Towards objective quantification of affective experience: integrating cognitive tasks, neuroimaging and large language models

Jiazhou Chen

Supervised by
Prof. Quentin J.M. Huys (Primary), Prof. Argyris Stringaris (Secondary), and Dr. Dylan Nielson

Submitted in partial fulfillment of the requirements for a
Doctor of Philosophy (PhD) in Computational Neuroscience

Division of Psychiatry
University College London (UCL)
2025

Dedicated to my grandfather, CHEN QIAO NIAN

Doctoral Candidate Thesis Declaration Form

I, Jiazhou Chen, confirm that:

- This thesis is a presentation of original work.
- This work has not previously been presented for a degree or other qualification at this University or elsewhere.
- All input or assistance in the creation of the academic work other than from the supervisory team (including AI) has been acknowledged below.
- Any text or data in the thesis that has been presented for publication (including in review) elsewhere is declared in Part B below.
- Where information has been derived from other sources, this has been indicated in the thesis. the thesis.

Declarations

- **Funding Support**: J.C. was supported by the Intramural Training Program of the National Institute of Mental Health (NIMH, part of the National Institutes of Health, Bethesda, MD, U.S.A.), as part of the UCL-NIMH Joint Doctoral Program in Neuroscience. D.M.N was supported by the Intramural Research Program of the NIMH (grant no. ZICMH002968, to Francisco Pereira). Q.J.M.H has received consultancy fees and options from Aya Health and Alto Neuroscience and a research grant from Koa Health. J.C. A.S. and D.M.N. have no conflict of interests to report. The work was supported by the UCL NIH BRC (London, U.K.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. The views expressed in this article do not necessarily represent the views of the NIMH, the Department of Health and Human Services or the United States Government.

  The work was supported by the UCL NIH BRC (London, U.K.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

- **Software code**: customized R and python scripts were used for data analysis. JavaScript, HTML and python scripts were used for data representation and collection. All external packages used were properly cited throughout this thesis. Contribution and assistance were received from Tore Erdmann, Jade Serfaty,

Anna Hall, Jakub Onysk, Anahit Mkrtchian, Zeguo Qi, Haoyang Lu, Yaniv Abir, Yu Zhou, and Sepehr Shirani in addition to the members of the supervisory team. ChatGPT and Github Copilot were used to troubleshoot JavaScript and HTML codes.

- **AI**: generative Transformer Models, including LLaMA2, GPT-2, GPT-2X, BERT, and DistillBERT, were used as part of the experimental work to objectively quantify emotional states from text. Gemini 2.5 Pro model was used during the editing of this chapter to fix grammar errors and improve readability. It was not used during the drafting nor was it used to generate novel content.
- **Proofreading and editorial support**: Evan Taylor Tegley and Janet Bui provided substantial proofreading and editing support. Gemini 2.5 Pro model was used during the editing of this thesis to fix grammar errors and improve readability. It was not used during the drafting nor was it used to generate novel content.
- **Publications**: content in chapter 2 was uploaded to the PsyArXiv preprint server. Citation: Chen, J., Huys, Q. J., Stringaris, A., & Nielson, D. M. (2023). On the misery of cognitive effort. PsyArXiv Preprint. https://doi.org/10.31234/osf.io/r9e4j
- The views expressed in this thesis do not necessarily represent the views of the NIMH, the Department of Health and Human Services or the United States Government.

UCL Research Paper Declaration Form

For a manuscript prepared for publication but not yet published

- **Current title of the manuscript:** On the misery of cognitive effort
- **Has it been uploaded to a preprint server (e.g., medRxiv, arXiv)?** Yes, https://doi.org/10.31234/osf.io/r9e4j
- **Intended publication outlet:** Journal of Experimental Psychology: General
- **List all authors in the intended authorship order:** Jiazhou Chen, Quentin J.M. Huys, Argyris Stringaris, Dylan M. Nielson
- **Current stage of publication:** Pending
- **For multi-authored work, please provide a contribution statement detailing each author's role**: Conceptualization: J.C., Q.J.M.H, A.S., and D.M.N; Experiment Program written by J.C., and D.M.N; Data collection: J.C.; Data analysis: J.C. and D.M.N; Manuscript written and revised by J.C., Q.M.H, A.S., and D.M.N.
- **In which chapter(s) of your thesis can this material be found?**: Chapter 2

e-Signatures confirming accuracy of the above information

Candidate signature:

Jiazhou Chen

Date: Oct 30, 2025

Supervisor/ Senior Author signature:

Quentin Huys

Date: Oct 30, 2025

Abstract

Assessing affective states is difficult: self-report is intrusive, biased, and reactive, while objective correlates like heart rate are often one-dimensional. We conceptualize internal affective states as computational metareasoning heuristics. Given that Large Language Models (LLMs) share parallel computational structures, we investigate whether LLMs might have utility in objectively quantifying affective states.

Chapter 2 first examined the negative influence of cognitive effort on self-reported happiness. The results highlighted the limitations of embedded self-reports as a tool to investigate emotion dynamics in cognitive tasks.

We then approached the main aim of the thesis which stretches across the next four chapters. The aim was to examine LLM-derived affective neural representations, focusing on internal and external validity. We started by describing a novel pipeline to derive continuous, high-dimensional affective state quantifications from text. We then acquired magnetoencephalography (MEG) recordings of participants listening to stories, which allowed us to examine the extent to which LLM-derived affective quantifications align with neural representations. A particular focus was on the generality or context-specificity of the representations. To examine this, we created a novel battery of non-linguistic cognitive tasks that elicits a wide range of self-reported emotions.

Chapter 3 described results showing that LLM-derived quantifications satisfy key internal consistency requirements, and chapter 4 showed that the corresponding neural show some consistency and distinctiveness. Chapter 5 introduced the novel affective task battery, and chapter 6 finally examines the generalization properties of the neural affective state markers.

Overall, we found that affective state markers perform best within their training domain; generalize less to novel stories, and mostly fail to generalize from stories to cognitive tasks. These results indicated that there is an important contextual component to the neural representation of affective categories.

Impact statement

Impact on Academia

This research provided empirical support for a functional, computational view of affective states. The preregistered work demonstrating the negative influence of cognitive effort on self-reported momentary happiness provides support for the idea that affective states tracks appraisal of recent experiences. By providing open-access code and analysis pipelines, this work also promoted transparency and reproducibility.

Methodologically, the thesis introduced a novel, modular LLM pipeline that contrasts with ”black-box” approaches. By separating predictive generation from emotional appraisal, our pipeline increased interpretability and mechanistic transparency. It explicitly modeled the core metareasoning functions of prediction and appraisal, allowing independent validation of each computational stage and inspection of the predictive content being appraised. This approach provided a framework for injecting mechanistic hypotheses into LLM-based modeling.

By demonstrating that neural decoders trained on story-listening failed to cross-modally generalize to cognitive tasks, we provided empirical evidence for a constructionist view of emotion. This finding indicated that the neural representation of affect is dominated by context-dependent signals. This established a critical benchmark for the field: future claims of ”emotion decoding” must be validated through cross-modal generalization tests to prove that they were capturing an abstract affective state rather than a context-specific artifact. Finally, the battery of cognitive tasks offered a reusable, open-access resource for eliciting and studying granular, self-reported emotions.

Impact Outside Academia

The long-term impact of this work lies in clinical applications for mood disorders patients. The pipeline’s unique focus on prediction offered a potential new diagnostic avenue. With future development and refinement, one could utilize similar pipeline as a clinical tool. By asking patients to autocomplete some stories snippets and recording their neural signal, care providers might be able to extract objective quantified affective characteristics.

Acknowledgements

My biggest thanks go to Professor Quentin Huys, my primary mentor. I first met Quentin at a conference in Chicago, where I sweated through his tough questions, ranging from modeling human behavior to theories of emotions. I thought I did poorly, but days later, an email landed in my inbox filled with funding opportunities and encouraging words. This perfectly summarizes his mentorship: he asks the tough questions but is always more than willing to provide the resources to help you figure them out. Throughout my PhD, he guided me from vague curiosity to clear research questions, pushed me to improve my skills, kept me on track while letting me explore interesting ideas, and had an eye for detail, which refined the quality of our work. To all his dedication and mentorship, I am grateful.

I'd also like to thank Professor Argyris Stringaris and Dr. Dylan Nielson, who were instrumental to this work. As my supervisory team, they provided valuable feedback and analysis directions. It was always intellectually challenging to bounce ideas off them. Without their guidance, this work would not have been possible.

To everyone at the Applied Computational Psychiatry lab, the Max Planck UCL Centre for Computational Psychiatry and Ageing Research, and the NIMH Section on Development and Affective Neuroscience: thank you for including me as part of the team and making me feel so welcomed. I am grateful to Anna, Jakub, Jade, Quentin D, Michaela, Tore, Zeguo, Haoyang, Muzhi, Taekwan, Nikki, and Sepehr, who have been not only incredible scholars from whom I learned so much, but also great friends. Sincere thanks to Hanna, Lucrezia, Song, Charles, Agnes, Anahit, Yaniv, and Yu for sharing your expertise.

I am tremendously grateful to be part of my program, the UCL-NIMH joint doctoral program in Neuroscience. Being able to do a PhD is a rare privilege, and even fewer people have the opportunity to do that in a different country. I want to especially acknowledge the people behind the program: the directors from both institutions, Dr. Janet Clark at NIMH and Prof. Jonathan Roiser, and the Program Specialist Aneka Reid. For a PhD with a surplus of turbulence from the external world-COVID and government shutdowns, to name a couple-I was never made to feel alone.

To my sister/cousin, Shengxin, and my aunt, Li: This thesis would have been impossible without your help. LITERALLY. I lived in your home for years without rent, but there's more to it than money. Shengxin: I am so happy we got to live together again. Thank you for all the adventures, taking care of me and just being my family. My aunt Li: you care so much and take on so many responsibilities, me included, even when

you didn't have to. I am truly indebted for both of your support.

To my mom, Yuanmei: You are the most lively and vibrant person I know. When I was a kid, you had an ambitious goal: we'd step foot on all of China. And we did, right before my 18th birthday. From that, I learned that impressive goals take time, careful planning, and lots of effort. During high school, I was always so upset and really wanted a new start. When you saw how much I was hurting, you just decided that it is time to pack up and move. You left behind your professorship, a title that others cling onto for life. Four years later in Texas, we graduated the same year, you with your second master's and a brand new career. From that, you showed me that nothing in life is that scary and while changes might seem daunting, if you are determined, they are just opportunities in disguise. I thank you for your parenting and shaping me. On the same note, I'd also like to thank my step-dad Yanan, who loved us and gave us opportunities to thrive in a new country which was once foreign but we now call home.

To Evan, my husband: Thank you for always being there for me, no matter where you are. Midway through this PhD, I proposed. I stumbled over my words and made some flower arrangement that looked more like a crime scene than decoration. We also never held a wedding, so I wanted to write something for you here. Not as a romantic gesture but an expression of how grateful I am for you. Your care and love have become the one constant in this crazy world that I can lean on. I'd wanted to list all the ways you made me feel luck. But just like I've done many times in 2022, when I sat down and tried to write, nothing could perfectly describe how lucky I feel everyday. I realized that naming all the things I am thankful for about you is, just like deliberating all possible future outcomes, computationally intractable. I also thought of how tacky writing this here might seem, but then I turn around and see you shaping cane with Kiwi on your lap and Gunther napping (I am sure Riku is munching hay somewhere), and then all I feel is that I want to remember how grateful I feel to have shared this part of my life with you. I love you. Also thank you to my in-laws, Barbara and Tom, for accepting me into your family and being there for both of us during our doctorates.

To all my friends: Thank you to Yingxin, Zhiting, and Yinru for being there through all of life's ups and downs. To my college roommates and best friends, Xavier, Nathan, and Seth: thank you for being the chillest friends and for still checking up on me. Thank you to Paulina; your friendship means a lot, and I was so happy you got to visit London while I am here. I am so proud of you of all that you're doing and what you have accomplished. To the night management group, Lisa, Carly, Lysette, and Janet: thank you for tolerating my rambling thoughts and multi-line messages. Special thanks to Janet, who tolerated not just horrible me but also my horrible writing. Also to Lisa,

thank you for doing that one day L.A. trip with me, I needed that and will cherish that forever. To my UPMC friends, Thandi, Laura, and Morgan: thank you for being the cohort I didn't have. To my Gamers friends, Christina, Gabe, Celine, and Matthew: thank you for the adventures, virtual and in irl. Thank you for letting me rant even thought we were fighting bosses. There were millions of people online and somehow we met and became friends, can't say it is not a little magical.

Abbreviations

AI : Artificial Intelligence

RL : Reinforcement Learning

LLM(s) : Large Language Model(s)

MEG : Magnetoencephalography

(f)MRI : (functional) Magnetic Resonance Imaging

EEG : Electroencephalography

Contents

Chapter 1

Introduction

A happy infant is eager to explore its world, while an angry banker may spend money to harm another's economic prospect. The affective state we are in shapes the allocation of attention, biases memory retrieval, and fundamentally guides behavior, sometimes even outside of conscious awareness[Frijda, 1986, Bower, 1981]. How we feel can alter what we say [Rude et al., 2004], how we interpret ambiguous social cues [Curby et al., 2012], and how we perceive the world [Zadra and Clore, 2011, Becker and Leinenger, 2011]. Affective states are so important because they are often honest signs of the innermost motivation driving our behavior. When sadness prevails without an appropriate cause or when anxiety paralyses us in the safety of our neighborhood, we cease to function.

Studying and understanding affective states is hence of great importance for psychology, neuroscience, psychiatry and brain sciences more broadly. It is also increasingly important for artificial intelligence [Chakriswaran et al., 2019, Zhao et al., 2022, Assunção et al., 2022]. However, affective states are ephemeral and richly complex entities or phenomena. Despite their importance and ubiquity, they are phenomenally difficult to study. There are great disputes around the very definitions of affective states. Numerous phenomenological labels emphasize slightly different aspects of affective states and even the choice of referring to affective states positions us in a very particular corner of these debates. Affective states are richly intertwined not only with observable behavior, but also with the hidden domains of subjectivity and consciousness. In fact, it is this latter aspect, the notion that emotional states are only really accessible through conscious report, that serves as the most demanding barrier in studying affective states.

This thesis grounds the affective state in a view that is rooted in computational neuroscience. From this perspective, we lay out the basis for a novel approach to the charac-

terization and assessment of the affective states through their computational structure and neurobiology. While this is an ambitious long-term goal, the work presented here represents an initial step. We argue affective states exist primarily because the brain must manage its limited computational resources [Huys and Renz, 2017], but cannot do so optimally. Specifically, in the setting of planning, the brain cannot consider all possible sequences of future actions for there are too many. Critically, choosing which of the sequences to evaluate is itself a decision problem, with greater computational demands than the original problem. This recursive curse, the so-called metareasoning problem, means that wise management of computational resources is itself so resource intensive that only approximations remain feasible. These approximations, we will assume, are at the center of affective states' functions.

We will now first review theories of emotion with a specific emphasis on this notion of the computational nature of affective states. A key aspect of this notion is that it may lend itself to a novel measurement approach, and hence we will focus also on different approaches to measuring and quantifying these important but elusive affective states.

1.1 Theories of Affective States

To better understand the challenges in assessing affective states, one must look at how they are defined. Terms like affect, mood and emotion are often used interchangeably, which can complicate efforts to measure them precisely. Most past efforts to distinguish affective states have been descriptive: separating affect, mood and emotion by their core characteristics, such as duration, intensity, or relation to a specific cause[Lange and James, 1922, Frijda, 1986, Ekman, 1999]. These theories have long sought to explain what affective states are, why they exist, and how they function. These diverse perspectives, from evolutionary biology to cognitive science, each provide a crucial piece of the puzzle that is the affective state.

One of the key consideration starts with the putative evolutionary advantage that must be conferred by emotions. Darwin argued that emotions are not arbitrary phenomena but are evolved adaptations that promote survival [Darwin, 1872]. Their primary function is to solve recurring, high-stakes situations by rapidly selecting advantageous actions. Fear mobilizes a flight response, while anger prepares for a fight. This notion of usefulness, that affective states serve a functional purpose, is a cornerstone of some, but not all, modern theories. This tight link between emotion and action was also central to early theorists like William James and Carl Lange, who proposed that the subjective experience of an emotion is the perception of the physiological changes

accompanying an action [James, 1948, Lange and James, 1922]. In this view, we do not run because we are afraid, we are afraid because we run. Indeed, relaxation methods such as meditation are thought to be helpful in stress management [Goyal et al., 2014]. While the strict causality of this model has been debated, later theorists like Frijda have reaffirmed the connection between action tendencies and affective states, viewing them as flexible states of readiness to engage with the world in particular ways [Frijda, 1986].

Other theories emphasize the social and communicative importance of affective states [Fridlund, 1994, Keltner and Haidt, 1999, Van Kleef et al., 2004]. Expressing an emotion can signal intentions, motivations, and needs to others, facilitating social coordination and bonding. This social functionality demonstrates that affective states operate as complex cognitive heuristics, not just simple action tendencies or urges. Emotions can also be expressed to strategically influence others [Fridlund, 1994]. This behavioral-ecological view posits that emotional displays are communicative tools tailored to a social audience. For instance, the state of anger deploys a social heuristic that goes beyond a simple urge to fight: its expression serves as an implicit warning to alter another's behavior, thereby planning a more favorable social outcome [Van Kleef et al., 2004]. Similarly, an expression of sadness is not just a passive signal of loss, but a planned heuristic for soliciting support and recruiting aid from one's social group [Keltner and Haidt, 1999]. Therefore, another key aspect of emotional states is its influence on planning and strategy.

For expressed emotions to be effective, they must be widely understood. Paul Ekman proposed a theory of "basic emotions," arguing for a limited set of discrete, universal states—such as fear, anger, and joy—each with a unique and genetically determined neural and physiological signature [Ekman, 1999, Ekman et al., 1987]. While this work provided compelling early evidence, it was later criticized on methodological grounds, such as cueing participants and restricting their choices [Russell, 1994]. More recent, data-driven approaches offer a nuanced view. For example, Cowen and Keltner identified 27 distinct categories of emotional experience from self-reports after participants viewed thousands of short videos [Cowen and Keltner, 2017]. These findings suggested that discreteness is a key aspect of emotional state, but also revealed that these emotional states exist along continuous gradients, connecting related feelings.

This seemingly paradoxical discreet-yet-continuous view may reflect the innate-but-nuanced (or innate-but-constructed) aspect of emotional states. The discrete nature of these categories likely reflects the finite set of distinct, recurring common problems we as human face. Just as fear evolved as a heuristic for threats and disgust for contamination, other emotional states likely emerged as solutions to other specific challenges

like forming alliances, caring for offspring, or navigating social hierarchies. There are not infinite core survival problems, and thus, there are not infinite basic universal emotional states. The continuous gradients between these categories, however, reflect the complexity of the modern world, where single events can simultaneously involve elements of multiple core challenges. Therefore, the emotional landscape is neither a handful of rigid, basic modules nor an infinite, undifferentiated sea of feelings, but rather a high-dimensional, structured space of related states.

A different perspective emerged from appraisal theories, which addressed a key limitation of stimulus-bound theories like Ekman's. Appraisal theories posit that emotions are not just innate reactions to events, but are the result of a highly cognitive process of evaluation, or appraisal, along dimensions like novelty, valence, and goal conduciveness [Frijda, 1986, Scherer, 1984]. This explains how the same stimulus can elicit vastly different emotions in different people, or in the same person at different times. It is the meaning of the event being interpreted in a certain context, not the event itself, that matters. Building on this cognitive view, constructionist theories argue that the subjective experience of emotion is itself a cognitive construction [Barrett, 2017a]. In this view, the brain makes an inference about what emotion it is in, based on incoming sensory information, physiological signals, past experience, and social context. The reported affective experience is therefore not a readout of a set of cached parameters, but a concept the brain applies to make sense of a particular pattern of internal and external information. When a person reports that they are anxious, it is because that they have inferred this based on the bodily signal, internal thoughts and various contexts. This aligns with the notion that some aspects of affective states may not be consciously accessible, further reinforcing the distinction between an underlying emotional state and the introspective process used to access it.

These diverse theories highlight key aspects of affective states: they are evolutionarily advantageous heuristics that impose action tendencies, facilitate strategic planning, possess a structured yet nuanced landscape, and are cognitively constructed from an appraisal of a situation's contextual meaning. If we synthesize these aspects, we can conceptualize affective states as a solution to the metareasoning problem. The evolutionary usefulness of affective states then is their function as a fast, efficient heuristic that solves the otherwise intractable problem of deciding under the pressure of finite cognitive resources. The cognitive appraisal is thereby recast as the cognitive process that assesses a situation's parameters to select the most appropriate metareasoning heuristic for that context. The resulting action tendencies and physiological changes are the direct results of this selected heuristic, which reconfigures the brain and body to execute resource-efficient strategies. Furthermore, the discrete-yet-continuous structure of

the affective landscape reflects a core set of evolved heuristics for recurring problems, with the gradients between them allowing for flexible adaptation to novel or complex situations. Consequently, the constructed affective experience is the brain's conscious inference about which of these computational strategies is currently deployed. This act of inference is not automatic but is an active cognitive process of introspection. When centered on cognitive appraisals and contextual meaning, it results in the report of an affective experience like emotions, but when centered on interoceptive cues like a racing heart, it yields a report more akin to generic "feelings". This distinction highlights the core challenge in affective neuroscience: our understanding of the underlying neurocomputational state is fundamentally constrained by our reliance on these subjective reports. Overcoming this barrier is critical, not only for building a comprehensive model of the brain but also for addressing the societal burden of mood disorders, which are characterized by dysregulated affective states [World Health Organization, 2022, GBD 2019 Mental Disorders Collaborators, 2022]. Therefore, the central goal of this work is to develop a method that moves beyond subjective reports to objectively quantify the high-dimensional affective state itself.

Language is a powerful tool for both expressing and inducing affective states. Intriguing stories, engaging conversations, and stirring speech all rely on the bilateral connection between affective states and language. Language is also a deeply computational process, involving building complex, predictive models of the world based on a rich understanding of how events and actions function [Heilbron et al., 2022, Caucheteux et al., 2023]. We now possess tools capable of capturing these computational processes: neuroimaging can measure the brain's internal states during language comprehension, while LLMs explicitly model the predictive computations that underlie it [Vaswani et al., 2023]. This convergence raises the central question of this thesis: if affective states are computational in nature, and language is a primary means of inducing them, can we leverage the tools used to study the computations of language to objectively measure and model the affective states themselves? In the following sections, we review existing methods on quantifying affective states, and introduce a body of work that combines neuroimaging, cognitive tasks, and LLMs to take a step toward this goal.

1.2 The Gold Standard: Subjective Reporting Affective Experience

It is true that no one knows better how a person feels other than the person themselves. So naturally, if one were to assess how someone is feeling, why not just ask? Indeed, the gold standard for assessing a person's affective experience is through subjective self-report. Through introspection and communication, self-reports directly capture the conscious, subjective affective experience, which is the brain's constructed inference about the affective state [Barrett, 2004, Cowen and Keltner, 2017, Robinson and Clore, 2002]. It is through self-report that researchers have mapped the structured yet nuanced landscape of emotion [Cowen and Keltner, 2017, Russell et al., 1989]. Subjective report is embedded in decades of affective neuroscience research, including investigations into the structure of affect [Russell et al., 1989], the influence of emotion on decision-making [Isen, 2001, Lerner et al., 2004], the mechanisms of emotion regulation [Gross, 1998, Ochsner et al., 2002], the impact of affect on memory [Bower, 1981, Talarico and Rubin, 2003], and the computational mechanistic investigations of subjective feelings [Rutledge et al., 2014, Blain and Rutledge, 2020]. This method is also the backbone of validated clinical instruments such as the Beck Depression Inventory [Beck et al., 2011], the Positive and Negative Affect Schedule [Watson et al., 1988], and the cornerstone of psychiatric diagnosis, the DSM-5-TR [American Psychiatric Association, 2022].

Despite its foundational role, the validity of self-report is constrained by limitations that arise from its nature as a complex constructive process rather than a simple readout of the internal affective state. The first limitations is its cognitive demand. Using fMRI and contrasting different aspects of self-reporting, Satpute and colleagues [Satpute et al., 2013] argued that the process of self-reporting an affective experience involves three systems: first, the dorsomedial prefrontal cortex was involved in directing attention toward affective responses and their attributions; second, the activities in ventrolateral prefrontal cortex was related to verbal labeling; third, ventral anterior insula and amygdala were linked to reporting of intensity. These regions have been found to be associated with emotion regulation [Ochsner and Gross, 2005] and self-referential processing [Northoff et al., 2006]. Further, ERP studies showed that the presentation of emotionally salient stimuli elicited early negativity in posterior relating to attention and late positive linked to evaluation[Schupp et al., 2006]. While the studies provided insights on brain systems associated with self-report of affective experience, they are constrained by a fundamental confound. By design, these experiments simultaneously

elicit an affective state and engage the similar cognitive components required to report it. This makes it difficult to disentangle the neural signature of the affective state itself from the neural activity of reporting affective experience. Additionally, engagement of this complex cognitive process can be considered very effortful (our experimental examination on this is reported in Chapter 2). Consequently, the reliability and validity of this process thus is constrained and influenced by available cognitive resources, such as time[Otto and Daw, 2019] and memory capacity [Sandra and Otto, 2018]. Further, because of the need for attention and activate engagement, this process is not only cognitively demanding, but also interruptive of the task at hand. Then an individual's ability to accurately and consistently report their affective experience is heavily influenced by one's ability to task switch, which will contribute to inter-personal variance and reduce internal consistency. Both of the aforementioned cognitive factors are exacerbated when self-report are administered frequently. In this case, there's additional burden of task-switching, potentially introduce more intra- and inter-individual variance.

Second, the process of self-report is susceptible to cognitive biases. Because introspection is considered effortful, individuals may default to less cognitively demanding heuristics, such as applying common or easily accessible labels even when their affective state is complex or ambiguous [Kool et al., 2010].

Third, social pressures influence what individuals are willing to report. In clinical and research settings that cover sensitive topics, the desire for favorable impression can lead to the under-reporting of socially undesirable experiences [Sherman et al., 1975, Walsh, 1969, King et al., 2018, Hart et al., 2019]. This is particularly detrimental in clinical practice, where stigma may cause patients to conceal diagnostically crucial information, such as suicidal ideation, thereby increasing risk of suicide [Cai et al., 2021]. This is further complicated by individual differences in the ability to differentiate and label emotions, as seen pathologically in alexithymia [Taylor et al., 2024].

Finally, and perhaps most fundamentally, the self-report process itself is subject to measurement reactivity, which describes how the measurements can alter the state being measured. Introspecting upon and labeling an emotion is not an affectively neutral process. For example, Lieberman and colleagues found that the act of labeling affective experience was associated with increased activity in the right ventrolateral prefrontal cortex which, in turn, correlated with dampened amygdala activity [Lieberman et al., 2007]. Because the pathway is identified to be associated with emotion regulation, this suggests that the process of reporting is neurally intertwined with emotion regulation. Contradicting the assumption that because the administra-

tion of self-reports is cognitively demanding, it could result in patient burden, recent meta analysis showed preliminary intervention utility of self-reporting internal affective states in psychiatry, such as affect labeling and EMA [Torre and Lieberman, 2018, Bell et al., 2017]. Regardless of the valence, self-report can induce change in affective experience. That is the cognitive processes involved in self-report of affective experience nevertheless changed the very state they meant to measure. This confound challenges the validity of the quantification of affective states via self-report.

Taken together, these limitations, the cognitive demands, susceptibility to bias, and inherent reactivity, can severely undermine the results of research using them and create barriers for clinical practice. While acknowledging the utility of self-report, there is a pressing need for a less cognitive demanding, more passive and direct way to objectively assess or approximate internal affective states.

1.3 Objective Methods for Measuring Affective States

The limitations of self-report have motivated a long line of research in affective science: the search for objectively quantifiable correlates of emotional experience. The main idea is because affective experiences are not only subjective phenomena but also cognitive processes that impose widespread influence in behaviors, physiology and neural activities, therefore, by measuring signals relating to these objective changes, an indirect assessment of the states might be possible.

1.3.1 Behavioral Correlates

Human behavior is a rich source of information from which one can infer affective states. Early work focused on facial expressions, which can be viewed as the behavioral output of heuristics for strategic social planning [Ekman et al., 1987]. However, this approach is limited by a many-to-many mapping, as context heavily influences the meaning of an expression [Barrett et al., 2019]. Mechanistic approaches offer another path by using computational modeling to decompose decision-making. Frameworks like Reinforcement Learning can phenotype maladaptive behaviors in dysregulated affective states by showing how computational parameters, such as learning rates or decision temperature, are systematically altered. In doing so, these models provide a powerful method for quantifying how an affective state functions as a heuristic for strategic planning, systematically altering the parameters of decision-making [Huys et al., 2015, Pike and Robinson, 2022].

Ekman and colleagues argued for a set of discrete and universal recognizable facial expressions that can be mapped onto different emotional categories [Ekman et al., 1987, Ekman and Friesen, 1971]. In a 1971 experiment, Ekman and Friesen showed tribal participants in New Guinea with limited exposure to modern culture associate the same emotional concepts to the same facial behaviors as those who lived in both western and eastern societies, indicating a universal association between emotional representation and facial expressions. Further, recent advancement with machine learning vision models demonstrated impressive ability to analyze complex facial expression in real-time and capture dynamic features rapidly to recognize affective states beyond basic emotions [Arora et al., 2024]. That is, these data-driven models do more than simple classification. Instead, they utilize the high-dimensionality to quantify the intensity of an expression and the durations as well, providing a much richer estimation of the quantified affective states.

