifecycle: **(1) Proactive Governance.** Establishing robust regulatory frameworks like the OECD Neurotechnology Recommendation[1] to ensure safe integration that prioritizes human rights and the public good [53]. **(2) Neural Data Protection.** Fortifying privacy through techniques like federated learning and establishing legal "neuro rights" to safeguard cognitive integrity and user control. **(3) Trustworthy Design.** Building transparent, explainable AI systems [112] with continuous human oversight and user-centric error correction mechanisms that foster appropriate reliance [79]. **(4) Equitable Access.** Promoting broad availability through public-private partnerships and open-source initiatives, with clear policies for long-term support to prevent a "neuro-divide" and foster beneficial human-AI symbiosis.

4 Existing Pursuit: Empowering Brain Activity Analysis with LLMs

The demonstrated success of LLMs in transforming natural language processing (NLP) and the increasing capabilities of AI agents in executing complex tasks across various domains provide strong motivation for their application in the BCI field [70, 122]. In this section, we introduce recent exploration works as follows.

4.1 LLMs in Enhancing Brain Activity Analysis

The advent of LLMs offers powerful tools to tackle long-standing challenges in neural signal processing and interpretation [51].

Neural Signal Processing Enhancement. A significant impact of LLMs in BCI lies in their ability to process and interpret complex neural signals with greater sophistication, particularly in addressing issues of noise, variability, and signal alignment and decoding [123]. For example, Liu et al. [69] employ LLMs for denoising and extracting subject-independent semantic features from noisy signals. It mitigates the detrimental effects of *cross-subject variability*, which arises from individual differences in brain anatomy, neural dynamics, as well as signal acquisition conditions. It thereby enhances generalization and enables robust zero-shot predictions on unseen data. To systematically

[1] https://www.oecd.org/content/dam/oecd/en/topics/policy-sub-issues/emerging-technologies/neurotech-toolkit.pdf

address the challenges of signal noise and inter-subject variability, the research community is shifting towards building domain-specific EEG foundation models. A recent work WaveMind [124] leverages pre-training on extensive EEG data and introduces an instruction-tuning dataset. This approach aims to learn robust, subject-independent semantic features, thereby laying the foundational groundwork for the reliable BAC system development.

Natural Language Generation. LLMs are fundamentally changing how BCIs decode meaning and generate language from brain activity. Traditional BCIs often focus on recognizing a limited set of commands or characters. However, LLMs are enabling a shift towards more direct and expressive neural language decoding [47, 18], such as EEG-to-text [110, 68, 104, 4, 100, 16, 32, 120]. An innovative system, Thought2Text [75] uses instruction-tuned LLMs fine-tuned with EEG data to decode brain activity into textual outputs. BrainLLM [122] addresses language reconstruction directly from functional magnetic resonance imaging (fMRI) data. It maps neural representations, decoded from fMRI signals, to the LLM's text embedding space to generate continuous language. Another approach, namely Neural Spelling [50], utilizes a non-invasive EEG-based BCI combined with LLMs to enhance this spell-based neural language decoding. The integrated LLMs enable the model to effectively perform tasks like generative error correction and sentence completion.

4.2 LLM-based Agents in Reshaping Intelligent Assistive Technologies

Beyond the direct application of LLMs for signal processing and language decoding, the broader capabilities of LLM-based agents offer the potential for more adaptive, personalized, and collaborative BCI experiences.

Personalized and Adaptive Interaction. LLM-based agents are being developed to create BCI systems that can actively adapt to the user's real-time cognitive or affective state. An innovative example is a BCI system that deeply integrates a steady-state visual evoked potential (SSVEP) speller with an LLM API [51]. In this system, users can input natural language commands or text through the SSVEP speller, i.e., a common BCI paradigm where users focus their gaze on visual stimuli flickering at different frequencies. The LLM agent then processes this input and dynamically generates or adapts the SSVEP paradigms themselves, as well as the task interfaces presented to the user. This LLM-driven agent system offers a high degree of *personalization and adaptability*, overcoming traditional limitations of single functionality and low levels of intelligence. NeuroChat [6] is another example of such a system. It functions as a neuroadaptive AI tutor that integrates real-time EEG-based engagement tracking with an LLM agent. The system continuously monitors a learner's level of cognitive engagement, which is computationally derived from specific EEG frequency bands (alpha, beta, and theta power). Based on this inferred engagement level, NeuroChat dynamically adjusts the complexity of the educational content it presents, the style of its responses, and the overall pacing of the learning interaction.

LLM-driven Agents as Collaborators. The utility of LLM-driven agents extends beyond their role within the BCI system itself; they are also emerging as valuable human collaborators in the research and development process of BCI technology. A recent work [54] exemplifies such an application, utilizing LLMs like GPT-4o to foster human-AI collaboration specifically within BCI research projects. This includes brainstorming research ideas, generating codes for implementing neural network decoders for EEG signals, performing exploratory data analysis, and interpreting complex results and analytical plots. Its agent interaction framework follows a set of "Janusian Design Principles," which emphasize bidirectional transparency between the human and AI, the development of a shared knowledge base, and the concept of adaptive autonomy, where the AI agent can adjust its level of independence based on the task at hand. Furthermore, CorText framework [8], for instance, represents a step forward by integrating neural activity directly into the latent space of a large language model, enabling open-ended, natural language interaction with brain data. The system not only decodes information but can also answer follow-up questions about the decoded content, forming a dynamic, conversational loop.

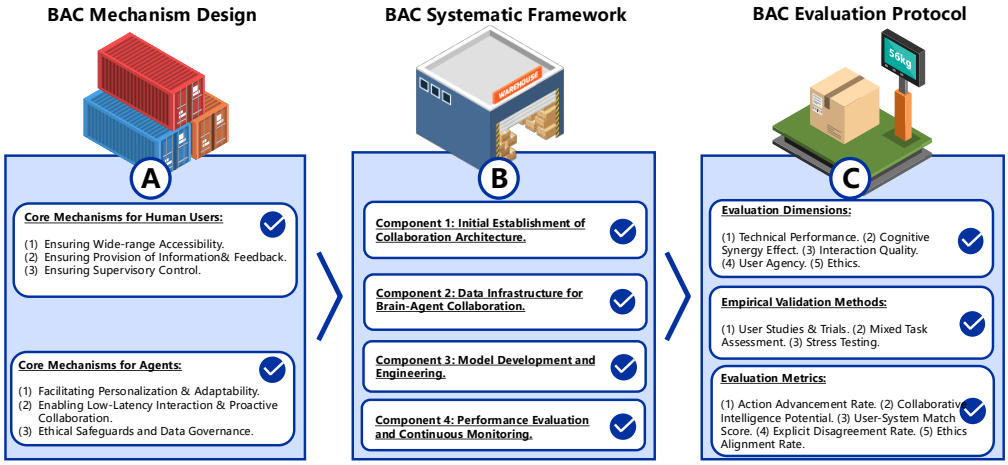

Figure 2: Brain-Agent Collaboration implementation guidelines and evaluation protocol.

