Beyond Workload: Paving the Road for the Next Generation of Implicit Prefrontal Cortex Based Brain-Computer Interfaces

A dissertation submitted by

Matthew P. Russell

In partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

TUFTS UNIVERSITY

May 2025

Adviser: Robert J.K. Jacob

Abstract

The rapidly evolving field of Human-Computer Interaction (HCI) faces a fundamental constraint: the limited bandwidth of information exchange between users and computing systems. One promising approach to increasing this bandwidth is *implicit interaction*: a paradigm in which applications modify their state based on information gleaned from users, rather than direct input. Within the context of reading such information from human neural signals, this concept is formally recognized as *implicit Brain-Computer Interfaces* (implicit BCI). My work focuses on implicit BCIs which measure the prefrontal cortex (PFC); early prototypes have successfully leveraged the PFC to approximate mental workload, but much is left to be understood about the full potential of this region. Through three research projects spanning two brain measurement modalities, this dissertation makes targeted contributions to this area of research. With functional Near-Infrared Spectroscopy (fNIRS), I explore two facets of PFC activation demonstrated in functional Magnetic Resonance Imaging (fMRI)-based neuroscience research which are underexplored in applied contexts: episodic memory and brain-network based classification; in the first project, I study the measurable effects of episodic and working memory within the context of using Large Language Models (LLMs), and in the second project I develop a real-time implicit BCI designed to differentiate between different brain networks. The third project benchmarks low-cost EEG in three studies which distinguish brain states based on different factors: quality of moves made during chess playing, workload levels within standard cognitive psychology tasks, and cognitive states during the tasks. For all studies I use Linear Mixed Models (LMM) to observe macro patterns in the data, and machine learning to explore potential for implicit BCI. Results indicate that, in addition to the well-understood concept of measuring singular aspects of consciousness across a gradient (e.g. workload), promising potential exists for leveraging the PFC towards classification across tasks which engage different cognitive processes, both with fNIRS and low-cost EEG. Further, careful consideration of "noise" in implicit BCI introduces a new idea: Human-Sensor-Computer Interaction (HSCI). Taken together, this dissertation provides relevant context to inform the next generation of Human-Sensor-Computer systems, including PFC-based interfaces stretching past workload, and beyond.

Acknowledgements

Dedication

This work is dedicated to my daughters, Autumn and Moon.

Table of Contents

List of Figures

List of Tables

Beyond Workload: Paving the Road for the Next Generation of Implicit Prefrontal Cortex Based Brain-Computer Interfaces

PART I

Introduction and Background

The introductory part of this dissertation provides the context and background for the research which follows. Chapter 1 introduces the context and scope of this dissertation, and provides an outline of the parts and chapters to follow. Chapter 2 establishes the foundational background for this research: relevant neuroanatomical structures, measurement methodologies, psychophysiological signal analyses, and brain-computer interface applications. Chapter 3 covers the hardware details, data processing techniques, statistical and machine learning methods used in this dissertation.

Chapter 1

Introduction

1.1 Brain-Computer Interfaces

Increasing the bandwidth of the information exchange between humans and machines will undoubtedly shape the future of human-computer interaction. Whereas traditional interaction techniques between humans and computers - keyboards, mice, touchscreens, and more recently, voice - all are similar in the significant limitation of the upper bounds of their interaction bandwidth, the promise of directly interfacing with our neural circuitry presents a much more expansive set of possibilities. However, despite numerous explorations of BCI throughout modern science fiction, the reality is that today, the bandwidth of information exchange undergirding the current state-of-the-art BCIs, particularly those which are non-invasive, is limited to *less* than what is currently capable with a mouse and keyboard [6]. Such limitation has largely constrained the use of such BCI tools to users who are otherwise limited in their ability to communicate with machines - for instance, individuals with inability do move their limbs due to amyotrophic lateral sclerosis (ALS) or spinal cord injury [7]. While recognizing the current limitations of BCI, this work aims to advance progress towards adaptive, real-time BCIs which have sophisticated contextual understanding of the human user, potentially enabling a wide range of applications that could transform human-computer interaction.

1.2 Implicit Interaction

The methodology of interaction that bridges the communication between brain and machine in this dissertation is known as *implicit*, or passive, interaction. This interaction style is defined by the machine's monitoring of the human state of the user in real-time, and by making subtle changes to the interface based on that state [8]. That is, rather than the user actively inputting information into the machine, the machine can read information about the user and leverage that information towards helpful interface adjustments; Zander describes it as "an unconscious action that is integrated in another action" [9]. By moving from exclusively explicit command-driven interfaces to systems that can also interpret and respond to users' unconscious signals and states, this approach represents a significant paradigm shift in HCI. Potential applications of implicit interfaces span numerous domains, from adaptive environments that adjust to cognitive load [10], to assistive technologies that can anticipate user needs before explicit requests are made [11]. More specifically, this dissertation studies *implicit BCI* - that is, interfaces which leverage neural states to adaptations beneficial to the user.

1.3 Prefrontal Cortex

Naturally, the potential scope of any implicit BCI in terms of its capacity to transmit information to (or from) a machine is directly influenced by the measurement location used to interface with the brain. To that end, the context of this work is non-invasive, read-only measurements of the prefrontal cortex (PFC) at the approximate regions of Brodmann 9 and 10 (see Figure 1.1)[1].

1.4 Measurement Devices

Similarly, the scope of any implicit interface is likewise limited by the nature of the measurement used to read information from the human. For any physiological measurement device, there exists a stark trade-off between the resolution of the information provided by the device and the expense, portability, and applicability of the device in more ecological settings; although functional magnetic resonance imaging (fMRI) has exceptional spatial resolution and can measure activity across the

[1]The rationale for the contextualization to these regions is largely the hardware constraints we have at the HCI lab.

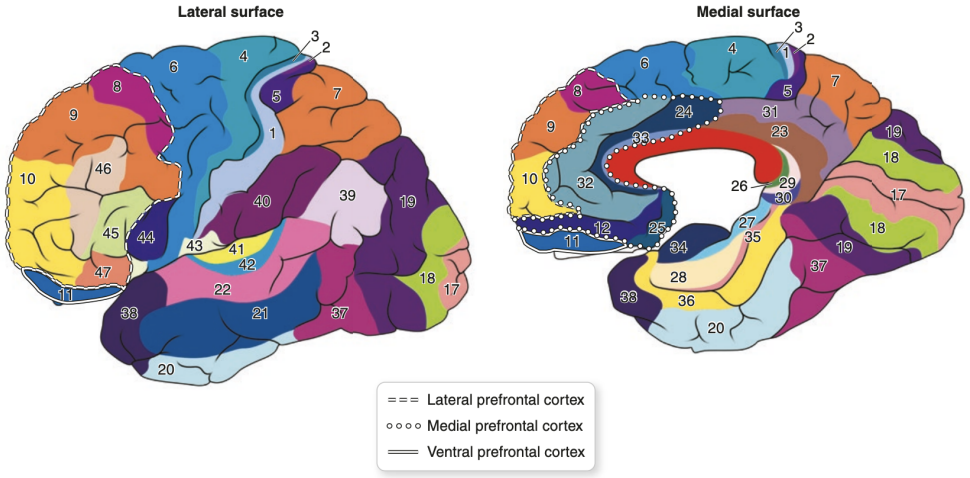


Figure 1.1: Lateral and medial views of the human brain, separated by Brodmann regions. The approximate regions of interest in our work are Brodmann areas 9 and 10 [1]. Adapted from **Neurobiological Sciences: Neuroanatomy, Neurophysiology, and Neurochemistry**, by A. J. Steiner, L. Aguilar-Hernandez, R. Abdelsalam et al., 2023, Springer Nature [2]. Reproduced with permission from Springer Nature.

entire brain, its prohibitive operational complexity and cost make it a wholly impractical choice for research studies intending to mimic more realistic human-computer interaction settings. To that end, my work leverages two such systems: functional Near-Infrared Spectroscopy (fNIRS) and a consumer-grade Electroencephalography (EEG), as they are both are usable in settings with participants sitting at a computer performing ordinary tasks. Whereas the fNIRS device used has relatively bulky form factor, it is reasonable to project that eventual development of the technology will enable comfort similar to wearing a baseball-cap; the consumer-grade Muse 2 EEG device that I use is extremely light-weight and functional, and can be worn without discomfort for extended periods.

1.4.1 Non-Neural Measurements

Measurements outside of the brain are also used in some cases: in one study in this work I use the Empatica E4 device - effectively similar in form factor to a wristwatch - for measurements of heart rate (HR), heart rate variability (HRV), and electrodermal activity (EDA); although these measurements cannot be contextualized within the scope of implicit BCI, they certainly could be leveraged to supplement implicit BCI or to develop implicit interaction on their own. Lastly, although self-reported measurements are not typically leveraged for real-time interfaces, the information that

they provide can be valuable towards contextualizing 'ground-truth' aspects of human experience during tasks: for this I most commonly use the NASA Task Load Index (NASA-TLX), a workload questionnaire common in HCI settings. Additionally, for one study, I use self-reports within the circumplex model of affect.

1.5 Beyond Workload in the Prefrontal Cortex

Initial fNIRS/EEG applications based on measurements made of the prefrontal cortex have thus far largely leveraged machine learning to develop inferences of cognitive load [12]. This dissertation highlights multiple aspects of prefrontal function which combine to compose the foundation undergirding the frontier of applied BCIs which can leverage PFC activity in ways that can stretch beyond workload. Although I explore in detail the bases for the ideas discussed here in Chapter 2, and while this is not a complete overview of the studies within this dissertation, a brief overview here provides the broad context of neurocognitive aspects by which I stretch "beyond workload".

1.5.1 Episodic Memory

Notable activation in the prefrontal cortex has been observed beyond just working memory: episodic memory has also been reliably correlated with prefrontal cortex activation [13]. And, although some research has been done on episodic memory with fNIRS in the PFC, it is largely underexplored within the context of BCI. To that end, the first part of this dissertation studies PFC activity with fNIRS during LLM use across a set of tasks designed to interface in differing ways with the PFC, spanning from working memory to episodic memory. This work develops insights into the neurocognitive states which interrelate with successful applications of current LLMs, and provides context and insight into the development of future interfaces based on episodic and working memory.

1.5.2 Leveraging Brain Networks for BCI: Application within Hincks' Paradigm

Hincks details that, while the relation of prefrontal activation to cognitive load is certainly not an invalid assumption, it is perhaps incomplete. His outline of the anti-correlated Default Mode Network (DMN) and Task-Positive Network (TPN) as identified in fMRI as the basis for implicit BCI [12] provides a new approach towards classification of neural activity which allows us to expand

beyond traditional applications: whereas the TPN has been demonstrated to activate during task activities such as working memory [14], DMN activity has been related to resting state activity, social cognition, mind wandering, and self-referential processes such as thinking about the past or future [15]; these networks are further shown to be anti-correlated: as one increases in activity, the other decreases. This dissertation takes steps to concretely explore the premise posed by Hincks that next-generation BCIs can leverage the complex interaction between these brain networks with a real-time implicit BCI, and helps to expand our conceptualization of what is possible towards this broader vision of high-fidelity information exchange between the human and machine.

1.5.3 Complex State Classification Using Low-Cost EEG Devices

Although the excellent spatial resolution of fNIRS represents a net positive in terms of the measurement of certain brain regions, the lack of precise spatial resolution from EEG may actually be useful in some contexts; when applied to the EEG signal, machine learning methods may be able to extrapolate patterns of information spanning multiple brain regions, potentially enabling the development of interfaces leveraging more complex human state information. However, while it might be possible to develop strong classification from extremely expensive and cumbersome EEG setups, what is actually possible now with ergonomic devices? To explore these ideas, I use the Muse 2, a low-cost, low-sensor count EEG device. My specific contributions in this domain are threefold: firstly, I explore the relationship between neural states and chess move quality. Secondly, I provide benchmarks of within-task workload classifications during standard cognitive psychology tasks (N-Back, Stroop, Mental Rotation), as well as in an ecological chess puzzles task. These results allow us to understand the capability of such low-cost devices to be leveraged in applied settings towards mental workload classification within multiple different contexts, and importantly, stretches from the standard tasks used in cognitive psychology into the ecological and applied realm with the chess puzzles task. And thirdly, between the tasks mentioned above, I consider classification *across tasks*. In this context, we ignore the workload levels, and instead consider our ability to differentiate between brain states as a function of the tasks themselves. This work establishes the capability of low-cost EEG sensors to reliably differentiate among a variety of complex cognitive states, and can enable the next generation of interfaces intended to move into classification geared towards differentiating more nuanced and complex facets of human state.

1.6 Outline

This work is composed of three primary projects which each push the state-of-the-art of noninvasive prefrontal-based BCIs in unique ways. Specifically, these works attempt to both unpack and extend the notion of brain-based classification vis-à-vis activation of the PFC, both through statistical methods to understand physiological effects, as well as through machine learning to understand operational contexts for future applications. Lastly, based on ideas learned throughout the completion of this work, the idea of Human-Sensor-Computer Interaction is introduced. I continue here with a brief outline of the chapters which follow.

Part I: Research Context and Technical Details

Chapter 2 covers background, including measurement tools used in this work: EEG, fNIRS, Empatica E4, and NASA-TLX, and the related research which contextualizes this dissertation.
Chapter 3 describes detailed data processing methodologies for the sensors mentioned above, as well as common statistical methods and processes used in this dissertation for data analysis.

Part II: lPFC and LLMs with fNIRS

Part II describes work which evaluates the physiological and self-reported responses of human users both with and without assistance from Large Language Models (LLMs). The context of this work within HCI as applies to LLM tools is first explained in Chapter II.
Chapter 5 presents an analysis of participant responses both with and without a LLM tool (Copilot for MS Word) during a series of tasks intended to target cognitive states which were intended to interact with the PFC in different ways. Tasks range from intending to induce cognitive load (e.g. SAT reading comprehension) to others intending to induce episodic memory (e.g. personal reflection). In this study, we measure participant lPFC brain data with fNIRS, emotional response with Empatica E4, self-reported workload with NASA-TLX, self-reported enjoyment, and objective quality of output. Our work develops useful insights into the benefits and drawbacks of LLM use which can be leveraged for real-time BCI, as well as contextualizing beneficial situations for LLM use within the context of brain function.
Chapter 6 reviews a study exploring the effects of LLM use on longer complex decision-making

tasks with fNIRS. This study sheds insights into the effects of prior LLM experience on patterns of prefrontal activity during LLM use, while demonstrating limitations in the context of applied PFC-based BCIs with fNIRS.

Part III: Real-Time lPFC-mPFC Based BCI with fNIRS

This part reviews the development of a real-time implicit fNIRS-based BCI which attempts to approximate prefrontal activation within the context of the anti-correlated paradigm of BCI proposed by Hincks [12]. Specifically, we expand from the notion of prefrontal-based fNIRS tasks requiring a gradient of high vs. low workload to instead explore the relative contextualization of anti-correlated TPN/DMN networks within the context of a mental workload task vs. a creative visualization task. Within this neurological context, we develop an application which prototypes the idea of a memory prosthesis introduced by Rhodes [16] and Lieberman [17], where from moment to moment the interface presents information appropriate to a user's state - in this case, with passively measured brain state as the storage and retrieval tag.

Part IV: Inferring Multidimensional Neural State Information from the PFC with Low-Cost EEG

Here, I cover a project which expands our understanding of complex human-state classification in the prefrontal cortex for BCI through measurements garnered through the Muse 2, a consumer-grade EEG device.

Chapter 9 reviews a user study exploring the relationship between EEG data and quality of chess moves.

Chapter 10 Explores what it means to understand workload as measured in the prefrontal cortex with low-cost EEG within the context of multiple standard tasks from cognitive psychology: Stroop (cognitive inhibition), Mental Rotation (spatial reasoning), N-Back (working memory), as well as within a Chess puzzles task, which requires a rich combination of mental activities. I then take this exploration further by demonstrating robust classification across-task states, which implies that rich neural features associated with complex mental processes can be distinguished by low-cost EEG systems towards future real-time BCI applications.

Part V: Conclusion

Chapter 11 Key findings of the projects in work are summarized, and ideas for future work are abstracted from the findings.

Chapter 12 Human-Sensor-Computer Interaction (HSCI) is introduced. Relevant background and implications for future work are discussed.

Chapter 13 Limitations of this work are covered.

Chapter 14 I conclude this dissertation.

1.7 Contributions of this Work

This body of work deepens our understanding of more nuanced human state classification via the PFC in applied tasks both in fNIRS and EEG devices; to that end, the primary contribution of the work is a breadth-first expansion of our understanding of what is possible "beyond workload" in the PFC not just in a purely abstract neuroscientific sense, but in terms of the equipment used in our lab and applied tasks similar to what ordinary people would experience in at-home use of such devices. Multiple promising avenues of further study are highlighted, including episodic memory, neural activation in the context of LLM-use, interfaces based on brain networks, and complex human-state classification using frequency domain features extracted from low-cost EEG. This work both explores the data both using LMMs to understand macro patterns, as well as with machine learning to determine the applicability of these data in real-world interfaces. Lastly, consideration of the nuances of interpreting human state from a physiological signal are explored within a framework of HSCI. Taken together, this work helps to bridge the gap from the confines of neuroscience into the next generation of real-world applications that leverage noninvasive neural sensors observing information from the PFC, and helps to pave the road towards adaptive, context-aware systems that can extend the frontier of human-machine interaction.

Chapter 2

Background

2.1 The Prefrontal Cortex

Detailed understanding of the full complexity of the activation patterns within the PFC as they relate to specific elements of a broader understanding of human state are not fully known [18]. However, research in this area has been completed done with a bevy of neuroimaging tools, including fMRI [19, 20], fNIRS [21], EEG [22], and Positron Image Topography (PET) [13]; substantial work has resulted in findings that contextualize and ground our ability to develop interfaces that interact with the complex neurological processes at play. To that end, the most notable and conclusive activation patterns in the PFC have been the associated with **working memory** and **episodic memory** [13]. Working memory is the cognitive system that enables short-term storage and recall of information during complex cognitive activities [23], and episodic memory refers to the process of conscious recollection of personal past experiences, and is characterized by a distinct sense of reliving events from one's own history [24].

In addition to these broad characterizations of the prefrontal region, network-based activation patterns have been well documented within the prefrontal area which comprise the opposing endpoints of Hincks' anti-correlated network BCI framework; that is, distinct patterns have been shown to arise within the medial (mPFC) and lateral (lPFC) aspects of the PFC[1]), which relate

[1]Although subdivisions related to ventromedial, dorsomedial, ventrolateral, and dorsolateral areas of the PFC are well-established in the neuroanatomical literature, the spatial resolution limitations of EEG and fNIRS make such fine-grained distinctions less relevant for the purposes of this work; we will therefore constrain our discussion to the lateral and medial regions of the PFC.

to activation of the DMN and TPN, respectively. These regions are visualized in Figure 2.1 and discussed in detail below.

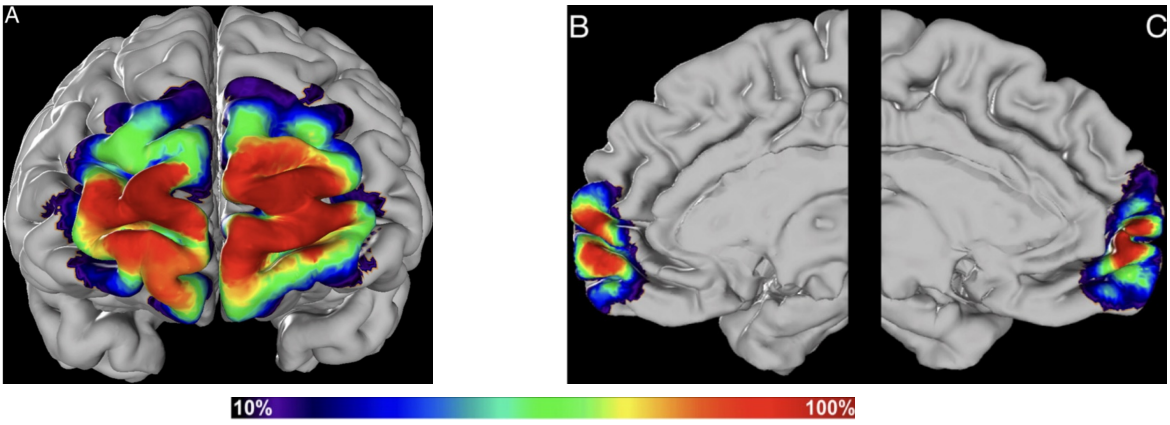


Figure 2.1: Visualization of the the prefrontal cortex on the MNI reference brain [3]. Color represents a continuous probability mapping of activation regions across 10 human brains, where deepest red indicates fully shared regions of activation across all brains at the given voxel. A) shows the frontal aspect, and B/C show medial aspects. The lateral aspects of the frontal area have been shown to activate based on workload tasks (e.g. n-back) and episodic memory, whereas the medial aspect has been shown to relate to emotion and social cognition tasks [4]. Adapted from **Cytoarchitecture, probability maps and functions of the human frontal pole**, by S. Bludau, S. B. Eickhoff, H. Mohlberg, S. Caspers, A. R. Laird, P. T. Fox, A. Schleicher, K. Zilles, and K. Amunts, 2014, **NeuroImage, 93**, Elsevier [4] with permission from Elsevier.

2.1.1 lPFC

Activation in the lPFC has been associated with the TPN through a wide variety of cognitive tasks associated with including problem-solving, planning, reasoning, working memory, and cognitive flexibility for creative processing and thinking [1, 4, 18, 25, 26, 27, 28]. This substantial association allows for the use of prefrontal cortex activation as a measurement of user mental workload when completing a variety of tasks. lPFC activation has also been discovered during episodic memory tasks [4, 29], although this contextualization is less-often explored in BCI.

2.1.2 mPFC

In contrast to lPFC, activation in the mPFC has been measured to indicate processing related to DMN activity. Eickhoff, in a comprehensive work detailing the functional segregation of the dorsomedial prefrontal cortex [30], provides examples of relevant patterns of activation, which include social processing [31, 32], processing of uncertainty [33], social cognitive processes related to theory

of mind [34, 35], moral reasoning [36], "nonsocial" semantic processing [37], and autobiographical memory retrieval [35, 38]. Significantly, there is distinct commonality in both the mPFC and lPFC in terms of their activation during episodic memory, however, the DMN (and mPFC) has been measured more often while participants are engaged in self-reflection, mentalization (focusing attention to one's own or others' emotional or mental states), or tasks designed to elicit emotional responses [29].

2.2 fNIRS

fNIRS uses diffuse optical imaging of near-infrared light to non-invasively measure changes in oxygenated Δ[HbO] and deoxygenated Δ[HbR] hemoglobin concentrations in the human brain [39]. Given that fNIRS only requires the user to wear a noninvasive headband, it promises to become a common technology used by researchers for task-related measurement of human subjects [39]. fNIRS functions by sending light pulses through the forehead; detectors capture the amount of light which returns, and these raw values are transformed into concentrations of oxygenated and deoxygenated hemoglobin using the Modified Beer-Lambert Law [40]. Although the depth of fNIRS is limited to the outer regions of the cortex, relatively shallow in comparison with fMRI, the precision of the localization of the measurement is approximately 2-3cm [41], much higher than that of EEG [42].

2.2.1 Removing Extracerebral Noise from fNIRS

Given that the light from the fNIRS devices passes through the skull, a significant amount of information in the resulting signal is from non-neural tissue. Whereas bandpass filters are commonly used to remove some high frequency information such as heartbeat (1 Hz – 1.5 Hz) and respiration (0.2 Hz – 0.5 Hz), and other low-frequency information related to blood pressure fluctuations called Mayer waves (0.1 Hz), there is still a large component of the signal which represents extracerebral hemodynamic activity, namely scalp hemodynamics [43]. Both fNIRS projects in this research leverage distinct approaches towards the removal of these signals: in the first, we leverage the dual-slope phase model pioneered by Blaney, et al. [44]. In the second, we use short-source detector pairs, in which light from sources near the detectors is filtered from the pairs further away from the

detectors by means of a Recursive Least Squares (RLS) adaptive filter [43][2]. These methods are discussed in more detail in the sections associated with those projects.

2.2.2 Neurovascular Coupling and the fMRI BOLD Response

An important contextualizing factor in fNIRS research is the question of interpreting neural activation from the signal itself; that is, how does one infer a relationship between Δ[HbO] and Δ[HbR] and neural activity? The most common basis for the inference of neural activation from hemodynamic activity is known as neurovascular coupling: that metabolic demand associated with neural activity leads to measurable changes in the cerebral vasculature [45]. In fMRI, magnets are used to measure this response based on the paramagnetic properties of [HbR] [45]; that is, biomagnetic fluctuations induced by changing concentrations of [HbR] due to neural activation result in a darkening of fMRI image voxels [46]. Measurement of changes in this way is known as the fMRI blood-oxygen-level-dependent (BOLD) response [46]. Figure 2.2 shows the BOLD response as measured in research investigation using both fMRI and fNIRS together [47]: the BOLD signal indicates a peak at the moment of increased [HbO] concentration and combined with decreased [HbR] concentration compared to baseline 5-10 seconds after neural activation, followed by an slight decrease from baseline 15-40 seconds after neural activation, after which the signal returns to baseline.

2.2.3 Calculating Cerebral Oxygenation with fNIRS: Δ[HbD]

Whereas the BOLD signal detected from fMRI is developed from magnetic perturbations in the vasculature, fNIRS measures changes in hemoglobin concentrations based on the scattering and absorption of light; although this enables fNIRS to to analyze both Δ[HbO] and Δ[HbR], it does not provide the same exact information captured by the fMRI BOLD response. To that end, the optimal method of inferring neural activation from the metrics of Δ[HbO] and Δ[HbR] is still an open question [48]. Following Kreplin, for statistical analyses of the fNIRS signal, I focus on a measure of cerebral oxygenation defined as Δ[HbO] - Δ[HbR] [48], what I will refer to as Hemoglobin Difference (Δ[HbD]). This signal effectively captures the pattern of simultaneous Δ[HbO] increase and Δ[HbO] decrease which occurs in concurrence with the BOLD response as measured by fMRI as

[2]These projects are actually presented out-of-order from when they were completed; the dual-slope phase method is a newer processing paradigm.

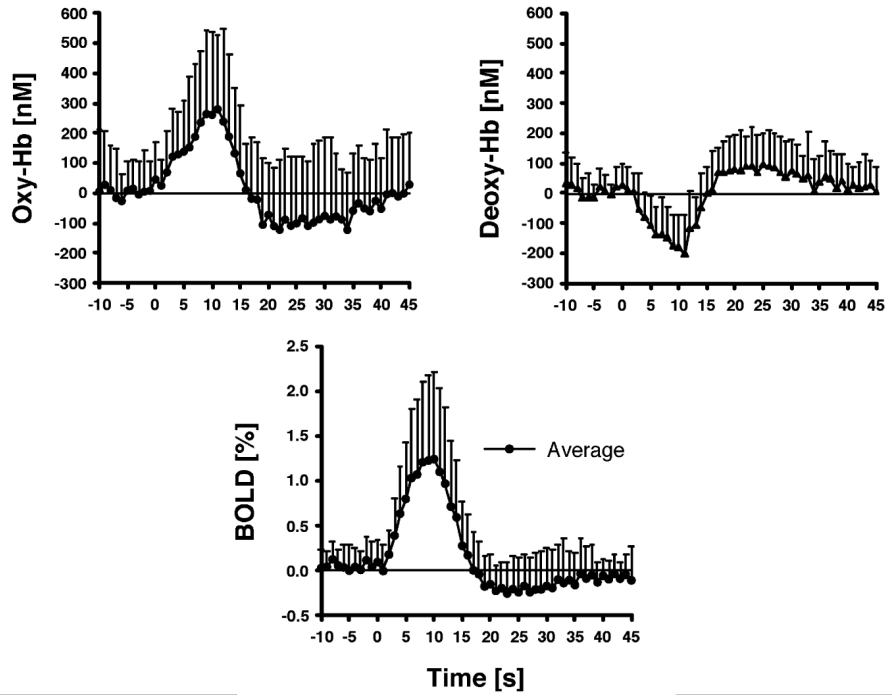


Figure 2.2: The BOLD signal. Upper left image depicts changes in Oxygenated Hemoglobin, upper right depicts changes in Deoxygenated Hemoglobin, and lower image depicts BOLD response. Adapted from **Investigating the post-stimulus undershoot of the BOLD signal—a simultaneous fMRI and fNIRS study**, by M. L. Schroeter, T. Kupka, T. Mildner, K. Uludağ, and D. Y. von Cramon, 2006, **NeuroImage, 30**(2), Elsevier. Reproduced with permission from Elsevier.

a consequence of neural activation, while also controling for changes in overall blood volume [48, 49].

2.2.4 Very Low Frequency Oscillations (VLF) in the fNIRS Signal

Research in fMRI and fNIRS has highlighted the accessibility and value of observing Low Frequency (LF) [0.07 Hz - 0.2 Hz] and Very Low Frequency (VLF) [0.02 Hz - 0.07 Hz] oscillations as correlates of cerebral hemodynamics [40, 50]. Notably, a *decrease* in the VLF band has been shown to correspond with task-based cortical activation [50]. While a comprehensive understanding of the physiological mechanisms underlying this VLF correspondence to neural activity remains incomplete [40], the prevailing theoretical basis for the physiological underpinning of these oscillations relates to cerebral autoregulation (CA), which is the body's ability to maintain consistent blood supply to the brain [51]. In terms of application of the methodology, task-based cortical activation in the prefrontal cortex has been successfully detected with fNIRS using this method [21], however the association with VLFO activity and neural activity is somewhat underexplored. To provide additional empirical

support for this relationship, and to help clarify for the reader the implications of this association, I conducted an analysis using the 68-participant Tufts Mental Workload dataset [52]. The key findings of this analysis are demonstrated in Figure 2.3; the full details of the analysis are written in Appendix A, and will be presented as an extended abstract at NAT 2025. The studies in parts II and III of this dissertation explore prefrontal hemodynamics with fNIRS using this metric relating a relative decrease of total power in the VLFO band of Δ[HbD] to increased neural activation.

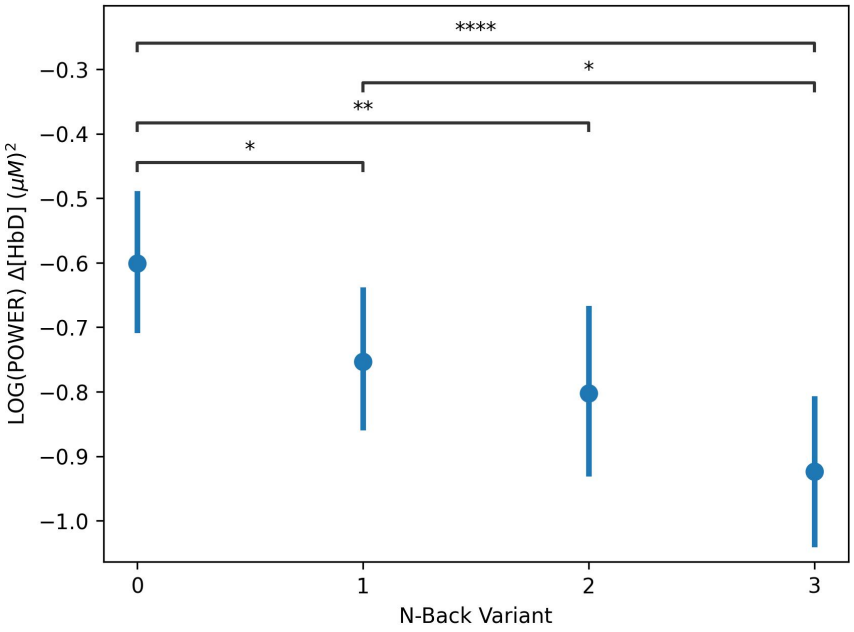


Figure 2.3: Log total power of Δ[HbD] in the VLFO band in the left lateral PFC as a function of N-Back level. The increase of N-Back level, corresponding with an increase in left lPFC function (and mental workload), results in a significant decrease of total power in the VLFO band. Note that, although many pairs of N-Back tasks are differentiable in this way, not all are significant (e.g. 1 vs. 2, 2 vs. 3)[3].

2.2.5 fNIRS in Implicit BCI Using the PFC

A considerable number of PFC-based implicit BCI studies using fNIRS have been done which use PFC activity to approximate mental workload. Some have been real-time tasks [53, 54, 55, 56, 57, 58, 59], while others are offline studies attempting to distinguish brain states [60, 61, 62]. Most studies infer mental workload by first training a model based on an N-Back task which later is used to modulate the difficulty of a separate task in real time [54, 55, 56, 61, 63, 64]. Workload-based interfaces have

[3]This finding is for Intensity data, and more precisely Dual-Slope Intensity (DSI). See Appendix A for more detail.

been used in the context of difficulty identification in games [65], interface modulation to optimize user state [54,61], interruption timing [55,63], adaptive learning [64], and multitasking identification to adapt target selection [56]. PFC-based fNIRS has been employed in a variety of other contexts as well: usability testing [57], drink preference in consumer product selection [66], self-reported positive vs. negative valence in listening to emotional music [67], and tracking operator fatigue and cognitive overload in aviation and air traffic control [62].

2.2.6 fNIRS and Episodic Memory in the lPFC

fNIRS has been used in multiple studies in measuring activation in the prefrontal cortex to measure epsodic memory in healthy human subjects[4]: the most common finding has been an increase in prefrontal activity during episodic memory tasks. This has been observed in both the left and right lPFC [68,69,70,71], but has been highlighted in the right lPFC [13]. In addition, encoding processes related to episodic memory have been studied in the prefrontal cortex using fNIRS, although the results are less consistent in that encoding has been shown to in some studies to increase activity in both left and right lateral lPFC [71,72,73], whereas others have shown decreases in lPFC activation during encoding [69]; music has been shown to decrease lPFC activation relative to silence during encoding processes [71]. Machine learning has also been used to distinguish between the processes of encoding and decoding of episodic memory [74]. BCIs which leverage fNIRS in the domain of episodic memory, however, are not yet available[5].

2.2.7 fNIRS and the DMN localized in the mPFC

In contrast to the many studies observing episodic memory with fNIRS and DMN-mPFC relationship as measured by fMRI, few fNIRS studies have been done observing mPFC (or indeed, PFC) activation in relation to the DMN on healthy human subjects. That said, activations in the DMN as measured by the mPFC have been measured with fNIRS specifically in the contexts of mind wandering [76]. No studies could be found which attempt to relate fNIRS and the DMN as measured by the mPFC to a BCI, however as mentioned elsewhere it has been proposed by Hincks [77].

[4]While many studies studies on episodic memory focus on Alzheimer's disease and elderly populations, this brief review specifically examines findings from healthy human subjects, as they represent the primary population of interest for our research.

[5]I note that, although this has not yet been studied in the fNIRS literature, some BCI applications leveraging episodic memory based on PFC activity have been done with EEG [75]

2.3 EEG

By contrast to fNIRS, EEG measures the electrical activity of the neuronal activation directly. More precisely, the fundamental signal measured by EEG originates from postsynaptic potentials: localized electrical potential changes that occur when neurotransmitters bind to receptors on the postsynaptic membrane [78]. These potentials, when generated synchronously by large populations of vertically aligned pyramidal neurons, sum to create electrical fields of sufficient magnitude to be detected at the scalp by detectors in a process called volume conduction [79]. In contrast to the temporal constraint of fNIRS signal related to the hemodynamic response, the temporal resolution of these neural activations is extremely fast, on the order of milliseconds [80, 81]. However, due to the projection of the three-dimensional physical reality of neural sources onto a two-dimensional representation of the scalp surface, it is mathematically impossible to determine the precise localization of neural activity from scalp EEG measurements alone, a fundamental limitation known as the inverse problem [79]. More specifically, the low fidelity of EEG's spatial resolution stems from multiple factors: the orientation of synchronously firing neurons may not be parallel to the scalp, there may be fields generated in multiple directions, and resistive elements like the skull can cause signal dispersion through volume conduction [80]. Therefore, although EEG can detect measurements with far more temporal precision than fNIRS, its ability to localize the measured effects is not nearly as strong [82]. Given that the prefrontal region has such a strong relationship to mental workload, part of my EEG work is to benchmark low-cost EEG systems in that context. The more interesting EEG work in this dissertation, however, aims to directly infer complex human state information dependent on combinations of brain various networks which are implicitly combined in the EEG signal.

2.3.1 Interpreting the EEG Signal

Although the raw signal of EEG relates to the firing of grouped neurons, there are different ways to interpret the activity of those neurons in terms of application. This section briefly describes some of the most common forms of EEG-based signal use for BCIs.

Event-Related Potentials

One primary form of neural interpretation is event-related potentials (ERPs), where many (often hundreds) of rapid (<1s) time-locked trials are collected per-participant, and averaged at the participant level, and then across participants, to produce a single waveform approximating the neural response of interest [83, 84]. In addition to neuroscience research, ERPs have likewise been used for explicit BCIs. One implementation is a P300 speller, which rapidly flashes a sequence of random letters to the user: a pronounced and detectable ERP response is elicited when the user sees the letter of interest [85].

Steady State Visual Evoked Potentials

Another technique is steady-state visual evoked potentials (SSVEP), which are interfaces where users are presented with stimuli flashing at different frequencies; the users' attention to a given stimulus produces detectable signatures in the EEG data of the same frequency, suitable for instance to move a mouse cursor on a computer screen [85, 86].

Frequency Domain Features

The most common method of interpreting neural activity from EEG for BCI is the rhythmic activity across multiple frequency bands. The most common frequency bands (δ: 1-4 Hz, θ: 4-8 Hz, α: 8-12 Hz, β: 13-30 Hz, and γ: 30-150 Hz) have relationships to distinct biological processes at the cellular level [87, 88], and have been shown to roughly correspond to distinct psychological processes. And, although these bands certainly do not define rigid boundaries through which all phenomena in the brain are expressed [80], transformation into the frequency domain provides useful measurements by which we can develop BCIs. The most common prefrontal EEG-based studies using frequency-domain activity have focused on cognitive load [22, 89], but the frequency components of the EEG signal have been used to help understand and distinguish between concentration vs. rest in a Sudoku task [90], emotional states [91], motor imagery detection [92], cognitive performance [93], and drowsiness [94]. Given the broad use of prefrontal-based EEG for BCI, I chose frequency-domain features as the metrics of interest for the EEG work in this dissertation.

2.3.2 Noise Removal for EEG

Similar to fNIRS, the raw EEG signal is likewise susceptible to extracerebral artifacts. Common sources of contamination include interference from the power source, environment, eye blinks, heart rate, and muscle movements [95]. A similar situation exists in EEG as in fNIRS in that, although bandpass filters and notch filters are commonly used to remove some sources of noise, other treatments are commonly applied to further enhance the data [96]: Independent Component Analysis (ICA) used with eye movement channels (EOG) to remove blink noise, statistical thresholding, and wavelet decomposition are three commonly used methods to remove extracerebral noise [95]. However, debate exists regarding the trade-offs of extensive filtering methods, with some recent research indicating that the benefits gained from the removal of artifacts and noise that could otherwise lead to spurious results or mask genuine effects may be outweighed by the loss of statistical power to detect effects after reduction of data incurred during noise removal [97]. Different approaches are made to signal processing in different aspects of our EEG-related work: relevant details are discussed in these sections.

2.3.3 Low-Cost, Low-Sensor EEG for BCI

In recent years, advancements in consumer-grade, low-cost EEG companies—such as InteraXon, OpenBCI, NeuroSky, and Emotiv—have positioned them as viable alternatives to traditional, expensive EEG systems [98]. These devices offer an accessible means of monitoring brain activity, enabling applications in dynamic, real-world settings. Their portability, affordability, and ease of use have made them practical tools for investigating mental workload and cognitive state classification, with demonstrated applications in stress detection [98], drowsiness detection [99], emotion classification [100], and adaptive human-computer interaction [101]. Despite initial promising research, these systems often face limitations in spatial resolution and signal fidelity when compared to research-grade devices [102]. And, while some of these systems support data collection over large areas of the brain (OpenBCI and Emotiv), others are single-probe (NeuroSky) or four-probe (Muse 2) systems. Given our choice of the Muse 2 device, we contextualize our focus on human-state classification to devices with lower probe counts, rather than whole-head systems, which, despite lower fidelity of information, represent systems which are more affordable and easier to set up

for home consumers [103]; and, although a full detailing of the research done with the Muse 2 device follows below, we briefly review the work of NeuroSky and others here. With the NeuroSky, preliminary work has been done in terms of basic state-differentiation between analytical reasoning and focused attention states, however classification has proven difficult [104]; other work has been done in the differentiation of reading and resting states [105]; preliminary work has also been done with this device in manipulating a prosthetic arm [106]. Most related to the EEG work discussed in this dissertation, however, is an evaluation of a non-consumer, single-probe frontal EEG system, in which the authors demonstrated its effectiveness in distinguishing tasks such as arithmetic operations, finger tapping, mental rotation, and lexical decisions. Notably, they observed an increase in θ band power during these tasks and successfully applied support vector machines (SVMs) to frequency-transformed 2.5-second windows, achieving classification accuracies exceeding 70% for arithmetic operations, finger tapping, and lexical decision [107]; this study provides a first-step towards our understanding of the possibilities and limitations of neural state classification with low-cost, low-sensor count EEG devices, and our work directly contributes a next-step towards the application of such tools for next-generation BCI.

2.3.4 Muse 2 in Research

The EEG system leveraged in this research is the Muse 2 headband. Initial research in multiple contexts has found that this low-cost, portable, and wireless EEG device is a suitable alternative to traditional more expensive and cumbersome devices; a visual oddball paradigm experiment and a reward-learning task experiment were conducted wherein researchers were able to observe and quantify the N200 and P300 ERP components, as well as frequency domain features in both tasks towards the identification of cognitive fatigue [108]. An investigation of baseball players during batting practice demonstrated decreased beta wave activation correlated to increased performance [109]. A longitudinal study of mindfulness treatment in sufferers of Obsessive Compulsive Disorder (OCD) demonstrated band power increases in α and β and decreases in δ and θ during mindfulness practice [110]. Preliminary studies have also been done with the Muse 2 which demonstrate the potential of low-cost EEG systems to classify mental states with machine learning. One study collected frequency domain features from five participants across three mental states (relaxing, neutral, and concentrating). Using feature selection techniques and machine learning classifiers

including Bayesian Networks, Support Vector Machines (SVM) and Random Forests, they achieved over 87% accuracy in state classification [111]. Another study investigated the development of a BCI system to enable drone control through consumer-grade EEG headsets: they focused on adapting drone thrust levels based on concentration, and likewise leveraged SVM and achieved 70% accuracy, however they also found wide variability among participants' individual scores [112]. Another study tested the reliability of the Muse 2 device against the TOBII Pro Nano eye tracking system. Although their work suggests that the eye tracking system performs better for classification was of N-Back levels 1 vs. 2, the relatively high cost of the TOBII device, puts it out of league with the Muse 2 in terms of current potential for consumer application [113].

2.4 Other Measurement Tools

In addition to measurements of the brain, I also employ other measurements to provide context into the internal state of the user. They are commonly used in HCI research, and are briefly described here.

2.4.1 Empatica E4

The Empatica E4 is a noninvasive, wristwatch-like wearable device that measures two physiological signals: Photoplethysmography (PPG) and Electrodermal Activity (EDA). PPG measures infrared light reflection from the skin [114], which enables the calculation of Heart Rate (HR), Inter-Beat-Interval (IBI), and Heart Rate Variability (HRV) [115, 116]. HRV, which tracks the timing variation between heartbeats, has been shown as a reliable indicator of autonomic nervous system activity and stress response [117]. EDA measures changes in skin conductance which are caused by activity in sweat glands; this measurement can be used to infer emotional intensity, stress levels, and attention [118]. Research has demonstrated that the E4's HR measurements align with gold-standard methods [115], and the device has been successfully employed in numerous research studies examining affect [119, 120] and stress [121, 122], though its EDA and IBI measurements show some limitation during physical activity [116].

2.4.2 NASA-TLX

In addition to physiological measurements, for some studies in this dissertation we also leverage self-reported measurements by way of the NASA Task Load Index (NASA-TLX), a widely used, multidimensional self-report scale of mental and physical workload that consists of six subscales [123]. Three of the subscales relate to the demands imposed on the subject (mental, physical, and temporal demand), whereas the other three subscales focus on the user's interactions with the task (performance, effort, and frustration). While initially designed for aviation, it has since been shown to be effective in a variety of other fields; the combination of its dimensions is generally effective in representing the workload experienced by most people when performing a task [124].

Chapter 3

Materials and Methods

This section details the hardware and technical tools which are employed throughout this dissertation.

3.1 fNIRS Hardware and Probe Design

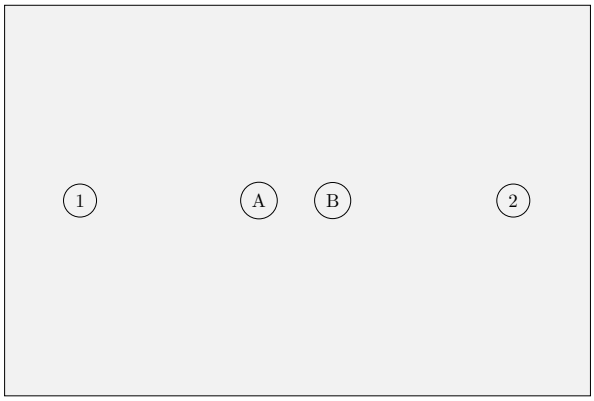
The specific fNIRS device employed throughout this research is a frequency domain near-infrared spectroscopy device produced by ISS Imagent, Champaign, IL USA, operating at a modulation frequency of 110MHz and with wavelengths of 690nm and 830nm. Light was delivered to custom probes via 400 µm diameter multi-mode fibers and collected by 5 mm diameter fiber bundles. These fibers were held in-place by a flexible plastic mesh and were encapsulated in black silicone. The work of the professor Fantini's lab in the biomedical engineering department at Tufts has enabled us to leverage the newest probe designs; therefore, two different fNIRS probes are used for the fNIRS portion of this research[1].

3.1.1 Probe Geometry Used in Part II Chapter 5

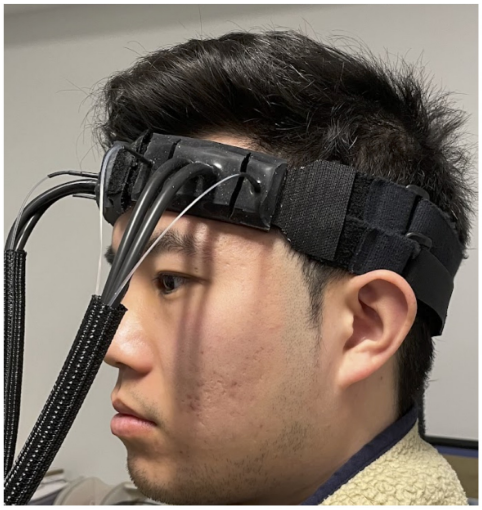
The fNIRS study in Part II Chapter 5 of this dissertation used two probes, both of which had optode geometry designed for the dual-slope (DS) method [44]: each probe consisted of two source positions, each with two wavelengths and two detectors. See a visualization of the optode geometry in Figure 3.1a, and an image of the two probe placement over the left and right prefrontal areas as

[1]The probe geometry used in Part II Chapter 6, which applies the dual slope-phase method, is newer than the probe used in Part II Study 5. However, the context and results of each of these projects lend themselves to presenting the results in 'reverse' order

in Figure 3.1b. Note that the DS method uses self-calibration and paired short and long probes to account for movement artifacts and other extracerebral noise - for details, see [44]. Although the probe in question has sources over the relatively medial and lateral areas, analyses of the DS method indicate the highest sensitivity over the central region between the probes [44], which in this study would be the relatively lateral aspect of the prefrontal area. After DS calculations, a single value for each of Dual-Slope Intensity (`DSI`) and Dual-Slope Phase (DSϕ) for each of Δ[HbO] and Δ[HbR] are produced for both the left and right probes.



(a) fNIRS probe geometry configured for the dual-slope method [44]. It has 2 source locations and 2 detector locations (A and B). Each source location contains two light sources, one at 830nm and the other 690nm. Long source-detector distances (1-B, 2-A) are 35mm from each detector, and short source-detector distances (1-A, 2-B) are 25mm.



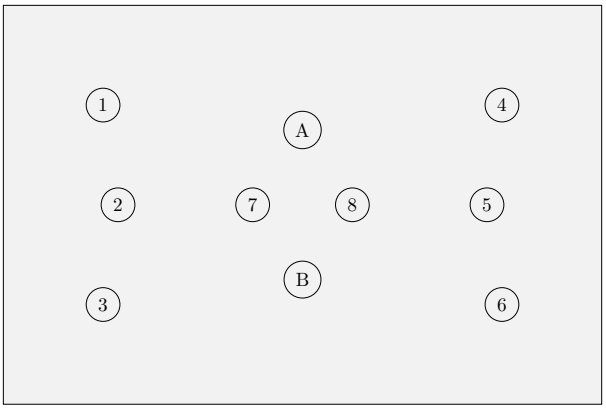
(b) User wearing a functional near-infrared spectroscopy (fNIRS) device[2].

Figure 3.1: fNIRS probe used for Part II Chapter 5. Left (a) is an image of the probe geometry, and right (b) is an image of a user wearing two probes, one over the left eyebrow, and the other over the right eyebrow (b).

3.1.2 Probe Geometry Used in Part II Chapter 6

The fNIRS study in Part II Chapter 6 of this dissertation has probe geometry as depicted in Figure 3.2a, and was positioned as illustrated in Figure 3.2b. This probe setup was not configured for the dual slope method. Although the bulk of the probe covers relatively lateral region of the PFC, the spatial resolution of fNIRS (2-3cm; [41]) suggests that measurements from the medial source locations (positions 4, 5, and 6 in Figure 3.2a) likely capture hemodynamic activity from the medial prefrontal cortex as well.

[2]In practice, the fiber optic cables are routed up and over the head, not blocking the eyes.



(a) fNIRS probe geometry used in the II Chapter 6 study of this dissertation. The probe has 8 source locations and 2 detector locations (A and B). Each source location contains 2 light sources, one at 830nm and the other 690nm. Source locations 7 and 8 are used only for short source-detector pair based adaptive filtering to remove extracerebral noise from the neural signal. Short sources are 1.5cm from each detector; for each detector, the nearest 4 source locations (outside of the short sources) are each 3cm from the detector, and the furthest 2 source locations are 3.61cm from the detector.

(b) Lab member wears the fNIRS headband used in Part II Chapter 6. Probe is placed over the left eyebrow.

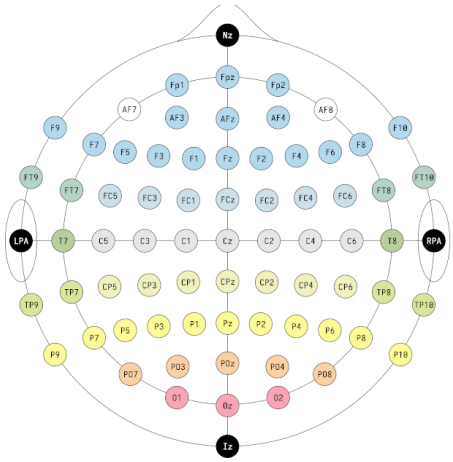
3.2 The Muse 2 Device

The Muse 2 device (see Figure 3.3) [125] is a consumer-grade EEG system capable of transmitting data over Bluetooth. It is light-weight, and comfortable to wear for extended periods. The device uses an online reference electrode to remove global noise from the signal, and then transmits data from the AF7, AF8, TP9, and TP10 electrodes. Due to the combination of the focus of this dissertation being on the PFC, as well as due to noise from the TP9 and TP10 channels, I focus my analyses and machine learning results only on the AF7 and AF8 probes. I use different programs to extract data from the Muse 2 for each of the two studies in this dissertation, MindMonitor [126], and MuseJS [127]; these tools are discussed in more details in their respective sections.

3.3 EEG Data Processing

Because the two projects that leverage the Muse 2 interface with the device and data differently, the specific data processing methods associated with these are discussed in their own Sections: 9.2.2 and 10.2.6.

(a) The Muse 2 device.



(b) EEG 10-10 montage (reproduced from [128]). The Muse 2 uses the central FpZ electrode as an online reference; the white AF7 and AF8 probes here are the locations of the Muse 2 used in this work.

Figure 3.3: Left, the Muse 2 device (a) and right, the 10-10 eeg montage with AF7 and AF8 positions in white.

3.4 Frequency Domain Transformations

Frequency domain transformations are commonly used in this research. Specifically, I use these transformations to analyze fNIRS, EEG, and Empatica E4 data. All frequency domain transformations in this dissertation leverage the Multitaper method. This method, originally created by David J. Thompson in 1982 [129], is a robust nonparametric method for spectral estimation that has been shown to provide a favorable balance between narrow-band bias, broad-band bias, and variance [130, 131, 132, 133]. Although the details of this methodology are outside the scope of this dissertation, in brief, the Multitaper method improves spectral estimation by averaging multiple spectral estimates, which are each created using a different orthogonal taper. These tapers, known as discrete prolate spheroidal sequences (DPSS), are designed to maximize energy concentration within a given frequency band. By averaging across the estimates of these orthogonal tapers, this approach reduces variance in the spectral estimate while minimizing spectral leakage compared to single taper methods. For more details, the reader is encouraged to review to Betash Babadi's excellent summary of the method [132]. For implementation, we leverage the mne library [134] in Python [135]. Following frequency-domain transformation, Simpson's rule is used to extract total power by integrating over the frequency bands of interest [136]. The resulting values are

log-transformed, which results in approximately normally distributed data suitable for Linear Mixed Model analyses (see 3.5 below for details).

3.5 Statistical Methods

The majority of our effects analysis relies on Linear Mixed-Effects Models (LMMs). Given that this approach may be less familiar to some readers than the traditional Analysis of Variance (ANOVA), we provide a brief overview of the methodology here. Readers already experienced with LMM analyses may skip to the next section.

3.5.1 Linear Model

Let us start by reviewing the Linear Model (LM). See equation 3.1 [137].

$$Y_i = \beta_0 + \sum_{k=1}^{n} \beta_k X_{ki} + \epsilon_i \tag{3.1}$$

In this model, Y_i represents the ith observation of the the dependent variable; β_0 is the intercept; there are n fixed effects, where each predictor X_k is associated with a coefficient β_k, and for each observation i, the corresponding predictor X_{ki} is multiplied by its respective coefficient B_k; the error term for observation i is denoted ϵ_i. The critical assumptions related to the Linear Model include: a linear relationship between independent and dependent variables, normality and independence of errors, and equal variances among residuals (homoscedasticity), [138]. Crucially, the assumption of independence of errors is violated whenever we have multiple measurements from a given individual, thus rendering this model insufficient when performing within-subjects analyses. And, although it is possible to use a straightforward within-subjects ANOVA on the data, this necessitates reducing each participant's set of multiple observations to a single mean value, resulting in a substantial loss of statistical power [139].

3.5.2 Linear Mixed Model

To solve this problem of the violation of independence of errors, let us consider the simplest implementation of a LMM [140]:

$$Y_{ij} = \beta_0 + \sum_{k=1}^{n} \beta_k X_{ki} + \alpha_j + \epsilon_{ij} \tag{3.2}$$

where $\alpha_j \sim \mathcal{N}(0, \sigma_\alpha^2)$ and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$ are independent. The LMM equation extends the original LM by adding a new term α_j, which represents a unique intercept for a grouping category j. For example, provided individual participant identification as a grouping factor, this extension explicitly models separate intercepts per participant, with the assumption of normally distributed intercepts across levels of j. LMMs further offer the flexibility to accommodate multiple random intercepts, multilevel structures which contain nested blocks within participants [141], random slopes[3], and can handle unbalanced data [143]. Therefore, the LMM, while elegantly addressing the independence violation of the basic linear model under repeated measures, also provides multiple benefits over the constraints of within-subjects ANOVAs.

3.5.3 Model Fitting and Validation

All statistical modeling in this dissertation is done in the `R` programming language [144]. Unless otherwise specified, models used are LMMs; these models are fit using the `lmerTest` package; factor effect significance was assessed using Type III F-tests. Kenward-Roger approximation is used to determine degrees of freedom [145]. Specific random effects structures for any formula were determined via likelihood ratio tests (LRTs), including significance as well as comparison of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values [146]. Necessary assumptions were validated for all fitted models using the `performance` package in `R` [147], including visual tests for linearity, homoscedasticity, and normality of residuals; for models containing multiple fixed effect factors, I also assessed multicollinearity by confirming all Variance Inflation Factor (VIF) values were below 10 [148].

3.5.4 Post-Hoc Tests

To determine the direction and magnitude of significant within-factor effects from the LMMs, I use Estimated Marginal Means (EMMs) [149]. EMMs provide adjusted means that account for the

[3]Although it is possible to model random slopes, the sample sizes of the datasets within this dissertation were not large enough to allow for convergence or appropriate theoretical interpretation [142]; given this constraint, this study focuses on random intercept models.

model's structure, which thereby offer more accurate representations of factor-level differences than raw means [150]. Post-hoc pairwise t-tests are then performed on these EMMs. Calculations of EMMs and pairwise comparisons are performed with the `emmeans` library [151] in `R`.

3.5.5 Correction for Multiple Comparisons

For all studies α is by default set to 0.05. Correction is made for omnibus tests where multiple LMMs are made for tests close in scope (e.g. one model per-EEG waveband, or two models for DSI and DSϕ in the same region, etc). The specific adjustment method used varies by study, but if Bonferroni correction [152] is used, the column **p** represents the uncorrected p-value, and the column **sig.** represents significance under the adjusted α as specified; if the Benjamini-Hochberg [153] procedure is used, then a new column **p.adj** represents adjusted p-values, and **p.adj.sig** represents significance under α=0.05 given the adjusted p-value. For all post-hoc pairwise comparisons, Tukey's Honest Significant Difference (HSD) correction is applied to control the Family-Wise Error Rate (FWER) [154].