However, using facial expressions as an objective proxy for affective states is limited by a fundamental many-to-many mapping: a single expression can relate to numerous states, and vice versa. This ambiguity is a critical drawback. Given that affective states are not just simple reflexes but also highly nuanced, context-dependent heuristics for navigating the complexities of the world. A measurement tool that lacks specificity fails to capture the rich computational and cognitive differences between nuanced affective states that might share a similar expressive appearance. This issue can be framed as a need for contextual awareness, a facial expression can mean different affective states in different situations or paired with different body language. Indeed, Barrett and colleagues showed evidence for a variable and many-to-many relationship between a facial expression and underlying affective states [Barrett et al., 2019]. For example, a scowl is not a just a sign for anger but it can signify confusion, concentration, or even just a reaction to bright light. Conversely, an angry person appears to be a neutral based on facial expression because people might be in certain affective state but without making any discernible expression. Further, individuals who are on the Autism Spectrum might exhibit atypical or attenuated facial expression [Song and Hakoda, 2018]. In their cases, an over-reliance on facial expression to assess affective states could lead to flawed or erroneous conclusions.

Mechanistic approaches that use computational modeling provide a powerful method for phenotyping individuals in different persistent affective states. Frameworks like Bayesian Decision Theory elegantly reframe core symptoms of depression, not as simple lethargy, but as systematic changes to the parameters of decision-making. Pessimistic priors or a decrease in vigor can be rational if the rate of future reward is estimated to be low [Huys et al., 2015]. Empirical findings support this view, show-

ing that clinically depressed individuals have altered reward and punishment learning rates [Pike and Robinson, 2022]. By capturing these stable, trait-like differences in how individuals learn and make choices, differentiation of people in different persistent affective state can be done. However, this strength in identifying persistent traits introduces a rigidity. The method is optimized to detect stable patterns and is less suited for identifying temporary, dynamic states. An individual experiencing transient sadness after a loss might immediately exhibit decision patterns that temporarily mimic those of clinical depression. This creates a risk of mischaracterizing. This limitation underscores the need for methods that can track the dynamics of affective states on a much faster timescale.

1.3.2 Physiological Correlates

Affective states induce widespread changes in the body through the autonomic nervous system. Physiological signals aim to capture these bodily signal, which correspond directly to the arousal component of the action tendencies proposed by early evolutionary theories of affective state. This arousal reflects the body's state of energy mobilization and readiness to engage in adaptive behaviors. The methods reviewed below provide a non-invasive and continuous objective index of subjective arousal, approximating affective state.

Electrodermal activity (EDA), or galvanic skin conductance, tracks fluctuations in skin conductance, which is directly related to sweat gland activity innervated by the sympathetic nervous system (a branch of autonomic nervous system) [Boucsein, 2012]. This system is the primary pathway for mobilizing the body for action, making EDA a sensitive index of the arousal that underpins action tendencies. EDA has two components: the tonic Skin Conductance Level, representing general arousal, and the phasic Skin Conductance Response, capturing event-related changes [Dawson et al., 2007]. Empirical evidence shows both components increase during affective experiences, particularly stress and emotional face processing (for Skin Conductance Level: [Fernández et al., 2012, Wang et al., 2018]; for Skin Conductance Response: [Christopoulos et al., 2016, Banks et al., 2012, Khalfa et al., 2002, Matejka et al., 2013, Nava et al., 2016]). However, these measures are not exclusively affective. They are also elicited by any novel or task-relevant event, complicating their interpretation as a pure measure of an emotional state's action-oriented arousal [Dawson et al., 2007].

Cardiovascular measures such as heart rate (HR) and heart rate variability (HRV) offer another non-invasive window into the autonomic nervous system's role in preparing

the body for action. These metrics reflect the interplay between the sympathetic (mobilizing) and parasympathetic (rest-and-digest) branches. While HR provides a general index of arousal [Lang et al., 1993, Brosschot and Thayer, 2003], HRV is considered a more nuanced marker of self-regulatory capacity and the flexibility to adapt an action tendency to the situation [Lane et al., 2009, Zhu et al., 2019]. Lower HRV, indicating reduced parasympathetic control, is linked to negative affective states and deficits in emotion regulation, suggesting a body stuck in a state of high-arousal readiness [Zhu et al., 2019, Appelhans and Luecken, 2006, Thayer and Lane, 2000]. However, like EDA, these measures suffer from a critical lack of specificity. An accelerated heart rate signals a high-energy action tendency, but it cannot distinguish the fight tendency of anger from the flight tendency of fear or the approach tendency of joy.

Eye-tracking approaches assess visual attention, which is a key component of action preparation. The core idea is that an affective state's action tendency will systematically bias where and how long we look. Empirically, individuals with anxiety disorders show an increased attentional bias towards threatening stimuli, which can be interpreted as part of a readiness to respond to potential danger [Clauss et al., 2022]. Similarly, individuals with depression often exhibit sustained attention on negative information [Duque and Vázquez, 2015]. Pupil dilation, another eye-tracking metric, reflects autonomic activation and serves as a reliable correlate of arousal, similar to EDA [Bradley et al., 2008]. Therefore, these measures suffers from the same drawback: a change in pupil size signals an increase in the intensity of an action tendency, but it could also reflect a non-affective change in cognitive effort or attention.

The fundamental challenge with these peripheral physiological measures is a problem of mismatched dimensionality. Affective states are high-dimensional constructs, while these methods capture low-dimensional correlates, such as the arousal component of feelings. Attempting to map a complex psychological state onto a simple signal inevitably results in a massive loss of information. This explains the critical lack of specificity; distinct emotional states like fear and excitement are collapsed onto the same point on the single dimension of arousal. Even when multiple physiological signals are combined, they often reflect activity within the same underlying autonomic nervous system. This system also showed correlation to non-affective processes, such as effort exertion. Therefore, it is difficult to isolate variance unique to an emotional state. Because affective experiences are neural states that originate in the brain, peripheral physiological signals are therefore indirect, downstream reflections of this central activity. To more fully capture the high-dimensional neurophysiological state, one must turn to the source: the brain.

1.3.3 Neural Correlates

Neuroimaging techniques like functional magnetic resonance imaging (fMRI) and magneto/electroencephalography (M/EEG) offer the non-invasive means to measuring the cognitively constructed affective state at its source. The thousands of data points from voxels or sensors provide a high-dimensional signal space that is a much better match for the complex nature of these states than low-dimensional physiological signals. This approach aims for a granular characterization of the affective state, moving beyond simple measures of arousal.

Early neuroimaging studies used a univariate approach, analyzing each brain region independently to find the neural basis for the discrete categories of emotion proposed by theories of basic emotions. This search for one-to-one mappings yielded influential findings, such as implicating the amygdala in fear [LeDoux, 2002, Ousdal et al., 2008, LaBar et al., 1998] and the insula in disgust [Phillips et al., 1997]. Meta-analyses further linked specific regions to distinct emotions, such as the subcallosal cingulate to sadness [Phan et al., 2002]. The promise of this localized view was attractive: if a consistent neural signature for an emotion could be found, quantifying a person's affective state could be as simple as measuring activity in a specific brain region.

This simple mapping, however, was challenged by growing evidence of functional heterogeneity. For example, evidence suggest that the amygdala also responds robustly to positive and novel stimuli, leading to a revised view of it as a more general hub for detecting biological salience or relevance [Sander et al., 2003, Adolphs, 2008, Ousdal et al., 2012, Cunningham and Brosch, 2012]. The definitive shift in perspective came from large-scale meta-analyses that reported inconsistent neural correlates of emotion, such that there is no evidence that any single brain region is exclusively dedicated to processing a specific emotion [Kober et al., 2008, Lindquist et al., 2012, Kragel and LaBar, 2016]. This body of neuroimaging evidence echoes the Psychological Constructionist Theory on the distributed neural representation of affective states. This view proposes that emotions, or more broadly affective states, are not biologically innate kinds but are constructed in the moment from the interaction of more basic, domain-general brain networks [Barrett, 2017b]. According to this framework, the brain does not have a dedicated "sadness circuit", rather, the brain constructs an instance of sadness by recruiting core cognitive networks in a context-dependent manner.

The shift from a localized to a distributed and constructed view of emotion necessitated a more sophisticated analytical approach. Studies now employ a decoding approach when attempting to quantify affective experience with neural signal through

Multivariate Pattern Analysis [Haxby, 2012, Norman et al., 2006]. This method leverages information contained in the distributed patterns of activity across the entire brain, treating each pattern as a high-dimensional data point. The core promise of the decoding approach is to move beyond simply localized brain activity and begin to make full use of the high-dimensional neuroimaging data to match the equally high-dimensional affective experiences. Combining this approach with fMRI's strength in spatial resolution, studies showed viability of decoding self-induced emotional states [Kassam et al., 2013], and picture induced affective states [Baucom et al., 2012]. With EEG and its high temporal resolution, decoding of the temporal dynamic of affective experiences is possible. Studies demonstrated successful EEG classifiers on decoding movie-induced positive emotion labels [Du et al., 2023, Liu et al., 2018].

While these advanced decoding methods have shown that brain activity carries rich information about a person's affective state, one key limitation remains. The constructive nature of affective states means they are deeply intertwined with the cognitive context in which they are elicited. For example, the recorded neural pattern of sadness elicited by listening to a story consists of not only the neural representation of sadness, it is a representation of sadness in the context of language and auditory networks engagement. Similarly, recording the neural signal of fear induced by a threatening image will actually result in neural correlates of the state of fear while engaging visual and attentional networks. The high specificity of neuroimaging becomes a double-edged sword: while it captures rich detail, the resulting neural decoders often learn the features of the elicitation task as much as the features of the abstract affective state. This raises the question, can a neural decoder trained to distinguish happiness from fear using static pictures of faces be useful in identifying affective state elicited by listening to a story or recalling a personal memory? The constructive nature of affective states implies a great degree of contextual dependency, that is the neural pattern of sadness might differ when it is induced by a sad story versus a surprise lost in gambling. This context-specificity means that a decoder trained in one paradigm typically might fail to generalize to another, presenting a major barrier to identifying a truly abstract, context-independent neural signature of emotion. Experimental studies in the literature failed to test whether their successful decoders can also generalized across context, a feat that no assessment methods other than self-report can achieve.

1.4 Integrating large language models, neuroimaging, and cognitive task

The preceding review of measurement methods reveals a central dilemma in affective science: a misalignment between the high-dimensional affective state and available assessment tools. While self-report may be indispensable for capturing the subjective and introspective aspects of an affective state, it has important inherent limitations: the cognitive load [Ochsner and Gross, 2005, Northoff et al., 2006, Schupp et al., 2006, Satpute et al., 2013], susceptibility to bias [Kool et al., 2010], dependence on willingness and ability to report [Sherman et al., 1975, Walsh, 1969, King et al., 2018, Hart et al., 2019, Cai et al., 2021, Taylor et al., 2024] and measurement reactivity [Lieberman et al., 2007, Torre and Lieberman, 2018]. These limitations preclude its use as a continuous, non-intrusive tool for affective state assessment. Objective measures from the body and brain, while promising, have been plagued by a fundamental lack of specificity[Du et al., 2023, Liu et al., 2018, Baucom et al., 2012]. Therefore, it is still unknown if they truly captured the neural affective states that can generalize across different contexts. The ideal solution, therefore, is to develop a new form of measurement, one which captures affective states and can be validated by its ability to predict self-reported affective experience.

The desiderata for an ideal objective assessment of affective states could look as follows. First, it must be fast, with the capacity to track the dynamics of emotional states on a rapid timescale. While periodic questionnaires might be appropriate for mood state assessments, the event-elicited emotional states can change within seconds [Kragel et al., 2022]. For example, a surprise reveal in a story might drastically alter the listener's emotional state quickly [Gagne and Dayan, 2023]. Therefore, this temporal precision is necessary to capture the dynamics of the event-driven emotional states, which cannot be adequately assessed with periodic self-report. Second, the measure must be passive and non-interruptive. It should operate in the background, allowing for the continuous assessment of affective states during complex, ongoing cognitive activities without consuming valuable cognitive resources or altering affective states in the process. Finally, and most critically, the affective states determined by the objective assessment must be highly correlated with self-reported ones. To be considered a valid proxy for internal experience, it must demonstrate that the objective signal it captures is meaningfully and reliably correlated with the "ground truth" of a person's reported feelings.

Below, we propose a method that combines the strengths of LLMs, cognitive tasks and

neuroimaging in an effort to take a step toward an objective assessment of affective states that is fast, passive, and predictive of self-reported states. LLMs represent a major technological advancement in recent years. These transformer models are trained on enormous natural language corpora [Vaswani et al., 2023, Touvron et al., 2023]. They are not only able to generate text but also to perform a range of sophisticated tasks by utilizing the many complex and meaningful latent structures that exist in language [Vaswani et al., 2023, Kumar et al., 2024, Lopopolo et al., 2024]. Furthermore, when guided, LLMs can generate text that captures some features of text produced by individuals with depression, suggesting that LLMs can mimic affective states representations in the semantic space [Onysk and Huys, 2025, Low et al., 2020]. These emergent abilities suggest that LLMs are not merely engineering tools for text manipulation, but may also serve as powerful computational models for human cognition.

The core of this thesis builds on the notion of emotional states as metareasoning strategies, which helps address the problem of optimally allocating finite cognitive resources.[Russell and Wefald, 1991, Ackerman and Thompson, 2017, Huys and Renz, 2017] In a world where computations are free, a rational agent would allocate all of its available resources, such as attention, memory capacity, or time, to solve any problem at hand. That means engaging in the most deliberative decision strategies to maximize desirable outcomes, such as exhaustive tree search [Kaindl, 1990], where the agent will simulate all outcomes of all possible actions in a state. However, for humans, this maximizing solution is computationally intractable [Huys and Renz, 2017]. There are simply too many possibilities to evaluate in real life, extensive opportunity costs, and very limited cognitive resources available. In fact, given the limitations on capacity and time, deliberative strategies are often sub-optimal because their computational cost is too high. Our brain must therefore rely on approximations or decision heuristics to take resource efficient actions. Given emotional states' vast influence on cognitive processes and their ability to integrate past experiences, they could act as neurocomputational heuristics that solve the resource allocation problem by rapidly reconfiguring cognitive priorities in response to an appraisal of the situation. The feeling of anxiety, for instance, is not just a subjective state but a rational strategy aimed at mitigating uncertainty [Grupe and Nitschke, 2013]. One might enter this state after experiencing consecutive unforeseeable negative events. Under the anxiety strategy, attention is allocated to sample and plan for potential unfavorable outcomes, the likelihood of negative outcomes might be perceived as higher, and the threshold for taking risky action is increased. This metareasoning perspective does not discard existing emotion theories but rather reframes them in computational terms. For instance, the process of cognitive appraisal can be seen as the

rapid evaluation that determines which metareasoning strategy is the appropriate one to deploy. The core set of Basic Emotions can be understood as evolutionarily ancient, highly effective strategies for solving recurring and critical survival-related resource allocation problems, such as those involving threat, loss, or reward.

The adaptive value of these emotional strategies is grounded in a fundamental statistical property of the natural world: temporal consistency [Diener and Larsen, 1984]. Consecutive events are not independent. Rather, the recent past is often the best predictor of what is to come in the immediate future. An emotional state can act as the mechanism to capture and leverage this environmental regularity by imposing a stable cognitive decision strategy over time. It represent the past through active appraisals of the recent past, weighted by recency and significance, to bias behavior and provide action tendencies that are appropriate for the current context. This same principle is precisely what allows predictive models like LLMs to excel. The statistical structure of language, where past words provide powerful context for future ones, mirrors the temporal structure of events in the world. A powerful computational parallel to this process exists at the core of the LLMs' underlying modeling structure: the attention mechanism [Vaswani et al., 2023], which enables LLMs to achieve long-distance conceptual consistency. When predicting the next word, the self-attention mechanism allows the model to dynamically weigh the importance of all prior words in the context, no matter how far back they appeared. This creates an internal representation that is exquisitely sensitive to context and maintains a coherent thread of meaning over long passages.

The temporal consistency that an affective state imposes on human cognition may be statistically related to the conceptual consistency that the self-attention mechanism imposes on language generation in LLMs. The sensitivity of human affect to external stimuli, such as emotional responses to movies, is analogous to how a transformer's internal state is dynamically shaped by its input context. Both are fundamentally systems for integrating the past to make sense of the present and guide predictions about the future. This computational analogy allows us to construct a framework for objective affective measurement, one designed specifically to address the critical challenge of generalization. The central hypothesis is that the long-range contextual dependencies that LLMs captures during the processing of narratives can serve as a high-dimensional, abstract, and context-sensitive proxy for the human affective state. The ultimate test of this framework will be to determine if this objective model of the state can reliably predict the self-reported affective experience, even across different contexts.

While this framework can in principle be applied to all affective states, the work in

this thesis will focus specifically on emotional states. Mood, conceptualized as tracking long-range average of recent outcomes [Eldar et al., 2016], might be misaligned with LLMs in terms of timescale and ineffectively elicited by the event-driven structure of cognitive tasks. Core affect, while fundamental, is an intentionally low-dimensional construct defined by valence and arousal. Attempting to map the high-dimensional representations from a large language model onto such a simple space would severely underutilize the LLMs' strengths. Emotional states, in contrast, are both event-oriented and high-dimensional. Their granular and context-specific nature provides a perfect match for the nuanced, high-dimensional semantic space that LLMs are designed to capture, making them the ideal target for this investigation.

To elicit these rich emotional states, we employed cognitive tasks and naturalistic stories. The constructive, appraisal-based view of emotion suggests that emotional states are not simple reactions, but are built from the evaluation of complex, evolving events in relation to one's goals. Simple, static stimuli are insufficient to evoke the granular emotional states, such as remorse or admiration. Rich, event-driven tasks by contrast, mimic the structure of real-world scenarios. They present a continuous stream of events that engage the appraisal and constructive processes, making them the ideal method for eliciting emotional dynamics. Naturalistic stories are similar to cognitive tasks such that they demand listeners' attention and predictive processes [Ryskin and Nieuwland, 2023], which are in alignment with what LLM does during stories. Therefore, these two methods of inductions are perfect for this. Naturalistic stories engage the same predictive processing mechanisms in the listener that are fundamental to the attention mechanisms of LLMs [Heilbron et al., 2022, Caucheteux et al., 2023]. Joint use of these two emotion eliciting methods creates a tight alignment between the elicitation method and the proposed framework.

This integrated framework is more than just a novel method for quantification, it can also provide insight on fundamental questions in affective neuroscience. The high-dimensional nature of the LLM-derived affective states label provides a quantitative testbed for theories that have been difficult to probe. For example, it allows for a formal investigation of emotion granularity. We can ask whether fine-grained states, like nostalgia and admiration, occupy distinct neural representation across contexts. This thesis work takes a step toward bridging the gap between subjective experience and objective measurements, opening up new avenues for understanding both healthy and disordered affective processing.

1.5 Thesis Overview

This dissertation is structured as a multi-stage investigation to develop the proposed computational linguistic framework and rigorously test its neural validity and generalizability.

Chapter 2 showed two online studies that use self-report to investigate how subjective happiness is influenced by cognitive effort, task difficulty, cognitive effort efficacy and choice. This work establishes a baseline affect-cognition link but also suggests that self-reported happiness in these contexts may reflect a narrow, meta-reasoning mental state tied to task performance rather than a generalized affective experience.

Chapter 3 demonstrated the development of a novel quantification pipeline that uses LLMs and a Hidden Markov Model to transform raw text of narratives into a continuous, word-by-word time series of multi-dimensional emotional state labels.

Chapter 4 establishes the emotional state labels have within label neural consistencies and that they correspond to distinct patterns of MEG activity.

Chapter 5 describes the design and validation of a battery of non-linguistic cognitive tasks, involving math, prisoner's dilemma, gambling, and navigation. This chapter shows that these tasks reliably elicit a wide range of self-reported emotional responses, paving ground for the stringent generalization tests in the final chapter.

Chapter 6 presents the empirical test of the dissertation's main hypotheses. Using data from participants who listened to stories and performed the cognitive tasks, MEG neural decoders were trained on the story data using the LLM-quantified labels. We then conduct a test of within-modality generalization on held-out stories and a test of cross-modal generalization on cognitive task data. The aim of this final stage is to empirically examine whether this framework can identify a truly abstract, context-independent neural signature of emotional states, or if, as constructionist theories predict, the neural representation of affect remains anchored to its eliciting context.

Finally, chapter 7 discusses the findings, suggesting that they provide valuable insight into the neural representation of emotional states and important implications for future work on their objective quantification.

Chapter 2

Cognitive effort has a negative impact on subjective happiness

2.1 Introduction

Whether we plan our wedding, think through a problem or engage in fun activities, our brain has to perform complex computational operations. Such neural computations incur a variety of costs. While devoting our neural machinery to one problem, we forego opportunities that may have been reaped by focusing on alternatives. Furthermore, neural computations incur hard energy costs. Whether to engage in neural computation is hence a complex problem, one that requires potential gains to be traded off against potential costs [Russell and Wefald, 1991]. Indeed, humans appear to be sensitive to cognitive effort costs, and often avoid cognitive effort both implicitly (e.g. [Huys et al., 2012, Callaway et al., 2022]) and explicitly (e.g. [Vogel et al., 2020, Inzlicht et al., 2018, Botvinick et al., 2009, Westbrook et al., 2013]). In fact, individuals can show preference for pain over a task demanding a high level of cognitive effort [Vogel et al., 2020]. This strong level of avoidance suggests that expending cognitive effort, or the prospect of doing so, is both implicitly and explicitly unpleasant and punishing in some form [Aw et al., 2011, Mintz, 2010, Grèzes et al., 2021, Ellis et al., 2022, Het and Wolf, 2007].

However, this simplistic view of cognitive effort as computationally costly and aversive belies a more complex picture. First, the experienced feeling of expending cognitive effort does not necessarily reflect true energetic costs. The amount of calories consumed during a cognitive task is negligible [Kurzban, 2010], and measures of brain blood glucose do not change with extensive cognitive effort [Madsen et al., 1995]. The

experienced feeling of cognitive effort rather appears to reflect a subjective and biased assessment of the value of resource investment [Shenhav et al., 2017].

Secondly, studies on work productivity and sports training have also shown that people seek out and appear to enjoy effortful activities, ranging from mountain climbing to a variety of intellectual pursuits [Wang et al., 2017, Inzlicht et al., 2018, Kurniawan et al., 2013, Shenhav et al., 2017]. Here, the expenditure of cognitive effort, and the prospect of doing so, are associated with a positive affective experience. Although, in part, this may arise when the immediate costs of cognitively effortful behaviors are offset by the promise of longer-term or alternative rewards. For instance, during the effortful process of planning, current expenditure of cognitive effort is associated with expected gain in future reward [Ho et al., 2022, Ferstl et al., 2022, Daw et al., 2011]. In social situations, individuals might opt to engage in cognitively effortful pro-social behavior to pursue better social outlook or others' reciprocation [Batson, 1987, Godman et al., 2014, Vaish et al., 2016]. Indeed, planning tendencies reduce when without longer-term benefits [Kool et al., 2016, Callaway et al., 2022], and individuals reduce their pro-social behavior with increasing cognitive effort demands [Lockwood et al., 2017].

Lastly, there are strong suggestions of a bidirectional relationship between mood and effort. On one hand, mood clearly affects willingness to exert cognitive and physical effort. In the extreme, psychopathological states such as mania feature increased willingness to expend cognitive and physical effort.[American Psychiatric Association, 2013]. Similarly, low mood and pathological states such as depression reduce the willingness to exert cognitive and physical effort [Anderson et al., 2019, Ganesan, 2020]. Indeed, fatigue and difficulties in concentration are diagnostic symptoms of depression, and both speak to a reduced propensity to expend effort [American Psychiatric Association, 2013]. Outside of psychopathology, studies have shown that experimentally induced mood impacts cognitive effort expenditure as well [Treadway et al., 2009, Joana, 2009, Brinkmann and Gendolla, 2008, Gendolla and Krüsken, 2002]. On the other hand, while mood states are responsive to appetitive and aversive events [Philip, 1971, Diener et al., 2009, Grosscup and Lewinsohn, 1980, Stone and Neale, 1984], it is unknown how cognitive effort influences mood at the moment of expenditure. There is strong evidence for the influence of appetitive events on mood, and current evidence suggests that momentary mood reflects the statistics of recent rewards [Keren et al., 2021, Liuzzi et al., 2021]. Considering that cognitive effort is both aversive and closely linked to reward, one could speculate that cognitive effort itself should have an impact on mood.

Furthermore, there is a question as to what the direction of this effect might be. As mentioned above, both a positive (mood-enhancing) direction, and a negative (mood-lowering) direction are conceivable. To our knowledge, this has not been examined. While previous studies examined how cognitive effort changes affective experience [Ferstl et al., 2022, Robinson and Morsella, 2014, Erber and Tesser, 1992, Larsen et al., 1986, Larsen and Berenbaum, 2014, Sonnentag and Grant, 2012], there are no parametric manipulations of cognitive effort while assessing momentary mood in an experimental setting.

Here, we therefore examined the impact of parametrically varied cognitive effort demand on the momentary affective experience at a fine temporal scale. Specifically, we were interested in how changes in cognitive effort exertion over time relate to temporal fluctuations in affective states.

To examine the impact of cognitive effort, we combined this momentary mood assessment with a letter sorting paradigm commonly used for the study of the neurobiology of cognitive effort [Jansma et al., 2006]. This task allowed a parametric manipulation of cognitive effort by varying short-term memory load. The combination of these two tasks allowed us to study how variations in cognitive effort impact on momentary affective states.

2.2 Methods

2.2.1 Ethics approval

The experimental protocol was approved by the National Institutes of Health Office of Human Subjects Research Protection (Protocol Number: P194594).

2.2.2 Tasks

Two cognitively demanding tasks were used to investigate the effects of cognitive effort, as indexed by difficulty, on momentary mood. In both tasks, we asked participants to perform letter sorting tasks with varying difficulty and report their trial-by-trial mood ratings. A large body of literature also demonstrated the influence of reward processes on affective experience [Rutledge et al., 2017, Rutledge et al., 2014, Keren et al., 2021, O'Callaghan and Stringaris, 2019]. Reward process is also closely linked to the motivation to expend cognitive effort [Frömer et al., 2021, Yang et al., 2014, Massar et al., 2018]. Therefore, we also included

an explicit manipulation on reward magnitude as a control condition to account for possible reward effects. In both tasks, participants were asked to respond with keyboard and provide happiness rating with mouse.

Multi-Attempt Letter Task

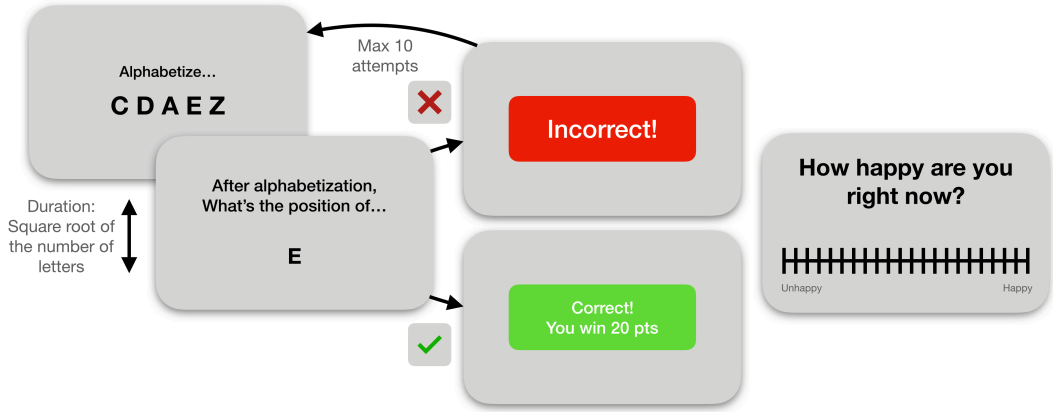


Figure 2.1: The Multi-attempt Letter Task. In each trial, participants viewed a string of letters of length n for $\sqrt{n}$ seconds, such that participants had more time to encode the letters in more difficult conditions. Participants then had to mentally order these letters alphabetically and were asked to respond about the position of one specific randomly chosen letter. For all conditions, participants had 3 seconds to make a response. The difficulty of each trial was manipulated by the length of the string presented on that trial, number of letters (n; ranging from 3 to 9) . Feedback was provided after each response. If the response was incorrect, participants were returned to the same problem for up to a maximum of 10 attempts. After 10 failed attempts, the trial would end without bonus points being awarded. If the response was correct, participants saw the bonus points earned. The magnitude of the bonus points ranged from 30 to 90 in 10 increments. If no responses were provided, the trial is counted as missed and participants would proceed as if it was incorrect. At the end of each trial, participants were asked to self-report their momentary mood level using a visual-analog scale ranging from 0 to 10 (0 labeled unhappy and 10 labeled happy), as done in previous work [Rutledge et al., 2014, Keren et al., 2021, Liuzzi et al., 2021, Rutledge et al., 2015]. The visual-analog scale reseted each trial and provided no starting point and registers input as participants click on one of the ticks. Participants had 5 seconds to provide their momentary happiness ratings.

An overview of the Multi-attempt Letter Task is shown in Fig. 2.1. There were 168 trials, each with a unique strings. The difficulty and reward magnitude were fully randomized and temporally orthogonal to each other, such that the participants experienced the full range of reward magnitude for each level of difficulty.

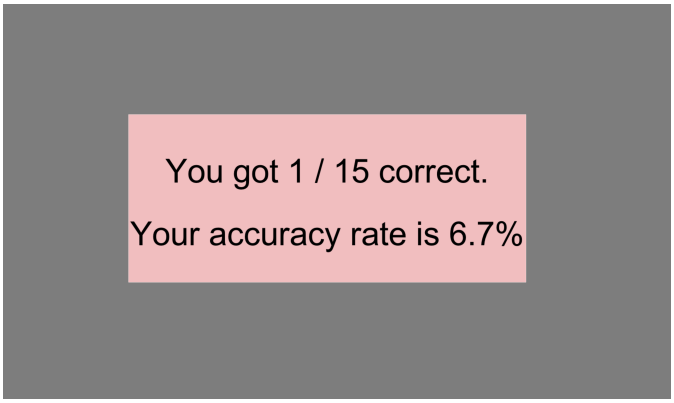


Figure 2.2: An example of the feedback on participant performance in the Single-Attempt Letter Task.

Single-Attempt Letter Task

The Single-Attempt Letter Task was identical to the Multi-Attempt Letter Task except that, as the name suggests, participants were allowed only one attempt for each trial. In addition, the task also only provided sparse performance feedback. Feedbacks were only provided randomly on a total of 7 trials, with a between feedback interval ranging from 15 and 45 trials in 5 increments. Feedback consisted of both percentage and counts of correct trials (Fig. 2.2). To ensure identical overall duration between the Single-Attempt and the Multi-Attempt Letter Tasks, the number of trials was increased to 210. The Single-Attempt Letter Task was preregistered prior to data collection and analysis (osf link).