5 Key Challenges in Employing LLMs for Intelligent Assistive Technologies

The integration of LLMs and agents into intelligent assistive technologies presents significant challenges. Here we highlight several key challenges:

Robust Neural Signal Interpretation and LLM Integration. A core challenge is the inherent nature of neural signals, which are noisy, non-stationary, and highly variable across and within individuals [87]. LLMs, typically trained on structured text, may struggle with these low signal-to-noise ratio inputs.

Navigating Profound Ethical and Privacy Dilemmas. Neural data is extremely sensitive, raising concerns about mental privacy and neurosecurity [92]. LLMs may potentially memorize and leak this information or infer unintended mental states. Existing legal frameworks are often insufficient for the complexities of neural data, highlighting the critical need for effective ethical data handling with privacy-preserving [117].

Ensuring Safety, Security, and Adversarial Robustness. LLMs are susceptible to adversarial attacks like jailbreaking and prompt injection [62], which could lead to dangerous outcomes in a brain activity analysis context, such as incorrect control of assistive devices. The unreliability of current agent safety evaluations and the risk of misinterpreting ambiguous neural signals demand robust ethical safeguards [43, 126, 113].

Maintaining User Agency, Control, and System Transparency. The "black box" nature of LLMs [115] impedes user trust and understanding of system decisions, which is crucial for safety-critical Brain-Agent applications [88]. If users fail to comprehend the LLM's reasoning, their sense of agency is diminished. This challenge underscores the need to ensure supervisory control and evaluation dimensions focusing on user agency and interaction quality.

6 Brain-Agent Collaboration Implementation Guidelines

The development of aforementioned works demonstrates that the integration of LLMs is not just a processing tool, but also an active participant and human collaborator that can facilitate more complex interactions with *reasoning*, *planning*, and *acting* capabilities. This catalyzes a fundamental paradigm extension from conventional BCI towards what can be termed **Brain-Agent Collaboration (BAC)**. This extension fundamentally alters the dynamic between human users and external devices, establishing humans as essential contributors who provide supplementary information, feedback, and

interactive control to LLM-powered agents, thereby improving system performance, reliability, and safety [28, 96]. We summarize our advocacy in Figure 2.

6.1 Mechanism Designs of Brain-Agent Collaboration

BAC builds upon core LLM agent components and further places critical emphasis on the brain activity understanding. To build an effective BAC framework, we discuss several key mechanisms for agents and human users, as follows.

Core Mechanisms for Human Users. Different from many BCI paradigms that particularly rely on the user learning to voluntarily modulate their brain activity [20, 48], we emphasize that BAC should be more user-centered. Therefore, BAC should satisfy the following requirements for engaged users: **(1) Ensuring Wide-range Accessibility.** To achieve widespread adoption, BAC systems must be designed to be accessible to a diverse range of users, including those with varying selections of data modality, levels of cognitive and physical abilities [38]. **(2) Ensuring Provision of Information and Feedback.** Humans contribute essential information, such as credential information and domain expertise, that agents cannot reliably deduce independently [78, 58]. Additionally, human evaluation of agent outputs through feedback mechanisms, from basic ratings to sophisticated critiques, demonstrations, or corrections, serves as crucial guidance for agent refinement [30, 24]. **(3) Ensuring Supervisory Control**. A fundamental requirement for a BAC system is that users must perceive and maintain a strong sense of control over the collaborative process, even when partnered with a highly intelligent agent. This includes unambiguous neural intent decoding and direct supervisory control with minimal cognitive load [97, 57].

Core Mechanisms for Agents. We then underpin the following mechanisms for agents: **(1) Facilitating Personalization and Adaptability.** Given the inherent variability in brain signals and cognitive styles, a one-size-fits-all approach is insufficient for BAC systems. Personalization and adaptability of agents are key to ensuring the effective of BAC systems [71, 52]. **(2) Enabling Low-Latency Interaction and Proactive Collaboration.** It is crucial to ensure low-latency, high-bandwidth data exchange between the neural interface and agent for real-time interaction [118]. Furthermore, agents must exhibit a degree of intelligence to explore, understand the deeper context, and actively contribute to the collaborative goals. This requires agents to be equipped with capabilities of memorizing, reasoning, and goal anticipation [102, 125]. **(3) Ethical Safeguards and Data Governance.** These considerations are not supplementary add-ons but are foundational to address the previous challenges. The paradigm extension from BCI as a tool to BAC as a partner is predicated on establishing trust between the user and the agent. Given the data sensitivity, it must be engineered through transparent and verifiable mechanisms. Therefore, robust ethical safeguards and clear data governance are not simply afterthoughts [129]; they are a core functional prerequisite for enabling the safe and effective human-agent partnership that defines BAC.

6.2 Systematic Framework of Brain-Agent Collaboration

Following the above mechanism designs, we introduce the following four key components.

Component 1: Initial Establishment of Collaboration Architecture. Brain-Agent Collaboration centers on human brain signals decoding to eventually execute agent actions, typically powered by Large Language Models (LLMs) [59]. Agents are expected to deconstruct complex problems into manageable sub-tasks, reason over available data, leverage appropriate tools, and learn from interactional feedback. Here, the focus is on human-AI teaming rather than full automation, with architectures designed to incorporate human oversight. Notably, architectural configurations can vary from the single-agent setting, suitable for well-delineated tasks, to multiple-agent settings, where multiple AI agents collaborate or compete, pooling diverse capabilities to address intricate challenges. These agents can operate on different foundation models, each tailored to specific roles within the collaborative endeavor. Generally, BAC architectures follow an *Interpretation–Communication–Interaction* paradigm, where the integration of agent roles improves the effectiveness of human brain signal cognition [86].