3.5.6 Effect Sizes

Effect size calculations are valuable metrics which provide complementary information to the usual p-values; rather than simply determine the whether an observation is improbable under the null, effect size calculations quantify the magnitude of differences by expressing them as a proportion of variance explained [155]. Although effect sizes are difficult to calculate for LMMs given the complex variance partitioning between fixed and random effects, it is possible to produce them from the F and t test statistics given approximations of the degrees of freedom in the model [156]. In this dissertation I report partial Epsilon squared (ϵ_p^2), also known as adjusted partial eta squared (adj. η_p^2)[4], which quantifies the proportion of variance associated with a given effect while controlling for other variables in the model, and reduces the bias introduced by the usual η_p^2 calculation [157][5]. Specifically, I calculate ϵ_p^2 from F-statistics for main effects and interactions, and from t-statistics resulting from pairwise comparisons, as follows.

[4]Although the traditional means by which the ϵ_p^2 value is calculated is different from the adj. η_p^2 calculation shown here, their equivalence is proven in [157]; we use the term ϵ_p^2 simply for the convenience of the naming convention.
[5]Despite that it is less often reported than ω_p^2 it has been shown that ϵ_p^2 is less biased [158].

From F-statistics

$$\epsilon_p^2 = \frac{F \times df_{\text{effect}}}{F \times df_{\text{effect}} + df_{\text{error}}} \tag{3.3}$$

From t-statistics[6]:

$$\epsilon_p^2 = \frac{t^2}{t^2 + df_{\text{error}}} \tag{3.4}$$

Then,

$$\epsilon_p^2 = \epsilon_p^2 - (1 - \epsilon_p^2) \times \frac{df_{\text{effect}}}{df_{\text{error}}}. \tag{3.5}$$

[157]. To perform these analyses, I use the `effectsize` library [156] in `R`. For interpretation, I use Field's guidelines [156, 159]:

- $\epsilon_p^2 < 0.01$ — Very small
- $0.01 \leq \epsilon_p^2 < 0.06$ — Small
- $0.06 \leq \epsilon_p^2 < 0.14$ — Medium
- $\epsilon_p^2 \geq 0.14$ — Large

Effect size confidence intervals are based on test directionality. For ANOVA results (one-tailed F tests), I report 90% confidence intervals for effect sizes, aligning with $\alpha = 0.05$, as recommended by Steiger [160]. For post-hoc t-test results (two-tailed tests), I report 95% confidence intervals. To avoid confusion, I use the same ϵ_p^2 CI column header in both tables.

3.6 Machine Learning Methods

3.6.1 Machine Learning Models

A variety of machine learning models are used throughout the studies in this dissertation. As we will see, many of these models perform differently for different participants and in different scenarios; indeed a "best model" has certainly not been found for BCI. Following the most commonly used models for fNIRS and EEG classification from [161, 162] and [52], I use K-Nearest Neighbors (KNN) [161, 163], Linear Discriminant Analysis (LDA) [164], Quadratic Discriminant Analysis

[6]Note that this is equivalent to the formula for F-statistics given $t^2 = F$ and $df_{\text{effect}} = 1$.

(QDA) [165], Artificial Neural Networks (ANN) [166], Support Vector Machines (SVM) [167] and Random Forests (RF) [168]. I also add to this collection SVM with an applied Radial Basis Function (RBF) kernel (SVC) [169]. And, although deep learning approaches have demonstrated promising results for fNIRS-based [170] and EEG-based BCI [171, 172], I decided against using these methods due to the substantial time investment required for model preparation and training; although I use a variety of models and propose some insights towards the machine learning applications in this work, it is important to contextualize the work done herein as outside the scope of machine learning as a field: this work applies basic machine learning techniques to establish baselines for classification, rather than advanced techniques to improve dataset accuracy. However, I recognize the potential value of deep learning methods for future research, and urge other researchers to apply such techniques to the datasets associated with the studies done herein. All machine learning classification is done within the `scikit-learn` [173] library in Python.

3.6.2 Leave-One-Out Cross-Validation

The most rigorous approach to classification patterns for BCI studies is known as Leave-One-Out Cross-Validation [162]. This is an approach where participants are segmented individually as their own test set. It ensures that no overfitting is possible to the data, which is particularly challenging with BCI data, given high variability across participants [52]. This cross-validation methodology will be used for all except the final study this dissertation, which opts for a separate approach given very high variance in data samples per-participant.

3.6.3 Result Presentation: `STANDARD` and `OPTIMIZED`

Throughout the preparation of this work, it became quite clear to me that many participants respond differently to different classifiers. Therefore, for most studies I will report two forms of classification. In the first, what I will label `STANDARD` classification, metrics (see below) are averaged across participants for a given model. In the second, what I will call `OPTIMIZED` classification, I will select the best performing model per-participant and report the average across these. These different strategies are presented with different intentions: the `STANDARD` result indicates an upper-bound on the best possible application that one could deploy now, and the `OPTIMIZED` result indicates the possibility for future interface development based on the criteria at hand. In short, `OPTIMIZED`

indicates whether the given criteria are worth pursuing in the future.

3.6.4 Machine Learning Metrics for Classification

I report Macro Precision, Recall, and F1 scores for machine learning results. Formally, for a K-class problem, these macro-averaged metrics are defined as:

$$\text{Macro-Precision} = \frac{1}{K}\sum_{i=1}^{K}\frac{TP_i}{TP_i + FP_i} \tag{3.6}$$

$$\text{Macro-Recall} = \frac{1}{K}\sum_{i=1}^{K}\frac{TP_i}{TP_i + FN_i} \tag{3.7}$$

$$\text{Macro-F1} = \frac{1}{K}\sum_{i=1}^{K}\frac{2 \times P_i \times R_i}{P_i + R_i} \tag{3.8}$$

where TP_i, FP_i, and FN_i denote true positives, false positives, and false negatives for class i; by giving equal weight to each class, macro-averaging highlights model performance across all classes [174].

3.7 Data Visualization

All data visualizations in this dissertation, unless otherwise specified, were created in `Python` using the `seaborn` library [175]. Error bars represent 95% confidence intervals derived through a multilevel bootstrap to account for repeated measures [139, 175].

3.8 Institutional Review Board

All studies in this dissertation were individually approved by the Tufts University Institutional Review Board (IRB).

3.9 Moving Forward

Having established the theoretical foundations of PFC measurement and its role in cognitive processes, we now turn to an innovative application domain: exploring how the PFC responds

during human interaction with Large Language Models (LLMs). This initial research project helps to advance our understanding of PFC activity in complex and applied human-AI collaborative tasks, and also provides insight into how neural signals might be leveraged for adaptive interfaces in these emerging interaction contexts; the following work represents an initial exploration into the measurable effects of neural activation related to complex human states

PART II

Prefontal Cortex Activation During Interaction with LLMs

We now begin our exploration into the measurable effects of neural activation related to complex human states 'beyond workload' through one project with two studies in which users interface with LLMs. In the studies, completed over a year-long collaboration with Microsoft research, participants engage in a variety of tasks, both with and without the assistance of LLM-based tools. Throughout their work, we monitor changes in prefrontal activation with fNIRS. The two studies in this project represent initial exploratory analyses confirming measurable effects; they do not directly result in applied interfaces, but rather, provide usable insights into potential domains for future fNIRS-based BCI applications which leverage the PFC, and further provide valuable information related to the effects of LLM tools on human users, both in terms of neural states and through other metrics. I begin in Chapter 4 by discussing the studies within this project in more detail, and by providing background information on LLMs in HCI. The studies themselves follow in Chapters 5 and 6.

Chapter 4

Overview and Background on LLMs in HCI

4.1 Part II Chapter 5 Overview

In the first study of this project we explore the effect of using Copilot for Microsoft Word within the context of tasks designed along a *gradient of subjectivity*. This gradient of tasks was designed to target aspects of human state which we believed would differentially reflect the benefits and drawbacks of engaging with a LLM-tool: namely, the more subjective the task, the less beneficial the tool. This gradient of subjectivity also has its basis in neural function, in that we intended the subjective endpoint of the gradient to express different activation patterns in the prefrontal cortex than the objective endpoint of the gradient. We specifically relate this pattern in terms of episodic memory, and more broadly, to the DMN; thus, we explore the effects of LLM use on prefrontal activation through a variety of tasks intended to themselves target the prefrontal cortex in different ways. In addition to observing effects on the prefrontal cortex, in this study we also explored this context within multiple physiological and self-reported metrics, including information from HR, HRV, and EDA, objective performance measures, self-reported measures of NASA-TLX and enjoyment, and user comments.

4.2 Part II Chapter 6 Overview

In the second study of this project, we observe the changes in prefrontal activation during a longer (20m) complex decision-making task. This study is less focused on the effects of Copilot across tasks

targeting differential aspects of prefrontal function, but rather, more specifically on the effect of Copilot during a complex cognitive task intended to mimic a real-life working environment. That is, we designed this individual task to induce cognitive load, and this study explores the effects of Microsoft 365 Copilot on self-reported workload and prefrontal cortex activation. Findings are also considered in this study as a consequence of LLM-tool experience. This study suggests practical insights related to real-world applications of fNIRS-based BCI on human subjects while interfacing with LLM-tools. Preliminary findings from this study will be presented as an extended abstract at the Neuroadaptive Technology (NAT) 2025 conference.

4.3 Background: Large-Language Models and HCI

Prior to reporting the details of these studies in the next two chapters, we first discuss the background in the literature specifically regarding LLM-based HCI studies. Human-computer interaction research on user impact from LLMs is still in its beginning stages. Much of the current research is still based in analyzing user output and using qualitative methods to understand user preferences [176, 177]. However, in recent years, quantitative methods have played a more significant role with studies looking at how user performance and time spent on a task changes with the use of LLMs [178, 179]. Notable areas of application where research has been conducted to understand the effects LLM tools have on users include writing, computer programming, and decision making.

Writing

Yuan tested an LLM story writing tool with professional authors to gain insights into the effectiveness of LLMs in supporting creative writing [177]. Singh examined the potential and challenges of LLM use for creative writing [180], and Reza produced ABScribe, a novel interface for more easily integrating human and machine-generated work in Human-AI co-writing tasks [181]. Other researchers have explored whether there is a difference between quality in AI and human-generated literary short texts [182]. Both Yuan and other studies have, using both qualitative and quantitative methods, demonstrated a productivity boost when using LLMs for work-related tasks, especially for novice and low-skilled workers [177, 178, 183]. However, the complexity of these systems reduces their benefits for novice users who don't know how to use them effectively, especially in light of the

sophistication required for prompt design [184, 185].

Programming

Computer programming has also proven an effective testing ground for studying the effects of LLM tools on users. Ziegler [186] performed a comprehensive study investigating the effects of Github Copilot on users, with a specific interest in productivity while Nguyen studied the challenges that non-expert users face when using LLM-based tools to assist in programming [187].

Decision Making

Researchers have also investigated the benefits, drawbacks, and limitations of using LLM tools as an integral component of decision making processes. Lawless investigated the combination of LLMs with Constraint Programming to facilitate decision making [188]. Chiang studied the use of AI tools to help decision making specifically in group-based settings [189]. Buçinca has studied intrinsic motivation in Human-AI decision making, and Lakkaraju investigated the fairness and efficacy of LLM tools used in the context of financial decision making [190].

Other LLM-based User Studies

Other avenues of approach for investigating the effects of LLM-based tools on users include Arakawa's work on adapting an LLM chatbot towards executive coaching [191], Huang's work exploring the use of LLM assistants to help prevent driver fatigue [192], Suh's work on LLM-based tools for structured design space exploration [193] and multilevel sensemaking [194], and Tankelevitch's work on mapping the underlying metacognitive load while using AI tools [195].

4.3.1 Gaps

Notably, little significant research could be found that utilized physiological measures or analyzed a user's cognitive workload directly via self-report. Microsoft's own early user research into the effectiveness of their Copilot LLM reflects this trend [196]. More research is necessary to establish design practices for LLM interfaces that maximize the tool's benefit for the user.

Chapter 5

Effects of LLMs on Humans Across a Gradient of Subjectivity

The first study in this work investigates neural responses of participants during 4 tasks which comprise a gradient of subjectivity intended to illustrate a vector of challenge in terms of the AI-tool's ability to benefit the user which coincides with varying aspects of activation in the prefrontal cortex. In this within-subjects study, using tasks ranging from objective (SAT reading comprehension) to subjective (personal reflection), and with measurements including fNIRS, Empatica E4, NASA-TLX, and questionnaires, I measure Copilot's effects on users. I also evaluate users' performance with and without Copilot across tasks. In objective tasks, participants reported a reduction of workload and an increase in enjoyment, which was paired with objective performance increases. Participants reported reduced workload and increased enjoyment with no change in performance in a creative poetry writing task. However, no benefits due to Copilot use were reported in a highly subjective self-reflection task. Although no physiological changes were statistically significant due to Copilot use, task-dependent differences in prefrontal cortex activation offer complementary insights into the cognitive processes associated with successful and unsuccessful human-AI collaboration: specifically, that AI assistants' effectiveness varies with task type—particularly showing decreased usefulness in tasks that engage episodic memory—and presents a brain-network based hypothesis of human-AI collaboration. Further, machine learning analyses indicate operational information related to reading comprehension during with vs. without LLM-tool use. This study presents interesting

insights regarding the benefits, drawbacks, effects, and usability of LLM tools, while demonstrating operational capabilities of neural signals with vs. without LLM tool use for future real-time implicit BCIs.

5.1 Copilot for Microsoft Word

Copilot for Word is an extension of the standard Microsoft Word interface which leverages AI to assist users throughout a variety of tasks. Although the Copilot ecosystem in Word allows users a wide array of functionality through multiple contexts, in order to minimize training time for our users as well as the potential for interface-based confounds, we focused the user's interaction with Copilot to a single chat window on the side of the Word screen (see Figure 5.1). This chat allows users to interact with the Copilot assistant, and it in turn interfaces with a LLM to produce relevant responses. While the specifics of which LLM is used are abstracted from the Word interface, Microsoft's documentation specifies that it leverages a variant of GPT-4 along with the text-to-image model DALL-E [197]. For this research, the relevant tasks that Copilot can perform are: text generation and refinement, answering queries related to the current document, or queries requesting general information or answers to specific questions.

5.2 Research Questions

The primary aim of this study is to explore the measurable effects of using an interactive LLM, in this case Copilot for Microsoft Word, on human users. Our specific research questions regarding this aim follow below. For each question **RQX** we are interested in **RQX-A**: overall effect, **RQX-B**: effects within each task, and **RQX-C**: effects that differ along the gradient of subjectivity[1].

RQ1-`TLX`

Does the use of the Copilot assistant change users' workload levels as measured by NASA-TLX?

[1]For **RQX-C**, we consider effects of `CONDITION` across `TASK` if the interaction effect is significant, otherwise if the interaction is not significant we consider effects simply across `TASK`.

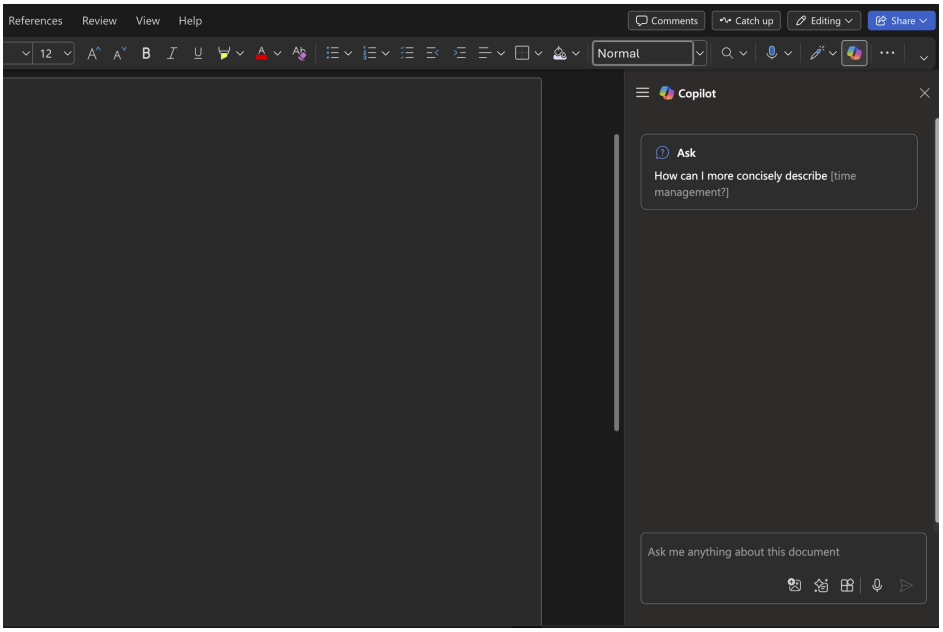


Figure 5.1: Microsoft Word with the integrated Copilot sidebar on the right-hand side of the screen. Copilot has access to the context window of the open document.

RQ2-`fNIRS`

Does using the Copilot assistant change users' levels of prefrontal cortex activation as measured by fNIRS? Are there patterns within this data which can be leveraged for real-time implicit BCI applications?[2]

RQ3-`E4`

Does the use of the Copilot assistant change users' levels of stress as measured by HR, HRV, and EDA?

RQ4-`QUALITY`

Does using the Copilot assistant change the quality of users' output?

[2]These two questions will be approached in different ways: the first with statistical analyses using LMMs, and the second with machine learning.

RQ5-FEELING

How do users feel about using the Copilot assistant?

5.3 Materials and Methods

5.3.1 Study Tasks

We modeled our tasks along a *gradient of subjectivity*. We designed this gradient along theoretical considerations of neurological systems, and developed tasks with practical experimental constraints in mind. At one end of the gradient are highly structured tasks with objectively clear and correct answers: we hypothesized that these tasks would engage participants in mental workload typically associated with prefrontal cortex activity; we expected these tasks would allow Copilot to meaningfully assist users, and would result in a corresponding decrease of prefrontal activation relating to decreased workload. At the opposite end of the gradient are open-ended tasks with highly subjective elements: we hypothesized that these tasks would engage participants in prefrontal activation associated with episodic memory; we expected that these tasks would present significant challenges for the AI assistant and would not affect brain function.

Determining the specific tasks that we would have our users engage in required much care and several iterations to strike a balance between tasks that were easy enough for the LLM that it could perform them perfectly with a single click and tasks that were too lengthy and involved for users to accomplish in a reasonable amount of time. A particular challenge we discovered from prior research and our own tests is that large language models are most effective in tasks with high complexity and low ambiguity [198]; that is, Copilot produces highly detailed and effective output in direct proportion to the level of detail and structure of the task: the more structured and detailed the task, the more structured and detailed the output from Copilot. After iterative refinement, we settled on a set of four task groups: **reading comprehension** (objective, fact-based, requires working memory), **event planning** (structured, but creative), **poetry writing** (creative with personal elements), and **personal reflection** (highly subjective and directly connected to personal experience and episodic memory). For each task type, we created two subtasks designed to be equally difficult. The full subtask details are located in Appendix B; statistical analysis showing

no differences in workload between subtask variants, as well as no differences in workload over time, are located in Appendix C.

Reading Comprehension [SAT]

These questions were slightly modified versions of examples taken from the CollegeBoard's Scholastic Aptitude Test (SAT), and were easily answered by Copilot. This task served as a baseline, representing highly objective problem-solving with minimal subjectivity. We anticipated standard cognitive demands on users without Copilot, and minimal cognitive demand when assisted by the LLM.

Event Planning [PLANNING]

These tasks asked the user to design and plan an event with structured and detailed to-do checklists of the event-related information. While still structured, these tasks were more open-ended than those in SAT, and required more subjective, personal, and creative input. We hypothesized that Copilot would be helpful to the user in completing this task, but that it would require more work from users in the Copilot condition as compared to SAT.

Poetry Writing [POEM]

These tasks asked the user to write a short poem of 10-15 lines on a broad theme such as joy or nature. This task represents a substantial shift toward subjective material, requiring purely creative expression that, at least in the without-LLM assistance condition, would necessarily draw on subjective personal experience. We believed that this task would engage more fully with the episodic memory than the first two, and that the LLM assistant would enable users to quickly produce output, but that it also would struggle to assist them given the inherently subjective nature of a poem.

Personal Reflection [REFLECTION]

These tasks asked the user to reflect on their favorite album or movie and discuss why it was their favorite based on their personal experiences. This task was designed to maximally engage purely

subjective, autobiographical episodic memory; we therefore hypothesized that it would be quite challenging for the LLM tool to meaningfully assist the user during these tasks.

5.3.2 Study Structure

All users signed informed consent documents prior to beginning the study, which was approved by the Tufts University Institutional Review Board. We provided an initial survey regarding familiarity with AI tools. Participants then did a 5 minute training task to familiarize them with the Copilot assistant. This included a variety of prompts for the user to use with the assistant to better help them understand what it could and could not do. Participants were able to ask questions prior to beginning the tasks if they needed help with Copilot. Users then completed each of the four tasks in a randomized order counterbalanced across participants. The choice of which subtask would be completed with the LLM assistant was likewise counterbalanced. After each task participants filled out post-task surveys including the NASA-TLX, a space for users to write any comments they would like, and a follow-up question rated on a scale of [0-10]: "How would you rate your overall experience with this task? (0=Terrible, 10=Amazing)". The participants were compensated with an Amazon gift card ($25) for their time.

5.3.3 Data Collection and Preprocessing

Demographics

We recruited 20 healthy individuals (7 men, 10 women, 3 opted not to disclose) for the study, ranging from 18-25 years old (mean 21).

Exclusions from Physiological Data

Four participants were excluded due to excessive noise across multiple trials seen through visual inspection of the fNIRS data. One fNIRS participant was excluded because an experimenter incorrectly marked the data and one was excluded because the user refused to wear the fNIRS headband. Within otherwise used fNIRS data, frequency domain Δ[HbD]ϕ data of the left prefrontal cortex for two participants in one task session exceeded 1.5 times the Interquartile Range (IQR) across all participants: the data were also excluded. From the Empatica data, three users had

invalid signal connection issues between the E4 and our collection device during collection time (Google Pixel 6 Phone), two users were excluded due to manual marker input errors, and one user declined to wear the wristband.

fNIRS System

For this study, we utilized the fNIRS setup discussed in Section 3.1.1. Each optical probe had optode geometry designed for the dual-slope (DS) method [44]. After fitting the headband and before starting collection data, nominal gains for each detector were found using BOXY for the user. The resulting I and ϕ data for each source-detector pair was processed DS methods, resulting in measurements of Δ[HbO] (μM) and Δ[HbR] (μM) for both DS Intensity (DSI) or DS Phase (DSϕ) [44].

fNIRS Data Preprocessing

A 5th order Butterworth bandpass filter was applied of the range [0.02, 0.2] Hz [43], and baseline correction for each measurement in each trial was performed with the initial 15 seconds from that trial. Two separate processing pipelines were then performed: one for statistical analysis and the other for machine learning.

fNIRS Preprocessing for Statistical Analyses For statistical analysis, Δ[HbD] was calculated by Δ[HbO]-Δ[HbR] [48], frequency domain transformation was performed using the Multitaper method [130, 131], Simpson's rule was used to integrate over the VLF frequency band [136], and the resulting values were log-transformed. Statistical analyses were then performed on the `DSI` and `DS`ϕ data [21, 50], with separate models created for each probe and measurement value. For convenience, I refer to the log total power in the VLFO band of the fNIRS signal (encompassing both `DSI` and `DS`ϕ) as `fNIRS` in the text below.

fNIRS Preprocessing for Machine Learning For each measurement, we apply aggregation functions on the data for each combination of participant, task, and condition. This resulted in a feature vector of length 32: 4 aggregation functions [`mean`, `standard deviation`, `skew`, and `slope of the linear regression`] x 2 probe positions [`L`, `R`] x 2 wavelengths [`HbO`, `HbR`] x 2 measurements

[DSI, DSϕ]. Although aggregation could be performed over varying window sizes, we select 100 samples for this analysis. See the Section 5.3.5 for further details on the methodology used for machine learning.

Empatica E4

Our preprocessing steps for the various Empatica E4 data was as follows.

- HR: We extracted the mean HR for each trial.
- HRV: Because Empatica's inter-beat-interval recording has preprocessing of the signal applied in advance of the point of measurement from the device that removes most of the non-normal beats in the RR interval, we used Empatica's IBI to represent the IBI of normal sinus beats (NN [199]) [115], and used the standard deviation of the Empatica IBI data as SDNN for our HRV calculation. We excluded trials with an IBI value outside of the range [1, 125] ms (5/81 trials were excluded).
- EDA: Each of the trials were bandpass filtered with a 4th order Butterworth filter of the range [0.01, 0.8] Hz [200]. We then transformed the signal into the frequency domain using the same process as with the fNIRS data; the frequency band extracted was [0.045 0.25] Hz which has been shown to produce a reliable inference of sympathetic EDA [200].

Separate statistical modeling was performed on each of the data streams.

NASA-TLX

TLX is defined as the unweighted average NASA-TLX scores for each participant's response for each subtask [123].

ENJOYMENT

ENJOYMENT for each participant for each instance of each subtask is the value reported in the survey after the task asking them to rate their overall experience with that task [0-10].

Task Evaluation Scores

We also measured and evaluated the actual quality of participants' output. This was difficult for the same reasons that designing the tasks themselves was challenging: moving down the gradient of subjectivity, it becomes more difficult to assign objective grades to the tasks. That is, while it is trivial to determine grades for SAT, and perhaps slightly less so for the PLANNING, it is much ore challenging to do so for POEM and REFLECTION. Our solution for the three non-objective tasks was to use a two-dimensional methodology of grading: *breadth* and *depth*. The goal of this methodology was not to grade based on the potentially subjective nature of the users' output but instead to quantify their success in performing the task at hand. Three members of our research team graded each of the submissions provided for PLANNING, POEM and REFLECTION independently, rating each submission on a [1-5] scale for both of *breadth* and *depth* (details of the methodology can be found in the Supplementary Material. Consistency of the graders' output was measured with Intraclass Correlation ICC [201], specifically using a two-way mixed-effects model considering consistency over the mean of k raters (ICC3k) [202]. Quality scores for *breadth* and *depth* were averaged across graders, and the resulting scores were then averaged to produce a single score value for each user for each task. Quality scores for each task were then normalized across users to a 0-1 scale. SAT quality scores were simply defined as the percent of correct answers total per task: for this reason, SAT data were modeled separately from the other three tasks. These evaluations are referred to as PERFORMANCE.

5.3.4 Statistical Methods

To account for the repeated measures design of our study we analyzed our data using Linear Mixed Models (LMMs) [203]. We created separate models for each research question using the following R formula as a template:

$$\texttt{DV} \sim \texttt{CONDITION} * \texttt{TASK} + \texttt{(1|PID/TASK)} \tag{5.1}$$

Where DV represents the measured dependent variable of interest (TLX, fNIRS, HR, HRV, IBI, PERFORMANCE, or ENJOYMENT), CONDITION is a factor with two levels indicating use of Copilot (with-Copilot (AI) or without-Copilot (NAI)), and TASK is a factor with four levels indicating the type of task performed (SAT, PLANNING, POEM, or REFLECTION). Random intercepts are specified for

each participant (PID), with nested intercepts within participant for each TASK. For each model, likelihood ratio tests (LRTs) were used to refine the random effects structure [204]; models that showed better fit without the nested random effect of TASK within PID had this term removed. ANOVA results from the LMMs for CONDITION are used to determine significance for all RQX-A. If interaction of CONDITION $\times$ TASK demonstrates significance, post-hoc contrasts are performed among the emmeans for CONDITION within levels of TASK to answer all RQX-B. To answer all RQX-C questions respective of Copilot (done if CONDITION $\times$ TASK is significant), custom emmeans contrasts are performed to test the effect of CONDITION across different pairs of TASK levels. To answer all RQX-C questions irrespective of Copilot (done if CONDITION $\times$ TASK is not significant, but TASK is), post-hoc contrasts are performed among the emmeans comparing levels of TASK; For all tests, α is set at 0.05, except in the case of omnibus testing for fNIRS data and empatica data, where I apply Bonferroni correction; for fNIRS, we consider the measures of DSI and DSϕ for each side of L and R as related, and thus α is adjusted to 0.025; for Empatica, I consider HR and HRV related, so α is adjusted to 0.025 for those tests.

5.3.5 Machine Learning

Given that we wish to consider the potential for future real-time implicit BCI applications to be developed based on these data, we also present machine learning results for RQ2-fNIRS. We first prepare the dataset as discussed in Section 5.3.3, and select a window size of 100 samples (19.2s). This window size was chosen with a balance of theoretical justification of neural activation within the windows, likely usable data in an online context. Due to the study design allowing participants to complete tasks early if desired, there is some class imbalance in the dataset at the individual participant level: for context, we report the average support levels across classes. Additionally, the reported metric, F1 score, remains robust to such minor imbalances in class distribution. For models, I selected RF, SVM, SVC, ANN, and KNN, and perform Leave-One-Out Cross-Validation (LOO-CV), where each participant's data is in turn exclusively used for testing a model trained on the other participants' data. Reported results are both Macro-F1 scores averaged across participants per-model (STANDARD), as well as best scores per-participant (OPTIMIZED). Different classification labels are used dependent on the research question:

- RQ2-fNIRS-A: Use data from all tasks with `CONDITION` [`AI`, `NAI`] as the label.
- RQ2-fNIRS-B: Use data for one task at a time with `CONDITION` [`AI`, `NAI`] as the label.
- RQ2-fNIRS-C:
  - Irrespective of `CONDITION`: Select each set of pairs within `TASK`, and classify with `TASK` as the label.
  - Within levels of `CONDITION`: Do the above within levels of `AI` or `NAI`.[3]

All classification problems are binary classification.

5.4 Results

5.4.1 RQ1-TLX Results

Does the Copilot assistant change overall workload levels as measured by `TLX`?

RQ1-TLX-A Results

The Copilot condition resulted in overall lower `TLX` scores ($F_{1,133} = 60.42, p < 0.001, \epsilon_p^2 = .31$). Results are visible in Table 5.1 and visualized in Figure 5.2.

Table 5.1: ANOVA result from a model with `WORKLOAD` as the `DV` in Formula 5.1. Although overall self-reported workload decreased with Copilot, differences were found with an interaction with `CONDITION`.

Factor	**df1**	**df2**	**F**	**p**	**sig.**	ϵ_p^2	ϵ_p^2 **CI**
CONDITION	1	133	60.42	<0.001	***	0.31	[0.20,0.40]
TASK	3	133	9	<0.001	***	0.15	[0.06,0.23]
CONDITION × TASK	3	133	7.07	<0.001	***	0.12	[0.03,0.20]

[3]Although more complex machine learning equivalents of our statistical analyses are possible, translating our approach of post-hoc contrasts studying the effect of one factor within levels of another to a machine learning framework would require multiple layers of modeling and comparison: I therefore opted for this simplified approach to only study RQ2-fNIRS-C irrespective of and within `CONDITION`.

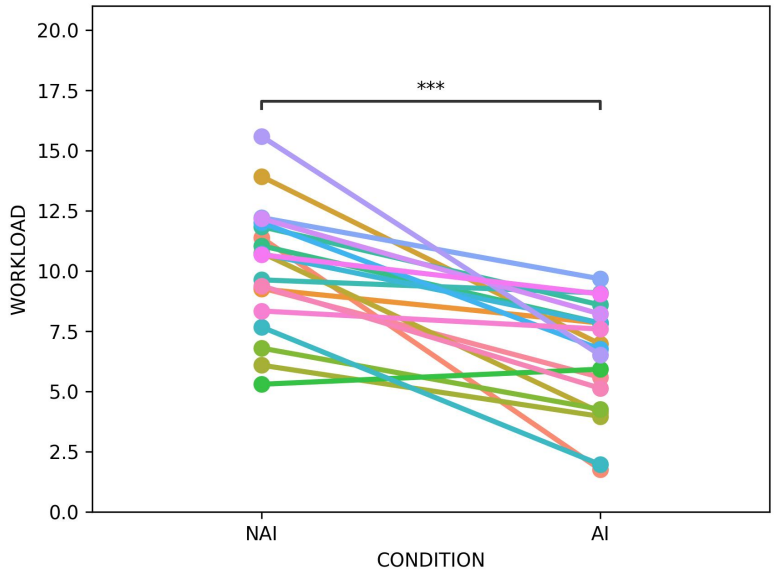


Figure 5.2: `TLX` scores in the `NAI` (without Copilot) and `AI` (with Copilot) conditions over all tasks. Each line represents a unique user. Self-reported workload generally decreased when using Copilot. Further discussion of separate effects across levels of `TASK` is below.

RQ1-TLX-B Results

`CONDITION` × `TASK` demonstrated a strong effect ($F_{3,133} = 7.07, p < 0.001, \epsilon_p^2 = .12$). Pairwise contrasts shown in Table 5.2 and visualized in Figure 5.3 show that the `AI` was significantly less than `NAI` for all levels of `TASK` with the notable exception of `REFLECTION` ($t_{133} = 0.17, p = 0.864, \epsilon_p^2 = 0.00$), which did not show a significant change. This result is as-expected in terms of decreases in workload decreases for the more objective `SAT` and `PLANNING`, and in terms of no change for `REFLECTION`, but it is somewhat surprising that the participants reported a large decrease in workload with Copilot in the more subjective `POEM`.

Table 5.2: Effects of Copilot use on self-reported mental workload within levels of `TASK`. Copilot reduced self-reported workload for all tasks except `REFLECTION`.

Task	**Contrast**	**Est.**	**SE**	**df**	**t**	**p**	**sig.**	ϵ_p^2	ϵ_p^2 **CI**
SAT	NAI - AI	5.46	0.97	133	5.63	<0.001	***	0.19	[0.08, 0.30]
POEM	NAI - AI	5.80	0.97	133	5.98	<0.001	***	0.21	[0.10, 0.32]
PLANNING	NAI - AI	3.66	0.97	133	3.77	<0.001	***	0.09	[0.02, 0.19]
REFLECTION	NAI - AI	0.17	0.97	133	0.17	0.864	ns	0.00	[0.00, 0.00]

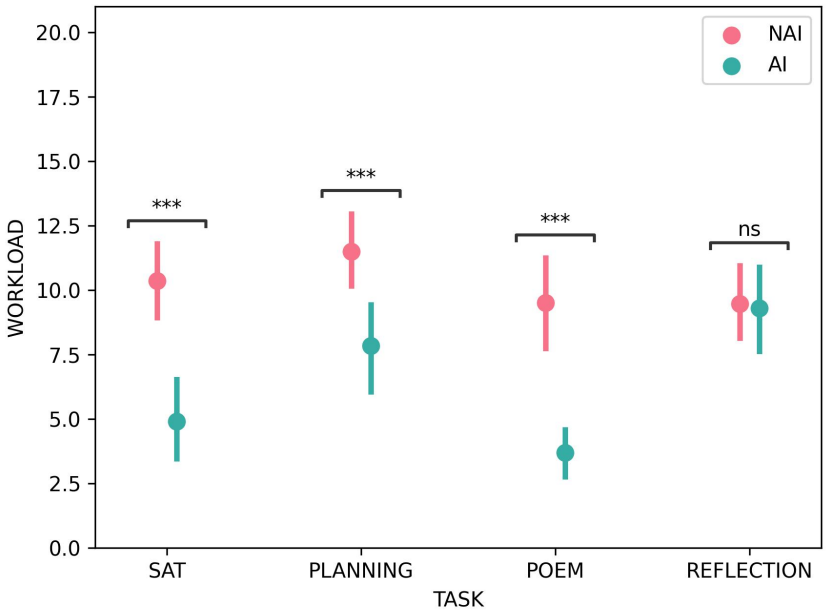


Figure 5.3: Self-reported workload levels were lower with Copilot for all levels of `TASK` except `REFLECTION`, which shows no change.

RQ1-TLX-C Results

Results regarding RQ1-TLX-C in consideration of changes due to Copilot use are shown in Figure 5.4 and Table 5.3. Copilot significantly reduced workload in all tasks in relation to `REFLECTION`: `POEM - REFLECTION` ($t_{133} = 4.11, p < 0.001, \epsilon_p^2 = 0.11$), `SAT - REFLECTION` ($t_{133} = 3.86, p < 0.001, \epsilon_p^2 = 0.09$), and `PLANNING - REFLECTION` ($t_{133} = 2.54, p = 0.012, \epsilon_p^2 = 0.04$).

Table 5.3: Contrast results comparing the effect of `AI` versus `NAI` across levels of `TASK` on self-reported workload. The decrease in workload accounted for by Copilot was significantly larger in `SAT`, `POEM`, and `PLANNING` than in `REFLECTION`.

Contrast	Effect	Est.	SE	df	t	p	sig.	ϵ_p^2	ϵ_p^2 CI
POEM - REFLECTION	AI - NAI	5.63	1.37	133	4.11	<0.001	***	0.11	[0.03,0.21]
SAT - REFLECTION	AI - NAI	5.29	1.37	133	3.86	<0.001	***	0.09	[0.02,0.20]
PLANNING - REFLECTION	AI - NAI	3.49	1.37	133	2.54	0.012	*	0.04	[0.00,0.12]
PLANNING - POEM	AI - NAI	-2.14	1.37	133	-1.56	0.121	ns	0.01	[0.00,0.07]
PLANNING - SAT	AI - NAI	-1.80	1.37	133	-1.31	0.192	ns	0.01	[0.00,0.06]
POEM - SAT	AI - NAI	0.34	1.37	133	0.25	0.804	ns	0.00	[0.00,0.00]

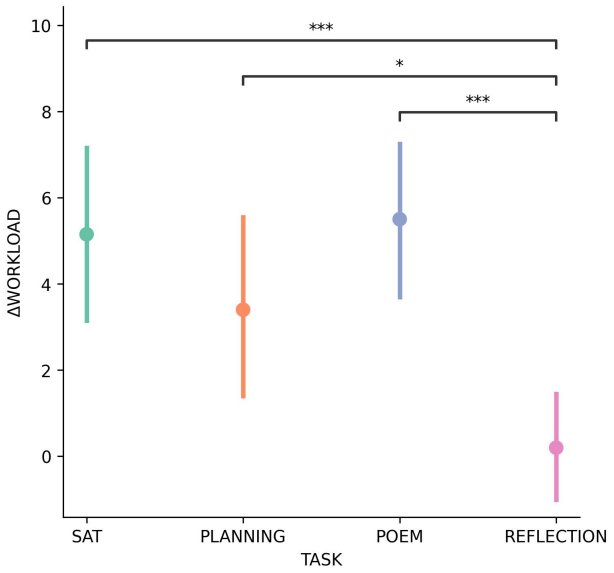


Figure 5.4: Effect of Copilot use on self-reported workload across tasks. Larger values indicate that Copilot decreased workload by a larger amount. Self-reported TLX scores were significantly lowered by Copilot in all tasks as compared to `REFLECTION`.

RQ1-TLX Results Summary

As expected, self-reported workload decreased with Copilot in relation to the gradient of subjectivity: `SAT` and `PLANNING` exhibited large decreases, whereas `REFLECTION` did not. Surprisingly, we also noted the largest overall decrease in self-reported workload during `POEM`. These results indicate that LLM-use may be helpful to users during subjective tasks which are purely creative, but not in subjective tasks which engage episodic memory.

5.4.2 RQ2-fNIRS Results

RQ2

Does the use of the Copilot assistant change users' levels of prefrontal cortex activation as measured by `fNIRS`?

RQ2-fNIRS-A and RQ2-fNIRS-B LMM Results

Detailed results are in Table 5.4. The use of Copilot did not effect `fNIRS` for either `DSI` or `DSϕ` either the left (`DSI`: $F_{1,52} = 2.60, p = 0.113, \epsilon_p^2 = 0.03$; `DSϕ`: $F_{1,51.04} = 2.14, p = 0.150, \epsilon_p^2 = 0.02$)

or right (DSI: $F_{1,52} = 0.61, p = 0.437, \epsilon_p^2 = 0.00$; DSϕ: $F_{1,39.0} = 0.17, p = 0.683, \epsilon_p^2 = 0.00$) sides. Similarly, no effects were found among the interaction of CONDITION×TASK for either measure in the left (DSI: $F_{1,52} = 1.25, p = 0.302, \epsilon_p^2 = 0.01$; DSϕ: $F_{1,50.98} = 1.90, p = 0.142, \epsilon_p^2 = 0.05$) or right (DSI: $F_{1,52} = 0.54, p = 0.655, \epsilon_p^2 = 0.00$; DSϕ: $F_{1,52} = 0.39, p = 0.736, \epsilon_p^2 = 0.00$). These results indicate that, despite self-reported workload changes, there were not large measurable changes in PFC activity due to differential VLFO patterns as a consequence of Copilot use.

Given the large self-reported workload differences with and without AI, a lack of significance in the statistical data in this regard is somewhat surprising. One possible explanation is that the differences in any difficulty levels between the tasks' baselines and the Copilot use was not extreme, for example as in similar levels of the N-Back task [205]. Another possible consideration is the nature of the sample size - these possibilities are discussed in more detail later, in Part V.

Table 5.4: Results of modeling formula 5.1 with fNIRS as the DV for all four combinations of [L, R], and [DSI, DSϕ]. These results indicate significant activation changes in DSI of the right PFC based on TASK. No effect on prefrontal activity in either the left or right PFC, or in relation to DSϕ, is shown under CONDITION. Note that **sig.** considers adjusted α of 0.025, correcting across tests for DSI and DSϕ with each of L and R, separately.

Side	Meas	Factor	df1	df2	F	p	sig.	ϵ_p^2	ϵ_p^2 CI
L	DSI	CONDITION	1	52.0	2.60	0.113	ns	0.03	[0.00,0.14]
L	DSI	TASK	3	39.0	1.40	0.256	ns	0.03	[0.00,0.09]
L	DSI	CONDITION x TASK	3	52.0	1.25	0.302	ns	0.01	[0.00,0.04]
L	DSϕ	CONDITION	1	51.04	2.14	0.150	ns	0.02	[0.00,0.13]
L	DSϕ	TASK	3	39.08	1.19	0.330	ns	0.01	[0.00,0.03]
L	DSϕ	CONDITION x TASK	3	50.98	1.90	0.142	ns	0.05	[0.00,0.13]
R	DSI	CONDITION	1	52.0	0.61	0.437	ns	0.00	[0.00,0.00]
R	DSI	TASK	3	39.0	3.52	0.024	*	0.15	[0.00,0.30]
R	DSI	CONDITION x TASK	3	52.0	0.54	0.655	ns	0.00	[0.00,0.00]
R	DSϕ	CONDITION	1	52.0	0.17	0.683	ns	0.00	[0.00,0.00]
R	DSϕ	TASK	3	39.0	0.20	0.898	ns	0.00	[0.00,0.00]
R	DSϕ	CONDITION x TASK	3	52.0	0.39	0.763	ns	0.00	[0.00,0.00]

RQ2-fNIRS-C LMM Results

Irrespective of CONDITION, TASK showed significance with a strong effect size as measured on the right aspect of the PFC in the DSI measurement ($F_{3,39} = 3.52, p = 0.024, \epsilon_p^2 = 0.15$): post-hoc contrasts

were therefore run for TASK within the right probe. Results are shown in Table 5.5, and visualized in Figure 5.5. Of note are differences between PLANNING - REFLECTION ($t_{39} = 2.82, p = 0.036, \epsilon_p^2 = 0.15$) and SAT - REFLECTION ($t_{39} = 2.76, p = 0.042, \epsilon_p^2 = 0.14$); and although not significant, given the effect size we also note POEM - REFLECTION ($t_{39} = 2.20, p = 0.142, \epsilon_p^2 = 0.09$). These results indicate a difference in PFC activity as a consequence of TASK, specifically indicating that the episodic memory task REFLECTION induced higher prefrontal cortex activation as compared to the other tasks.

Worth consideration here is the lack of significance in the DSϕ finding. However, given that the DSϕ results from the baseline example N-Back data only show sensitivity to the strongest differentiation of levels of 0-3 back (see Appendix A for details), it is reasonable to expect that the DSϕ data may also not be noticeably different in the contexts which produce the most extreme differentiations of neural activation.

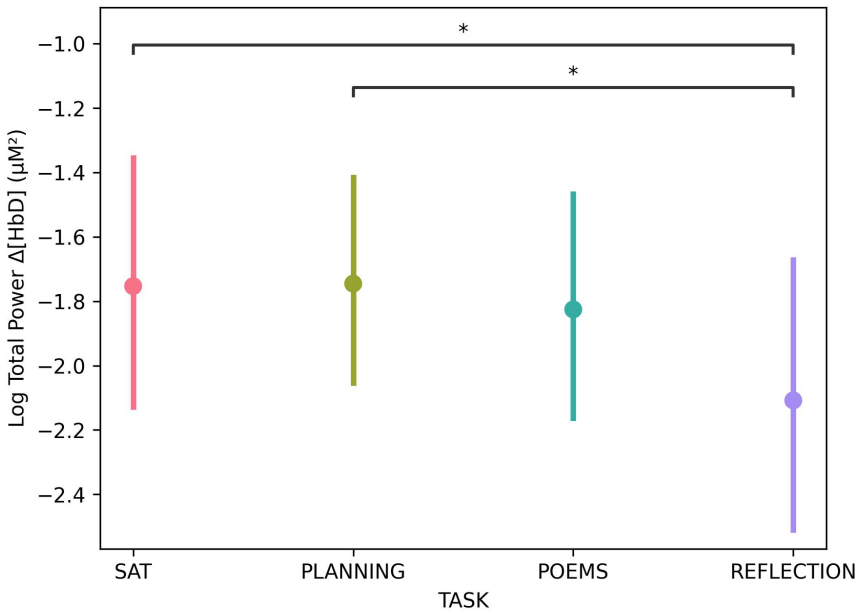


Figure 5.5: Log total power of Δ[HbD] of the VLF band in the right prefrontal probe compared across tasks, irrespective of CONDITION. Note that lower total power indicates higher prefrontal activation. The REFLECTION task demonstrated higher levels of activation as compared to SAT and PLANNING, likely due to its engagement of episodic memory.

Table 5.5: VLF Δ[HbD] contrast results for the TASK factor. REFLECTION showed decreased activity in the VLF band, indicating increased prefrontal activation, as compared to the SAT and PLANNING tasks, irrespective of CONDITION.

Side	Contrast	Est.	SE	df	t	p	sig.	ϵ_p^2	ϵ_p^2 CI
R	PLANNING - REFLECTION	0.36	0.13	39.00	2.82	0.036	*	0.15	[0.01,0.35]
R	SAT - REFLECTION	0.35	0.13	39.00	2.76	0.042	*	0.14	[0.01,0.35]
R	POEM - REFLECTION	0.28	0.13	39.00	2.20	0.142	ns	0.09	[0.00,0.28]
R	PLANNING - POEM	0.08	0.13	39.00	0.62	0.924	ns	0.00	[0.00,0.00]
R	POEM - SAT	-0.07	0.13	39.00	-0.56	0.942	ns	0.00	[0.00,0.00]
R	PLANNING - SAT	0.01	0.13	39.00	0.06	1.000	ns	0.00	[0.00,0.00]

RQ2-fNIRS-A and RQ2-fNIRS-B Machine Learning Results

See Table 5.6: regarding classification of brain data in relation to Copilot use, the LOO-CV procedure indicates promising performance for SAT (STANDARD: 68.0%, OPTIMIZED: 69.0%) and reasonable performance overall (STANDARD: 62.7%, OPTIMIZED: 64.3%). By contrast, although the other models did not perform well for STANDARD classification (PLANNING (50.7%), POEM (54.6%), and REFLECTION (52.3%)), their OPTIMIZES models did much better (PLANNING (64.2%), POEM (71.1%), and REFLECTION (64.8%)). Contrary to the statistical finding, these results indicate that there are subtle patterns within the measured hemodynamic activity that differ dependent on LLM use in reading comprehension tasks that are detectable and operational based on machine learning techniques; however, SAT presents the only straightforwardly deployable model as of now. That said, the relatively large score of overall performance, and the reasonable scores for the OPTIMIZED models indicates that more data and/or more advanced analytical method may be useful in terms of machine learning classification being able to be deployable "in the field".

RQ2-fNIRS-C Machine Learning Results

See Tables 5.7 and 5.8. Of the possible comparisons between tasks irrespective of condition, REFLECTION vs. POEM scored the hightest (STANDARD: 58.1%, OPTIMIZED: 63.9%). And, despite poorer STANDARD score, REFLECTION vs. SAT in the OPTIMIZED score is also moderate (61.0%). Within the other comparisons across levels of TASK, however, the results are not as good, with all results $< 60\%$.

Within levels of CONDITION, however, a slightly different pattern emerges. Although each of the

Table 5.6: Machine learning classification performance for CONDITION, presented for the entire dataset (RQ2-fNIRS-A) and per individual task (RQ2-fNIRS-B). F1-scores represent averages across all participants, while "Support" columns indicate the average number of test samples per participant for AI and NAI conditions. Given that results are Macro F1, the slight imbalances between AI/NAI within each row less relevant than the number of samples used for testing between ALL tasks and each individual task. "Standard" results are within the listed best-case Model overall, whereas "Optimized" results are the collection of the best-scoring model per-participant.

Task	**Standard**		**Optimized**	**Support**	
	Model	**F1-score**	**F1-score**	**AI**	**NAI**
ALL	SVM	0.627	0.643	67.8	61.4
SAT	RF	0.680	0.690	14.1	16.7
PLANNING	KNN	0.507	0.642	18.0	16.9
POEM	SVM	0.546	0.711	15.9	12.9
REFLECTION	RF	0.523	0.648	17.8	18.3

STANDARD results are $< 60\%$, nearly all of the OPTIMIZED results are $> 60\%$, with the best results per-group in AI: POEM vs. SAT (69.5%), and NAI: REFLECTION vs. SAT (70.5%). Although more work is required in order to translate these best-results to usable real-time BCI applications, they are nevertheless promising. Further, it is worth mentioning that although the best results come from tasks on opposing sides of the gradient of subjectivity, a strict pattern as such cannot be established, as for instance NAI PLANNING vs. REFLECTION shows (STANDARD: 46.6%, OPTIMIZED: 51.5%).

Table 5.7: Machine learning classification performance for TASK pairs, irrespective of CONDITION. Standard results show the best single model for all participants, while optimized results represent the average when selecting the best model per participant. Support columns indicate the average number of test samples per participant for each task.

Task A	**Task B**	**Standard**		**Optimized**	**Support**	
		Model	**F1-score**	**F1-score**	**A**	**B**
REFLECTION	POEM	MLP	0.581	0.639	36.1	28.7
REFLECTION	SAT	MLP	0.528	0.610	36.1	29.6
PLANNING	SAT	MLP	0.491	0.543	34.8	29.6
REFLECTION	PLANNING	MLP	0.480	0.562	36.1	34.8
POEM	SAT	KNN	0.450	0.583	28.7	29.6
PLANNING	POEM	KNN	0.450	0.548	34.8	28.7

Given the large disparity in self-reported workload between REFLECTION and the other tasks, and the additional distinctions in the statistical analyses of the neural data between REFLECTION - PLANNING and REFLECTION - SAT as discussed above, it is somewhat disappointing that the machine

Table 5.8: Machine learning classification for TASK pairs within levels of CONDITION. Standard results show the best single model for all participants, while optimized results represent the average when selecting the best model per participant. Support columns indicate the average number of test samples per participant for each task.

Condition	Task A	Task B	Standard		Optimized	Support	
			Model	F1-score	F1-score	A	B
AI	POEM	REFLECTION	SVC	0.596	0.601	12.9	18.3
AI	PLANNING	SAT	SVM	0.523	0.632	16.9	14.4
AI	POEM	SAT	RF	0.506	0.695	12.9	14.4
AI	PLANNING	POEM	MLP	0.455	0.655	16.9	12.9
AI	PLANNING	REFLECTION	KNN	0.428	0.649	16.9	18.3
AI	REFLECTION	SAT	SVM	0.393	0.635	18.3	14.4
NAI	PLANNING	SAT	SVC	0.594	0.613	17.9	16.2
NAI	REFLECTION	SAT	SVC	0.578	0.705	17.8	16.2
NAI	PLANNING	POEM	MLP	0.518	0.538	17.9	15.9
NAI	POEM	REFLECTION	KNN	0.484	0.658	15.9	17.8
NAI	PLANNING	REFLECTION	SVC	0.466	0.515	17.9	17.8
NAI	POEM	SAT	KNN	0.391	0.630	15.9	16.2

learning results for RQ2-fNIRS-C do not suggest that this finding is presently actionable in terms of an applied BCI context. However, it is important to note the differences in methodology between the statistical and the machine learning analyses: namely, that of temporal scope. In the statistical paradigm, the data from each approximately 6-minute task is condensed into a single value for each measurement and probe location; the machine learning method, by contrast, with an eye towards potential application in real-time settings, broke up each task block into sub-segments, on which the analyses were done. Therefore, although the statistical result presents the best-effort to improve the signal-to-noise ratio, its suffers in terms of an inability to be used in real-time contexts; by contrast, the machine learning results are more usable in real-time scenarios, and can leverage high-dimensional patterns, but lack explainability in terms of the physiological response[4]. Further, there is promising potential within the OPTIMIZED results which indicates opportunities for more customized classification approaches across participants.

RQ2-fNIRS Results Summary

In terms of the statistical analyses, no changes were found in prefrontal activation as related to Copilot use; however, significant differences were seen across TASK in the right PFC: namely, between

[4]More is considered on this topic in Part V

REFLECTION and SAT/PLANNING. Given REFLECTION's engagement of episodic memory, it is sensible to detect an effect in this regard.

However, the machine learning results present an interesting and useful counterpoint to the statistical analyses. Firstly, although not an exceptional result, CONDITION was indeed separable within SAT in a manner which is likely usable in real-time BCIs. And secondly, reasonable cross-task differentiations were visible in the OPTIMIZED results. While this effect is likely not yet operational in a BCI context, it nevertheless shows promise for future applications dependent on newer methodologies or perhaps more data.

These divergences between statistical and machine learning results in these cases may themselves be attributed to distinct factors: for SAT, the sample size may be insufficient to detect the effect in statistical analyses, while the machine learning methods may be capable of detecting subtle patterns that the VLFO transformation might eliminate. Regarding cross-TASK comparisons, the hemodynamic differences appear most detectable when aggregating across entire task blocks, suggesting that while broad patterns of VLFO activity exist at the macro level, they become less distinguishable when examining at the level of the temporal specificity necessary for real-time applications.

5.4.3 RQ3-E4 Results

RQ3

Does the use of the Copilot assistant change users' levels of stress as measured by HR, HRV, and EDA? To answer this we first developed separate initial models where we use each of the signals of interest as defined in section 5.3.3 as the DV in Formula 5.1. Results shown in Table 5.9.

RQ3-E4-A, RQ3-E4-B, and RQ3-E4-C Results

A marginal effect with low effect size of CONDITION on HR was observed ($F_{1,77} = 3.29, p = 0.074, \epsilon_p^2 = 0.03$); no significant changes in any of the Empatica E4 measures were observed either within or across tasks. These findings suggest that stress as measured by cardiovascular and electrodermal activity is unchanged by Copilot use, tasks along the gradient of subjectivity, and the interaction of these factors.

Table 5.9: Results from separate models created from Formula 5.1 with each measurement type as DV. No physiological measurements from the Empatica E4 device showed significant changes as a consequence of TASK, CONDITION, or their interaction. Note that for HR and HRV tests α is set to 0.025 due to similarity of the research question underlying the tests.

Measure	Factor	df1	df2	F	p	sig.	ϵ_p^2	ϵ_p^2 CI
HR	CONDITION	1	77	3.29	0.074	ns	0.03	[0.00,0.11]
HR	TASK	3	77	0.33	0.801	ns	0.00	[0.00,0.00]
HR	CONDITION×TASK	3	77	0.46	0.712	ns	0.00	[0.00,0.00]
HRV	CONDITION	1	56.31	0.34	0.561	ns	0.00	[0.00,0.00]
HRV	TASK	3	57.02	1.18	0.324	ns	0.00	[0.00,0.00]
HRV	CONDITION×TASK	3	56.29	0.41	0.743	ns	0.00	[0.00,0.00]
EDA	CONDITION	1	77	0.03	0.858	ns	0.00	[0.00,0.00]
EDA	TASK	3	77	0.27	0.844	ns	0.00	[0.00,0.00]
EDA	CONDITION×TASK	3	77	0.46	0.710	ns	0.00	[0.00,0.00]

5.4.4 RQ4-QUALITY Results

RQ4

Does the use of the Copilot assistant affect the quality of users' output?

Table 5.10: Quality ANOVA results. Note that, due to the varying distribution of data, SAT was put in a separate model from the other levels of TASK. CONDITION showed significance for SAT, and CONDITION, TASK, and their interaction all showed significant effects for the other model.

Model	Factor	df1	df2	F	p	sig.	ϵ_p^2	ϵ_p^2 CI
SAT	CONDITION	1	20	15.00	<0.001	***	0.40	[0.13,0.60]
OTHERS	CONDITION	1	100	6.90	0.010	**	0.06	[0.01,0.14]
OTHERS	TASK	2	100	7.18	<0.001	**	0.11	[0.02,0.20]
OTHERS	CONDITION×TASK	2	100	3.53	0.033	*	0.05	[0.00,0.12]

RQ4-QUALITY-A and RQ4-QUALITY-B Results

There is a significant effect of CONDITION for both SAT ($F_{1,20} = 15, p < 0.001, \epsilon_p^2 = .40$) and the other tasks ($F_{1,100} = 6.9, p = 0.01, \epsilon_p^2 = 0.06$). For the three other tasks there is likewise an effect of CONDITION × TASK ($F_{2,100} = 3.53, p < 0.033, \epsilon_p^2 = .05$), but Figure 5.6 and Table 5.11 show that within these three tasks the only significant task is PLANNING ($t_{100} = 3.68, p < 0.001, \epsilon_p^2 = 0.11$). These results indicate an increase in QUALITY for the more objective tasks, but not the more

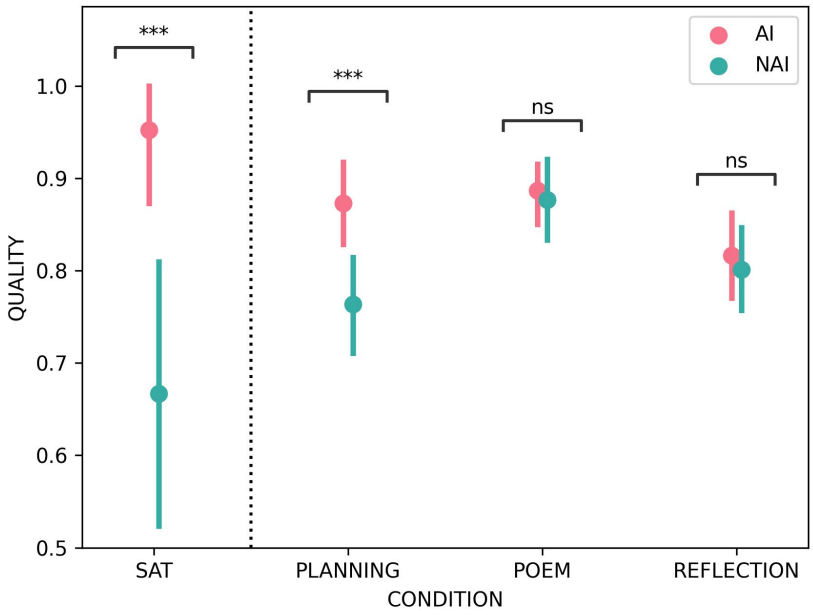


Figure 5.6: SAT and PLANNING tasks had significantly higher QUALITY scores in the AI condition. POEM and REFLECTION showed no change. Note that the SAT data was trained on a separate model because of distinctions in grading methodology.

subjective ones.