2.2.3 Participants and Procedure

Participants were recruited through online recruitment platforms Amazon Mechanical Turk (MTurk) [Amazon Inc, 2023] and Prolific [Prolific Inc, 2023] following standard online recruitment protocols [Aguinis et al., 2021, Mortensen and Hughes, 2018].

Before their participation, recruited individuals were presented with informed consent on a web page, where they must fully read and agree before continuing. Because both MTurk and Prolific mask personally identifiable information, including email, and that we only collected behavioral responses and no clinical assessments, our studies were determined to be exempt from institutional review board review by the National Institutes of Health (NIH) Office of Human Subjects Research Protections. The protocol on the consent processes and collection of behavioral responses was approved by the NIH

Office of Human Subjects Research Protections.

At the beginning of the study, participants were first instructed on the task design and that their bonus earning would be correlated with their task performance. Participants were reimbursed for their time via respective recruitment platforms, along with a performance-based bonus payment (later assigned). For the Multi-Attempt Letter Task, 119 participants were recruited and 109 completed the full experiment. For the Single-Attempt Letter Task, we recruited 360 participants who had not participated in the Multi-Attempt Letter Task previously. 330 completed the full experiment.

2.2.4 Inclusion and Exclusion

We excluded participants based on three criteria:

1. **Percentage of missing data:** a trial was considered missing if no response was provided within the time limit. In the case of the Multi-Attempt Letter Task, where multiple attempts were allowed, a trial was considered missing if all 10 attempts had no responses. A percentage was calculated based on the number of trials missed and participants who missed more than 20% were excluded.

2. **Number of consecutive identical key-presses:** the number of consecutive key-presses was defined as the number of repetition of the same key. A high number of repeated key-presses was indicative of a lack of engagement. Thus, we excluded participants who had consecutive identical key presses on more than 10% of the total trials.

3. **Probability of the average number of errors per trial / accuracy count:** for participants who were actively engaged, they should perform better than chance in this task. To determine if a participant was performing better than chance, we implemented a simulation-based method to generate synthetic behaviors as if responses were made randomly. Based on the probability density function of generated data, we calculated metrics that a participant needed to exceed to be considered statistically unlikely that they were providing answers randomly (significance level = 0.05). Because each task has different design elements, we employed different metrics for each tasks. For the Multi-Attempt Letter Task, the metric was the average number of errors per trial, and its threshold was 5.23. For the Single-Attempt Letter Task, the metric was total accuracy count, and the threshold was 49.

Using these criteria, we included 107 participants in the Multi-Attempt Letter Task sample and 206 participants in the Single-Attempt Letter Task sample. For detail break down of the count of excluded participants based on criteria, see Table. 2.1. The seemingly high count of exclusion in the Single-Attempt Letter Task is partially due to that some participants had lower than acceptable frame rates. An update in the online experiment hosting platform consisted a bug that reduces frame rate for certain machines. 100 out of the 124 excluded participants had lower than 10 frames per second, which could potentially has effect on participants' performance. In comparison, only 1 participants in the whole Multi-Attempt Letter Task sample, which was not affected by this bug, had lower than 10 frames per seconds.

	Completed	Missing	Repeat	Performance	% excluded
Multi-Attempt Letter Task	109	0	1	1	2%
Single-Attempt Letter Task	330	51	4	69	37%

Table 2.1: Breakdown of the excluded participants based on criteria. "Missing" represent criteria 1: percentage of missed responses; "Repeat" for criteria 2: Number of consecutive identical key-presses; and "Performance" for criteria 3: Probability of the average number of errors per trial / accuracy count. If a participant matched multiple criteria, they would be only counted once.

2.2.5 Data Analysis

We calculated Pearson correlation to investigate the association between difficulty (n) and number of errors (e, number of repeats before a correct response). We calculated the correlation estimate, test statistics and probability using the cor.test function in the R (version 4.1.3) stat package [R Core Team, 2020]. To examine to what extent trial-by-trial mood ratings m could be predicted by preceding trial events we built linear mixed-effects regression models using the lme4 [Bates et al., 2015] package from R [R Core Team, 2020]. We built four linear mixed effects models as follows. The full model explained the mood $m_{t,i}$ reported by subject i on trial t as:

$$m_{t,i} = (\alpha_0 + \beta_{0,i}) + (\alpha_n + \beta_{n,i})n_{t,i} + (\alpha_e + \beta_{e,i})e_{t,i} + (\alpha_r + \beta_{r,i})r_{t,i} + (\alpha_\tau + \beta_{\tau,i})t + \epsilon_{t,i} \quad (2.1)$$

where $n_{t,i}$ is the i'th subject's difficulty level on trial t, $e_{t,i}$ indicates whether the trial was an error (1 = error, 0 = no error), $r_{t,i}$ indicates reward bonus obtained on that trial. Fixed effects are denoted by α and random effects varying for each subject by β. Three

component models were examined, containing either difficulty, error or reward terms in addition to the intercept (α_0, β_0) and time passage effects (α_τ, β_τ), such as the mood drift over time effect [Jangraw et al., 2023]. In each model, mood ratings were the dependent variable, and task conditions were independent variables.

To examine the effects of task events in the Single-Attempt Letter Task, we constructed another similar model:

$$m_{t,i} = (\alpha_0 + \beta_{0,i}) + (\alpha_n + \beta_{n,i})n_{t,i} + (\alpha_\lambda + \beta_{\lambda,i})\lambda_{t,i} + (\alpha_f + \beta_{f,i})f_{t,i} + (\alpha_\tau + \beta_{\tau,i})t + \epsilon_{t,i} \quad (2.2)$$

where the term $f_{t,i}$ indicates the magnitude of performance feedback (in percentage), $\lambda_{t,i}$ indicates the number of trials passed since last feedback was displayed. We included the $f^i_{t,i}$ and $\lambda_{t,i}$ terms to capture the effect of feedback as well as its temporal influence on momentary mood.

Because we did not include measure of cognitive effort that's uncorrelated with difficulty, it is possible that any task manipulation effects on mood we observed could potentially due to only difficulty and not cognitive effort. In attempt to disentangle this issue, we employ the following explorative analysis. First, we assume the following:

$$P_{Correctness} = f(\text{difficulty, cognitive effort exertion, time-related factors, noise}) \quad (2.3)$$

Under this assumption, the probability of someone correctly perform the letter sorting task is a function of how difficult the trial is, how much cognitive effort they have exerted, how much one learn, how tired they are, and other factors that are time-related. Then, we constructed a correctness predicting logistic regression model with number of letters, square-root of number of letters, trial, trial interact with number of letters as predictors. The residual of such model could be a noisy representation of the difficulty independent cognitive effort exertion. We further include this correctness residual in our mood predicting model (as mentioned above):

$$m_{t,i} = (\alpha_0 + \beta_{0,i}) + ... + (\alpha_\gamma + \beta_{\gamma,i})\gamma_{t,i} + \epsilon_{t,i} \quad (2.4)$$

where the $\gamma_{t,i}$ represents the correctness residual for a given trial and participant. We only applied this model to trials where participants responded correctly as the incorrect trial will violate our assumption, i.e. we do not know if the residual would represent (the lack of) cognitive effort expenditure for incorrect trials.

All models included trial number as a predictor to capture time-related mood changes. Tests for linearity are shown in the supplemental results.

The fixed effect size d was estimated as:

$$d = \frac{\Delta\mu}{\sigma^2_{\text{rand}} + \sigma^2_{\text{res}}} \tag{2.5}$$

where $\Delta\mu$ is the difference between the means, σ^s_{rand} is the random effects and σ^2_{res} the residual variance.

2.3 Results

2.3.1 Baseline Mood

To ensure that our samples did not contain a disproportionately high number of depressed or elated individuals, we first calculated the summary statistics of the baseline momentary mood rating collected prior to the task, and conducted two tailed t tests on whether the true mean of the ratings is different than average (ratings = 5). Additionally, we also conducted a non-paired two sample t-test on the rating difference between the two sample. We found that our participants had reported significantly higher than average momentary mood ratings and that there is no significant difference in starting mood between the two samples. The descriptive and t-test statistics are summarized in the table below:

Sample	N	1st Qu.	Median	Mean	3rd Qu.	t	D.F.	P
LT	107	5.040	7.070	6.904	8.890	8.574	104	< 0.001
LTSF	210	6.100	7.980	7.472	9.020	11.966	209	< 0.001
Difference						0.175	206	0.861

Table 2.2: Summary statistics and one tailed t test results for the Multi-Attempt Letter Task and the Single-Attempt Letter Task samples. The summary statistics include the 1st quartile (1st Qu.), median, mean, and 3rd quartile (3rd Qu.).

Therefore, we conclude that our two samples were equivalent in terms of baseline mood. Although, it does suggest that, in both of our samples, participants on average have elevated mood to begin with.

2.3.2 Multi-Attempt Letter task

We first performed a manipulation check and examined whether longer strings were indeed more difficult and led to more errors. Figure 2.3B shows that a higher number

of letters n indeed led to a higher number of errors e (correlation estimate = 0.225, 95% confidence interval (C.I.) = [0.211, 0.239], $t_{17617} = 30.69$, p < 0.001), suggesting that longer strings were more difficult for participants to order alphabetically. This is also confirmed by a mixed linear effects model, using n, t and their interaction to predict e. We found that, controlling for t, higher n leads to more errors (estimate = 0.977, std. error = 0.049, $t_{106} = 19.366$, p < 0.001, effect size = 0.746). We did not find a significant interaction between n and t (estimate = 0.025, std. error = 0.064, $t_{185} = 0.395$, p = 0.693, effect size = 0.019)).

Next, we examined which aspects of the trials influenced momentary mood ratings. The component models showed that number of letters (n), errors (e), and reward (r) were individually associated with the mood ratings (Table 2.3, figure 2.3A). Importantly, while reward had a positive effect, both difficulty and errors had a negative effect. The effect of time was always negative, such that with increased in trial number, self-reported trial-by-trial mood rating decreased, which is consistent with previous studies showing a mood drift effect in task [Jangraw et al., 2023].

However, number of letters (n), error (e) and reward term (r) were partially correlated, in part by design. To control for this, we examined the full model containing all effects (Eqn. 2.1).

This model showed that while the error and the reward terms remained significant positive and negative predictors of momentary mood, respectively, the number of letters term was no longer significant (Table 2.3).

Model	$R^2_{marginal}$	$R^2_{conditional}$	Effect	Estimate	t-value	p-value	Effect size
Difficulty	0.023	0.827	n	-0.300	-5.823	<0.001	0.082
			t	-0.894	-4.815	<0.001	0.244
Error	0.040	0.859	e	-0.784	-8.705	<0.001	0.207
			t	-0.970	-5.313	<0.001	0.256
Reward	0.027	0.840	r	0.491	6.193	<0.001	0.135
			t	-0.917	-4.965	<0.001	0.253
Full	0.043	0.873	n	0.028	0.662	0.509	0.007
			e	-0.758	-8.538	<0.001	0.203
			r	0.417	5.679	<0.001	0.111
			t	-0.974	-5.337	<0.001	0.260

Table 2.3: Linear mixed effect models predicting trial-by-trial mood ratings. n denotes the difficulty or number of letters; e indicates the number of errors; r represents the reward magnitude; t is the trial number. The degree of freedom for all the independent variables in the models is 104.

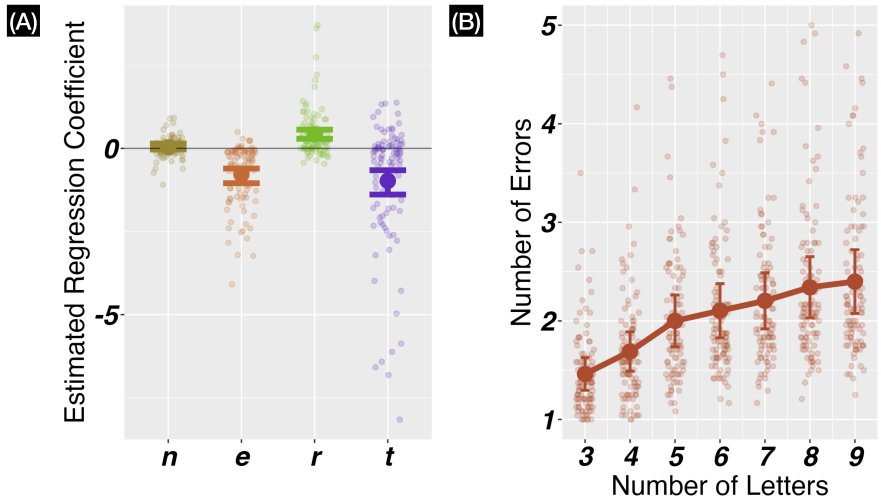


Figure 2.3: Multi-Attempt Letter Task Results. A): The estimated fixed effects of number of letters (n), number of errors (e), reward magnitude (r), and trial (t) on trial-by-trial mood rating. Each data point represents the estimated random effects varying for each subject. Error bar illustrates the standard error of the mean. B): Number of error (e) for difference levels of number of letters (n). Longer strings were more difficult. The number of errors increased with the length of the letter strings. Dots show the mean number of errors for each participant for a particular number of letters.

2.3.3 Single-Attempt Letter Task

In the first version of the task, the effect of cognitive effort appeared to be accounted for by the more frequent experience of accuracy negative feedback. However, this may have been a statistical artifact due to issues of collinearity, such that participants both made increased number of errors and experience higher degrees of cognitive effort in more difficult trials . To test this, we designed a new version of the letter task with sparser feedback that allowed a better distinction between the negative impact of cognitive effort and negative accuracy feedback on mood.

Participants' accuracy rate in the sparse feedback version was again lower in trials with more letters (higher n) (Figure 2.4B). This was confirmed by a mixed effects logistic regression model, showing that the effects of number of letters on trial-by-trial accuracy is significant (estimate = -0.484, std. error = 0.028, z = -17.145, $p < 0.001$, effect size = 0.484), controlling for trial number (t). We also identified a significant negative interaction between number of letters and trial number, such that participants' accuracy were lower later in the task (estimate = -0.063, std. error = 0.013, z = -4.957, $p < 0.001$, effect size = 0.063).

To investigate if participants' accuracy is impacted by the joint effect of time-in-task and difficulty, we included an interaction between trial number and number of letters. The mixed effects logistic regression model also revealed a significant negative interaction (estimate = -0.061, std. error = 0.013, z = -4.831, p<0.001, effect size = 0.061), such that trial-by-trial accuracy is even lower in later and more difficult trials. These effects persist even after adjusting for chance probability of each difficulty level (Figure 2.4C).

With the removal of trial-by-trial feedback in this task, we found that difficulty (n) was a significant negative predictor of momentary mood, (Table 2.4, Figure 2.4A). We also examined the effect of the sparse feedbacks (f) on momentary mood. Mood ratings were lower after feedback regardless of its valence. It is important to note that 64.63% of the feedback showed below average accuracy rate. However, participants' mood slowly recovered from the negative impact of feedback afterwards, as indicated by the significant positive effect of the term number of trial since last feedback (λ).

Model	$R^2_{marginal}$	$R^2_{conditional}$	Effect	Estimate	t-value	p-value	Effect size
Full	0.005	0.875	n	-0.168	-7.826	<0.001	0.166
			f	-0.174	-2.449	0.015	0.172
			λ	0.028	2.510	0.013	0.027
			t	-0.146	-3.645	<0.001	0.144

Table 2.4: Linear mixed effect models predicting trial-by-trial mood ratings. n denotes the difficulty or number of letters; f indicates whether performance feedback was displayed; λ represents the number of trials elapsed since last feedback was displayed; t is the trial number, which we use to infer the mood drift over time effect and time in task effect. The degree of freedom for all the independent variables in the models is 209.

To disentangle the effects of task difficulty and cognitive effort on mood, we implemented a logistic regression model predicting correctness. Our model showed that indeed all predictors were statistically significant (Table 2.5).

We extracted the residual from the correctness model and included as another independent variable in predicting happiness rating in correct trials only. We showed that the correctness residual, which is a noisy representation of task difficulty independent cognitive effort expenditure, still showed negative impact on mood.

2.4 Discussion

We examined the relationship between cognitive effort and momentary mood. In two tasks, a parametric increase in the exertion of cognitive effort reduced momentary

Model	$R^2_{marginal}$	$R^2_{conditional}$	Effect	Estimate	z-value	p-value
Correctness	0.079	0.251	n	-8.800	-13.168	<0.001
			n^2	5.723	-12.941	<0.001
			t	5.919	7.956	<0.001
			$n:t$	-5.351	-5.123	<0.001

Table 2.5: Linear mixed effect models predicting trial-by-trial mood ratings. n denotes the difficulty or number of letters; n^2 indicates the square root of n, included to capture non-linear effects of difficulty; t is the trial number, which we use to infer the mood drift over time effect and time in task effect. the $n : t$ interaction was included to capture potential learning effect

Model	$R^2_{marginal}$	$R^2_{conditional}$	Effect	Estimate	t-value	p-value
Full+Residual	0.011	0.908	n	-0.211	-2.045	0.042
			f	1.182	3.229	0.001
			γ	-0.743	-3.411	<0.001
			λ	1.070	3.020	0.003
			$f:\lambda$	-1.727	-2.485	0.014
			t	-3.802	-5.310	<0.001

Table 2.6: Linear mixed effect models predicting trial-by-trial happiness ratings. This model only applied on trials where participants had answered correctly. n denotes the difficulty or number of letters; f indicates the magnitude of performance feedback; γ is the residual extracted from the correctness model (see Table 2.5); λ represents the number of trials elapsed since last feedback was displayed; $f : \lambda$ interaction was included to capture the feedback magnitude effect that progressed with time; t is the trial number, which we use to infer the mood drift over time effect and time in task effect. The degree of freedom for all the independent variables in the models is 209.

mood. In the first study, increased cognitive effort led to an increase in errors, and subsequent repeat of the same trial. Therefore, in this task, effort was partially confounded by other factors such as duration, negative feedback and possibly fatigue. These confounds were removed in the second study allowing a direct estimate of the impact of cognitive effort on momentary mood. This confirmed that effort exertion has an overall negative effect on momentary affective states on short time scales.

A negative effect of cognitive effort on mood is consistent with both the view of cognitive effort as computational cost, and of momentary mood ratings as running estimates of reward rates [Rutledge et al., 2014, Bennett et al., 2021]. In terms of the former, we have noted previous failures to show an obvious energetic correlate of cognitive effort expenditure [Madsen et al., 1995, Kurzban, 2010]. As such, the negative affective consequence of cognitive effort could be due to opportunity costs. Opportunity costs arise when the execution of one activity means that the rewards for other activities cannot be earned. Indeed, when comparing the two studies, the impact of cognitive effort on mood in the first study was particularly marked when subjects were forced to repeat the same task after errors. While this result is obviously confounded by the negative feedback itself, the fact that errors in this task cost participants additional time for each attempt support the idea that opportunity costs contribute to the negative affective response to cognitive effort. In the absence of a clear energetic cost for cognitive effort, the opportunity cost itself is then likely to derive from the limitations on cognitive capacity [Sandra and Otto, 2018, Inzlicht et al., 2018]. The existence of a opportunity cost on cognitively effortful actions is necessary for the effective allocation of cognitive resources. Otherwise, prioritization of cognitive resources for one task would not mean that alternative tasks cannot be processed.

As for the latter, the question is why should momentary mood or affective judgements reflect the opportunity costs of cognitive effort, i.e., why should exerting cognitive effort change how we feel? Numerous studies over the past decade have shown that momentary mood ratings quantitatively reflect recent and history of rewards and losses [Emanuel and Eldar, 2023, Eldar et al., 2018, Rutledge et al., 2014, Rutledge et al., 2015, Rutledge et al., 2017, Keren et al., 2021, Liuzzi et al., 2021]. Specifically, momentary mood is predicted by not only reward history but also an average of recent positive and negative prediction errors, i.e. whether current experience exceeded expectations or failed to meet them [Rutledge et al., 2014, Keren et al., 2021, Liuzzi et al., 2021]. Monitoring this rate of reward has been argued to be useful for inferring underlying environmental changes [de Boer et al., 2017, Packheiser et al., 2021, Kumar et al., 2018] and to facilitate future behavior [Schultz et al., 1997, Farrell et al., 2022, Rouhani and Niv, 2021, Gläscher et al., 2010]. This is because the average re-

ward rate is a measure of the ongoing opportunity cost: an environment with high opportunity cost is one with high reward rate, i.e., one in which acting slowly causes many rewards to be missed [Niv et al., 2007]. Thus, this motivates high vigor and fast action. A similar argument can be made for cognitive effort [Brinkmann and Gendolla, 2008, Treadway et al., 2009, Treadway et al., 2012]. In this case, low effort—or slow computations—would lead to a major reduction in reward, hence, motivating high cognitive effort expenditure.

In our study, we assumed a linear relationship between cognitive effort and task difficulty, and we did not directly measure the sensation of cognitive effort. There is a possibility that mapping between cognitive effort and difficulty is not linear. First, while difficulty creates demand for cognitive effort, it is possible that various factors, such as individual ability, disengagement, and fatigue, could all lead to low cognitive effort exertion in the face of high difficulty. However, we demonstrated that reaction time, which is often used as a approximation of cognitive effort exertion [Robinson and Morsella, 2014, Ganesan, 2020, Robles and Johnson, 2017], is indeed modulated by task difficulty (see supplemental material: Reaction Time Analysis).

The study has some limitations. It relies on a single subjective mood rating to measure momentary mood, and we cannot exclude that other aspects of momentary mood could be affected differently. However, mood ratings to appear to index a general state of well-being [Rutledge et al., 2014, Keren et al., 2021, Liuzzi et al., 2021, Eldar et al., 2016], suggesting that the findings may generalize to a certain extent. Additionally, initial mood ratings in both studies showed an above average score. Although, we showed that initial mood has no impact on the effects of cognitive effort on mood.

Our study also has strengths. One of the main strengths is the large sample size based on detailed power analyses; c.f. supplemental material). Such large samples have become more common with online cognitive testing [Gillan and Daw, 2016]. Importantly, the demographics of online recruitment platforms are similar to the general population [McCredie and Morey, 2019, Huff and Tingley, 2015, Redmiles et al., 2019], with the main discrepancies in terms of lower affect and social engagement [McCredie and Morey, 2019, Shapiro et al., 2013]. The design involving two studies allowed us to replicate our main findings, and we note that the second study (the Single-Attempt Letter Task) was pre-registered. Importantly, the second study also showed that the effects remain when controlling for (or removing) explicit rewards and losses.

In conclusion, across two studies, cognitive effort induced by task difficulty resulted

in reduced self-report moment-to-moment happiness ratings. Viewed from a formal setting of reinforcement learning, the impact of cognitive effort on mood is in keeping with the putatively normative roles of both effort and mood, providing insight into potential mechanism for cognitive resource allocation.

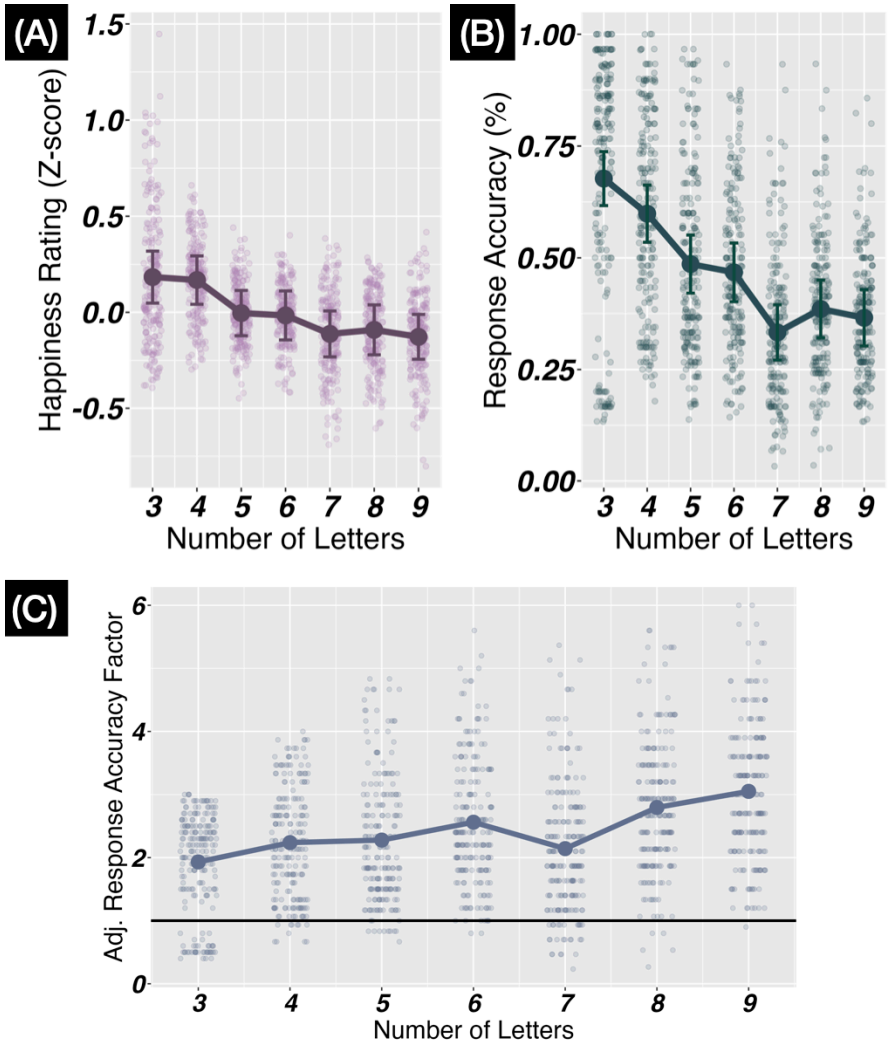


Figure 2.4: A): The estimated fixed effects of number of letters (n) on trial-by-trial mood rating (z-scored). Each data point represents the estimated random effects varying for each subject. Error bar illustrates the standard error of the mean. B): Responses to longer strings were less accurate. Accuracy decreases with the length of the letter strings. Dots indicate the mean accuracy rate given participant and number of letters. C): Participants accuracy are all above chance across all difficulty conditions. On the y axis is the adjusted response accuracy factor which was calculated as the accuracy rate divided by chance. Chance is different for each number of letter. As the number of letter increases, the chance probability will decrease, it can be calculated as $1/n$.

Chapter 3

Quantifying emotion time-course with large language models

3.1 Introduction

Language is the primary medium for emotional expression. We use language to communicate our current emotional states and the prime example of this is the subjective self-report, the (potentially problematic) cornerstone of affective neuroscience and clinical practice [Robinson and Clore, 2002, Cowen and Keltner, 2017, American Psychiatric Association, 2022]. Self-report is the process by which an individual introspects on their internal feelings and translates the results of that inference into a communicable, linguistic form. This can range from selecting a label on a rating scale to providing a detailed verbal description during a clinical interview. Emotion induction is as such heavily rooted in and shaped by the use of language. Naturalistic stories and vignettes can build a rich context over time, reliably inducing nuanced emotional states [Zupan and Babbage, 2017]. If emotional states consistently shape the language we produce, and are influenced by the language we are exposed to, then language constitutes an inherently valid and informative domain for investigating emotional states. This validity is supported by a large body of empirical work demonstrating that distinct affective states, such as depression or anxiety, are associated with reliable and quantifiable differences in linguistic patterns, from word choice to grammatical structure [Low et al., 2020, Seabrook et al., 2018].

3.1.1 Emotions as Metareasoning Heuristics

In the introductory chapter, we introduced the notion of metareasoning. Metareasoning concerns the problem of optimally allocating finite cognitive resources, such as attention and deliberation time [Ackerman and Thompson, 2017]. For humans, engaging in exhaustive, deliberative strategies to find a truly optimal solution for every decision is computationally intractable because there are simply too many possibilities to evaluate in a complex world [Huys and Renz, 2017]; and optimally assigning resources in this situation is yet more demanding and hence less feasible. The brain must therefore rely on approximations and heuristics to guide not only behavior, but also the internal allocation of resources, efficiently.

Our working hypothesis is that affective states are heuristics to address this computational problem [Russell and Wefald, 1991]. In this view, an affective state is a computational state in which resources are prioritized in a concerted manner, linking evaluations of different components into a coherent whole akin to previous suggestions [Lazarus, 1991, Scherer, 1984]. However, our suggestion is that the selection of a specific emotional heuristic arises from an approximate inference process about the potential value of a discrete, relatively small, set of metareasoning heuristics. If one metareasoning heuristic appears valuable, then it is engaged and thereby focuses resources on a small part of the much broader and vast decision and metareasoning problem. The emotional state is then a state in which the assignment of resources to a subset of behaviors and evaluations is favored. In this framework, emotional states act as approximate neurocomputational strategies by representing past context and influencing future information processing and behavior. Broadly, this suggestion is in keeping with evidence demonstrating that emotional states have a widespread influence, systematically biasing a host of cognitive domains including perception [Siegel et al., 2018b], social judgment [Niedenthal and Showers, 1991], memory retrieval [Dehon et al., 2010], and decision-making strategies [Seymour and Dolan, 2008].

The adaptive value of such metareasoning strategies is grounded in a fundamental statistical property of the natural world: temporal consistency. A state at one moment is often the best predictor of the subsequent ones, such that consecutive events in a time series are not statistically independent but instead highly correlated [Diener and Larsen, 1984]. Therefore, rather than re-evaluating every moment from scratch, the brain can capitalize on this autocorrelation by engaging in recognition of a general context, linking it to an appropriate subset of actions, and representing it as a state that persists in time. This persistent state then in turn influences future behavior, ensuring behavioral consistency over time and alignment with similar contexts. This

temporally persistent, context-representing state that ensures behavioral consistency is precisely the functional role that emotional states are thought to fulfill. Then emotional states are not just subjective feelings, but rather a computationally efficient mechanism that represents the current context, persists over time to leverage the temporal consistency, and shapes cognition and behavior accordingly. Indeed, when a surprise breaks this temporal consistency, it signals a change in context and possibly the need to engage a new emotional state [Eldar et al., 2018].