Component 2: Data Infrastructure for Brain-Agent Collaboration. The data infrastructure of the BAC systems may be with multi-modality [11, 14, 5, 17], including EEG, fMRI, fNIRS, and ECoG,

etc. Based on these data, a robust data pre-processing pipeline is essential, where AI-powered noise reduction and feature extraction techniques can be employed for better signal quality [54]. Lastly, a critical, yet less detailed, requirement for advanced BAC systems is the explicit synchronization and integration of diverse data streams, particularly for real-time brain-agent engagement [86].

Component 3: Model Development and Engineering. This component employs LLM-based agents as fundamental building blocks for reasoning and interactive functionalities. These agents are typically augmented with specialized modules for planning, memory management, tool utilization, and machine learning algorithm coordination [34, 59]. A critical prerequisite for model development is self-improvement capability driven by human feedback. Key techniques include reinforcement learning from human feedback (RLHF) [109, 23], reinforcement learning from AI feedback (RLAIF) [61], and direct preference optimization (DPO) [46]. Furthermore, future model engineering endeavors to bridge the "semantic gap" between low-level, often noisy brain signal data and the high-level cognitive states or intentions required for human interpretation.

Component 4: Performance Evaluation and Continuous Monitoring. Evaluating BAC systems requires a comprehensive approach that examines not only task completion rates but also the quality of brain-agent collaboration and the specific effects on human users [3]. We detail the evaluation protocol in the next section. Additionally, robust BAC systems must incorporate continuous monitoring capabilities to address the inherent non-stationarity of brain signals, fluctuations in human cognitive states (including fatigue and attention variations), and the dynamic behavioral patterns of AI agents, particularly large language models [44].

6.3 Illustrative Scenarios

To illustrate how these components combine in practice, we provide the following condensed examples across different use cases:

- **Daily Living Support**: A user with severe motor impairments wants to manage their environment. Instead of spelling commands, they form a high-level goal. The BAC agent, built on our described framework, interprets this and then proactively asks clarifying questions. This interaction demonstrates the core mechanisms (Section 6.1) of personalization and supervisory control, allowing the user to provide simple neural “yes/no” approvals with minimal cognitive load.
- **Neuro-rehabilitation**: A stroke survivor utilizes a robotic exoskeleton. A conventional BCI would detect motor imagery to trigger a discrete and pre-programmed action. The BAC paradigm extends this by additionally monitoring affective and cognitive states, such as neural markers of fatigue or frustration. Upon detecting a suboptimal state, the agent adaptively modulates the rehabilitation task, thus establishing a collaborative feedback loop optimized for recovery.

These scenarios illustrate the paradigm shift from BCI to BAC as a reasoning and adaptive partner, integrating the mechanisms and framework components previously discussed. Ensuring that such complex, collaborative systems are effective, safe, and robustly aligned with user goals necessitates a comprehensive evaluation strategy. We will now detail this evaluation protocol.

6.4 Evaluation Protocol of Brain-Agent Collaboration

Building upon the previous discussion in our systematic framework, establishing a dedicated evaluation protocol is crucial for validating the efficacy and safety of BAC systems. Such a protocol must extend beyond traditional technical benchmarks to holistically capture the nuances of the human-agent partnership. Therefore, we propose a multi-faceted protocol structured around the following key dimensions, metrics, and empirical validation methods.

Evaluation Dimensions. A comprehensive evaluation of BAC systems requires examining both *technical performance* and *human-centric* aspects of collaboration. Our protocol focuses on five key dimensions across diverse applications and user populations: **(1) Technical Performance.** It assesses the core operational efficiency and accuracy of the integrated BAC system, focusing on how effectively the agent decodes brain signals and executes actions to achieve shared goals and maintain system stability [81, 86]. **(2) Cognitive Synergy Effect.** This evaluates how effectively the agent augments human cognitive processes [108], leading to enhanced thinking, problem-solving, and the

emergence of collective intelligence [66]. **(3) Interaction Quality.** It focuses on the seamlessness and effectiveness of the brain-agent interaction [29, 9], encompassing communication quality, feedback mechanisms, and the overall user experience [85]. **(4) User Agency.** This examines the extent to which the human user retains control, autonomy, and strategic oversight within the collaborative system, mitigating risks of over-reliance and ensuring alignment with human values [99]. **(5) Ethics.** This dimension assesses the system's adherence to predefined ethical safeguards. It evaluates aspects like data privacy, transparency in agent reasoning, and the robustness of mechanisms that prevent neural data misuse or unintended influence, ensuring the collaboration remains verifiably aligned with human values.

Evaluation Metrics. Building on the core evaluation dimensions, we propose a diverse set of metrics to quantify the performance and impact of BAC systems. **(1) Action Advancement Rate (AAR).** This measures how effectively the agent advances the user's goals. It is calculated as the percentage of agent-driven interactions or outputs that are factually accurate, directly relevant to user objectives, and consistent with overall system parameters. A higher AAR indicates that the agent is an effective collaborator, meaningfully contributing to task success. **(2) Collaborative Intelligence Potential (CIP).** CIP is a dynamic score that assesses how well humans think with agents, examining iteration, metacognitive engagement, creativity, and refinement loops. CIP is calculated as: $\text{CIP} = f(\text{Iteration}, \text{Metacognitive Engagement}, \text{Creativity}, \text{Refinement Loops})$, where f represents a function that integrates these qualitative dimensions, often derived from a detailed transcript analysis or human evaluation. A high CIP score would indicate that users are leveraging agents to explore solutions and refine concepts in ways that enhance creative output beyond what either could achieve alone. **(3) User-System Match Score (USMS).** USMS evaluates the alignment between the BAC system and user needs, preferences, and overall experience [85, 81]. It is typically derived from standardized questionnaires, where users rate various aspects of the system on a Likert scale (e.g., 1 to 5). USMS is calculated as: $\text{USMS} = \boldsymbol{Avg}(\text{Likert Scale Scores})$. Scores between 4 and 5 (with 5 being the highest) indicate a good alignment, suggesting high user acceptance and perceived utility. **(4) Explicit Disagreement Rate (EDR).** EDR tracks how often users override or challenge agent decisions [73], calculated as $\text{EDR} = (\frac{\text{\# Explicit Disagreements}}{\text{\# Interactions}}) \times 100\%$. Higher rates indicate users are actively monitoring and critically evaluating agent outputs rather than passively accepting them, demonstrating robust human-in-the-loop dynamics. **(5) Ethics Alignment Rate (EAR).** EAR is a composite metric that can combine user-reported trust and safety scores with the system's success rate on standardized ethical stress tests, e.g., probing for data leakage, undue influence, or privacy violations.