Table 5.11: QUALITY contrast results for all levels of TASK excluding SAT. Only PLANNING increased in the AI condition as compared to the NAI condition.

Task	**Contrast**	**Est.**	**SE**	**df**	**t**	**p**	**sig.**	ϵ_p^2	ϵ_p^2 **CI**
PLANNING	AI - NAI	0.11	0.03	100	3.68	<0.001	***	0.11	[0.02,0.23]
REFLECTION	AI - NAI	0.02	0.03	100	0.53	0.600	ns	0.00	[0.00,0.00]
POEM	AI - NAI	0.01	0.03	100	0.34	0.735	ns	0.00	[0.00,0.00]

RQ4-QUALITY-C Results

The largest effect is seen in SAT. Post-hoc contrasts observing effects across tasks for the changes between quality of AI versus NAI, shown in Table 5.12 and visualized in Figure 5.7, showed that the effect of the increase in QUALITY score of Copilot use is significantly higher in PLANNING as compared to POEM ($t_{100} = 2.36, p = 0.020, \epsilon_p^2 = 0.04$) and REFLECTION ($t_{100} = 2.23, p = 0.028, \epsilon_p^2 = 0.04$). These results indicate that Copilot may be beneficial in terms of quality output for more objective tasks.

Table 5.12: Contrast results comparing the effect of `AI` versus `NAI` across levels of TASK on Quality scores. The increase in quality accounted for by Copilot was larger in `PLANNING` than in `POEM` or `REFLECTION`.

Contrast	Effect	Est.	SE	df	t	p	sig.	ϵ_p^2	ϵ_p^2 CI
PLANNING - POEM	AI - NAI	0.10	0.04	100	2.36	0.020	*	0.04	[0.00,0.14]
PLAN - REFLECTION	AI - NAI	0.09	0.04	100	2.23	0.028	*	0.04	[0.00,0.14]
POEM - REFLECTION	AI - NAI	-0.01	0.04	100	-0.13	0.896	ns	0.00	[0.00,0.00]

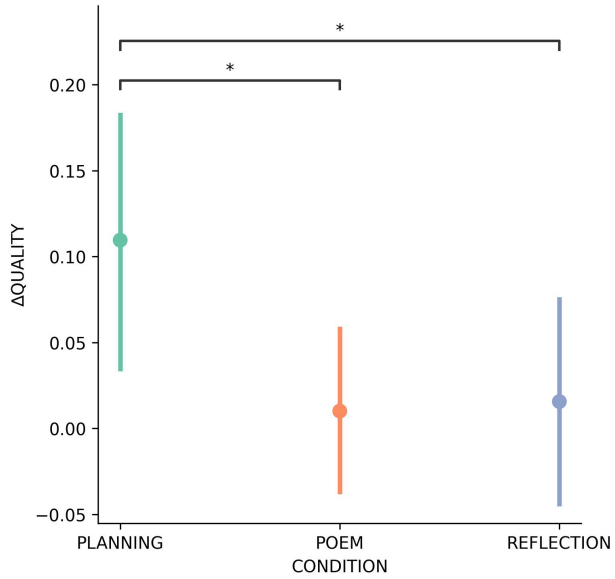


Figure 5.7: Effect of Copilot use on `QUALITY` scores across tasks. `QUALITY` increased significantly with Copilot in the `PLANNING` as compared to `POEM` and `REFLECTION`.

ICC Results

Scores for `OVERALL` (ICC $= 0.774, 95\%$ CI $= [0.7, 0.83]$), `PLANNING` (ICC $= 0.817, 95\%$ CI $= [0.69, 0.9]$), `REFLECTION` (ICC $= 0.751, 95\%$ CI $= [0.58, 0.86]$), and `POEM` (ICC $= 0.652, 95\%$ CI $= [0.42, 0.8]$) were all moderate. Within this range, however, we observed the expected behavior regarding our `ICC` measurement in that the more open-ended and subjective tasks demonstrated lower consistency scores, with the 95% lower CI for the `POEM` task rating as poor.

RQ4-QUALITY Results Summary

In summary, `QUALITY` scores for `SAT` and `PLANNING` increased with Copilot use, and the increase in quality score with Copilot use significantly differed between `PLANNING` and `POEM`/`REFLECTION`.

These results indicate that for more objective tasks, Copilot use can increase QUALITY, whereas for more subjective tasks, it is less likely to do so.

5.4.5 RQ5-FEELING Results: Quantitative Evaluation

RQ5-FEELING

Did participants enjoy using the Copilot assistant?

RQ5-FEELING-A and RQ5-FEELING-B Results

See Table 5.13. Participants reported higher ENJOYMENT when using Copilot ($F_{1,133} = 15.06, p < 0.001, \epsilon_p^2 = 0.05$). Contrast results (see Table 5.14 and Figure 5.8) indicate that, with the exception of REFLECTION ($t_{133} = -0.88, p = 0.380, \epsilon_p^2 = 0.00$), this is likewise true for each individual task. These results directly parallel the self-reported results for TLX, and indicate that, in addition to objective tasks, participants enjoyed using the Copilot assistant for subjective tasks which did not require self-reflection, and did not enjoy its use during reflective tasks.

Table 5.13: ANOVA results of Formula 5.1 with ENJOYMENT as the DV; significant results were found for CONDITION, TASK, and their interaction.

Factor	df1	df2	F	p	p.sig	ϵ_p^2	ϵ_p^2 CI
CONDITION	1	133	15.06	<0.001	***	0.10	[0.03,0.18]
TASK	3	133	4.66	<0.001	**	0.07	[0.01,0.14]
CONDITION×TASK	3	133	3.88	0.01	*	0.06	[0.00,0.12]

Table 5.14: Contrast results of ENJOYMENT (AI-NAI) within levels of TASK. All levels showed significant increases in ENJOYMENT during AI, with the notable exception of REFLECTION, which showed no significant change.

Task	Contrast	Est.	SE	df	t	p	sig.	ϵ_p^2	ϵ_p^2 CI
SAT	AI - NAI	2.25	0.62	133	3.60	<0.001	***	0.08	[0.02,0.18]
POEM	AI - NAI	1.80	0.62	133	2.88	0.005	**	0.05	[0.00,0.14]
PLANNING	AI - NAI	1.35	0.62	133	2.16	0.033	*	0.03	[0.00,0.10]
REFLECTION	AI - NAI	-0.55	0.62	133	-0.88	0.380	ns	0.00	[0.00,0.00]

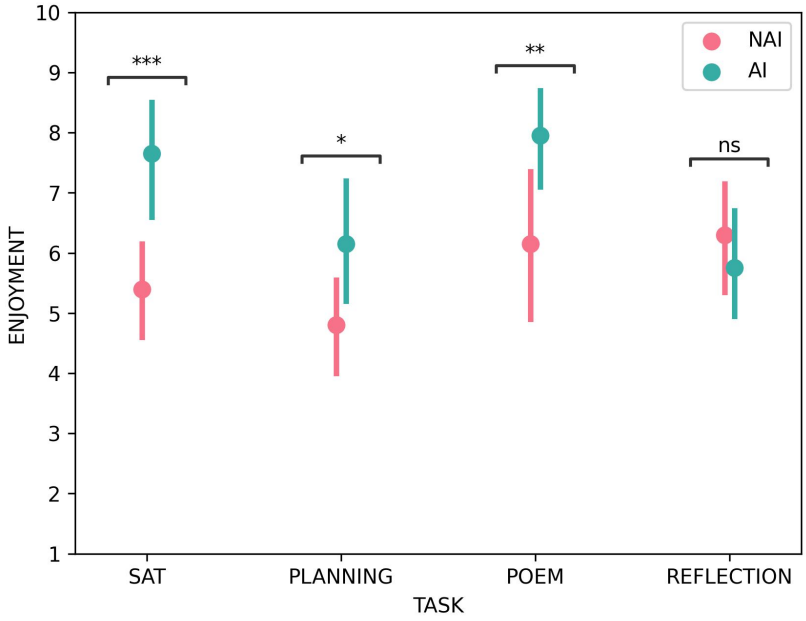


Figure 5.8: ENJOYMENT between CONDITION across TASK. While SAT and POEM demonstrated increases in ENJOYMENT with Copilot, no change was found for PLANNING or REFLECTION.

RQ5-FEELING-C Results

Results are shown in Table 5.15 and Figure 5.9: change in self-reported enjoyment with Copilot was higher for the all of the tasks as compared to REFLECTION (SAT - REFLECTION: $t_{133} = 3.17, p = 0.002, \epsilon_p^2 = 0.06$; POEM - REFLECTION: $t_{133} = 2.66, p = 0.009, \epsilon_p^2 = 0.04$, PLANNING - REFLECTION: $t_{133} = 2.15, p = 0.033, \epsilon_p^2 = 0.03$).

RQ5 Results Summary

These results mirror those of TLX, indicating that, although Copilot provided tangible benefits both in the purely objective tasks (SAT, PLANNING) as well in a creative task (POEM), it did not have any benefits during the episodic memory task (REFLECTION).

5.4.6 RQ5-FEELING Results: Qualitative Evaluation

After each task, users were asked to write optional comments response to their overall experience with the task and the usefulness of the AI tool.

Table 5.15: Contrast results comparing AI versus NAI across TASK. All levels of TASK showed higher ENJOYMENT in AI versus NAI as compared to REFLECTION.

Contrast	Effect	Est.	SE	df	t	p	sig.	ϵ_p^2	ϵ_p^2 CI
SAT - REFLECTION	AI - NAI	2.80	0.88	133	3.17	0.002	**	0.06	[0.01,0.16]
POEM - REFLECTION	AI - NAI	2.35	0.88	133	2.66	0.009	**	0.04	[0.00,0.13]
PLANNING - REFLECTION	AI - NAI	1.90	0.88	133	2.15	0.033	*	0.03	[0.00,0.10]
PLANNING - SAT	AI - NAI	-0.90	0.88	133	-1.02	0.310	ns	0.00	[0.00,0.03]
PLANNING - POEM	AI - NAI	-0.45	0.88	133	-0.51	0.611	ns	0.00	[0.00,0.00]
POEM - SAT	AI - NAI	-0.45	0.88	133	-0.51	0.611	ns	0.00	[0.00,0.00]

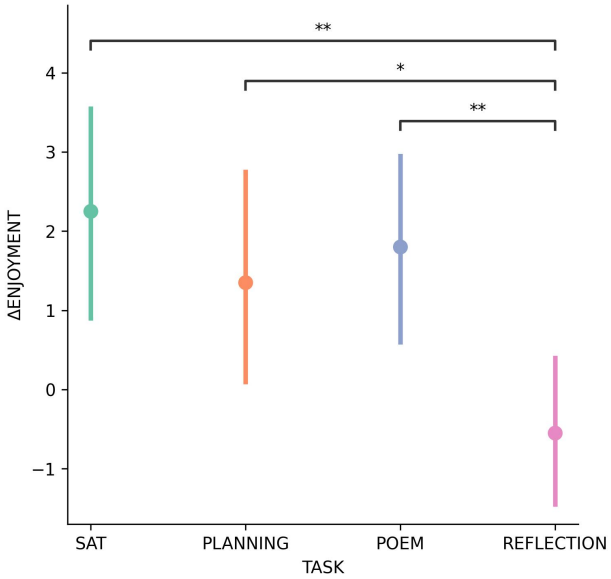


Figure 5.9: Effect of Copilot on ENJOYMENT scores compared across TASK. Similar to the changes in TLX, ENJOYMENT increased significantly with Copilot in the all tasks as compared REFLECTION.

Reading comprehension

As expected, most users found Copilot exceptionally helpful in completing the SAT reading comprehension questions. LLMs perform well with highly structured tasks such as reading a passage and answering multiple choice questions about it. However, not all users trusted that Copilot would be accurate, with user 3 stating that "*I would not want to use the AI tool for such a task because I feel like I would then not put in the effort of checking if the answers given are correct and then I would later on be in self doubt about whether or not the answers were correct*". This lack of trust reduced the likelihood that they might benefit from access to an LLM, even for tasks in which the tool shines.

Planning

User 19 succinctly puts it: "*the AI helps a lot with idea generation that can be worked on*", essentially saying that Copilot was especially helpful in generating ideas and content that could then be refined by the user. However, as other users found, in order to benefit from the generative capabilities of the LLM, a basic understanding of its functionality was necessary. User 12 found that "*the tool refuses to look up specific information I requested and repeatedly came back with generic responses despite being asked to 'be specific'. It was more frustrating than helpful after adopting its initial response as I end up combating with AI to get the information I want*".

Poem

Most people had little experience writing poems or didn't like writing them, meaning that Copilot was especially useful in helping them complete the task given the strict time constraints. However, some users felt that they were of a lower quality, with user 15 stating that "*Having AI for this task was helpful but made the whole ordeal quite boring and the poem, in the end, was not representative of my own feelings and emotions. While it was easier, I did feel like using AI for this kind of assignment yields quite ordinary pieces of work*".

Reflection

Similar to the poem task, users found that Copilot was ineffective in helping them write about their personal experiences and feelings in relation to art. However, one unique advantage the LLM tool provided was the ability to access information when writing the personal reflection, with user 10 finding that "*The tool definitely helped in giving a brief introduction to the album which would have required additional research on my part*".

Trends

These comments reveal that Copilot was especially helpful in a generative capacity, creating drafts or providing information that could then be refined when completing the task. However, multiple factors mitigated the potential benefits of Copilot: a lack of trust in Copilot's answers, a lack of understanding of its functionality, difficulties with iterating on content, and its inability to interact

with or produce personal content. The six minute time constraint for each task also played a role in mitigating Copilot's positive impact on a user. Many users felt that the significant time lag between prompting the tool and receiving a response slowed them down, such that the time-saving benefit of Copilot was diminished by the system's technological constraints given the short time-span of each trial.

5.5 Discussion

Self-reported WORKLOAD, QUALITY, and ENJOYMENT

Regarding the self-reported measures, Copilot's overall effect on users was as-expected for the objective tasks within the gradient of subjectivity: with Copilot, users reported decreased TLX workload and increased ENJOYMENT in SAT and PLANNING; this was coupled with increases in QUALITY. On the opposing end of the subjectivity gradient we likewise found expected results: for REFLECTION, participants reported no tangible changes as a consequence of Copilot use, nor was there a measured change in PERFORMANCE.

Compared to the other results, POEM produced a set of somewhat unexpected findings: namely, a large decrease in TLX workload coupled with an increase in ENJOYMENT. Were initially surprised with these results given the high degree of subjectivity in POEM. However, based on user comments, we believe that this result is partially due to the fact that our users were not used to writing poems; that is, Copilot's ability to produce a significant quantity of reasonable output nearly instantly made the task both easier and more enjoyable. This finding mirrors other work that has indicated that AI-related tools provide the most benefit to the least experienced users [206]. Given that the participants were novice poetry writers, we would caution extrapolation of this finding to the full set of creative domains, and encourage follow-up studies exploring the population of creative users in more depth. Further, no change in output quality was observed in POEM.

fNIRS

Given the decreases in TLX workload for three of the tasks when using Copilot, we were slightly surprised to see a disparity in terms of no findings in the fNIRS data to a similar regard. Of note, however, is that although our study tasks certainly required users' effort, none of them required an

extreme amount of mental workload (along the lines of the NBack task, for instance [205]); that is, tasks which require higher levels of mental effort under the baseline condition may be necessary in order to distinguish levels of prefrontal cortex activation as reflected in VLFO measurements. However, machine learning results presented a more nuanced picture in that actionable patterns are measurable between `AI` and `NAI` within `SAT`.

Another interesting finding regarding `fNIRS` was an increase in activation of the right PFC during `REFLECTION` as compared to `SAT` and `PLANNING`, irrespective of Copilot use. This result likely stems from the `REFLECTION` task's engagement of different underlying psycho-physiological state: that of self-reflection and autobiographical episodic memory retrieval. As discussed earlier, these states have been shown to increase prefrontal activation [29], and specifically have been linked to right prefrontal activation [13, 207]. And more broadly, self-reflection, self-referential states, and episodic memory activation have been linked to the larger Default Mode Network (DMN) [208]. Thus, in conjunction with our `TLX`, `QUALITY`, and `ENJOYMENT` results, our neural finding implies that the helpfulness of AI assistants decreases in response to increased levels of activation of episodic memory; it is also possible that this link is related more broadly to DMN activation. However, it is likewise important to note that, despite the strong statistical finding, more work is necessary towards developing operational distinctions related to episodic memory within a BCI context using fNIRS measuring the PFC.

Other Physiological Results

Given that there were no significant effects related to `HR`, `HRV`, or `EDA`, we can conclude that neither the effects of Copilot use, nor tasks across the gradient of subjectivity, are extreme in the physiological domain outside of the brain.

5.5.1 Limitations

Trial Time

Constraints related to the fNIRS device require our experimental time frame to be relatively short (75 minutes total per user). Given that our study design involved four tasks and post-task surveys, this meant a limitation of six minutes per task, which users reported as a relatively tight window in

which to complete them. For comprehensive follow-up work, future studies might therefore provide more in-depth tasks, selecting only one or two tasks from the gradient of subjectivity.

Copilot Training

Similar to trial time issues, our initial training task was only roughly six minutes in length. We further found during the study that even after some users had claimed to have completed the training task, they were still unfamiliar with some of the basic functionality of the Copilot assistant, which decreased their ability to perform well with it. Therefore, a more formalized test to evaluate subjects' understanding of the capabilities of the LLM model would be quite useful after the training phase of the study.

`QUALITY` Evaluation

While `QUALITY` score for `SAT` is objective, it was subjective for the other three tasks and so is invariably a less reliable metric for these. We discovered that the more precise and clear we designed the task rubric - which would be essential for completely objective grading - the more advantage was given to Copilot in terms of making the task trivial. Therefore, for future studies testing more subjective tasks, more comprehensive metrics of evaluation should be designed.

Only Prefrontal Measures

Although having a window into the brain is invaluable for this kind of work, we are limited in the fact that our measures are only on the PFC - further studies with broader capacity to investigate fully the expression in the brain of using LLM tools would be more beneficial to understanding their effects on users.

5.6 Implications for BCI

Our work explores the premise that hemodynamic activity in the prefrontal cortex might be altered sufficiently by LLM-use such that we might have the basis for accessing information related to human state in this task; based on our results, future work may explore reading comprehension with the context of LLM use as a potential aspect for state-classification; likewise, episodic memory

in the right aspect of the prefrontal cortex may present a viable vector of classification for future interfaces.

5.6.1 Information Processing

At one end of the gradient `SAT` is done immediately by Copilot and with perfect accuracy. Not surprisingly, users' workload decreased, and comments show that this change. More broadly, however, the formalization of this result allows us to infer that the future of human effort related to reading-related tasks will shift from the need to comprehend aspects of a text on one's own to instead determining the questions one wants to ask about a given text. As such, the need to quantify the specific effects on the human of the transformational aspect of moving from wading through a deluge of detailed information into spaces of higher abstraction when using LLMs will take center stage in future human-centric studies. Further, the machine learning finding from the neural data indicates that the information processing domain and interaction with LLM tools may be a fertile ground for future research involving applied, real-time BCIs. But to fully realize the potential of these tools, designers will also need to address the issue of trust, ensuring that participants trust that the answers provided by the LLM are accurate as many users indicated in this study that they did not.

5.6.2 Idea Generation

In the middle of the gradient, `PLANNING` and `POEM` relied on the user answering a prompt after generating ideas that were not necessarily related to their personal experiences. The decrease in `TLX` when assisted by Copilot signals that LLMs could be useful in idea generation tasks. These findings, however, did also not establish themselves with a concrete physiological basis, either through statistical methods or machine learning. And, even though the utility of LLMs in assisting users with generating ideas due to their access to massive libraries of content is well-understood, further investigation is necessary into how users can best leverage those capabilities. Some participants signaled a lack of understanding in how to generate ideas that were specific to their needs but also novel, showing a significant gap in the usability of LLMs today.

5.6.3 Subjective Experiences

At the other end of the gradient, it is clear that tasks that depend largely on the individual experiences of the user are as-of-yet effectively impenetrable for Copilot, which knows nothing about the details of a given user's experience. Although we expect that future versions of such systems will be able to be trained with user-specific data, it remains to be seen if such 'personalized' tools will be able to be able to meaningfully assist users in tasks requiring self-reflection and DMN activation. Such contexts include therapy, creative writing based on personal history, or highly individualized decision making. Testing the effectiveness of these personalized tools in such contexts would hinge on their ability to integrate and understand nuanced personal experiences, emotions, and the individual's cognitive patterns, and remains a fertile domain for future work.

5.7 Conclusion

I tested Copilot, an interactive LLM-based AI assistant, using a multimodal set of measurement techniques including prefrontal cortex activation via fNIRS in terms of its effects on user states through a variety of tasks designed along a gradient of subjectivity intended to become increasingly difficult for the assistant. Results indicate that for tasks which are challenging yet tightly constrained overall in terms of objectivity, users benefit in terms of decreases in self-reported mental workload and increases in reported enjoyment and objective performance. For creative tasks for new users with more subjective criteria for success (`POEM`), Copilot produced very similar gains to the more objective tasks, despite our expectations; however, these results should be interpreted with caution as participants may not have approached this purely creative task with the same level of rigor as the others. In purely reading-comprehension tasks (`SAT`), the distinction between neural activation as measured by fNIRS was not statistically significant, but is nevertheless operational for real-time implicit BCI. Lastly, we found that Copilot was not able to assist users meaningfully in tasks which require primarily subjective material (`REFLECTION`), and that while brain measurement via fNIRS indicated larger prefrontal cortex activation during this task than the others, likely due to episodic memory retrieval and potentially broader DMN activation, based on classification f1-scores the finding is not currently usable for real-time implicit BCI, but potentially might be in the future.

I concretely specify the activation of neural states related to episodic memory as a shortcoming of artificial agents[5], and more tentatively indicate that the lack of the assistant's ability to help users may align with a broader activity of the DMN. While this is an initial study with a single LLM-based AI tool, more will be required in the domain of evaluation of effects of AI assistants on human users.

5.8 Transitioning to Complex Decision-Making

The next study begins by establishing just such an evaluation by considering an applied complex decision-making task. In this study, we exploit the relatively objective end of the subjectivity gradient; although the current study was able to discern some meaningful differences of neural data using machine learning between with vs. without Copilot conditions within a relatively "extreme" differentiation in the reading comprehension task, the following study takes a step forward towards applying such a task in a larger, more complex task that is intended to mimic a real-world office scenario.

[5]But potential boon for BCI studies

Chapter 6

LLM Tools in Complex Decision-Making

The core idea behind this study was to follow-up from the first Copilot study with an applied, ecological task design that intended to mimic a real-world office scenario. In this environment, how does PFC activation change with Copilot use? Is this change operationalizable in an implicit BCI context? And, more broadly, are any potential prefrontal changes related to changes in self-reported mental workload, to users' feelings in terms of the valence-arousal space, or perhaps some combination of the two? Lastly, does previous experience with AI tools have any effect across measures? To explore these ideas, the second study in this project assesses the effects of LLMs on human users during a complex decision-making task. In this study Microsoft 365 Copilot is the LLM interface by which participants have AI assistance. For the study tasks, participants were instructed to imagine that they were the CEO of a small business dealing with significant problems, and their job was to rank-order three provided possible solution strategies for the company's issues in multiple dimensions, including user impact, financial and temporal costs, and risk.

6.1 Materials and Methods

6.1.1 Microsoft 365 Copilot

For this study we used Copilot for Microsoft 365. For convenience, I will refer to this tool also as "Copilot", however, despite that this is similar to the Copilot tool used in the previous study, is not the same. In the previous study, that Copilot tool was integrated into the Word interface; here,

it is a standalone interactive prompt, similar to ChatGPT. Furthermore, this version of Copilot has more comprehensive abilities than the former version; with custom commands, one can have Copilot reference files in one's OneDrive folder. This is where the files for the study were located, so Copilot could effectively "read" the documents. However, this functionality was not perfect. Often, Copilot would fail to find documents. Or, if a document was found, Copilot might not "read" it in its entirety. Typically, prompting the tool again with the same initial prompt would work to solve these usability issues. Given that these issues were likely to cause problems for participants, we encoded the training information into a document which the participant had the opportunity to engage with; a quiz was presented which they were required to pass prior to beginning the study.

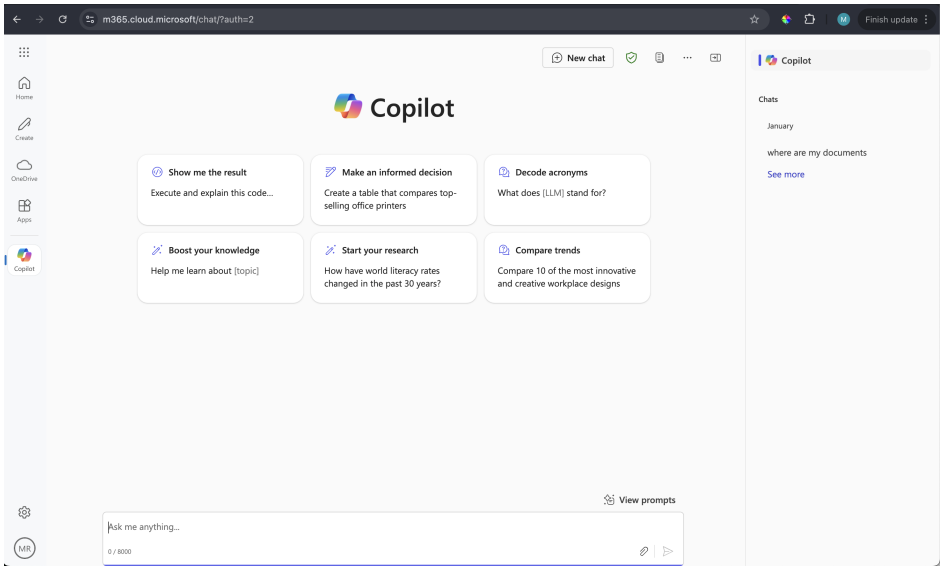


Figure 6.1: Microsoft 365 Copilot

6.1.2 Study Tasks

We developed a custom set of two study tasks, both variants of the same task designed to be equally difficult. Although these tasks are related to a business environment, they were intended to be effectively complex decision-making tasks approachable to any undergraduate student. For each task, the participant was given a set of four documents. The first document was a background and overview of the (fictitious) company of which they were the new Chief Executive Officer (CEO). This document outlined the struggles facing the company in recent years, and contained space for the participant to rank-order the proposals in a variety of categories. The other three documents

were proposals with potential strategies for solving the company's problems. These proposals all followed the same outline: summary of the specific problem that the proposal is attempting to address, a summary of the strategy chosen, an overview of implementation plan, expected temporal, financial, and labor costs, anticipated impact on users, and potential risks and mitigation strategies. Participants were instructed to rank order each of the three proposals in terms of: user impact, financial cost, temporal cost, risk, and overall. For each category, in addition to the rank ordering, participants were instructed to write a brief (1-2 sentence rationalization of their choice). The full text of all documents provided to users is located in Appendix E.

Task Evaluations

We designed the tasks such that, although the delineations of some rank-orderings within specific categories were not extremely clear, the overall evaluations for the two tasks have concrete answers. Therefore, we were able to grade participants as correct or incorrect for their rank-orderings. Participants who did not manage to fill out the rank orderings for a given task were marked as incorrect.

6.1.3 Valence-Arousal Analysis

We created a custom assessment based on the standard Valence-Arousal model. After each task, we provided participants with an image of the Valence-Arousal model used by Gerber [209]. Participants were asked to click whichever location on the image best represented their state during the previous task. This data was used to analyze differences in state between tasks.

6.1.4 Study Structure

All users signed informed consent documents prior to beginning the study, and completed an initial form regarding their prior experience with LLM tools (0-100). Participants then did a 5 minute training to familiarize themselves with the Copilot 365 assistant. As mentioned above, nuances of the interface (file access, issues with information finding in documents, etc.) were presented in this training session; participants were given a chance to ask questions of the experimenter if anything was unclear. After this, participants were required to successfully complete a quiz indicating their understanding of the functionality was sufficient to continue; those who failed the

quiz were instructed to go back to the training document and/or discuss with the experimenter any confusion. After passing the quiz, participants continued with the tasks. They did two tasks, each of which included a rest period of 1 minute, 25 minutes of task time, and then post-task surveys. In the with-Copilot condition, participants were instructed to use Copilot as an assistant towards producing the best possible output that they could; in the without-Copilot condition, participants were simply told to open the documents and begin the task. After successful completion of the study, participants were debriefed and compensated with a $25 Amazon gift card. Both the tasks, and which condition would allow for the Copilot assistant, were randomized for each participant. Given the long (25m) task length, this enables us to compare if and how strongly order effects confound the other findings.

6.1.5 Data Collection and Preprocessing

Demographics

We ran our study on 37 healthy participants (age μ=23.7, σ=8.6).

Exclusions from Physiological Data

Seven users were excluded from data analysis: one due to difficulty fitting the headband, one refused to wear the headband, one did not understand the tasks, one only copy-pasted data in the with-AI condition, and 3 due to poor data quality.

`EXPERIENCE` Factor

Based on the initial survey of experience with AI tools, which asked participants to rate their familiarity with AI tools from 0-100, we subdivided the population into two groups: 15 participants (μ=9.0, σ=8.5) were classified as novice AI users, and the remaining 15 participants (μ=46.9, σ=22.8) were categorized as experienced AI users. This between-subjects factor will be referred to in the results as `EXPERIENCE` (`EXP`).

fNIRS System and Preprocessing

The same hardware and methods were used to collect and preprocess this data as in 5.3.3. Note that the same methods were used both for preprocessing in terms of statistical analysis, as well as for machine learning.

NASA-TLX

The `TLX` preprocessing was performed as in 5.3.3.

6.1.6 Statistical Methods

In contrast to the previous study, in this case we have a 2×2×2 mixed-design. Factors are: `CONDITION` (`COND`), whether the participant was using Copilot for the given task, a within-subjects factor with two levels of `AI` and `NAI`; `ORDER`, a between-subjects factor with two levels indicating the ordering of the `AI` presentation, (`NAI` → `AI` and `AI` → `NAI`); and `EXPERIENCE` (`EXP`), a between subjects factor with two levels relating to AI use (`Q1` for novice AI users, and `Q2` for those more experienced)[1]. I use LMMs to analyze the data, with random intercepts for `PID`. No random effects were done across trials given that each participant only did two trials. Therefore, our template `R` formula for is

$$\texttt{DV} \sim \texttt{CONDITION} * \texttt{ORDER} * \texttt{EXPERIENCE} + (\texttt{1}|\texttt{PID}) \tag{6.1}$$

where `DV` is one of `TLX`, `fNIRS`, or `VA` (Valence-Arousal). Post-hoc tests are performed within levels of between-subjects grouping factors (e.g. test for `AI` vs. `NAI` within levels of `ORDER`); post-hoc tests which showed significance in both ordering categories were followed up with custom contrasts comparing the difference between them, to be able to determine if ordering levels produced substantially different effects. Lastly, given that `CORRECTNESS` is a binary variable, for this `DV` we used Generalized Linear Mixed Model (GLMM) analysis [210, 211], with the same factors as above. As in the prior study, Bonferroni correction is used where appropriate: α is set to 0.025 accounting for two measures of `DSI` and `DS`ϕ within each of the `L` and `R` prefrontal cortices. I likewise set α to 0.025 for `x` and `y` coordinate models for the `VA` analysis.

[1]With only 30 participants, I note that the sample size for the interaction among all factors is relatively small; given the exploratory nature of the analysis, I present these higher-order interactions, but acknowledge the need for caution in interpretation: our primary focus remains on main effects and two-way interactions.

6.1.7 Machine Learning

To test the operational ability to distinguish between with and without Copilot conditions, I performed LOO-CV on the dataset. Models used in this study were the same as in 5.3.5: SVC, SVM, RF, MLP, and KNN; I likewise report the best results overall per-model and within levels of `ORDER`. Similarly, results when averaging across participants are specified as `STANDARD`, and when considering the best results per-participant regardless of model as `OPTIMIZED`. Given the shorter overall results section, in this study the machine learning results are presented at the end of the overall results, rather within the results discussing the neural data.

6.2 Results

6.2.1 NASA-TLX Results

See Table 6.1. Although none of the individual factors of the model showed significance, the interaction of `CONDITION` × `ORDER` did ($F_{1,25.0} = 5.39, p = 0.029, \epsilon_p^2 = 0.14$). Clearly, the indication is that there is an order effect, however, that this effect depends on levels of `CONDITION`. Therefore,

Table 6.1: `TLX` self-reported workload.

Factor	df1	df2	F	p	sig.	ϵ_p^2	ϵ_p^2 CI
COND	1	25.0	0.71	0.408	ns	0.00	[0.00,0.00]
ORDER	1	25.0	0.42	0.524	ns	0.00	[0.00,0.00]
EXP	1	25.0	1.49	0.234	ns	0.02	[0.00,0.18]
COND × ORDER	1	25.0	5.39	0.029	*	0.14	[0.00,0.36]
COND × EXP	1	25.0	2.00	0.170	ns	0.04	[0.00,0.22]
ORDER × EXP	1	25.0	0.29	0.596	ns	0.00	[0.00,0.00]
COND × ORDER × EXP	1	25.0	1.77	0.196	ns	0.03	[0.00,0.20]

we ran post-hoc contrasts on this interaction effect. Specifically, we tested the effect of `CONDITION` within levels of `ORDER`. The contrast was only significant when participants experienced the `ANI`→`AI` order ($t_{25.00} = 2.11, p = 0.045, \epsilon_p^2 = 0.12$), with lower workload reported in `AI` as compared to `NAI`.

These results indicate that self-reported workload remains unchanged when experiencing the `AI` condition first, but it decreases if `AI` comes second. These results suggest that the workload-reducing effect of the `AI` condition is substantial, however it also suggests that without prior familiarity of

Table 6.2: TLX ΔAI within ORDER.

Order	Contrast	Estimate	SE	df	t	p	sig.	ϵ_p^2	ϵ_p^2 CI
AI → NAI	AI - NAI	1.14	1.02	25.00	1.12	0.275	ns	0.01	[0.0,0.19]
NAI → AI	AI - NAI	-2.43	1.15	25.00	-2.11	0.045	*	0.12	[0.0,0.37]

such a task, AI may introduce more complexity to the extent that it does not sufficiently help the user.

6.2.2 fNIRS Results

See Table 6.3. Initial inspection of the data shows a similar picture to that of NASA-TLX. That is, none of the main factors on their own were significant across measures. Further, COND × ORDER demonstrated significance in both L measures: L DSI ($F_{1,28.0} = 20.38, p < 0.001, \epsilon_p^2 = 0.40$), and L DSϕ ($F_{1,28.0} = 7.15, p = 0.012, \epsilon_p^2 = 0.18$), as well as in R DSI ($F_{1,28.0} = 17.67, p < 0.001, \epsilon_p^2 = 0.37$). Further, although after α correction to 0.025 considering the DSI and DSϕ measurements per-probe in the same research question R DSϕ is not significant, its medium effect size still warrants consideration ($F_{1,28.0} = 4.35, p = 0.046, \epsilon_p^2 = 0.10$). In addition to these omnibus results mirroring the effects seen in TLX, the fNIRS data also show a significant interaction effect between CONDITION × EXPERIENCE ($F_{1,28.0} = 10.67, p = 0.003, \epsilon_p^2 = 0.25$. I will continue by discussing the post-hoc results related to these findings first, and then follow by considering the finding of CONDITION × EXPERIENCE[2].

I initially ran the same post-hoc tests checking for the effects of AI - NAI within levels of ORDER as was done for TLX (Table 6.4). Although the effect was significant within L DSI for the AI→NAI order ($t_{26.0} = 3.49, p = 0.002, \epsilon_p^2 = 0.29$, it was *also* significant in the opposing direction, for L DSI in the NAI→AI order ($t_{26.0} = -0.44, p = 0.017, \epsilon_p^2 = 0.17$. To test the effects directly, I ran custom contrasts between the change in the second and first task within each. The result was not significant ($t_{26.0} = -0.97, p = 0.341, \epsilon_p^2 = 0.00$), indicating that the order effect was similarly strong in terms of the left PFC activity regardless of CONDITION. As well, no effects was visible in terms of $DS\phi$.

The right probe, however, showed a different result. Firstly, neither the DSI nor DSϕ measurements showed significance within the AI→NAI order. However, *both* did for the NAI→AI order (R DSI: $t_{26.00} = 4.69, p < 0.001, \epsilon_p^2 = 0.44$; R DSϕ: $t_{26.0} = 2.40, p = 0.024, \epsilon_p^2 = 0.15$). Recall that an

[2]Although there is a similar finding related to ORDER × EXPERIENCE, this factor is unrelated to our research question, and is not duplicated outside of L DSI, thus it is not analyzed further.

Table 6.3: Results of modeling fNIRS for all four combinations of [L, R], and [DSI, DSϕ]. These results indicate significant activation changes related to the interaction of CONDITION × ORDER for both the L (DSI and DSϕ) and R (only DSI) PFC. In the left PFC, the interaction of CONDITION × EXPERIENCE is likewise significant. Note that p.sig considers adjusted α of 0.025, correcting across tests for DSI and DSϕ within each side of L and R, separately.

Side	Meas	Factor	df1	df2	F	p	sig.	ϵ_p^2	ϵ_p^2 CI
L	DSI	COND	1	28.0	1.68	0.206	ns	0.02	[0.00,0.18]
L	DSI	ORDER	1	28.0	0.57	0.457	ns	0.00	[0.00,0.00]
L	DSI	EXP	1	28.0	0.11	0.740	ns	0.00	[0.00,0.00]
L	DSI	COND × ORDER	1	28.0	20.38	<0.001	***	0.40	[0.17,0.58]
L	DSI	COND × EXP	1	28.0	10.67	0.003	**	0.25	[0.05,0.45]
L	DSI	ORDER × EXP	1	28.0	4.57	0.041	ns	0.11	[0.00,0.31]
L	DSI	COND × ORDER × EXP	1	28.0	2.79	0.106	ns	0.06	[0.00,0.24]
L	DSϕ	COND	1	28.0	0.09	0.764	ns	0.00	[0.00,0.00]
L	DSϕ	ORDER	1	28.0	0.73	0.399	ns	0.00	[0.00,0.00]
L	DSϕ	EXP	1	28.0	0.31	0.584	ns	0.00	[0.00,0.00]
L	DSϕ	COND × ORDER	1	28.0	7.15	0.012	*	0.18	[0.02,0.38]
L	DSϕ	COND × EXP	1	28.0	2.21	0.148	ns	0.04	[0.00,0.21]
L	DSϕ	ORDER × EXP	1	28.0	2.45	0.129	ns	0.05	[0.00,0.22]
L	DSϕ	COND × ORDER × EXP	1	28.0	0.21	0.650	ns	0.00	[0.00,0.00]
R	DSI	COND	1	28.0	3.69	0.065	ns	0.08	[0.00,0.28]
R	DSI	ORDER	1	28.0	3.20	0.084	ns	0.07	[0.00,0.26]
R	DSI	EXP	1	28.0	1.13	0.296	ns	0.00	[0.00,0.11]
R	DSI	COND × ORDER	1	28.0	17.67	<0.001	***	0.37	[0.14,0.55]
R	DSI	COND × EXP	1	28.0	1.85	0.184	ns	0.03	[0.00,0.19]
R	DSI	ORDER × EXP	1	28.0	3.68	0.065	ns	0.08	[0.00,0.28]
R	DSI	COND × ORDER × EXP	1	28.0	6.71	0.015	ns	0.16	[0.01,0.37]
R	DSϕ	COND	1	28.0	1.99	0.169	ns	0.03	[0.00,0.20]
R	DSϕ	ORDER	1	28.0	2.45	0.129	ns	0.05	[0.00,0.22]
R	DSϕ	EXP	1	28.0	1.28	0.268	ns	0.01	[0.00,0.14]
R	DSϕ	COND × ORDER	1	28.0	4.35	0.046	ns	0.10	[0.00,0.30]
R	DSϕ	COND × EXP	1	28.0	0.61	0.441	ns	0.00	[0.00,0.00]
R	DSϕ	ORDER × EXP	1	28.0	0.83	0.370	ns	0.00	[0.00,0.00]
R	DSϕ	COND × ORDER × EXP	1	28.0	0.08	0.778	ns	0.00	[0.00,0.00]

increase in total power of the Δ[HbD] waveband indicates an decrease in neural activation. Therefore, we can conclude that the NAI→AI order produced a significant decrease in activation of the right PFC as measured by both DSI and DSϕ. This distinction is a unique counterpoint to the left PFC change largely occurring as a consequence of the order effect. Although firm delineations between L and R PFC activation have not yet been drawn, Goel has posited the potential for certain aspects of R PFC function to interrelate with indeterminacy tolerance, specifically in the context of an information representation system (contrasting with the L PFC, which decodes the representations), and/or

Table 6.4: Post-hoc contrasts of AI - NAI within levels of ORDER. Given the lack of significant omnibus test for R DSϕ, no tests were done on that combination. Results show significant L and R DSI for the ORDER level NAI→AI. Note that α is set to 0.025, given the comparisons of DSI and DSϕ within each side.

Side	Meas	Order	Contrast	Est.	SE	df	t	p	sig.	ϵ_p^2	ϵ_p^2 CI
L	DSI	AI→NAI	AI - NAI	-0.44	0.17	26.00	-2.56	0.017	*	0.17	[0.00,0.42]
L	DSI	NAI→AI	AI - NAI	0.70	0.20	26.00	3.49	0.002	**	0.29	[0.04,0.53]
L	DSϕ	AI→NAI	AI - NAI	-0.37	0.16	26.00	-2.36	0.026	ns	0.14	[0.00,0.39]
L	DSϕ	NAI→AI	AI - NAI	0.40	0.18	26.00	2.24	0.034	ns	0.13	[0.00,0.38]
R	DSI	AI→NAI	AI - NAI	-0.35	0.18	26.00	-1.94	0.064	ns	0.09	[0.00,0.34]
R	DSI	NAI→AI	AI - NAI	0.99	0.21	26.00	4.69	<0.001	***	0.44	[0.15,0.64]
R	DSϕ	AI→NAI	AI - NAI	-0.04	0.29	26.00	-0.15	0.882	ns	0.00	[0.00,0.00]
R	DSϕ	NAI→AI	AI - NAI	0.80	0.33	26.00	2.40	0.024	*	0.15	[0.00,0.40]

that it might have a role in restraining the left PFC from settling on preliminary inferences [212]. Within that context, we might consider that the decrease in R PFC activity during the AI condition (within the NAI→AI order) could be due to a relaxing of the processes which underlie hesitance to settle on inferences from knowledge representations; that is, the AI allows one to more easily make inferences about the information at hand. This finding suggests important implications for how AI tools may fundamentally alter cognitive processing strategies during complex decision-making tasks, potentially by reducing cognitive inhibition processes that normally prevent premature conclusions; rather than a net-negative, this might be considered within the context of more easily enabling users to come to conclusions - whether those are reasonable ones, however, is outside of the scope of the neural information measured, and will be discussed more in the finding related to CORRECTNESS.

Returning to the finding of CONDITION x EXPERIENCE within the L DSI measure, post-hoc tests were run within levels of EXPERIENCE of the contrast AI - NAI. Results are in Table 6.5. Overall, irrespective of the effect of ORDER, this contrast indicates that those in the less-experienced cohort experienced lower-prefrontal activation in the L PFC during AI as compared to NAI ($t_{26.00} = 2.63, p = 0.014, \epsilon_p^2 = 0.18$). This finding may indicate more reliance on the AI tool on behalf of those who have used AI less frequently. That is, although it is inconclusive *why* this is the case, it is possible that less-experienced users re-frame their approach to the task when AI assistance is available; rather than consider it as a problem to solve independently, they may instead consider it as a process where they primarily need to interpret AI outputs, which then engages different cognitive processes.

Table 6.5: Post-hoc contrasts of `AI - NAI` within levels of `EXPERIENCE` for `L DSI`. There is only a significant result within the group of lesser-experienced users of AI tools.

Side	Meas	Experience	Contrast	Estimate	SE	df	t	p	sig.	ϵ_p^2	ϵ_p^2 CI
L	DSI	Q1	AI - NAI	0.51	0.19	26.00	2.63	0.014	*	0.18	[0.0,0.43]
L	DSI	Q2	AI - NAI	-0.25	0.18	26.00	-1.37	0.183	ns	0.03	[0.0,0.24]

6.2.3 Valence-Arousal Results

See Table 6.6. Significance was found for the `y` (Arousal) axis only, with again the factor interaction of interest as `CONDITION` × `ORDER`. Within-`ORDER` results shown in Table 6.7 and Figure 6.2 indicate that participants who experienced the `AI` condition second had a decreased arousal as compared to when they experienced `NAI` first ($t_{26.00} = 3.56, p = 0.001, \epsilon_p^2 = 0.30$). This is by contrast with the group who experienced `AI` first; in this case there was no effect between conditions. Although a visualization of this grouping is not shown here, it is visualized for completion in Appendix D.

Table 6.6: Results comparing `x` and `y` coordinates. Note that α is set to 0.025 for these values, given that x and y were separate models under the same research question.

Axis	Factor	df1	df2	F	p	sig.	ϵ_p^2	ϵ_p^2 CI
x	COND	1	26.0	0.11	0.740	ns	0.00	[0.00,0.00]
x	ORDER	1	26.0	0.24	0.627	ns	0.00	[0.00,0.00]
x	EXP	1	26.0	0.14	0.712	ns	0.00	[0.00,0.00]
x	COND × ORDER	1	26.0	0.08	0.776	ns	0.00	[0.00,0.00]
x	COND × EXP	1	26.0	3.30	0.081	ns	0.08	[0.00,0.28]
x	ORDER × EXP	1	26.0	1.48	0.234	ns	0.02	[0.00,0.17]
x	COND × ORDER × EXP	1	26.0	0.04	0.843	ns	0.00	[0.00,0.00]
y	COND	1	26.0	2.26	0.145	ns	0.04	[0.00,0.23]
y	ORDER	1	26.0	0.03	0.867	ns	0.00	[0.00,0.00]
y	EXP	1	26.0	0.43	0.520	ns	0.00	[0.00,0.00]
y	COND × ORDER	1	26.0	15.05	0.001	***	0.34	[0.11,0.54]
y	COND × EXP	1	26.0	4.01	0.056	ns	0.10	[0.00,0.30]
y	ORDER × EXP	1	26.0	5.18	0.031	ns	0.13	[0.00,0.34]
y	COND × ORDER × EXP	1	26.0	0.85	0.365	ns	0.00	[0.00,0.00]

6.2.4 Correctness Results

See Table 6.8. Results indicate no significant effects on correctness of using the `AI` tool. Despite the observed differences in `fNIRS`, `TLX`, and `VA` measures (under certain circumstances), results show that

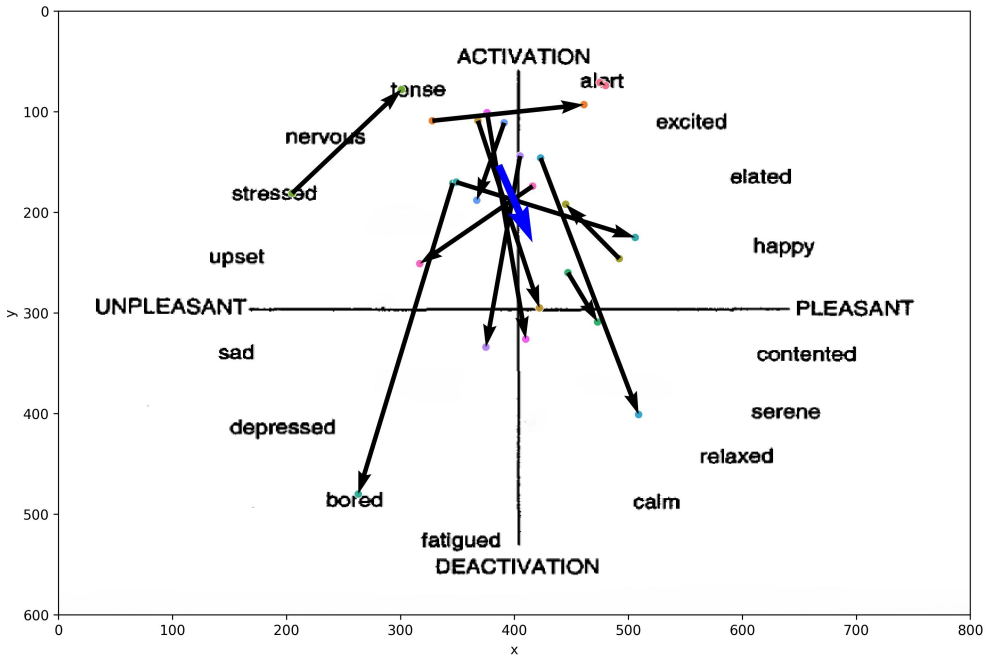


Figure 6.2: After each task, participants selected the spot on the circumplex Valence-Arousal model which they felt best related to their state of mind during the task. This group of participants is those who experienced `AI` second; arrows begin in the `NAI` condition and point to the `AI` condition. Blue arrow represents the average of all normalized arrows across participants', scaled to the mean length across participants.

Table 6.7: Within `ORDER` comparisons of `AI - NAI` in the arousal dimension of the circumplex model of affect. When participants performed the `AI` task second, their arousal decreased slightly, indicating a relaxation response.

Axis	Order	Contrast	Estimate	SE	df	t	p	sig.	ϵ_p^2	ϵ_p^2 CI
y	AI→NAI	AI - NAI	-48.61	26.77	26.00	-1.82	0.081	ns	0.08	[0.0,0.32]
y	NAI→AI	AI - NAI	110.19	30.97	26.00	3.56	0.001	**	0.30	[0.05,0.54]

participants' ability to correctly rank-order the proposals remained similarly equivalent regardless of Copilot assistance or the effect of `ORDER`. This result suggests that, while the Copilot may alter the cognitive processes and subjective experience of decision-making, it does not necessarily translate to measurable performance improvements or drawbacks in this specific complex decision-making task.

6.2.5 Machine Learning Results

Machine learning results are shown in Tables 6.9 and 6.10. Despite the promising statistical results indicating differential brain patterns between tasks, no `STANDARD` results are present which could

Table 6.8: Correctness results.

Factor	Chisq	df	p	sig.
COND	0.00	1	1.000	ns
ORDER	0.06	1	0.803	ns
EXP	2.16	1	0.142	ns
COND × ORDER	0.00	1	1.000	ns
COND × EXP	1.40	1	0.237	ns
ORDER × EXP	1.15	1	0.285	ns
COND × ORDER × EXP	1.08	1	0.299	ns

Table 6.9: `STANDARD` LOO-CV machine learning results per-model, where each model's results per-participant is averaged. Support columns indicate the average number of samples for each class in the testing set.

Model	F1-score	Support A	Support B
RF	0.505	72.033	72.267
KNN	0.501	72.033	72.267
MLP	0.491	72.033	72.267
SVM	0.464	72.033	72.267
SVC	0.447	72.033	72.267

indicate usable patterns for implicit BCI work. The best overall result per-model is RF, with 50.5% F1 across participants; within groups of `ORDER`, the results are still similar, with the best in `AI`→`NAI` as 52.7%, and `NAI`→`AI` as 48.9%. The `OPTIMIZED` results are slightly better: across each group of 15 participants (`AI`→`NAI`: $\mu = 57.8\%, \sigma = 0.08$; `NAI`→`AI`: $\mu = 55.0\%, \sigma = 0.08$), and across all participants: $\mu = 56.4\%, \sigma = 0.08$, however all results still fall below 60%, and thus are likely not usable in implicit BCI contexts.

A variety of reasons are possible in terms of why these machine learning results are what they are. Firstly, in the frame of reference of the previous study, the difficulty of this task with and without AI is still relatively high; the level of distinction in workload as potentially changed by Copilot isn't dramatic as it was in the SAT task. By implication, in terms of mental workload differentiations based on AI vs. NAI tools, it is possible that such interventions may be challenging to deploy in the field except for tasks for which AI has an extreme effect.

Table 6.10: STANDARD LOO-CV machine learning results within levels of ORDER (given order effects as a potential confound, participant groups are split based on task order). Results are Macro F1 scores averaged across participants. Support columns indicate the average number of samples for each class in the testing set.

Model	Order	F1-score	Support A	Support B
KNN	AI→NAI	0.527	75.2	70.1
RF	AI→NAI	0.517	75.2	70.1
SVM	AI→NAI	0.500	75.2	70.1
MLP	AI→NAI	0.497	75.2	70.1
SVC	AI→NAI	0.468	75.2	70.1
RF	NAI→AI	0.489	67.9	75.2
MLP	NAI→AI	0.483	67.9	75.2
KNN	NAI→AI	0.467	67.9	75.2
SVC	NAI→AI	0.419	67.9	75.2
SVM	NAI→AI	0.416	67.9	75.2

6.3 Conclusion

These results indicate that LLM tools like Copilot may help reduce cognitive load and slightly relax users once they overcome the initial task learning curve, without sacrificing performance. However, the usability of this result towards real-time implicit BCI work uncertain. Overall, in conjunction with the findings from the previous study, there are interesting statistical patterns in the data which shed light on the activation of the PFC in various conditions, both with and without LLM-assistance. And, in this study, the patterns of neural hemodynamics more closely paralleled the self-reported information from TLX and the VA data. However, these patterns remain challenging to consider towards real-time applied BCIs.

6.4 Looking Ahead

While the previous two studies examined the effects of LLM tools on PFC activation in various contexts, they primarily focused on understanding these effects rather than directly applying them in real-time interfaces. Building upon these insights - particularly the observed distinctions between task types and their neural correlates - we transition to the development and evaluation of a fNIRS-based functional real-time implicit BCI system. This next study moves from observation to application, implementing the theoretical framework proposed by Hincks [12] that leverages the

anticorrelated TPN/DMN brain networks as the vector of classification to drive a real-time implicit BCI system.

PART III

Real-Time lPFC-mPFC Based BCI with fNIRS

fNIRS has proven in recent time to be a reliable detector of workload vis-à-vis the PFC for real-time implicit BCIs. But what can be done in terms of application of neural measurements of the PFC beyond mental workload? Whereas the first project in this dissertation focused on understanding LLM use in terms of working episodic memory through multimodal measurements including the PFC, in this study I trained and tested a prototype fNIRS-based BCI interface intended to leverage anticorrelated brain networks towards a first implementation of a memory prosthesis which presents information appropriate to a user's current brain state from moment to moment. This prototype implementation used data from two tasks designed to interface with different brain networks: a creative visualization task intended to engage the Default Mode Network (DMN), and a complex knowledge-worker task to engage the Dorsolateral Prefrontal Cortex (DLPFC). Performance of 71% from leave-one-out cross-validation across participants indicates that such tasks are differentiable, which is promising for the development of future applied fNIRS-based BCI systems. Further, analyses across lateral and medial prefrontal areas indicates potential approaches for future classification.

Chapter 7

Background

This project specifically intends to push the boundaries of implicit BCI with prefrontal cortex measurements using fNIRS. In contrast to other systems which specifically attempt to capture high-workload states [53, 54, 55, 57, 61, 63, 64], we designed our tasks in the paradigm proposed by Hincks [12] attempting to interface with different brain networks: a creative visualization task intended to engage the DMN through the mPFC [213], and a complex knowledge-worker task to engage the TPN through the lPFC [26]. Performance from simulated real-time classification of the data indicates that such tasks are differentiable, which is promising for the development of future applied fNIRS-based BCI systems, both in the context of network-dependent brain-state classification.

In terms of application, our prototype system is motivated to step toward the notion of a general brain-based "assistant" that helps its user recall items by indexing them according to their mental state and presenting relevant information automatically, rather like the "memory prosthesis" first introduced in the work of Rhodes [16] and Lieberman [17], but with passively measured brain state as the storage and retrieval tag. We can envision in the future a brain-based interface which is able to recognize and adapt fluidly to a user's brain state - such a system, as a memory prosthesis, would both be able to store brain states associated with important information, and to provide such information when the user requires it.

Such an associative memory assistant could be useful in a variety of common knowledge worker research tasks. Examples range from examining and organizing text a body of legal documents

for a lawyer, to surveying papers for an academic survey or policy analysis, to businesses analysis for acquisition or valuation. In a conventional filing system, the user could store such items in a bookmarked list as they are reviewed and then retrieve them from it later.

Furthermore, bookmark creation will be able to be automatically generated based on analysis of the brain signal. Bookmarks will then be ordered by how well each one matches the current brain state. Thus, whenever a user sees the list, they will first see those items that they entered while in the same brain state as they are in currently. The rationale is that these might be the most relevant items for the user at the current moment. The benefit is that the system would display them automatically and continuously, without any user effort, without scrolling through a variety of previously stored bookmarks nor having to enter tags explicitly. The filing system index is simply the user's passively measured brain state. Of course, in a more practical system, the filing system would permit other indexes as well.