3.1.2 Quantifying Emotional States with Large Language Models

The specific starting point for the current argument is that the temporal consistency also extends to language. We say different things when we are feeling happy, sad, or frustrated, such that our language is consistent with our emotional state [Jackson et al., 2022, Cohn et al., 2004, Hutto and Gilbert, 2014]. For example, people use different language when they are depressed [Wang et al., 2013, Raffaelli et al., 2021, Seabrook et al., 2018] or anxious [Low et al., 2020]. This emotional state-congruent language use reflects the broader function of an emotional state as a heuristic that shapes and constrains all cognitive processes, including language production. We can rephrase this by saying that the underlying emotional state dictates the probability distribution over what people would express in words. This, of course, is consistent with the proposed role of emotional state.

Our question, then, is whether affective states, which are detectable in language, might be related to, and informative about, the underlying metareasoning state introduced above. Given the conceptual link between affective states and language, LLMs might provide a path to the objective quantification of affective states through language. LLMs have demonstrated a striking capability not only to consistently generate natural language [Colombatto and Fleming, 2024, Dillion et al., 2023, van Duijn et al., 2023, Hagendorff et al., 2024] but also to engage in tasks beyond language, such as playing video games [Wang et al., 2023] or arguing over philosophical constructs [Ye et al., 2024, Asghari and Bialy, 2025, Ashwani et al., 2024]. One of the most defining features of LLMs is the attention-based transformer architecture, which allows for rich representations of long contexts that in turn constrain the probability of the next words [Brown et al., 2020, Vaswani et al., 2023]. This functional parallelism raises a question and an opportunity for objective assessment of emotional states: can the long-range contextual correlations, as captured by an LLM, be used to generate a meaningful and objective quantification of a human emotional state? While recent work has shown that LLMs can successfully capture affective features from language for classification

tasks [Giachanou and Crestani, 2016, Demszky et al., 2020, Farruque et al., 2024], it remains unknown whether their output is sufficiently nuanced and valid to serve as a proxy for human emotional state.

3.1.3 Current Study

If LLMs are to be used for objective assessment of emotional states, then LLM-derived emotional state classifications (henceforth 'labels') must exhibit some basic properties. The key criterion we focused on first is that the labels must demonstrate semantic consistency, meaning they must be meaningfully grounded in the unfolding narrative content. Recent work has demonstrated the importance of this forward-looking "inner sentiment," showing that the predictive trajectory of the next word captures how conjunctions and intensifiers alter meaning within a sentence's course [Gagne and Dayan, 2023]. To ensure our labels were sensitive to this dynamic, we based our classification not just on the text itself, but on forward-looking predictions, or continuations, generated from an LLM.

This specific approach was chosen because the emotional state of a listener is often defined not by the words they have just heard, but by the prediction of what might come next; and because this conceptualization links word choice with metareasoning. For example, the emotional experience of suspense in a narrative is partially dependent on predictive inference[Lehne et al., 2015]. By having the LLM generate a distribution of likely continuations for every moment in the story, we capture this predictive content. This means our subsequent classification step is not labeling the semantics of the past, but rather the features of the predicted future. We hypothesize that this method provides a more faithful proxy for the listener's emotional or metareasoning state.

An alternative approach would involve applying the emotion classifier directly to the raw story text preceding each word. However, it is challenging to know what the right window might be. Depending on the architecture used, it might introduce an arbitrary hyperparameter due to the classifier's limited input window (for example, 512 tokens for BERT models). For longer contexts, one must decide how much preceding text to truncate, potentially discarding relevant long-range information. Our continuation-based method elegantly avoids this arbitrary decision. The initial large-context LLM processes the full available narrative history and distills the relevant predictive information into short continuation segments. The emotion classifier then operates on these context-rich, yet brief, continuations, effectively leveraging long-range dependencies without requiring arbitrary truncation.

Another key criterion is that the labels should also be structurally and temporally consistent. Empirically, emotions co-occur in meaningful clusters and persist over time. We assessed this by examining the correlation structure among the emotion probabilities and then used a Hidden Markov Model (HMM) to explicitly model the temporal persistence of states.

3.2 Methods

3.2.1 Naturalistic Stories

To elicit a wide range of emotions, eighteen spoken-word stories featuring personal storytelling were selected. These stories originated from a public radio program and were previously described in a published dataset[LeBel et al., 2023]. Stories were chosen for their emotional evocativeness and thematic diversity (see Table 3.1 for summaries, overall emotion valence, duration, and thematic categorization).

3.2.2 Generating Emotional State Labels from Story Text with LLM

To derive a word-by-word time series of emotional states from the narrative text, we implemented a three-step pipeline involving continuation, classification and extraction of temporally coherent states.

Step 1: Generating Predictive Continuations.

First, we used a pre-trained LLM to generate possible continuation sentences at each word in the story. Specifically, we generated 100 continuations for each word in the original story text. For each word in the text, all of the preceding text was used as the prompt. For the pre-trained LLM, we used the LLaMA2 model with 7 billion parameters [Touvron et al., 2023]. We implemented this with the text-generation pipeline from the transformers library in Python [Huggingface, 2025]. We configured the model to generate 100 continuations with a maximum of 50 new tokens in each, stopping earlier if an end-of-sequence token was produced. To balance performance and computational cost, we utilized a 6-bit quantized version of the model. Inference was performed on a system with an NVIDIA Quadro RTX 4000 GPU, accelerated with CUDA [Nickolls et al., 2008]. A grahpical example of this process is shown in Figure 3.1.

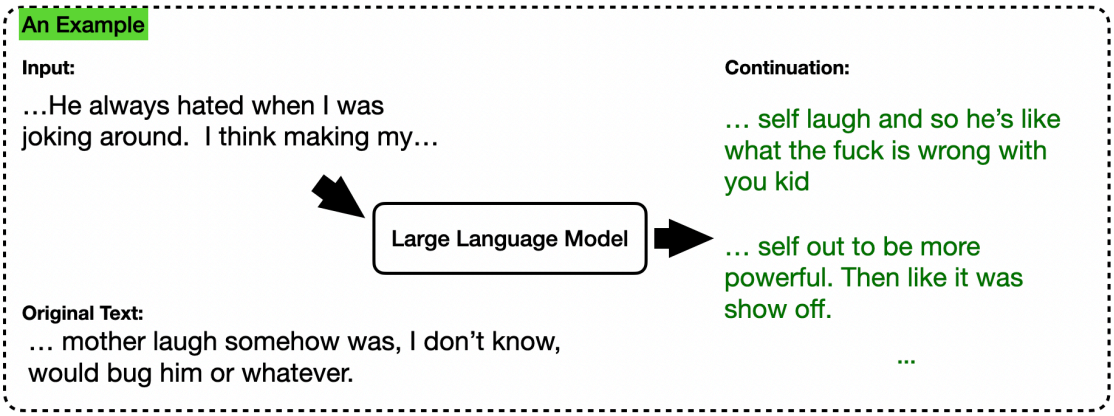


Figure 3.1: **Example of the continuation process.** The story text up to a given word (e.g., "...making my") was used as a prompt. A large language model (Llama 2 7B) then generated 100 continuations based on this prompt.

Step 2: Emotion Classification.

Each of the continuations was then passed into a fine-tuned emotion classification model to produce a 28-dimensional emotion probability vector for each word. By averaging across the different continuations, we established a probability distribution of likely emotion states at each point in time. Specifically, we used a DistilBERT model [Sanh et al., 2020] that was fine-tuned on the GoEmotions dataset [Demszky et al., 2020].

The GoEmotions dataset contains 58,009 Reddit comments annotated for 27 emotion categories plus a neutral category (emotion categories: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise). Using the text-classification pipeline from the transformers library [Huggingface, 2025], we passed each of the 100 continuations through the fine-tuned model and produced 100 separate 28-dimensional probability vectors.

Step 3: Temporal Smoothing and State Discretization

Empirically, some of the emotion labels are often highly correlated. Indeed, while the stories were chosen to capture a wide range of emotional states, it is unlikely that a (relatively short) set of stories would comprehensively and selectively elicit all 27 di-

mensions of emotions. We hence examined the dimensionality of the LLM-generated emotion space.

We employed a HMM to model these dynamics. This approach was chosen based on that we were a) interested in relatively discrete and distinct emotions; and b) reasoned that emotional states would be relatively temporally consistent [Bagnara et al., 2025]. HMMs capture both of these assumptions.

To implement this, the sequence of 28-dimensional emotion probability vectors served as the observed data for the HMM. We trained separate HMMs with the number of hidden states varying from 1 to 28. To select the optimal number of states for each story, we computed the Bayesian Information Criterion (BIC) for each model and chose the model that provided the best fit to the data. We used the GaussianHMM implementation from the hmmlearn library in Python, configured with a full covariance matrix[hmmlearn, 2025]. This configuration allows the emission probability for each hidden state to be modeled by a multivariate Gaussian distribution, which is critical for capturing the observed correlations between different emotions within a single latent state (e.g., a state that represents a mix of ”anger” and ”annoyance”).

3.2.3 Comparing Semantic Representation of Continuations

To validate that the LLM-generated continuations were semantically meaningful and predictive of the actual narrative, we analyzed their semantic similarity. We used the all-MiniLM-L6-v2 sentence transformer model [Reimers and Gurevych, 2019] to encode all continuations and segments of the original story text into sentence embeddings. The semantic similarity between any two text segments was calculated as the cosine similarity between their respective embedding vectors. We performed four comparisons:

- *Within*: The average pairwise similarity among the 100 continuations generated from the same prompt (at $word_t$). This measures the continuations' coherence.
- *Original*: The similarity between continuations (generated at $word_t$) and the actual 50-token segment that followed in the story ($word_{t+1}$ to $word_{t+51}$). This measures predictive accuracy.
- *Next 50*: The similarity between each continuation (generated at $word_t$) and the story segment further in the future (50 words down, $word_{t+51}$ to $word_{t+101}$). This measures longer-range predictive accuracy.

- *Random*: The similarity between each continuation and a randomly selected 50-token segment from a different story. This serves as a baseline.

We then compared the average cosine similarity between the first three comparisons against the *Random* baseline using two-sample t-test.

3.2.4 Validating Emotion Classification

To assess the validity of the emotion classification model, we tested whether the predicted emotion probability aligned with the expected overall valence of the stories. Specifically, stories classified a priori as generally positive should elicit higher predicted probabilities for positive-valence emotions, while negative stories should elicit higher likelihoods for negative-valence emotions.

Quantifying the precise high-dimensional emotional trajectory of complex narratives is challenging. Therefore, we adopted a simplified approach by grouping both the stories and the predicted emotions into broad valence categories. First, based on their overall narrative arc and themes (see Table 3.1), the 18 stories were manually categorized as Positive, Negative, or Mixed. Second, the 28 GoEmotions labels were grouped into Positive, Negative, and Ambiguous valence based on the original dataset publication [Demszky et al., 2020] (Neutral was treated as a separate category and was therefore removed from this analysis).

We then used a linear mixed-effects model to statistically test for the expected congruence. We constructed a model predicting the word-by-word emotion probabilities with fixed effects of the story category (Positive, Negative, Mixed), the emotion valence category (Positive, Negative, Ambiguous), and their interaction. Story identity was included as random effect to account for baseline differences in likelihood across stories and the non-independence of word-level data within each story. A significant interaction effect, indicating that the likelihood pattern across emotion valences depends on the story category, would support the validity of our labeling approach.

3.3 Results

3.3.1 Continuations Predicts Narrative Context

To validate the semantic quality of the generated text, we examined the sentence embeddings across four comparisons: Within, Original, Random, and Next 50 (Figure 3.2).

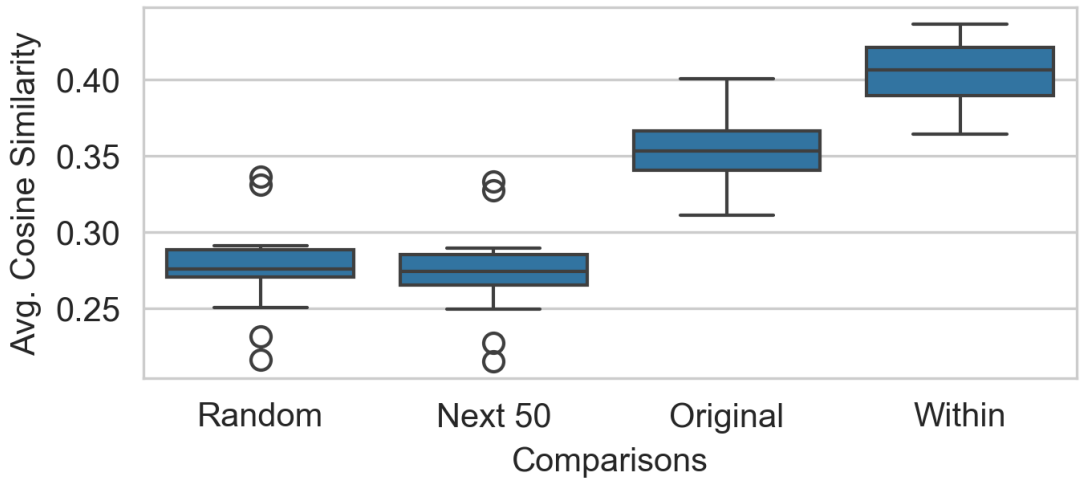


Figure 3.2: **Mean cosine similarity across four different comparisons.** Comparisons in *Random* have a mean of 0.277 and standard error of 0.030; *Next 50* ones have a mean of 0.274 and standard error of 0.030; the mean for *Original* comparison is 0.352 and standard error is 0.023; and for *Within*, the mean is 0.403 and standard error is 0.023.

As a measure of the continuations' consistency, the *Within* comparison yielded the highest mean similarity, and it was significantly higher than *Random* baseline (t_{15} = 7.967 , $p < 0.001$). The similarity score for the *Original* comparison was the second highest and also significantly higher than the Random baseline (t_{15} = 13.381 , $p < 0.001$). However, the *Next 50* comparison similarities were not significantly different from the random baseline (t_{15} = 0.271 , $p = 0.788$).

3.3.2 LLM Quantified Emotion Probabilities Aligned with Story Valence

To assess the validity of the emotion classifications, we examined whether the average likelihoods for specific emotions corresponded to the overall valence category of the stories. Qualitatively, a clear pattern emerged (see Figure 3.3). Positive-valence emotions (e.g., amusement, admiration, excitement, gratitude, joy) generally exhibited higher average likelihoods during positive stories compared to negative or mixed stories. Conversely, negative-valence emotions (e.g., annoyance, disapproval, anger, disappointment, disgust, sadness, fear, remorse) showed higher average likelihoods during negative stories. Surprisingly, some emotions typically considered positive (optimism, caring, love) showed higher average likelihoods in negative stories compared to positive ones.

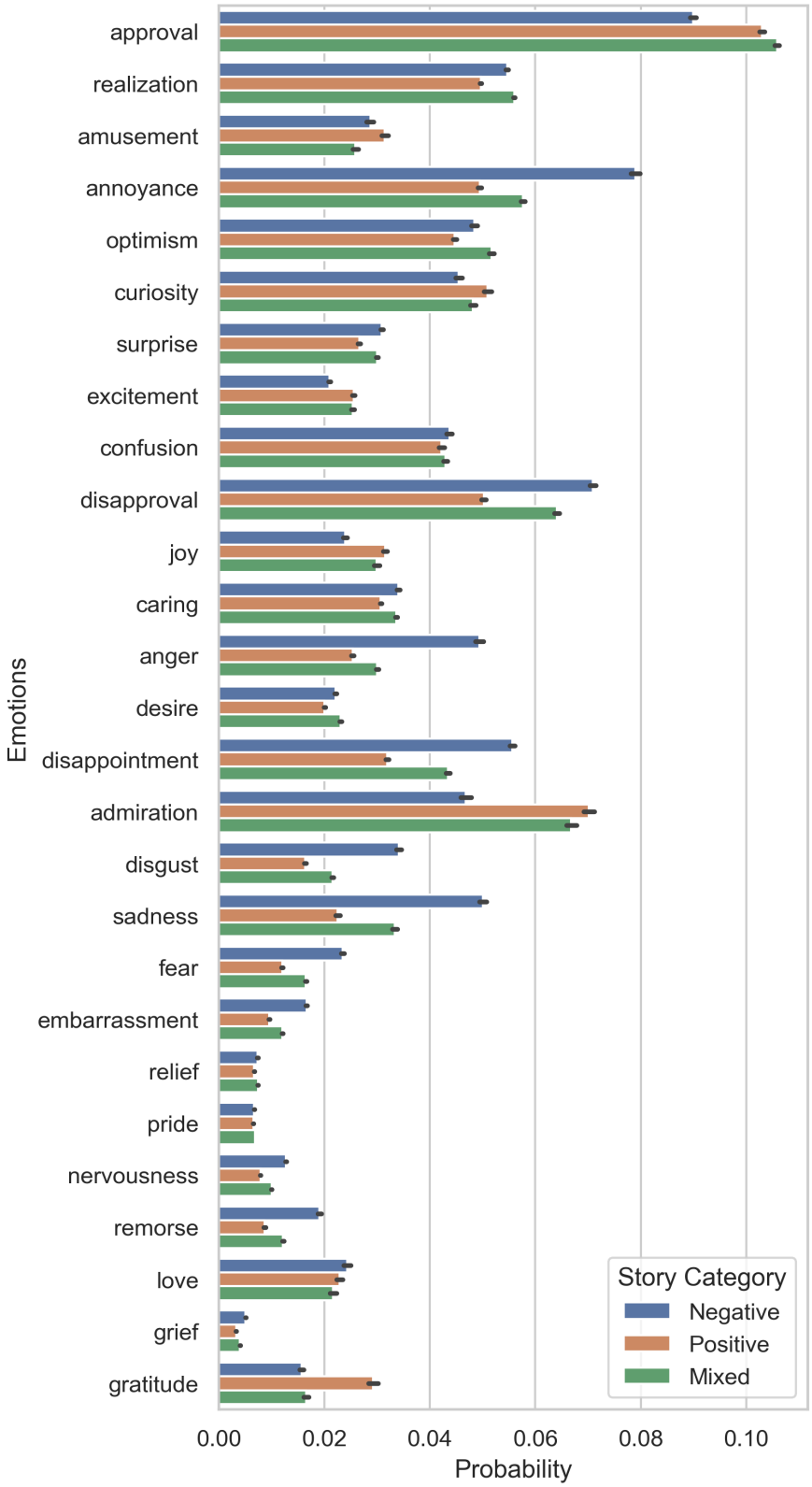


Figure 3.3: **Average emotion probability by story category.** Each bar represents the average probability given emotion and story category. Error bar showed 95% confidence interval.

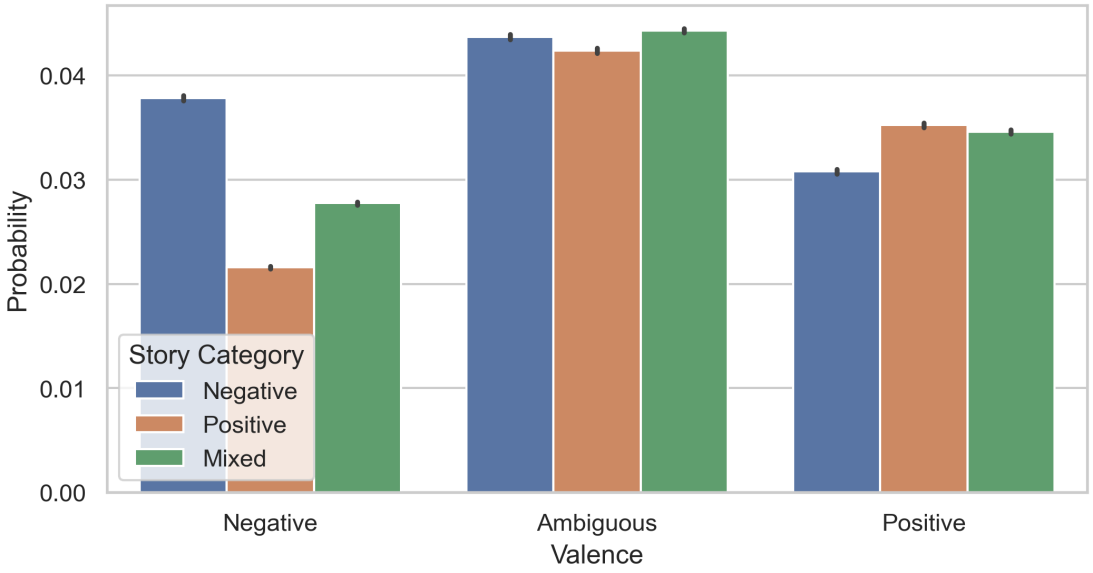


Figure 3.4: **Average valence probability by story category** Each bar represents the average probability given valence and story category. Error bar showed 95% confidence interval. Emotion to valence mapping was done according to the GoEmotion dataset publication [Demszky et al., 2020].

Statistically, the results from the linear mixed-effects model confirmed a significant interaction between the emotion valence category and the story category in predicting word-by-word emotion likelihood (ANOVA, Interaction Term: $F_{4,676791}$ = 2893.013, χ^2 = 11572.052, p $<$ 0.001). As illustrated in Figure 3.4 and confirmed by post-hoc tests, Positive stories were characterized by significantly higher likelihoods for positive-valence emotions compared to negative (Est. = 0.021, $T_6 76791 = 105.277$, p <0.001) or ambiguous ones (Est. = 0.006, $T_6 76791 = 21.254$, p <0.001).

3.3.3 Model Quantified Emotion Probabilities Showed Clustering

We next examined the temporal relationship between the 28 emotion probability time series by computing their pairwise Pearson's correlations within each story and then averaging the results. The correlation matrix, visualized in Figure 3.5, reveals a high degree of structure, indicating that many emotions consistently co-occurred throughout the narratives.

For example, strong positive correlations were observed among groups of related emotions, such as the cluster of embarrassment, disgust, disapproval, and disappointment.

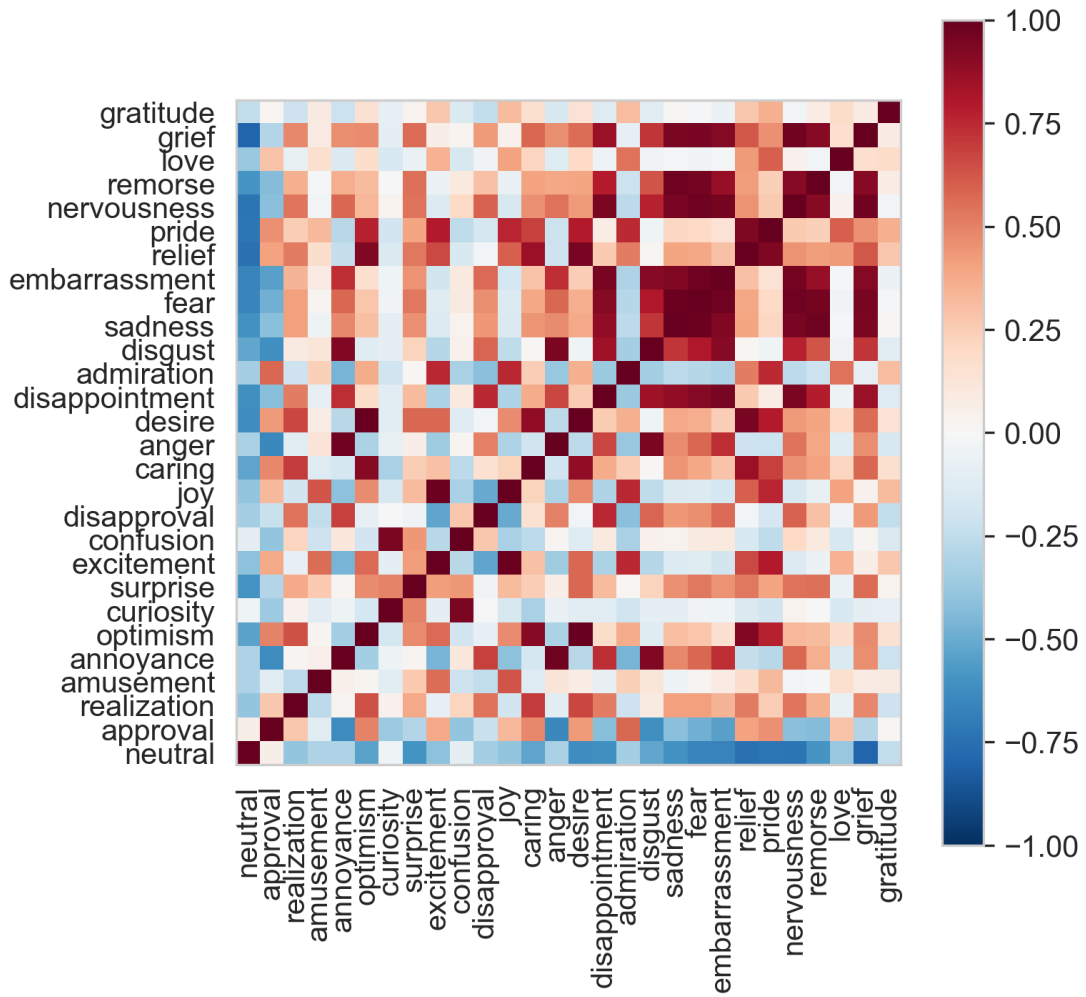


Figure 3.5: **Correlation Heatmap for Emotion Probabilities** Pearson correlations for pairs of emotion probabilities. Brighter means higher correlation and darker means lower. Emotions do co-occur: for example, the embarrassment, disgust, disapproval and disappointment are highly correlated.

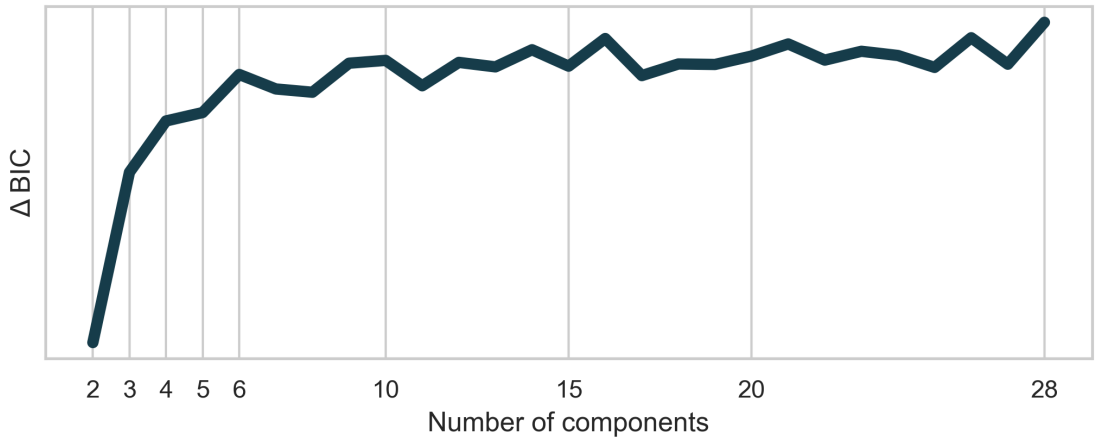


Figure 3.6: **Model Improvement (measured as increase in BIC) by number of HMM states.** As the number of HMM state increases, BIC increase. The biggest improvement happened when increasing the number of states from 2 to 3. The improvement stayed relatively stable after 10 HMM states.

This high multicollinearity suggests that the 28-dimensional emotional space can be effectively represented by a lower-dimensional set of latent states. This finding provides a direct rationale for our use of the Hidden Markov Modeling.

To model the temporal persistence of these states and identify a lower-dimensional set of latent emotional categories, we fitted HMMs with different number of states (min = 1, max = 28). We first examined relationship between the amount of variance explained and the number of states. As shown in Figure 3.6, model improvement slowed at 6 and reached asymptote at around 10. As expected, the biggest improvement happened when the state number increased from 2 to 3. With 3 states, the model can account for positive, negative and neutral emotions. Finally, we examined the correlational heatmap between the each emotional probability time series and the HMM predicted states' time series in Figure 3.7.

3.4 Discussion

In this chapter, we developed and provided initial validation for a computational pipeline to quantify emotional states from naturalistic stories. Our three-step process: generating predictive continuations, classifying their emotional content, and modeling their temporal structure, produced continuous and high-dimensional time series of emotional state labels. We demonstrated two validations for this pipeline. First, we

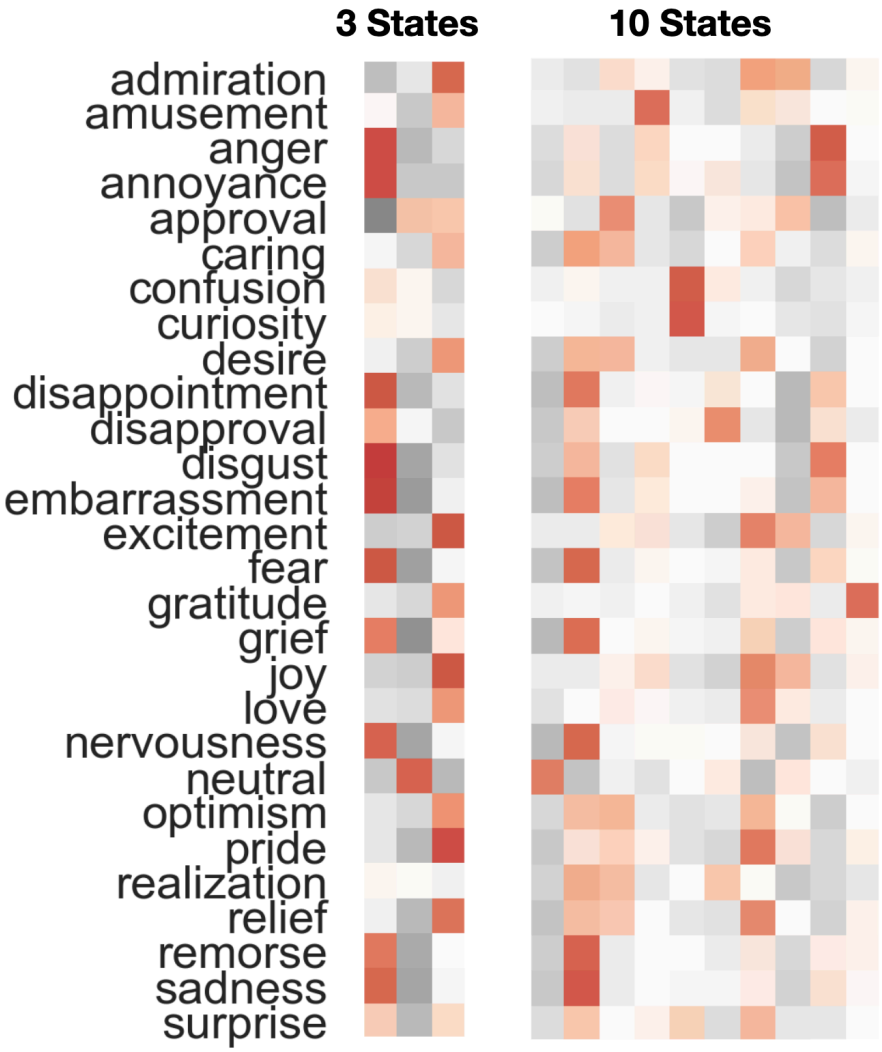


Figure 3.7: **Correlational Heatmap for HMM states and Emotions** The two heatmaps here showed the correlation between emotion probability time series and the predicted state time series from two HMM models (number of state = 3 and 10). In the 3 states model, we can see that the model nicely split into negative, neutral and positive, while the 10 states model showed more granularity in differentiating different emotions.

illustrated the semantic validity of the LLM-generated continuations, which captured the immediate narrative trajectory. Second, we established the construct validity of the classified emotion probabilities, which aligned with the stories' overall valence and exhibited a coherent, psychologically plausible correlational structure.