Empirical Validation Methods. To assess the real-world effectiveness and safety of BAC systems, we further propose a multi-method validation strategy that combines quantitative and qualitative measures. **(1) User Studies & Trials:** Conduct rigorous user studies and clinical trials with diverse populations in real-world scenarios. Longitudinal studies are particularly important for assessing long-term effects and system stability. **(2) Mixed Task Assessment.** Design experiments that involve a variety of tasks, ranging from automated decision-making to creative problem-solving, to comprehensively evaluate the BAC system versatility and adaptability [29]. **(3) Stress Testing.** Simulate real-world disruptions and challenging conditions to assess the BAC system's robustness, reliability, and effectiveness of its responses under duress [33].

7 Conclusion

This position paper argues for a paradigm extension from Brain-Computer Interface (BCI) to Brain-Agent Collaboration (BAC), driven by the integration of LLM-based agents. We advocate for the agent roles as active assistants rather than passive data processors. The proposed implementation guidelines and evaluation frameworks provide a starting point for developing trustworthy and effective BAC systems. It is imperative that the research community, developers, and stakeholders proactively adopt these measures to ensure that, as we integrate AI agents with human cognition, the resulting technologies are safe, ethical, and truly beneficial in real-world applications.

Acknowledgments and Disclosure of Funding

The authors thank MBZUAI startup fund for supporting part of this research.

References

[1] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

[2] Swati Aggarwal and Nupur Chugh. Review of machine learning techniques for eeg based brain computer interface. *Archives of Computational Methods in Engineering*, 29(5):3001–3020, 2022.

[3] Abeer Alabbas and Khalid Alomar. A weighted composite metric for evaluating user experience in educational chatbots: Balancing usability, engagement, and effectiveness. *Future Internet*, 17(2):64, 2025.

[4] Hamza Amrani, Daniela Micucci, and Paolo Napoletano. Deep representation learning for open vocabulary electroencephalography-to-text decoding. *IEEE Journal of Biomedical and Health Informatics*, pp. 1–12, 2024. ISSN 2168-2208. doi: 10.1109/jbhi.2024.3416066. URL http://dx.doi.org/10.1109/JBHI.2024.3416066.

[5] Naseem Babu, Jimson Mathew, and AP Vinod. Large language models for eeg: A comprehensive survey and taxonomy. *arXiv preprint arXiv:2506.06353*, 2025.

[6] Dünya Baradari, Nataliya Kosmyna, Oscar Petrov, Rebecah Kaplun, and Pattie Maes. Neurochat: A neuroadaptive ai chatbot for customizing learning experiences. *arXiv preprint arXiv:2503.07599*, 2025.

[7] Marcel Binz, Stephan Alaniz, Adina Roskies, Balazs Aczel, Carl T Bergstrom, Colin Allen, Daniel Schad, Dirk Wulff, Jevin D West, Qiong Zhang, et al. How should the advancement of large language models affect the practice of science? *Proceedings of the National Academy of Sciences*, 122(5):e2401227121, 2025.

[8] Victoria Bosch, Daniel Anthes, Adrien Doerig, Sushrut Thorat, Peter König, and Tim Christian Kietzmann. Brain-language fusion enables interactive neural readout and in-silico experimentation. *arXiv preprint arXiv:2509.23941*, 2025.

[9] Francesco Bossi, Francesca Ciardo, and Ghilès Mostafaoui. Neurocognitive features of human-robot and human-machine interaction, 2024.

[10] Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

[11] Alessio Paolo Buccino, Hasan Onur Keles, and Ahmet Omurtag. Hybrid eeg-fnirs asynchronous brain-computer interface for multiple motor tasks. *PloS one*, 11(1):e0146610, 2016.

[12] José J Cañas. Ai and ethics when human beings collaborate with ai agents. *Frontiers in psychology*, 13:836650, 2022.

[13] Vinay Chamola, Ankur Vineet, Anand Nayyar, and Eklas Hossain. Brain-computer interface-based humanoid control: A review. *Sensors*, 20(13):3620, 2020.

[14] Jiafa Chen, Kaiwei Yu, Yifei Bi, Xing Ji, and Dawei Zhang. Strategic integration: A cross-disciplinary review of the fnirs-eeg dual-modality imaging system for delivering multimodal neuroimaging to applications. *Brain Sciences*, 14(10):1022, 2024.

[15] Junhong Chen, Ziqi Yang, Haoyuan G Xu, Dandan Zhang, and George Mylonas. Multi-agent systems for robotic autonomy with llms. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4194–4204, 2025.

[16] Qiupu Chen, Yimou Wang, Fenmei Wang, Duolin Sun, and Qiankun Li. Decoding text from electroencephalography signals: A novel hierarchical gated recurrent unit with masked residual attention mechanism. *Engineering Applications of Artificial Intelligence*, 139:109615, 2025. ISSN 0952-1976. doi: https://doi.org/10.1016/j.engappai.2024.109615. URL https://www.sciencedirect.com/science/article/pii/S0952197624017731.

[17] Sitong Chen, Beiqianyi Li, Cuilin He, Dongyang Li, Mingyang Wu, Xinke Shen, Song Wang, Xuetao Wei, Xindi Wang, Haiyan Wu, et al. Chineseeeg-2: An eeg dataset for multimodal semantic alignment and neural decoding during reading and listening. *arXiv preprint arXiv:2508.04240*, 2025.

[18] Xiaoyu Chen, Changde Du, Che Liu, Yizhe Wang, and Huiguang He. Bp-gpt: Auditory neural decoding using fmri-prompted llm, 2025. URL https://arxiv.org/abs/2502.15172.

[19] Yanxiao Chen, Fan Wang, Tianwen Li, Lei Zhao, Anmin Gong, Wenya Nan, Peng Ding, and Yunfa Fu. Considerations and discussions on the clear definition and definite scope of brain-computer interfaces. *Frontiers in Neuroscience*, 18:1449208, 2024.

[20] Jennifer Chmura, Joshua Rosing, Steven Collazos, and Shikha J Goodwin. Classification of movement and inhibition using a hybrid bci. *Frontiers in neurorobotics*, 11:38, 2017.