Beyond the basic low-level interaction speed advantage of having the top bookmarks preselected effortlessly, Gray and Boehm-Davis [214] provide experimental evidence of a direct impact of such rapid, low-level, lightweight interaction on a user's higher level strategy and behavior; it can produce changes well beyond the actual speedup of the improved low-level interaction. They observe that a slight change in an interface can shift subjects from a trial-and-error problem solving approach to a plan-based one. Instead of displaying content near the user's current state, some work suggests that it might be better to display content semantically far removed, in a creative ideation task [215]. Our system could directly support either approach.

As described below, this prototype is designed to take a step toward this higher level vision while initially reducing some of its complexities. We assume the user is alternating between only two specific tasks; and for now, we use task-classification as a proxy for bookmarking process. The prototype runs in real-time to demonstrate the general feasibility of the memory assistant design. We defined two tasks that could be done by an experimental subject without particular domain expertise and that were intended to elicit two different measurable brain states, and I investigated our ability to distinguish them passively and in real time.

7.1 Materials and Methods

7.1.1 Task Design

After iterative pilot testing, I chose to work with a broad task that an at-home user might experience: designing a room in their home or apartment. Participants were given three rooms to design: a Living Room, Bedroom, and Dining Room. We subdivided the broad task of room design into two phases - the *inspiration and visualization phase* (`Task A`), in which the the goal was to observe images of a room similar to the one they were being asked to design - and the *furniture selection and workload* phase (`Task B`), wherein they chose furniture for their room. We chose these two tasks precisely in an attempt to interface with the prefrontal cortex in different ways, and for their similarity with real-world tasks a user might perform in their home. Both of our experimental tasks are open-ended by design, and require a complex set of thought processes that are unscripted and non-trivial.

Inspiration and Visualization Phase

During the *inspiration and visualization phase*, participants were provided a sidebar of small image links of example photos of the room they were assigned to design. During this phase, the participant's task was explicitly limited to clicking on the sidebar image links, observing the larger images that would appear as a result of clicking on the links, and considering what they would like for their own room. Although visual prompts were provided, participants were instructed to use these images as inspiration for their own internal thoughts of what they would like to create; that is, this task was specifically designed to engage spontaneous cognition and internally directed thought [216, 217], and therefore is a proxy for applied DMN-based tasks.

Furniture Selection and Workload Phase

During this phase, participants were tasked with browsing items from the Ikea website. They were further responsible for keeping track of items they would like to purchase in a Google Sheets spreadsheet. During this task, participants were assigned a budget of $750 USD per room, which they were trying to maximize use of. Similar to `Task A`, we also provided a sidebar with photo

links, but these were of Ikea furniture items that linked to the corresponding items on the website (instead of to an image viewer program) - each of these items, we mentioned to participants, were to be discounted by 50%. They were to calculate prices using a calculator or spreadsheet calculator if desired, and keep track of the totals in the spreadsheet. Participants were asked to choose at least five items during each phase. Through the combination of multitasking, numeric calculation, and high time pressure, this task was specifically designed to engage with the lPFC [218, 219, 220, 221].

7.1.2 Equipment

We used a single fNIRS probe setup as discussed in Section 3.1.2: the probe was positioned over the left eyebrow at approximately at Brodmann region 10. Due to the probe geometry, one set of 3 light source positions was over a relatively lateral aspect of Brodmann 10, and the other set of the light sources were over a relatively medial aspect; the detector itself was in the center of source positions. Outside of the near sources, the closest 4 source positions were each 3cm from the detectors, and the furthest two source positions were 3.61cm from the detectors. Each light source position had two sources which emit infrared light at one of two near-infrared wavelengths (830nm and 690nm) [222]. Raw Alternating Current (AC), Direct Current (DC), and Phase values were converted via the Modified Beer-Lambert Law to Delta Oxygenated and Deoxygenated Hemoglobin values (Δ[HbO] and Δ[HbR][1]) [222]. Data was acquired at approximately 5.8Hz; real-time data were bandpass filtered from 0.1 to 0.4Hz [223], which enables us to isolate the physiologically relevant hemodynamic response signals from cardiac and Mayer waves [7, 224]. Unfortunately, we encountered data acquisition issues in one of the two detectors, therefore only a single detector was used for this study. Although the single detector was able to capture prefrontal data from relatively lateral and medial aspects of the prefrontal cortex, the limitation of the single probe reduced our ability to capture the vertical spatial distribution of hemodynamic responses over the prefrontal area. The second detector would have allowed for more comprehensive mapping of activation patterns and potentially improved classification accuracy by providing both redundancy in some aspects of measurement and broader spatial coverage in others.

[1] In contrast to the first study in this dissertation, this project does not employ the Dual-Slope method, as this project was run prior to that in Part II. Further, the ϕ data were not recorded: therefore, I report here only `I`, rather than `DSI` and `DS`ϕ

7.1.3 Participants

After initially prototyping our study with 6 participants (4 male, aged 18 to 23 years, mean age 20.1, sd 1.2), we recruited 8 participants for the study (6 male, aged 18 to 27 years, mean age of 20.6, sd 2.8). All participants reported being right-handed. None reported having had either traumatic head injury or learning or reading disability. All reported normal/corrected-to-normal vision.

7.1.4 Experiment Design

Participants sat in a comfortable chair in front of a computer terminal running Red Hat Enterprise Linux 7.7. [225], read and signed a consent form, and filled out a demographic questionnaire. We then explained the tasks to the participants and fitted the fNIRS headband. They then completed two groups of room design tasks, where each group contained one trial of `Task A`, a rest period of 2 minutes, then one of `Task B`, then a rest period of 2 minutes. I chose an extended rest period length of 2 minutes for two reasons: first, to ensure complete dissipation of post-stimulus overshoot of from the BOLD signal [47], and second, to provide participants with a substantive break time to mentally relax between tasks. See Figure 7.1 for a visual representation of the task flow. Brain data from the first two groups of tasks were used to train a machine learning model; during the last group of tasks the model was used in real time to classify the user's brain state every 20 seconds - the user-interface would update to show the links corresponding with the brain state predicted by the model. After the three sets of tasks participants filled out a post-survey questionnaire and were compensated with $25 USD.

7.1.5 Interface Details

During the start of each task a window would appear on the user's screen with two buttons - *Images* and *Catalog* (see Figure 7.2). Users were instructed to press the *Images* button during `Task A`, and to press the *Catalog* button in `Task B`. The appropriate images or catalog links would only appear upon pressing the button. During `Task A`, clicking on the image links would pop out a larger image into an image viewer - users could zoom in to more closely observe the inspiration image if desired. At the beginning of `Task B`, we opened a Firefox window with two tabs - a Google sheets spreadsheet tab for the participant to keep track of their purchases, and a basic Google web

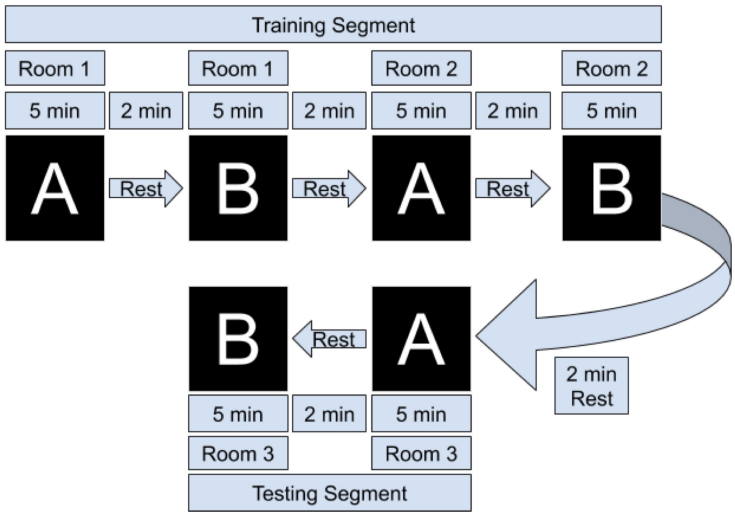


Figure 7.1: Overview of task flow. `Task A` refers to the inspiration and visualization phase (looking at images), and `Task B` refers to the furniture selection and workload phase. Each set of A/B consisted of designing a Living Room, Dining Room, or Bedroom. The Adaptive Filter coefficients, Scaling Coefficients, and SVM Model are learned/trained after the first set of tasks, and during the final group of two tasks they are tested in real-time by a Python thread that extracts data for classification every 20 seconds.

calculator tab (see Figure 7.3). The sidebar contained images which were links that would open a new tab in the same browser window which would go directly to the Ikea website to an item that was on sale. Participants were given an incentive to select the sidebar links by being instructed to maximize the number of items selected while staying under budget; these sidebar items were discounted at 50% off. During `Task B`, users were freely allowed to browse the entire Ikea web interface, but they were not allowed to depart from it and the other tabs we had opened. During the last set of trials, machine learning was used to automatically select the sidebar option of interest.

7.1.6 Data Filtering and Preprocessing

For each trial I used a Recursive Least Squares adaptive filter with our near-channels to remove the effects of neurovascular coupling and movement artifacts [226, 227]. Per [227], I filtered Δ[HbR] and Δ[HbO] separately. Further, one filter was used for each of the left and right sides of the

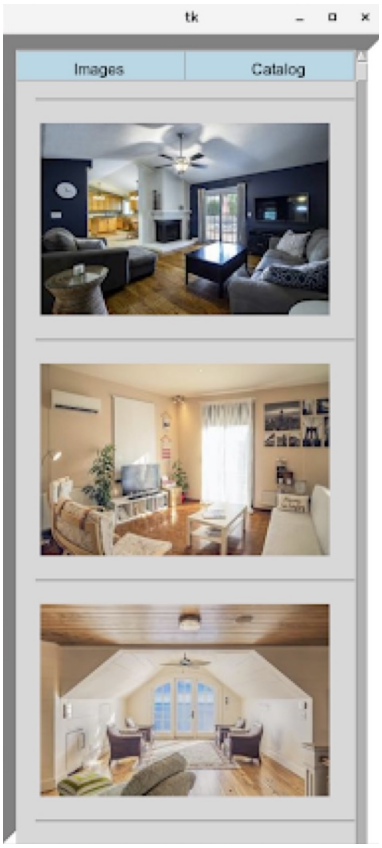


Figure 7.2: Example Sidebar presented in `Task A` (inspiration and visualization phase). Clicking on an image would expand it full-screen.

probe, where each filter associated the near-source with the outer three sources nearest it. Data from the first two groups of trials were used to train coefficients of the RLS filters for Δ[HbR] and Δ[HbO] [227]. After filtering the data, we scaled each channel by removing mean and scaling to the unit variance [173].

I then divided the data into segments of 100 (17.24 seconds). This window frame size was chosen in consideration of the balance between temporal resolution and data quality. That is, I use a window which should be able to capture the full dynamics of a single hemodynamic response [228], while maintaining a time-window length within which to perform useful classification in the context of a real-time interface. I then extracted the `max` and `mean` values of each channel for each window [7]. Feature selection and machine learning were implemented via the `scikit-learn` library [173]. Specifically, the `SelectKBest` algorithm [173], which leverages F-test results to identify the K most statistically significant features from the original feature set, was used for features selection with the parameter K=10. We then input the data to a Support Vector Machine (SVM) with a Linear

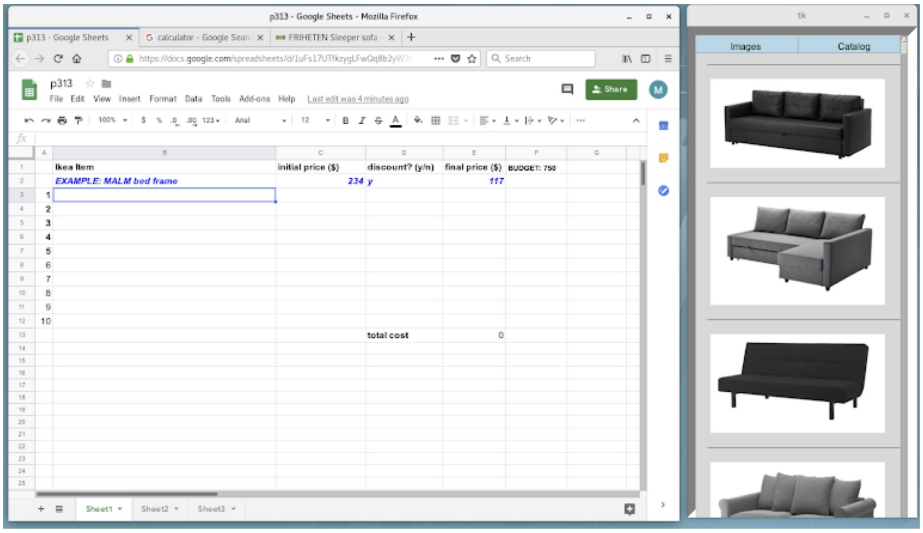


Figure 7.3: `Task B`; task tabs on the left, and images on the right would link to the Ikea website; these images were of furniture which would be discounted by 50%.

Kernel [229], using default parameters including L2 regularization (C=1.0), squared hinge loss, and 1000 maximum iterations.

7.1.7 Online Classification

The final two trials were used to perform real-time ('online') classification. During these trials data was extracted every 100 frames (17.24 seconds), adaptive filter coefficients learned from the training data were used to filter the data, scaling coefficients from the training data were used to scale the data, feature set size was reduced to the same features used in the training set, and data was then classified by the pre-trained SVM. The result of classification led to immediate presentation of the stimulus the participant was attempting to work on: that is, correct classification would show the 'correct' links, and incorrect classification required the user to scroll to the top of the list and press the button corresponding to the task they were completing.

7.1.8 Offline Classification

I also conducted additional offline analyses to provide more comprehensive results. This analysis allows the use of the entire dataset, rather than have the classifier be limited to a single participant's data for training the model.

Leave-One-Out Cross-Validation

I implement participant-level Leave-One-Out Cross-Validation (LOO-CV) to evaluate classification performance. This validation procedure consists of multiple folds, where in each fold we exclude one participant's complete dataset for testing and exclusively use the training cohort's data to determine the preprocessing pipeline parameters, including adaptive filter coefficients and scaling factors. Following preprocessing, for this study model hyperparameters were optimized through an inner cross-validation procedure conducted solely on the training participants, where the best cross-validated model was selected for the test set. Model results were then generated for the test set. Most results for this study will be `STANDARD` results, where I aggregate across all participant-specific test sets for a set of grouping factors, however I also discuss `OPTIMIZED` results in Section 7.2.4.

Brain-Network Dependent Classification

To investigate the functional specificity of probe locations in conjunction with our tasks I conducted analyses within the context of reduced probe sets: specifically, in addition to all probe data, I also tested removing either the relatively medial or lateral probes. This analysis enables exploration of the utility of relatively lateral and medial aspects of PFC as associated with the tasks at hand.

Feature Selection

For offline classifications the feature selection strategy was modified with a number of improvements. First, an expansion of the feature set to include standard deviation, skew, and slope of the linear regression. Second, with varying data window sizes of [50, 100, 150, 200, 250, 300] samples, equivalent to [8.62, 17.24, 25.86, 34.48, 43.1, 51.72] seconds, respectively. Unlike the previous approach, the `SelectKBest` function was not utilized for feature selection.

Model Selection

For the offline analysis the following models were chosen: SVM, QDA, LDA, KNN, ANN, and RF. For hyper-parameter tuning, SVM regularization parameter C was evaluated at levels (0.1, 1, and 10); QDA regularization parameter was tried with levels (0.1, 0.5, 1); LDA was tried with different solvers ('svd', 'lsqr'); KNN was implemented with varying neighborhood sizes (3, 5, 7, and 9); ANN

was assessed one internal layer of either 10 or 50 nodes, and used fixed maximum iteration count of 5000 to ensure convergence; and RF classifier was tested with varying numbers of decision trees (10, 50, 100, and 200).

Group-Specific Analyses

I also consider best-case-scenario results for each grouping of participant, probe set, window size, and model. This enables us to be maximally optimistic about potential classification in future scenarios.

7.1.9 Statistical Analyses

Statistical methods were also used for offline analyses. Log total power of the VLFO band for Δ[HbD] for each probe was calculated in the same manner as in Section 5.3.3 (excluding the ϕ data). A LMM was fit with `REGION` (factor with two levels: `L`, `R`), `TASK` (factor with two levels: `TASK A`, `TASK B`), and their interaction as terms, and log total power of VLFO Δ[HbD] `I` (labeled as `fNIRS`) as the dependent variable. Random intercepts were specified for `PARTICIPANT`, and nested intercepts specified for both `TRIAL ID` (factor with 6 levels, one per trial, used to distinguish separate sections of data collection for each participant) and `PROBE` (factor with 6 levels, representing location of fNIRS sources) within `PARTICIPANT`, as these are expected sources of random variance. Thus, the `R` formula for the model is

$$\texttt{log(VLF}\Delta\texttt{HbD I)} \sim \texttt{TASK} * \texttt{REGION} + \texttt{(1|PID)} + \texttt{(1|PID:TRIAL ID)} + \texttt{(1|PID:PROBE)} \quad (7.1)$$

Separate models were not created per region due to no multicollinearity (VIFs < 10) among factors in the original model.

7.2 Results

7.2.1 Real-time Results

See Table 7.1 and Figure 7.4. Notably, the model's performance varies significantly across different participants. The most successful results were achieved with PID 3, with F1-scores of 0.966 and

0.963 for visualization and workload classes respectively, and a macro average F1-score of 0.964. However, overall performance across participants was inconsistent. Several participants (1, 2, 4, and 6) showed particularly poor results, with F1-scores of 0.000 for one or both classes. The average macro performance across all participants was 0.516 with a substantial standard deviation of 0.282, highlighting the high variability in the model's effectiveness. The confusion matrix in Figure 7.4 shows that the model overall performed better in classifying Visualization (class 0) compared to Workload (class 1) tasks, although the results table indicates that this pattern was not consistent across all participants. Overall, the real-time results indicate that model used was not sufficiently robust for reliable real-time classification across different users. I believe that the lack of substantial training data is the largest factor in the low overall scores.

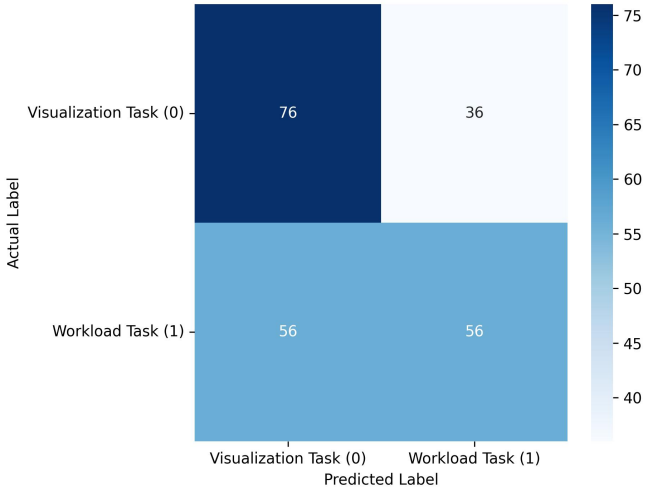


Figure 7.4: Confusion matrix across all participant predictions in the real-time experiment. Although the model classified a relatively high number of 76 visualization tasks correctly, it struggled to correctly identify workload tasks with only 56 correct classifications. Likewise, it incorrectly classified 56 workload samples as visualization and 36 visualization samples as workload. These results suggest that the model is not capable for usable real-time classification.

7.2.2 LOO-CV Results

See Table 7.2 and Figures 7.5 and 7.6 for LOO-CV results. The RF classifier demonstrated the best performance overall per-window of 0.710 in the largest window size of 300, and achieved a mean F1-score of 0.667 across all window sizes. The model that produced this performance included a best overall F1 classification for the visualization task of 0.721, and similarly strong classification

Table 7.1: Per-participant results for the online real-time classification. Despite strong results for some participants, these data suggest that our initial model paradigm does not sufficiently capture patterns within the data suitable for real-time classification.

Participant	Class	Precision	Recall	F1-Score
0	Visualization (0)	1.000	0.786	0.880
	Workload (1)	0.824	1.000	0.903
	Macro Average	0.912	0.893	**0.892**
1	Visualization (0)	0.000	0.000	0.000
	Workload (1)	0.364	0.571	0.444
	Macro Average	0.182	0.286	**0.222**
2	Visualization (0)	0.500	1.000	0.667
	Workload (1)	0.000	0.000	0.000
	Macro Average	0.250	0.500	**0.333**
3	Visualization (0)	0.933	1.000	0.966
	Workload (1)	1.000	0.929	0.963
	Macro Average	0.967	0.964	**0.964**
4	Visualization (0)	0.481	0.929	0.634
	Workload (1)	0.000	0.000	0.000
	Macro Average	0.241	0.464	**0.317**
5	Visualization (0)	0.500	0.929	0.650
	Workload (1)	0.500	0.071	0.125
	Macro Average	0.500	0.500	**0.388**
6	Visualization (0)	0.000	0.000	0.000
	Workload (1)	0.417	0.714	0.526
	Macro Average	0.208	0.357	**0.263**
7	Visualization (0)	0.733	0.786	0.759
	Workload (1)	0.769	0.714	0.741
	Macro Average	0.751	0.750	**0.750**
		Participant Macro		**0.516 ± 0.282**

for the workload task of 0.704. However, SVM, KNN, and QDA all showed comparable overall effectiveness, with mean F1-scores of (0.669, 0.663, 0.665), respectively, across window sizes. The SVM classifier exhibited its best performance of 0.700 with a 200-sample window, KNN and QDA both performed best at the 250 sample window, each with top scores of 0.690. The ANN and LDA classifiers demonstrated the lowest overall effectiveness, with mean F1-scores of 0.623 and 0.580, respectively. While the ANN showed occasionally stronger performance with a maximum

score of 0.660 at window size 100, LDA consistently underperformed compared to other methods, showing a maximum score of 0.630 at window size 100. Across models there is a slight trend of better performance in visualization task detection compared to workload classification. Additionally, although most models showed improved performance with larger window sizes, this relationship is not strictly monotonic; further, the classification accuracies were similar enough such that a trade-off of slight accuracy for faster classification may be preferred in real-time contexts.

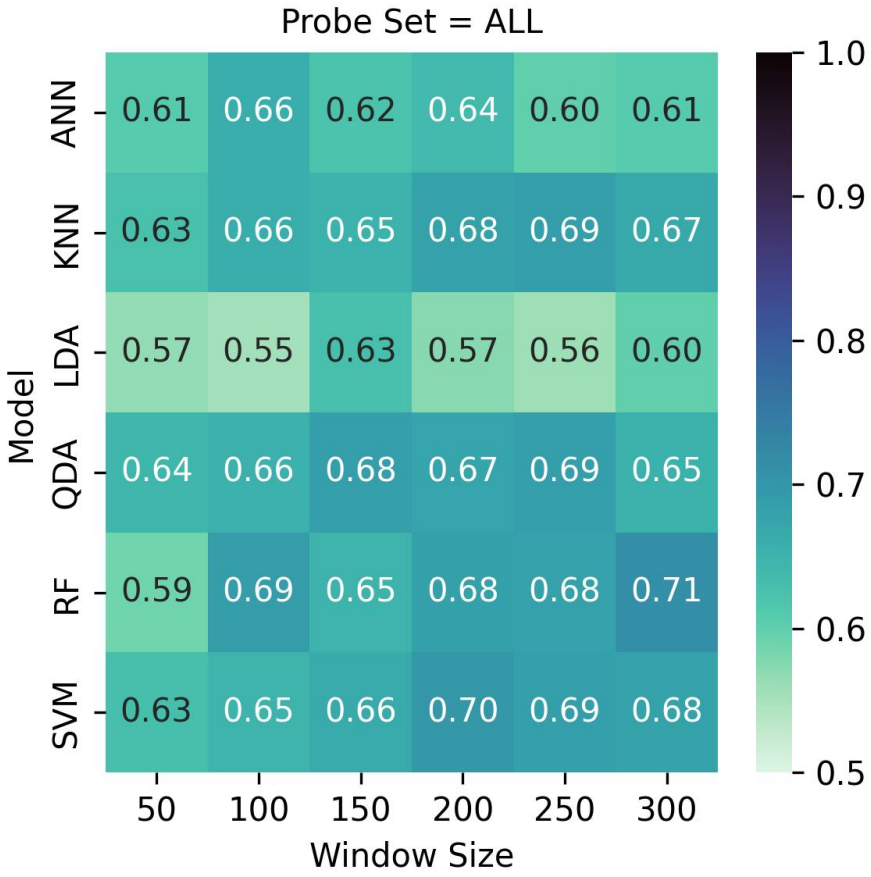


Figure 7.5: LOO cross validation results. Test set scores are produced based on the best model per-participant after inner hyperparameter optimization. Although RF with a window size of 300 performed best with 71%, it showed similar results across multiple window sizes. QDA, KNN, and SVM likewise performed well overall.

7.2.3 Brain-Network Dependent Classification

Comparative analysis of left and right probe sets are presented in Table 7.3 and visualized in Figures 7.7 and 7.8. Several notable patterns are visible across both probe locations and temporal windows

Model	Metric	Window Size						Mean
		50	100	150	200	250	300	
RF	Visualization (0)	0.618	0.702	0.684	0.708	0.695	0.721	0.688
	Workload (1)	0.560	0.677	0.616	0.657	0.664	0.704	0.646
	Average	0.590	0.690	0.650	0.680	0.680	0.710	**0.667**
KNN	Visualization (0)	0.632	0.665	0.669	0.690	0.704	0.680	0.673
	Workload (1)	0.622	0.653	0.632	0.668	0.669	0.652	0.649
	Average	0.630	0.660	0.650	0.680	0.690	0.670	**0.663**
SVM	Visualization (0)	0.659	0.663	0.669	0.707	0.694	0.691	0.681
	Workload (1)	0.601	0.638	0.660	0.688	0.681	0.667	0.656
	Average	0.630	0.650	0.665	0.700	0.690	0.680	**0.669**
QDA	Visualization (0)	0.639	0.658	0.701	0.685	0.698	0.664	0.674
	Workload (1)	0.648	0.654	0.668	0.663	0.676	0.644	0.659
	Average	0.640	0.660	0.680	0.670	0.690	0.650	**0.665**
LDA	Visualization (0)	0.592	0.570	0.636	0.584	0.581	0.619	0.597
	Workload (1)	0.539	0.537	0.622	0.561	0.535	0.579	0.562
	Average	0.570	0.550	0.630	0.570	0.560	0.600	**0.580**
ANN	Visualization (0)	0.632	0.691	0.663	0.670	0.623	0.618	0.650
	Workload (1)	0.584	0.626	0.582	0.606	0.576	0.598	0.595
	Average	0.610	0.660	0.620	0.640	0.600	0.610	**0.623**

Table 7.2: Performance comparison across six machine learning models over varying window sizes in LOO-CV using all probe data, showing F1-scores per-class for visualization and workload. Results indicate promising performance across multiple models, including RF, KNN, SVM, and QDA. LDA and ANN showed the least effectiveness in classification.

in the data.

The left probe data showed particular sensitivity to window size selection, with the best performing models demonstrating improved performance at larger temporal windows: SVM had the strongest overall performance, with a maximum F1-score of 0.705 at a 250-sample window, and an average F1-score of 0.671 across window sizes. This performance was closely matched by RF with a score of 0.695 at 300 samples, and a slightly lower overall performance across windows of 0.642. QDA presented an interesting departure from the trend of higher classification accuracies with larger window sizes, demonstrating its best performance of 0.690 at window size of 150 - while its overall accuracy of 0.662 was better than RF, its maximum performance was not as good. As with the full-probe data, LDA and ANN underperformed by comparison to the other models, with

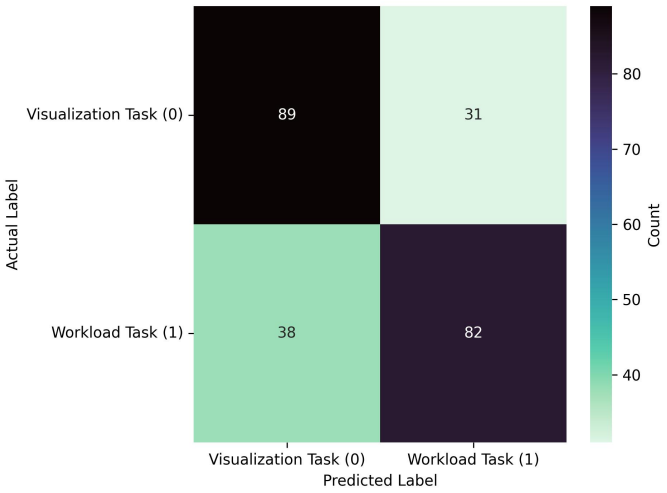


Figure 7.6: Confusion matrix for the best performing model from the LOO-CVV using all probes: RF, which achieved .710 at a window size of 300. The model predicted visualization and workload similarly well, correctly predicting 89 visualization samples, and 82 workload samples. The matrix also reveals relatively balanced misclassification patterns, with 31 visualization tasks misclassified as workload and 38 workload tasks misclassified as visualization.

accuracies of 0.622 and 0.648, respectively.

Results from the right probe revealed different patterns from the left. At lower window sizes of 100 and 150, RF demonstrated the best performance overall of 0.717 and 0.723, besting the best performance from all models trained across both datasets. RF also demonstrated the highest classification accuracy across window sizes, with an average of 0.689. Performance disparity between probe sets was largest for QDA, which maintained 0.63-0.69 with the left probe but decreased to 0.55-0.59 with the right probe. Although LDA showed the most consistent performance across both probe sets, maintaining F1-scores between 0.59-0.65 regardless of location, the overall performance of this classifier was lower than other methods. The asymmetry in classifications among the better-performing models suggests that there are significant differences in the underlying data distributions between probe locations as related to task-based activation.

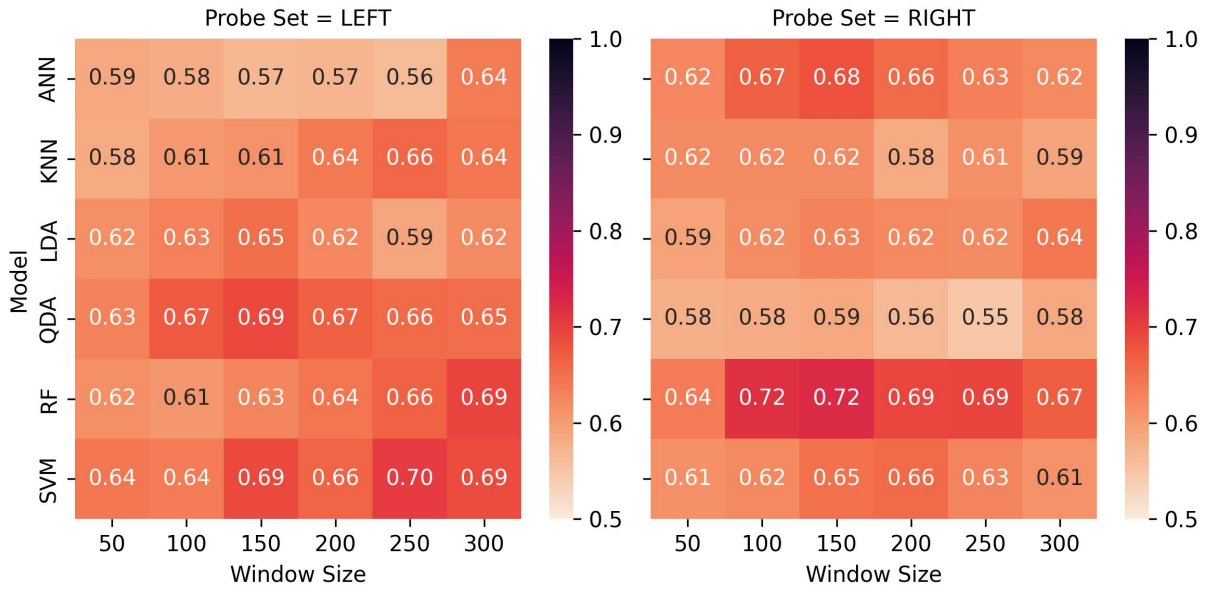


Figure 7.7: Performance comparison of models across window sizes using left and right probe sets, showing F1-scores for task classification. The heatmaps reveal distinctions in performance across probes for some models, with particular increases in performance for RF at lower window sizes for the right probes, whereas SVM performed notably better using the left probes' data at higher window sizes.

Model	**Probe Set**	**Window Size**						**Mean**
		50	100	150	200	250	300	
RF	LEFT	0.616	0.607	0.626	0.643	0.664	0.695	0.642
	RIGHT	0.635	0.717	0.723	0.693	0.694	0.670	0.689
KNN	LEFT	0.583	0.605	0.612	0.641	0.659	0.642	0.624
	RIGHT	0.621	0.620	0.621	0.583	0.614	0.592	0.609
SVM	LEFT	0.641	0.636	0.691	0.664	0.705	0.687	0.671
	RIGHT	0.613	0.622	0.650	0.656	0.628	0.612	0.630
QDA	LEFT	0.630	0.673	0.690	0.667	0.656	0.654	0.662
	RIGHT	0.578	0.582	0.586	0.565	0.547	0.584	0.574
LDA	LEFT	0.616	0.632	0.651	0.624	0.589	0.621	0.622
	RIGHT	0.592	0.619	0.630	0.622	0.621	0.641	0.621
ANN	LEFT	0.586	0.580	0.568	0.572	0.563	0.637	0.584
	RIGHT	0.625	0.668	0.682	0.656	0.635	0.622	0.648

Table 7.3: Model performance when trained on subsets of probe data across varying window sizes.

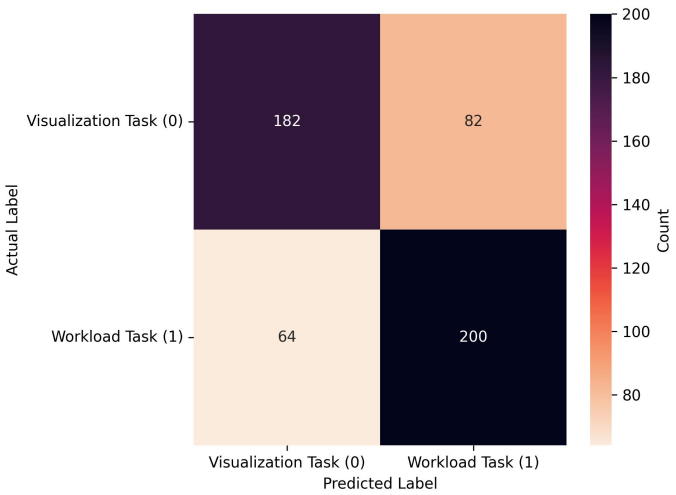


Figure 7.8: Confusion matrix for the best performing model from the LOO-CVV data when using limited probe sets: RF, which achieved 0.723 at a window size of 150 when using data only from the right probe locations. As seen in the confusion matrix for the all-probe data, proportions of classifications of both visualization and workload were similar, however in this case the workload task was correctly classified more often (200) than the visualization task (182), and workload was more often incorrectly classified (82) than visualization (64). Note that, given that the window size is half of that in Figure 7.6, the sample size is approximately double. In fact it is slightly larger, given that one extra sample could be gathered per-trial with this smaller window size.

I believe that these findings suggest an opportunity for enhanced performance through meta-classification approaches. Specifically, the data shows that each probe location exhibits unique strengths in capturing task-related neural states: of most notable distinction, the right probe set demonstrates exceptional performance with RF, reaching F1-scores of 0.72 at moderate window sizes, while the left probe set shows particular strength with SVM, achieving F1-scores of 0.70 with larger windows. This performance asymmetry supports the importance of considering network interactions as hypothesized by Hincks [230], but suggests a novel approach to leveraging these interactions; rather than rely on simultaneous bilateral measurements for direct network comparison, the results indicate that independent classification streams from each probe location and temporal samples could be combined through a meta-classifier architecture. This approach would capitalize on the complementary strengths observed from different models based on probes: RF with the right probe, and SVM and QDA with the left probe. Further, the distinct temporal window preferences between probe sets (right probe performing optimally at 100-150 samples, left probe at 250-300 samples) could further be supported by a meta-classification approach over multiple window lengths.

7.2.4 `OPTIMIZED` Performance Across Grouping Factors

Another consideration within the data is the best performing classifier per grouping factor of participant, window size, probe set, and model. These data are in Table 7.4. Significantly, the overall mean score is 83.5%. This finding indicates that, although there may be classifiable patterns within the data, those patterns are highly variable across participants. And, the implication is that future interfaces may be benefit through cross-validation procedures which perform some element of model selection and/or dataset selection.

7.2.5 Statistical Findings

See Table 7.5. None of the factors in the model are statistically significant. Given the examples demonstrated in the analysis of the Tufts Mental Workload dataset shown in Appendix A, the lack of a significant result is possibly due to the study being underpowered; despite the promising machine learning classification, these results indicate that sample size limitations may be obscuring meaningful patterns in the data. This will be discussed in more detail in Part V.

Table 7.4: `OPTIMIZED` ML results within each combination of participant, probe set, model, and window size. Results here indicate promising potential for future applications whereby these factors are not singularly considered.

Participant	Probe Set	Model	Window Size	F1-Score
0	RIGHT	RF	100	0.800
1	LEFT	QDA	300	0.900
2	RIGHT	MLP	150	0.803
3	LEFT	MLP	250	0.858
4	ALL	KNN	250	0.721
5	LEFT	KNN	250	0.858
6	RIGHT	RF	200	0.808
7	ALL	RF	300	0.933
-	-	-	mean	0.835

Table 7.5: Statistical results. No factors showed significance in the model, indicating that patterns based on prefrontal VLFO are not visible in the data. I believe that study being underpowered is the main issue for these results.

Factor	df1	df2	F	p	sig.	ϵ_p^2	ϵ_p^2 CI
TASK	1	39.00	0.03	0.876	ns	0.00	[0.00, 0.00]
REGION	1	39.00	0.02	0.898	ns	0.00	[0.00, 0.00]
TASK × REGION	1	1351.00	1.53	0.217	ns	0.00	[0.00, 0.00]

7.3 Conclusion

I developed a real-time implicit fNIRS-BCI study based on the vision of the leveraging of brain networks towards a next-generation memory prosthesis interface using fNIRS. Although online real-time classification was not superb, offline simulations of real-time classification which leveraged a larger dataset, an inner cross-validation loop, longer window times, multiple classifiers, and a wider feature set show great promise for future tasks leveraging brain-network based tasks for applied BCI which are suitable for cross-participant classification across multiple classifiers. Further, results suggest multiple promising directions for future system development. Firstly, parallel classification streams that independently process signals from each probe location, and potentially at different window lengths, then combine these predictions through a higher-level meta-classifier. Such an architecture could maintain the benefits of bilateral monitoring while accounting for distinct information patterns to be captured at each location. And, secondly, classification approaches which

can leverage larger datasets and perform model-based cross-validation with participant data in order to select optimal models or parameters could provide useful approaches to handling cross-participant variability. I believe that future extension of such interfaces with broader access to the brain will be able to provide wider and more comprehensive interfaces based on more complex sets of human state information. While real-time performance requires further optimization, and more participants are needed to effectively determine the task effects on prefrontal hemodynamic activity, this study represents a significant advancement in brain network-based BCIs, potentially leading to more intuitive interfaces that facilitate information access based on mental states.

7.4 Shifting Modalities

The prototype real-time fNIRS-based BCI system presented in Part III builds upon the earlier investigations of PFC activity during LLM interaction in Part II, demonstrating the potential for classification of brain states based on relatively precise localization of brain regions and networks. In doing so, these studies illustrate unique approaches of leveraging PFC measurement using fNIRS for implicit BCI applications. However, fNIRS represents just one approach to measuring PFC activity. To further explore the possibilities of applied PFC-based state classification in real-world contexts, we now shift our focus to EEG. Whereas in parts III and II I exploit the spatial resolution of fNIRS towards next-generation implicit BCI development, in this project I consider leveraging the relative weakness of spatial resolution in EEG as a potential strength for complex state classification. To that end, this set of studies pivots to exploring the application of broader aspects of brain activity across multiple networks and regions within the frequency domain features of EEG towards detection of both workload, as well as complex human-state discrimination. Further, the following studies examine what can be achieved with low-cost, low-sensor count EEG devices in classifying complex cognitive states.

PART IV

Inferring Multidimensional Neural State Information from the PFC with Low-Cost EEG

Whereas the last two parts of this dissertation focused on novel applications of fNIRS towards next-generation implicit BCI applications, this part focuses on EEG. Although EEG has been more commonly used than fNIRS in the implicit BCI domain, there are nevertheless many areas on which to improve; in particular, this work constrains itself within the dynamics of using the Muse 2 device, a low-cost, low-sensor count EEG system. In particular, I ask what is capable now of low-cost/low-sensor EEG devices towards robust BCI interfaces, and what future applications might we be able to project from such devices? Particular contributions include multiple benchmarks of the performance of the Muse 2, including: a first study investigating the Muse 2 in classifying neural states related to chess move quality, a follow-up study exploring difficulty levels within multiple standard cognitive psychology tasks and an ecological chess puzzles task, and lastly, classification *across tasks*. I begin here in Chapter 8 with a brief transition into EEG, and provide more detailed overviews of the studies which will follow in Chapters 9 and 10.

Chapter 8

EEG Background and Project Overview

Electroencephalography (EEG) has emerged as a powerful tool in Brain-Computer Interface (BCI) research. In particular, EEG has an established capability to measure cognitive workload [231], which has proven valuable in ergonomics and human-computer interaction studies for differentiating task difficulties [232]. And recent advances in consumer-grade devices have expanded EEG's accessibility beyond traditional clinical settings. This democratization has enabled researchers across diverse fields, from engineering to cognitive neuroscience, to leverage EEG technology in novel ways [233].

However, bridging the gap between clinical neuroscience findings, initial validation of low-cost devices, and practical applications for general users remains an active area of research [234]. Further, while EEG technology continues to advance, significant work is still needed to establish reliable real-time signal processing methods suitable for BCI applications [235]. These challenges necessitate empirical studies to both validate the technological developments of lower-cost devices, and to inform the design of future user interfaces.

The third project of this dissertation employs the Muse 2, a wireless and portable EEG device initially validated for cognitive workload measurement in BCI applications [111, 112]. Despite this initial validation, there remains limited understanding of its practical application in realistic BCI scenarios [234]. To address this gap, I investigated this device in three studies. For all of the studies, I apply both statistical analyses of the data and attempt machine learning for classification with an eye towards potential avenues for real-time state classification in future interfaces.

First, I evaluate the potential for the MUSE to capture neural signatures related to move-quality

during chess games. Results include significant statistical increases in beta and gamma wavebands during higher-quality moves, and, although leave-one-out cross-validation indicates that these data are not sufficient for the development of real-time interfaces, best-scoring models per-participant indicate potential for future developemnts.

Second, I collected data with MUSE while participants engaged in three standard cognitive workload tasks (N-Back, Stroop, and Mental Rotation) and a Chess Puzzle playing task. Similar to the first study, most statistical analyses here demonstrate robust distinctions in cortical activation dependent on task level, however machine learning classification results indicate a large gap towards the application of such findings in the real world. The notable exception in machine learning is the N-Back task, which performs well in a machine learning context.

Thirdly, we collected data for participants who did both the cognitive workload tasks *and* the chess playing task in the same sitting, and evaluated the neural distinctions across tasks. This approach shows a large number of interesting statistical results, as well as promising preliminary machine learning classification results, that indicate subtle cross-task classification across tasks engaging participants different aspects of human psychological activity is possible with even low-cost, low-sensor devices like the Muse 2.

In summary, this project demonstrates that the Muse 2 device is capable of evaluating patterns in the brain relating to within-task workload, as well as across tasks; machine learning evaluation demonstrates promising predictive performance of within-task workload during the N-Back task, and during cross-task classification, which opens the door for future work to explore more subtle human state classification leveraging the PFC for real-time, implicit BCI.

Chapter 9

Neural Correlates of Move Quality During Chess Games

9.1 Introduction

In the first study of this project, I evaluate the potential of the Muse 2 device to be used in an application of which assesses the neural correlates of move quality during chess moves. In this work, we both explore the statistics related to the band-based neurological responses to quality of chess moves, and also apply machine learning to attempt prediction of move quality based on brain state. Although initial results are promising in terms of statistical inferences, low scores of machine learning for prediction demonstrates that there is more work to be done prior to neurological state-based differentiation dependent on chess move quality with low-cost EEG.

9.2 Materials and Methods

To evaluate the quality of chess moves as they relate to neural state measured by low-cost EEG, we first built a local copy of the Lichess open source engine [236]. We then recruited 17 healthy participants (mean age 20.5 $\pm$ 2.03 years, 1 female) to play five 5-minute chess games against a computer opponent with adjustable difficulty (levels 1-8). Participants selected the difficulty level of the computer opponent based on their own skill level, and could adjust the difficulty by one after each game; they were advised to maintain an optimal challenge level for them, which would

maximize enjoyment while minimizing frustration or boredom.

9.2.1 Data Collection

Data was collected from the Muse 2 device using the Mind Monitor [126] application. Connection to the Mind Monitor application was a source of difficulty: frequent disconnections occurred during game-play, after which moves afterwards for that game were lost. Out of a maximum of 5 possible games, a median of 4 games' worth of data were collected per participant, with the mean of 3.64 and standard deviation of 1.15 games; overall, 4 participants were excluded due to lack of data, resulting in a final dataset of 1038 moves across 13 participants.

9.2.2 EEG Preprocessing

The Mind Monitor application applied a 60Hz notch filter prior to collection, and recorded waveform data (in Bels) of each of the following frequency bands: δ [0.4-4Hz], θ [4-7Hz], α [8-12Hz], β [13-30Hz], and γ [30-80Hz] at 10Hz [126, 237]. EEG samples were filtered based on features collected by MindMonitor to exclude eye-blinks, jaw clenches, or invalid connection to the forehead (10% of samples) [126].

9.2.3 Evaluation of Move Quality

Move quality was evaluated using Stockfish 15 [238]. With the engine, the game positions were evaluated both before and after the move in centipawns. These evaluations were the converted to win probabilities (W) using a logistic function

$$W = \frac{1}{1 + 10^{\frac{-p}{400}}} \tag{9.1}$$

where p is the centipawn position evaluation [239]. I then define move quality score was as the difference between the post-move and pre-move win probabilities, thus measuring deviation from theoretically optimal play.

9.2.4 Data Labeling

Move quality was normalized per-participant by transformation into three categories based on tertiles (quantile cuts) of each participant's evaluation scores. This participant-specific normalization accounted for individual differences in skill level, such that the "Higher Quality" and "Lower Quality" designations reflect each player's relative performance rather than absolute evaluation. The middle tertile ("Medium Quality") was removed from the analysis to create a clearer distinction between higher and lower quality moves. This preprocessing step results in 694 total moves, 347 per class.

9.2.5 Statistical Analyses

For statistical analyses, frequency domain data from the frontal AF7 and AF8 probe were averaged together to increase the signal-to-noise ratio of the Muse 2. Data were further averaged per-move to eliminate potential for autocorrelated samples biasing the models. One LMM was created per wavelength, with move quality as a fixed effect and games nested within participants as random effects. Bonferroni adjustment is made across 5 wavelengths (α=0.01).

9.2.6 Machine Learning

For machine learning we employ a leave-one-out cross-validation (LOO-CV) methodology to classify EEG brainwave data related to move quality, where on each iteration a separate subject is used for testing. Data from all 4 probes was used for classification. Summary features for each probe-wavelength were extracted: mean, median, max-min difference, 75th-25th percentile difference, and variance. Analyses were then performed analyses over the same set of models used in Section 7.1.8; `STANDARD` and `OPTIMIZED` analyses are likewise reported. Also, for this classification, Extreme Gradient Boosting (XGBoost) was also added, a scalable and efficient tree-based boosting algorithm known for its high performance in machine learning tasks [240, 241].

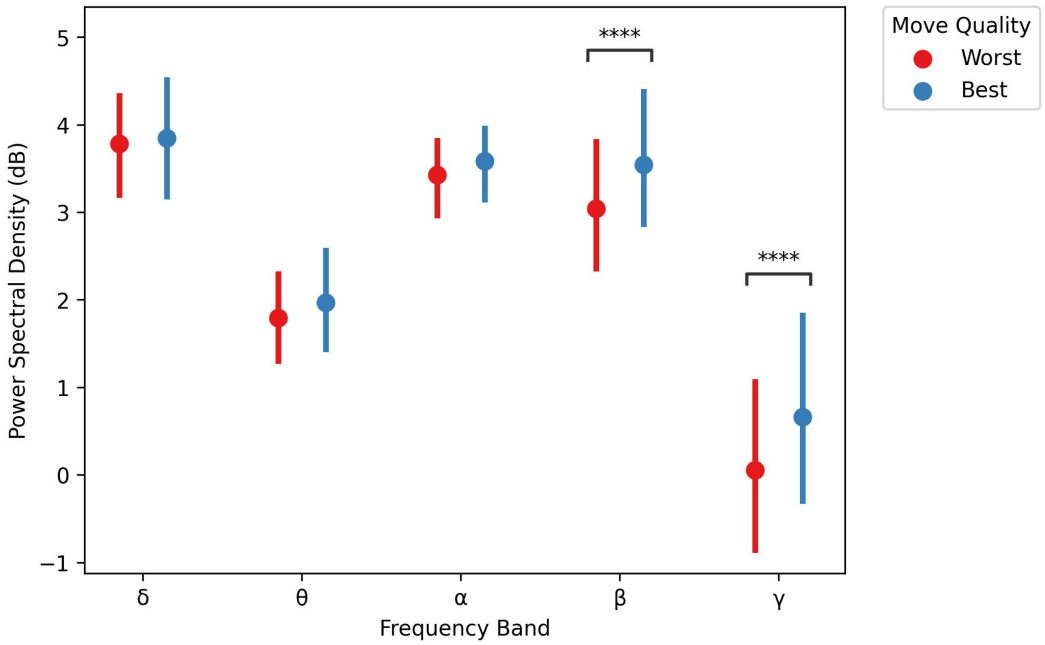


Figure 9.1: EEG band power comparing best and worst moves. β and γ band power increased with an increase in move quality across participants.

9.3 Results

9.3.1 Statistical Results

Results of the statistical analysis are shown in Figure 9.1 and Table 9.1 demonstrated that both β ($F_{1,667.58} = 32.25, p < 0.001, \epsilon_p^2 = 0.04$) and γ ($F_{1,666.24} = 35.56, p < 0.001, \epsilon_p^2 = 0.05$) brain wave activity increased significantly during higher quality chess moves. No other wavebands showed significant effects as a consequence of move quality.

Table 9.1: Statistical analyses of waveband data. Due to multicollinearity between wavebands, separate models were trained for each band. β and γ bands showed significant effects. The **sig.** column reflects α=0.01, with Bonferroni correction adjusting for 5 wavelengths.

Band	**df1**	**df2**	**F**	**p**	**sig.**	ϵ_p^2	ϵ_p^2 **CI**
δ	1	669.38	0.21	0.651	ns	0.00	[0.00,0.00]
θ	1	671.44	1.63	0.205	ns	0.00	[0.00,0.01]
α	1	673.97	2.21	0.140	ns	0.00	[0.00,0.01]
β	1	667.58	32.25	<0.001	***	0.04	[0.02,0.07]
γ	1	666.24	35.56	<0.001	***	0.05	[0.03,0.08]

9.3.2 Machine Learning Results

See Table 9.2. On initial glance, the `STANDARD` LOO-CV scores averaged across participants show a relatively grim picture: the best performing model was SVM, with 54% F1 score. On their own, these results are not be good enough to consider for deployment in real-time BCI interfaces.

Table 9.2: `STANDARD` LOO-CV F1 averages across participants per-model. Overall results are not viable for application in real-time interfaces.

KNN	LDA	MLP	QDA	RF	SVM	XGB
0.532	0.503	0.516	0.494	0.472	0.540	0.503

However, a closer look into the `OPTIMIZED` data (Figure 9.2 and Table 9.3) indicates a more nuanced picture: when considering not the averages across participants, but rather the `OPTIMIZED` score of each participant's best model, the average score across participants increases to 60%.

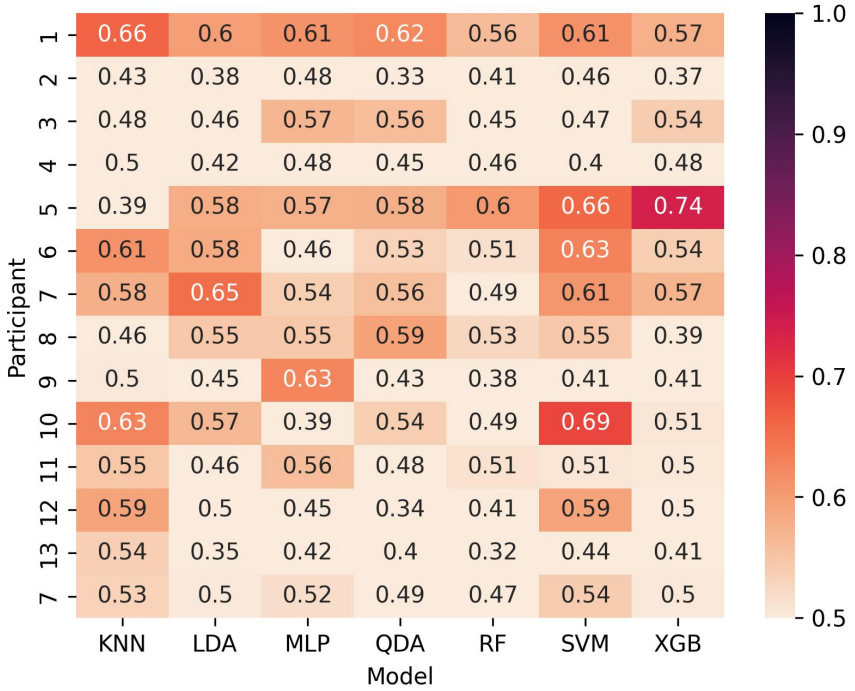


Figure 9.2: Machine learning scores per-participant across models. Darker boxes indicate higher scores. Although classification overall is not superb, variability across models and participants indicates potential application spaces may exist within such data.

The high variability across participants and models indicates potential for meta-classification techniques, or potentially model-based cross-validation on data from a given participant to determine

optimal classifier or meta-classifier with which to work.

Table 9.3: Best scoring model per-participant, sorted in descending order from left to right. The mean score across `OPTIMIZED` results is better, with a mean score of 60%.

Participant	**5**	**10**	**1**	**7**	**6**	**9**	**8**	**12**	**3**	**11**	**13**	**4**	**2**
Model	**XGB**	**SVM**	**KNN**	**LDA**	**SVM**	**MLP**	**QDA**	**KNN**[1]	**MLP**	**MLP**	**KNN**	**KNN**	**MLP**
Macro F1	0.737	0.693	0.662	0.652	0.630	0.626	0.593	0.590	0.570	0.559	0.536	0.495	0.485

9.4 Conclusion

This study demonstrates that the Muse 2 device can detect meaningful variations in neural activity during a complex cognitive task like chess. The consistent increases in β and γ band activity during higher quality moves suggest specific neural signatures associated with optimal decision-making. And, whereas the modest effect sizes and variable machine learning results highlight current limitations in translating these neural patterns into reliable real-time predictions, high variance across participants and models indicates potential for new classification techniques to be applied for future interfaces. These findings contribute to our understanding of the potential and constraints of consumer EEG devices in brain-computer interface applications.

9.5 Next Steps

Although this study has established a clear relationship between EEG signals and chess move quality, given the difficulty of current application for BCI, a natural progression is to take a "step backward" towards a slightly more constrained domain. To that end, the next study does so, while in consideration of several questions arise. Firstly, although EEG has been reliably shown to be able to detect cognitive load, what about with the low-cost MUSE II device? And secondly, can we distinguish between fundamentally different cognitive activities, regardless of difficulty levels? The following chapter addresses these questions directly through an investigation of both within-task workload levels and cross-task state differentiation. Further, to maximize the potential for successful classification, I initially constrain the context of the study to examining established cognitive paradigms. However, to begin to bridge the gap from the sterile world of cognitive neuroscience out

[1]This participant performed equally well with SVM.

into real world ecological settings, I also introduce an ecological chess puzzle task. This progression from domain-specific workload assessment to broader cognitive state differentiation represents an important step toward understanding the practical capabilities of low-cost, low-sensor EEG for implicit BCI applications.

Chapter 10

Within-Task Workload and Cross-Task Neurocognitive State Classification with Low-Cost EEG for BCI

Having examined the neural correlates of chess move quality in the Chapter 9, we now extend our exploration of the Muse 2 device's capabilities by investigating both within-task workload classification and cross-task state differentiation. While the previous study focused on quality distinctions within a single domain (chess), the following research broadens scope to examine how effectively this consumer-grade EEG can distinguish between different levels of cognitive load within standardized tasks, as well as between fundamentally different cognitive activities. Specifically, I study a combination of established cognitive paradigms (N-Back, Stroop, and Mental Rotation) and an ecological task (chess puzzles). To that end, this work focuses on three primary research questions, all in relation to the Muse 2 device:

1. Can we reliably detect and classify different levels of cognitive workload within individual tasks (N-Back, Stroop, Mental Rotation, and chess puzzles)?

2. To what extent can we distinguish between different cognitive tasks, independent of their difficulty levels?

3. Are the neural signatures identified in questions 1 and 2 robust enough to support real-time

classification for BCI applications?

Through this project, I demonstrate successful statistical distinctions of workload levels within some tasks, as well as differentiation between task types irrespective of difficulty level. With machine learning I further show reliable predictive power to differentiate of workload levels in the N-Back task, while also achieving effective cross-task classification. These findings demonstrate that consumer-grade EEG devices can effectively detect and differentiate some forms of cognitive workload, and that they can be leveraged with some success towards real-time classification distinguishing workload in some tasks, as well as in differentiating between nuanced cognitive states, supporting their potential use in adaptive BCI applications.