A key methodological decision was to use a modular pipeline: a Llama 2 7B model for continuation generation and a separate fine-tuned DistilBERT model for emotion classification. This stands in contrast to an alternative approach, such as prompt engineering with a single LLM to directly output an emotion label for a given text segment [Mayer et al., 2023, Botunac et al., 2024, Xu et al., 2024]. We argue this modular approach, despite using a set of older foundation models, offers advantages in both interpretability and robustness. Theoretically, our pipeline explicitly models the two processes central to the metareasoning function of emotional state: prediction and appraisal. The Llama 2 model acts as a predictive model, generating a distribution of possible future directions that a listener might anticipate. The DistilBERT model then acts as an appraisal model, evaluating the emotional content of those predicted futures.

This separation allows us to independently validate each component. As our semantic analysis showed, we can inspect the content of the generated continuations, offering transparency into the information being fed to the appraisal stage. In contrast, the prompt-engineered approach functions more like a "black box" in this regard. While it is theoretically possible to probe monolithic models' internal states [Bilal et al., 2025, Bills et al., 2023], this process is notoriously difficult and less accessible. Additionally, the final classification remains vulnerable to prompt-specific biases [Mao et al., 2023], making the output less reliable. Our modular approach allows for more targeted analysis of each sub-process, paving the way for future work that can systematically test how specific features of the predictive content (e.g., its uncertainty, language structure, or certain semantic themes) drive the final emotional state appraisal.

Recent LLM advances have focused on superior long-form reasoning, complex instruction-following, and multi-modal integration [Islam and Moushi, 2025, Patil and Jadon, 2025], but our pipeline's goal is more specific: to model the moment-to-moment predictive processing underlying language comprehension. For this, even the older LLMs (such as the LLaMA2) are already highly proficient [Heilbron et al., 2022, Caucheteux et al., 2023]. This perspective also reframes our finding on the LLM's "limited predictive horizon." The generated continuations were highly similar to the immediate narrative future but not the distant future. We argue this does not necessarily reflect deficiency in the model's

capability but feature of the stimuli themselves. These narratives are not linear and fully predictable proofs. Rather, they are personal stories chosen for their emotional power, which often relies on subverting expectations with a surprising twist. In fact, a model that could "correctly" predict these surprising turns might actually be a poor proxy for the human listener's subjective state of surprise. The limited horizon we observed suggests our pipeline is capturing a psychologically plausible, local predictive window. This local window is in alignment with moment-to-moment affective appraisal, which operates on immediate "what happens next" predictions rather than complete, long-range ones.

Assessing emotion from naturalistic stimuli is notoriously difficult [Greasley et al., 2000, Pólya and Csertő, 2023, Weninger et al., 2015, Saarimäki, 2021]. Here, we demonstrated that a relatively simple pipeline can extract a valid and continuous affective signal from the complex naturalistic stories, such that our word-by-word emotion probabilities, when aggregated, successfully showed significant differences between stories of different valence. The analysis of the emotion correlation structure and the subsequent HMM fitting provides strong evidence for the psychological plausibility of the labels. The 28 emotion categories are not independent but co-occur in meaningful clusters, aligning with recent landmark work showing data-driven maps of human emotional experience [Cowen and Keltner, 2017]. The HMM leverages this structure to identify a lower-dimensional set of persistent, latent states.

A key aspect of our work described here is the absence of moment-to-moment subjective ratings from participants. While a common practice in research on affective experiences, we omitted them based on the concern that such inquiries would create a disruption in the continuous semantic processing we aimed to measure. Indeed, it is impossible to generate a single individual's full high-dimensional emotional trajectory. Without this "ground truth," it is not possible to definitively validate that the distinct affective states quantified by our model are subjectively experienced by participants; but it is also the case that this full construct is not observable either by an experimenter, or, arguably, even by individuals themselves. Averaging over multiple ratings by individuals has its own set of conceptual limitations.

One obvious current challenge with the use of LLMs is that the different models change rapidly. As such, the results presented here may change with the use of different models. Furthermore, this study utilized a specific set of emotionally charged, first-person narratives. It remains an open question how well this method generalizes to other forms of stimuli, such as fictional stories, less emotional texts, and most importantly non-linguistic contexts like cognitive tasks. Exploring the robustness of this approach

across different stimulus types will be crucial for establishing it as a domain-general tool for the quantification of affective states.

S	Story	Dur	Brief Summary	Valence	Theme	Pos
1	**A Doll House**	503	Emotional abusive moments inflicted by father	Negative	Parental Issues	Before
	The Closet that Ate Everything	649	Sorting late mother's closet, finding a box of greeting cards.			
	Have You Met Him Yet	1013	White house staff recalled the path to meet Obama	Positive	Special Taskforce	Between
	Adventures in Saying Yes	804	Bitter-sweet moments of having adopted children	Mixed	Being Parents	
	Legacy	820	A father realized his young daughter shares his life struggles			After
	Naked	865	Empowerment and self-confidence in striping	Positive	Special Taskforce	
2	**It's a Box**	730	Experience of being assaulted and harmed further by stigmatism	Negative	Hurt and Hurt Again	Before
	Under the Influence	627	Hope turned into denial after years of parents rage over coming out			
	Hang Time	668	Hilarious recount of first helicopter training session went south	Positive	New experiences	Between
	Buck	685	Overwhelmed by everything anew after being jailed for 26 years			
	Alternate Ithaca Tom	707	Wondering what life would be like if some decisions were made differently	Mixed	Mid-life Crisis	After
	How to Draw	728	Rebirth and embrace creativity and artistry after decades of number crunching job			

Table 3.1: **Emotional valence and general theme of the selected stories.** Stories were also presented to participants in experiment sessions. *S* denote session number, *Dur* showed the story duration in minutes. *Valence* indicate experimenter determined overall emotional valence. *Pos* showed the story position relative to other components in the session.

Chapter 4

Neural correlates of LLM-identified emotional states

4.1 Introduction

In the preceding chapter, we described and initially validated a computational pipeline to quantify emotional states from narrative text. Rooted in the emotional-state-as-metareasoning-heuristic framework, our approach explicitly modeled the listener's predictive and appraisal processes to produce a continuous, high-dimensional time series of emotional state labels. We characterised aspects of the construct validity of the resulting states, demonstrating that the labels are semantically grounded, align with the overall valence of the stories, and exhibit a psychologically plausible correlational structure. However, these model-quantified emotional state labels remain a purely in silico construct, disconnected from the neural processes it aims to capture.

The primary goal of this chapter is to provide validation at a neural level. If these model-derived emotional states are meaningful, they should correspond to consistent and distinctly identifiable patterns of neural activity. Establishing such a neural consistency is hence a key prerequisite for the more ambitious decoding and generalization analyses in chapter 6.

The search for this neural correlate demands an imaging modality that can balance the competing demands of temporal and spatial resolution. Our LLM-based pipeline was specifically designed to capture the rapid, word-by-word dynamics of emotion, which can shift on a sub-second timescale [Kragel et al., 2022, Gagne and Dayan, 2023]. This rules out methods with low temporal resolution, such as fMRI, as they would average over these critical, fine-grained temporal changes. This need for sub-second temporal

precision is echoed by that language comprehension itself operates on this timescale. For instance, the brain processes semantic surprisal within 400ms (the N400) and detects syntactic violations at approximately 600ms (the P600) [Kutas and Hillyard, 1980, Osterhout and Holcomb, 1992].

However, this temporal focus cannot come at the complete expense of spatial information. Studies using fMRI, despite their temporal limitations, have successfully decoded emotional states, demonstrating that these states are represented in distributed, high-dimensional patterns across the brain [Kassam et al., 2013, Baucom et al., 2012, Saarimäki et al., 2016]. This confirms that spatial pattern of the neural activity is also critical for differentiating emotional states.

We therefore employed magnetoencephalography (MEG). MEG combines high temporal resolution with some spatial resolution. Unlike electroencephalography (EEG), magnetic fields are only minimally distorted by the hair, skull and soft tissues [Hämäläinen et al., 1993]. This fundamental physical difference means MEG signals are less likely to be distorted spatially, providing both spatially and temporal precision to capture dynamic cognitive processes [Liu et al., 2019, Youssofzadeh et al., 2023].

4.1.1 Current Study

The central goal of this chapter is to test the neural consistency of the LLM-derived emotional state labels. We hypothesize that if these labels capture neurally meaningful emotional states, then different moments in time assigned the same label should evoke consistent and distinct patterns of neural signal.

To test this, we analyzed MEG data from participants listening to the same naturalistic stories. We first performed a replication analysis of neural responses to word surprisal [Heilbron et al., 2022]. This served as a crucial validation step to confirm that participants were actively engaged in the predictive processing during story comprehension, a prerequisite for our appraisal-based model. We then used two complementary analyses to test for state-specific neural signatures: 1) a time-domain ANOVA to identify spatio-temporal clusters where MEG activity significantly differed between emotional states, and 2) a frequency-domain topographical similarity analysis to determine if these states correspond to stable and distinct spatial patterns of neural power.

4.2 Methods

4.2.1 Ethical approval

Experimental protocol was approved by the University College London Research Ethics Committee (Approval: 27121/001)

4.2.2 Participants

13 participants were recruited from local communities and the University College London research participant registry as part of a larger study. All participants were UK residents, native English speakers, and at least 18 years old. Interested individuals underwent screening based on inclusion/exclusion criteria. Qualified participants provided informed consent before the experimental sessions.

Inclusion / Exclusion Data from nine participants were included in the final analysis. Four participants were excluded due to: intoxication during screening, excessive MEG signal noise caused by dental retainers, excessive head movement during recording, and failure in recording the trigger signal during one session. All procedures were approved by the UCL's Research Ethics Committee.

4.2.3 Procedure

Participants attended three separate MEG recording sessions, spaced at least 24 hours apart. Before the first session, participants were screened for MEG safety to ensure removal of all metal content. Inside the magnetically shielded room, participants were seated comfortably in the MEG scanner. Head position indicator coils were placed on the participant's head (left/right preauricular points and nasion) for head localization and continuous head tracking. Auditory stimuli were delivered binaurally via MEG-compatible insert earphones. Visual stimuli were presented on a screen positioned in front of the participant.

Each session consisted of six story blocks, each corresponding to one story. The details of the stories, including their session membership, duration, emotion valence, semantic theme and position in the block are described in chapter 3 table 3.1. At the start of each block, the participant's initial head position was recorded. Head movement was monitored throughout the block. Participants were instructed to remain still and listen

attentively to the stories. Optional short breaks (maximum 5 minutes) were offered between blocks. No behavioral responses or subjective ratings related to the stories were collected during the MEG sessions.

4.2.4 Emotional State Labels

An emotional state label for each word was inferred as detailed in chapter 3. Briefly, this consisted of a three-step process relying on a Large Language Model. To balance emotion granularity and the number of available labels per class, we chose 10 emotional states as our classification targets for training and validating the neural decoders (See Figure 4.1 for state labels and their emotion composition). A higher number of emotional states only yielded minimal model performance increase (as showed in Figure 3.6).

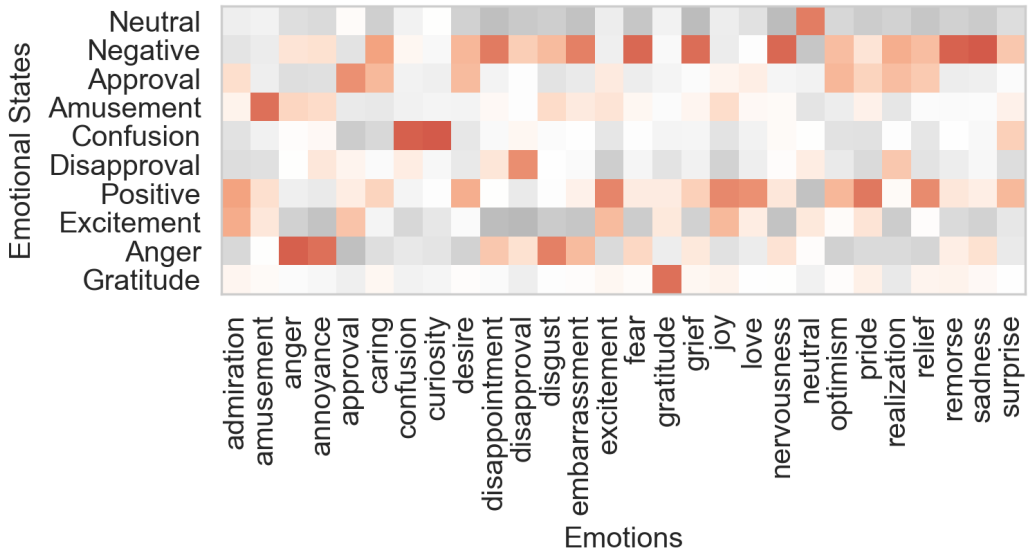


Figure 4.1: **10 Emotional States and Corresponding Emotion Composition** The 10 emotional states that were used for the analysis and a correlation map showing the correlation between each emotional state time series to emotion probability time series. Deeper red indicate a higher correlation.)

4.2.5 MEG Data Processing

MEG data preprocessing and analysis were performed using the MNE-Python library [Larson et al., 2025] and custom Python scripts. All analyses shown here only included

data from session 1 and session 2, therefore, 12 total stories per person. The last session data was withheld for future validation.

Data Preprocessing

The continuous raw MEG data for each session underwent three preprocessing steps: cropping, filtering and independent component analysis.

Data Cropping: Raw data were segmented to include only when participants were listening to the stories, using recorded trigger signals marking the start and end of each story block.

Filtering and Downsampling: Data were band-pass filtered between 0.5 Hz and 200 Hz using a finite impulse response filter. The 0.5 Hz high-pass cutoff was chosen to remove slow drifts while preserving low-frequency content relevant for language processing [Youssofzadeh et al., 2020]. The data were then downsampled to 400 Hz to reduce computational load while retaining frequencies up to the new Nyquist frequency of 200 Hz. This range adequately covers lower frequencies and the high gamma band (around 55-100 Hz), which showed significance in emotion classification [Li and Lu, 2009]. Power line noise was removed using notch filters at 50 Hz and its harmonics (100 Hz, 150 Hz) using the MNE function "notch_filter".

Independent Component Analysis (ICA) for Artifact Detection: ICA was used to identify and remove artifacts related to eye blinks and cardiac activity. ICA decomposition was performed using the FastICA algorithm [Ablin et al., 2018], computing 200 components. Artifactual components were identified based on their correlation with electrooculogram (EOG) and electrocardiogram (ECG) signals. EOG signals were approximated by the two preauricular point electrodes. An ECG signal was approximated through cross-channel averaging. Components exhibiting high correlation (Pearson's r > 0.9) were marked for removal, using automated detection functions "find_bads_eog" and "find_bads_ecg". The cleaned MEG signal was then reconstructed from the remaining non-artifactual components.

Word Onset Detection

Narrative audio stimuli were delivered to participants using a dedicated audio presentation system, which simultaneously keeps an copy of the auditory analog signal alongside the MEG data, sampled at the same rate. Word-by-word timing annotations (onsets and durations) for the stories were obtained from the original dataset source

[LeBel et al., 2023]. To synchronize these annotations with the MEG signal, we aligned the original stimulus audio file with the audio signal recorded by the MEG system. This was accomplished by computing the cross-correlation between the two signals. We identified the optimal temporal lag that maximized this correlation. This lag value was then used to shift all word-level timing markers from the original annotations. This procedure yielded precise, word-level event timings (onsets and durations) that were time-locked to the MEG data.

MEG Data Acquisition

MEG data were acquired using a 275-channel CTF whole-head MEG system (CTF MISL, Coquitlam, BC, Canada) housed in a magnetically shielded room at the Wellcome Centre for Human Neuroimaging, UCL. Data were recorded continuously with a sampling rate of 1200 Hz. Head position was tracked continuously using the HPI coils.

MEG Data Epoching

To match the word-level emotional state labels we quantified from the stories, we created word-by-word epochs, aiming to decode the emotion label for each word.

One important consideration is that words have short durations, on average 250ms per word in our dataset. If epochs were aligned with word onset, then either the epoch length would be limited to the shortest word duration or the epoch number would be reduced because shorter words would be removed. The first is unreasonable as the words can be as brief as 100ms, which is insufficient in capturing most human cognitive processes. While the latter is more logical, we might be introducing biases by eliminating words that are shorter. Additionally, different stories feature speakers whose delivery proceeds at different paces, which could further introduce story-based biases. Therefore, we created two types of epochs for training our decoders: *overlapping* and *non-overlapping*. Each epoch is then paired with one emotional state label. To align with the introspection phase in task blocks, the duration of all the epochs was set to be 2000ms.

Overlapping Epochs These epochs were constructed for event-related analysis using raw MEG signal. We avoid the issue of short word duration by allowing the epochs to overlap with each other. For most raw signal analysis, time after onset is a significant feature. Therefore, the same data point will have different time after word onset for

different words.

Non-overlapping Epochs These epochs were constructed for average power spectral density analyses. Because the average power spectral density averages over time, overlapping the epochs will create data leakage if the peak of the power spectral information exists in the overlapping window. Therefore, for analysis using this measure, we created epochs that are fixed in duration and sequential in a story but without overlap. These epochs can hence cover multiple words, and the emotional state label for each epoch was determined by the emotion with the longest cumulative duration within that epoch.

Power Spectral Density (PSD) Calculation

Event-related analyses are ill-suited for naturalistic paradigms because they assume a fixed processing latency relative to a stimulus onset. The time required for semantic comprehension can vary significantly from word to word, making it difficult to obtain a consistent time-locked neural signal across trials [Dell'Acqua et al., 2010, Brennan, 2016, Hauk et al., 2004]. To address this temporal variability, we calculated MEG PSD. By averaging spectral power over a time window, PSD topographies capture the spatial distribution of oscillatory activity associated with an emotional state, making them less sensitive to precise peak latencies.

We calculated the PSD for each of the epochs and calculate the power for different frequency bands by averaging the PSD across frequencies within each power band. Here, we constructed six power bands: delta (0.5-4Hz), theta (4-8Hz), alpha (8-12Hz), beta (12-30Hz), gamma (30-45Hz) and high gamma (55-100Hz). We used the compute_psd method (with multi-taper method) implemented in the MNE package[Larson et al., 2025] to calculate PSD.

4.2.6 Analysis

Replication of Word Surprisal Neural Effects

To confirm that participants were actively engaged in predictive processing during story listening, we conducted a replication analysis of a previous finding on the neural correlates of word surprisal [Heilbron et al., 2022]. We quantified the surprisal of each word in the stories using a pre-trained GPT-2 model. For each word, surprisal was defined as the next-word probability given the preceding text.

We then performed a univariate regression analysis to model the MEG signal at each sensor and time point. The MEG signal was predicted using word surprisal as the primary regressor, while controlling for word-level acoustic features. The resulting beta coefficients for the surprisal regressor were tested for significance using a spatio-temporal cluster permutation test as implemented in the MNE package [Larson et al., 2025]. We used a cluster-forming threshold of $p < 0.001$ (two-tailed) with 10,000 permutations. Clusters were considered significant if their FDR corrected p-value was less than 0.05.

Neural Consistency: Time-Domain ANOVA

To test our main hypothesis that LLM-derived emotional states correspond to distinct neural patterns, we examined the consistency of MEG signals within each model quantified emotional state. First, we segmented the continuous MEG data into 2000 ms epochs, time-locked to the onset of each word. Epochs containing signal amplitudes exceeding a peak-to-peak threshold of $4 * 10^{-12}$T were discarded. Each remaining epoch was assigned the discrete emotional state label determined by the HMM for the corresponding word.

We then performed univariate one-way ANOVAs at each sensor and time point across epochs, with the HMM emotional state label as the independent variable and accounting for sensor adjacency. This tested whether the mean MEG signal differed significantly between the inferred emotional states. To correct for multiple comparisons across space and time, we used the permutation-based spatio-temporal clustering function as implemented in MNE [Larson et al., 2025]. We set a cluster-forming threshold corresponding to an F-statistic with $p < 0.001$. The significance of the resulting spatio-temporal clusters was assessed via 10,000 permutations, with a final cluster-level significance threshold of $p < 0.05$.

Neural Consistency: Frequency-Domain Topographical Similarity

To complement the time-domain analysis and mitigate potential confounds from variable word durations, we performed a spectral analysis using non-overlapping epochs. For each participant and story block, the continuous MEG data were segmented into consecutive, non-overlapping 2000 ms epochs. The first epoch was aligned to the onset of the first word in the block, and subsequent epochs began immediately where the previous one ended. Each epoch was then assigned the single model-derived emotional state label that occupied the longest duration within that 2000 ms window.

For each non-overlapping epoch, we calculated the power spectral density (PSD) across a frequency range of 0.5 to 100 Hz using Welch's method, as implemented in MNE [Larson et al., 2025]. The resulting power values were then averaged across five canonical frequency bands: Delta (1–4 Hz), Theta (4–8 Hz), Alpha (8–12 Hz), Beta (12–30 Hz), Low Gamma (30-45Hz) and High Gamma (55–100 Hz). We then calculated the individual-level topographies by averaging. We finally generated a cross-individual grand-average power topography for each emotional state within each frequency band, allowing for qualitative visualization of the spatial topography.

To quantify the spatial consistency of these topographies within each emotional state, we used a 5-fold cross-validation procedure. For a given individual, state and frequency band, epochs were split into 5 folds. Iteratively, 4 folds were used to compute an average "template" topography for that state/band. The cosine similarity was then calculated between this template and the topography of each individual epoch in the held-out test fold. This process yielded a topographical cosine similarity score for each epoch. To determine the chance level for the consistency score, we randomly re-assigned the labels for each epoch and repeated this process 1000 times. With this permutation-based method, we obtained the null distribution of cosine similarity scores across emotional states. To determine the statistical significance, we conducted one sample t-tests against the permutation-generated null similarity. We then corrected the p values with FDR method to account for multiple comparisons.

4.3 Results

4.3.1 Neural Word Surprisal Signal

We first sought to confirm that participants were actively engaged in predictive processing. To do this, we replicated results on LLM quantified word surprisal in MEG signal. The cluster-based permutation test on the regression coefficients revealed significant neural responses to word surprisal, as shown in Figure 4.2. We identified a positive cluster peaking between 400 and 500 ms after word onset, localized to bilateral temporo-parietal sensors. The timing and topography of this effect are highly consistent with the N400 event-related field. This result demonstrated that our participants were actively predicting upcoming words.

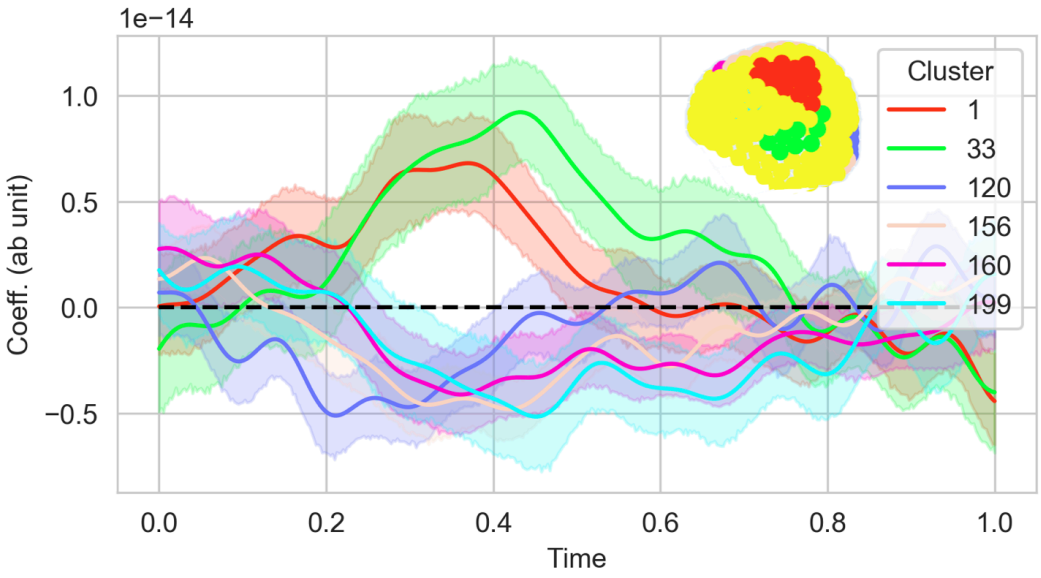


Figure 4.2: **Spatial Temporal Clusters Corresponding to the Next-Word Probability** Spatial Temporal Clustering was done on the coefficient of linear regression models predicting next-word probability with MEG signal. Main graph showed the coefficients time course (from 0-1000ms since word onset) of the significant clusters. Auxiliary plot showed the location of the significant clusters and the color corresponds to the main graph's clusters.

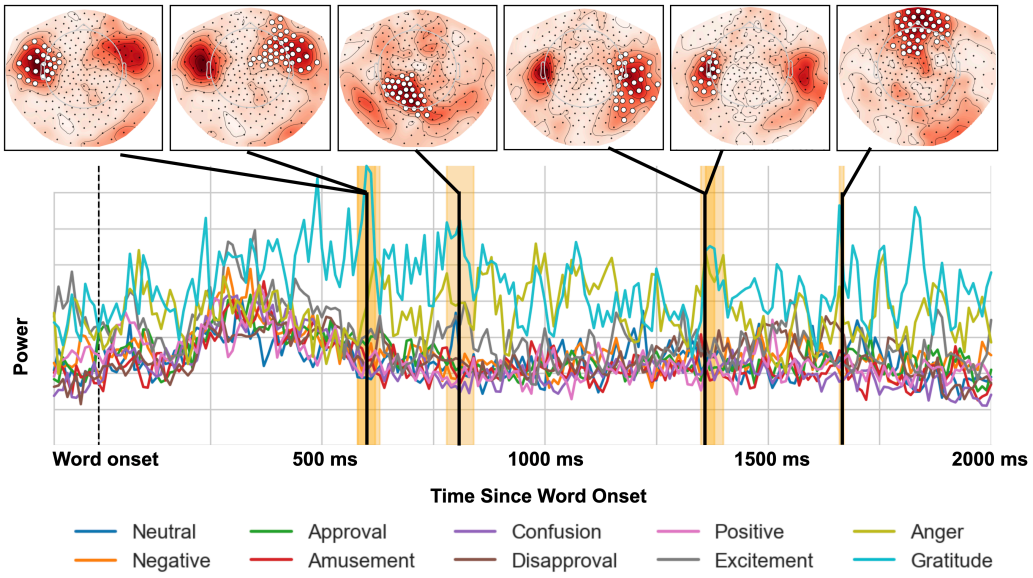


Figure 4.3: **Significant Spatial Temporal Clusters for Differentiating 10 Emotional States** Average MEG signal aggregated of each 10 emotional states between 0 to 2 seconds after word onset. Yellow shades time when the F test is significant, black line links the topography of the significant clusters.

4.3.2 Time-Domain Consistency within Quantified Emotional States

We next asked whether the distinct LLM-derived emotional states correspond to distinct neural patterns. The univariate ANOVA, corrected for multiple comparisons using a spatio-temporal cluster test, showed significant differences in MEG activity across the 10 emotional states (Figure 4.3). The analysis identified two primary spatio-temporal clusters where neural activity reliably discriminated between the states. An early cluster emerged from approximately 200 to 500 ms after word onset with a fronto-central topography. A later cluster from roughly 800 to 1400 ms localized to a right temporo-parietal sensor group.

4.3.3 Spectral Topographical Consistency within Quantified Emotional States

Next, we analyzed the spectral topographies of non-overlapping MEG epochs to see whether the LLM-quantified emotional states are associated with distinct and stable spatial patterns across frequency bands. Qualitatively, the template topographies generated for each state showed unique spatial distributions of power, particularly for the states derived from the 10-state HMM model (Figure 4.4).

While these grand-average topographies suggest that each emotional state was associated with distinct neural patterns, this qualitative observation requires quantitative validation. A meaningful average topography must be representative of the individual epochs it is averaged from. Therefore, to formally test this, we calculated the cosine similarity between within-state epochs pairs across all the affective states and frequency bands using 5-fold cross-validation. The null distribution of the cosine similarity scores for each emotional state are: Amusement (-0.015), Anger (-0.027), Approval (-0.050), Confusion (-0.034), Disapproval (-0.060), Excitement (-0.035), Gratitude (-0.011), Negative (-0.047), Neutral (-0.071), and Positive (-0.016). Generally, the scores were all slightly negative and centered near zero (ranging from -0.071 to -0.011), establishing a baseline of no meaningful similarity against which the true within-state consistency was compared. For statistical significance for each of the emotional state and power band, see Table 4.1.