[21] Wikipedia contributors. Brain–computer interface — Wikipedia, the free encyclopedia, May 2025. URL https://en.wikipedia.org/wiki/Brain%E2%80%93computer_interface. [Online; accessed 21-May-2025].

[22] Nikhil J Dhinagar, Sophia I Thomopoulos, and Paul M Thompson. Leveraging a vision-language model with natural text supervision for mri retrieval, captioning, classification, and visual question answering. *bioRxiv*, pp. 2025–02, 2025.

[23] Shangheng Du, Jiabao Zhao, Jinxin Shi, Zhentao Xie, Xin Jiang, Yanhong Bai, and Liang He. A survey on the optimization of large language model-based agents. *arXiv preprint arXiv:2503.12434*, 2025.

[24] Subhabrata Dutta, Timo Kaufmann, Goran Glavaš, Ivan Habernal, Kristian Kersting, Frauke Kreuter, Mira Mezini, Iryna Gurevych, Eyke Hüllermeier, and Hinrich Schuetze. Problem solving through human-ai preference-based cooperation. *arXiv preprint arXiv:2408.07461*, 2024.

[25] Austen El-Osta, Mahmoud Al Ammouri, Shujhat Khan, Sami Altalib, Manisha Karki, Eva Riboli-Sasco, and Azeem Majeed. Community perspectives regarding brain-computer interfaces: A cross-sectional study of community-dwelling adults in the uk. *PLOS Digital Health*, 4(2):e0000524, 2025.

[26] Walaa H Elashmawi, Abdelrahman Ayman, Mina Antoun, Habiba Mohamed, Shehab Eldeen Mohamed, Habiba Amr, Youssef Talaat, and Ahmed Ali. A comprehensive review on brain–computer interface (bci)-based machine and deep learning algorithms for stroke rehabilitation. *Applied Sciences*, 14(14):6347, 2024.

[27] Monica Fabiani, Gabriele Gratton, and Kara D Federmeier. Event-related brain potentials: Methods, theory, and applications. *Handbook of psychophysiology*, 3:85–119, 2007.

[28] Xueyang Feng, Zhi-Yuan Chen, Yujia Qin, Yankai Lin, Xu Chen, Zhiyuan Liu, and Ji-Rong Wen. Large language model-based human-agent collaboration for complex task solving. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1336–1357, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.72. URL https://aclanthology.org/2024.findings-emnlp.72/.

[29] George Fragiadakis, Christos Diou, George Kousiouris, and Mara Nikolaidou. Evaluating human-ai collaboration: A review and methodological framework. *arXiv preprint arXiv:2407.19098*, 2024.

[30] Jie Gao, Simret Araya Gebreegziabher, Kenny Tsu Wei Choo, Toby Jia-Jun Li, Simon Tangi Perrault, and Thomas W Malone. A taxonomy for human-llm interaction modes: An initial exploration. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–11, 2024.

[31] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1), 2023.

[32] Mostafa El Gedawy, Omnia Nabil, Omar Mamdouh, Mahmoud Nady, Nour Alhuda Adel, and Ahmed Fares. Bridging brain signals and language: A deep learning approach to eeg-to-text decoding, 2025. URL https://arxiv.org/abs/2502.17465.

[33] Evgenia Gkintoni, Hera Antonopoulou, Andrew Sortwell, and Constantinos Halkiopoulos. Challenging cognitive load theory: The role of educational neuroscience and artificial intelligence in redefining learning efficacy. *Brain Sciences*, 15(2), 2025.

[34] Joshua I Glaser, Ari S Benjamin, Raeed H Chowdhury, Matthew G Perich, Lee E Miller, and Konrad P Kording. Machine learning for neural decoding. *eneuro*, 7(4), 2020.

[35] Moshe Glickman and Tali Sharot. How human–ai feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, 9(2):345–359, 2025.

[36] Diego Gosmar and Deborah A Dahl. Hallucination mitigation using agentic ai natural language-based frameworks. *arXiv preprint arXiv:2501.13946*, 2025.

[37] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.

[38] Dorit Hadar Souval, Yuval Haber, Amir Tal, Tomer Simon, Tal Elyoseph, and Zohar Elyoseph. Transforming perceptions: exploring the multifaceted potential of generative ai for people with cognitive disabilities. *JMIR Neurotechnology*, 4:e64182, 2025.

[39] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*, 2023.

[40] Dong Huang, Jie M Zhang, Michael Luck, Qingwen Bu, Yuhao Qing, and Heming Cui. Agentcoder: Multi-agent-based code generation with iterative testing and optimisation. *arXiv preprint arXiv:2312.13010*, 2023.

[41] Jinghui Huang, Lele Huang, Ying Li, and Fanfu Fang. A bibliometric analysis of the application of brain-computer interface in rehabilitation medicine over the past 20 years. *Journal of Multidisciplinary Healthcare*, pp. 1297–1317, 2025.

[42] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, January 2025. ISSN 1558-2868. doi: 10.1145/3703155. URL http://dx.doi.org/10.1145/3703155.

[43] Wei-Chieh Huang, Henry Peng Zou, Yaozu Wu, Dongyuan Li, Yankai Chen, Weizhi Zhang, Yangning Li, Angelo Zangari, Jizhou Guo, Chunyu Miao, et al. Deepresearchguard: Deep research with open-domain evaluation and multi-stage guardrails for safety. *arXiv preprint arXiv:2510.10994*, 2025.

[44] Syed Abu Huraira Hussain, Imran Raza, Syed Asad Hussain, Muhammad Hasan Jamal, Tauseef Gulrez, and Ali Zia. A mental state aware brain computer interface for adaptive control of electric powered wheelchair. *Scientific Reports*, 15(1):9880, 2025.

[45] Md Kafiul Islam and Amir Rastegarnia. Recent advances in eeg (non-invasive) based bci applications. *Frontiers in Computational Neuroscience*, 17:1151852, 2023.

[46] Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A Smith, Yejin Choi, and Hanna Hajishirzi. Unpacking dpo and ppo: Disentangling best practices for learning from preference feedback. *Advances in neural information processing systems*, 37:36602–36633, 2024.

[47] Dulhan Jayalath, Gilad Landau, and Oiwi Parker Jones. Unlocking non-invasive brain-to-text, 2025. URL https://arxiv.org/abs/2505.13446.