10.1 Materials and Methods

10.1.1 Tasks

N-Back

The conceptualization of the N-Back task traces its roots to the work of Kirchner in 1958 [242], establishing a cornerstone cognitive exemplar for the probing of working memory and executive function. This task is paramount in the realm of cognitive investigation due to its unique capacity to measure the retention and manipulation of information across brief intervals. Through the employment of diverse stimuli—ranging from auditory-verbal to visuospatial—and the increasing complexity introduced by 1-back, 2-back, and 3-back variations, the N-Back task presents an intricate measure of working memory capacity, executive function, and attentional control aptitudes. The operational framework of this task is meticulously designed to assess accuracy, response latencies, and error rates, thereby illuminating the rapidity and proficiency of cognitive operations by requiring participants to simultaneously store and manipulate information [243, 244, 245].

Stroop

As a widely used practice in psychology, the Stroop Test is a task designed to test the measurement of a user's cognitive load [246]. Created by John Ridley Stroop in 1935, the task focuses on specific cognitive processes such as selective attention, memory, learning, cognitive load, and processing

speed [247]. The test requires the participant to identify the color of a printed word while ignoring the actual word itself. For instance, if the word 'BLUE' is printed in a yellow font, the correct response would be 'yellow'. The task measures the user's accuracy and speed to give insight into the user's cognitive ability and efficiency [248].

Mental Rotation

The Mental Rotation Task (MRT) is a cognitive psychological test devised by Roger Shepard and Jacqueline Metzler in 1971, aimed at assessing spatial reasoning ability. In this test, participants are presented with two 3-dimensional objects in different perspectives positioned in space. The objective is to determine whether the objects are identical. Studies utilizing the MRT have revealed that reaction time tends to increase linearly as the angular disparity between the orientations grows, which indicates that participants mentally rotate one of the objects to determine if it matches the other [249]. Moreover, EEG work has demonstrated that the MRT elicits activation in the superior parietal lobule and the intraparietal sulcus, indicating their involvement in spatial processing during the task [250].

Chess

Several studies have explored the effects of chess playing on the brain. Increased theta power in posterior brain regions was observed when chess players engaged in faster-paced games, suggesting a link to long-term memory retrieval and chunk processing [251]. High-level chess players have also been shown to exhibit increased alpha EEG power in the occipital area during chess playing, indicating a greater adaptive response to the cognitive demands of the task [252]. Another study found that winning players exhibited higher theta power in frontal, central, and posterior brain regions as the difficulty of the opponent gradually increased; higher alpha power was observed in more challenging games, suggesting the engagement of creative thinking and the exploration of alternative solutions. By contrast, losing players demonstrated a decrease in beta and alpha power as the opponent's difficulty escalated, indicating a lack of adaptive response to challenging situations and an inability to formulate effective solutions [253].

Summary

The selection of these four cognitive tasks: N-Back, Stroop, Mental Rotation, and Chess puzzles—creates a comprehensive framework for examining distinct yet interrelated aspects of cognitive function: the N-Back task primarily assesses working memory and information manipulation, the Stroop test measures inhibitory control and attention, the Mental Rotation task evaluates spatial reasoning, and the chess puzzle task introduces a unique dimension of strategic thinking and pattern recognition that bridges and extends these cognitive domains. Chess puzzles are particularly valuable in this battery because they require the simultaneous engagement of multiple cognitive processes: working memory (similar to N-Back, but in a more complex context), inhibitory control (as in Stroop, but applied to move selection), and spatial reasoning (complementing Mental Rotation, but with added strategic complexity). More detail on the tasks follows below.

10.2 Materials and Methods

10.2.1 Study Outline

We performed our study in two phases

1. In our initial phase I we ran participants *either* on the set of mental workload tasks *or* on the chess puzzle task. These data are leveraged for within-task workload analyses only.

2. In phase II we ran a second set of participants *both* on the mental workload tasks *and* the chess puzzle task in the same sitting. These data are both added to the within-subjects workload dataset from phase I, and also leveraged separately for cross-task analyses.

The tasks used for both phases I and II were the same. For each participant in both studies we administered a preliminary consent form, fit them with the MUSE device, had them perform the tasks, and informally debriefed them. For phase I, participants for the chess study were sent a $10 Amazon gift card, and participants for the mental workload study were sent a $20 Amazon gift card. For phase II, participants were sent a $30 Amazon gift card. No participants from phase I were enlisted for phase II. The study was approved by our organization's Institutional Review Board. While we use the data from both phases I and II to model within-task difficulty levels, we only use

the data from phase II for our cross-task analyses.

10.2.2 Program Implementation

We implemented our experiments using JavaScript, and leveraged several key libraries and tools. We built the core experimental framework with `jspsych` [254], which provided a flexible foundation for our study design. For the chess-related components, we employed `Chess.js` [255] to handle game logic and `ChessBoard.js` [256] to create an interactive visual interface. We interfaced our code with the Muse 2 device by way of the `muse-js` [127] library.

10.2.3 Chess Puzzles Database

To create a database of chess puzzles we began by downloading the `lichess.org` open chess puzzle database, which contains millions of chess puzzles [257]. To minimize confounding factors due to puzzle differences, we filtered the puzzles to only contain those that result in checkmate, which resulted in a total of 686,559 puzzles. Each puzzle had an associated Glicko2 rating to quantify its difficulty [258]. We further filtered the total puzzle set based on Glicko2 difficulty ratings within the range [600, 2250], where across the `lichess.org` site a rating of 600 represents the skill level of a complete beginner and the rating of 2250 represents the skill level of an expert [259]. We then binned the puzzles into Glicko2 rating ranges of 50; the first bin contained puzzles with ratings from [600-650), and so on. The lower rating bins (600-1750) each contained over 10,000 puzzles. While the higher rating bins (1800-2250) contained fewer puzzles, each bin still maintained a minimum of 197 puzzles, ensuring sufficient availability for high-performing participants.

10.2.4 Task Details

Chess Task

In the chess task, participants played a total of 6 rounds with 30 puzzles each. Each puzzle consisted of an arrangement of pieces such that the participant could solve the puzzle by making a certain sequence of moves to reach a checkmate. An example of a chess puzzle is shown in Figure 10.1. The first puzzle of each round started with a Glicko2 rating from the 800-850 bin. If a participant successfully completed a given puzzle, the next puzzle would come from the next hardest bin.

Conversely, if the participant made an incorrect move while solving a puzzle, the next puzzle would come from the next easiest bin. After each move, participants were shown whether their move was correct. Participants were instructed to perform the puzzles as quickly as possible without sacrificing accuracy. The maximum time allotted per-puzzle was 30 seconds.

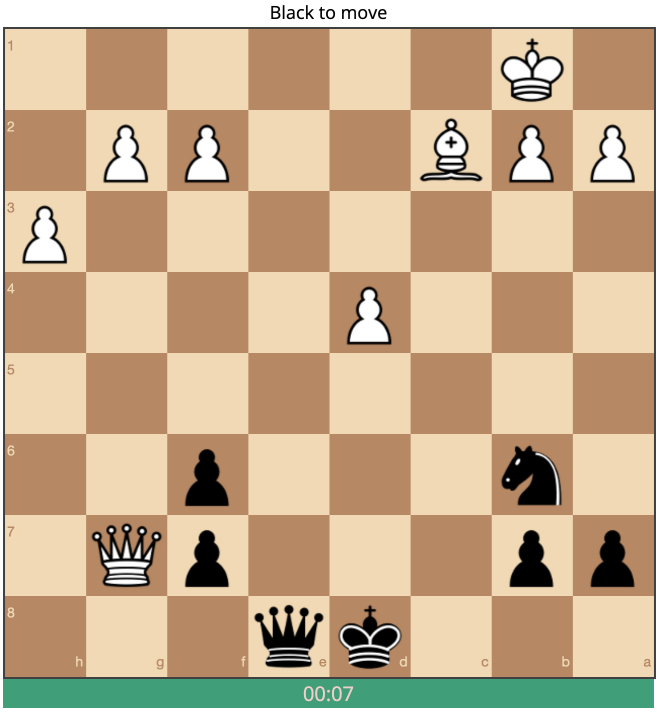


Figure 10.1: A chess puzzle where the solution is to move the black queen to square e1, and then to square d1, which delivers checkmate by taking the white bishop that will move to defend the white king.

Cognitive Neuroscience Tasks

We ran the three cognitive tasks (N-Back, Stroop, and Mental Rotation) in a random order. Before completing each cognitive task, the participants were provided with a written explanation of the task. Next, they performed a task demo where they practiced the tasks and were shown after each response whether or not their answers were correct. During the training period, participants were encouraged to ask questions to ensure they understood the task; if they were confused about how to complete the task, they had the opportunity to perform the training session again. No participants performed the training session for any task more than 2 times. For each trial across all task types, a fixation cross was displayed for 500ms prior to trial onset, the inter-trial interval was 250ms, and

a keyboard response ended the trial.

N-Back Task For the N-Back task, participants were shown a series of letters. Each letter was shown for 500ms, and the participant had 2500ms to respond, after which the trial ended automatically. For a given trial, they were instructed to press the `y` key if the current letter was the same as the letter they saw N letters previously (known as a N-Back trial), or to press the `n` key otherwise; N-Back trials occurred with 30% frequency in all workload conditions. See Figure 10.2 for an example of a series of N-Back trials.

Figure 10.2: The top row is a series of possible characters presented during a N-Back task. The bottom row represents the correct sequence if the task was a 2-Back, specifically. That is, on seeing all of the first four letters, the participant would press the 'N' key. However, the 5th character 'G' was also seen two characters prior, thus the participant would press 'Y'. During the study, the characters are only shown one at a time, so the participant must remember N characters continuously.

Stroop Task For the Stroop task, participants were shown a series of words that each spelled the name of a color – these words were also written in a text color which was not necessarily the same as the color they spelled. The participant pressed the key corresponding to the first letter of the word's text color. Congruent trials, in which the color of the word matches the color that the word spelled, occurred with 50% frequency. Each stimulus was shown for a maximum of 2000ms, after which the trial ended automatically. See Figure 10.3 for an example of an incongruent trial.

Mental Rotation Task For the Mental Rotation task, participants were shown a series of pairs of 3D blocks. We used the block paradigm and images developed by [5]. Participants were asked to press the `y` key if the blocks were the same but rotated, and the `n` key if they were not the same block. Blocks were the same with 50% probability. Each stimulus was shown for a maximum of 7500ms, after which the trial ended automatically. See Figure 10.4 for an example of the Rotation task.

Figure 10.3: Example of an incongruent Stroop stimulus; the word 'red' is written in blue. In this case the participant would press the 'b' key on their keyboard, indicating the color the word is written in, not the color that is spelled. Red, yellow, blue, and green were all used for the study.

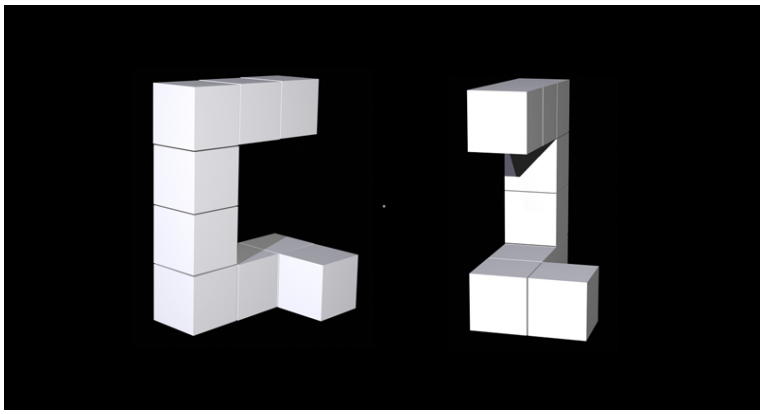


Figure 10.4: Example of a trial used for the Mental Rotation task (images used from work of [5]). This is not a valid rotation, because the second object is not a rotated version of the first object.

Task Block Information For all tasks, four blocks of trials were run. Each block of the Stroop task contained 75 trials; each block of the Mental Rotation task contained 24 trials; each block of the N-Back task contained four sub-blocks of 25 trials each, one per N-Back variant, in a randomized order. See Figure 10.5 for a visual representation of the experimental paradigm. After each block, there was a 30-second rest. Note that the combination of different maximum time per-trial, along with different number of trials-per-block, resulted in different total numbers of trials and epochs-per-trial, collected across participants.

10.2.5 Participant Information and Chess Player Skill

Our subject population is composed by healthy graduate and undergraduate students from Tufts University (overall ages range from 18-26, mean 20.4). In phase I, our subject pool involved 8 chess participants (1 female) and 13 mental workload task participants (4 female). For phase II, we recruited 9 participants (all male) to do both tasks in one sitting. Our workload-specific dataset

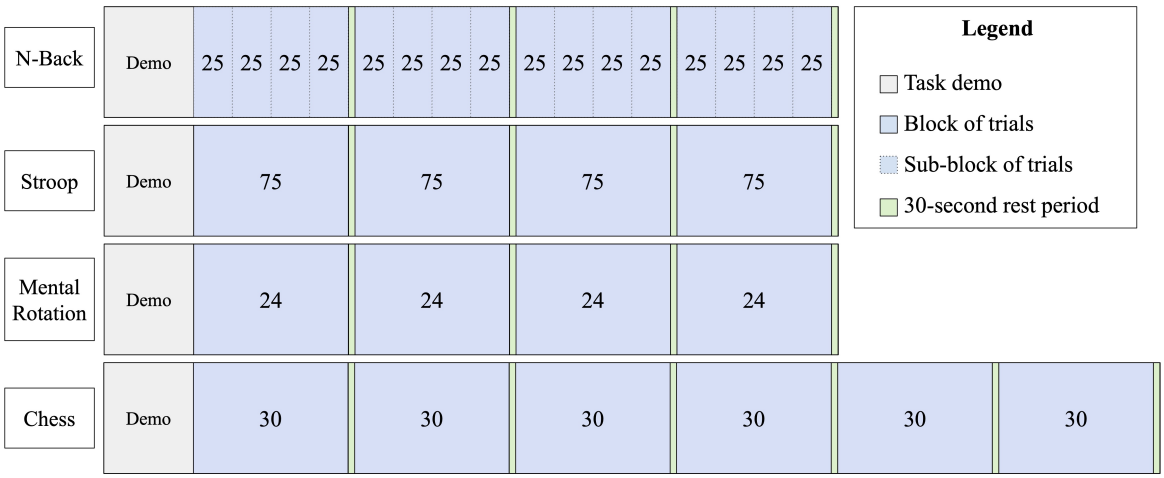


Figure 10.5: Experimental procedure per task. In the N-Back task, each set of 4 sub-blocks are a random permutation of 25 N-Back trials of the same N, where N is selected from {0,1,2,3}. Stroop blocks are 75 trials each, and Mental Rotation blocks are 24 trials each. The chess blocks are each 30 puzzles in length. Note that this graphic only indicates the number of trials collected; due to data loss from the Muse 2, time-per-trial, and number of trials per-block, our final dataset resulted in imbalanced samples (see Table 10.1).

therefore begins with 17 chess participants and 22 mental workload task participants, and our cross-task analysis dataset contains 9 participants.

Chess participants also filled out a form indicating the frequency of their chess play. The distribution of responses is shown in Figure 10.6 - a substantial majority (75%) of participants responded that they play at least once per week. After gathering our data, we assessed skill distribution of Chess players by examining the maximum puzzle difficulty each participant successfully completed, as illustrated in Figure 10.7. For reference, puzzles rated at 1400 Elo are the most frequently attempted puzzles in the lichess.org database [257]. While participant skill level could theoretically influence neural responses during puzzle solving, our preliminary analyses found no significant effects. Given these results, subsequent analyses do not investigate skill-based effects. Future research examining skill-based neural responses may benefit from recruiting a larger participant pool with broader representation across skill levels.

10.2.6 EEG Data Collection and Preprocessing

Raw data was collected via `muse-js` at 256 Hz. We focus our analysis only on the prefrontal cortex data, accessed through the Muse 2's AF7 and AF8 probes. Our primary objective in the EEG

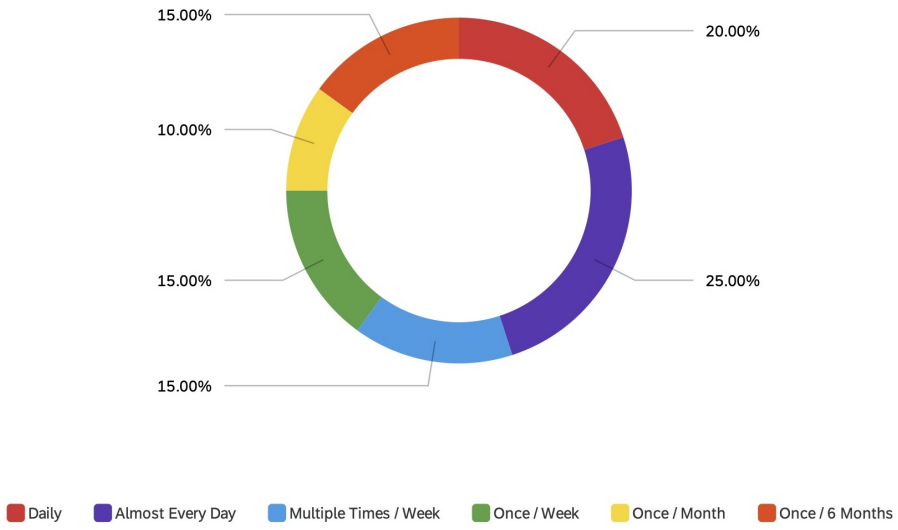


Figure 10.6: Frequency of Chess play across all Chess participants (N=17). Most players (75%) played chess at least once per week.

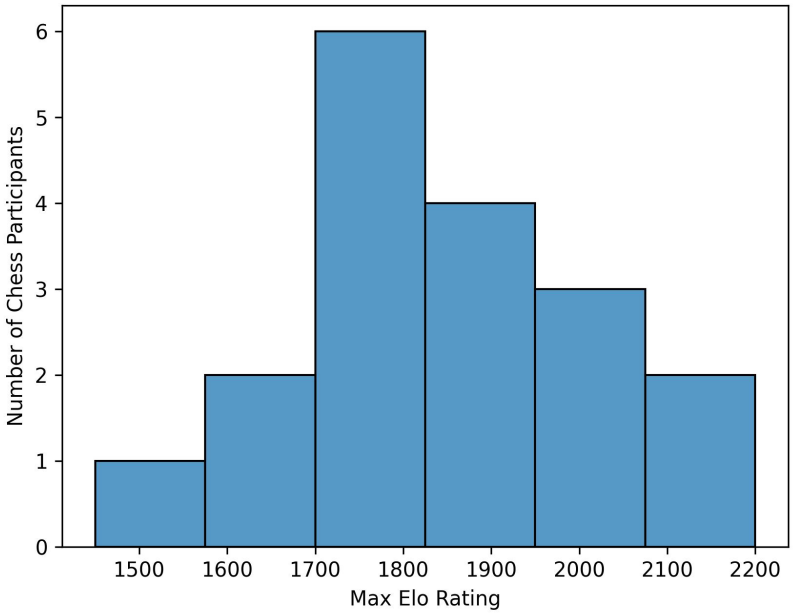


Figure 10.7: Distribution of maximum Chess puzzle difficulty level achieved across participants, measured in Elo Glicko2 rating (see Section 10.2.3). Preliminary quantile-based analyses revealed no significant associations between maximum Elo rating and EEG response patterns (analyses not reported in main results).

analysis was to preserve the maximum amount of usable data while developing a methodology which could be leveraged in the future for online processing. Data pre-processing and feature extraction were both conducted in Python. Data preprocessing are visualized in Figure 10.8.

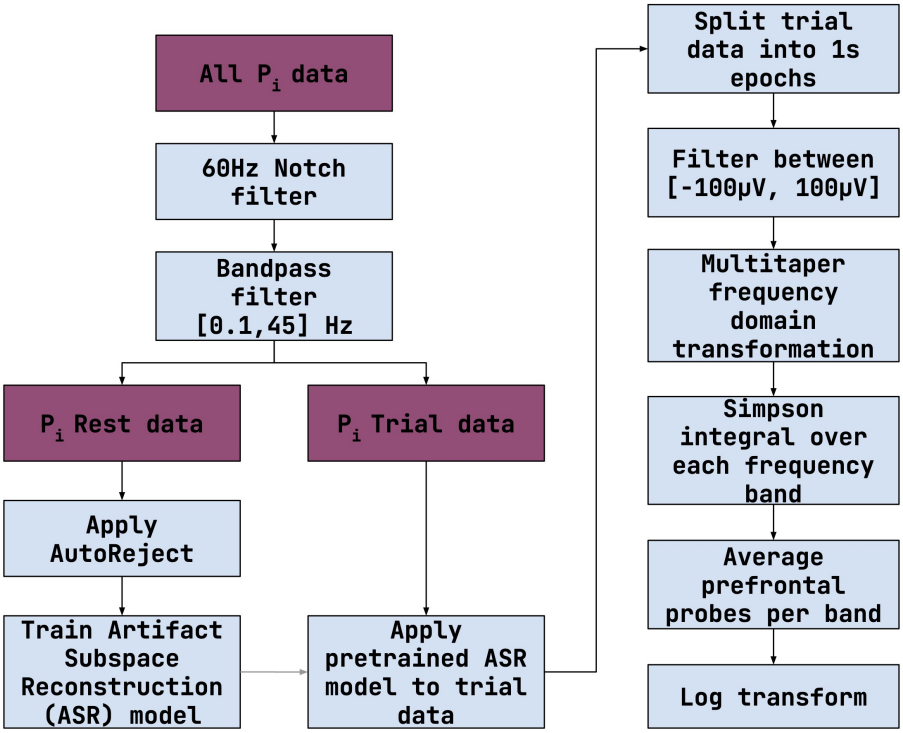


Figure 10.8: Visualization of the EEG processing pipeline showing data flow from raw signal through filtering, artifact removal, and spectral analysis stages for a single participant P_i. Rest data is cleaned with the AutoReject algorithm, and the cleaned data is used to train an ASR model, which is then applied to trial data before frequency domain transformation and averaging.

I first utilized the `mne` package [134] to implement a notch filter at 60 Hz and a bandpass filter spanning 0.1 to 45 Hz[1]. For each participant, we applied the AutoReject algorithm [260] on their resting-state data, which we automatically split into epochs of 2 seconds for preprocessing. AutoReject fully automates the correction of momentary and transient movement artifacts by establishing sensor-specific thresholds based upon effective cross-validation techniques across multiple datasets (the MNE sample data [134], the multimodal faces dataset [261], and the EEGBCI motor imagery data [262, 263]) to help distinguish genuine neural signals from artifacts caused by muscle movements, eye blinks, or technical issues. Once thresholds are defined, AutoReject evaluates each

[1]I initially included frequencies as low as 0.1Hz to accommodate potential ERP analyses [96], but ultimately focused on $\geq$4Hz for our frequency analysis. Consequently, frequencies below 4Hz were not used in our final analytical pipeline.

EEG trial and either repairs by interpolating clean data from adjacent sensors (also known as channels) or excludes them entirely from the dataset to ensure only the cleanest, artifact-free data are to be used in the later stages of analysis. We used the resulting cleaned resting-state data to train an Artifact Subspace Reconstruction (ASR) model [264] on their task data. This model effectively identifies and reconstructs artifact-contaminated subspace segments, thereby ensuring minimal loss of brain signal integrity while preserving a maximum amount of data. The ASR algorithm plays a crucial role by intelligently isolating and reconstructing the data segments contaminated with artifacts, not only enhancing the quality of EEG data by minimizing interference but also preserving the valuable brain signals needed for accurate analysis. We then split the data into epochs (discussed in Section 10.2.6 below). To optimize signal-to-noise ratio we applied a thresholding procedure [265] where we excluded any epochs containing values outside the range [-100 µV, 100 µV] in either of the frontal probes. The median data inclusion rate per-participant is 98% across both chess and mental workload datasets, indicating that most participants were largely unaffected by filtering. However, the mean data inclusion rate per-participant is 64%, indicating that the process disproportionately affected a subset of participants. Participants with excessive data loss (>=60% epochs excluded, n=8) were omitted from the analysis. Of these participants, five were from phase I during the workload tasks, one was excluded from phase I during the chess tasks, and two were excluded from the phase II data. Although the specific causes of these grouped losses cannot be definitively determined from the available data, the clustering of exclusions suggests potentially systematic factors in data collection. For future work, we recommend implementing real-time data quality monitoring to identify issues promptly. Our established preprocessing pipeline can also be adapted for real-time analysis to detect batches of invalid samples as they occur, enabling immediate adjustments to headband positioning or facilitating the collection of additional resting-state data for cleaning procedures.

Epoching the Data

The selection of epoch length presents a methodological trade-off between signal quality and practical utility. While longer epochs enhance signal-to-noise ratio and provide more reliable spectral estimates, they pose implementation challenges for real-time neural interfaces and provide less information for machine learning models to understand underlying patterns. Additionally, the temporal structure

of our experimental paradigm constrains the maximum feasible epoch length: the Stroop, Rotation, and Chess tasks contained multiple workload conditions within single experimental blocks. To optimize for potential real-time classification applications while maintaining adequate signal quality, we extracted epochs from the continuous data stream using fixed time windows. Each epoch was labeled according to the concurrent workload condition. We selected 1-second epochs to maximize the number of available samples for machine learning while maintaining sufficient duration for reliable spectral estimation of frequencies as low as 4 Hz.

Frequency Domain Transformation and Prefrontal Averaging

Each epoch was transformed into the frequency domain with `mne`'s Multitaper library [130, 131]. Following the work of [107], we delineated frequency bands as follows: θ (4, 8), α_1 (8, 11), α_2 (11, 14), β_1 (14, 25), β_2 (25, 35), γ_1 (35, 40), and γ_2 (40, 45). The Simpson integral [136] was used to extract single values representing the total power in each frequency band. Given our focus on overall pre-frontal activation patterns, we averaged the frequency-domain data from both probes into a single value per frequency bin, optimizing signal-to-noise ratio while maintaining measurement validity for our research objectives. Resulting power values were log-transformed.

Data Collection Challenges and Solutions

Throughout our study we experienced connectivity issues with the MUSE device. In our initial implementation of the experimental protocol, this would cause complete data loss for a participant. To mitigate data loss, we implemented a reconnection protocol prior to the start of each task block (e.g., one block of N-Back, Stroop, mental workload, or chess puzzles). This approach significantly reduced data loss overall, but some blocks were still lost for some participants. Table 10.1 presents detailed information on the number of captured trials and epochs captured per-participant per-task after data loss was taken into account.

For future research using the Muse 2 device, we propose several methodological refinements to enhance data quality and reliability. First, we strongly recommend utilizing a wired connection for data transmission rather than wireless connectivity. While this may introduce some constraints for ecological validity in real-time applications, our experience indicates that data loss represents a critical limitation that must be addressed to ensure the proper collection of data. Second, we

Table 10.1: Detailed dataset sizes after exclusions and filtering. Data here are reported from both phases I and phase II. Features per Task, are the number of participants and (Num. P) mean Blocks per Participant (B/P). For each level of Workload (W-Load) within Task, the features are: mean Trials per Block per Participant (T/B/P), total number of Trials (T), mean Epochs per Block per Participant (E/B/P), and total Epochs (E). Note that, although the number of trials is similar across workload levels, the total number of epochs increases as workload increases, because there will be more 1 second samples in higher workload conditions. For statistical modeling we average all epochs for each participant at the block level (and within levels of workload for within-task workload classification); for machine learning, epochs within each block, each of which contain all workload levels, are grouped together and used exclusively in the training or testing set.

Task	**Num. P**	**Mean B/P**	**W-Load**	**Mean T/B/P**	**Total T**	**Mean E/B/P**	**Total E**
Chess	16	5.4	0	5.7	556	45.3	3109
			1	6.4	631	68.2	4690
			2	10.2	989	129.6	8981
			3	8.3	758	146.9	9582
			Total	30.65	2934	389.86	26362
N-Back	11	4	0	22.3	1397	27.0	1058
			1	22.4	1345	29.0	1050
			2	23.1	1480	33.7	1328
			3	90.9	1438	36.4	1332
			Total	22.7	5660	126.1	4768
Rotation	11	3.9	0	5.8	388	11.1	433
			1	6.1	411	14.2	579
			2	6.2	409	17.2	717
			3	5.8	387	16.7	715
			Total	23.9	1595	59.2	2444
Stroop	10	3.9	0	33.9	2260	31.9	1302
			1	35.2	2317	36.6	1391
			Total	69.1	4577	68.6	2793

advocate for the implementation of signal quality assessment protocols. These should include initial verification of proper headband positioning and continuous monitoring of signal integrity throughout the experimental session as discussed above. Such proactive quality control measures would allow researchers to identify and address technical issues promptly, thereby minimizing data loss and reducing reliance on post-hoc exclusion criteria. It is important to note that the current iteration of the `muse-js` acquisition framework lacks native support for automated signal quality assessment. Consequently, researchers may need to develop custom solutions to implement these recommended quality control measures. Despite this limitation, we believe these procedural modifications will substantially improve data integrity in future studies utilizing this device.

10.2.7 Workload Labels

For our initial statistical analyses we observe the distinctions among reaction time and correctness for the tasks with multi-level classification where possible: for N-Back, difficulty is the level of the N-Back (0-back, 1-back, 2-back, or 3-back). For Rotation, difficulty is the degree of rotation (0 degrees, 50 degrees, 100 degrees, or 150 degrees). For Stroop, difficulty is congruent (0) or incongruent (1), and for chess, difficulty of a given puzzle for a given participant is the quartile expression (labeled as 0-3) of that puzzle's difficulty rating within all of the unique puzzle difficulty ratings encountered by that participant. For the machine learning analyses we simplify our labels by grouping all workload samples within a trial type in [0, 1] workload range as low workload (labeled 0), and grouping all workload samples within a trial type in [2, 3] workload range as high workload (labeled 1). Although in this orientation of the data we run the risk of classifying samples distinctly from categorizations of workload classification which are rather close together, we believe that, particularly given the data loss suffered from the Muse 2 device, the larger sample size per group is a better choice for this analysis. We encourage others to apply their own techniques to the data, which will be made publicly available.

10.2.8 Statistical Methodology

The initial part of our analysis employs statistical methods to examine the measured neural data for both within-task workload and between-tasks. Except where otherwise stated, we utilize LMMs to achieve this aim. In each model, we incorporate Participant ID and Task Block per-participant as grouping factors for random effects.

10.2.9 Modeling of Reaction Time and Correctness

To investigate behavioral metrics we modeled reaction time as a function of difficulty for all of the trials over all participants. For the N-Back, Stroop, and Rotation tasks, time is defined as the reaction time taken to press a key following the stimulus presentation. For the chess puzzle task, time is defined as the time required to solve the puzzle. Reaction time data was log-transformed for statistical modeling [266]. Omnibus test p-values for per-task models were corrected with Benjamini-Hochberg (BH) correction, as were all post-hoc comparisons. Prior to correction, pairwise

comparisons were Tukey-adjusted to control the family-wise error rate. Correctness was also modeled as a function of difficulty. Because correctness of a given trial is binary, Generalized Linear Mixed-Effect Models (GLMMs) are employed with a binomial family function for the analysis; random effects structure is the same as described in Section 10.2.8; p-values were adjusted using the same method as for the Reaction Time tests.

10.2.10 Statistical Modeling of EEG Data

Prior to beginning our EEG analysis, data for participant's participant, task type, block, and workload level were averaged. The decision to apply this averaging was due to two primary concerns: first, it addresses potential autocorrelation issues arising from the temporal dependencies inherent in time-series EEG data. Second, it enhances the signal-to-noise ratio, which is particularly important when working with consumer-grade EEG devices that typically exhibit higher baseline noise levels than research-grade systems. To analyze workload data within each task, data for each combination of participant, trial type, and wavelength were scaled to zero mean and unit variance. Individual models were then created for each unique combination of wavelength and task type. BH correction is used for the omnibus tests across wavelengths within each task type. For each significant model, pairwise comparisons were generated, correcting within wavelength-task model with Tukey adjustment. BH correction is likewise used to correct post-hoc comparisons within a given task type. To analyze EEG data across tasks, individual models were created for each wavelength. For these models, workload was ignored. BH correction is used both across all omnibus tests. For each significant model, post-hoc tests are run and initially corrected with Tukey adjustment, followed by further BH correction across all such post-hoc tests.

10.2.11 Machine Learning Analyses

To simulate real-world application in an applied BCI context, machine learning analyses were also performed on the data. For the machine learning analyses, all the available 1-second frequency-domain samples are used without averaging: this is with an eye towards future applications involving real-time classification. Given the overall complexity of the analyses in this study, the only machine learning model selected is RF: this is to create a baseline from which to work - future work can be done to explore classification accuracies of other models. To focus the analysis further, a binary

classification problem for each task is performed: specifically, the data from the N-Back, Chess, and Rotation tasks are relabeled such that the lower two classification levels are considered "Low" workload, and the higher two classification levels are considered "High" workload. There are seven features, one for each of the log-transformed power band spectra averaged between the two frontal probes. As with the statistical modeling, I approach the machine learning analysis in two contexts: within-task workload, and cross-task classification.

The machine learning modeling for this study here works under the assumption that a real-time BCI application will be developed with some training data collected from each individual - that is, rather than attempt LOO-CV, some data is used from each participant for the training set. Crucially, however, in order to prevent the models from learning auto-correlated features within each time block, training and testing sets are based on blocks of trials as shown in Figure 10.5. That is, any given block of trials taken together will only be either in the test or training set. Notably, each block contains approximately equal samples for each workload class.

Grouped and Subsampled Monte Carlo Cross-Validation

Due to data loss from the MUSE device, the fact that the chess task was collected for as long as the other three tasks combined, and due to varying time-spans of the other three tasks, our dataset contains imbalanced samples both within and across participants. However, this imbalance is not representative of the brain data itself; it stems solely from the experimental paradigm. To develop an online algorithm for differentiating between brain-states or workload levels in practice, the protocol would be modified to create a balanced dataset. Therefore, to test the data with machine learning in the fairest way possible a grouped and subsampled Monte Carlo Cross-Validation methodology was employed for both the within-task workload machine learning models and for the between-task model. The data preparation pipelines are shown in detail in Figures 10.9 and 10.10. This method involves 1,000 iterations for the cross-task workload model, and 1,000[2], where each iteration follows these steps:

1. For each participant

    (a) Split their data into train and test sets, where 20% of blocks as described in Figure 10.5

[2]For all of our models, after conducting 1,000 iterations of Monte Carlo Cross-Validation we determined that the 95 confidence interval of the mean fell within ±0.5 percentage point.

are used for testing. This grouping is crucial to avoid the model learning auto-correlated feature patterns from within a given trial block.

(b) For both test and train sets: subsample all but the minority classes randomly and without replacement. For within-class workload, separate models are produced for each trial type, so a participant's data is subsampled based on the number of samples for each workload level; for cross-task classification, the data from all trial types for a given participant are subsampled based on the number of samples for each trial type.

(c) Scale the data by calculating μ and σ for each feature in the training set; subtract μ and divide by σ for both training and testing data.

2. Balance the number of training and testing samples by subsampling all of the training and testing data (separately and without replacement) based on the combination of participant id and class label.

3. Initialize a Random Forest classifier using the `ensemble.RandomForest` function from `scikit-learn` [173] on the training data using default parameters [3]; record both overall results as well as results of per-participant classification.

Within-Task Workload Machine Learning

For within-task workload each task is modeled separately. Because in this case cross-task classification is not necessary, the entire dataset is used, including all valid data from participants in both phases I and II. See Table 10.2 for detailed information on average train and test set sizes across the Monte Carlo iterations.

Cross-Task Machine Learning

For cross-task analysis data is required for each participant from all tasks, therefore only the data collected in phase II is used for this machine learning set. For this machine learning set, there were a total of 6 participants. Across all Monte Carlo iterations there was an average of 2,723 training

[3] Model configuration included 100 trees (n_estimators), unlimited maximum depth with nodes expanded until reaching pure leaf or minimum split threshold, minimum samples for split of 2, minimum samples per leaf of 1, and Gini criterion for measuring split quality. We opted for default parameters to establish a robust baseline implementation while avoiding potential overfitting from extensive parameter tuning

Table 10.2: Number of participants per-model, and average train and test set sizes (rounded down) over all Monte Carlo iterations of within-task workload machine learning analysis. Due to the data splitting process discussed in 10.2.11, for both test and train sets the labels in the model are equally balanced for each participant. Participants are likewise equally represented in both train and test sets.

Trial type	Number of participants	Average train set sample size	Average test set sample size
Chess	17	3,983	1987
N-Back	11	1,964	622
Rotation	10	842	220
Stroop	10	1,476	404

samples and 776 testing samples. Due to the data splitting process discussed in 10.2.11, the training and testing samples were each evenly balanced within each participant for samples per label, and were likewise balanced for samples per-participant. `scikit-learn` [173] was used to implement the grouping strategy, Random Forest model, and correctness statistics. `imblearn` [267] was used for subsampling of the data. Averages are reported over all 1,000 models of the precision, recall, and macro F1-score.

10.3 Results

10.3.1 Reaction Time and Correctness

See Figure 10.11 for a visualization of reaction time and correctness data. Regarding temporal statistics, for all tasks, time to solve was significant: N-Back ($F_{3,6117.05} = 795.57, p < 0.001, \epsilon_p^2 = 0.28$)[4], Chess ($F_{3,300.01} = 300.37, p < 0.001, \epsilon_p^2 = 0.24$), Rotation ($F_{3,1554.39} = 127.54, p < 0.001, \epsilon_p^2 = 0.20$), and Stroop ($F_{1,4894.41} = 521.26, p < 0.001, \epsilon_p^2 = 0.10$). Likewise, with the exception of N-Back 2 vs. 3 ($t_{6114.86} = -2.18, p = 0.128, \epsilon_p^2 = 0.0$), the time required to solve each task increased significantly as difficulty increased. See Table 10.3 for full post-hoc contrast results. In-person informal conversations with many participants indicated that they were feeling extremely challenged by 3-back such that even 'giving up' during the task at parts was common; to this end, the finding of similar reaction times during 2-back and 3-back is sensible.

Correctness was likewise significantly correlated with workload for all tasks N-Back ($\chi^2(3, N = 6225) = 263.89, p < 0.001$), Chess ($\chi^2(3, N = 2938) = 244.88, p < 0.001$), Rotation ($\chi^2(3, N =$

[4]All p-values reported inline reflect adjustments for multiple comparisons. Tables present both unadjusted (p) and adjusted (p.adj) values

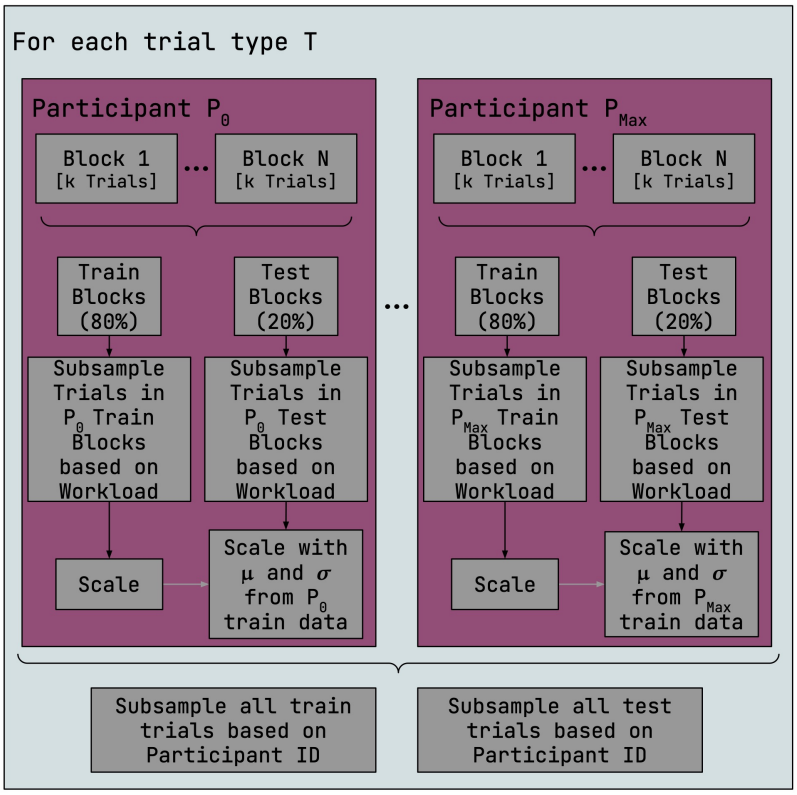


Figure 10.9: Visualization of one iteration of the data splitting steps for the Monte Carlo pipeline for within-task workload classification. Blocks of grouped trials for each participant are split into training and testing sets randomly, with 80% of blocks in training and the remaining 20% of blocks in testing. Trials within training and testing blocks are subsampled randomly to the lowest frequency of workload label. Training data are scaled, and μ and σ from the training data are used to scale the test data for that participant. All participants' training sets are then combined and subsampled to the lowest frequency participant identifier; likewise happens for the test sets. This pipeline is done separately for each Monte Carlo simulation for each trial type.

$1608) = 31.50, p < 0.001$), and Stroop ($\chi^2(1, N = 4950) = 56.20, p < 0.001$). However, correctness did not have as strong of an effect across workload within tasks. See Table 10.4 for post-hoc contrast results. Correctness was significantly different with $p < 0.001$ across all contrasts of Chess, for the Stroop contrast, all of N-Back outside of 0-1 ($z = 0.97$, $p = 0.765$), and for Rotation contrasts 0-3 and 2-3. Rotation contrast of 1-3 was significant ($z = 3.09$, $p = 0.014$). However, Rotation contrasts of 0-1 ($z = 2.12$, $p = 0.172$), 0-2 ($z = 1.12$, $p = 0.754$), and 1-2 ($z = -1.04$, $p = 0.765$) were not significant. The results for correctness were as-expected for Chess and Stroop, and largely the

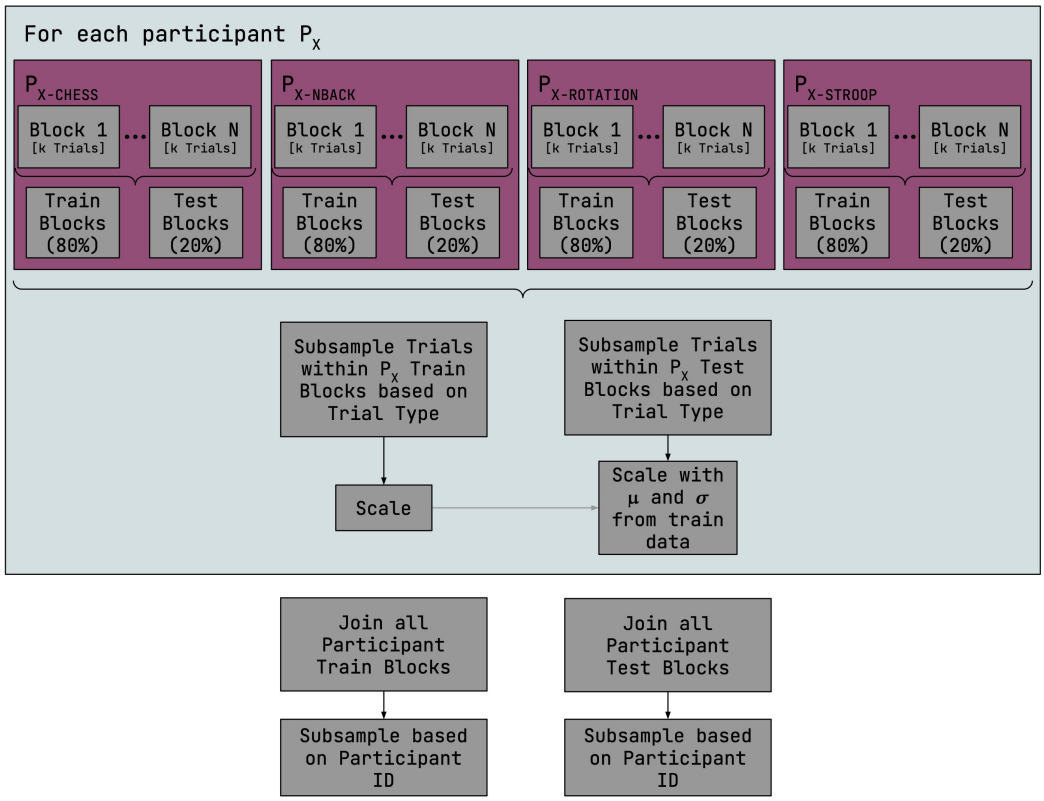


Figure 10.10: Visualization of one iteration of the data splitting steps for the Monte Carlo pipeline for cross-task classification. Blocks of grouped trials within each trial type for each participant are split into training and testing blocks randomly, with 80% of blocks in training and the remaining 20% of blocks in testing. All such training and testing blocks are combined for a given participant, and trials within the training and testing blocks are then subsampled (separately) to the lowest frequency label, and scaled. All such training and testing blocks for all participants and trial types are combined into single train and test sets, which are separately subsampled to the lowest frequency participant ID. F1 results are reported from both the precision and recall scores on the test set overall, and per-participant.

case for N-Back. However, given that 0-back and 1-back are effectively trivial tasks, near-perfect accuracy is sensible here. The Rotation task data, however, indicates that the largest degree of rotation (hardest difficulty) was the only one truly distinct from the others in terms of correctness.

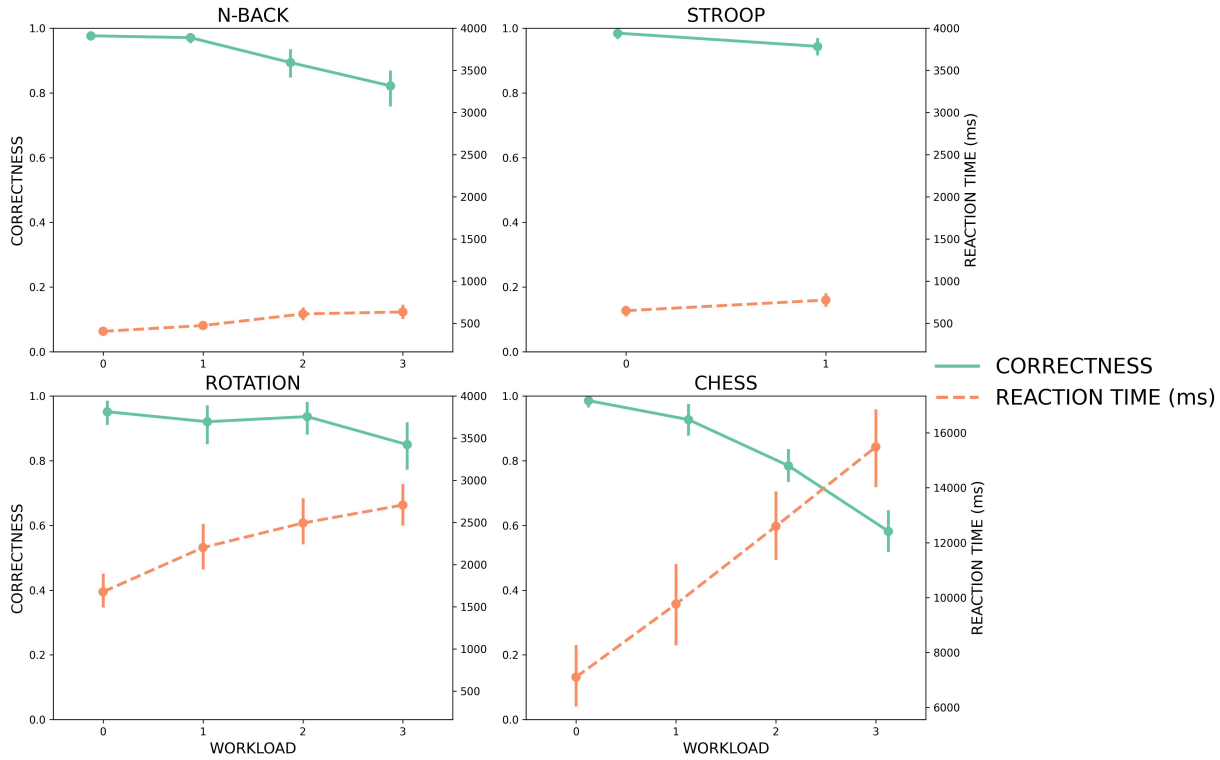


Figure 10.11: Performance and reaction time as a function of workload level for each of the four tasks. In general, as workload increases, correctness decreases and reaction time increases. Note that the temporal (right-side) scale y-scales for the three cognitive neuroscience tasks are the same (0-4000ms), however the same axis for the Chess task is from (6000-17000ms).

Table 10.3: Post-hoc contrast results for modeling reaction time as a function of workload level for each task. Separate models were created for each task type; p-values are corrected using the Benjamini-Hochberg procedure. Outside of levels 2-3 in the N-Back task, reaction time increased significantly across difficulty levels of all tasks.

Task	**Contrast**	**Est.**	**SE**	**df**	**t**	**p**	**p.adj**	**sig.**	ϵ_p^2	ϵ_p^2 **CI**
Chess	0 - 3	-0.34	0.01	2913.17	-27.70	<0.001	<0.001	***	0.21	[0.18,0.23]
Chess	0 - 2	-0.23	0.01	2886.14	-20.14	<0.001	<0.001	***	0.12	[0.10,0.15]
Chess	1 - 3	-0.23	0.01	2911.83	-19.71	<0.001	<0.001	***	0.12	[0.10,0.14]
Chess	0 - 1	-0.10	0.01	2868.08	-8.26	<0.001	<0.001	***	0.02	[0.01,0.03]
Chess	1 - 2	-0.13	0.01	2909.87	-11.55	<0.001	<0.001	***	0.04	[0.03,0.06]
Chess	2 - 3	-0.10	0.01	2821.98	-9.82	<0.001	<0.001	***	0.03	[0.02,0.05]
N-Back	0 - 3	-0.17	0.00	6118.84	-41.79	<0.001	<0.001	***	0.22	[0.20,0.24]
N-Back	0 - 2	-0.16	0.00	6124.60	-40.03	<0.001	<0.001	***	0.21	[0.19,0.22]
N-Back	1 - 3	-0.10	0.00	6111.06	-24.76	<0.001	<0.001	***	0.09	[0.08,0.10]
N-Back	0 - 1	-0.07	0.00	6115.00	-16.82	<0.001	<0.001	***	0.04	[0.03,0.05]
N-Back	1 - 2	-0.09	0.00	6117.60	-22.74	<0.001	<0.001	***	0.08	[0.07,0.09]
N-Back	2 - 3	-0.01	0.00	6114.86	-2.18	0.128	0.128	ns	0.00	[0.00,0.00]
Stroop	0 - 1	-0.07	0.00	4894.41	-22.83	<0.001	<0.001	***	0.10	[0.08,0.11]
Rotation	0 - 3	-0.23	0.01	1552.22	-18.09	<0.001	<0.001	***	0.17	[0.14,0.21]
Rotation	0 - 2	-0.19	0.01	1557.22	-15.37	<0.001	<0.001	***	0.13	[0.10,0.16]
Rotation	1 - 3	-0.11	0.01	1558.35	-8.51	<0.001	<0.001	***	0.04	[0.03,0.07]
Rotation	0 - 1	-0.12	0.01	1548.43	-9.85	<0.001	<0.001	***	0.06	[0.04,0.08]
Rotation	1 - 2	-0.07	0.01	1548.64	-5.70	<0.001	<0.001	***	0.02	[0.01,0.04]
Rotation	2 - 3	-0.04	0.01	1562.05	-2.90	0.020	0.021	*	0.01	[0.00,0.01]

Table 10.4: Post-hoc contrast results for modeling correctness as a function of workload level for each task. Separate models were created for each task type; p-values are corrected using the Benjamini-Hochberg procedure. Outside of N-Back levels 0-1 and differences in Rotation difficulty among 0, 1, and 2, workload level significantly predicts correctness in differences for all tasks.

Task	**Contrast**	**Est.**	**SE**	**df**	**z**	**p**	**p.adj**	**sig.**
Chess	0 - 3	3.96	0.37	inf	10.76	<0.001	<0.001	***
Chess	0 - 2	2.86	0.36	inf	7.87	<0.001	<0.001	***
Chess	1 - 3	2.33	0.19	inf	12.55	<0.001	<0.001	***
Chess	0 - 1	1.62	0.39	inf	4.19	<0.001	<0.001	***
Chess	1 - 2	1.24	0.17	inf	7.15	<0.001	<0.001	***
Chess	2 - 3	1.09	0.13	inf	8.69	<0.001	<0.001	***
N-Back	0 - 3	2.28	0.18	inf	12.61	<0.001	<0.001	***
N-Back	0 - 2	1.63	0.19	inf	8.74	<0.001	<0.001	***
N-Back	1 - 3	2.06	0.17	inf	12.19	<0.001	<0.001	***
N-Back	0 - 1	0.22	0.23	inf	0.97	0.765	0.765	ns
N-Back	1 - 2	1.41	0.18	inf	8.05	<0.001	<0.001	***
N-Back	2 - 3	0.66	0.11	inf	6.02	<0.001	<0.001	***
Stroop	0 - 1	1.38	0.18	inf	7.50	<0.001	<0.001	***
Rotation	0 - 3	1.38	0.28	inf	4.88	<0.001	<0.001	***
Rotation	0 - 2	0.36	0.32	inf	1.12	0.675	0.754	ns
Rotation	1 - 3	0.74	0.24	inf	3.09	0.011	0.014	*
Rotation	0 - 1	0.65	0.30	inf	2.12	0.145	0.172	ns
Rotation	1 - 2	-0.29	0.28	inf	-1.04	0.726	0.765	ns
Rotation	2 - 3	1.03	0.26	inf	4.03	<0.001	<0.001	***

10.3.2 Statistical Analysis of Task-Specific Cognitive Load in EEG Signals

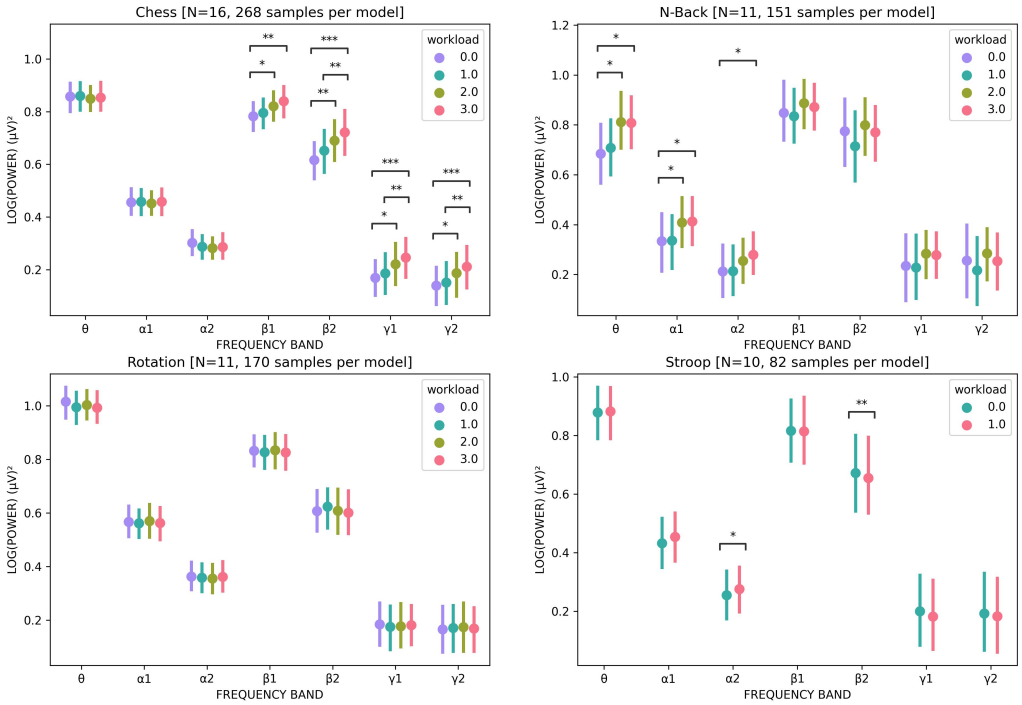


Figure 10.12: EEG power band data as compared across levels of workload within each task. Across workload levels Chess showed significant differences in high-frequency bands (β_1, β_2, γ_1, γ_2). Conversely, N-Back showed significant differences in the lower-frequency bands (θ, α_1, α_2). Stroop showed significance in α_2 and β_2, and Rotation did not show any significant differences across power bands.

Chess

Significant workload effects are observed across all relatively high frequency bands in the Chess task: β_1 ($F_{3,197.30} = 6.27, p = 0.001, \epsilon_p^2 = 0.07$), β_2 ($F_{3,197.39} = 13.39, p < 0.001, \epsilon_p^2 = 0.16$), γ_1 ($F_{3,197.40} = 10.29, p < 0.001, \epsilon_p^2 = 0.12$), and γ_2 ($F_{3,197.41} = 8.35, p < 0.001, \epsilon_p^2 = 0.10$). Post-hoc analyses (see Table 10.6) revealed the most substantial effects in all frequency bands between the largest gaps in workload levels. For the contrast (0-3): β_1: $t_{197.61} = -3.96, p = 0.002, \epsilon_p^2 = 0.07$; β_2: $t_{197.78} = -6.01, p < 0.001, \epsilon_p^2 = 0.15$; γ_1: $t_{197.80} = -5.12, p < 0.001, \epsilon_p^2 = 0.11$; γ_2: $t_{197.82} = -4.49, p = 0.001, \epsilon_p^2 = 0.09$), and the contrast (0-2): β_1: $t_{197.01} = -3.01, p = 0.038, \epsilon_p^2 = 0.04$;

Table 10.5: ANOVA results modeling EEG band data as a function of within-task workload. Separate LMER models were created for each unique combination of wavelength level and task. P-values are corrected with Benjamini-Hochberg procedure across all models. With increased workload levels, Chess showed changes in the upper range of frequency bands, and N-Back showed increased power in the lower range of frequency bands. Only β_2 was the only significant wavelength in the Stroop task, although there is a notably high effect size for α_2. No changes were observed in the Rotation task.