As shown in Figure 4.5, the resulting cosine similarity scores were highest for topographies in the gamma band, indicating that these templates were representative of individual epochs for most emotional states. In contrast, consistency was generally lower in other frequency bands. This frequency-domain analysis corroborates our time-domain results, suggesting that the distinct LLM-inferred states are associated with neurally

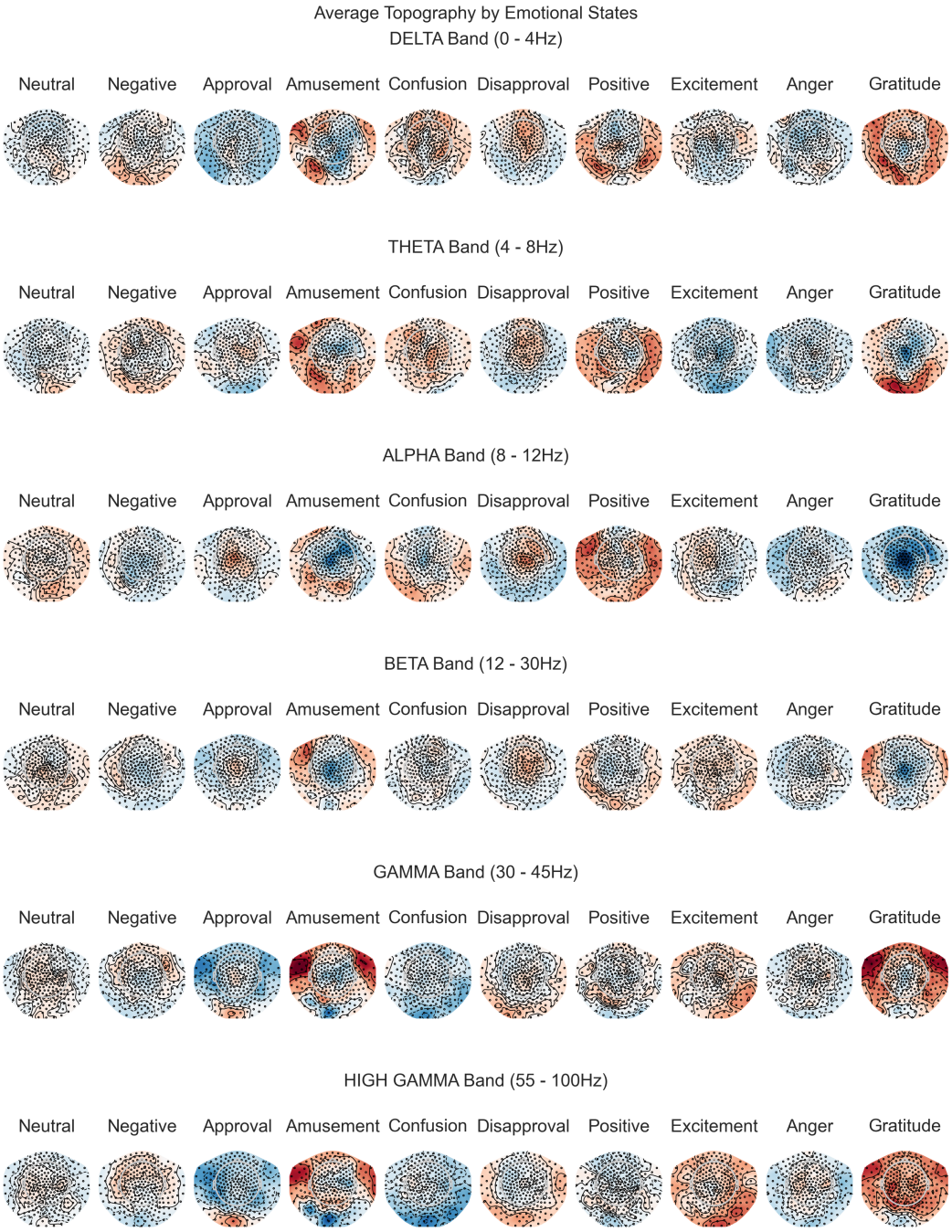


Figure 4.4: **Distinct spectral topographies across 10 Emotional States and frequency bands** Non-overlapping epochs' PSD topography averaged across participants. Distinctive spatial patterns exist for each model quantified emotional states.

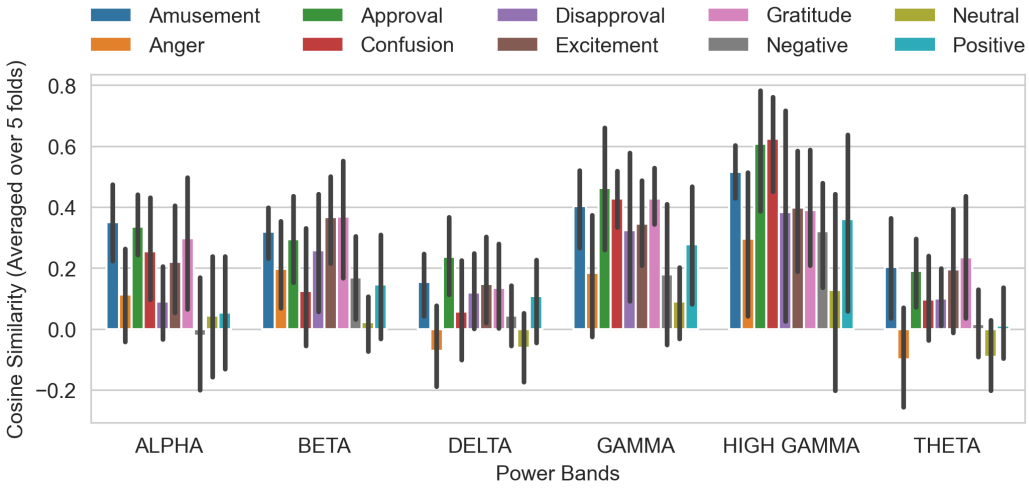


Figure 4.5: **Cosine Similarity Across Epochs for 10 Emotional States** Each bar showed the average Cosine Similarity for a frequency band and emotional states. Error bar denotes 95% confidence interval. High-frequency bands, the gamma and high gamma, showed higher level of consistency. However, variance exists across emotional states, for example, Gratitude showed high consistency regardless of power band.

distinct signatures.

Band	Emotional States	T_8	$p_{corrected}$	Significant?
DELTA	Neutral	0.145	0.467	
DELTA	Negative	1.726	0.083	
DELTA	Approval	4.219	0.006	*
DELTA	Amusement	3.057	0.021	*
DELTA	Confusion	1.032	0.188	
DELTA	Disapproval	2.588	0.032	*
DELTA	Positive	1.691	0.083	
DELTA	Excitement	2.381	0.038	*
DELTA	Anger	-0.561	0.717	
DELTA	Gratitude	1.834	0.073	
THETA	Neutral	-0.323	0.644	
THETA	Negative	1.047	0.188	
THETA	Approval	3.910	0.008	*
THETA	Amusement	2.482	0.036	*
THETA	Confusion	1.696	0.083	
THETA	Disapproval	2.880	0.024	*
THETA	Positive	0.423	0.373	

THETA	Excitement	2.080	0.055	
THETA	Anger	-0.797	0.776	
THETA	Gratitude	2.185	0.048	*
ALPHA	Neutral	1.056	0.188	
ALPHA	Negative	0.252	0.433	
ALPHA	Approval	7.281	0.001	*
ALPHA	Amusement	5.288	0.002	*
ALPHA	Confusion	3.112	0.021	*
ALPHA	Disapproval	2.464	0.036	*
ALPHA	Positive	0.680	0.286	
ALPHA	Excitement	2.759	0.026	*
ALPHA	Anger	1.648	0.086	
ALPHA	Gratitude	2.602	0.032	*
BETA	Neutral	1.976	0.062	
BETA	Negative	2.865	0.024	*
BETA	Approval	4.449	0.005	*
BETA	Amusement	7.050	0.001	*
BETA	Confusion	1.583	0.093	
BETA	Disapproval	3.068	0.021	*
BETA	Positive	1.719	0.083	
BETA	Excitement	5.017	0.003	*
BETA	Anger	2.844	0.024	*
BETA	Gratitude	3.513	0.013	*
GAMMA	Neutral	2.472	0.036	*
GAMMA	Negative	1.874	0.070	
GAMMA	Approval	4.805	0.004	*
GAMMA	Amusement	6.307	0.001	*
GAMMA	Confusion	9.651	< 0.001	*
GAMMA	Disapproval	2.879	0.024	*
GAMMA	Positive	2.862	0.024	*
GAMMA	Excitement	4.755	0.004	*
GAMMA	Anger	1.968	0.062	
GAMMA	Gratitude	8.187	< 0.001	*
HIGH GAMMA	Neutral	1.143	0.172	
HIGH GAMMA	Negative	3.814	0.009	*
HIGH GAMMA	Approval	5.789	0.002	*
HIGH GAMMA	Amusement	11.499	< 0.001	*

HIGH GAMMA	Confusion	7.918	< 0.001	*
HIGH GAMMA	Disapproval	2.278	0.042	*
HIGH GAMMA	Positive	2.288	0.042	*
HIGH GAMMA	Excitement	4.076	0.007	*
HIGH GAMMA	Anger	2.426	0.037	*
HIGH GAMMA	Gratitude	3.886	0.008	*

Table 4.1: **One-sample T tests Across Emotional States and Power Bands** To determine whether the cosine similarities are significantly higher than the permutation-based null distributions, we conducted multiple one sample t-tests and corrected for multiple comparisons using the FDR method.

4.4 Discussion

This chapter's primary goal was to neurally validate the computational pipeline described in Chapter 3. We hypothesized that if LLM-derived emotional state labels capture a meaningful psychological construct, then they should correspond to consistent and discriminable patterns of neural activity.

To start, we successfully replicated the established neural signature of word surprisal [Heilbron et al., 2022]. This result showed that predictive processing, one of the core cognitive process our pipeline captures, was active during listening of stories. Demonstrating that participants' brains were indeed generating and updating these predictions in real-time further confirms the validity of our emotional state quantification process and establishes a firm foundation for our subsequent analyses.

Building on this, our primary analyses confirmed that the model-quantified emotional states are neurally consistent and discriminable. The time-domain ANOVA identified significant spatio-temporal clusters where the MEG signal reliably differed between the 10 HMM states. We demonstrated that the labels generated by our pipeline captures affectively meaningful signals in the human brain.

The specific timing of the observed neural clusters provides further insight into how these emotional states may be constructed during language comprehension. We identified an early cluster (approx. 200-500ms) and a later, sustained cluster (approx. 800-1400ms). This temporal progression aligns remarkably well with cognitive and appraisal-based theories of emotion. The early fronto-central cluster (200-500ms) over-

laps in time with classic event-related components of language processing, such as the N400 [Kutas and Hillyard, 1980]. The temporal alignment suggests that the initial appraisal and differentiation of the emotional state may be tightly intertwined. The later, right-lateralized temporo-parietal cluster (800-1400ms) is consistent with the timing of late positive potentials, a well-established neural marker of sustained emotional salience and elaborative processing [Schupp et al., 2006]. Its presence suggests that our HMM-derived labels are capturing more than just the immediate appraisal of a word, and might also be tracking the sustained affective state that follows.

The frequency-domain analysis provides a converging line of evidence. We found that the emotional states were associated with quite distinct power topographies, with the highest consistency observed in the gamma and high gamma band. This finding is particularly intriguing. In cognitive neuroscience, gamma oscillations are widely implicated in information integration [Drebitz et al., 2025]. Emotional states under the constructive view are the product of an integration of mulitiple cognitive systems (e.g., sensory, memory, interoception) [Barrett, 2017b]. Our finding that gamma power topographies are the most stable identifier of our emotional labels aligns with this constructive view. Furthermore, this result resonates with a growing body of recent EEG decoding studies that have repeatedly identified high-gamma activity as a highly informative, feature-rich signal for classifying emotional states across different induction modalities, such as music [Yang et al., 2025] and videos [Liu et al., 2018, Du et al., 2023].

As we had discussed in chapter 3, the experimental design lacked moment-to-moment subjective ratings from participants. While this was a deliberate methodological choice to avoid disrupting the continuous semantic and affective processing that we aimed to measure, without this "ground truth," we cannot definitively validate that the neurally distinct states we identified were consciously experienced by the participants. However, the neural findings support the notion that aspects of the LLM quantification may be useful and valid. However, it remains an open question whether the neural signatures identified here are abstract and domain-general, or if they are specific to the linguistic context in which they were elicited. This will be the focus of the next two chapters.

Chapter 5

Inducing rich emotional states with cognitive tasks

5.1 Introduction

A wide range of emotion induction procedures have been developed and validated to reliably induce emotional states in laboratory settings. Among the most common are exposure to evocative media, such as viewing images, listening to music, or watching film clips [Zupan and Babbage, 2017, Jääskeläinen et al., 2021, Saarimäki, 2021]. Significant effort has gone into the creation of standardized stimulus sets to maximize effectiveness and replicability across laboratories. These have resulted in well-characterized and substantial corpora of stimuli, such as the International Affective Picture System [Lang et al., 1997], the Geneva Affective Picture Database [Dan-Glauser and Scherer, 2011], Complex Affective Scene Set [Weierich et al., 2019], and Image Stimuli for Emotion Elicitation [Kim et al., 2018]. Using these standardized stimulus sets, affective neuroscience researchers have investigated how affective states influence cognition and behaviors in healthy and ill populations [Davidson, 1998, Davidson, 2010, Cacioppo et al., 1999, Mogg and Bradley, 1998, Sheline et al., 2009, Bonanno et al., 2004]; identified neural correlates of affective states [Gusnard et al., 2001, Raichle, 2015]; and used them to develop emotion decoding procedures from neural or physiological signals [Shu et al., 2018, Torres et al., 2020, Zhang et al., 2020]. These tightly controlled and experimentally validated sets of stimulus have been foundational for reproducible affect research as they enable study of a reliable mapping between specific stimuli and affective states in the general population.

Methods that make use of naturalistic stimulus, such as movies or stories, relax experimental control in exchange for improved ecological validity [Zupan and Babbage, 2017, Jääskeläinen et al., 2021]. The primary advantage of the use of naturalistic stimuli is their ability to elicit a continuous, dynamic, and contextually consistent stream of affective states that unfold over time [Saarimäki, 2021]. They can evoke more complex and nuanced emotions, such as suspense or bittersweetness, which are difficult to capture with static images [Jääskeläinen et al., 2021]. Recent advancements in augmented reality devices allows affective elicitation through creation of immersive experiences, offering a unique opportunity to safely study affective states like stress in complex real-world contexts, such as driving, which are otherwise difficult to investigate [Somarathna et al., 2023, Baltodano et al., 2018].

A robust model of emotion must be able to generalize beyond the context in which it was built. While passive paradigms are invaluable for emotion induction, cognitive tasks provide an alternative setting to test this generalization. These tasks elicit affective responses through active engagement with an experimental paradigm. The key strength lies in the paradigms' ability to study parametrically induced emotional states: the experimenter can design the interactions and contingencies in the tasks, such as reward probability, task difficulty, or social feedback, to systematically induce emotional states.

This active engagement more closely mimics real-life situations, as the participant's own choices, combined with task contingencies and performance feedback, jointly determine their emotional experience. This, counterintuitively, also means the resulting emotional state might be less controlled by the experimenter than in a passive paradigm. There are two primary sources of this variability. First, participants make different choices, which can lead them down unique experimental paths and create different emotional trajectories. Second, even when faced with the same objective outcome, individuals appraise it differently based on their personal history, goals, or prior experiences in the task. This variability should not be considered as a flaw but a way to elicit a more diverse range of emotional states.

A prominent design for investigating affective experience within computational tasks was established by Rutledge and colleagues [Rutledge et al., 2014]. The key part of this design is not the task itself, but the integration of a parametric task with repeated, momentary subjective self-reports. In the original study, participants performed a probabilistic reward task and were intermittently asked, "How happy are you right now?" This method allowed the authors to build computational models that directly linked objective task variables, such as reward prediction errors and expected values, to mo-

mentary happiness ratings.

The work in Chapter 2 of this thesis also followed this design, linking cognitive effort related variables to momentary happiness ratings. However, human affective experience is not limited to a single dimension of happiness or valence. As outlined in the introduction chapter, affective states, and hence emotional states, are high-dimensional, granular constructs. This granularity is functionally critical: two different negative states, like anger and sadness, may have similar valence but represent entirely different computational heuristics that promote vastly different actions. A one-dimensional momentary happiness scale is insufficient to capture anger and contempt in a social game [Fischer and Roseman, 2007] or the positive feeling of pride from solving a difficult math problem[Ashcraft, 2002, Richardson and Suinn, 1972]. To build a comprehensive model of emotional state, it is essential to use a measurement tool that can capture this rich, granular landscape.

This chapter introduces a battery of four cognitive tasks, each designed to probe a distinct cognitive and affective domain. The goal is to create an experimental battery with wide coverage of the emotional landscape. This includes the probabilistic reward task to elicit emotions tied to prediction errors, a social interaction task to generate states contingent on cooperation and fairness, a competence-based task to probe feelings related to self-perception and skill, and a navigation task to tap into the dynamic emotions of goal-directed behavior. We adopted the momentary reporting design but instead replaced the 1 dimensional happiness self-report with a high-dimensional emotion selection. Throughout the experiment, participants were presented with a screen displaying a wide array of emotion categories, allowing them to select ones that best matched their current affective experience. Here, we hope to capture a more granular affective states through a variety of cognitive tasks and complement them with richer, high-dimensional self-reported data.

5.1.1 Probabilistic Reward Task

The first task in the battery is designed to elicit core outcome-based emotions, such as joy, excitement, disappointment, and frustration. These affective states are tightly coupled to the reward processing. We therefore selected a probabilistic reward task, which is a foundational paradigm in affective and computational neuroscience. Most notable of this type of paradigm is the one implemented by Rutledge and colleagues [Rutledge et al., 2014] as we described before. These task allows us to precisely manipulate the key drivers of these emotions: reward magnitude, risk, reward expectations, and reward prediction error.

The core affective mechanism in this paradigm is the generation of trial-by-trial reward expectation and prediction errors. While previous work has robustly linked RPEs and reward expectations to a single dimension of momentary happiness, these computational signals are also thought to drive more granular affective experiences. A positive RPE can elicit happiness or joy [Rutledge et al., 2014]. Conversely, negative outcomes or negative RPEs are expected to elicit distinct negative states. Disappointment arises from the appraisal of an expected reward failing to materialize, whereas frustration is linked to the appraisal of being blocked from a goal[Pekrun, 2006].

5.1.2 Game Theory Paradigm

The second task targets a set of complex social emotions, including anger, guilt, gratitude, and trust. These states are fundamentally different from simple reward-based feelings because they are contingent on the perceived intentions and actions of others, which is why a social setup is necessary. In probabilistic reward tasks, outcomes are determined by chance, but in these paradigms, the outcome is jointly determined by the participant's own actions and the choices of a partner. This social contingency, which allows for the quantitative study of fairness, trust, and cooperation, is the essential ingredient required to elicit affective states related to interpersonal experiences.

Remorse and gratitude might be elicited using the Prisoner's Dilemma, where participants must choose between mutually beneficial cooperation or self-interested behavior. Because outcomes are contingent on inferences about a partner's intentions, these games are exceptionally good for eliciting guilt from betraying a cooperator or gratitude from reciprocated trust [Ketelaar and Tung Au, 2003, Ashlock and Rogers, 2008]. Anger and disappointment might be elicited through violation of social norms. For example, in the Ultimatum Game, one player proposes a monetary split that another can accept or reject. Participants are known to reject unfair splits, and they may feel anger toward a partner if consistent unfair splits are offered [Burnell et al., 1999, Gilam et al., 2019]. Neuroimaging investigation on social cooperation also showed that mutual cooperation is associated with increased reward circuitry activities [Rilling et al., 2002].

5.1.3 Competence-based Task

The third task is designed to induce emotional states related to self-evaluative emotions, such as anxiety, frustration, pride, and shame. These states are not necessarily tied to external rewards, but to an internal appraisal of one's own performance and

competence, particularly in a context of social evaluation. To elicit them, we employ a competence-based task. The principle of using evaluative threat to induce affect is well-established, most notably in the Trier Social Stress Task [Kirschbaum et al., 1993], which uses social-evaluative pressure to reliably induce stress. While seemingly neutral, performance-based tasks are potent elicitors of emotions because they engage these constructs of self-esteem and social evaluation.

The anticipation and execution of a math problem under pressure is a well-documented method for inducing anxiety and stress. "Math anxiety" describes a state of stress and worry about failure in the context of solving math problems [Ashcraft, 2002, Richardson and Suinn, 1972]. This anxious state is theorized to consume crucial working memory resources, which in turn impairs performance and creates a vicious cycle of anxiety and poor outcomes [Ashcraft, 2002]. Successfully solving a difficult math problem is usually attributed to one's own competence and effort, can elicit pride [Pekrun, 2006, Weiner, 1985]. On the other hand, a failure in doing so might be coupled with frustration and shame. Frustration arises from the appraisal of being blocked from a goal [Pekrun, 2006]. Shame, a more intense self-evaluative emotion, can be elicited by repeated failures that are attributed internally to a lack of ability or competence [Weiner, 1985].

5.1.4 Navigation Task

Spatial navigation tasks provide a powerful paradigm for studying the affective dynamics of motivated, goal-directed behavior. In these tasks, participants navigate a complex space from a starting point toward a goal. One can model this process with a RL framework, such that each step a participant inches closer to the goal generates a positive RPEs, while setbacks like heading down a path towards a dead end generate negative RPEs. Navigation tasks can also be framed within a goal-directed behavior framework as a sequence of actions and the cognitive process begins with the anticipation of a goal. Making clear progress is associated with positive anticipatory emotions like hope and excitement, while encountering obstacles induces frustration and anxiety. Consummatory reward at the goal elicits joy or relief upon success, or disappointment upon failure. Furthermore, this type of tasks can also illuminate the bidirectional relationship between affective states and goal appraisal. With the objective appraisal of one's progress changing affective states, the current affective state in turn biases how that progress is appraised.

For example, a participant in a positive affective state may appraise a small step forward as significant, boosting their mood further and engage in explorative behavior,

which might lead to discovery of more rewards. Conversely, a participant in a negative state may appraise the same step as trivial and interpret a minor setback as a catastrophic failure, leading to a downward spiral of frustration and inaction. The navigation task thus provides a useful environment for studying this self-reinforcing feedback loop, capturing the moment-to-moment coupling between goal pursuit, cognition, and affective states.

Together, these paradigms are designed to elicit a broad and diverse range of human affective experiences within a single, controlled experimental session. This chapter details the specific implementation of each task and presents empirical data from an online validation study. The central aim is to demonstrate that this task battery is an effective tool for eliciting the wide and granular range of emotional states necessary for the fine-grained study of human affect.

5.2 Methods

5.2.1 Ethical Approval

Experimental protocol was approved by the University College London Research Ethics Committee (Approval: 16639/001)

5.2.2 Participants

We used the Prolific platform for online participant recruitment [Prolific Inc, 2023]. Three separate studies were advertised on the Prolific platform to residents of the United Kingdoms who were at least 18 years old and had access to a desktop computer. In total, we obtained a total of 165 informed consents and 129 completed the experiment.

Inclusion / Exclusion We excluded 19 participants who had more than 50% missing task responses. We also excluded 1 participant who had 50% missing emotion selection data. Total of 108 participants' data were used in data analysis.

5.2.3 Procedure

Eligible individuals were invited to our study website to first review the informed consent. Participants must select checkboxes for each of the inclusion / exclusion criteria

item-by-item. They then were asked to make a quick response using keyboard and mouse to ensure both input devices were available. Afterwards, participants were given instructions of the four tasks as well as the emotion selection grid. They were then given a chance to a practice trial of each of the tasks. Afterwards, all participants were asked to engage in trials of the four tasks. The trials were interleaved such that no consecutive trials were the same task. The order of the trials for all participants were pre-determined and randomized. Each participant participated in 8 trials for each of the four tasks, totaling 32 trials.

5.2.4 Trial Strcuture and Emotion Selection

We introduce a battery of cognitive tasks that consists of four components: Gamble, Trust, Math and Maze (for an overview, see Table 5.1). Our goal is to fully cover the emotion options that we used in our accompanying examination on emotional state decodability and generalization. There are three components that the tasks all shared: 1) experimental timing, 2) harmonized feedback and 3) post-trial emotion selection (see 5.1 for details).

Task	Type	Task Descriptions
Gamble	Probabilistic Reward Task	Choose between safe and gamble option
Trust	Game Theory Paradigms	Solve Prisoner's Dilemma with an artificial agent
Math	Competence-based Task	Solve math equations
Maze	Navigation Tasks	Solve mazes to reach destinations

Table 5.1: **Task Overview**

5.2.5 Emotion Selection

For the emotion selection, participants were given a grid of 28 options to report their current emotional state. The options are: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, surprise. Neutral is included here to ensure that participants were not forced to select any emotions. Multiple selection was allowed. Participants were allowed to make selection using either keyboard or mouse.

During practice, participants also completed an additional mini game that was designed to familiarize participants with the layout of the grid of emotions. Participants

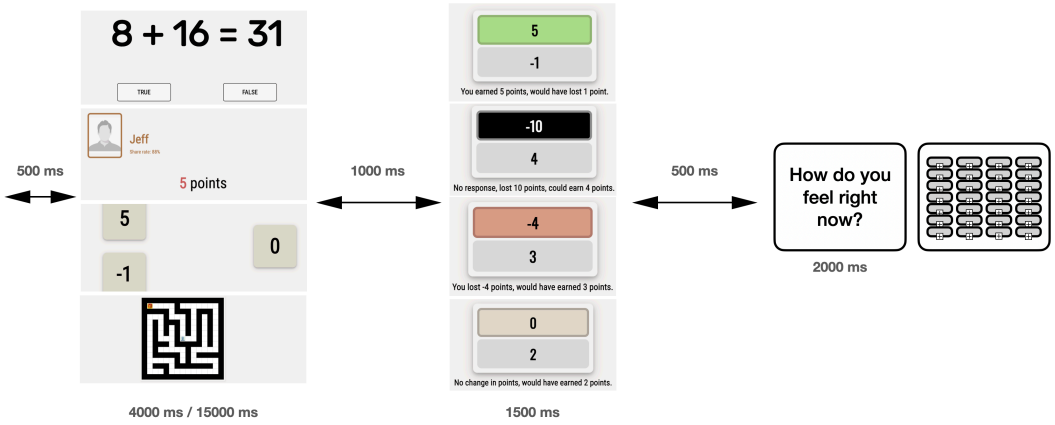


Figure 5.1: **Overall Experimental Design** All tasks start with 500 ms of blank screen, followed by a maximum of 4000 ms (or 15000 ms for Maze) stimuli display time, when participants can make responses. The trial terminates when an acceptable response is provided or the maximum time is reached. 1000 ms of feedback anticipatory period follows the trial termination, and the actual feedback is shown for 1500 ms. The feedbacks of all the tasks are harmonized such that no semantic information was presented (the black texts were for illustration purposes only, now shown during the actual trials). The feedback has two parts: top actual feedback and bottom counterfactual feedback. The top part is the actual feedback, where the background color corresponds to feedback type: green indicates winning points, red for losing points, beige indicates no points changed, and black for no valid responses. The counterfactual feedback shows what the outcome would have been if the participant had chosen the opposite actions (see each task subsection below for detail). All task feedback were harmonized to be between -10 to 10 points. After the feedback and 500 ms of blank screen, a mandatory 2000 ms introspection period was shown with the text "How do you feel right now?" The participants were then given a grid of alphabetized emotions to report their emotions. Participants had at most 20000 ms to report their emotional states.

were asked to find and select two or three emotions in the grid at once. If they correctly selected the emotions, they were allowed to move on or they were asked to deselect the incorrect emotions and re-select the correct ones. Participants were asked to do this repeatedly until all emotions on the grid were sampled at least once.

5.2.6 Gamble task

In each of the Gamble task trials, participants were given two options: gamble and safe (for an illustration, see figure 5.2). The gamble option had two outcomes stack vertically and the safe option had one outcome. The probability of winning the gamble is always set to 50% and participants were made aware of that. Therefore, there should be no element of uncertainty about win probabilities. The expected values of both the gambling and safe options are set to be the same across all trials and conditions. Participants learned the outcome of their choice immediately after and we provide counterfactual outcome as well, that is they learned whether they won the gamble if they chose to gamble or if they would have won the gamble if they chose the safe option.

We expected joy and excitement to be elicited by positive gambling outcomes, while annoyance, disapproval, and sadness to be elicited by negative outcomes. With the counterfactual outcome, we expect participants to report relief and remorse. Further, we included an action shift in random trials, during which participants' chosen action was switched toward the other. We expect this component to elicit surprise, disapproval, disappointment and confusion.

5.2.7 Math task

In each of the Math task trials, participants were shown a math equation that involves addition, subtraction and multiplication between two numbers (for an illustration, see figure 5.3). We set the numbers to always be two digits and the first digits were never 1 (i.e. the numbers ranged from 21 to 99). The participants were asked to determine if the equation is true or false by using pre-trained keys on the keyboard. The participants won more points for answering correctly or lost more points for answering incorrectly if the trails had multiplication, compared to addition or subtractions. The counterfactual feedback here would be the points they would have lost or won for their counterfactual action. For example, if a participant answered a multiplication correctly, they would have seen a green top box with 10 points and -5 points for counterfactual. Conversely, if a participant answered an addition equation incorrectly, they would have seen a red top box with -10 points and 5 points for counterfactual.

Figure 5.2: **The Gamble Task** Participants were given two options: gamble or safe. Gamble options had two outcomes stacked vertically and safe option had one outcome.

We expect pride, joy, excitement to be elicited when participants answered correctly, while confusion, disappointment, embarrassment, realization to be elicited for incorrect trials.