[48] Camille Jeunet, Emilie Jahanpour, and Fabien Lotte. Why standard brain-computer interface (bci) training protocols should be changed: an experimental study. *Journal of neural engineering*, 13(3):036024, 2016.

[49] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating LLM hallucination via self reflection. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1827–1843, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.123. URL https://aclanthology.org/2023.findings-emnlp.123/.

[50] Xiaowei Jiang, Charles Zhou, Yiqun Duan, Ziyi Zhao, Thomas Do, and Chin-Teng Lin. Neural spelling: A spell-based bci system for language neural decoding. *arXiv preprint arXiv:2501.17489*, 2025.

[51] Jing Jin, Yutao Zhang, Ruitian Xu, and Yixin Chen. An innovative brain-computer interface interaction system based on the large language model. *arXiv preprint arXiv:2502.11659*, 2025.

[52] Wenjie Jin, XinXin Zhu, Lifeng Qian, Cunshu Wu, Fan Yang, Daowei Zhan, Zhaoyin Kang, Kaitao Luo, Dianhuai Meng, and Guangxu Xu. Electroencephalogram-based adaptive closed-loop brain-computer interface in neurorehabilitation: a review. *Frontiers in Computational Neuroscience*, 18:1431815, 2024.

[53] Himanshu Joshi. Ai governance by design for agentic systems: A framework for responsible development and deployment. 2025.

[54] Maryna Kapitonova and Tonio Ball. Human-ai teaming using large language models: Boosting brain-computer interfacing (bci) and brain research. *arXiv preprint arXiv:2501.01451*, 2024.

[55] Aleksandra Kawala-Sterniuk, Natalia Browarska, Amir Al-Bakri, Mariusz Pelc, Jaroslaw Zygarlicki, Michaela Sidikova, Radek Martinek, and Edward Jacek Gorzelanczyk. Summary of over fifty years with brain-computer interfaces—a review. *Brain sciences*, 11(1):43, 2021.

[56] Philipp Kellmeyer. Big brain data: On the responsible use of brain data from clinical and consumer-directed neurotechnological devices. *Neuroethics*, 14(1):83–98, 2021.

[57] Fahad Khan, Yufeng Wu, Julia Dray, Bronwyn Hemsley, and A Baki Kocaballi. Conversational agents to support people with communication disability: A co-design study with speech pathologists. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–9, 2025.

[58] JiWoo Kim, Minsuk Chang, and JinYeong Bak. Beyond turn-taking: Introducing text-based overlap into human-llm interactions. *arXiv preprint arXiv:2501.18103*, 2025.

[59] Naveen Krishnan. Ai agents: Evolution, architecture, and real-world applications. *arXiv preprint arXiv:2503.12687*, 2025.

[60] Pieter Kubben et al. Invasive brain-computer interfaces: A critical assessment of current developments and future prospects. *JMIR Neurotechnology*, 3(1):e60151, 2024.

[61] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. 2023.

[62] Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, Jie Huang, and Yangqiu Song. Multi-step jailbreaking privacy attacks on chatgpt. *ArXiv*, abs/2304.05197, 2023. URL https://arxiv.org/pdf/2304.05197.pdf.

[63] Minglun Li, Dianning He, Chen Li, and Shouliang Qi. Brain–computer interface speller based on steady-state visual evoked potential: a review focusing on the stimulus paradigm and performance. *Brain sciences*, 11(4):450, 2021.

[64] Siyang Li, Hongbin Wang, Xiaoqing Chen, and Dongrui Wu. Multimodal brain-computer interfaces: Ai-powered decoding methodologies. *arXiv preprint arXiv:2502.02830*, 2025.

[65] Yangning Li, Weizhi Zhang, Yuyao Yang, Wei-Chieh Huang, Yaozu Wu, Junyu Luo, Yuanchen Bei, Henry Peng Zou, Xiao Luo, Yusheng Zhao, et al. Towards agentic rag with deep reasoning: A survey of rag-reasoning systems in llms. *arXiv preprint arXiv:2507.09477*, 2025.

[66] Bang Liu, Xinfeng Li, Jiayi Zhang, Jinlin Wang, Tanjin He, Sirui Hong, Hongzhang Liu, Shaokun Zhang, Kaitao Song, Kunlun Zhu, et al. Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems. *arXiv preprint arXiv:2504.01990*, 2025.

[67] Bingchuan Liu, Xiaogang Chen, Nanlin Shi, Yijun Wang, Shangkai Gao, and Xiaorong Gao. Improving the performance of individually calibrated ssvep-bci by task-discriminant component analysis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1998–2007, 2021.

[68] Hanwen Liu, Daniel Hajialigol, Benny Antony, Aiguo Han, and Xuan Wang. Eeg2text: Open vocabulary eeg-to-text decoding with eeg pre-training and multi-view transformer. *arXiv preprint arXiv:2405.02165*, 2024.

[69] Yifei Liu, Hengwei Ye, and Shuhang Li. Llms help alleviate the cross-subject variability in brain signal and language alignment. *arXiv preprint arXiv:2501.02621*, 2025.

[70] Xiaoliang Luo, Akilles Rechardt, Guangzhi Sun, Kevin K Nejad, Felipe Yáñez, Bati Yilmaz, Kangjoo Lee, Alexandra O Cohen, Valentina Borghesani, Anton Pashkov, et al. Large language models surpass human experts in predicting neuroscience results. *Nature human behaviour*, 9 (2):305–315, 2025.

[71] Yixin Ma, Anmin Gong, Wenya Nan, Peng Ding, Fan Wang, and Yunfa Fu. Personalized brain–computer interface and its applications. *Journal of Personalized Medicine*, 13(1):46, 2022.

[72] Sergio Machado, Fernanda Araújo, Flávia Paes, Bruna Velasques, Mario Cunha, Henning Budde, Luis F Basile, Renato Anghinah, Oscar Arias-Carrión, Mauricio Cagy, et al. Eeg-based brain-computer interfaces: an overview of basic concepts and clinical applications in neurorehabilitation. *Reviews in the Neurosciences*, 21(6):451–468, 2010.

[73] Rahul Tushar Mehta. Intent visualization in human-agent teams. 2024.

[74] Chunyu Miao, Henry Peng Zou, Yangning Li, Yankai Chen, Yibo Wang, Fangxin Wang, Yifan Li, Wooseong Yang, Bowei He, Xinni Zhang, et al. Recode-h: A benchmark for research code development with interactive human feedback. *arXiv preprint arXiv:2510.06186*, 2025.