Task	**Band**	**df1**	**df2**	**F**	**p**	**p.adj**	**sig.**	ϵ_p^2	ϵ_p^2 **CI**
Chess	θ	3.00	197.53	0.18	0.911	0.911	ns	0.00	[0.00, 0.00]
Chess	α_1	3.00	197.45	0.35	0.791	0.911	ns	0.00	[0.00, 0.00]
Chess	α_2	3.00	197.36	2.58	0.055	0.077	ns	0.02	[0.00, 0.06]
Chess	β_1	3.00	197.30	6.27	<0.001	0.001	***	0.07	[0.02, 0.13]
Chess	β_2	3.00	197.39	13.39	<0.001	<0.001	***	0.16	[0.08, 0.23]
Chess	γ_1	3.00	197.40	10.29	<0.001	<0.001	***	0.12	[0.05, 0.19]
Chess	γ_2	3.00	197.41	8.35	<0.001	<0.001	***	0.10	[0.04, 0.16]
N-Back	θ	3.00	113.44	4.88	0.003	0.011	*	0.09	[0.01, 0.17]
N-Back	α_1	3.00	111.70	6.46	<0.001	0.003	**	0.12	[0.03, 0.21]
N-Back	α_2	3.00	113.45	4.39	0.006	0.014	*	0.08	[0.01, 0.15]
N-Back	β_1	3.00	107.80	2.13	0.101	0.157	ns	0.03	[0.00, 0.08]
N-Back	β_2	3.00	109.17	1.22	0.306	0.306	ns	0.01	[0.00, 0.02]
N-Back	γ_1	3.00	109.49	1.75	0.161	0.188	ns	0.02	[0.00, 0.06]
N-Back	γ_2	3.00	107.74	2.05	0.112	0.157	ns	0.03	[0.00, 0.07]
Stroop	θ	1.00	40.00	0.00	0.980	0.980	ns	0.00	[0.00, 0.00]
Stroop	α_1	1.00	40.00	3.15	0.084	0.147	ns	0.05	[0.00, 0.19]
Stroop	α_2	1.00	40.00	6.58	0.014	0.050	*	0.12	[0.01, 0.29]
Stroop	β_1	1.00	40.00	0.08	0.772	0.901	ns	0.00	[0.00, 0.00]
Stroop	β_2	1.00	40.00	9.32	0.004	0.028	*	0.17	[0.03, 0.34]
Stroop	γ_1	1.00	40.00	4.43	0.042	0.097	ns	0.08	[0.00, 0.23]
Stroop	γ_2	1.00	40.00	0.17	0.682	0.901	ns	0.00	[0.00, 0.00]
Rotation	θ	3.00	124.32	0.86	0.465	0.926	ns	0.00	[0.00, 0.00]
Rotation	α_1	3.00	124.47	0.17	0.917	0.926	ns	0.00	[0.00, 0.00]
Rotation	α_2	3.00	124.72	0.19	0.901	0.926	ns	0.00	[0.00, 0.00]
Rotation	β_1	3.00	124.18	0.26	0.852	0.926	ns	0.00	[0.00, 0.00]
Rotation	β_2	3.00	124.18	1.31	0.273	0.926	ns	0.01	[0.00, 0.02]
Rotation	γ_1	3.00	124.22	0.34	0.796	0.926	ns	0.00	[0.00, 0.00]
Rotation	γ_2	3.00	124.20	0.15	0.926	0.926	ns	0.00	[0.00, 0.00]

β_2: $t_{197.01} = -3.79, p = 0.004, \epsilon_p^2 = 0.06$; γ_1: $t_{197.01} = -3.26, p = 0.019, \epsilon_p^2 = 0.05$; γ_2: $t_{197.01} = -2.95, p = 0.040, \epsilon_p^2 = 0.04$). Further, all except the β_1 band showed significance for the contrasts (1-3): β_2 ($t_{197.78} = -4.26, p = 0.001, \epsilon_p^2 = 0.08$); γ_1 ($t_{197.80} = -4.00, p = 0.002, \epsilon_p^2 = 0.07$); and γ_2 ($t_{197.82} = -3.71, p = 0.005, \epsilon_p^2 = 0.06$).

These findings suggest a strong distinction among the larger separation of workload levels for

Table 10.6: Post-hoc contrast results modeling changes in EEG band data as a function of within-task workload for the Chess task. Increases in power were found in among the most separable workload levels (0-3 and 0-2) for all the frequency bands. Increases in power between workload levels 1-3 were likewise found in all except β_1.

Task	Band	Contrast	Est.	SE	df	t	p	p.adj	sig.	ϵ_p^2	ϵ_p^2 CI
Chess	β_1	0 - 3	-0.45	0.11	197.61	-3.96	0.001	0.002	**	0.07	[0.02, 0.15]
Chess	β_1	0 - 2	-0.33	0.11	197.01	-3.01	0.016	0.038	*	0.04	[0.00, 0.10]
Chess	β_1	1 - 3	-0.31	0.11	197.61	-2.72	0.035	0.070	ns	0.03	[0.00, 0.09]
Chess	β_1	0 - 1	-0.14	0.11	197.01	-1.27	0.586	0.669	ns	0.00	[0.00, 0.04]
Chess	β_1	1 - 2	-0.19	0.11	197.01	-1.74	0.305	0.386	ns	0.01	[0.00, 0.05]
Chess	β_1	2 - 3	-0.12	0.11	197.61	-1.02	0.740	0.772	ns	0.00	[0.00, 0.02]
Chess	β_2	0 - 3	-0.72	0.12	197.78	-6.01	<0.001	<0.001	***	0.15	[0.07, 0.24]
Chess	β_2	0 - 2	-0.44	0.12	197.01	-3.79	0.001	0.004	**	0.06	[0.01, 0.14]
Chess	β_2	1 - 3	-0.51	0.12	197.78	-4.26	<0.001	0.001	**	0.08	[0.02, 0.16]
Chess	β_2	0 - 1	-0.21	0.12	197.01	-1.79	0.282	0.376	ns	0.01	[0.00, 0.06]
Chess	β_2	1 - 2	-0.23	0.12	197.01	-2.00	0.193	0.289	ns	0.01	[0.00, 0.06]
Chess	β_2	2 - 3	-0.27	0.12	197.78	-2.30	0.102	0.187	ns	0.02	[0.00, 0.08]
Chess	γ_1	0 - 3	-0.63	0.12	197.80	-5.12	<0.001	<0.001	***	0.11	[0.04, 0.20]
Chess	γ_1	0 - 2	-0.39	0.12	197.01	-3.26	0.007	0.019	*	0.05	[0.01, 0.12]
Chess	γ_1	1 - 3	-0.49	0.12	197.80	-4.00	0.001	0.002	**	0.07	[0.02, 0.15]
Chess	γ_1	0 - 1	-0.14	0.12	197.01	-1.14	0.662	0.722	ns	0.00	[0.00, 0.03]
Chess	γ_1	1 - 2	-0.25	0.12	197.01	-2.12	0.151	0.242	ns	0.02	[0.00, 0.07]
Chess	γ_1	2 - 3	-0.23	0.12	197.80	-1.92	0.223	0.315	ns	0.01	[0.00, 0.06]
Chess	γ_2	0 - 3	-0.56	0.12	197.82	-4.49	<0.001	0.001	***	0.09	[0.03, 0.17]
Chess	γ_2	0 - 2	-0.36	0.12	197.01	-2.95	0.018	0.040	*	0.04	[0.00, 0.10]
Chess	γ_2	1 - 3	-0.46	0.12	197.82	-3.71	0.002	0.005	**	0.06	[0.01, 0.13]
Chess	γ_2	0 - 1	-0.10	0.12	197.01	-0.80	0.856	0.856	ns	0.00	[0.00, 0.00]
Chess	γ_2	1 - 2	-0.26	0.12	197.01	-2.16	0.140	0.239	ns	0.02	[0.00, 0.07]
Chess	γ_2	2 - 3	-0.20	0.12	197.82	-1.60	0.382	0.459	ns	0.01	[0.00, 0.05]

the Chess task in the high frequency bands. With increased power in the β and γ frequencies generally associated with active cognitive processing, attention, decision making, and mental effort [268,269], the progressive increase in high-frequency band power from lower to higher workload conditions supports the interpretation that participants were engaging more cognitive resources as task complexity increased, particularly when comparing across 2 or 3 difficulty levels.

N-Back

In the N-Back task significant workload effects are observed in the lower frequency bands: θ ($F_{3,113.44} = 4.88, p = 0.011, \epsilon_p^2 = 0.09$), α_1 ($F_{3,111.70} = 6.46, p = 0.003, \epsilon_p^2 = 0.12$), and α_2 ($F_{3,113.45} = 4.39, p = 0.014, \epsilon_p^2 = 0.08$). Similar to the differentiations seen with the Chess task, post-hoc

Table 10.7: Post-hoc contrast results modeling EEG band data as a function of within-task workload for the N-Back task. Increases in power were found between workload levels 0-3 for all three bands, and between workload levels and 0-2 for both the θ and α_1 bands.

Task	**Band**	**Contrast**	**Est.**	**SE**	**df**	**t**	**p**	**p.adj**	**sig.**	ϵ_p^2	ϵ_p^2 **CI**
N-Back	θ	0 - 3	-0.69	0.22	112.79	-3.09	0.013	0.048	*	0.07	[0.01, 0.18]
N-Back	θ	0 - 2	-0.70	0.22	112.63	-3.20	0.010	0.048	*	0.08	[0.01, 0.18]
N-Back	θ	1 - 3	-0.45	0.23	112.24	-1.97	0.207	0.339	ns	0.02	[0.00, 0.11]
N-Back	θ	0 - 1	-0.24	0.22	113.63	-1.08	0.703	0.953	ns	0.00	[0.00, 0.04]
N-Back	θ	1 - 2	-0.46	0.22	115.15	-2.05	0.177	0.319	ns	0.03	[0.00, 0.11]
N-Back	θ	2 - 3	0.01	0.22	114.33	0.06	1.000	1.000	ns	0.00	[0.00, 0.00]
N-Back	α_1	0 - 3	-0.67	0.21	111.28	-3.22	0.009	0.048	*	0.08	[0.01, 0.19]
N-Back	α_1	0 - 2	-0.70	0.20	111.20	-3.42	0.005	0.048	*	0.09	[0.01, 0.20]
N-Back	α_1	1 - 3	-0.58	0.21	110.28	-2.76	0.034	0.077	ns	0.06	[0.00, 0.16]
N-Back	α_1	0 - 1	-0.08	0.21	111.94	-0.41	0.977	1.000	ns	0.00	[0.00, 0.00]
N-Back	α_1	1 - 2	-0.61	0.21	113.12	-2.93	0.021	0.063	ns	0.06	[0.00, 0.17]
N-Back	α_1	2 - 3	0.03	0.21	112.48	0.15	0.999	1.000	ns	0.00	[0.00, 0.00]
N-Back	α_2	0 - 3	-0.70	0.22	112.79	-3.14	0.012	0.048	*	0.07	[0.01, 0.18]
N-Back	α_2	0 - 2	-0.47	0.22	112.64	-2.15	0.143	0.286	ns	0.03	[0.00, 0.12]
N-Back	α_2	1 - 3	-0.63	0.23	112.25	-2.75	0.034	0.077	ns	0.05	[0.00, 0.15]
N-Back	α_2	0 - 1	-0.07	0.23	113.64	-0.33	0.988	1.000	ns	0.00	[0.00, 0.00]
N-Back	α_2	1 - 2	-0.40	0.23	115.16	-1.77	0.291	0.436	ns	0.02	[0.00, 0.09]
N-Back	α_2	2 - 3	-0.23	0.22	114.33	-1.01	0.741	0.953	ns	0.00	[0.00, 0.03]

analyses (see Table 10.7) revealed the strongest effect between levels 0-3, represented in the θ ($t_{112.79} = -3.09, p = 0.048, \epsilon_p^2 = 0.07$), α_1 ($t_{111.28} = -3.22, p = 0.048, \epsilon_p^2 = 0.08$), and α_2 ($t_{112.79} = -3.14, p = 0.048, \epsilon_p^2 = 0.07$) frequencies. Effects between workload levels 0-2 were also represented in two out of three bands: θ ($t_{112.63} = -3.20, p = 0.048, \epsilon_p^2 = 0.08$), and α_1 ($t_{111.20} = -3.42, p = 0.048, \epsilon_p^2 = 0.10$). These findings suggest a distinct pattern in the N-Back task with workload effects primarily manifesting in lower frequency bands. The prominence of theta and alpha band activity aligns with established research linking these frequencies to working memory operations, particularly the theta band's association with repetitive task load [270], and the alpha band's association selective attention, inhibition, and controlled access to stored information [271]. Similar to the Chess task, the clearest differentiation between workload levels for the N-Back task showed similar effect sizes between the largest workload contrasts, suggesting that the subtlest distinctions in workload are not visible in the data. The overall pattern of increased lower frequency power with heightened working memory load reflects the task's core demand on memory maintenance and manipulation processes, distinguishing it from the more complex cognitive processing demands

observed in the Chess task.

Stroop

Refer to Table 10.5. In the Stroop task we found significant distinction between congruent and incongruent stimuli (0-1) in the β_2 ($F_{1,40.00} = 9.32, p = 0.004, \epsilon_p^2 = 0.17$) and α_2 bands ($F_{1,40.00} = 6.58, p = 0.050, \epsilon_p^2 = 0.12$); we therefore ran post hoc tests on both. This revealed significance for both, showing α_2 increasing during the incongruent task ($t_{40.00} = -2.57$, $p = 0.014$, $\epsilon_p^2 = 0.12$), and β_2 decreasing during incongruent stimuli ($t_{40.00} = 2.57$, $p = 0.008$, $\epsilon_p^2 = 0.17$). Previous work has linked the Stroop task to a decline in beta band power, which may reflect cognitive control mechanisms involved in conflict resolution [272]. Supporting this finding, [273] observed a similar decrease in β band power during Stroop task performance, suggesting a consistent neural signature of conflict processing. The combination of this observation, along with an increase in α power consistent with the inhibitory mechanisms previously described, suggests a coordinated neural response during conflict resolution in the Stroop task.

Rotation

Notably, analysis revealed no significant within-task workload differences across difficulty levels in the Rotation task. This consistent pattern of neural engagement suggests that cognitive demands may not increase linearly with rotation angles, indicating that participants likely engaged similar neural mechanisms regardless of the rotation magnitude required. It is also possible, however, that the objectivity of the rating scale used may have confounded 'true' difficulty. In [107]'s similar work, difficulty levels were instead deduced from subjective ratings - it is possible that such difficulty scales might yield better results for this task.

10.3.3 Statistical Analysis of Cross-Task EEG Signal Differentiation

ANOVA results (Table 10.8) revealed significant differences between tasks across all frequency bands ($F_{3,97.81-98.86} = 6.31 - 24.45, allp < 0.001, \epsilon_p^2 = 0.14 - 0.34$), with the strongest effects observed in lower frequency ranges: α_1 ($\epsilon_p^2 = 0.41$), α_2 ($\epsilon_p^2 = 0.34$), and θ ($\epsilon_p^2 = 0.34$). Post-hoc analyses (Table 10.9 and Figure 10.13) revealed distinct patterns across tasks which are discussed below.

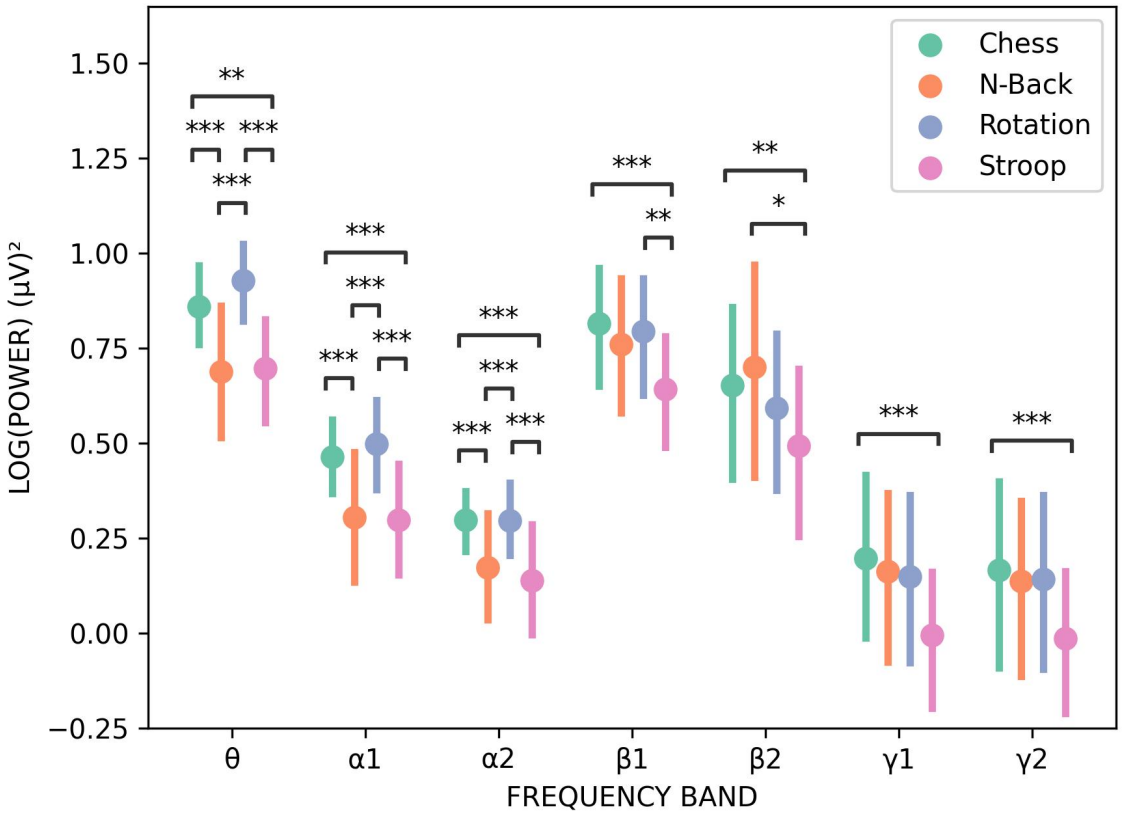


Figure 10.13: EEG spectral power compared across tasks, irrespective of workload level within each task. Differences in spectral power were observed across all frequency bands. Most notably, Chess shows significantly higher power than Stroop across all bands. Rotation likewise shows higher power than Stroop across multiple bands, but not in significantly in the high-frequency ranges. Both Chess and Rotation show higher power than N-Back in the lower frequency bands.

Table 10.8: ANOVA results from modeling EEG waveband data as a function of task type. To avoid multicollinearity between frequency bands, separate models were created for each waveband. P-values were adjusted using the Benjamini-Hochberg procedure to control for multiple comparisons. Significant differences in neural activity among tasks were observed across all frequency bands analyzed.

Band	**df1**	**df2**	**F**	**p**	**p.adj**	**sig.**	ϵ_p^2	ϵ_p^2 **CI**
θ	3.00	98.86	18.32	<0.001	<0.001	***	0.34	[0.21, 0.44]
α_1	3.00	98.26	24.45	<0.001	<0.001	***	0.41	[0.28, 0.51]
α_2	3.00	98.30	18.78	<0.001	<0.001	***	0.34	[0.21, 0.45]
β_1	3.00	98.00	9.22	<0.001	<0.001	***	0.20	[0.08, 0.30]
β_2	3.00	97.81	6.86	<0.001	<0.001	***	0.15	[0.04, 0.24]
γ_1	3.00	97.83	8.53	<0.001	<0.001	***	0.18	[0.07, 0.28]
γ_2	3.00	97.91	6.31	0.001	0.001	***	0.14	[0.03, 0.23]

Table 10.9: Post-hoc contrast results from modeling EEG waveband data as a function of task. Significant differences were discovered across a variety of frequency bands for varying tasks, indicating that differing neural signatures associated with the tasks are measurable with the Muse 2 device. Among wavebands, effect sizes in the relatively low bands (θ, α_1, α_2) are the largest, with a notable exceptions of increased power of Chess as compared to Stroop across all bands.

Band	**Task1**	**Task2**	**Est.**	**SE**	**df**	**t**	**p**	**p.adj**	**sig.**	ϵ_p^2	ϵ_p^2 **CI**
θ	Chess	N-Back	0.17	0.04	94.73	4.62	<0.001	<0.001	***	0.18	[0.06, 0.31]
θ	Chess	Rotation	-0.06	0.04	92.75	-1.51	0.436	0.555	ns	0.01	[0.00, 0.09]
θ	Chess	Stroop	0.17	0.04	106.26	4.13	<0.001	0.001	**	0.13	[0.03, 0.25]
θ	N-Back	Rotation	-0.23	0.04	93.06	-6.04	<0.001	<0.001	***	0.27	[0.13, 0.41]
θ	N-Back	Stroop	-0.00	0.04	106.47	-0.12	0.999	0.999	ns	0.00	[0.00, 0.00]
θ	Rotation	Stroop	0.23	0.04	104.01	5.58	<0.001	<0.001	***	0.22	[0.10, 0.35]
α_1	Chess	N-Back	0.15	0.03	94.17	5.50	<0.001	<0.001	***	0.24	[0.10, 0.37]
α_1	Chess	Rotation	-0.03	0.03	92.21	-1.20	0.626	0.773	ns	0.00	[0.00, 0.07]
α_1	Chess	Stroop	0.16	0.03	104.68	5.27	<0.001	<0.001	***	0.20	[0.08, 0.33]
α_1	N-Back	Rotation	-0.19	0.03	93.12	-6.62	<0.001	<0.001	***	0.31	[0.17, 0.45]
α_1	N-Back	Stroop	0.01	0.03	105.60	0.20	0.997	0.999	ns	0.00	[0.00, 0.00]
α_1	Rotation	Stroop	0.19	0.03	103.52	6.46	<0.001	<0.001	***	0.28	[0.15, 0.41]
α_2	Chess	N-Back	0.13	0.03	94.23	5.10	<0.001	<0.001	***	0.21	[0.08, 0.35]
α_2	Chess	Rotation	0.01	0.03	92.28	0.25	0.995	0.999	ns	0.00	[0.00, 0.00]
α_2	Chess	Stroop	0.16	0.03	104.76	5.69	<0.001	<0.001	***	0.23	[0.10, 0.36]
α_2	N-Back	Rotation	-0.12	0.03	93.13	-4.76	<0.001	<0.001	***	0.19	[0.07, 0.32]
α_2	N-Back	Stroop	0.03	0.03	105.62	1.00	0.750	0.900	ns	0.00	[0.00, 0.00]
α_2	Rotation	Stroop	0.15	0.03	103.52	5.50	<0.001	<0.001	***	0.22	[0.09, 0.35]
β_1	Chess	N-Back	0.11	0.04	95.88	2.70	0.041	0.082	ns	0.06	[0.00, 0.17]
β_1	Chess	Rotation	0.07	0.04	94.70	1.76	0.299	0.406	ns	0.02	[0.00, 0.11]
β_1	Chess	Stroop	0.23	0.04	101.68	5.15	<0.001	<0.001	***	0.20	[0.08, 0.33]
β_1	N-Back	Rotation	-0.04	0.04	94.86	-0.88	0.817	0.953	ns	0.00	[0.00, 0.00]
β_1	N-Back	Stroop	0.12	0.04	101.85	2.66	0.044	0.084	ns	0.06	[0.00, 0.16]
β_1	Rotation	Stroop	0.15	0.04	100.42	3.50	0.004	0.009	**	0.10	[0.02, 0.22]
β_2	Chess	N-Back	0.05	0.06	96.49	0.80	0.854	0.970	ns	0.00	[0.00, 0.00]
β_2	Chess	Rotation	0.15	0.06	95.68	2.58	0.054	0.095	ns	0.06	[0.00, 0.17]
β_2	Chess	Stroop	0.26	0.06	100.33	4.17	<0.001	0.001	**	0.14	[0.04, 0.27]
β_2	N-Back	Rotation	0.11	0.06	95.60	1.82	0.272	0.380	ns	0.02	[0.00, 0.11]
β_2	N-Back	Stroop	0.21	0.06	100.26	3.45	0.004	0.010	*	0.10	[0.02, 0.22]
β_2	Rotation	Stroop	0.11	0.06	99.21	1.70	0.327	0.429	ns	0.02	[0.00, 0.10]
γ_1	Chess	N-Back	0.13	0.05	96.27	2.35	0.094	0.147	ns	0.04	[0.00, 0.15]
γ_1	Chess	Rotation	0.14	0.06	95.35	2.50	0.067	0.113	ns	0.05	[0.00, 0.16]
γ_1	Chess	Stroop	0.29	0.06	100.64	5.06	<0.001	<0.001	***	0.19	[0.07, 0.33]
γ_1	N-Back	Rotation	0.01	0.06	95.38	0.21	0.997	0.999	ns	0.00	[0.00, 0.00]
γ_1	N-Back	Stroop	0.17	0.06	100.69	2.89	0.024	0.053	ns	0.07	[0.00, 0.18]
γ_1	Rotation	Stroop	0.15	0.06	99.55	2.69	0.041	0.082	ns	0.06	[0.00, 0.17]
γ_2	Chess	N-Back	0.12	0.06	96.35	2.07	0.170	0.255	ns	0.03	[0.00, 0.13]
γ_2	Chess	Rotation	0.11	0.06	95.42	1.85	0.257	0.373	ns	0.02	[0.00, 0.11]
γ_2	Chess	Stroop	0.27	0.06	100.85	4.34	<0.001	0.001	***	0.15	[0.04, 0.28]
γ_2	N-Back	Rotation	-0.01	0.06	95.36	-0.17	0.998	0.999	ns	0.00	[0.00, 0.00]
γ_2	N-Back	Stroop	0.15	0.06	100.81	2.42	0.079	0.128	ns	0.05	[0.00, 0.15]
γ_2	Rotation	Stroop	0.16	0.06	99.61	2.59	0.053	0.095	ns	0.05	[0.00, 0.16]

From these data a few notable trends arise. Firstly are the similarities within the power disparities in the low-to-middle frequency bands (θ, α_1, and α_2) between two groups of two tasks: Chess and Rotation, both significantly higher in all such bands (see Table 10.9) than the N-Back and Stroop tasks. This clustering suggests these tasks may share underlying cognitive mechanisms, particularly in how they engage spatial processing and mental manipulation of visual information. Significantly, N-Back and Stroop tasks involve rapid, sequential trials, each of which involves less cognitive power than the overall activity of the Chess and rotation tasks.

The pattern between Chess, Rotation, and Stroop persists in the β_1 band, with a notable exception: the N-Back task demonstrates distinct characteristics. This divergence likely stems from the N-Back task's intensive working memory demands, which appear to engage higher frequency bands more substantially. Indeed, N-Back shows higher power than Stroop in the β_2 band, potentially a consequence of the decreased β_2 activation due to conflict inhibition discussed in the within-task results for Stroop. Chess continues to have significantly higher power than Stroop in the highest bands (β_2, γ_1, and γ_2).

Taken as a whole, these findings demonstrate the Muse 2 device's capability to differentiate between complex neural activation patterns across varied cognitive tasks. The device's ability to capture these distinct neural signatures suggests promising applications for cognitive task classification and analysis.

10.3.4 Within-Task Workload Machine Learning Results

As shown in Table 10.10, the within-task workload classification presented significant challenges. The N-Back task achieved a 63% F1-score, suggesting some promise, while other tasks performed near chance levels when using the Muse 2 device for online classification. Several factors may explain these results. First, our limited dataset size may have constrained the models' learning capabilities. Additionally, classification accuracy might improve through either the incorporation of additional features or the use of alternative models that process raw data directly rather than using frequency-domain transformations. These results demonstrate promising potential for applying the Muse 2 device and similar consumer-grade EEG systems to distinguish between high and low workload levels as defined by the working-memory based cognitive load induced by the N-Back task. With additional data collection and refinement, this capability could potentially extend to

chess-based cognitive load classification. However, our findings suggest that the Muse 2 may not be suitable for BCI applications that rely on cognitive inhibition or mental rotation tasks, and we recommend careful consideration before deploying it in such contexts.

Table 10.10: Task-specific classification of cognitive workload using machine learning. N-Back performs the best, with 63% F1-score over 2-class classification. Chess performs just better than chance, with 54%. The Rotation and Stroop task classifiers do not perform better than chance. Note that the results are averages over all Monte-Carlo iterations.

Task	Workload	Precision	Recall	F1	Support
Chess	0	0.540	0.553	0.543	994
	1	0.540	0.522	0.528	994
	Macro Average	0.539	0.538	0.535	1987
N-Back	0	0.622	0.662	0.641	311
	1	0.638	0.596	0.616	311
	Macro Average	0.630	0.629	0.628	622
Rotation	0	0.492	0.513	0.501	110
	1	0.492	0.471	0.481	110
	Macro Average	0.492	0.492	0.491	220
Stroop	0	0.507	0.572	0.538	202
	1	0.509	0.443	0.474	202
	Macro Average	0.508	0.508	0.506	404

10.3.5 Cross-Task Machine Learning Results

Full results are shown in Table 10.11. The machine learning model for between-task classification demonstrated robust performance, achieving a macro-average F1-score of 49% (95% confidence interval of [0.493, 0.495]) compared to the expected random chance performance of 25%. The Rotation task showed the strongest classification metric with an F1-score of 53%, while Chess, Stroop, and N-Back tasks achieved F1-scores of 50%, 49%, and 45%, respectively. This performance is particularly noteworthy given the inherent challenges of EEG classification. The balanced precision and recall scores across all tasks (ranging from approximately 46% to 56%) indicate consistent and reliable classification capabilities. The relatively similar performance levels of the tasks align with our earlier findings regarding unique combinations of neural activation patterns discussed earlier. These classification results demonstrate that the Muse 2 device can effectively differentiate between

distinct cognitive tasks, with performance approximately twice that of random chance across all task types.

Table 10.11: Cross-task classification of EEG signals using machine learning. These results are averages across all Monte Carlo iterations.

Task	**Precision**	**Recall**	**F1**	**Support**
Chess	0.552	0.460	0.499	195
N-Back	0.485	0.425	0.452	195
Rotation	0.493	0.580	0.531	195
Stroop	0.474	0.519	0.494	195
Macro Average	0.501	0.496	0.494	780

10.4 Future Directions and Methodological Considerations

10.4.1 Signal Processing Enhancements

Our current epoching and frequency-domain transformation approach, while effective for real-time applications, could be complemented by alternative analytical methods. Event-related potential (ERP) analysis would enable precise examination of temporal relationships between task events and neural responses, potentially revealing workload-sensitive components like the P300 or N400. This approach would be particularly valuable for the Stroop and Mental Rotation tasks, where specific cognitive processes occur in response to stimulus presentation. Advanced time-frequency methods, such as wavelet transforms or variable-window short-time Fourier transforms, could provide finer temporal resolution at higher frequencies while maintaining adequate frequency resolution for lower bands. These methods could capture subtle workload-related changes and transient spectral patterns that may be averaged out in our fixed-window approach. These might be combined with longer-epoch windows, enabling finer-grained insights into the temporal dynamics of cognitive workload across different task phases.

10.4.2 Further Machine Learning Applications

As noted, this study only focused on RF as the machine learning model of choice. However, future work could attempt multiple classification models as shown throughout the other studies in this

dissertation; potentially, deep learning methods could likewise be applied to the data.

10.5 Conclusion

This study yields several significant findings regarding the capabilities of the Muse 2 device for cognitive analysis and brain-computer interface applications. I demonstrated that the Muse 2 can detect subtle variations in differing forms cognitive load across both traditional experimental paradigms (N-Back and Stroop tasks) and more ecologically valid scenarios (Chess puzzles task). The cross-task analysis revealed distinct neural signatures that show promise for BCI applications. Our Monte Carlo cross-validation procedure provided robust testing of the system's predictive capabilities for real-time BCI implementations. The within-task workload classification results indicate that while the Muse 2 can effectively track cognitive load gradients in the N-Back task, it may have limitations for fine-grained distinctions within Chess, Rotation, or Stroop tasks for online classification. This suggests opportunities for future research exploring advanced preprocessing techniques and alternative machine learning approaches. Notably, cross-task classification results demonstrate particularly promising applications for the Muse 2 in BCI research. Rather than focusing on within-task workload assessment, the device shows significant potential for broader task differentiation, indicating that real-time BCIs can be developed for meaningful human state classification. In contrast to the strong potential for specific localization-based classification as shown with fNIRS in parts II and III, this work demonstrates that leveraging the "weakness" of EEG can be particularly useful for complex state differentiation. In conclusion, this research establishes that consumer-grade EEG devices like the Muse 2 offer viable solutions for detecting mental workload states and classifying neurological patterns across diverse cognitive tasks. While challenges remain, these findings open up exciting possibilities for future development of adaptive brain-computer interfaces.

PART V

Conclusions

Across the three major research projects - examining PFC activity during LLM use, implementing a real-time fNIRS-based BCI leveraging brain networks, and exploring complex state classification with consumer-grade EEG - we have investigated the potential of PFC for next-generation implicit BCIs. Each study has revealed both promising directions for understanding and future classification, as well as help to define meaningful boundaries which limit and contextualize the potential for future work. We now turn to synthesizing the findings, extrapolating thematic patterns, and discussing their implications for the future of implicit BCI that moves 'beyond workload'. To that end, in Chapter 11 I summarize the collective findings of this work, provide concrete insights abstracting their importance towards the next generation of implicit BCIs which leverage the PFC, and consider the nuances of human-based signal interpretation via statistics and machine learning; in Chapter 12 I synthesize these ideas towards the introduction of the concept of Human-Signal-Computer Interaction; in Chapter 13 I discuss limitations of this dissertation; and in Chapter 14 I conclude this work.

Chapter 11

Key Findings

I will initially limit the discussion here to findings in this work related to the brain. To that end, the primary findings from each of the studies are summarized below.

11.1 Summary of BCI-related findings

Part II: PFC during LLM use

- Study I: PFC patterns during LLM-use across the gradient of subjectivity
  - LLM-use during reading comprehension tasks elicits measurable and actionable PFC activity with fNIRS as compared to baseline reading comprehension when using machine learning; however, this finding is not reflected the statistical analysis of the data.
  - Episodic memory linked to right prefrontal activation of the PFC irrespective of LLM-tool use is visible with LMM modeling; initial machine learning results indicate that this may be a potential application for implicit BCI systems leveraging the PFC with fNIRS.
- Study II: Complex decision-making
  - LLM use during complex decision-making elicits measurable patterns of hemodynamic activity in the right PFC as compared to baseline complex decision-making during long (25m) tasks if one has already been familiar with the baseline; although these patterns are not currently operationalizable with machine learning in BCI contexts, there may be

potential for future work leveraging such classification with tasks that are more separable in-terms of mental workload with and without the application of the LLM tool.

Part III: real-time lPFC-mPFC based BCI with fNIRS

- Real-time implicit BCI is possible with brain-network based machine learning classification as the basis for differentiation, however this finding was not reflected in statistical analyses of the data. Classification techniques which leverage different meta-classifiers for different regions or networks of interest may be beneficial to real-time applications.

Part IV: Muse 2 EEG for BCI

- Chess move quality
  - β and γ band power increase with increasing quality of chess moves. Leveraging these insights towards real-time BCIs with machine learning remains a challenge, but potential could be improved with meta-classifiers, or potentially model-based cross-validation per-participant.
- Within-task workload and cross-task classification
  - β and γ band power increase with increasing difficulty of chess puzzles. θ and α band power increase with N-Back difficulty. A mild increase in $\alpha 2$ and decrease in $\beta 2$ band power were observed in the Stroop incongruent condition as compared to the congruent.
  - N-Back performs reasonably well with machine learning, but new techniques and larger datasets may need to be developed to help deepen the accuracy of the classification.
  - No result related within-task mental rotation levels was discovered in the statistical or machine learning findings.
  - Neither Stroop nor Chess were successful in terms of machine learning classification, although more powerful methods or other models may be able to find improvements.
  - Despite mixed results for within-task classification, cross-task classification performed very well overall.

11.2 Extrapolated Core Findings

From the main findings discussed above, and extrapolating from other details throughout the studies above, I note a few key ideas worthy of consideration.

- Brain-network based classification and episodic memory are promising vectors for next-generation fNIRS-based BCI which measures the PFC.
- Low-Cost EEG systems measuring the PFC can distinguish levels of workload in the N-Back task, and have great promise in terms of cross-task classification between a variety of complex tasks engaging different neural states.
- Classification within the `OPTIMIZED` criteria across multiple studies show that different participants respond better to different machine learning models; likewise, the Part III study indicates that activation patterns within brain regions may be sensitive to different machine learning models and temporal windows for classification.
- Machine learning results do not necessarily align with the best-effort statistical analyses of neural signals.

11.3 Abstracted Findings for Future Work

We can further develop a few concrete abstractions from the ideas above towards the broader contextualization of these results for the development of future BCI systems.

- The strength of fNIRS in terms of spatial localization, even within the limitation of PFC-based interfaces, can be exploited towards new designs. That is, this dissertation has established the possibility of application of ideas from modern neuroscience based on the PFC, but actual interfaces using them towards real-time interfaces is yet to be proven. At the very least, a similar mPFC-lPFC (or DMN/TPN) based interface could be developed with the insights learned from the study presented here. Episodic memory likewise seems to be a viable candidate for classification.

- The weakness of EEG in terms of spatial localization is in some sense a strength for BCI when considering that measured spectral patterns can reflect broad-based neural activity, even with a low-sensor count consumer-grade device. The core finding of cross-task state-based differentiation expands beyond modern neuroscientific evidence in terms of precise rationalization of differentiation of neural activation, and needs to be explored further, particularly within the low signal-to-noise ratio of EEG, which requires careful interpretation of the results from within a neuroscientific context.

- Combining these EEG and fNIRS together presents a sensible approach as the foundation for next-generation implicit interfaces.

- Based on differences across participants and brain regions, there is great opportunity to explore meta or ensemble classification approaches which use multiple models per-region, per-individual, or per-measurement modality. This idea is concrete, and it is very possible to apply it to currently available datasets.

- Significant patterns of neural activation related to task state may be measurable through statistical analysis, but operationalizing measurements from neural sensors is a fundamentally different problem.

The last point, in particular, evokes a nuanced (yet somewhat controversial) dynamic worth further consideration. Prior to diving into the direct discussion of this point, I'd like to first explore in more detail the discrepancies in the data regarding statistical and machine learning.

11.4 ML 'vs.' Statistics

A common theme encountered in the data analyses has been a mismatch between the statistical analyses and machine learning results. To that end, see Table 11.1, which shows a brief overview of results which are "mismatched" in this way. Among these results, there are patterns in two directions: insignificant statistics with notable ML results, and significant statistics with not-notable ML results[1] .

[1]It is worth noting that the cutoff for "notability" is being set here at 60%. This is discussed more in Section 13
[2]Post-hoc analyses not run for these data, because omnibus test yielded an insignificant result.

Table 11.1: Summary of results across studies for which machine learning and statistical analyses are in conflict. The results above the double line have insignificant statistics but usable ML results, and the results below the double line are the inverse.

Study	**N Participants**	**Meas.**	**Factor**	**p**	ϵ_p^2	**ML**
Gradient of Subjectivity	15	L DSI	CONDITION	0.113	0.03	65%
Gradient of Subjectivity	15	L DSI	CONDITION (within SAT)	nr^2	nr	69%
Ikea	8	[all probes]	TASK	0.876	0.00	85%
Complex Decision-Making	30	R DSI	NAI→AI (AI - NAI)	<0.001	0.44	55%
Complex Decision-Making	30	R DSϕ	NAI→AI (AI - NAI)	0.024	0.15	55%
EEG/Workload: Chess	18	β and γ bands	WORKLOAD	<=0.001	[0.07-0.16]	54%
EEG/Workload: N-Back	12	θ and α bands	WORKLOAD	[0.003-0.014]	[0.08-0.12]	54%
EEG/Workload: Stroop	11	$\beta 2$ and $\alpha 2$ bands	WORKLOAD	[0.028;0.050]	[0.17;0.12]	51%

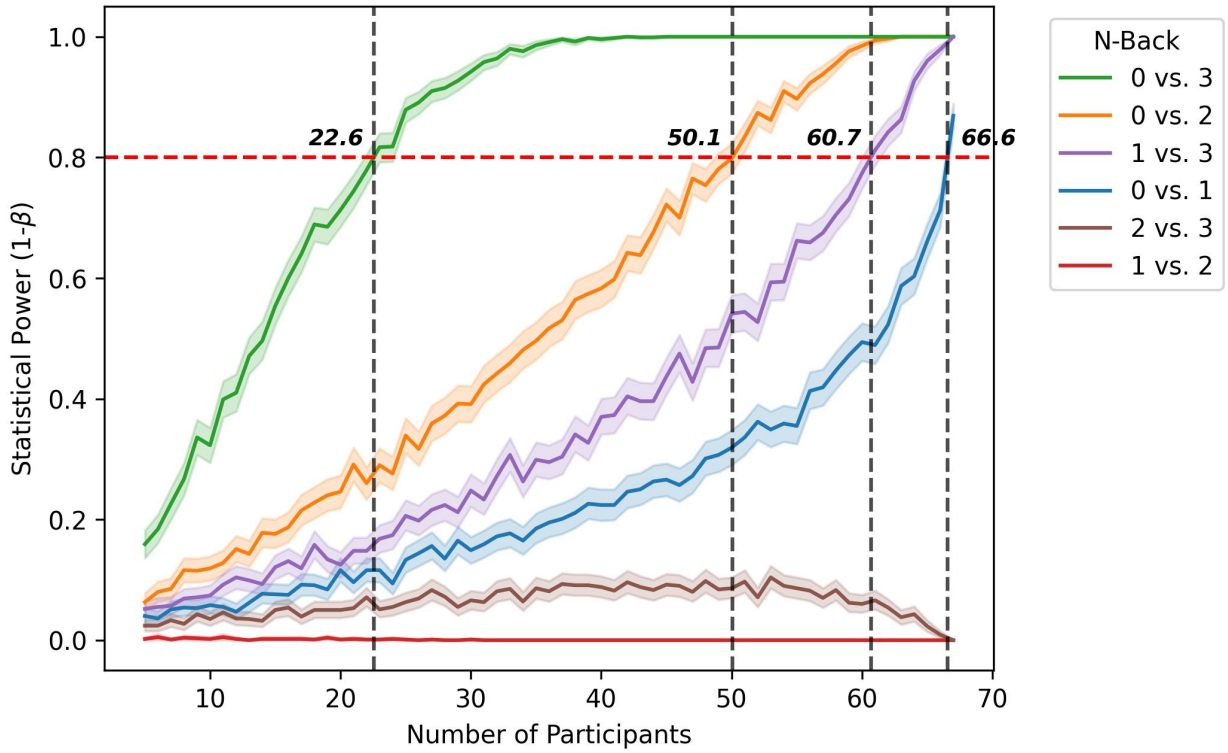
Insignificant Statistics and Notable ML

Two studies demonstrated that the statistical analyses of the data did not show the whole picture: in the gradient of subjectivity results, the effect of the `AI` vs. `NAI` condition on the Left PFC clearly was discernible by machine learning, yet the statistical analyses was not significant; similarly, in the Ikea study, distinguishing between conditions was certainly possible with machine learning, yet no significant values were found in the statistical analysis. Despite the insignificant result in the gradient of subjectivity task, this study with 15 participants did uncover a small effect size of 0.03 found; in the Ikea study, with only 8 participants, no significant effects were found. I hypothesize that the most likely consideration here in both cases for the 'lack' of results is the sample size of the datasets. To confirm the possibility of this being in fact the case, I revisited the Tufts dataset used to show the VLFO calculations shown briefly in the Introduction (2.2.4) and in more detail in Appendix A. Specifically, for each grouping of binary workload differentiators (e.g. 0 vs. 3 back, 1 vs. 2 back, etc), I ran a power analysis with a Monte Carlo simulation, where, for a given number of participants `k`, ranging from `k` from 5-68, I performed 1000 iterations wherein each iteration involved fitting a distinct model with the data from `k` randomly selected participants (without replacement)[3]; for each `k`, the proportion of significant results below $\alpha = 0.05$ were recorded. Results are in Figure 11.1.

The results of these analyses greatly help to clarify the mismatches discovered in the fNIRS data. Based on these data, more participants than perhaps thought otherwise are be necessary to uncover workload-based distinctions in VLFO band power (of the left PFC with the DSI measure).

[3]Procedure is the same as in Appendix A.

Figure 11.1: Power analysis of LLM results across levels of N-Back comparison. For each number of participants on the x axis, 1000 models were used with a random set of participants taken from the total dataset. The y-axis indicates the proportion of significant results at that level.

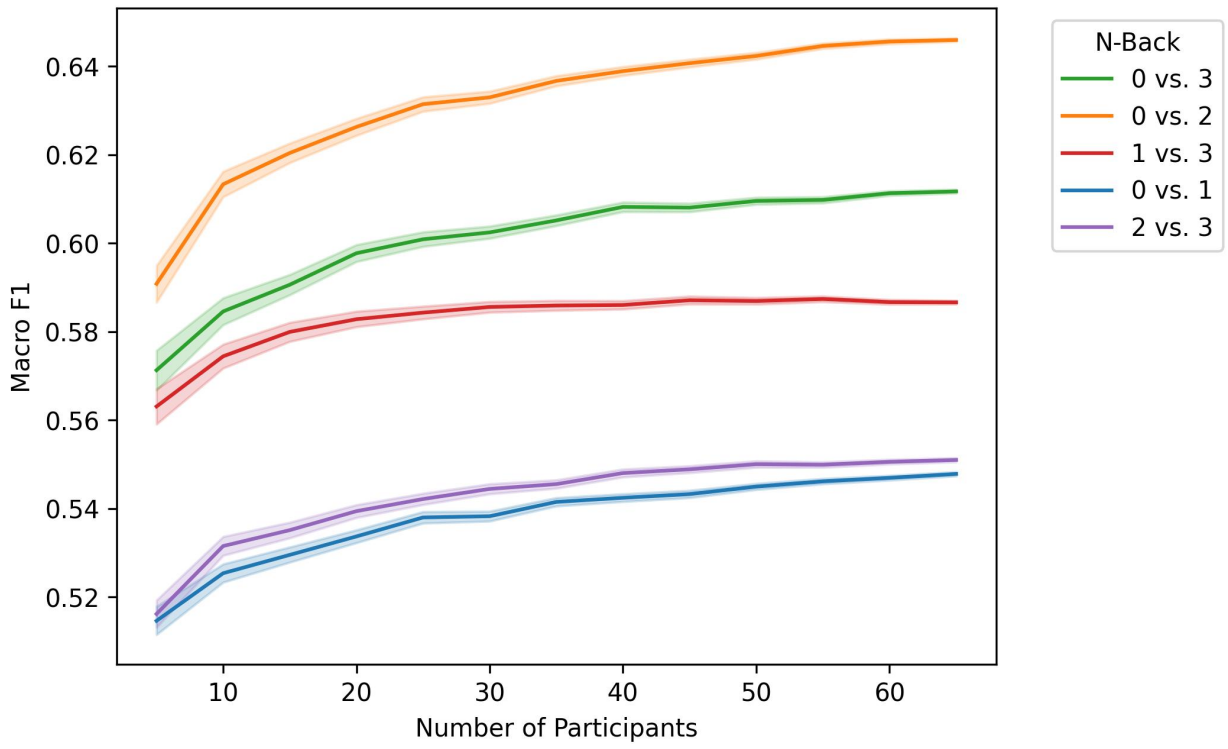


Of note, the number of participants required increases dramatically given "difficulty" of delineation. For example, 0 vs. 3 Back requires approximately 22-23 participants to reach 80% power, whereas 0 vs. 1 Back, a much harder delineation, requires 66-67 participants! The implication is rather sobering for those who are used to running such human subjects studies: to have 80% power to determine statistically significant effects of more subtle distinctions of workload within the Left PFC as related to the N-Back task, large numbers of participants are required than perhaps previously thought necessary. In consideration of this finding, however, it is very much also worth highlighting that the R PFC result between levels of `TASK` in the gradient of subjectivity study is worth further consideration; that is, the strength of the change in PFC activation due to episodic memory vs. more objective tasks is very worthy of note for future BCI work, as we detected a significant result with large effect size with a much smaller number of participants (15); however some caution is urged given that we do not have a similar dataset with larger numbers of participants and analyses for that metric of interest.

Significant Statistics and Not-Notable ML

In other studies, we found the inverse issue: for Complex Decision-Making, and in workload considerations for all of Chess, N-Back, and Stroop, significant results were clearly presented in the statistical findings, however the machine learning results were all <=55%, not indicating usefulness for BCI. For some understanding in terms of the dynamics at play in the context of fNIRS, I ran a similar Monte-Carlo analysis as above, but with machine learning instead of statistical analyses. I selected Random Forest as a baseline model. Given that the ML models take longer to train than the statistical models take to fit, I only ran models for which `k` was divisible by 5. Again, I performed 1000 iterations for each `k`, wherein each iteration involved randomly selecting `k` participants (without replacement), performing LOO-CV across those participants, and tracking the average Macro F1 score. Results are shown in Figure 11.2.

Figure 11.2: Analysis of LOO-CV RF runs across levels of N-Back comparison. For each number of participants to be used in the training set (x axis), 1000 models were used with a random set of this many participants taken from the total dataset; the y-axis displays the grand average over the 1000 LOO-CV result averages for each number of participants.



The results show slight improvement as the number of participants increases, but it is also clear

that the upper bound of Macro F1 is not likely to improve much past 5 percentage points from the baseline of 5 participants. Also of note, in terms of the "null" result in the complex decision-making task, the machine learning performance is approximately similar to what is discovered between levels of 0 vs. 1 Back or 2 vs. 3 Back; that is, as discussed earlier, the differentiation of the cognitive state between the with-Copilot and without-Copilot conditions is simply not "extreme" enough to be identifiable with machine learning. That said, the patterns therein *were* substantial enough to be statistically significant. And, for the EEG results, although we do not have a similar benchmark by which to test them, it is likely a similar pattern to what we are experiencing with the fNIRS data in this instance. For future such interfaces, integrating these two paradigms together may hold significant promise towards improving classification accuracy, in addition to the other suggestions suggested earlier, including ensemble learning with multiple classifiers being used together across locations, participants, and measurement modalities.

Chapter 12

Human-Sensor-Computer Interaction (HSCI)

At this point I hope the reader will permit me a brief and somewhat informal digression to introduce a concept which abstracts some broader concepts from the core lessons learned throughout the creation of this dissertation as a whole.

12.1 Analytical Context in BCI

Regarding the method of determining distinctions in the neural signal within the context of sometimes conflicting methodological results (e.g. machine learning vs. statistical analyses discussed above), it is worth stepping back for a moment and considering the intentions underlying the modeling approaches themselves.

12.1.1 The Forward Problem

When considering the standard statistical approach to "understanding" the patterns in a neural signal, we engage with what is known as the *forward problem*: we design (hopefully) well-formulated experiments where factor manipulations are modeled under distributional[1] and other assumptions, and where sources of variance (including error) are explicitly accounted for. This process can be extremely useful: understanding β or γ increases with increasing quality of chess moves or puzzle difficulty gives us reasonable information by which to conduct future work and/or to understand the nature of the brain in the context of the experimental manipulations. However, the scope of these

[1]Yes, this ignores the Bayesian approach. More discussed in Limitations.

findings in BCI can also be quite limited. Consider, for instance, the finding that a self-reflection task decreases Δ`HbD Intensity` in the VLFO band of the right lateral prefrontal cortex implying decreased cerebral autoregulation likely due to increases in localized hemodynamic activity due to neuronal activation associated with episodic memory. If interfaces using machine learning (or other means) cannot operationalize these results for real-time application of BCI, they are not usable. So for the neuroscientist, these may be useful findings, but for the HCI researcher, they are somewhat worthless!

12.1.2 The Inverse Problem

By contrast, in terms of the application of the neural data into interfaces, we are confronted with a related, yet distinct challenge: our aim is to interpret the signal that we measure and apply that interpretation in an interface; in mathematics, we have what is known as an inverse problem: we are tasked with calculating the causal factors that produce differences in the signal. The essence of a causal problem has most famously been expressed as: "Can one hear the shape of a drum?" [274]. And in our case: **given data from the brain, can one "hear" the shape of the state of the human?** Naturally, machine learning is a suitable fit in this scenario. Most often, we therefore transform a set of time series signal data into a single feature vector using a variety of statistical features, and watch the optimization methods work their magic. In theory, this works great. But what happens when the results aren't as expected? Why don't the machine learning methods always "just work"? When staring down the barrel of 50% accuracy results, what is one to do?

12.1.3 The Case for Statistics in BCI

Any researcher who has observed signals derived from a neural sensor will be familiar with the feeling of "what am I really looking at?". That is, there are distinct questions related to the nature of the signal. Identification of components of such a signal by eye outside of extreme movement is close-to impossible. Questions at this point can abound, particularly when the magic of optimization fails to perform well. Is the data too noisy? Are there problems with the measurement device? Errors in the data preprocessing or machine learning pipeline? And, irrespective of the measurement tool itself, is there anything actually different actually happening in the brain? Such questions in the face of consistent 50% results can drive one to the point of madness. So what is the remedy? In

this case, it seems to me that the statistical analysis which is otherwise "useless" for BCI becomes *extremely* valuable: by delineating patterns in the data under carefully thought-through analyses, under best-efforts to remove extracerebral "noise", there is or is not indeed something occurring "in the brain". However, the requirements to make such determinations with any degree of confidence, as we have seen, necessitate relatively large numbers of participants (at least in the context of workload differentiation), so this does not come cheaply.

12.1.4 Embracing Brain + Extracerebral Information

And yet, there is another nuance to consider here. **When the "magic" of ML *does work*, with what confidence can we be sure that neural activity is the sole cause of the measured effects?** Let's consider the machine learning application of a relatively simple model like SVM. Such a model, which determines the hyperplane that maximizes the length of the support vectors to the nearest samples of the classes in question, does not "care" in one way or another whether the factors at play have any relation to human neural function. And, given the near-impossibility to completely remove noise from noninvasive sensors, it is effectively certain that such noise will find its way into the parameters of the model. *But what if this information helps classification accuracy?* Perhaps the movement of participants is slightly higher in higher workload conditions. Maybe participants are blinking more during N-Back tasks than in Chess puzzles. Perhaps brain entrainment to specific frequencies within the data presentation are eliciting patterns which are easily recognizable to ML models. In this sense, the things which to the neuroscientist present the highest degree of transgression present a wonderful vector of opportunity for anyone studying HCI. And yet, at what point does this become problematic if we remain forcibly constrained within the domain of "Brain-Computer Interfaces"?

12.2 HSCI

12.2.1 Introduction

To that end, I believe it useful to re-frame our way of thinking about the problem of implicit BCI in terms of this inverse problem of "hearing the shape of the human". The fundamental principle is straightforward: rather than consider the locus of measurement (e.g. the brain) as the core principle

by which we frame the development of the technology driving the future interaction of the human and machine, we should instead highlight the interrelationship of the human and the sensor. **That is, rather than attempting to hear the shape of the human from brain activity, can we hear the shape of the human from the signal?** I call this idea Human-Sensor-Computer Interaction (HSCI).

With this idea in mind, in terms of the sensors commonly used in BCI, rather than avoid the elephant in the room of extracerebral information "corrupting" the data, we can instead leverage the strengths of machine learning in tandem with the realities of reading data from sensors which measure "brain + extracerebral information" towards more complex inferences of information about the human towards the expansion of the communication bandwidth between human and machine. Notable primary examples of such information that relate directly to the rest of this dissertation are hemodynamic information from within the scalp incorporated into the fNIRS signal, and muscle movements like eye blinks "accidentally" measured by EEG.

Indeed, within the context of HSCI, information measured from the brain through best-effort application of techniques to reduce the influence of extracerebral information becomes a *limitation*, rather than a boon[2]. Of chief interest, therefore, is to take any and all sensors at face value: that is, one does not start with the principle that EEG or fNIRS measure "the brain"; but instead, we start with the proposition that EEG measures electrical activity, that fNIRS measures light, and that some combination of neural activity and extracerebral activity are present in the measured signal. By taking this lower-level approach to the conceptualization of the sensor, we then become empowered to capture all of the usable information it provides as it relates to the human state.

12.2.2 Expanding Our Perspective Beyond the Physical

But what is the catch? What is the trade-off with this design decision? In some sense, the application of such interfaces trained on non-disabled users may not be applicable to such users, as movement artifacts from healthy human users would indeed create noise unable to be of contextual use towards disabled human subjects. In other contexts, however, such as blinking or scalp noise, such information could indeed be usable by traditional "BCI" systems. Instead, at least within the

[2]This is *not* to say that we should not apply statistical analyses under rigorous conditions attempting to maximally remove noise towards understanding the nature of a given application's effects on neural function, but that we should do both

context of heathy human subjects, I believe that by taking a lower-level approach to the notion of a sensor which measures the human, *the trade-off is that we can take a more expansive view of what we define as a sensor*. That is to say, in the context of HSCI, anything which provides information related to a human's state is valid and valuable information, and is thus a sensor. This might be physical information, or nonphysical information. It could be test scores, reading habits, browsing history, etc. It might be interactive, or non-interactive. The question then for a given application becomes which sensors provide the optimal trade-off between complexity and richness of the human state information and the difficulty required to extract that information.

12.3 Background

It is worth stopping for a moment to consider the context of research within which this broader idea of HSCI fits. In the domain of expanding sensor information to enable the modeling of noise, this concept is effectively a reformulation of ideas discussed in Desney Tan's book on BCI [275], specifically in chapters by Girouard [276] and Plass-Oude Bos [277], however the conceptualization of the idea moves past simply considering the brain as the localization of primary measurement with extracerebral noise added onto the set of input signals. HSCI is also conceptually related to the idea of Physiological Computing championed by Fairclough [278, 279], but there are a few notable distinctions in that regard worth discussion.