5.2.8 Trust task

In a Trust task trial, participants were given two options: share or keep as well as information about the partner they are playing with (for an illustration, see figure 5.4). In line with the classic Prisoner's Dilemma game, when both chose to cooperate (share), they gain double the points, and when only one chose to share while the other chose to keep, the keeper gets points but not the sharer. The optimal strategy for this game is always to not cooperate. They were also informed about the partners that they were playing with during practice. We informed the participants that they would be playing against AI models that were trained with real human data to mimic unique human play styles. After participation, the participants were debriefed on that all partner behaviors were pre-determined and part of the task contingencies. There were good or bad partners: good partners have higher probability to share while bad partners have lower. During the practice, the participants were given a chance to play the game unlimited times to try and figure out which partner (identified by name and color) was good or

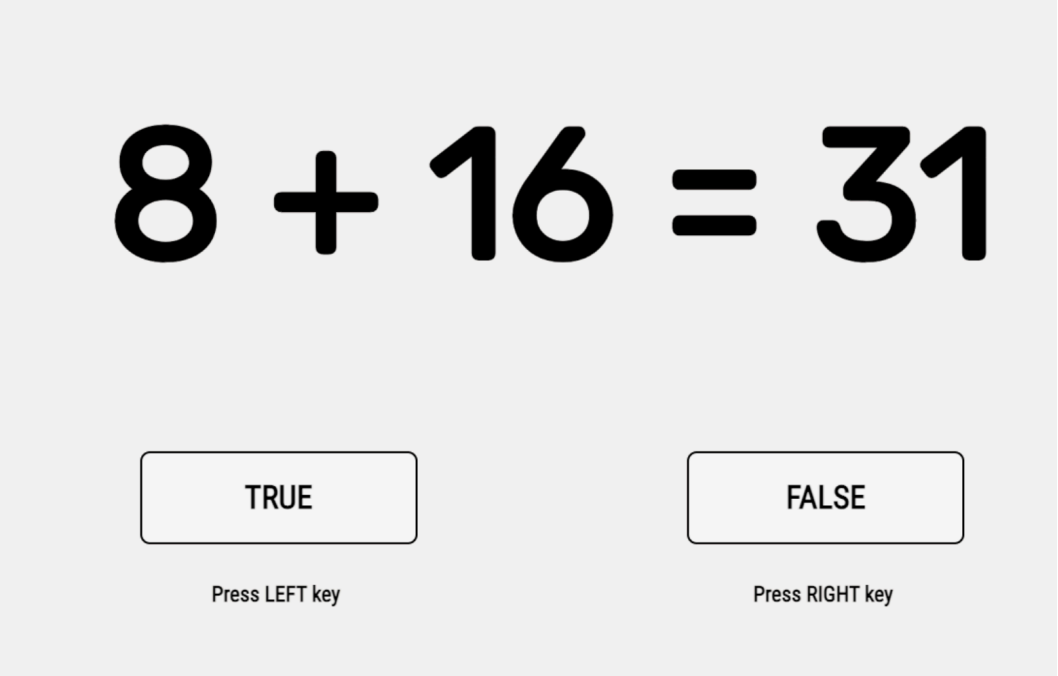


Figure 5.3: **The Math Task** Participants were shown a math equation and two options: True or False. Participants will use the options to indicate the correctness of the math equation.

bad. During the actual task, partner share probability was shown to assist the participants' decision (good partner will approach 80% share rate while a bad partner around 20%). If the participants chose to keep, they can learn about what the partner would have done if they had shared through the counterfactual feedback. Occasionally, the partners would "act" uncharacteristically to further elicit more emotional categories.

We expected the participants to report approval, caring, joy, love, relief for when they and the partner both chose to share. Anger, disapproval, surprise, grief, realization, remorse were expected for when good partners decided to keep when the participants shared. For bad partners, they might report disgust, anger, disapproval and annoyance, and gratitude, joy, surprise, relief during unexpected share.

5.2.9 Maze task

In the Maze task trials, participants were placed at the center of a maze (start), along with a gift box (goal) randomly located in the maze (for an illustration, see figure 5.5). The maze can be easy or hard depending on the number of correct steps needed to reach the goal from the start. Participants will see different mazes throughout the ex-

Figure 5.4: **The Trust Task** Participants were shown points at stake and two options: share or keep. Top left showed partner information, including name, share percentage and unique color.

periments and we designed the maze such that they were generally equal in difficulty. The participants were told that the goal can be good or bad: they would need to approach the good goals and avoid the bad goals. The task has two major tricks: 1) participants did not know whether the goal is good or bad until the end of the trial; and 2) participants must navigate the maze "blindly," meaning they needed to key in all actions without seeing the movements within the allocated time. At the end of the trial, participants were presented with their actual movement and revelation of the goal. The participants earned more points if their final distance to the good goals was shorter or longer for the bad goals.

We expected participants to feel relief, joy, excitement for when they correctly predicted the nature of the goal and completed appropriate actions, or anger, disappointment, sadness for mismatch between the goal and actions. Because the participants key in actions "blindly," they might also report confusion, annoyance, anger, disapproval, realization, embarrassment if the movement did not matched what they had planned.

5.2.10 Analysis

We performed three main quantitative analyses to validate the task battery. All analyses used participant identity as a random effect to account for subject-level variance, and all statistical tests were conducted with a significance level of $\alpha = 0.05$.

First, to demonstrate that participants were actively engaged and reporting affective states, we calculated the proportion of neutral-only self-reports for each participant (relative to all of their emotional trials). We then ran a one-sample t-test on this distribution of proportions, testing against a null hypothesis of 0.5 (random chance).

Second, to test whether the task type (Gamble, Math, Trust, Maze) significantly predicted which emotions were reported, we ran a series of Mixed-Effect Logistic Regression models. A separate model was fit for each of the 28 emotion categories. All models predicts the emotion response (binarized for each unique emotion) with task type as main effect participant identity as mixed effect. The significance of the Task variable for each model was determined using an ANOVA (Type III Wald Chi-square test). To control for these 28 separate tests, the resulting p-values were corrected for multiple comparisons using the Benjamini-Hochberg False Discovery Rate (FDR) method.

Third, to test whether the emotional response to an outcome was dependent on the task, we ran a second series of Mixed Effect Logistic Regression Models. These models added an interaction term of task and feedback type, including only positive or negative feedback (no response trials were excluded). The p-values for the interaction term

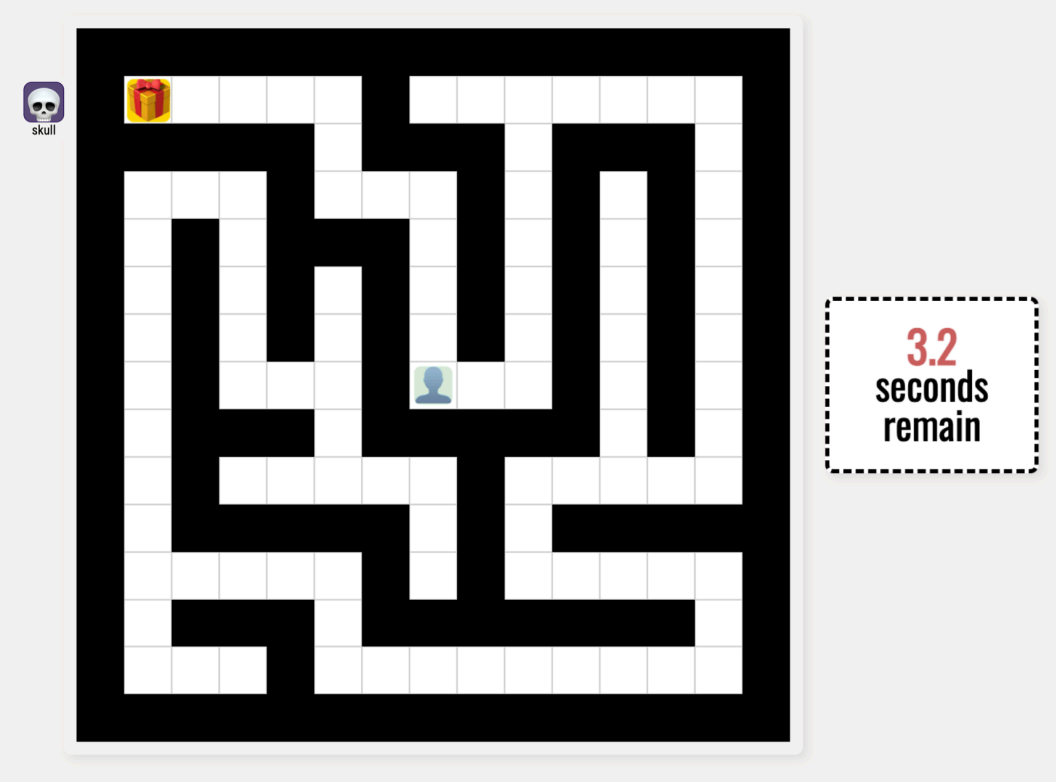


Figure 5.5: **The Maze Task** Participants were placed at the center of a maze, along with a gift box, which is the goal. Right middle box showed the reminder time allowed for movement. At the end, the goal might changed to a skull indicating that the goal was bad. In such case, participants earned more points if their final distance to the goal was longer.

were similarly corrected for multiple comparisons using FDR.

5.3 Results

We first confirmed that participants reported specific emotions far more often than 'Neutral'. On average, participants selected only neutral on 13.1% of the trials. A one-sample t-test confirmed this proportion was significantly lower than the 0.5 chance level ($t(107) = -18.55$, $p < 0.001$, $d = -1.79$), demonstrating that the tasks reliably elicited specific affective states.

5.3.1 The Main Effect of Task

The ANOVA analysis confirmed that the task context had a significant effect on the majority of reported emotions, as shown in Table 5.2. After FDR correction, 16 of the 28 emotion categories showed a significant main effect of Task ($p < 0.05$). This result quantitatively rejects the hypothesis that the tasks are interchangeable and confirms they elicit distinct emotions (visualized in Figure 5.6).The astronomical F_3 and χ^2 values for 'Caring', 'Disappointment', and 'Optimism' are statistical artifacts of complete separation. This occurs when an emotion is reported 0 times in one or more tasks.

5.3.2 Pairwise Task Comparisons

To decompose main effect of task, we ran pairwise post-hoc comparisons for the significant emotions, with results displayed in Table 5.3. This analysis allows us to pinpoint what emotion is being elicited by what task.

Emotion	Comparison	T-value	dof	Cohen's d	$p_corrected$	Sig.
Anger						
	Gamble vs. Math	1.015	1290.752	0.056	0.414	
	Gamble vs. Maze	-2.594	1459.903	-0.133	0.025	*
	Gamble vs. Trust	-0.111	1346.123	-0.006	0.932	
	Math vs. Maze	-3.700	1449.814	-0.186	0.001	*
	Math vs. Trust	-1.153	1377.433	-0.062	0.347	
	Maze vs. Trust	2.527	1509.404	0.128	0.029	*
Annoyance						
	Gamble vs. Math	3.521	1254.338	0.194	0.002	*

	Gamble vs. Maze	-3.081	1440.617	-0.160	0.007	*
	Gamble vs. Trust	4.157	1228.252	0.228	< 0.001	*
	Math vs. Maze	-6.928	1471.666	-0.350	< 0.001	*
	Math vs. Trust	0.636	1384.313	0.034	0.646	
	Maze vs. Trust	7.642	1458.884	0.384	< 0.001	*
Approval						
	Gamble vs. Math	-3.091	1210.437	-0.168	0.007	*
	Gamble vs. Maze	-1.228	1454.648	-0.063	0.315	
	Gamble vs. Trust	-2.979	1259.398	-0.159	0.009	*
	Math vs. Maze	2.032	1305.315	0.107	0.080	
	Math vs. Trust	0.151	1389.016	0.008	0.928	
	Maze vs. Trust	-1.900	1362.642	-0.099	0.101	
Caring						
	Gamble vs. Math	-1.302	1037.772	-0.070	0.281	
	Gamble vs. Maze	0.160	1308.633	0.009	0.928	
	Gamble vs. Trust	-4.034	797.478	-0.211	< 0.001	*
	Math vs. Maze	1.457	926.288	0.080	0.221	
	Math vs. Trust	-3.117	1013.166	-0.165	0.007	*
	Maze vs. Trust	-4.130	764.841	-0.226	< 0.001	*
Confusion						
	Gamble vs. Math	2.760	1224.118	0.152	0.016	*
	Gamble vs. Maze	-2.946	1456.537	-0.152	0.010	*
	Gamble vs. Trust	2.270	1256.402	0.124	0.050	*
	Math vs. Maze	-5.935	1414.361	-0.297	< 0.001	*
	Math vs. Trust	-0.526	1391.411	-0.028	0.710	
	Maze vs. Trust	5.427	1446.332	0.272	< 0.001	*
Disappointment						
	Gamble vs. Math	2.908	1231.523	0.160	0.010	*
	Gamble vs. Maze	-1.439	1430.711	-0.075	0.224	
	Gamble vs. Trust	1.534	1302.128	0.084	0.197	
	Math vs. Maze	-4.579	1467.026	-0.231	< 0.001	*
	Math vs. Trust	-1.435	1380.730	-0.077	0.224	
	Maze vs. Trust	3.108	1514.975	0.158	0.007	*
Disgust						
	Gamble vs. Math	-1.070	1253.212	-0.058	0.391	
	Gamble vs. Maze	-3.225	1227.022	-0.158	0.006	*

	Gamble vs. Trust	-3.033	1041.549	-0.160	0.008	*
	Math vs. Maze	-2.191	1422.107	-0.110	0.058	
	Math vs. Trust	-2.069	1233.602	-0.110	0.076	
	Maze vs. Trust	0.023	1492.247	0.001	0.982	
Embarrassment						
	Gamble vs. Math	-3.619	1035.690	-0.195	0.002	*
	Gamble vs. Maze	-3.593	1289.111	-0.177	0.002	*
	Gamble vs. Trust	0.166	1330.683	0.009	0.928	
	Math vs. Maze	0.288	1436.987	0.015	0.853	
	Math vs. Trust	3.792	990.849	0.205	< 0.001	*
	Maze vs. Trust	3.789	1242.906	0.187	< 0.001	*
Gratitude						
	Gamble vs. Math	0.368	1319.538	0.020	0.796	
	Gamble vs. Maze	2.947	1088.415	0.162	0.010	*
	Gamble vs. Trust	-1.926	1342.968	-0.104	0.098	
	Math vs. Maze	2.643	1199.702	0.141	0.022	*
	Math vs. Trust	-2.322	1345.537	-0.124	0.044	*
	Maze vs. Trust	-4.920	1080.944	-0.262	< 0.001	*
Joy						
	Gamble vs. Math	-2.420	1325.754	-0.132	0.037	*
	Gamble vs. Maze	-0.135	1391.843	-0.007	0.932	
	Gamble vs. Trust	-3.111	1348.140	-0.168	0.007	*
	Math vs. Maze	2.405	1383.474	0.126	0.037	*
	Math vs. Trust	-0.676	1391.990	-0.036	0.622	
	Maze vs. Trust	-3.126	1406.361	-0.162	0.007	*
Neutral						
	Gamble vs. Math	0.501	1320.221	0.027	0.713	
	Gamble vs. Maze	5.356	1160.468	0.295	< 0.001	*
	Gamble vs. Trust	-0.375	1244.528	-0.021	0.796	
	Math vs. Maze	4.955	1256.221	0.266	< 0.001	*
	Math vs. Trust	-0.873	1259.188	-0.049	0.490	
	Maze vs. Trust	-5.618	1068.502	-0.318	< 0.001	*
Optimism						
	Gamble vs. Math	-0.841	1331.771	-0.046	0.506	
	Gamble vs. Maze	1.036	1349.660	0.055	0.406	
	Gamble vs. Trust	-0.593	1349.536	-0.032	0.664	

	Math vs. Maze	1.935	1394.983	0.101	0.098	
	Math vs. Trust	0.258	1389.265	0.014	0.869	
	Maze vs. Trust	-1.687	1444.967	-0.087	0.152	
Pride						
	Gamble vs. Math	-5.106	1011.296	-0.275	< 0.001	*
	Gamble vs. Maze	-4.037	1327.907	-0.200	< 0.001	*
	Gamble vs. Trust	-1.705	1301.174	-0.092	0.149	
	Math vs. Maze	1.485	1369.357	0.078	0.213	
	Math vs. Trust	3.618	1182.884	0.195	0.002	*
	Maze vs. Trust	2.342	1493.466	0.118	0.043	*
Relief						
	Gamble vs. Math	-0.976	1331.997	-0.053	0.427	
	Gamble vs. Maze	2.057	1287.071	0.110	0.077	
	Gamble vs. Trust	-0.368	1348.115	-0.020	0.796	
	Math vs. Maze	3.106	1321.720	0.164	0.007	*
	Math vs. Trust	0.624	1385.685	0.033	0.647	
	Maze vs. Trust	-2.497	1394.111	-0.130	0.030	*
Remorse						
	Gamble vs. Math	0.989	1246.661	0.054	0.425	
	Gamble vs. Maze	-0.110	1398.732	-0.006	0.932	
	Gamble vs. Trust	-1.580	1318.704	-0.085	0.186	
	Math vs. Maze	-1.170	1493.209	-0.059	0.342	
	Math vs. Trust	-2.573	1196.140	-0.137	0.026	*
	Maze vs. Trust	-1.539	1339.730	-0.080	0.197	
Surprise						
	Gamble vs. Math	2.150	1179.865	0.119	0.063	
	Gamble vs. Maze	0.513	1346.778	0.027	0.712	
	Gamble vs. Trust	2.246	1165.115	0.124	0.052	
	Math vs. Maze	-1.791	1488.371	-0.091	0.126	
	Math vs. Trust	0.090	1388.929	0.005	0.938	
	Maze vs. Trust	1.896	1496.115	0.096	0.101	

Table 5.3: **Pair-wise comparisons for emotions with significant task main effects.** All comparisons between a pair of tasks for each of the emotions were showed here. Significant results were found for all emotions except for Optimism and Surprise, showing that tasks selectively elicited wide range of emotions.

5.3.3 Task-by-Feedback Interaction

Finally, we demonstrated that the emotional response to winning or losing was dependent on the task (see Table 5.4 for ANOVA results). After FDR correction, 11 of the 28 emotion categories showed a significant task by feedback type interaction (for a visualization, see Figure 5.7).

5.4 Discussion

The primary aim of this chapter was to develop and validate a battery of cognitive tasks as an effective tool for eliciting a diverse and granular range of emotional states. The results from our statistical analyses provide strong quantitative support for that our task battery is a robust tool for eliciting emotional responses.

We observed clear quantitive evidence of task specificity. The main effect analysis (Table 5.2) showed that the majority of emotions (16 of 28) were significantly dependent on the task. The post-hoc comparisons (Table 5.3) allowed us to verify our a priori hypotheses from the introduction with statistical rigor.

Broadly speaking, the results align with theoretical expectations. For instance, the Trust Task, as intended, was a primary elicitor of social emotions such as disgust, disappointment, grief, caring and gratitude. In contrast, the Math Task successfully induced self-evaluative states like pride, joy, and approval, while the Gamble Task reliably drove outcome-based feelings of curiosity, joy and sadness. The Maze task induced fear, confusion, sadness and embarrassment. This alignment between task design and reported emotion provides confidence that the battery is functioning as intended.

Perhaps the most important finding of this chapter is the significant task by feedback interaction (Table 5.4). This quantitatively demonstrates that the emotional response to a positive or negative feedback is task-dependent. A negative outcome in the gamble

Emotions	F_3	χ^2	$p_{corrected}$	Significant?
Admiration	0.043	0.129	0.988	
Amusement	0.502	1.506	0.706	
Anger	6.151	18.453	< 0.001	*
Annoyance	24.729	74.187	< 0.001	*
Approval	4.984	14.952	0.004	*
Caring	2021028.300	6063084.900	< 0.001	*
Confusion	19.685	59.055	< 0.001	*
Curiosity	1.506	4.518	0.268	
Desire	1.691	5.073	0.233	
Disappointment	192840.887	578522.661	< 0.001	*
Disapproval	2.620	7.860	0.076	
Disgust	5.816	17.448	0.001	*
Embarrassment	9.846	29.538	< 0.001	*
Excitement	0.910	2.730	0.488	
Fear	1.605	4.815	0.248	
Gratitude	6.790	20.370	< 0.001	*
Grief	0.581	1.743	0.676	
Joy	4.215	12.645	0.011	*
Love	1.162	3.486	0.392	
Nervousness	2.856	8.568	0.059	
Neutral	25.370	76.110	< 0.001	*
Optimism	157524.038	472572.114	< 0.001	*
Pride	14.648	43.944	< 0.001	*
Realization	1.855	5.565	0.199	
Relief	5.620	16.860	0.002	*
Remorse	3.230	9.690	0.037	*
Sadness	0.993	2.979	0.461	
Surprise	3.316	9.948	0.036	*

Table 5.2: **ANOVA showed Tasks selectively elicited wide range of emotions.** The table showed the F-ratio, χ^2 statistic, and FDR-corrected p-value for the main effect of Task on the log-odds of each emotion being reported. Each emotion was modeled separately. The astronomical F ratio and χ^2 values for Caring, Disappointment, and Optimism are statistical artifacts of complete separation, which is when one or more levels of the Task variable perfectly predict a zero outcome.

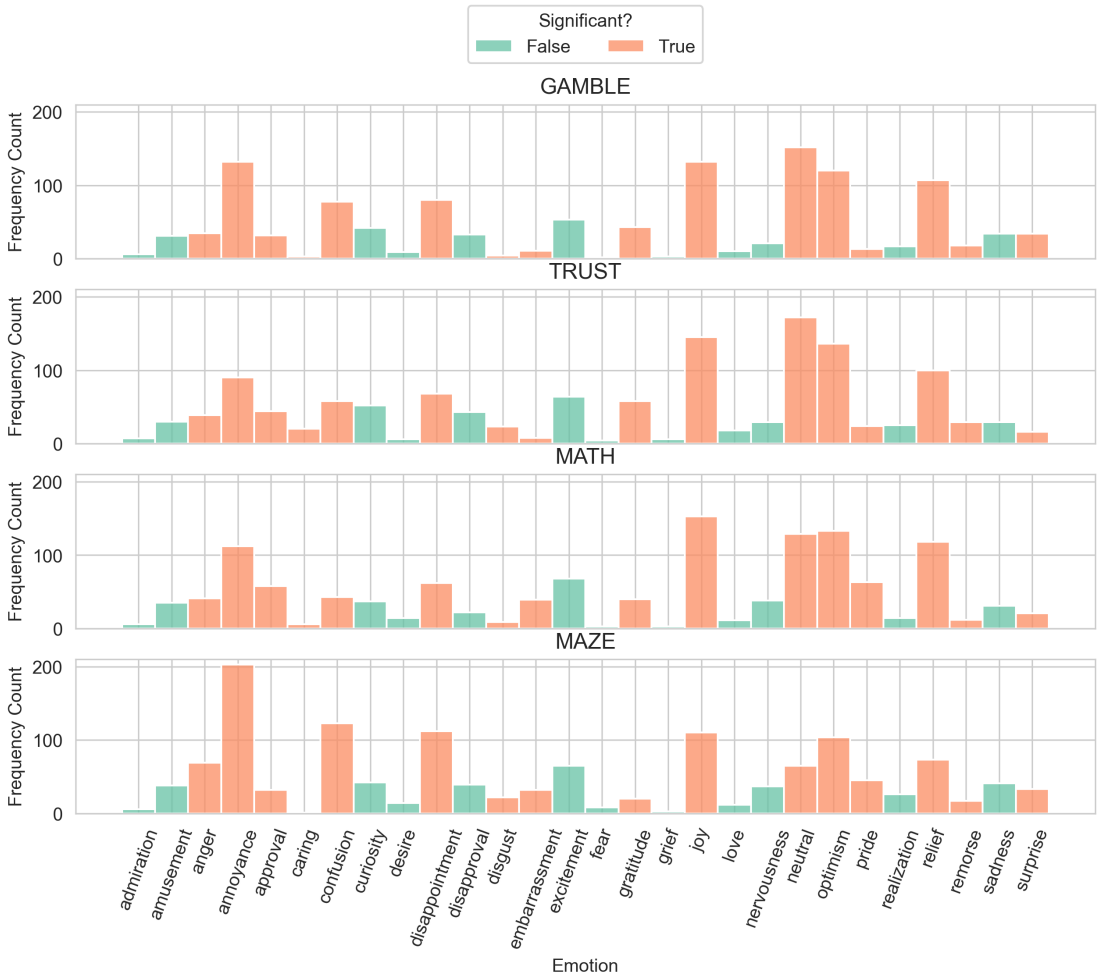


Figure 5.6: **Emotion Responses showed selectivity for task type.** Each panel represent a unique tasks. X axis showed different emotions, y axis showed frequency count of this emotion. Color denotes whether task type showed significant effect on emotion as determined by the ANOVA analysis.

Emotions	F_3	χ^2	$p_{corrected}$	Significant?	
Admiration	0.093	0.372	e	0.985	
Amusement	0.596	2.384	e	0.810	
Anger	7.236	28.944	e	< 0.001	*
Annoyance	12.452	49.808	e	< 0.001	*
Approval	1.559	6.236	e	0.300	
Caring	1.963	7.852	e	0.209	
Confusion	4.100	16.400	e	0.008	*
Curiosity	0.901	3.604	e	0.647	
Desire	0.622	2.488	e	0.810	
Disappointment	4.068	16.272	e	0.008	*
Disapproval	8.007	32.028	e	< 0.001	*
Disgust	1.237	4.948	e	0.455	
Embarrassment	0.289	1.156	e	0.985	
Excitement	3.024	12.096	e	0.042	*
Fear	0.238	0.952	e	0.985	
Gratitude	1.813	7.252	e	0.230	
Grief	0.096	0.384	e	0.985	
Joy	1309041.951	5236167.804	e	< 0.001	*
Love	0.123	0.492	e	0.985	
Nervousness	3.128	12.512	e	0.039	*
Neutral	4.561	18.244	e	0.004	*
Optimism	2.180	8.720	e	0.160	
Pride	0.943	3.772	e	0.645	
Realization	0.645	2.580	e	0.810	
Relief	5.132	20.528	e	0.002	*
Remorse	1.710	6.840	e	0.253	
Sadness	1.858	7.432	e	0.230	
Surprise	5.151	20.604	e	0.002	*

Table 5.4: **Emotional response to feedback is task-specific.** The table showed the F-ratio, χ^2 statistic, and FDR-corrected p-value for the interaction of task type and feedback type on the log-odds of each emotion being reported. Each emotion was modeled separately.

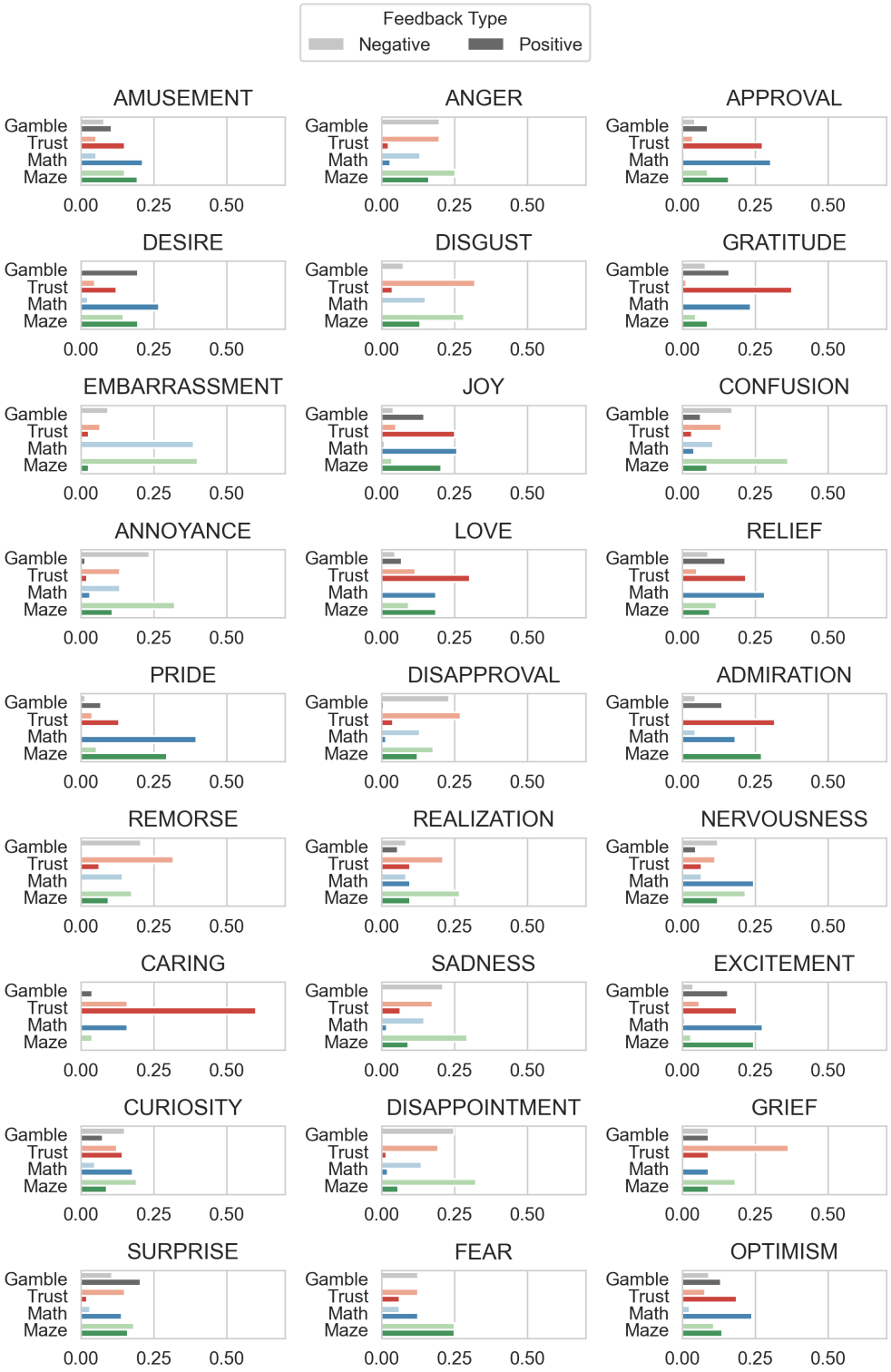


Figure 5.7: **Task and Feedback type influence emotion reports** Each panel represent a unique emotions. X axis showed percentage of emotion reported, y axis showed task types and color denotes feedback types. Lighter colors are for negative feedback trials and more solid colors are for positive feedback trials.

task might not elicit the same emotional responses as a negative outcome in the trust task. As task dependence is a specific aspect of context dependence, these finding directly supports the broader hypothesis of this thesis: a simple, one-dimensional label is insufficient to capture the high-dimensional emotional responses elicited by cognitive tasks.