[75] Abhijit Mishra, Shreya Shukla, Jose Torres, Jacek Gwizdka, and Shounak Roychowdhury. Thought2text: Text generation from eeg signal using large language models (llms). *NAACL*, 2025.

[76] Mahmoud Mohammadi, Yipeng Li, Jane Lo, and Wendy Yip. Evaluation and benchmarking of llm agents: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 6129–6139, 2025.

[77] Anderson Mora-Cortes, Nikolay V Manyakov, Nikolay Chumerin, and Marc M Van Hulle. Language model applications to spelling with brain-computer interfaces. *Sensors*, 14(4): 5967–5993, 2014.

[78] Riya Naik, Ashwin Srinivasan, Estrid He, and Swati Agarwal. An empirical study of the role of incompleteness and ambiguity in interactions with large language models. *arXiv preprint arXiv:2503.17936*, 2025.

[79] Ritesh Noothigattu, Djallel Bouneffouf, Nicholas Mattei, Rachita Chandra, Piyush Madan, Kush R Varshney, Murray Campbell, Moninder Singh, and Francesca Rossi. Teaching ai agents ethical values using reinforcement learning and policy orchestration. *IBM Journal of Research and Development*, 63(4/5):2–1, 2019.

[80] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

[81] He Pan, Peng Ding, Fan Wang, Tianwen Li, Lei Zhao, Wenya Nan, Yunfa Fu, and Anmin Gong. Comprehensive evaluation methods for translating bci into practical applications: usability, user satisfaction and usage of online bci systems. *Frontiers in Human Neuroscience*, 18: 1429130, 2024.

[82] Jing-Cheng Pang, Heng-Bo Fan, Pengyuan Wang, Jia-Hao Xiao, Nan Tang, Si-Hang Yang, Chengxing Jia, Sheng-Jun Huang, and Yang Yu. Empowering language models with active inquiry for deeper understanding, 2024. URL https://arxiv.org/abs/2402.03719.

[83] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

[84] Janis Peksa and Dmytro Mamchur. State-of-the-art on brain-computer interface technology. *Sensors*, 23(13):6001, 2023.

[85] Danny Plass-Oude Bos, Hayrettin Gürkök, Bram Van de Laar, Femke Nijboer, and Anton Nijholt. User experience evaluation in bci: mind the gap! *International Journal of Bioelectromagnetism*, 13(1):48–49, 2011.

[86] Yingyi Qiu, Han Liu, and Mengyuan Zhao. A review of brain–computer interface-based language decoding: From signal interpretation to intelligent communication. *Applied Sciences*, 15(1), 2025. doi: 10.3390/app15010392.

[87] Javad Rahimipour Anaraki, Antonina Kolokolova, and Tom Chau. Personalized classifier selection for eeg-based bcis. *Computers*, 13(7), 2024. ISSN 2073-431X. doi: 10.3390/computers13070158. URL https://www.mdpi.com/2073-431X/13/7/158.

[88] Param S. Rajpura, H. Cecotti, and Yogesh Kumar Meena. Explainable artificial intelligence approaches for brain–computer interfaces: a review and design space. *Journal of Neural Engineering*, 21, 2023. URL https://api.semanticscholar.org/CorpusId:266375223.

[89] Anand Ramachandran. Revolutionizing brain-computer interfaces the transformative impact of advanced ai across functional areas. 01 2025.

[90] Sumedh Rasal. Llm harmony: Multi-agent communication for problem solving, 2024. URL https://arxiv.org/abs/2401.01312.

[91] Dheeraj Rathee, Haider Raza, Sujit Roy, and Girijesh Prasad. A magnetoencephalography dataset for motor and cognitive imagery-based brain-computer interface. *Scientific Data*, 8(1): 120, 2021.

[92] Thorsten Rudroff. Decoding thoughts, encoding ethics: A narrative review of the bci-ai revolution. *Brain Research*, 1850, 2024. URL https://api.semanticscholar.org/CorpusId:274966814.

[93] Saeid Sanei and Jonathon A Chambers. *EEG signal processing*. John Wiley & Sons, 2013.

[94] Gerwin Schalk and Eric C Leuthardt. Brain-computer interfaces using electrocorticographic signals. *IEEE reviews in biomedical engineering*, 4:140–154, 2011.

[95] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36: 68539–68551, 2023.

[96] Yijia Shao, Vinay Samuel, Yucheng Jiang, John Yang, and Diyi Yang. Collaborative gym: A framework for enabling and evaluating human-agent collaboration. *arXiv preprint arXiv:2412.15701*, 2024.

[97] Jerry J Shih, Dean J Krusienski, and Jonathan R Wolpaw. Brain-computer interfaces in medicine. In *Mayo clinic proceedings*, volume 87, pp. 268–279. Elsevier, 2012.

[98] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

[99] Georg Starke, Felix Gille, Alberto Termine, Yves Saint James Aquino, Ricardo Chavarriaga, Andrea Ferrario, Janna Hastings, Karin Jongsma, Philipp Kellmeyer, Bogdan Kulynych, et al. Finding consensus on trust in ai in health care: Recommendations from a panel of international experts. *Journal of medical Internet research*, 27:e56306, 2025.

[100] Yitian Tao, Yan Liang, Luoyu Wang, Yongqing Li, Qing Yang, and Han Zhang. See: Semantically aligned eeg-to-text translation, 2024. URL https://arxiv.org/abs/2409.16312.

[101] David E Thompson, Stefanie Blain-Moraes, and Jane E Huggins. Performance assessment in brain-computer interface-based augmentative and alternative communication. *Biomedical engineering online*, 12:1–23, 2013.

[102] Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*, 2025.

[103] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

[104] Jiaqi Wang, Zhenxi Song, Zhengyu Ma, Xipeng Qiu, Min Zhang, and Zhiguo Zhang. Enhancing eeg-to-text decoding through transferable representations from pre-trained contrastive eeg-text masked autoencoder. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7278–7292. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.393. URL http://dx.doi.org/10.18653/v1/2024.acl-long.393.

[105] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

[106] Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. A survey of llm-based agents in medicine: How far are we from baymax? *arXiv preprint arXiv:2502.11211*, 2025.