12.3.1 Physiological Computing and HSCI

"Psychophysiological States" and "Human States"

In terms of Physiological computing, the first primary caveat detailed by the author in association with the biocybernetic loop[3] is that there is a distinct relation of the physiological measurements to "valid measures of psychological concepts" [279]. HSCI does not share this constraint, given that the potential for accessing states of consciousness beyond psychological contextualization is already possible. Although this idea may seem extreme, we can consider this it trivially clear in terms of the work of Semertzidis in "Brain-Computer Integration" [280], whereby the information

[3]This loop is the collection of physiological data, filtering the data/abstracting information from the data, and application of that abstracted information to the computing system. [279].

measured from EEG is directly displayed to the participant as means of communication of unspecified information. Indeed, in consideration of simply visualizing an input vector of current sensor systems back to the user, how can we conceptualize or comprehend the "meaning" of such information? We can certainly posit abstractions on top of this interaction, however the principle that the complexity of information within the nature of the interaction transcends ontological representation is straightforward. Similarly, unsupervised methods of categorizing extremely high-dimensional sensor data may very well lead to the development of interfaces which engage in direct interpretation extending past the confines of psychological frameworks; in my view, it is more likely that in the future the psychological frameworks will be derived from the post-hoc assessments of such unsupervised measurements, rather than the other way around.

Non-Physical Information

Physiological computing inherently limits itself to measurement from physiological sensors. By contrast, the definition of a "sensor" in HSCI is more abstract, instead encompassing any measurement that can reliably inform an interface about the state of the human. Indeed, throughout this dissertation, non-physical measurements have demonstrated remarkable reliability and utility in characterizing human states. However, this conceptualization of non-physical sensors extends beyond traditional self-reported metrics (`NASA-TLX`, `VA` metrics) and behavioral metrics (reaction time, accuracy) to also includes other sources of non-physical data (interaction patterns with digital media, typed language use, interface preferences, etc.). While there are trade-offs in using these metrics, particularly when they require active user participation - they represent valuable sources of human state information that should not be excluded from our interaction frameworks simply because they lack a direct physiological basis. And yet, each of these measurements, likewise, should be considered within the same context as a physical sensor in terms of their ability to provide a limited set of usable information for a computing system. As interaction designers, we should embrace this expanded measurement space, developing systems that intelligently combine various measurement types based on their contextual appropriateness, reliability, and specific requirements of the application domain.

12.4 Implications for Future Work

So what are the actual areas of study within HSCI? Coming from the context of (implicit) BCI, perhaps the most concrete question which immediately presents itself is how to identify adequate decision boundaries likely to be distinguishable by machine learning models within the broader space of high-dimensional, multi-modal sensor inputs relating to complex aspects of human state. In my estimation, however, there are a variety of considerations which step towards this broader vision that must be addressed first prior to being able to fully approach this idea.

- To what degree does traditional signal filtration affect the classification performance of machine learning models? Can we quantify and determine the sources of extracerebral effects (e.g. scalp hemodynamics, facial movements, etc.) during laboratory conditions which contribute to useful (or non-useful) signal classification of known human states? Comparative studies of classification outcomes using artifact-removed versus raw EEG data across various cognitive tasks would provide valuable insights. This research direction could fundamentally alter how we approach signal processing for human state classification.

- How can self-report measures and other non-physical data sources (e.g. reaction time, correctness, etc.) be used by HSCI systems to create more comprehensive interfaces? Can we stretch the nature of "implicit" interfaces to include instead quantified trade-offs between user interruption and usability of information gleaned from such interruptions?

- What are the limitations, nuances, and complexities of integrating self-report or other non-physical data with information from physiological sensors? How can the varying time-durations of such measures complement, reinforce, and limit each other? Although currently such self-report models may seem to provide more comprehensive information than the neural signals, at some point the trade-off will shift; when will that point occur? What do interfaces look like when we are validating self-report or other interface usability information structure based off of information from signal information?

- Naturally, the extension of the ideas presented here imply that the computing system has enough information from the user to know more about them than perhaps they do of themselves.

The interface designer of such systems will be faced with extreme and perhaps dangerous amounts of influence over their users. In such situations, we must design frameworks of interaction that protect users, and likewise design interfaces enabling users to maintain control over the interface, rather than the other way around. Applications related to the parameterization and interfaces to enable such control are necessary areas of study in their own right. Wizard-of-oz paradigms can be leveraged to explore such ideas prior to the development of fully-fledged HSCI.

12.5 Summary

This chapter introduces Human-Sensor-Computer Interaction. This idea represents a philosophical shift that recontextualizes our understanding of implicit systems (in particular, BCIs) with the understanding that to fully use the information provided by current state-of-the-art noninvasive sensors, we must re-characterize what we are trying to measure with them; rather than hold fast to the idea of "the brain", given the impossibility of removing extracerebral information from noninvasive "neural" sensors, it would be wise to consider the broader implication of leaning into the information we can extract towards understanding "the human". This idea is further extended from "neural" sensors to include both physical and non-physical sensors, which can be combined to create the next generation of interfaces stepping towards more complex interaction between the human and the machine. Preliminary ideas are discussed for future work within the domain of HSCI.

Chapter 13

Limitations

This work is limited in a variety of ways.

13.1 PFC

Firstly, the neural aspect of this work is limited in terms of the localization of the signal. Of course, the main focus on this work has been in the domain of the PFC, but it bears consideration that expanding the ideas presented herein beyond the PFC requires measurements from other areas of the brain. Indeed, one could argue that the idea of inferring complex high-dimensional state information from the human necessarily *requires* measurements across multiple regions, the poor localization of EEG notwithstanding.

13.2 Machine Learning

13.2.1 Lack of Deep Learning/Other Methods

Although I have used for most projects a few canonical algorithms, applications leveraging deep neural nets, with all the benefits of LSTMs, RNNs, or LLMs towards signal classification, and opportunities for unsupervised, semi-supervised, reinforcement, or other kinds of learning may provide opportunities for increased classification accuracy. Plenty of opportunity exists for improvement and refinement in this direction.

13.2.2 Baseline for "Usable" Results

Throughout this work, the general baseline for what is considered as an "interesting" or "usable" result is approximately 60% Macro F1 score. The degree to which this value approximates practical applications in the here-and-now, however, is somewhat limited. Instead, it is hoped that the reader can infer that, with some additional work implemented in larger datasets or with more complex forms of ML classification employed, it is likely that the patterns in the data recognized in this way could develop into usable results.

13.3 Statistical Modeling

13.3.1 Sample Sizes

As discussed in Section 11.4, the dataset sample sizes in the works here are relatively small. One source of this was the number of participant exclusions required due to technical constraint issues (e.g. bluetooth failures, etc.), particularly with the MUSE device. Relying on cables rather than bluetooth data collection is highly encouraged. Use of the participant numbers presented for "baseline" sample sizes presented in the power analysis shown in 11.1 may provide more useful sizes for researchers interested in statistical analyses.

13.3.2 LMM Constraint

Although the use of LMM modeling represents a certain step-up from the ANOVA, a variety of other methods are emerging as highly useful which were not employed here, ranging from methods like Generalized Estimating Equations (GEE) to modern statistical methods like Bayesian analytical approaches. The former case would enable use to explicitly perform time-series analysis, and the latter might enable us to encode priors related to workload measures effects on the PFC in worthwhile ways. Further analysis of all of the studies herein with these (or other) modeling approaches might provide useful insights into the nature of the signal under the tasks and conditions studied.

13.4 Device Constraints

Although I use both fNIRS and EEG, there is only one device considered in each category. Further, the EEG system used is a low-grade consumer device; although I present interesting findings related to this device, the company which produces it has already created a successor device - EEG work based on harder-to-use devices in the near-term can better inform our understandings of the capacity of to develop more complex interfaces in the long-term.

13.5 HSCI

Although the basic principles of HSCI are clearly defined in this work, this concept is brand-new. And, although consideration of the usability of the full spectrum of a given signal, in tandem with the ability to expand beyond neural signals into any signal measuring the human, is worthwhile, much more work is necessary to fully take the kernel of this philosophical approach and more concretely map it into the broader research space.

Chapter 14

Conclusion

It bears repeating that, despite current limitations, human-machine interfacing will inevitably, invariably, and without question play a considerable role in the evolution of human consciousness; this dissertation takes a small step towards the realization of this longer-term vision. By considering the untapped strengths of both EEG and fNIRS towards more-subtle human state classification in the PFC, this work has first set to establish a targeted set of possibilities which can support the next frontier of research. And, by a careful consideration of the *actual* information measured by the sensors commonly used in BCI, including what is commonly considered as noise, a conceptualization of HCI as mediated by physical and nonphysical sensors is introduced, through which it is hoped we can move more powerfully towards the development of systems which engage with increasingly complex aspects of the human experience.

PART VI

Appendices

Appendix A

VLFO Analysis of the Tufts Mental Workload Dataset

The Tufts Mental Workload Dataset is comprised of 68 participants' data collected during the N-Back task. For this study, done by Wang. et al. [52], participants wore the same probe setup as described in 3.1.1. In the N-Back task, one is required to store and recall N constantly changing pieces of information in real-time. Specifically, as the value of 'N' increases, the level of mental workload likewise increases. This relationship vis-à-vis neural activation of the left PFC during increasing levels of the N-Back task has been confirmed in fMRI studies [281]. In this dataset, each participant performed 16 3-minute trials of N-Back, grouped in 4 blocks each of 0-3-Back, done in a Latin square design. To analyze the data, for each block of N-Back, for each participant in this dataset, we applied the following procedure

1. Calculate Δ[HbD] = Δ[HbO] - Δ[Hb] for both FD Phase (ϕ) and Intensity (I) values from each of the left (L) and right (R) probes.

2. Apply a frequency domain transformation of the Δ[HbD] data, and extract the VLF band [0.02-0.07Hz] data.

3. Calculate the log of the Simpson integral over the VLF band.

We then used LMM modeling, specifying nested random intercepts with task block nested within participant ID.

A.1 Results

Results are shown in Table A.1 below. All factors showed significance. Post-hoc contrasts are shown in Table A.2.

Table A.1: Results over modeling fNIRS data for the N-Back task over 68 participants with log Δ[HbD] VLFO band power as the dependant variable. α is set to 0.025, grouping with two measures within each of the L and R probes.

Probe	**Meas**	**df1**	**df2**	**F**	**p**	**p.sig**	ϵ_p^2	ϵ_p^2 **CI**
L	DSI	3	813.0	11.31	<0.001	***	0.04	[0.02,0.06]
L	DSϕ	3	813.0	6.31	<0.001	***	0.02	[0.00,0.03]
R	DSI	3	813.0	3.96	0.01	**	0.01	[0.00,0.02]
R	DSϕ	3	813.0	4.00	0.01	**	0.01	[0.00,0.02]

The most notable result from the post-hoc contrasts is a clear pattern of differentiations within the left PFC, primarily within the measurement of `DSI`, which had four of the seven total significant contrasts: (0-3): $t_{813.00} = 5.75, p < 0.001, \epsilon_p^2 = 0.04$; (0-2) $t_{813.00} = 3.58, p = 0.002, \epsilon_p^2 = 0.01$; (1-3) $t_{813.00} = 3.03, p = 0.013, \epsilon_p^2 = 0.01$; (0-1) $t_{813.00} = 2.72, p = 0.034, \epsilon_p^2 = 0.01$. Aside from these, significant contrasts were also found for the workload contrasts of 0-3 in: the DSϕ measure in the left probe: $t_{813.00} = 4.34; p < 0.001; \epsilon_p^2 = 0.02$, for the DSI measure in the right probe: $t_{813.00} = 3.25; p < 0.007; \epsilon_p^2 = 0.01$, and the DSϕ measure in the right probe: $t_{813.00} = 3.20; p < 0.008; \epsilon_p^2 = 0.01$. These results suggest that the VLFO data reflects notable distinctions in the left prefrontal cortex, but not the right, except for the highest levels of workload distinction in the N-Back task.

Table A.2: Post-hoc contrasts for each of the significant models above, sorted in descending order by effect size (ties broken with p-values). P-values have been adjusted within each contrast using Tukey's HSD.

Probe	**Meas**	**Contrast**	**Est.**	**SE**	**df**	**t**	**p**	**p.sig**	ϵ_p^2	ϵ_p^2 **CI**
L	DSI	0 - 3	0.32	0.06	813.00	5.75	<0.001	***	0.04	[0.02,0.07]
L	DSϕ	0 - 3	0.18	0.04	813.00	4.34	<0.001	***	0.02	[0.01,0.05]
L	DSI	0 - 2	0.20	0.06	813.00	3.58	0.002	**	0.01	[0.00,0.03]
R	DSI	0 - 3	0.18	0.06	813.00	3.25	0.007	**	0.01	[0.00,0.03]
R	DSϕ	0 - 3	0.15	0.05	813.00	3.20	0.008	**	0.01	[0.00,0.03]
L	DSI	1 - 3	0.17	0.06	813.00	3.03	0.013	*	0.01	[0.00,0.03]
L	DSI	0 - 1	0.15	0.06	813.00	2.72	0.034	*	0.01	[0.00,0.02]
R	DSI	0 - 1	0.14	0.06	813.00	2.55	0.053	ns	0.01	[0.00,0.02]
R	DSϕ	0 - 2	0.11	0.05	813.00	2.40	0.078	ns	0.01	[0.00,0.02]
L	DSϕ	0 - 2	0.10	0.04	813.00	2.34	0.090	ns	0.01	[0.00,0.02]
R	DSI	0 - 2	0.13	0.06	813.00	2.29	0.101	ns	0.01	[0.00,0.02]
L	DSϕ	1 - 3	0.09	0.04	813.00	2.28	0.103	ns	0.01	[0.00,0.02]
L	DSI	2 - 3	0.12	0.06	813.00	2.17	0.133	ns	0.00	[0.00,0.02]
R	DSϕ	1 - 3	0.10	0.05	813.00	2.12	0.146	ns	0.00	[0.00,0.02]
L	DSϕ	0 - 1	0.08	0.04	813.00	2.06	0.167	ns	0.00	[0.00,0.02]
L	DSϕ	2 - 3	0.08	0.04	813.00	2.00	0.189	ns	0.00	[0.00,0.02]
R	DSϕ	1 - 2	0.06	0.05	813.00	1.33	0.546	ns	0.00	[0.00,0.01]
R	DSϕ	0 - 1	0.05	0.05	813.00	1.07	0.708	ns	0.00	[0.00,0.01]
L	DSϕ	1 - 2	0.01	0.04	813.00	0.28	0.992	ns	0.00	[0.00,0.00]
R	DSI	1 - 2	-0.01	0.06	813.00	-0.26	0.994	ns	0.00	[0.00,0.00]
R	DSI	1 - 3	0.04	0.06	813.00	0.70	0.898	ns	0.00	[0.00,0.00]
R	DSI	2 - 3	0.05	0.06	813.00	0.95	0.776	ns	0.00	[0.00,0.00]
L	DSI	1 - 2	0.05	0.06	813.00	0.86	0.824	ns	0.00	[0.00,0.00]
R	DSϕ	2 - 3	0.04	0.05	813.00	0.80	0.855	ns	0.00	[0.00,0.00]

Appendix B

Full Text of Gradient of Subjectivity Tasks

B.1 Planning Tasks

B.1.1 Planning Task A: Future Leaders Retreat

Construct a short ($\frac{1}{2}$ - 1 page) plan for a "Future Leaders Retreat" intended for emerging student leaders from REDACTED University. This retreat will focus on personal leadership development, resilience training, and introspection. Ensure that your plan includes:

1. A reflective name for the retreat that resonates with personal growth.
2. Agenda highlights such as mindfulness sessions, personal leadership journey sharing, and resilience building workshops.
3. A specific serene location (on or off campus) conducive to introspection and inner growth.
4. Considerations required for the holistic development and well-being of the attendees.
5. Plan for candidate selection for the retreat.

B.1.2 Planning Task B: Alumni Leadership Summit: REDACTED University Elite Networking Event

Draft a short ($\frac{1}{2}$ - 1 page) plan for an exclusive business networking event targeting REDACTED University alumni in leadership positions. Your plan should specify:

1. A dynamic event name that signifies industry leadership and networking.
2. Keynote speakers of interest, industry panel discussions, and insights into business trends.
3. A location near or on REDACTED University that embodies a business-centric environment.
4. Strategies to promote inter-industry networking and engagement between alumni and ambitious students.
5. Note that you may pick an area of expertise for the summit which relates to your field of study (or possible majors for you if undecided).

B.2 Poetry Tasks

B.2.1 Poetry Task A: Nature

Write a brief (10–15 line) poem on the beauty of nature.

B.2.2 Poetry Task B: Joy

Imagine a moment of unexpected joy on an ordinary day. Write a short (10-15 line) poem capturing the essence of that emotion.

B.3 Reflection Tasks

B.3.1 Reflection Task A: Movie

Pick your favorite movie released before 2020. Then draft a 2-paragraph reflection on how the movie resonates with your personal experiences or memories. Use as much detail as possible (quotes, scenes, etc).

B.3.2 Reflection Task B: Album

Pick your favorite album released before 2020. Then draft a 2-paragraph reflection on how the album resonates with your personal experiences or memories. Use as much detail as possible (song lyrics, album themes, etc).

B.4 SAT Tasks

The SAT tasks were slightly modified version of the 2016 SAT practice tests: numbers 5 [282] and 7 [283].

Appendix C

Gradient of Subjectivity: Potential Confound Analysis

C.1 Subtask Difficulty

For a given TASK, although we randomized whether SUBTASK A or B would be done with the Copilot assistant, it is neverthless important to determine whether or not the SUBTASKs for each TASK were of equal difficulty. To do this, we analyzed the data of only the NAI CONDITION in a between-subjects manner (as each subject only did each subtask once). Specifically, we performed independent-samples t-tests for each pair of subtasks. Results are listed in Table C.1 and Figure C.1. No significant results were found, indicating that the SUBTASKs within each TASK were of similar difficulty.

Table C.1: T-Test results for SUBTASK Difficulty Comparison

TASK	Df	t-value	p.adj	sig.
POEM	19	-0.693	0.497	ns
REF	19	-0.615	0.546	ns
SAT	19	0.004	0.997	ns
PLAN	19	-0.690	0.506	ns

C.2 Task Time

We also analyzed the potential confound of task time as it relates to mental workload. Specifically we were concerned that the task number would effect the change in workload scores between the AI

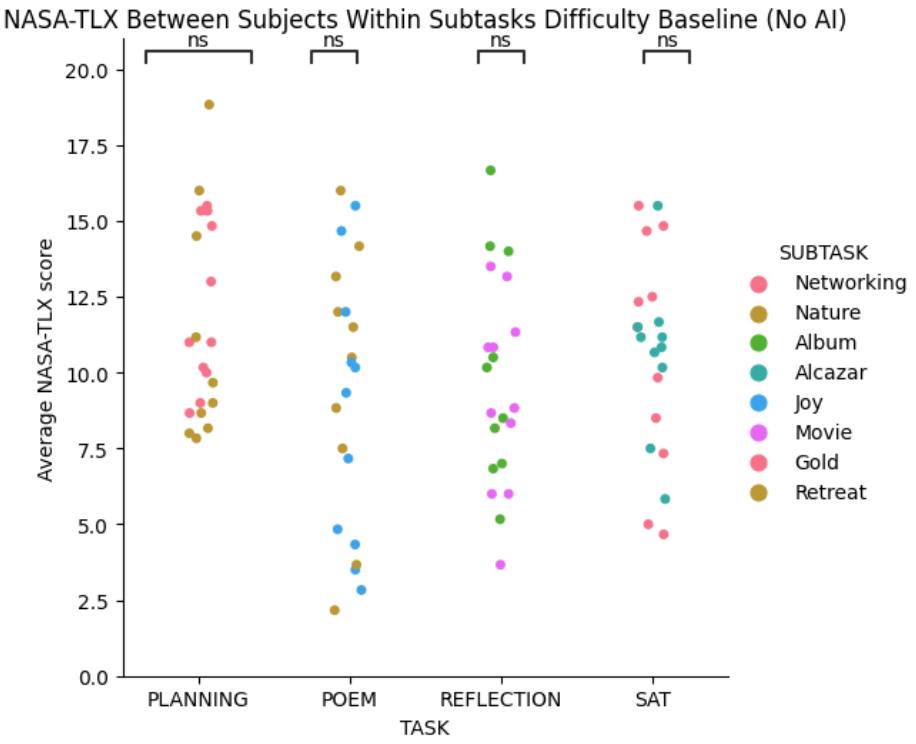


Figure C.1: `NASA-TLX` Mental Workload Score within each `SUBTASK`. Within each `TASK`, none of the `SUBTASK`s were significantly more difficult than the other.

and `NAI` levels of `CONDIITON`. To test this, we created a lmer model with the formula

$$\Delta SCORE \sim TASK\_NUM + (1|pid) \tag{C.1}$$

Where `TASK_NUM` was a number from 1-4, and Δ`SCORE` is the *change* in score defined as `NAI - AI`. The ANOVA for this model did not report a significant result ($F_{3,60}$=1.87, p=0.144, η_p^2=0.09, 95% CI=[0.00, 1.0]), although there was a moderate effect size. Contrast results are shown in Table C.2 and Figure C.2. None of the contrasts demonstrated significance.

Table C.2: Post-Hoc Contrast Results for `TASK_NUM`

Contrast	Estimate	SE	df	t.ratio	p.value	p.sig	η_p^2	95% CI
task_num1 - task_num3	3.76	1.66	60.00	2.26	0.119	ns	0.08	[0.0,1.0]
task_num2 - task_num3	2.81	1.66	60.00	1.69	0.339	ns	0.05	[0.0,1.0]
task_num0 - task_num3	2.62	1.66	60.00	1.57	0.401	ns	0.04	[0.0,1.0]
task_num0 - task_num1	-1.14	1.66	60.00	-0.69	0.902	ns	0.01	[0.0,1.0]
task_num1 - task_num2	0.95	1.66	60.00	0.57	0.940	ns	0.01	[0.0,1.0]
task_num0 - task_num2	-0.19	1.66	60.00	-0.11	0.999	ns	0.00	[0.0,1.0]

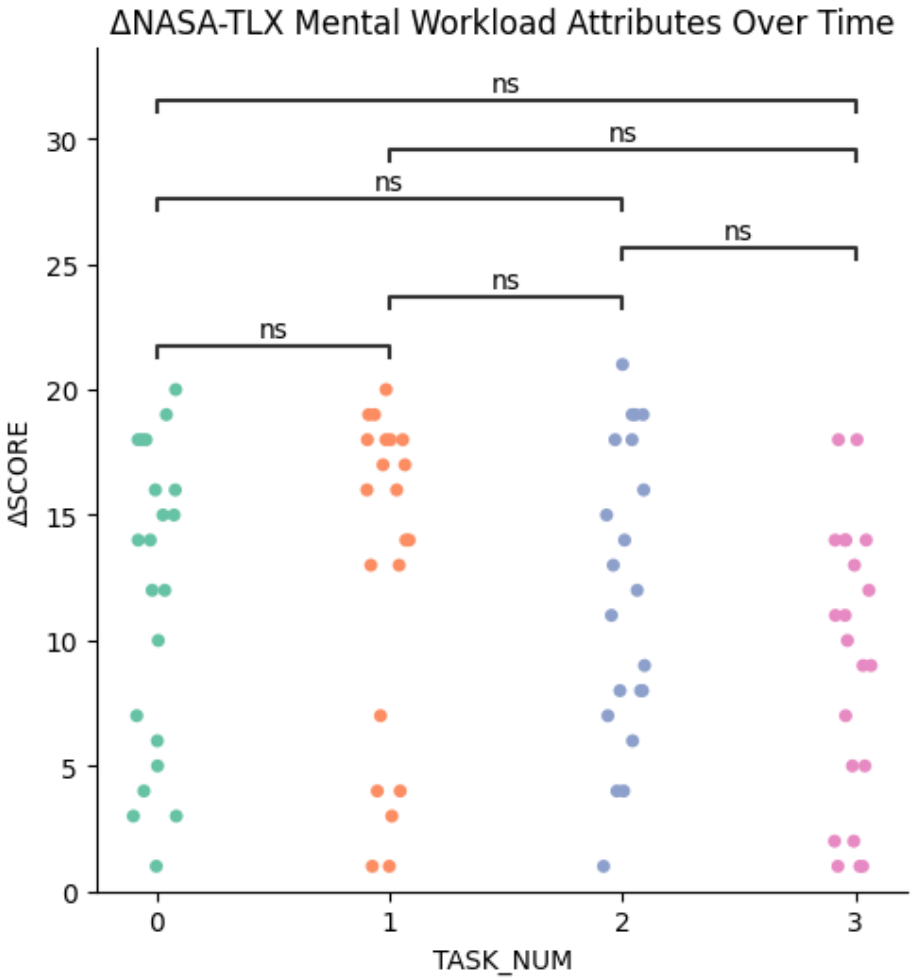


Figure C.2: Change in Workload Score (`NAI - AI`) as a Function of Task Number

Appendix D

Quiver Plot for AI First Cohort

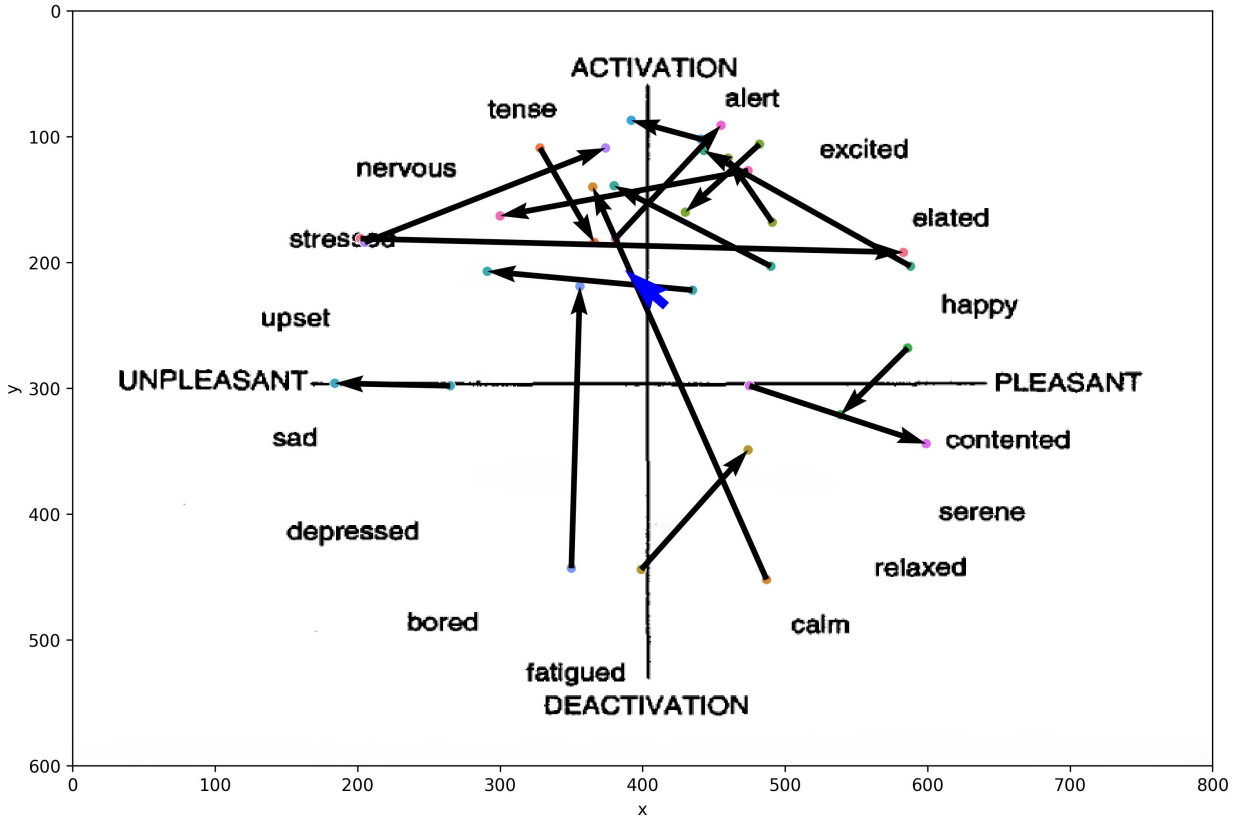


Figure D.1: After each task, participants selected the spot on the circumplex Valence-Arousal model which they felt best related to their state of mind during the task. This group of participants is those who experienced `AI` first; arrows begin in the `NAI` condition and point to the `AI` condition. Blue arrow represents the average of all participants' arrows after normalizing for direction, and then scaled to the mean distance across participants. Minimal differences are observed between conditions.

Appendix E

Full Text of Complex Decision-Making Tasks

Below follows the information provided to participants to complete the complex decision-making tasks used in 6. For each task, the task README is provided first, followed by the three documents each detailing a potential plan.

E.1 Task A: AtoZ Digital User-Interfaces for Automotive Systems: README

AtoZ is a car digital user interface company which is facing poor user experience reports for their software. Specifically, they are receiving negative user experience reports for their in-car digital interface, leading to decreased customer satisfaction and potentially impacting car sales.

1. Company Background: AtoZ Automotive Systems is a leading provider of digital interfaces for luxury and mid-range vehicles. Founded in 2010, they've been at the forefront of car infotainment systems but are now facing challenges with their latest software version.

2. Problem Info: The current in-car UI is reported to be unintuitive, slow to respond, and difficult to use while driving. User feedback indicates issues with complex menu structures, small touch targets, and difficulty in performing common tasks like adjusting climate controls or navigating to a destination.

3. Task Goal: You have been provided with multiple proposals for options to help solve the problem at hand. Specifically, you will rank-order the proposals for each of the criteria of

user impact, implementation costs, and risk. For each of the rank-orderings, you will also write a few sentences describing the rationale behind your rank ordering. Then, you will rank-order the proposals overall, and write a paragraph describing why you have chosen your number one option as such.

To complete the task, you will find in this document a page for each of the proposals. Insert your answers there. Remember that your work will be critical to the next steps for the success of AtoZ Automotive Systems!

User Impact Evaluation

Rank order:

1.
2.
3.

Reasoning

Financial Cost Evaluation

Rank order:

1.
2.
3.

Reasoning

Time Cost Evaluation

Rank order:

1.
2.
3.

Reasoning

Risk Evaluation

Rank order:

1.

2.

3.

Reasoning

Overall Evaluation

Rank order:

1.

2.

3.

Reasoning

E.2 Task A: AtoZ Digital User-Interfaces for Automotive Systems: Proposal 1 - Gamification

Problem Summary

Problem Description While AtoZ's current UIs are functional and reliable, user engagement levels are lower than desired. Users often find the interface to be purely utilitarian and lack incentives to explore its full range of features. This results in underutilization of the system's capabilities and lower overall satisfaction.

Quantitative Information

- User feedback indicates that 45% of users do not regularly engage with all available features.
- Customer satisfaction surveys show a 25% neutral or negative rating regarding UI engagement.
- There has been a 20% increase in feedback requesting more interactive and engaging elements in the UI.

Qualitative Information

- Users express that the UI lacks engaging elements that make daily interactions enjoyable.
- Frequent comments highlight a desire for more interactive and rewarding experiences.
- Competitor analysis shows that gamified interfaces are perceived as more engaging and user-friendly

Strategy Chosen

The strategy is to integrate gamification elements into the existing UI to boost user engagement, make interactions more enjoyable, and encourage users to explore all available features.

Implementation Plan

1. **Research and Planning**
    - Conduct research on successful gamification strategies and user preferences.
    - Develop a comprehensive plan outlining key gamification elements to be integrated.
2. **Feature Development**
    - Introduce a points and rewards system for using different UI features and achieving milestones.
    - Implement interactive challenges and missions that users can complete to earn rewards.
    - Develop a leaderboard and achievement system to foster a sense of competition and accomplishment.
3. **UI/UX Design**
    - Collaborate with UI/UX designers to seamlessly integrate gamification elements into the existing interface.
    - Ensure the design is intuitive and enhances the overall user experience without causing distraction.

4. **Testing and Iteration**

    - Conduct beta testing with a diverse group of users to gather feedback on the new gamified elements.
    - Make iterative improvements based on user feedback to ensure optimal engagement and satisfaction.

5. **Implementation and Launch**

    - Implement the gamification features across all platforms and vehicle models.
    - Launch a marketing campaign to introduce the new features and educate users on how to use them.

6. **Monitoring and Support**

    - Continuously monitor user engagement and satisfaction metrics.
    - Provide ongoing support and updates to keep the gamification elements fresh and engaging.

Costs

- **Time Cost**
    - Total time: 12 months
        * Research and Planning: 2 months
        * Feature Development: 4 months
        * UI/UX Design: 2 months
        * Testing and Iteration: 2 months
        * Implementation and Launch: 1 month
        * Monitoring and Support: 1 month
- **Financial Cost**
    - Total: $1.9 million

        * Research and Planning: $200,000
        * Feature Development: $800,000
        * UI/UX Design: $300,000
        * Testing and Iteration: $200,000
        * Implementation and Launch: $250,000
        * Monitoring and Support: $150,000

- **Labor Cost**
    - **Total: 23 personnel**
        * Research Team: 4 specialists
        * Development Team: 8 engineers
        * UI/UX Design Team: 4 designers
        * Testing Team: 4 engineers
        * Support Team: 3 specialists

Impact on Users

- **Increased Engagement:**
    - Expected to boost user engagement with the UI by 40%.
- **Enhanced User Satisfaction:**
    - Anticipated increase in customer satisfaction ratings by 30%.
- **Feature Utilization:**
    - Predicted 35% rise in the utilization of various UI features.

Risks and Mitigation

- **User Resistance**
    - Risk: Users may find gamification elements distracting or unnecessary.

- Likelihood: Moderate.
- Severity: Moderate.
- Mitigation: Ensure the gamification elements are optional and can be customized or turned off.

- **Over-Gamification**
  - Risk: The UI may become too focused on gamification, detracting from its primary functions.
  - Likelihood: Low.
  - Severity: High.
  - Mitigation: Maintain a balance between gamification and core functionalities through careful design and testing.

- **Technical Challenges**
  - Risk: Integration issues with existing systems.
  - Likelihood: Moderate.
  - Severity: High.
  - Mitigation: Conduct thorough testing and have a dedicated team for troubleshooting and support.

E.3 Task A: AtoZ Digital User-Interfaces for Automotive Systems: Proposal 2 - Voice Recognition

Problem Summary

Problem Description While AtoZ's current UI systems are highly functional, user feedback and industry trends indicate that the existing voice-recognition capabilities are suboptimal. Users report difficulties in getting accurate responses, especially in noisy environments or with diverse accents. This issue affects the overall user experience, leading to frustration and reduced use of voice commands.

Quantitative Information

- User feedback indicates a 30% error rate in voice command recognition.
- Customer satisfaction surveys show that 45% of users are dissatisfied with the voice-recognition feature.
- There has been a 20% decline in the use of voice commands over the past year.

Qualitative Information

- Users often complain about the system not understanding commands in noisy environments.
- Accents and dialects pose significant challenges, leading to repeated attempts to execute simple commands.
- Comparisons with competitors show that AtoZ's voice-recognition system lags behind in terms of accuracy and ease of use.

Strategy Chosen The strategy is to implement an advanced voice-recognition system leveraging AI and machine learning to improve accuracy, adaptability, and overall user satisfaction.

Implementation Plan

1. **Research and Development (R&D)**
    - Invest in R&D to explore state-of-the-art voice-recognition technologies.
    - Partner with AI experts and leading voice-recognition companies to incorporate the latest advancements.
2. **Data Collection and Training**
    - Collect a diverse dataset of voice samples, including various accents, dialects, and noisy environments.
    - Use machine learning algorithms to train the system for improved accuracy and adaptability.

3. **Integration and Testing**

    - Integrate the new system into existing UIs.
    - Conduct extensive testing in different scenarios to ensure robustness and reliability.

4. **User Training and Feedback**

    - Provide users with training materials to familiarize them with the new system.
    - Gather user feedback post-implementation to identify and address any issues.

Costs

- **Time Cost**
    - The project is expected to take 12 months:
        * R&D: 3 months
        * Data Collection and Training: 4 months
        * Integration and Testing: 3 months
        * User Training and Feedback: 2 months
- **Monetary Cost**
    - Total cost: $2 million
        * R&D: $1 million
        * Data Collection and Training: $500,000
        * Integration and Testing: $300,000
        * User Training and Feedback: $200,000
- **Labor Cost**
    - Total: 28 personnel
        * R&D Team: 10 engineers
        * Data Collection Team: 5 specialists
        * Integration and Testing Team: 8 engineers
        * User Training and Support Team: 5 specialists

Impact on Users

- **Improved Accuracy**
  - Expected to reduce the error rate in voice command recognition from 30% to below 5%.
- **Enhanced User Satisfaction**
  - Anticipated increase in customer satisfaction ratings by 40%.
- **Increased Usage**
  - Predicted 25% rise in the use of voice commands.

Risks and Mitigation

- **Technological Challenges**
  - Risk: Integration issues with existing systems.
  - Likelihood: Moderate.
  - Severity: High.
  - Mitigation: Conduct thorough testing and have a dedicated team for troubleshooting.
- **User Adaptation**
  - Risk: Users may struggle to adapt to the new system.
  - Likelihood: Low.
  - Severity: Moderate.
  - Mitigation: Provide comprehensive training and continuous support.
- **Data Privacy**
  - Risk: Concerns over the collection and use of voice data.
  - Likelihood: High.
  - Severity: High.

– Mitigation: Implement strict data privacy protocols and transparent communication with users.

E.4 Task A: AtoZ Digital User-Interfaces for Automotive Systems: Proposal 3 - Simplify User Interface

E.4.1 Problem Summary

Problem Description Although AtoZ's current UIs offer a wide range of features and functionalities, user feedback indicates that the interfaces are often too complex and difficult to navigate. This complexity detracts from the overall user experience, leading to confusion and decreased satisfaction.

Quantitative Information

- User feedback indicates that 40% of users find the current UI difficult to navigate.
- Customer satisfaction surveys show a 35% dissatisfaction rate with the complexity of the UI.
- There has been a 15% increase in customer support requests related to UI navigation over the past year.

Qualitative Information

- Users report that menus are cluttered and not intuitive.
- Common tasks require multiple steps, leading to frustration.
- Competitor analysis reveals that simpler UIs are perceived as more user-friendly and appealing.

Strategy Chosen

The strategy is to redesign and simplify the existing user interface, focusing on intuitive navigation, reduced clutter, and streamlined access to commonly used features.

Implementation Plan

1. **User Research and Analysis**

    - Conduct user surveys and focus groups to gather detailed feedback on UI pain points.
    - Analyze customer support data to identify common issues and areas for improvement.

2. **UI/UX Design**

    - Collaborate with experienced UI/UX designers to create simplified mockups.
    - Focus on minimalistic design principles, reducing the number of steps required for common tasks.

3. **Prototyping and Testing**

    - Develop interactive prototypes of the new UI design.
    - Conduct usability testing with a diverse group of users to gather feedback and make iterative improvements.

4. **Implementation and Rollout**

    - Implement the new UI design across all platforms and vehicle models.
    - Provide users with updated documentation and tutorials to ease the transition.

5. **Post-Implementation Review**

    - Monitor user feedback and usage data to assess the effectiveness of the new design.
    - Make further refinements based on ongoing user feedback.

Costs

- **Time Cost**
    - The project is expected to take 10 months:
        * User Research and Analysis: 2 months

        * UI/UX Design: 3 months
        * Prototyping and Testing: 2 months
        * Implementation and Rollout: 2 months
        * Post-Implementation Review: 1 month

- **Financial Cost**
    - Total: $1.5 million
        * User Research and Analysis: $300,000
        * UI/UX Design: $500,000
        * Prototyping and Testing: $200,000
        * Implementation and Rollout: $400,000
        * Post-Implementation Review: $100,000

- **Labor Cost**
    - **Total: 24 personnel**
        * User Research Team: 5 specialists
        * UI/UX Design Team: 6 designers
        * Prototyping and Testing Team: 4 engineers
        * Implementation Team: 6 engineers
        * Support and Review Team: 3 specialists

Impact on Users

- **Improved Usability**
    - Expected to reduce the complexity of UI navigation by 50%.

- **Enhanced User Satisfaction**
    - Anticipated increase in customer satisfaction ratings by 35%.

- **Decreased Support Requests**
  - Predicted 25% reduction in customer support requests related to UI navigation.

Risks and Mitigation

- **Resistance to Change**
  - Risk: Users may resist changes to the familiar UI.
  - Likelihood: Moderate.
  - Severity: Moderate.
  - Mitigation: Provide comprehensive training and clear communication about the benefits of the new design.
- **Design Oversimplification**
  - Risk: Oversimplifying the UI might remove essential functionalities.
  - Likelihood: Low.
  - Severity: High.
  - Mitigation: Conduct thorough testing and gather extensive user feedback during the design phase.
- **Implementation Delays**
  - Risk: Delays in implementing the new UI across all platforms.
  - Likelihood: Moderate.
  - Severity: Moderate.
  - Mitigation: Ensure strong project management and adherence to timelines.

E.5 Task B: EcoTech Solutions: Sustainable Energy Management System Implementation: README

EcoTech Solutions is a rapidly growing green technology company facing challenges with its energy management software for commercial buildings. They are receiving negative feedback from clients about the system's effectiveness in reducing energy consumption and its user interface, potentially impacting future sales and client retention.

1. Company Background: EcoTech Solutions, founded in 2015, is a leader in providing sustainable energy management systems for commercial buildings. Their software has been instrumental in helping businesses reduce their carbon footprint, but recent client feedback suggests they're falling behind competitors.

2. Problem Info: The current energy management system is reported to be inefficient in optimizing energy usage, has a complex user interface, and lacks integration with newer IoT devices. Clients are struggling to achieve their sustainability goals and find it difficult to interpret the data provided by the system.

3. Task Goal: You have been provided with multiple proposals to address the problem at hand. Your task is to rank-order the proposals for each of the criteria of **user impact, implementation costs, and risk**. For each rank-ordering, write a few sentences describing the rationale behind your decisions. Then, rank-order the proposals overall, and write a paragraph explaining why you've chosen your top option.

To complete the task, you will find in this document a section for each of the proposals. Insert your answers in the designated areas. Remember that your work will be crucial for the future success of EcoTech Solutions!

User Impact Evaluation

Rank order:

1.
2.

3.

Reasoning:

Financial Cost Evaluation

Rank order:

1.

2.

3.

Reasoning

Time Cost Evaluation

Rank order:

1.

2.

3.

Reasoning

Risk Evaluation

Rank order:

1.

2.

3.

Reasoning

Overall Evaluation

Rank order:

1.

2.

3.

Reasoning

E.6 Task B: EcoTech Solutions: Sustainable Energy Management System Implementation: Proposal 1 - AI-Powered Energy Optimization

Problem Summary

Problem Description While EcoTech's current energy management system is functional, it fails to optimize energy usage effectively across various building systems. Users report difficulties in achieving significant energy savings, especially during peak usage times. The system lacks the ability to predict and proactively adjust to changing energy demands.

Quantitative Information

- Client feedback indicates only a 15% average reduction in energy consumption, far below the promised 30%.
- 55% of users report dissatisfaction with the system's ability to optimize energy usage during peak times.
- There has been a 25% increase in support tickets related to inefficient energy management over the past year.

Qualitative Information

- Clients often complain about the system's inability to adapt to changing weather conditions and occupancy patterns.
- The current system struggles to balance energy savings with occupant comfort effectively.
- Competitor analysis shows that AI-powered systems are achieving better results in similar building environments.

Strategy Chosen

Implement an AI-powered energy optimization system that uses machine learning algorithms to predict energy demands and automatically adjust building systems for optimal efficiency.

Implementation Plan

1. Data Collection and Analysis
    - Gather historical energy usage data from existing clients.
    - Analyze patterns and identify key factors influencing energy consumption.
2. AI Model Development
    - Develop machine learning models to predict energy demand based on various factors (weather, occupancy, time of day, etc.).
    - Create optimization algorithms to balance energy savings with occupant comfort.
3. Integration and Testing
    - Integrate the AI system with existing building management systems.
    - Conduct extensive testing in various building types and conditions.
4. User Interface Enhancement
    - Develop an intuitive dashboard for real-time monitoring and manual overrides.
    - Create customizable reports for easy interpretation of energy savings.
5. Deployment and Training
    - Roll out the new system to existing clients in phases.
    - Provide comprehensive training to both clients and support staff.

Costs

- Time Cost
    - The project is expected to take 14 months
        * Data Collection and Analysis: 2 months
        * AI Model Development: 5 months

        * Integration and Testing: 3 months
        * User Interface Enhancement: 2 months
        * Deployment and Training: 2 months

- Financial Cost
    - Total: $2.5 million
        * Data Collection and Analysis: $300,000
        * AI Model Development: $1,000,000
        * Integration and Testing: $600,000
        * User Interface Enhancement: $400,000
        * Deployment and Training: $200,000

- Labor Cost
    - Total: 30 personnel
        * Data Science Team: 8 specialists
        * AI Development Team: 10 engineers
        * Integration Team: 6 engineers
        * UI/UX Design Team: 4 designers
        * Deployment and Training Team: 2 specialists

Impact on Users

- Improved Energy Savings
    - Expected to increase average energy reduction from 15% to 35%.

- Enhanced User Satisfaction
    - Anticipated increase in customer satisfaction ratings by 40%.

- Increased System Adoption
    - Predicted 30% rise in the use of advanced features.

Risks and Mitigation

- Data Privacy Concerns:
  - Risk: Clients may be concerned about the collection and use of detailed energy usage data.
  - Likelihood: High
  - Severity: High
  - Mitigation: Implement strict data anonymization and encryption protocols and provide transparent data usage policies.
- System Reliability:
  - Risk: The AI system may make incorrect decisions, leading to energy waste or discomfort.
  - Likelihood: Moderate
  - Severity: High
  - Mitigation: Implement safeguards and manual override options and conduct thorough testing in various scenarios.
- Integration Challenges
  - Risk: Difficulties in integrating with diverse existing building management systems.
  - Likelihood: High
  - Severity: Moderate
  - Mitigation: Develop flexible APIs and conduct extensive compatibility testing.

E.7 Task B: EcoTech Solutions: Sustainable Energy Management System Implementation: Proposal 2 - Modular IoT Integration Platform

Problem Summary

Problem Description EcoTech's current energy management system lacks compatibility with the latest IoT devices and sensors, limiting its ability to provide comprehensive and granular control over building systems. This incompatibility results in inefficient energy management and difficulty in scaling the system for different building types.

Quantitative Information

- 70% of clients report inability to integrate new IoT devices with the current system.
- There's been a 40% increase in requests for custom integrations over the past year.
- 50% of potential clients cite lack of IoT compatibility as a reason for choosing competitors.

Qualitative Information

- Clients express frustration with the limited range of compatible devices and sensors.
- The current system struggles to provide real-time data from diverse sources, hindering quick decision-making.
- Market analysis shows a trend towards more interconnected and responsive building management systems.

Strategy Chosen

Develop a modular IoT integration platform that allows easy connection of various IoT devices and sensors, providing a scalable and future-proof solution for energy management.

Implementation Plan

1. Platform Architecture Design
    - Design a flexible, scalable architecture for IoT device integration.
    - Develop standardized protocols for device communication.
2. Core Platform Development
    - Build the central platform for managing device connections and data flow.
    - Implement security measures for safe device communication.
3. Device Integration Development
    - Create integration modules for popular IoT devices and sensors.
    - Develop an SDK for third-party developers to create custom integrations.
4. User Interface Enhancement
    - Design an intuitive interface for device management and data visualization.
    - Implement customizable dashboards for different user roles.
5. Testing and Quality Assurance
    - Conduct extensive testing with various IoT devices and scenarios.
    - Perform security audits and penetration testing.
6. Deployment and Support
    - Roll out the platform to existing clients in phases.
    - Provide documentation and support for clients and third-party developers.

Costs

- Time Cost
  - The project is expected to take 12 months
    - Platform Architecture Design: 2 months
    - Core Platform Development: 4 months
    - Device Integration Development: 3 months
    - User Interface Enhancement: 1 month
    - Testing and Quality Assurance: 1 month
    - Deployment and Support: 1 month
- Financial Cost
  - Total: $2.2 million
    - Platform Architecture Design: $300,000
    - Core Platform Development: $800,000
    - Device Integration Development: $600,000
    - User Interface Enhancement: $200,000
    - Testing and Quality Assurance: $200,000
    - Deployment and Support: $100,000
- Labor Cost
  - Total: 25 personnel
    - Architecture Team: 4 senior engineers
    - Core Development Team: 8 software engineers
    - Integration Team: 6 developers
    - UI/UX Design Team: 3 designers
    - QA Team: 2 testers
    - Deployment and Support Team: 2 specialists

Impact on Users

- Increased Device Compatibility
  - Expected to support integration with 95% of popular IoT devices.
- Improved System Flexibility
  - Anticipated 50% reduction in time required for custom integrations.
- Enhanced Data Granularity
  - Predicted 40% improvement in the detail and accuracy of energy usage data.

Risks and Mitigation

- Security Vulnerabilities
  - Risk: Increased number of connected devices may introduce new security vulnerabilities.
  - Likelihood: High
  - Severity: High
  - Mitigation: Implement robust encryption, regular security audits, and automatic security updates.
- Complexity for End-Users
  - Risk: The increased options and customization may overwhelm some users.
  - Likelihood: Moderate
  - Severity: Moderate
  - Mitigation: Provide intuitive user interfaces, comprehensive training, and tiered access levels.
- Third-Party Reliability
  - Risk: Reliance on third-party IoT devices may introduce reliability issues.

- Likelihood: Moderate
- Severity: High
- Mitigation: Implement strict certification processes for third-party integrations and provide fallback mechanisms.

E.8 Task B: EcoTech Solutions: Sustainable Energy Management System Implementation: Proposal 3 - Predictive Maintenance and Fault Detection System

Problem Summary

Problem Description EcoTech's current energy management system lacks the ability to proactively identify and address equipment inefficiencies and failures. This results in unexpected downtime, energy waste, and increased maintenance costs for clients. The system's reactive approach to maintenance hinders its effectiveness in optimizing energy usage and ensuring consistent performance of building systems.

Quantitative Information

- Clients report an average of 3 unexpected equipment failures per year, each resulting in 15% increased energy consumption.
- 55% of maintenance activities are reactive rather than preventive or predictive.
- Energy waste due to undetected equipment inefficiencies is estimated at 20% of total consumption.

Qualitative Information

- Clients express frustration with the system's inability to prevent equipment failures and energy waste.
- Maintenance teams struggle to prioritize their activities due to lack of predictive insights.

- Market analysis shows a growing trend towards predictive maintenance in smart building management.

Strategy Chosen

Develop and implement a Predictive Maintenance and Fault Detection System that uses advanced analytics and machine learning to anticipate equipment failures, detect inefficiencies, and optimize maintenance schedules.

Implementation Plan

1. Data Integration and Sensor Network Enhancement:
    - Assess current sensor infrastructure and identify gaps.
    - Implement additional IoT sensors for comprehensive equipment monitoring.
    - Develop data integration pipelines for real-time equipment performance data.
2. Predictive Model Development:
    - Develop machine learning models for failure prediction and anomaly detection.
    - Create algorithms for optimal maintenance scheduling.
    - Design energy efficiency models to detect and quantify waste.
3. Alert and Workflow System:
    - Develop a real-time alert system for impending failures and detected inefficiencies.
    - Create customizable workflows for different types of alerts and maintenance activities.
    - Implement a prioritization system for maintenance tasks.
4. User Interface Development:
    - Design intuitive dashboards for equipment health monitoring and maintenance planning.
    - Develop mobile interfaces for on-the-go alerts and task management.

5. Integration and Testing:
    - Integrate the new system with existing EcoTech platforms.
    - Conduct extensive testing with historical data and in live environments.
6. Training and Deployment:
    - Develop comprehensive training materials for maintenance teams and facility managers.
    - Roll out the system to existing clients in phases, starting with pilot implementations.

Costs

- **Time Cost**
    - The project is expected to take 11 months:
        * Data Integration and Sensor Network Enhancement: 2 months
        * Predictive Model Development: 4 months
        * Alert and Workflow System: 2 months
        * User Interface Development: 1 month
        * Integration and Testing: 1 month
        * Training and Deployment: 1 month
- Financial Cost
    - Total cost: $2.1 million
        * Data Integration and Sensor Network Enhancement: $500,000
        * Predictive Model Development: $800,000
        * Alert and Workflow System: $300,000
        * User Interface Development: $200,000
        * Integration and Testing: $200,000
        * Training and Deployment: $100,000
- Labor Cost

  - Total: 22 personnel
    * Data Integration Team: 4 engineers
    * Machine Learning Team: 6 data scientists
    * Software Development Team: 5 developers
    * UI/UX Design Team: 3 designers
    * QA and Testing Team: 2 engineers
    * Training and Deployment Team: 2 specialists

Impact on Users

- Reduced Unexpected Downtime:
  - Expected to decrease unexpected equipment failures by 80%.
- Improved Energy Efficiency:
  - Anticipated 15% reduction in energy waste due to early detection of inefficiencies.
- Enhanced Maintenance Efficiency:
  - Predicted 40% increase in preventive maintenance activities and 30% reduction in overall maintenance costs.

Risks and Mitigation

- Data Quality and Availability:
  - Risk: Insufficient or poor-quality data may lead to inaccurate predictions.
  - Likelihood: Moderate
  - Severity: High
  - Mitigation: Implement rigorous data quality checks, provide guidelines for sensor deployment, and develop models that can handle data gaps.
- False Positives/Negatives:

  - Risk: System may generate false alarms or miss critical issues.
  - Likelihood: Moderate
  - Severity: High
  - Mitigation: Implement confidence levels for predictions, allow for human oversight, and continuously refine models based on feedback.

- Integration Complexity:
  - Risk: Challenges in integrating with diverse existing building management systems.
  - Likelihood: High
  - Severity: Moderate
  - Mitigation: Develop flexible APIs, create standardized integration protocols, and provide dedicated integration support for complex cases.

Bibliography

[1] S. C. Bunce, K. Izzetoglu, H. Ayaz, P. Shewokis, M. Izzetoglu, K. Pourrezaei, and B. Onaral, "Implementation of fNIRS for Monitoring Levels of Expertise and Mental Workload," in *Foundations of Augmented Cognition. Directing the Future of Adaptive Systems* (D. D. Schmorrow and C. M. Fidopiastis, eds.), (Berlin, Heidelberg), pp. 13–22, Springer Berlin Heidelberg, 2011.

[2] A. J. Steiner, L. Aguilar-Hernandez, R. Abdelsalam, K. Q. Mercado, A. M. Taran, L. E. Gelfond, and W. W. IsHak, "Neurobiological sciences: Neuroanatomy, neurophysiology, and neurochemistry," in *Atlas of Psychiatry*, pp. 91–146, Springer, 2023.

[3] D. L. Collins, P. Neelin, T. M. Peters, and A. C. Evans, "Automatic 3d intersubject registration of mr volumetric data in standardized talairach space," *Journal of computer assisted tomography*, vol. 18, no. 2, pp. 192–205, 1994.

[4] S. Bludau, S. B. Eickhoff, H. Mohlberg, S. Caspers, A. R. Laird, P. T. Fox, A. Schleicher, K. Zilles, and K. Amunts, "Cytoarchitecture, probability maps and functions of the human frontal pole," *NeuroImage*, vol. 93, p. 16, 2014.

[5] G. Ganis and R. Kievit, "A new set of three-dimensional shapes for investigating mental rotation processes: Validation data and stimulus set," *Journal of Open Psychology Data*, vol. 3, Mar. 2015.

[6] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain–computer interfaces for communication and control," *Clinical neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.

[7] K. H. Naseer, Noman, "Fnirs-based brain-computer interfaces: A review," *Frontiers in Human Neuroscience*, vol. 9, Jan. 2015.

[8] A. Mallas, M. Xenos, and C. Katsanos, "A descriptive model of passive and natural passive human-computer interaction," in *Human-Computer Interaction. Theoretical Approaches and Design Methods: Thematic Area, HCI 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings, Part I*, (Berlin, Heidelberg), pp. 104–116, Springer-Verlag, 2022.

[9] T. O. Zander, C. Kothe, S. Welke, and M. Rötting, "Utilizing secondary input from passive brain-computer interfaces for enhancing human-machine interaction," in *Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience: 5th International Conference, FAC 2009 Held as Part of HCI International 2009 San Diego, CA, USA, July 19-24, 2009 Proceedings 5*, pp. 759–771, Springer, Springer Berlin Heidelberg, 2009.

[10] P. Gerjets, C. Walter, W. Rosenstiel, M. Bogdan, and T. O. Zander, "Cognitive state monitoring and the design of adaptive instruction in digital environments: Lessons learned from cognitive workload assessment using a passive brain-computer interface approach," *Frontiers in Neuroscience*, vol. 8, p. 385, Dec. 2014.

[11] J. Borst, A. Bulling, C. Gonzalez, and N. Russwinkel, "Anticipatory human-machine interaction," *dagstuhl Seminar 22202*, 2022.

[12] S. W. Hincks, *A Physical Paradigm for Bidirectional Brain-Computer Interfaces*. PhD thesis, Tufts University, 2019.

[13] S. F. Nolde, M. K. Johnson, and C. L. Raye, "The role of prefrontal cortex during tests of episodic memory," *Trends in Cognitive Sciences*, vol. 2, no. 10, pp. 399–406, 1998.

[14] U. Basten, C. Stelzel, and C. J. Fiebach, "Intelligence is differentially related to neural effort in the task-positive and the task-negative brain network," *Intelligence*, vol. 41, no. 5, pp. 517–528, 2013.

[15] S. A. Akbar, A. T. Mattfeld, A. R. Laird, and D. L. McMakin, "Sleep to internalizing pathway in young adolescents (sipya): A proposed neurodevelopmental model," *Neuroscience & Biobehavioral Reviews*, vol. 140, p. 104780, 2022.

[16] B. J. Rhodes, "The wearable remembrance agent: A system for augmented memory," *Personal Technologies*, vol. 1, pp. 218–224, Dec. 1997.

[17] H. Lieberman, "Letizia: An agent that assists web browsing," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1* (C. S. Mellish, ed.), IJCAI'95, (San Francisco, CA, USA), pp. 924–929, Morgan Kaufmann Publishers Inc., 1995.

[18] N. Ramnani and A. M. Owen, "Anterior prefrontal cortex: Insights into function from anatomy and neuroimaging," *Nature Reviews Neuroscience*, vol. 5, pp. 184–194, Mar. 2004.

[19] M. D'Esposito, B. R. Postle, and B. Rypma, "Prefrontal cortical contributions to working memory: Evidence from event-related fMRI studies," *Experimental Brain Research*, vol. 133, pp. 3–11, July 2000.

[20] D. S. Manoach, G. Schlaug, B. Siewert, D. G. Darby, B. M. Bly, A. Benfield, R. R. Edelman, and S. Warach, "Prefrontal cortex fMRI signal changes are correlated with working memory load," *NeuroReport*, vol. 8, no. 2, pp. 545–549, 1997.

[21] A. Vermeij, A. S. Meel-van den Abeelen, R. P. Kessels, A. H. van Beek, and J. A. Claassen, "Very-low-frequency oscillations of cerebral hemodynamics and blood pressure are affected by aging and cognitive load," *NeuroImage*, vol. 85, pp. 608–615, 2014.

[22] S. Chikhi, N. Matton, and S. Blanchet, "Eeg power spectral measures of cognitive workload: A meta-analysis," *Psychophysiology*, vol. 59, no. 6, p. e14009, 2022.

[23] A. Baddeley, "Working memory," *Current biology*, vol. 20, no. 4, pp. R136–R140, 2010.

[24] E. Tulving, "Memory: Performance, knowledge, and experience," *European Journal of Cognitive Psychology*, vol. 1, pp. 3–26, Mar. 1989.

[25] E. Koechlin, G. Basso, P. Pietrini, S. Panzer, and J. Grafman, "The role of the anterior prefrontal cortex in human cognition," *Nature*, vol. 399, pp. 148–151, May 1999.

[26] A. K. Barbey, M. Koenigs, and J. Grafman, "Dorsolateral prefrontal contributions to human working memory," *Cortex*, vol. 49, pp. 1195–1205, May 2013.

[27] A. Dietrich, "The cognitive neuroscience of creativity," *Psychonomic Bulletin & Review*, vol. 11, pp. 1011–1026, Dec. 2004.

[28] C. Shah, K. Erhard, H.-J. Ortheil, E. Kaza, C. Kessler, and M. Lotze, "Neural correlates of creative writing: An fmri study," *Human Brain Mapping*, vol. 34, no. 5, pp. 1088–1101, 2013.

[29] S. J. Gilbert, S. Spengler, J. S. Simons, J. D. Steele, S. M. Lawrie, C. D. Frith, and P. W. Burgess, "Functional specialization within rostral prefrontal cortex (area 10): A meta-analysis," *Journal of Cognitive Neuroscience*, vol. 18, no. 6, pp. 932–948, 2006.

[30] S. B. Eickhoff, A. R. Laird, P. T. Fox, D. Bzdok, and L. Hensel, "Functional segregation of the human dorsomedial prefrontal cortex," *Cerebral Cortex*, vol. 26, no. 1, pp. 304–321, 2016.

[31] D. M. Amodio and C. D. Frith, "Meeting of minds: The medial frontal cortex and social cognition," *Nature Reviews Neuroscience*, vol. 7, p. 268, Apr. 2006.

[32] J. P. Mitchell, "Social psychology as a natural kind," *Trends in Cognitive Sciences*, vol. 13, no. 6, pp. 246–251, 2009.

[33] D. Bzdok, R. Langner, L. Schilbach, D. A. Engemann, A. R. Laird, P. T. Fox, and S. B. Eickhoff, "Segregation of the human medial prefrontal cortex in social cognition," *Frontiers in Human Neuroscience*, vol. 7, p. 232, 2013.

[34] R. N. Spreng and C. L. Grady, "Patterns of brain activity supporting autobiographical memory, prospection, and theory of mind, and their relationship to the default mode network," *Journal of Cognitive Neuroscience*, vol. 22, no. 6, pp. 1112–1123, 2010.

[35] R. N. Spreng, R. A. Mar, and A. S. N. Kim, "The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis," *Journal of Cognitive Neuroscience*, vol. 21, no. 3, pp. 489–510, 2009.

[36] D. Bzdok, L. Schilbach, K. Vogeley, K. Schneider, A. R. Laird, R. Langner, and S. B. Eickhoff, "Parsing the neural correlates of moral cognition: Ale meta-analysis on morality, theory of mind, and empathy," *Brain Structure and Function*, vol. 217, no. 4, pp. 783–796, 2012.

[37] J. R. Binder, R. H. Desai, W. W. Graves, and L. L. Conant, "Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies," *Cerebral Cortex*, vol. 19, no. 12, pp. 2767–2796, 2009.

[38] H.-Y. Chen, A. W. Gilmore, S. M. Nelson, and K. B. McDermott, "Are there multiple kinds of episodic memory? an fmri investigation comparing autobiographical and recognition memory tasks," *Journal of Neuroscience*, vol. 37, no. 10, pp. 2764–2775, 2017.

[39] S. Fantini and A. Sassaroli, "Frequency-domain techniques for cerebral and functional near-infrared spectroscopy," *Frontiers in neuroscience*, vol. 14, p. 519087, 2020.

[40] A. Sassaroli, M. Pierro, P. R. Bergethon, and S. Fantini, "Low-frequency spontaneous oscillations of cerebral hemodynamics investigated with near-infrared spectroscopy: A review," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 18, pp. 1478–1492, July 2012.

[41] P. Pinti, I. Tachtsidis, A. Hamilton, J. Hirsch, C. Aichelburg, S. Gilbert, and P. W. Burgess, "The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience," *Annals of the New York Academy of Sciences*, vol. 1464, no. 1, pp. 5–29, 2020.

[42] L. F. Nicolas-Alonso and J. Gomez-Gil, "Brain computer interfaces, a review," *Sensors*, vol. 12, no. 2, pp. 1211–1279, 2012.

[43] F. Klein and C. Kranczioch, "Signal processing in fnirs: A case for the removal of systemic activity for single trial data," *Frontiers in Human Neuroscience*, vol. 13, p. 331, 2019.

[44] G. Blaney, A. Sassaroli, T. Pham, C. Fernandez, and S. Fantini, "Phase dual-slopes in frequency-domain near-infrared spectroscopy for enhanced sensitivity to brain tissue: First applications to human subjects," *Journal of biophotonics*, vol. 13, no. 1, p. e201960018, 2020.

[45] Z. Chen and V. Calhoun, "Volumetric bold fMRI simulation: From neurovascular coupling to multivoxel imaging," *BMC Medical Imaging*, vol. 12, pp. 1–13, Apr. 2012.

[46] S. Ogawa, T.-M. Lee, A. R. Kay, and D. W. Tank, "Brain magnetic resonance imaging with contrast dependent on blood oxygenation.," *Proceedings of the National Academy of Sciences*, vol. 87, no. 24, pp. 9868–9872, 1990.

[47] M. L. Schroeter, T. Kupka, T. Mildner, K. Uludağ, and D. Y. von Cramon, "Investigating the post-stimulus undershoot of the bold signal—a simultaneous fMRI and fNIRS study," *NeuroImage*, vol. 30, no. 2, pp. 349–358, 2006.

[48] U. Kreplin and S. H. Fairclough, "Activation of the rostromedial prefrontal cortex during the experience of positive emotion in the context of esthetic experience. an fNIRS study," *Frontiers in Human Neuroscience*, vol. 7, p. 879, 2013.

[49] H. Ayaz, M. Izzetoglu, P. A. Shewokis, and B. Onaral, "Sliding-window motion artifact rejection for functional near-infrared spectroscopy," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pp. 6567–6570, IEEE, IEEE, 2010.

[50] H. Obrig, M. Neufang, R. Wenzel, M. Kohl, J. Steinbrink, K. Einhäupl, and A. Villringer, "Spontaneous low frequency oscillations of cerebral hemodynamics and metabolism in human adults," *NeuroImage*, vol. 12, no. 6, pp. 623–639, 2000.

[51] A. V. Andersen, S. A. Simonsen, H. W. Schytz, and H. K. Iversen, "Assessing low-frequency oscillations in cerebrovascular diseases and related conditions with near-infrared spectroscopy: A plausible method for evaluating cerebral autoregulation?," *Neurophotonics*, vol. 5, no. 03, p. 1, 2018.

[52] Z. Huang, L. Wang, G. Blaney, C. Slaughter, D. McKeon, Z. Zhou, R. Jacob, and M. C. Hughes, "The tufts fNIRS mental workload dataset & benchmark for brain-computer interfaces that generalize," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

[53] A. Girouard, E. T. Solovey, and R. J. Jacob, "Designing a passive brain computer interface using real time classification of functional near-infrared spectroscopy," *International Journal of Autonomous and Adaptive Communications Systems*, vol. 6, no. 1, p. 26, 2013.

[54] D. Afergan, E. M. Peck, E. T. Solovey, A. Jenkins, S. W. Hincks, E. T. Brown, R. Chang, and R. J. Jacob, "Dynamic difficulty using the brain metrics of workload," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, (New York, NY, USA), pp. 3797–3806, ACM, Apr. 2014.

[55] D. Afergan, S. W. Hincks, T. Shibata, and R. J. Jacob, "Phylter: A system for modulating notifications in wearables using physiological sensing," in *International Conference on Augmented Cognition*, pp. 167–177, Springer, Springer International Publishing, 2015.

[56] D. Afergan, T. Shibata, S. W. Hincks, E. M. Peck, B. F. Yuksel, R. Chang, and R. J. Jacob, "Brain-based target expansion," in *Proceedings of the 27th annual ACM symposium on User interface software and technology*, UIST '14, pp. 583–593, ACM, ACM, Oct. 2014.

[57] L. M. Hirshfield, R. Gulotta, S. Hirshfield, S. Hincks, M. Russell, R. Ward, T. Williams, and R. Jacob, "This is your brain on interfaces: Enhancing usability testing with functional near-infrared spectroscopy," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pp. 373–382, ACM, ACM, May 2011.

[58] L. M. Hirshfield, E. T. Solovey, A. Girouard, J. Kebinger, R. J. Jacob, A. Sassaroli, and S. Fantini, "Brain measurement for usability testing and adaptive interfaces: An example of uncovering syntactic workload with functional near infrared spectroscopy," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pp. 2185–2194, ACM, ACM, Apr. 2009.

[59] L. Hirshfield, K. Chauncey, R. Gulotta, A. Girouard, E. Solovey, R. Jacob, A. Sassaroli, and S. Fantini, "Combining electroencephalograph and near infrared spectroscopy to explore users' instantaneous and continuous mental workload states," in *HCI International*, vol. 2009, pp. 239–247, Springer Berlin Heidelberg, 2009.

[60] S. D. Power, T. H. Falk, and T. Chau, "Classification of prefrontal activity due to mental arithmetic and music imagery using hidden markov models and frequency domain near-infrared spectroscopy," *Journal of Neural Engineering*, vol. 7, p. 026002, 02 2010.

[61] M. S. Strait, M, "What we can and cannot (yet) do with functional near-infrared spectroscopy," *Frontiers in Neuroscience*, vol. 8, May 2014.

[62] H. Ayaz, P. A. Shewokis, S. Bunce, K. Izzetoglu, B. Willems, and B. Onaral, "Optical brain monitoring for operator training and mental workload assessment," *NeuroImage*, vol. 59, pp. 36–47, 01 2012.

[63] T. Shibata, A. Borisenko, A. Hakone, T. August, L. Deligiannidis, C. Yu, M. Russell, A. Olwal, and R. J. Jacob, "An implicit dialogue injection system for interruption management," in *Proceedings of the 10th Augmented Human International Conference 2019*, AH2019, (New York, NY, USA), pp. 1–9, ACM, Mar. 2019.

[64] B. Yuksel, K. B. Oleson, L. Harrison, E. M. Peck, D. Afergan, R. Chang, and R. J. Jacob, "Learn piano with bach: An adaptive learning interface that adjusts task difficulty based on brain state," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, (New York, NY, USA), pp. 5372–5384, ACM, May 2016.

[65] A. Girouard, E. T. Solovey, L. M. Hirshfield, K. Chauncey, A. Sassaroli, S. Fantini, and R. J. K. Jacob, *Distinguishing Difficulty Levels with Non-invasive Brain Activity Measurements*, pp. 440–452. Springer Berlin Heidelberg, 2009.

[66] S. Luu and T. Chau, "Decoding subjective preference from single-trial near-infrared spectroscopy signals," *Journal of Neural Engineering*, vol. 6, no. 1, p. 016003, 2008.

[67] S. Moghimi, A. Kushki, S. Power, A. M. Guerguerian, and T. Chau, "Automatic detection of a prefrontal cortical response to emotionally rated music using multi-channel near-infrared spectroscopy," *Journal of Neural Engineering*, vol. 9, p. 026022, Mar. 2012.

[68] M. Okamoto, Y. Wada, Y. Yamaguchi, Y. Kyutoku, L. Clowney, A. K. Singh, and I. Dan, "Process-specific prefrontal contributions to episodic encoding and retrieval of tastes: A functional NIRS study," *NeuroImage*, vol. 54, no. 2, pp. 1578–1588, 2011.

[69] Q. Yu, B. Cheval, B. Becker, F. Herold, C. C. H. Chan, Y. N. Delevoye-Turrell, S. M. R. Guérin, P. Loprinzi, N. Mueller, and L. Zou, "Episodic memory encoding and retrieval in face-name paired paradigm: An fNIRS study," *Brain Sciences*, vol. 11, no. 7, p. 951, 2021.

[70] S. Dong and J. Jeong, "Process-specific analysis in episodic memory retrieval using fast optical signals and hemodynamic signals in the right prefrontal cortex," *Journal of Neural Engineering*, vol. 15, no. 1, pp. 015001–015001, 2018.

[71] S. Jahani, A. L. Fantana, D. Harper, J. M. Ellison, D. A. Boas, B. P. Forester, and M. A. Yücel, "Fnirs can robustly measure brain activity during memory encoding and retrieval in healthy subjects," *Scientific Reports*, vol. 7, no. 1, p. 9533, 2017.

[72] M. Okamoto, M. Matsunami, H. Dan, T. Kohata, K. Kohyama, and I. Dan, "Prefrontal activity during taste encoding: An fNIRS study," *NeuroImage*, vol. 31, no. 2, pp. 796–806, 2006.

[73] L. Ferreri, J.-J. Aucouturier, M. Muthalib, E. Bigand, and A. Bugaiska, "Music improves verbal memory encoding while decreasing prefrontal cortex activity: An fnirs study," *Frontiers in Human Neuroscience*, vol. 7, p. 779, 2013.

[74] M. Abujelala, R. Karthikeyan, O. Tyagi, J. Du, and R. K. Mehta, "Brain activity-based metrics for assessing learning states in VR under stress among firefighters: An explorative machine learning approach in neuroergonomics," *Brain Sciences*, vol. 11, no. 7, p. 885, 2021.

[75] J. F. Burke, M. B. Merkow, J. Jacobs, M. J. Kahana, and K. A. Zaghloul, "Brain computer interface to enhance episodic memory in human participants," *Frontiers in Human Neuroscience*, vol. 8, 01 2015.

[76] G. Durantin, F. Dehais, and A. Delorme, "Characterization of mind wandering using fnirs," *Frontiers in Systems Neuroscience*, vol. 9, p. 45, 2015.

[77] S. W. Hincks, D. Afergan, and R. J. Jacob, "Using fNIRS for real-time cognitive workload assessment," in *International Conference on Augmented Cognition*, pp. 198–208, Springer, Springer International Publishing, 2016.

[78] H. Zhang, Q.-Q. Zhou, H. Chen, X.-Q. Hu, W.-G. Li, Y. Bai, J.-X. Han, Y. Wang, Z.-H. Liang, D. Chen, *et al.*, "The applied principles of eeg analysis methods in neuroscience and clinical neurology," *Military Medical Research*, vol. 10, no. 1, p. 67, 2023.

[79] M. Brienza and O. Mecarelli, "Neurophysiological basis of eeg," *Clinical electroencephalography*, pp. 9–21, 2019.

[80] P. A. Gable, M. W. Miller, and E. M. Bernat, "Introduction: Methods for collecting EEG data for frequency analyses in humans," in *The Oxford Handbook of EEG Frequency*, pp. 3–14, Oxford University Press, 09 2022.

[81] S. J. Luck, *An Introduction to the Event-Related Potential Technique*. A Bradford Book Ser., Cambridge, Mass. [u.a.]: MIT Press, 2nd ed. ed., 2014.

[82] R. Li, D. Yang, F. Fang, K.-S. Hong, A. L. Reiss, and Y. Zhang, "Concurrent fnirs and eeg for brain function investigation: A systematic, methodology-focused review," *Sensors*, vol. 22, no. 15, p. 5865, 2022.

[83] S. McWeeny and E. S. Norton, "Understanding event-related potentials (erps) in clinical and basic language and communication disorders research: A tutorial," *International Journal of Language and Communication Disorders*, vol. 55, no. 4, pp. 445–457, 2020.

[84] K. M. Jensen and J. A. MacDonald, "Towards thoughtful planning of ERP studies: How participants, trials, and effect magnitude interact to influence statistical power across seven ERP components," *Psychophysiology*, vol. 60, p. e14245, Dec. 2023.

[85] T. O. Zander, J. Brönstrup, R. Lorenz, and L. R. Krol, *Towards BCI-Based Implicit Control in Human–Computer Interaction*, pp. 67–90. Springer London, 2014.

[86] D. Zhu, J. Bieger, G. Garcia Molina, and R. M. Aarts, "A survey of stimulation methods used in SSVEP-based BCIs," *Computational Intelligence and Neuroscience*, vol. 2010, no. 1, p. 702357, 2010.

[87] G. Buzsaki, *Rhythms of the Brain*. Oxford university press, 2006.

[88] M. Cohen, *Analyzing Neural Time Series Data: Theory and Practice*. The MIT Press, Jan. 2014.

[89] R. N. Roy, S. Charbonnier, A. Campagne, and S. Bonnet, "Efficient mental workload estimation using task-independent EEG features," *Journal of Neural Engineering*, vol. 13, p. 026019, Feb. 2016.

[90] J.-S. Kang, A. Ojha, and M. Lee, "Concentration monitoring with high accuracy but low cost EEG device," in *Neural Information Processing: 22nd International Conference, ICONIP 2015, November 9-12, 2015, Proceedings, Part IV 22*, pp. 54–60, Springer, Springer International Publishing, 2015.

[91] M. Balconi, E. Grippa, and M. E. Vanutelli, "What hemodynamic (fNIRS), electrophysiological (EEG) and autonomic integrated measures can tell us about emotional processing," *Brain and Cognition*, vol. 95, pp. 67–76, Apr. 2015.

[92] F. M. Garcia-Moreno, M. Bermudez-Edo, M. J. Rodríguez-Fórtiz, and J. L. Garrido, "A CNN-LSTM deep learning classifier for motor imagery EEG detection using a low-invasive and low-cost BCI headband," in *2020 16th international conference on intelligent environments (IE)*, pp. 84–91, IEEE, IEEE, July 2020.

[93] B. Zoefel, R. J. Huster, and C. S. Herrmann, "Neurofeedback training of the upper alpha frequency band in EEG improves cognitive performance," *NeuroImage*, vol. 54, pp. 1427–1431, Jan. 2011.

[94] Y.-S. Jang, S.-L. Lee, and S.-A. Ryu, "Characteristics of frequency band on EEG signal causing human drowsiness," *The Journal of the Korea institute of electronic communication sciences*, vol. 8, no. 6, pp. 949–954, 2013.

[95] C. Q. Lai, H. Ibrahim, M. Z. Abdullah, J. M. Abdullah, S. A. Suandi, and A. Azman, "Artifacts and noise removal for electroencephalogram (EEG): A literature review," in *2018 IEEE Symposium on Computer Applications and Industrial Electronics (ISCAIE)*, pp. 326–332, IEEE, IEEE, 2018.

[96] D. Tanner, K. Morgan-Short, and S. J. Luck, "How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition," *Psychophysiology*, vol. 52, no. 8, pp. 997–1009, 2015.

[97] A. Delorme, "EEG is better left alone," *Scientific Reports*, vol. 13, no. 1, p. 2372, 2023.

[98] G. Vos, M. Ebrahimpour, L. van Eijk, Z. Sarnyai, and M. Rahimi Azghadi, "Stress monitoring using low-cost electroencephalogram devices: A systematic literature review," *International Journal of Medical Informatics*, vol. 198, p. 105859, 2025.

[99] J. LaRocco, M. D. Le, and D.-G. Paeng, "A systemic review of available low-cost eeg headsets used for drowsiness detection," *Frontiers in Neuroinformatics*, vol. 14, 10 2020.

[100] D. Girardi, F. Lanubile, and N. Novielli, "Emotion detection using noninvasive low cost sensors," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 125–130, IEEE, IEEE, 2017.

[101] P. Bashivan, I. Rish, S. Heisig, and R. Jean-Baptiste, "Mental state recognition via wearable EEG," *arXiv preprint arXiv:1602.00985*, 2016.

[102] S. Lee, M. Kim, and M. Ahn, "Evaluation of consumer-grade wireless EEG systems for brain-computer interface applications," *Biomedical Engineering Letters*, vol. 14, no. 6, pp. 1433–1443, 2024.

[103] S. Vashisht and B. Sharma, "A comparative analysis of neurosky and emotiv EEG systems for brain-computer interface applications," in *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*, pp. 1113–1118, IEEE, IEEE, 2024.

[104] S. Morshad, M. R. Mazumder, and F. Ahmed, "Analysis of brain wave data using neurosky mindwave mobile ii," in *Proceedings of the International Conference on Computing Advancements*, ICCA 2020, pp. 1–4, ACM, 2020.

[105] I. Gonzalez, T. Amin, R. Jones, and M. Mongia, "Characterizations of EEG signals from neurosky mindset device," *Houston, Texas: Rice University*, 2012.

[106] N. T. Vo, "Performance of a prosthetic arm using mindwave neurosky sensor," in *2023 8th International Scientific Conference on Applying New Technology in Green Buildings (ATiGB)*, pp. 52–56, IEEE, IEEE, 2023.

[107] W. K. Y. So, S. W. H. Wong, J. N. Mak, and R. H. M. Chan, "An evaluation of mental workload with frontal EEG," *PLoS One*, vol. 12, no. 4, p. e0174949, 2017.

[108] O. E. Krigolson, C. C. Williams, A. Norton, C. D. Hassall, and F. L. Colino, "Choosing muse: Validation of a low-cost, portable EEG system for ERP research," *Frontiers in Neuroscience*, vol. 11, Mar. 2017.

[109] A. Pluta, C. C. Williams, G. Binsted, K. G. Hecker, and O. E. Krigolson, "Chasing the zone: Reduced beta power predicts baseball batting performance," *Neuroscience Letters*, vol. 686, pp. 150–154, 2018.

[110] L. L. Hawley, N. A. Rector, A. DaSilva, J. M. Laposa, and M. A. Richter, "Technology supported mindfulness for obsessive compulsive disorder: Self-reported mindfulness and eeg correlates of mind wandering," *Behaviour Research and Therapy*, vol. 136, p. 103757, 2021.

[111] J. J. Bird, L. J. Manso, E. P. Ribeiro, A. Ekart, and D. R. Faria, "A study on mental state classification using eeg-based brain-machine interface," *2018 International Conference on Intelligent Systems (IS)*, pp. 795–800, 09 2018.

[112] P. Peining, G. Tan, and A. P. Wai, "Evaluation of consumer-grade eeg headsets for bci drone control," in *Proceedings of the IRC Conference on Science, Engineering, and Technology*, 2017.

[113] L. Zhang and H. Cui, "Reliability of muse 2 and tobii pro nano at capturing mobile application users' real-time cognitive workload changes," *Frontiers in Neuroscience*, vol. 16, 2022.

[114] C. McCarthy, N. Pradhan, C. Redpath, and A. Adler, "Validation of the empatica e4 wristband," in *2016 IEEE EMBS International Student Conference (ISC)*, pp. 1–4, IEEE, May 2016.

[115] A. A. Schuurmans, P. De Looff, K. S. Nijhof, C. Rosada, R. H. Scholte, A. Popma, and R. Otten, "Validity of the empatica e4 wristband to measure heart rate variability (HRV) parameters: a comparison to electrocardiography (ECG)," *Journal of Medical Systems*, vol. 44, pp. 1–11, Sept. 2020.

[116] N. Milstein and I. Gordon, "Validating measures of electrodermal activity and heart rate variability derived from the empatica e4 utilized in research settings that involve interactive dyadic states," *Frontiers in Behavioral Neuroscience*, vol. 14, p. 148, Aug. 2020.

[117] J. E. Peabody, R. Ryznar, M. T. Ziesmann, and L. Gillman, "A systematic review of heart rate variability as a measure of stress in medical professionals," *Cureus*, vol. 15, Jan. 2023.

[118] H. F. Posada-Quintero and K. H. Chon, "Innovations in electrodermal activity data collection and signal processing: A systematic review," *Sensors*, vol. 20, p. 479, Jan. 2020.

[119] S. Ba and X. Hu, "Measuring emotions in education using wearable devices: A systematic review," *Computers & Education*, vol. 200, p. 104797, 07 2023.

[120] P. Schmidt, A. Reiss, R. Dürichen, and K. Laerhoven, "Wearable-based affect recognition—a review," *Sensors*, vol. 19, p. 4079, Sept. 2019.

[121] B. Hickey, T. Chalmers, P. Newton, C.-T. Lin, D. Sibbritt, C. McLachlan, R. Clifton-Bligh, J. Morley, and S. Lal, "Smart devices and wearable technologies to detect and monitor mental health conditions and stress: A systematic review," *Sensors*, vol. 21, p. 3461, May 2021.

[122] J. Kim, J. Park, and J. Park, "Development of a statistical model to classify driving stress levels using galvanic skin responses," *Human Factors and Ergonomics in Manufacturing & Service Industries*, vol. 30, pp. 321–328, Apr. 2020.

[123] R. A. Grier, "How high is high? a meta-analysis of NASA-TLX global workload scores," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 59, pp. 1727–1731, Sept. 2015.

[124] S. G. Hart, "Nasa-task load index (NASA-TLX); 20 years later," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 50, no. 9, pp. 904–908, 2006.

[125] InteraXon, "Muse technical specifications," tech. rep., InteraXon Inc., 2023. Accessed: March 21, 2025.

[126] M. Monitor, "Mind monitor," n.d. Retrieved January 18, 2024.

[127] U. S. Urish, "muse-js: Javascript SDK for the muse EEG headset." `https://github.com/urish/muse-js`, 2023. GitHub repository.

[128] Wikipedia Contributors, "Eeg 10-10 system with additional information." Wikimedia Commons, 2022. File: EEG_10-10_system_with_additional_information.svg. Licensed under CC BY-SA 4.0.

[129] D. Slepian, "Prolate spheroidal wave functions, fourier analysis, and uncertainty—v: The discrete case," *Bell System Technical Journal*, vol. 57, no. 5, pp. 1371–1430, 1978.

[130] D. D. Cox, "Spectral analysis for physical applications: Multitaper and conventional univariate techniques," *Technometrics*, vol. 38, no. 3, pp. 294–294, 1993.

[131] J. Candy, "Multipaper spectral estimation: An alternative to the welch periodogram approach," tech. rep., Office of Scientific and Technical Information (OSTI), 2019.

[132] B. Babadi and E. N. Brown, "A review of multitaper spectral analysis," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 5, pp. 1555–1564, 2014.

[133] T. P. Bronez, "On the performance advantage of multitaper spectral analysis," *IEEE Transactions on Signal Processing*, vol. 40, no. 12, pp. 2941–2946, 1992.

[134] A. Gramfort, "Meg and eeg data analysis with mne-python," *Frontiers in Neuroscience*, vol. 7, 2013.

[135] Python Core Team, *Python: A dynamic, open source programming language*. Python Software Foundation, 2019.

[136] J. Fdez, N. Guttenberg, O. Witkowski, and A. Pasquali, "Cross-subject EEG-based emotion recognition through neural networks with stratified normalization," *Frontiers in Neuroscience*, vol. 15, p. 626277, 2021.

[137] M. A. Poole and P. N. O'Farrell, "The assumptions of the linear regression model," *Transactions of the Institute of British Geographers*, no. 52, pp. 145–158, 1971.

[138] P. S. University, "Lesson 12.3.2: Assumptions of simple linear regression," 2025. Accessed: 2025-02-18.

[139] V. Saravanan, G. J. Berman, and S. J. Sober, "Application of the hierarchical bootstrap to multi-level data in neuroscience," *Neurons, behavior, data analysis and theory*, vol. 3, Oct. 2020.

[140] D. G. Gomes, "Should i use fixed effects or random effects when i have fewer than five levels of a grouping factor in a mixed-effects model?," *PeerJ*, vol. 10, p. e12794, 2022.

[141] R. H. Baayen, D. J. Davidson, and D. M. Bates, "Mixed-effects modeling with crossed random effects for subjects and items," *Journal of Memory and Language*, vol. 59, pp. 390–412, Nov. 2008.

[142] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily, "Random effects structure for confirmatory hypothesis testing: Keep it maximal," *Journal of Memory and Language*, vol. 68, pp. 255–278, Apr. 2013.

[143] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017.

[144] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024.

[145] M. G. Kenward and J. H. Roger, "Small sample inference for fixed effects from restricted maximum likelihood," *Biometrics*, vol. 53, pp. 983–997, Sept. 1997.

[146] F. Lewis, A. Butler, and L. Gilbert, "A unified approach to model selection using the likelihood ratio test," *Methods in eEcology and Evolution*, vol. 2, no. 2, pp. 155–162, 2011.

[147] D. Lüdecke, M. S. Ben-Shachar, I. Patil, P. Waggoner, and D. Makowski, "performance: An R package for assessment, comparison and testing of statistical models," *Journal of Open Source Software*, vol. 6, no. 60, p. 3139, 2021.

[148] A. Alin, "Multicollinearity," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 3, pp. 370–374, 2010.

[149] S. R. Searle, F. M. Speed, and G. A. Milliken, "Population marginal means in the linear model: An alternative to least squares means," *The American Statistician*, vol. 34, no. 4, pp. 216–221, 1980.

[150] R. V. Lenth, "Least-squares means: The r package lsmeans," *Journal of statistical software*, vol. 69, pp. 1–33, 2016.

[151] R. V. Lenth, *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2024. R package version 1.10.1.

[152] T. J. VanderWeele and M. B. Mathur, "Some desirable properties of the bonferroni correction: Is the bonferroni correction really so bad?," *American Journal of Epidemiology*, vol. 188, pp. 617–618, 11 2018.

[153] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 57, pp. 289–300, 01 1995.

[154] H. Abdi and L. J. Williams, "Tukey's honestly significant difference (HSD) test," *Encyclopedia of Research Design*, vol. 3, no. 1, pp. 1–5, 2010.

[155] M. Brysbaert and M. Stevens, "Power analysis and effect size in mixed effects models: A tutorial," *Journal of Cognition*, vol. 1, no. 1, 2018.

[156] M. S. Ben-Shachar, D. Lüdecke, and D. Makowski, "effectsize: Estimation of effect size indices and standardized parameters," *Journal of Open Source Software*, vol. 5, no. 56, p. 2815, 2020.

[157] J. T. Mordkoff, "A simple method for removing bias from a popular measure of standardized effect size: Adjusted partial eta squared," *Advances in Methods and Practices in Psychological Science*, vol. 2, pp. 228–232, July 2019.

[158] R. M. Carroll and L. A. Nordholm, "Sampling characteristics of kelley's ε and hays' ω," *Educational and Psychological Measurement*, vol. 35, pp. 541–554, Oct. 1975.

[159] A. Field, *Discovering Statistics Using IBM Spss Statistics*. Sage publications limited, 2024.

[160] J. H. Steiger, "Beyond the f test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis.," *Psychological Methods*, vol. 9, no. 2, p. 164, 2004.

[161] N. Naseer, N. K. Qureshi, F. M. Noori, and K.-S. Hong, "Analysis of different classification techniques for two-class functional near-infrared spectroscopy-based brain-computer interface," *Computational Intelligence and Neuroscience*, vol. 2016, no. 1, p. 5480760, 2016.

[162] M. A. Khan, H. Asadi, L. Zhang, M. R. C. Qazani, S. Oladazimi, C. K. Loo, C. P. Lim, and S. Nahavandi, "Application of artificial intelligence in cognitive load analysis using functional near-infrared spectroscopy: A systematic review," *Expert Systems with Applications*, vol. 249, p. 123717, 09 2024.

[163] D. Bzdok, M. Krzywinski, and N. Altman, "Machine learning: Supervised methods," *Nature Methods*, vol. 15, p. 5, Jan. 2018.

[164] P. Xanthopoulos, P. M. Pardalos, T. B. Trafalis, P. Xanthopoulos, P. M. Pardalos, and T. B. Trafalis, "Linear discriminant analysis," *SpringerBriefs in Optimization*, pp. 27–33, 2013.

[165] Y. Qin, "A review of quadratic discriminant analysis for high-dimensional data," *WIREs Computational Statistics*, vol. 10, p. e1434, May 2018.

[166] N. Thanh Hai, N. Q. Cuong, T. Q. Dang Khoa, and V. Van Toi, "Temporal hemodynamic classification of two hands tapping using functional near—infrared spectroscopy," *Frontiers in Human Neuroscience*, vol. 7, p. 516, 2013.

[167] J. Gemignani, "Classification of fnirs data with lda and svm: a proof-of-concept for application in infant studies," *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 824–827, 11 2021.

[168] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[169] G. L. Prajapati and A. Patle, "On performing classification using svm with radial basis and polynomial kernel functions," in *2010 3rd International Conference on Emerging Trends in Engineering and Technology*, pp. 512–515, IEEE, 11 2010.

[170] C. Eastmond, A. Subedi, S. De, and X. Intes, "Deep learning in fnirs: A review," *Neurophotonics*, vol. 9, pp. 041411–041411, July 2022.

[171] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: A review," *Journal of Neural Engineering*, vol. 16, p. 031001, Apr. 2019.

[172] K. M. Hossain, M. A. Islam, S. Hossain, A. Nijholt, and M. A. R. Ahad, "Status of deep learning for EEG-based brain-computer interface applications," *Frontiers in Computational Neuroscience*, vol. 16, p. 1006763, Jan. 2023.

[173] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[174] J. Opitz, "From bias and prevalence to macro f1, kappa, and MCC: A structured overview of metrics for multi-class evaluation," *Heidelberg University*, 2022.

[175] M. Waskom, "Seaborn: Statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021.

[176] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie, "A survey on evaluation of large language models."

[177] A. Yuan, A. Coenen, E. Reif, and D. Ippolito, "Wordcraft: Story writing with large language models," in *27th International Conference on Intelligent User Interfaces*, IUI '22, pp. 841–852, ACM, 03 2022.

[178] F. Dell'Acqua, E. McFowland, E. R. Mollick, H. Lifshitz-Assaf, K. Kellogg, S. Rajendran, L. Krayer, F. Candelon, and K. R. Lakhani, "Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality," Working Paper 24-013, Harvard Business School Technology & Operations Mgt. Unit, 2023.

[179] S. Noy and W. Zhang, "Experimental evidence on the productivity effects of generative artificial intelligence," *Science*, vol. 381, pp. 187–192, 07 2023.

[180] N. Singh, G. Bernal, D. Savchenko, and E. L. Glassman, "Where to hide a stolen elephant: Leaps in creative writing with multimodal machine intelligence," *ACM Transactions on Computer-Human Interaction*, vol. 30, pp. 1–57, 09 2023.

[181] M. Reza, N. M. Laundry, I. Musabirov, P. Dushniku, Z. Y. M. Yu, K. Mittal, T. Grossman, M. Liut, A. Kuzminykh, and J. J. Williams, "Abscribe: Rapid exploration & organization of multiple writing variations in human-ai co-writing tasks using large language models," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 1–18, ACM, 05 2024.

[182] V. E. Gunser, S. Gottschling, B. Brucker, S. Richter, D. C. Çakir, and P. Gerjets, "The pure poet: How good is the subjective credibility and stylistic quality of literary short texts written with an artificial intelligence tool as compared to texts written by human authors?," in *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*, vol. 44, pp. 60–61, Association for Computational Linguistics, 2022.

[183] E. Brynjolfsson, D. Li, and L. R. Raymond, "Generative AI at work," Working Paper 31161, National Bureau of Economic Research, Apr. 2023.

[184] J. Prather, B. N. Reeves, P. Denny, B. A. Becker, J. Leinonen, A. Luxton-Reilly, G. Powell, J. Finnie-Ansley, and E. A. Santos, ""it's weird that it knows what i want": Usability and interactions with copilot for novice programmers," *ACM Transactions on Computer-Human Interaction*, vol. 31, pp. 1–31, Nov. 2023.

[185] J. Zamfirescu-Pereira, R. Y. Wong, B. Hartmann, and Q. Yang, "Why johnny can't prompt: How non-ai experts try (and fail) to design LLM prompts," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, (New York, NY, USA), pp. 1–21, Association for Computing Machinery, 2023.

[186] A. Ziegler, E. Kalliamvakou, X. A. Li, A. Rice, D. Rifkin, S. Simister, G. Sittampalam, and E. Aftandilian, "Measuring github copilot's impact on productivity," *Communications of the ACM*, vol. 67, pp. 54–63, Feb. 2024.

[187] S. Nguyen, H. M. Babe, Y. Zi, A. Guha, C. J. Anderson, and M. Q. Feldman, "Understanding how developers use large language models in programming tasks: A case study on github copilot," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 1–26, Association for Computing Machinery, 2023.

[188] C. Lawless, J. Schoeffer, L. Le, K. Rowan, S. Sen, C. St. Hill, J. Suh, and B. Sarrafzadeh, ""i want it that way": Enabling interactive decision support using large language models and constraint programming," *ACM Transactions on Interactive Intelligent Systems*, vol. 14, pp. 1–33, Aug. 2024. Just Accepted.

[189] C.-W. Chiang, Z. Lu, Z. Li, and M. Yin, "Enhancing ai-assisted group decision making through llm-powered devil's advocate," in *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, pp. 103–119, ACM, 03 2024.

[190] K. Lakkaraju, S. E. Jones, S. K. R. Vuruma, V. Pallagani, B. C. Muppasani, and B. Srivastava, "Llms for financial advisement: A fairness and efficacy study in personal decision making," in *4th ACM International Conference on AI in Finance*, ICAIF '23, pp. 100–107, ACM, 11 2023.

[191] R. Arakawa and H. Yakura, "Coaching copilot: Blended form of an LLM-powered chatbot and a human coach to effectively support self-reflection for leadership growth," in *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, CUI '24, (New York, NY, USA), pp. 1–14, Association for Computing Machinery, 2024.

[192] S. Huang, X. Zhao, D. Wei, X. Song, and Y. Sun, "Chatbot and fatigued driver: Exploring the use of LLM-based voice assistants for driving fatigue," in *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI EA '24, (New York, NY, USA), pp. 1–8, Association for Computing Machinery, 2024.

[193] S. Suh, M. Chen, B. Min, T. J.-J. Li, and H. Xia, "Luminate: Structured generation and exploration of design space with large language models for human-ai co-creation," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 1–26, ACM, 05 2024.

[194] S. Suh, B. Min, S. Palani, and H. Xia, "Sensecape: Enabling multilevel exploration and sensemaking with large language models," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, (New York, NY, USA), pp. 1–18, Association for Computing Machinery, 2023.

[195] L. Tankelevitch, V. Kewenig, A. Simkute, A. E. Scott, A. Sarkar, A. Sellen, and S. Rintel, "The metacognitive demands and opportunities of generative ai," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 1–24, ACM, 05 2024.

[196] A. Cambon, B. Hecht, B. Edelman, D. Ngwe, S. Jaffe, A. Heger, M. Vorvoreanu, S. Peng, J. Hofman, A. Farach, M. Bermejo-Cano, E. Knudsen, J. Bono, H. Sanghavi, S. Spatharioti, D. Rothschild, D. Goldstein, E. Kalliamvakou, P. Cihon, and M. Demirer, "Early LLM-based tools for enterprise information workers likely provide meaningful boosts to productivity a first update from microsoft's research initiative on AI and productivitywith additional support from the entire ai and productivity team at microsoft)," tech. rep., Microsoft, Inc, 2023.

[197] Microsoft, "Copilot overview - azure cognitive services." `https://learn.microsoft.com/en-us/copilot/overview`, 2023. Accessed: 2023-06-07.

[198] M. Haslberger, J. Gingrich, and J. Bhatia, "No great equalizer: Experimental evidence on AI in the UK labor market," 2023.

[199] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in Public Health*, vol. 5, p. 258, Sept. 2017.

[200] H. F. Posada-Quintero, J. P. Florian, A. D. Orjuela-Cañón, T. Aljama-Corrales, S. Charleston-Villalobos, and K. H. Chon, "Power spectral density analysis of electrodermal activity for sympathetic function assessment," *Annals of Biomedical Engineering*, vol. 44, no. 10, pp. 3124–3135, 2016.

[201] J. J. Bartko, "The intraclass correlation coefficient as a measure of reliability," *Psychological Reports*, vol. 19, no. 1, pp. 3–11, 1966.

[202] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, 2016.

[203] A. L. Oberg and D. W. Mahoney, *Linear Mixed Effects Models*, pp. 213–234. Humana Press, 2007.

[204] Q. H. Vuong, "Likelihood ratio tests for model selection and non-nested hypotheses," *Econometrica*, vol. 57, no. 2, pp. 307–333, 1989.

[205] C. Herff, D. Heger, O. Fortmann, J. Hennrich, F. Putze, and T. Schultz, "Mental workload during n-back task—quantified in the prefrontal cortex using fNIRS," *Frontiers in Human Neuroscience*, vol. 7, p. 935, 2014.

[206] J. Butler, S. Jaffe, N. Baym, M. Czerwinski, S. Iqbal, K. Nowak, R. Rintel, A. Sellen, M. Vorvoreanu, B. Hecht, and J. Teevan, "Microsoft new future of work report 2023," Tech Report MSR-TR-2023-34, Microsoft Research, 2023.

[207] E. Tulving, S. Kapur, F. Craik, M. Moscovitch, and S. Houle, "Hemispheric encoding/retrieval asymmetry in episodic memory: Positron emission tomography findings.," *Proceedings of the National Academy of Sciences*, vol. 91, pp. 2016–2020, Mar. 1994.

[208] V. Menon, "20 years of the default mode network: A review and synthesis," *Neuron*, vol. 111, no. 16, pp. 2469–2487, 2023.

[209] A. J. Gerber, J. Posner, D. Gorman, T. Colibazzi, S. Yu, Z. Wang, A. Kangarlu, H. Zhu, J. Russell, and B. S. Peterson, "An affective circumplex model of neural systems subserving valence, arousal, and cognitive overlay during the appraisal of emotional faces," *Neuropsychologia*, vol. 46, pp. 2129–2139, 07 2008.

[210] J. B. Gilbert, "Modeling item-level heterogeneous treatment effects: A tutorial with the glmer function from the lme4 package in r," *Behavior Research Methods*, vol. 56, pp. 5055–5067, 11 2023.

[211] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[212] V. Goel, *Hemispheric asymmetry in the prefrontal cortex for complex cognition*, pp. 179–196. Elsevier, 2019.

[213] R. E. Beaty, M. Benedek, R. W. Wilkins, E. Jauk, A. Fink, P. J. Silvia, D. A. Hodges, K. Koschutnig, and A. C. Neubauer, "Creativity and the default network: A functional connectivity analysis of the creative brain at rest," *Neuropsychologia*, vol. 64, pp. 92–98, Nov. 2014.

[214] W. D. Gray and D. A. Boehm-Davis, "Milliseconds matter: An introduction to microstrategies and their use in describing and predicting interactive behavior," *Journal of Experimental Psychology: Applied*, vol. 6, no. 4, pp. 322–335, 2001.

[215] J. Chan, P. Siangliulue, D. Qori McDonald, R. Liu, R. Moradinezhad, S. Aman, E. T. Solovey, K. Z. Gajos, and S. P. Dow, "Semantically far inspirations considered harmful?: Accounting for cognitive states in collaborative ideation," in *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*, C&C '17, (New York, NY, USA), pp. 93–105, ACM, June 2017.

[216] R. L. Buckner, J. R. Andrews-Hanna, and D. L. Schacter, "The brain's default network: Anatomy, function, and relevance to disease," *Annals of the New York Academy of Sciences*, vol. 1124, pp. 1–38, Mar. 2008.

[217] E. Bartoli, E. Devara, H. Q. Dang, R. Rabinovich, R. K. Mathura, A. Anand, B. R. Pascuzzi, J. Adkinson, Y. N. Kenett, K. R. Bijanki, *et al.*, "Default mode network electrophysiological dynamics and causal role in creative thinking," *Brain*, vol. 147, p. awae199, June 2024.

[218] C. Stelzel, H. Bohle, G. Schauenburg, H. Walter, U. Granacher, M. A. Rapp, and S. Heinzel, "Contribution of the lateral prefrontal cortex to cognitive-postural multitasking," *Frontiers in Psychology*, vol. 9, p. 1075, 2018.

[219] C. Herff, D. Heger, F. Putze, J. Hennrich, O. Fortmann, and T. Schultz, "Classification of mental tasks in the prefrontal cortex using fnirs," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2160–2163, IEEE, IEEE, 2013.

[220] H. N. Modi, H. Singh, A. Darzi, and D. R. Leff, "Multitasking and time pressure in the operating room: Impact on surgeons' brain function," *Annals of Surgery*, vol. 272, pp. 648–657, July 2020.

[221] D. Mahesan, D. Antonenko, A. Flöel, and R. Fischer, "Modulation of the executive control network by anodal tdcs over the left dorsolateral prefrontal cortex improves task shielding in dual tasking," *Scientific Reports*, vol. 13, p. 6177, Apr. 2023.

[222] L. Kocsis, P. Herman, and A. Eke, "The modified beer-lambert law revisited," *Physics in Medicine and Biology*, vol. 51, pp. N91–8, 02 2006.

[223] E. Kirilina, A. Jelzow, A. Heine, M. Niessing, H. Wabnitz, R. Brühl, B. Ittermann, A. M. Jacobs, and I. Tachtsidis, "The physiological origin of task-evoked systemic artefacts in functional near infrared spectroscopy," *NeuroImage*, vol. 61, pp. 70–81, May 2012.

[224] A.-K. Seghouane and D. Ferrari, "Robust hemodynamic response function estimation from fNIRS signals," *IEEE Transactions on Signal Processing*, vol. 67, pp. 1838–1848, Apr. 2019.

[225] E. T. Solovey, A. Girouard, K. Chauncey, L. M. Hirshfield, A. Sassaroli, F. Zheng, S. Fantini, and R. J. Jacob, "Using fnirs brain sensing in realistic HCI settings: Experiments and guidelines," in *Proceedings of the 22nd annual ACM symposium on User interface software and technology*, UIST '09, pp. 157–166, ACM, ACM, Oct. 2009.

[226] Y. Zhang, J. W. Sun, and P. Rolfe, "Rls adaptive filtering for physiological interference reduction in nirs brain activity measurement: a monte carlo study," *Physiological Measurement*, vol. 33, pp. 925–942, 05 2012.

[227] Q. Zhang, E. Brown, and G. E. Strangman, "Adaptive filtering to reduce global interference in evoked brain activity detection: A human subject case study," *Journal of Biomedical Optics*, vol. 12, no. 6, p. 064009, 2007.

[228] M. W. Voss, "Chapter 9 - the chronic exercise–cognition interaction: fMRI research," in *Exercise-Cognition Interaction* (T. McMorris, ed.), pp. 187–209, San Diego: Academic Press, 2016.

[229] E. T. Solovey, F. Lalooses, K. Chauncey, D. Weaver, M. Parasi, M. Scheutz, A. Sassaroli, S. Fantini, P. Schermerhorn, A. Girouard, *et al.*, "Sensing cognitive multitasking for a brain-based adaptive user interface," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, CHI '11, pp. 383–392, ACM, May 2011.

[230] S. W. Hincks, S. Bratt, S. Poudel, V. Phoha, D. C. Dennett, R. J. K. Jacob, and L. M. Hirshfield, "Entropic brain-computer interfacing: Using fNIRS and EEG to measure attentional states in a bayesian framework," in *PhyCS* (A. Pope, H. P. Silva, and A. Holzinger, eds.), ([Setúbal, Portugal]), pp. 23–34, SCITEPRESS - Science and Technology Publications, Lda., 2017. Literaturangaben.

[231] A. Gevins and M. E. Smith, "Neurophysiological measures of cognitive workload during human-computer interaction," *Theoretical Issues in Ergonomics Science*, vol. 4, pp. 113–131, 01 2003.

[232] P. Zarjam, J. Epps, and N. H. Lovell, "Beyond subjective self-rating: Eeg signal classification of cognitive workload," *IEEE Transactions on Autonomous Mental Development*, vol. 7, pp. 301–310, 12 2015.

[233] J. Sabio, N. S. Williams, G. M. McArthur, and N. A. Badcock, "A scoping review on the use of consumer-grade EEG devices for research," *PLOS ONE*, vol. 19, p. e0291186, Mar. 2024.

[234] K. Värbu, N. Muhammad, and Y. Muhammad, "Past, present, and future of eeg-based bci applications," *Sensors*, vol. 22, p. 3331, 04 2022.

[235] D. J. McFarland and J. R. Wolpaw, "EEG-based brain–computer interfaces," *Current Opinion in Biomedical Engineering*, vol. 4, pp. 194–200, Sept. 2017.

[236] Lichess, "Lichess open source project," n.d. Retrieved January 18, 2024.

[237] A. Kawala-Sterniuk, M. Podpora, M. Pelc, M. Blaszczyszyn, E. J. Gorzelanczyk, R. Martinek, and S. Ozana, "Comparison of smoothing filters in analysis of EEG data for the medical diagnostics purposes," *Sensors*, vol. 20, p. 807, Feb. 2020.

[238] The Stockfish developers, "Stockfish," 2024. Computer software.

[239] T. Biswas and K. Regan, "Measuring level-k reasoning, satisficing, and human error in game-play data," in *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pp. 941–947, IEEE, IEEE, Dec. 2015.

[240] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), pp. 785–794, ACM, Aug. 2016.

[241] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, K. Chen, R. Mitchell, I. Cano, T. Zhou, M. Li, J. Xie, M. Lin, Y. Geng, Y. Li, and J. Yuan, *XGBOOST: Extreme Gradient Boosting*, 2024. R package version 1.7.8.1.

[242] W. K. Kirchner, "Age differences in short-term retention of rapidly changing information.," *Journal of Experimental Psychology*, vol. 55, no. 4, pp. 352–358, 1958.

[243] S. M. Jaeggi, M. Buschkuehl, W. J. Perrig, and B. Meier, "The concurrent validity of the n-back task as a working memory measure," *Memory*, vol. 18, no. 4, pp. 394–412, 2010.

[244] A. R. A. Conway, M. J. Kane, M. F. Bunting, D. Z. Hambrick, O. Wilhelm, and R. W. Engle, "Working memory span tasks: A methodological review and user's guide," *Psychonomic Bulletin and Review*, vol. 12, no. 5, pp. 769–786, 2005.

[245] M. J. Kane and R. W. Engle, "The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective," *Psychonomic Bulletin & Review*, vol. 9, pp. 637–671, Dec. 2002.

[246] J. Gwizdka, "Using stroop task to assess cognitive load," in *Proceedings of the 28th Annual European Conference on Cognitive Ergonomics*, ECCE '10, pp. 219–222, ACM, 2010.

[247] C. M. MacLeod, "John ridley stroop: Creator of a landmark cognitive task.," *Canadian Psychology / Psychologie canadienne*, vol. 32, no. 3, p. 521, 1991.

[248] C. M. MacLeod, "The stroop effect," *The Corsini Encyclopedia of Psychology*, pp. 1–6, Jan. 2015.

[249] R. N. Shepard and J. Metzler, "Mental rotation of three-dimensional objects," *Science*, vol. 171, no. 3972, pp. 701–703, 1971.

[250] B. A. Osuagwu and A. Vuckovic, "Similarities between explicit and implicit motor imagery in mental rotation of hands: An EEG study," *Neuropsychologia*, vol. 65, pp. 197–210, Dec. 2014.

[251] S. Villafaina, D. Collado-Mateo, R. Cano-Plasencia, N. Gusi, and J. P. Fuentes, "Electroencephalographic response of chess players in decision-making processes under time pressure," *Physiology & Behavior*, vol. 198, pp. 140–143, Jan. 2019.

[252] S. Villafaina, M. A. Castro, T. Pereira, A. Carvalho Santos, and J. P. Fuentes-García, "Neurophysiological and autonomic responses of high and low level chess players during difficult and easy chess endgames – a quantitative EEG and HRV study," *Physiology & Behavior*, vol. 237, p. 113454, Aug. 2021.

[253] J. P. Fuentes-García, S. Villafaina, D. Collado-Mateo, R. Cano-Plasencia, and N. Gusi, "Chess players increase the theta power spectrum when the difficulty of the opponent increases: An EEG study," *International Journal of Environmental Research and Public Health*, vol. 17, pp. 46–46, Dec. 2019.

[254] J. R. de Leeuw, R. A. Gilbert, and B. Luchterhandt, "jspsych: Enabling an open-source collaborative ecosystem of behavioral experiments," *Journal of Open Source Software*, vol. 8, p. 5351, May 2023.

[255] J. Schmich, "Chess.js: A javascript chess library." `https://www.npmjs.com/package/chess.js`, 2023. Version 1.0.0.

[256] C. Oakman, "Chessboard.js: A javascript chessboard." `https://www.npmjs.com/package/chessboardjs`, 2023. Version 1.0.0.

[257] Lichess Team, "Lichess database." `https://database.lichess.org/`, 2024. Accessed: January 2023.

[258] M. E. Glickman, "Example of the glicko-2 system," *Boston University*, vol. 28, 2012.

[259] Lichess Team, "Weekly classical rating distribution." `https://lichess.org/stat/rating/distribution/classical`, 2024. [Accessed 21-10-2024].

[260] M. Jas, D. A. Engemann, Y. Bekhti, F. Raimondo, and A. Gramfort, "Autoreject: Automated artifact rejection for MEG and EEG data," *NeuroImage*, vol. 159, pp. 417–429, Oct. 2017.

[261] D. G. Wakeman and R. N. Henson, "A multi-subject, multi-modal human neuroimaging dataset," *Scientific Data*, vol. 2, p. 150001, Jan. 2015.

[262] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, pp. E215–E220, 06 2000.

[263] G. Schalk, D. McFarland, T. Hinterberger, N. Birbaumer, and J. Wolpaw, "Bci2000: A general-purpose brain-computer interface (BCI) system," *IEEE Transactions on Biomedical Engineering*, vol. 51, pp. 1034–1043, June 2004.

[264] C.-Y. Chang, S.-H. Hsu, L. Pion-Tonachini, and T.-P. Jung, "Evaluation of artifact subspace reconstruction for automatic EEG artifact removal," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1242–1245, IEEE, 07 2018.

[265] L. J. Gabard-Durnam, A. S. Mendez Leal, C. L. Wilkinson, and A. R. Levin, "The harvard automated processing pipeline for electroencephalography (HAPPE): Standardized processing software for developmental and high-artifact data," *Frontiers in Neuroscience*, vol. 12, Feb. 2018.

[266] S. Lo and S. Andrews, "To transform or not to transform: Using generalized linear mixed models to analyse reaction time data," *Frontiers in Psychology*, vol. 6, p. 1171, Aug. 2015.

[267] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017.

[268] X. Jia and A. Kohn, "Gamma rhythms in the brain," *PLoS Biology*, vol. 9, p. e1001045, Apr. 2011.

[269] J. D. Kropotov, "Beta and gamma rhythms," in *Functional Neuromarkers for Psychiatry* (J. D. Kropotov, ed.), pp. 107–119, San Diego: Elsevier, 2016.

[270] J. D. Kropotov, *Frontal Midline Theta Rhythm*, pp. 77–95. San Diego: Elsevier, 2009.

[271] W. Klimesch, "Alpha-band oscillations, attention, and controlled access to stored information," *Trends in Cognitive Sciences*, vol. 16, pp. 606–617, Dec. 2012.

[272] J. Zhao, W.-K. Liang, C.-H. Juan, L. Wang, S. Wang, and Z. Zhu, "Dissociated stimulus and response conflict effect in the stroop task: Evidence from evoked brain potentials and brain oscillations," *Biological Psychology*, vol. 104, pp. 130–138, Jan. 2015.

[273] A. Tafuro, E. Ambrosini, O. Puccioni, and A. Vallesi, "Brain oscillations in cognitive control: A cross-sectional study with a spatial stroop task," *Neuropsychologia*, vol. 133, p. 107190, Oct. 2019.

[274] F. D. Moura Neto and A. J. da Silva Neto, *An Introduction to Inverse Problems with Applications*, ch. 1, p. xi. Springer Berlin Heidelberg, 2013.

[275] D. S. Tan, *Brain-Computer Interfaces*. Human-Computer Interaction Ser., Springer London, Limited, 2010. Description based on publisher supplied metadata and other sources.

[276] A. Girouard, E. T. Solovey, L. M. Hirshfield, E. M. Peck, K. Chauncey, A. Sassaroli, S. Fantini, and R. J. K. Jacob, *From Brain Signals to Adaptive Interfaces: Using fNIRS in HCI*, pp. 221–237. Springer London, 2010.

[277] D. Plass-Oude Bos, B. Reuderink, B. van de Laar, H. Gürkök, C. Mühl, M. Poel, A. Nijholt, and D. Heylen, *Brain-Computer Interfacing and Games*, pp. 149–178. Springer London, 2010.

[278] S. H. Fairclough, "Fundamentals of physiological computing," *Interacting with Computers*, vol. 21, pp. 133–145, 01 2009.

[279] S. H. Fairclough, *Physiological Computing and Intelligent Adaptation*, pp. 539–556. Elsevier, 2017.

[280] N. Semertzidis, F. Zambetta, and F. Mueller, "Brain-computer integration: A framework for the design of brain-computer interfaces from an integrations perspective," *ACM Transactions on Computer-Human Interaction*, vol. 30, pp. 1–48, 09 2023.

[281] B. Lamichhane, A. Westbrook, M. W. Cole, and T. S. Braver, "Exploring brain-behavior relationships in the n-back task," *NeuroImage*, vol. 212, p. 116683, 2020.

[282] The College Board, "SAT Practice Test #5," 2016. Accessed: 2025-03-17.

[283] The College Board, "SAT Practice Test #7," 2016. Accessed: 2025-03-17.