[107] Yifan Wang, Cheng Jiang, and Chenzhong Li. A review of brain-computer interface technologies: Signal acquisition methods and interaction paradigms. *arXiv preprint arXiv:2503.16471*, 2025.

[108] Yingxu Wang and Vincent Chiew. On the cognitive process of human problem solving. *Cognitive systems research*, 11(1):81–92, 2010.

[109] Yuanhao Wang, Qinghua Liu, and Chi Jin. Is rlhf more difficult than standard rl? a theoretical perspective. *Advances in Neural Information Processing Systems*, 36:76006–76032, 2023.

[110] Zhenhailong Wang and Heng Ji. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 5350–5358, 2022.

[111] Jiaheng Wei, Yuanshun Yao, Jean-Francois Ton, Hongyi Guo, Andrew Estornell, and Yang Liu. Measuring and reducing llm hallucination without gold-standard answers, 2024. URL https://arxiv.org/abs/2402.10412.

[112] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. "let me explain!": exploring the potential of virtual agents in explainable ai interaction design. *Journal on Multimodal User Interfaces*, 15(2):87–98, 2021.

[113] Yaozu Wu, Jizhou Guo, Dongyuan Li, Henry Peng Zou, Wei-Chieh Huang, Yankai Chen, Zhen Wang, Weizhi Zhang, Yangning Li, Meng Zhang, et al. Psg-agent: Personality-aware safety guardrail for llm-based agents. *arXiv preprint arXiv:2509.23614*, 2025.

[114] Yaozu Wu, Dongyuan Li, Yankai Chen, Renhe Jiang, Henry Peng Zou, Wei-Chieh Huang, Yangning Li, Liancheng Fang, Zhen Wang, and Philip S. Yu. Multi-agent autonomous driving systems with large language models: A survey of recent advances, resources, and future directions. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025.

[115] Wenpeng Xing, Minghao Li, Mohan Li, and Meng Han. Towards robust and secure embodied ai: A survey on vulnerabilities and attacks. *ArXiv*, abs/2502.13175, 2025. URL https://api.semanticscholar.org/CorpusId:276449837.

[116] Hongshen Xu, Zichen Zhu, Lei Pan, Zihan Wang, Su Zhu, Da Ma, Ruisheng Cao, Lu Chen, and Kai Yu. Reducing tool hallucination via reliability alignment. *arXiv preprint arXiv:2412.04141*, 2024.

[117] Hong Yang and Li Jiang. Regulating neural data processing in the age of bcis: Ethical concerns and legal approaches. *Digital Health*, 11, 2025. URL https://api.semanticscholar.org/CorpusId:277408752.

[118] Yingxuan Yang, Huacan Chai, Yuanyi Song, Siyuan Qi, Muning Wen, Ning Li, Junwei Liao, Haoyi Hu, Jianghao Lin, Gaowei Chang, et al. A survey of ai agent protocols. *arXiv preprint arXiv:2504.16736*, 2025.

[119] Yiqian Yang. Neugaze: Reshaping the future bci. *arXiv preprint arXiv:2504.15101*, 2025.

[120] Yiqian Yang, Yiqun Duan, Hyejeong Jo, Qiang Zhang, Renjing Xu, Oiwi Parker Jones, Xuming Hu, Chin teng Lin, and Hui Xiong. Neugpt: Unified multi-modal neural gpt, 2024. URL https://arxiv.org/abs/2410.20916.

[121] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.

[122] Ziyi Ye, Qingyao Ai, Yiqun Liu, Maarten de Rijke, Min Zhang, Christina Lioma, and Tuukka Ruotsalo. Generative language reconstruction from brain recordings. *Communications Biology*, 8(1):346, 2025.

[123] Zaid Zada, Ariel Goldstein, Sebastian Michelmann, Erez Simony, Amy Price, Liat Hasenfratz, Emily Barham, Asieh Zadbood, Werner Doyle, Daniel Friedman, et al. A shared model-based linguistic space for transmitting our thoughts from brain to brain in natural conversations. *Neuron*, 112(18):3211–3222, 2024.

[124] Ziyi Zeng, Zhenyang Cai, Yixi Cai, Xidong Wang, Junying Chen, Rongsheng Wang, Yipeng Liu, Siqi Cai, Benyou Wang, Zhiguo Zhang, et al. Wavemind: Towards a conversational eeg foundation model aligned to textual and visual modalities. *arXiv preprint arXiv:2510.00032*, 2025.

[125] Weizhi Zhang, Yangning Li, Yuanchen Bei, Junyu Luo, Guancheng Wan, Liangwei Yang, Chenxuan Xie, Yuyao Yang, Wei-Chieh Huang, Chunyu Miao, et al. From web search towards agentic deep research: Incentivizing search with reasoning agents. *arXiv preprint arXiv:2506.18959*, 2025.

[126] Zhexin Zhang, Shiyao Cui, Yida Lu, Jingzhuo Zhou, Junxiao Yang, Hongning Wang, and Minlie Huang. Agent-safetybench: Evaluating the safety of llm agents. *arXiv preprint arXiv:2412.14470*, 2024.

[127] Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023.

[128] Tianyu Zhou, Pinqiao Wang, Yilin Wu, and Hongyang Yang. Finrobot: Ai agent for equity research and valuation with large language models. *arXiv preprint arXiv:2411.08804*, 2024.

[129] Xinliang Zhou, Chenyu Liu, Zhisheng Chen, Kun Wang, Yi Ding, Ziyu Jia, and Qingsong Wen. Brain foundation models: A survey on advancements in neural signal processing and brain discovery. *arXiv preprint arXiv:2503.00580*, 2025.

[130] Henry Peng Zou, Wei-Chieh Huang, Yaozu Wu, Yankai Chen, Chunyu Miao, Hoang Nguyen, Yue Zhou, Weizhi Zhang, Liancheng Fang, Langzhou He, Yangning Li, Dongyuan Li, Renhe Jiang, Xue Liu, and Philip S. Yu. A survey on large language model based human-agent systems, 2025. URL https://arxiv.org/abs/2505.00753.

[131] Henry Peng Zou, Wei-Chieh Huang, Yaozu Wu, Chunyu Miao, Dongyuan Li, Aiwei Liu, Yue Zhou, Yankai Chen, Weizhi Zhang, Yangning Li, et al. A call for collaborative intelligence: Why human-agent systems should precede ai autonomy. *arXiv preprint arXiv:2506.09420*, 2025.