The central aim here was to demonstrate that the battery of tasks is a suitable tool for eliciting a wide range of affective states in a context fundamentally different from a passive, narrative-driven paradigm. This goal is analogous to many neuroimaging or psychophysiology studies where a task's primary value is not in its standalone behavioral outcome [Demidenko et al., 2021, Zhang et al., 2013], but in its proven ability to reliably generate significant neuroimaging support.

In conclusion, the results of this validation study are successful by this standard. The qualitative findings clearly demonstrate that the task battery achieves its primary objective: it elicits a broad and granular spectrum of affective states, with clear specificity between the different task domains. We are therefore confident that this battery is an effective and robust tool for inducing the rich affective data needed for our subsequent investigations.

Chapter 6

Decoding and generalization of LLM-quantified emotional states from MEG signals

6.1 Introduction

Initially, research into affective representations in the brain followed localizationist views, attempting to identify distinct links between neuroanatomical regions and affective experiences. This perspective yielded important associations, such as that of the amygdala with fear [LeDoux, 2002] and the insula with disgust [Phillips et al., 1997]. This framework, however, was an oversimplification of how affective experience might be represented in the brain. It failed to account for the functional heterogeneity of brain regions [Poldrack, 2010] and the distributed neural representation of emotional processing [Pessoa, 2017]. For example, amygdala activities were associated with not only fear but also the processing of positive stimuli [Sergerie et al., 2010]. Recent research shifted toward a network-based view that posits the neural representation of affective experience to be the dynamic interaction between distributed brain systems [Barrett, 2017b, Lindquist et al., 2012].

A significant body of research has successfully employed machine learning classifiers to decode affective states from high-dimensional neural data. By training models on multivariate patterns of brain activity, numerous studies have demonstrated that it is possible to reliably distinguish between neural responses to stimuli of different affective experience, such as happy versus sad faces [Kragel and LaBar, 2013] or pleasant versus unpleasant sounds [Ethofer et al., 2009]. The applicability of this decoding approach

has been extended to naturalistic paradigms, including the classification of emotional experiences during movie viewing [Ke et al., 2025]. Collectively, this work has robustly established that, within a specific experimental context, distinct emotional states possess neurally decodable signatures.

Our experimental work in Chapter 4 established that LLM-quantified emotional states derived from narratives were associated with distinct MEG patterns across time and frequency domains. However, a critical limitation of these findings, both in the broader literature on affective decoding and in our own work, is whether the observed neural consistency generalizes beyond the specific context in which it was elicited. The predictive success of most affective decoders is circumscribed by the induction paradigm, and even the specific stimuli. A critical benchmark for the validity and utility of any objective quantification of affective experience is to demonstrate the capacity to generalize beyond its training paradigm. That is, a successful decoder of emotional states trained on narrative listening must show above-chance performance in decoding emotional states elicited by out-of-sample narratives and non-linguistic experiences. An emotional decoder that fails to generalize functions merely as a context-specific pattern classifier.

The study presented in this chapter was designed to test the cross-context generalizability of our LLM-based quantification of emotional states. We examined the generalization performance of decoders trained to identify LLM-quantified emotional states from the story-listening context by testing their performance on held-out stories (an out-of-sample test) and on self-reported emotional states during a battery of cognitive tasks (a cross-modal test). These validation tests directly examined whether our LLM-quantified emotional states capture a context-independent neural signature. Above-chance generalization would provide evidence for a generalizable neural representation of emotional states, and further validating our quantification approach. However, a failure to sustain above chance performance would support a constructionist perspective [Barrett, 2017b, Lindquist et al., 2012], suggesting that the neural representation of emotional states might be context-dependent and flexibly constructed from domain-general networks.

6.2 Methods

6.2.1 Ethical approval

The experimental protocol was approved by the University College London Research Ethics Committee (Approval: 27121/001)

6.2.2 Participants and procedure

We recruited participants from the local community (N = 13) via online advertisements on the UCL SONA platform. Inclusion criteria were confirmations by participants that they were 1) healthy, 2) over 18 years old, 3) native English speakers, 4) free of tattoos above their chest area, 5) not visually impaired (or corrected to normal vision), and 6) free of metal in their body. Written informed consent was obtained before any study procedures were undertaken.

Inclusion / Exclusion Four participants were excluded for the following reasons: one due to intoxication, one due to excessive noise caused by metal retainer, one due to excessive movement, and one due to lack of trigger signal.

6.2.3 Procedure

To balance participant fatigue and data consistency, we split the overall data collection into three 2-hour sessions. In each session, participants were asked to listen to six different stories and complete two task blocks (see Figure 6.1 for exact ordering). Breaks were offered right before each task block. The content of the stories is detailed in table 3.1. Each task block consisted of modified versions of the four tasks described in chapter 5. Trick conditions in the *Math* and *Trust* tasks were removed due to poor consistency found in the online study (detailed in chapter 5).

MEG data were acquired as in chapter 4 with a CTF MEG system equipped with a whole-head SQUID magnetometer with 275 channels, located in the Department of Imaging Neuroscience at the University College London. MEG data were collected with a sample rate of 1200Hz.

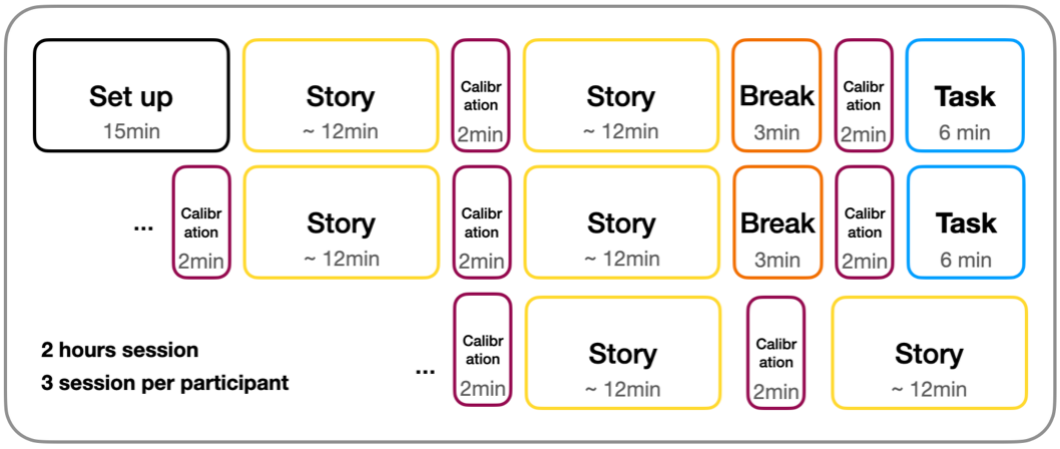


Figure 6.1: **Study Design** Schematic of one the three experimental sessions. Each 2-hour session included one setup period (black), six story-listening blocks (yellow), two cognitive task blocks (blue), two optional breaks (orange).

6.2.4 MEG Data Processing

We implemented the same MEG processing pipeline detailed in section 4.2.5. To summarize, we preprocessed the continuous data by band-pass filtering (0.5-200 Hz), applying notch filters (50 Hz and harmonics), and downsampling the data to 400 Hz. Artifacts related to eye blinks and cardiac activity were identified and removed using Independent Component Analysis (ICA). Word-level event timings were precisely synchronized to the MEG signal by computing the cross-correlation between the original stimulus file and the audio signal recorded with the MEG. Using these synchronized word timings, the cleaned data was then segmented into 2000ms epochs to match the word-level emotional labels. Depending on the specific analysis, we created either *overlapping* epochs (for event-related raw signal analysis) or *non-overlapping* sequential epochs (for PSD analysis). Finally, PSD was computed for non-overlapping epochs using the multi-taper method and averaged into six frequency bands (delta, theta, alpha, beta, gamma, and high gamma).

6.2.5 Emotional State Labels

An emotional state label for each word was inferred as detailed in chapter 3. Briefly, this consisted of a three-step process relying on a Large Language Model. To balance emotion granularity and the number of available labels per class, we chose 10 emotional states as our classification targets for training and validating the neural decoders (See Figure 4.1 for state labels and their emotion composition). A higher number of

emotional states only yielded minimal model performance increase (as showed in Figure 3.6).

6.2.6 Neural decoding models

To decode emotional state labels with MEG, we constructed logistic regression models that can be expressed in this general form, where w denotes words, c for different channels, and t denotes time after word onset (for MEG signal models) or frequency bands (for spectral models):

$$E_{w,t} \sim \beta_0 + \beta_1 x_{1,w,t} + ... + \beta_c x_{c,w,t} \tag{6.1}$$

where $E_{w,t}$ is the emotional states label assigned to word w at time t after word onset. We used β_0 to model the average of the MEG data across all channels. β_c is the regression coefficient parameter for channel c that our models fits. $x_{c,w,t}$ is the c channel MEG signal for a given word c and time after onset t. We used concatenated story data from two sessions (total of 12 stories) and used them as the primary training dataset. The third session was intentionally left-out for future work on validation and optimization of our method, which is beyond the scope of this thesis.

6.2.7 Measuring Decoder Performance

The study design incorporates two distinct emotion induction modalities: story-based and task-based, and within story-based induction, multiple narratives were included. To evaluate decoder performance when exposed to varying levels of context, we computed three ROC AUC (Area Under the Receiver Operating Characteristic Curve) scores, each corresponding to a different contextual exposure: **In-Context**, **Out-of-Sample**, and **Cross-Modal** scores (for an overview of the different levels, see Figure 6.2. These conditions reflect a reduction in the contextual information through removal of induction modality and content-specific context in the test set. We constructed multinomial logistic regression decoders for both 3- and 10-state emotional states classification. To evaluate the decoders' performance, we used the weighted one-vs-rest ROC AUC, where the AUC is calculated for each class against all others and then weighted by that class's support in the dataset. We assessed decoder performance using two validation methods as follows:

In-context Performance Epochs from all 12 stories were concatenated, shuffled and then split into five folds. For each fold, we trained on four folds of the data and tested the performance on the remaining data. With this 5-fold-cross-validation within stories, the decoders can exploit on neural representations of emotions that are induction modality-specific, content-specific, or the combination of the two. Therefore, the in-context scores reflects an estimate of the upper-bound of the decoders' ability to track emotional state-related neural representations.

Out-of-sample Performance With a leave-one-story-out approach, in each iteration, one of the 12 stories was held out for testing while the remaining stories were used for training. While the induction modality remained consistent (story), the narrative content of the test story was unseen during training. Therefore, the out-of-sample scores measure the decoders' ability to generalize across novel content-specific context within the same induction modality (story-listening).

Cross-modal Performance To evaluate the decoders' content-independent generalization and cross-modal performance, decoders trained on story data were applied to neural signal during the introspection phase in the task-based induction. First, a binarized correlation map was constructed (as shown in Figure 3.7) between model-inferred emotional states and the 27 self-reported emotion categories, using a threshold that ensured each emotion category corresponded to at least one of the LLM-quantified emotional states. Participants' self-reported emotion categories were then mapped onto emotional states based on this correspondence. This mapping from 27 self-reported categories to our 10 LLM-states resulted in a many-to-one mapping. Therefore, in cases where an emotion category aligned with multiple states, a prediction was considered correct if it matched any of the associated states. The cross-modal scores reflect the decoders' ability to transfer across both induction modality and content, representing the most abstract level of generalization.

Significance tests To determine if, on average, a decoder configuration achieved significantly above-chance performance, we took all the scores from all participants for that configuration and conducted a one-sample t-test against a theoretical chance level. The chance level for ROC AUC is 0.5.

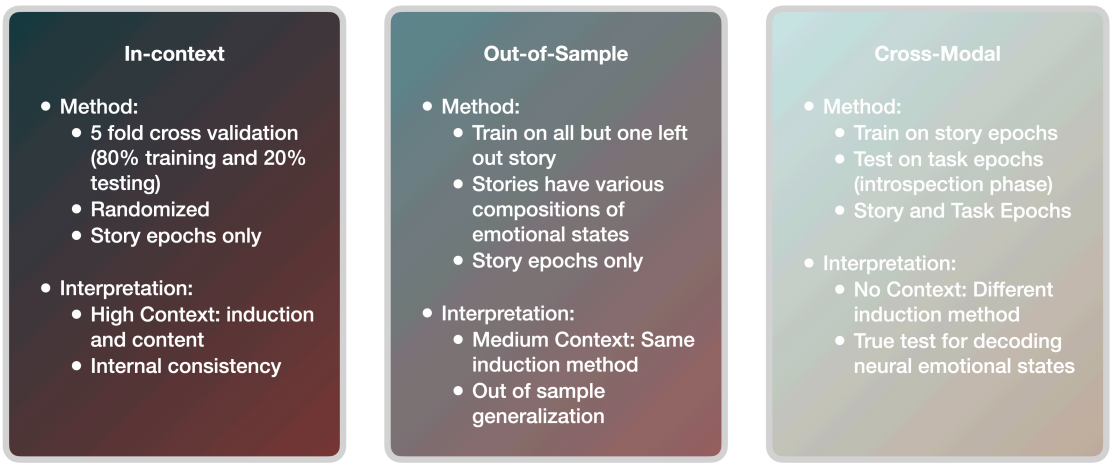


Figure 6.2: **Overview of the three context levels.** Each panels summarized the different levels of the context exposure.

6.3 Results

6.3.1 Multi-class Spectral Topography Decoders

The 10-state decoder achieved significantly above-chance in-context scores and, as expected, a lower but still above-chance out-of-sample score. However, the cross-modal score for the 10-state decoder was not significantly different from chance. For the 3-state decoder, the in-context score was significantly *below* chance, while both the out-of-sample and cross-modal scores were at chance (Figure 6.3). We also calculated top-2 accuracy for the 10 states to further demonstrate the decreasing performance as the generalization demand increased (in-context $>$ out-of-sample $>$ cross-modal; Figure 6.4).

6.3.2 Single Band Multi-class Spectral Topography Decoders

To examine which frequency bands contributed most to emotion classification, we constructed separate decoders using only the topography from a single power band. While decoders trained on each individual power band performed significantly above chance (Figure 6.5), we found no significant differences in performance between the bands.

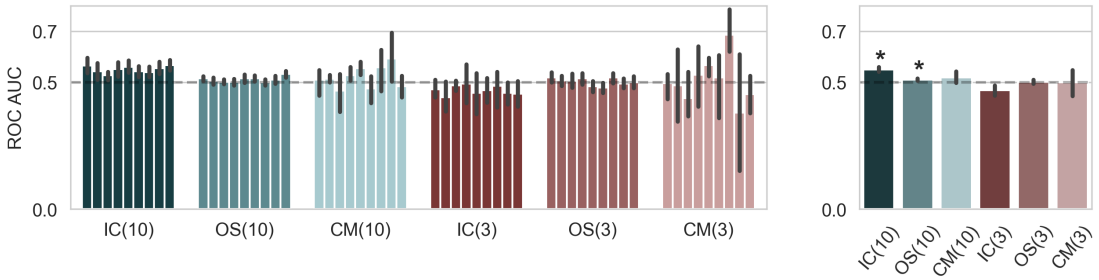


Figure 6.3: **Multi-class decoders performance across emotional state configurations and validation methods.** On average, we found above-chance performance for In-context scores for 10 states [IC (10); 95% c.i. 0.541-0.560, t_{43}=10.536, $p < 0.01$], Out-of-sample score for 10 states [OS (10); 95% c.i. 0.506-0.515, t_{43}=4.298, $p < 0.01$]. However, In-context scores for 3 states [IC (3); 95% c.i. 0.447-0.489, t_{43}=-3.068, $p = 0.007$] showed significantly below chance performance and no different than chance performance for OS scores for 3 states [OS (3); 95% c.i. 0.494-0.510, t_{43}=0.471, $p = 0.766$]. Neither of the CM scores for 10 states [CM (10); 95% c.i. 0.494-0.544, t_{43}=1.525, $p = 0.206$] or 3 states were significantly above chance [CM (3); 95% c.i. 0.446-0.554, t_{43}=0.001, $p = 0.999$]. All p values are adjusted for multiple comparison. We calculated 95% confidence interval for each mean estimation and illustrated as error bars. Chance level illustrated as grey dotted line (at 0.5).

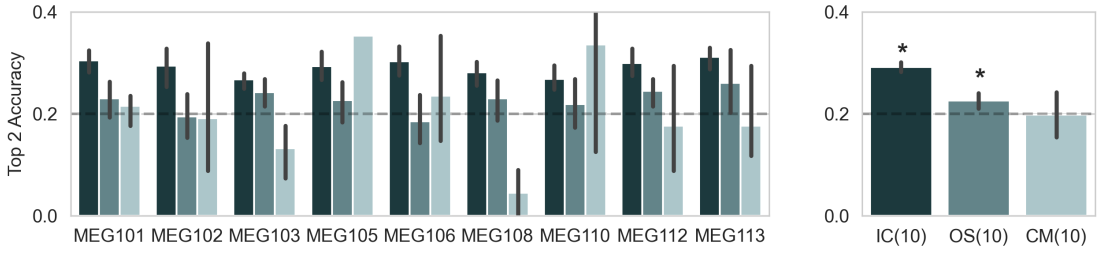


Figure 6.4: **Multi-class decoders' performance measured by Top 2 Accuracy** Only 10-state performances are shown, as top-2 classification for 3 states is meaningless. As expected, performance followed a similar pattern to ROC AUC, decreasing as the generalization demands increased (IC > OS > CM) [IC: 95% c.i. 0.281-0.301, t_{43}=18.079, $p < 0.01$; OS: 95% c.i. 0.211-0.241, t_{43}=3.445, $p < 0.01$; CM: 95% c.i. 0.150-0.246, t_{43}=-0.076, $p = 0.530$]. All p values are adjusted for multiple comparison. We calculated 95% confidence interval for each mean estimation and illustrated as error bars. Chance level illustrated as grey dotted line (at 0.2).

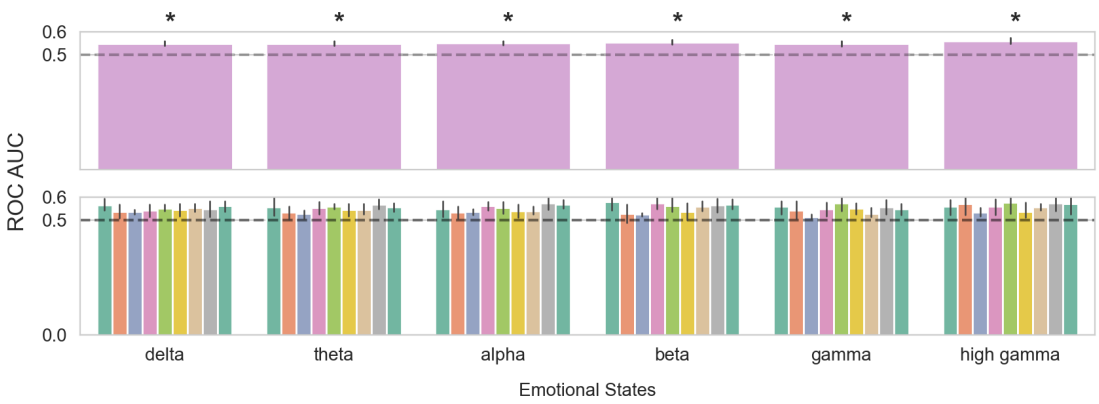


Figure 6.5: **Multi-class decoders performance across powerband configuration** On average, we found above chance performance for all power bands [Delta band (0.5-4Hz): 95% c.i. 0.540-0.558, t_{43}=10.754, $p < 0.01$; Theta (4-8Hz): 95% c.i. 0.540-0.559, t_{43}=10.435, $p < 0.01$; Alpha (8-12Hz): 95% c.i. 0.541-0.560, t_{43}=10.623, $p < 0.01$; Beta (12-30Hz): 95% c.i. 0.543-0.567, t_{43}=9.309, $p < 0.01$; Gamma (30-45Hz): 95% c.i. 0.536-0.558, t_{43}=8.372, $p < 0.01$; High Gamma (55-100Hz): 95% c.i. 0.547-0.573, t_{43}=9.269, $p < 0.01$].

6.3.3 Binary One-vs-Rest Spectral Topography Decoders

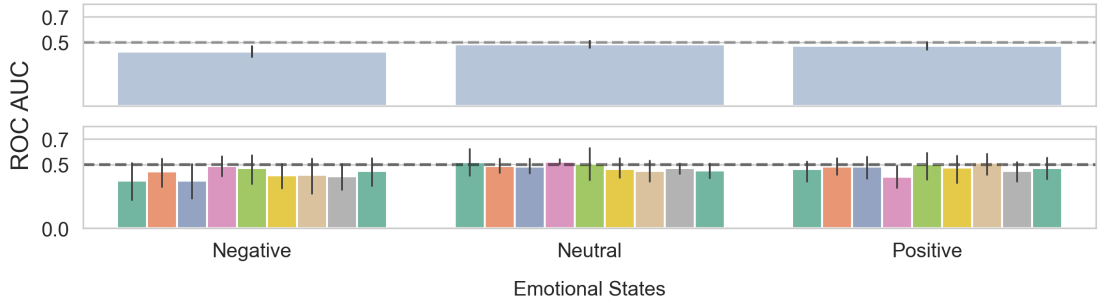
In-context Performance

The 5-fold cross-validation results on binary spectral decoders that decode individual emotional states showed that for 3 states, no decoders were significantly above chance (see Figure 6.6a). However, we found that for 10 states, decoders for Amusement (95% c.i. 0.533-0.622, t_{43}=3.496, $p < 0.01$), Anger (95% c.i. 0.514-0.604, t_{43}=2.649, p = 0.01), Gratitude (95% c.i. 0.510-0.563, t_{43}=2.729, $p < 0.01$), Negative (95% c.i. 0.513-0.590, t_{43}=2.670, p= 0.01), had significantly above chance performance (Figure 6.6b).

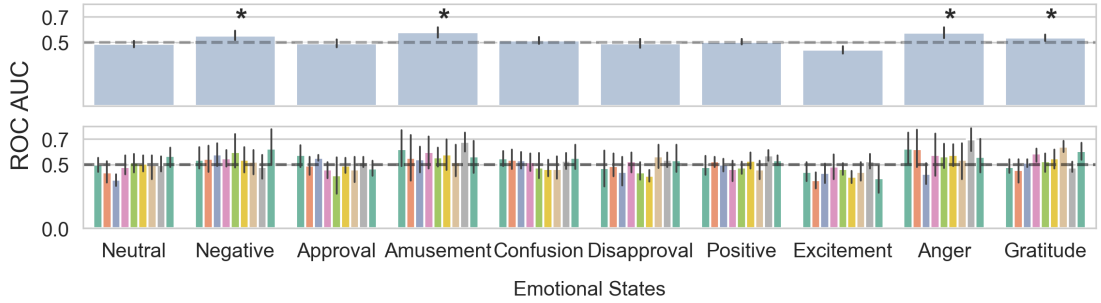
Out-of-sample Performance

To further validate the decoders, we examine their performance when we withhold one story as testing data and train on the rest of the stories. For 10 emotional states, We found that performance decreases across emotional states, and no decoders yielded significantly above chance performance (see Figure 6.7).

However, because the stories cover a wide range of themes and are thus emotionally diverse, the number of samples for each emotion category varies between stories. It appeared that as the number of samples increased in the held-out story, the mean ROC



(a)



(b)

Figure 6.6: **In-context Performance for One-vs-Rest Binary Decoders** Panel (a) showed the performance for 3 emotional states and Panel (b) showed the performance for 10 emotional states. In both panels, top figure showed the cross validated decoder performance (measured by ROC AUC) of each of the 10 emotional states. Error bars indicate 95% confidence intervals. Bottom figure demonstrated further split the decoders' performance by individuals. Each unique color represents a unique participant's decoders.

AUCs became more stable. Conversely, ROC AUC became somewhat unstable in stories with low sample sizes (Figure 6.8).

To account for these potential outliers, we calculated the minimal number of samples required to achieve a stable ROC AUC for each emotional state. We then re-evaluated performance, excluding test folds where the number of positive samples for a given emotion fell below this stability threshold. After this adjustment, the decoders for Negative (95% c.i. 0.508-0.552, t_{43}=2.77, $p < 0.01$) and Approval (95% c.i. 0.513-0.542, t_{43}=3.848, $p < 0.01$) were found to perform significantly above chance (Figure 6.9).

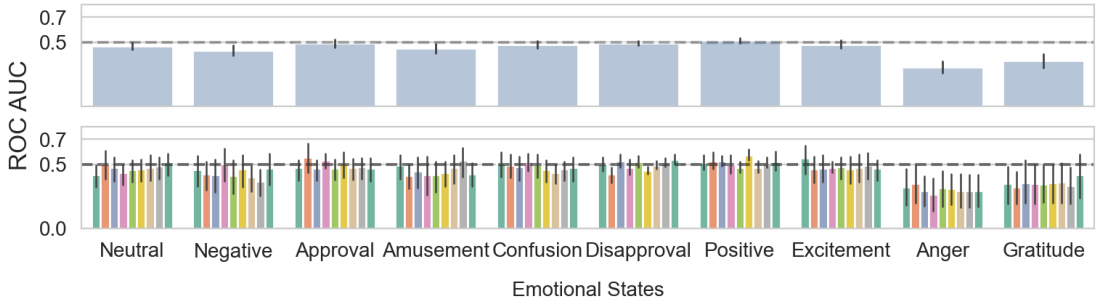


Figure 6.7: **Out-of-sample Performance for One-vs-Rest Binary Decoders** Top figure showed the averaged ROC AUC across 10 emotional states. Bottom figure showed the ROC AUC across emotional states and split by participant.

Cross-modal Performance

Generalization to out-of-sample stories was limited to a subset of emotions. We next tested the decoders' ability to generalize across induction modalities. Surprisingly, the binary decoder for the "Positive" emotional state performed significantly above chance (95% c.i. 0.540-0.646, t_{43}=3.542, $p < 0.01$, p value corrected for multiple comparison.) when classifying self-reported emotional states during the task introspection phase (Figure 6.10).

6.4 Discussion

The experimental study described in this chapter tested the context dependence of neural emotion representations by training decoders on story-listening data and evaluating their ability to generalize to novel stories and a different cognitive task. Our findings revealed a graded decline in performance as the demands for generalization increased. Decoders performed best when tested within the story-listening context (In-Context Performance), but performance dropped when generalizing to novel stories (Out-of-Sample Performance) and reduced to chance when crossing the induction modality to the cognitive task (Cross-Modal Performance). To examine specifically how well the decoders can decode individual emotional states, we constructed binary one-vs-rest decoders and found that the decoders for the labels Negative, Amusement, Anger and Gratitude decoders performed above chance in-context. Out-of-sample, binary decoders for the labels Negative and Approval were able to successfully generalize to novel stories. Finally, the decoder for the Positive label was the only decoder to suc-

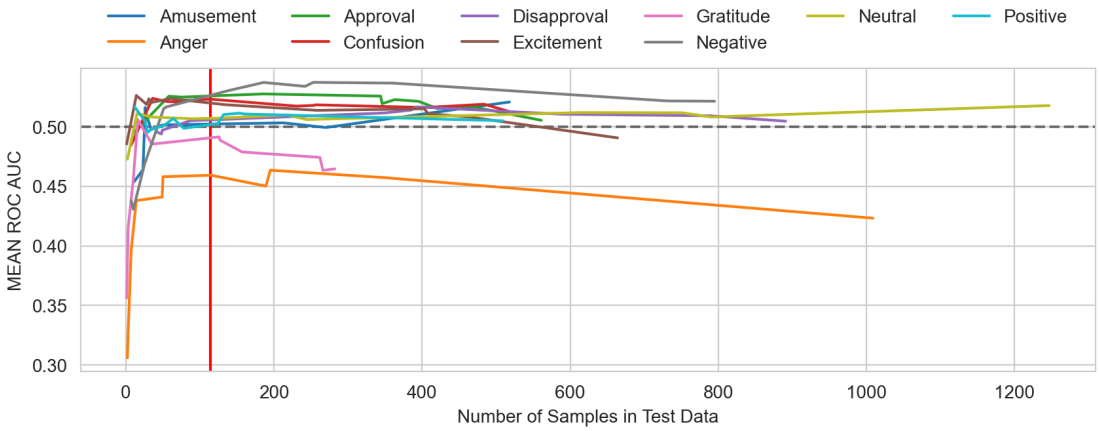


Figure 6.8: **Mean ROC AUC by the Number of Sample in Test Story** Mean ROC AUC changes drastically when the number of sample in test data is low. Red line denotes the average number of sample (114.1) that is required to stabilize ROC AUC across emotional states.

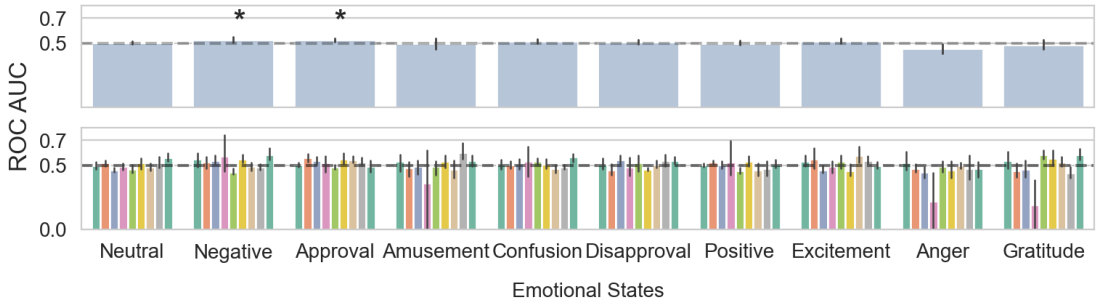


Figure 6.9: **Leave-One-Story-Out validation Performance After Outlier Removal for One-vs-Rest Binary Decoders** Compared to Figure 6.7, Negative and Approval decoders are significantly above the chance level.

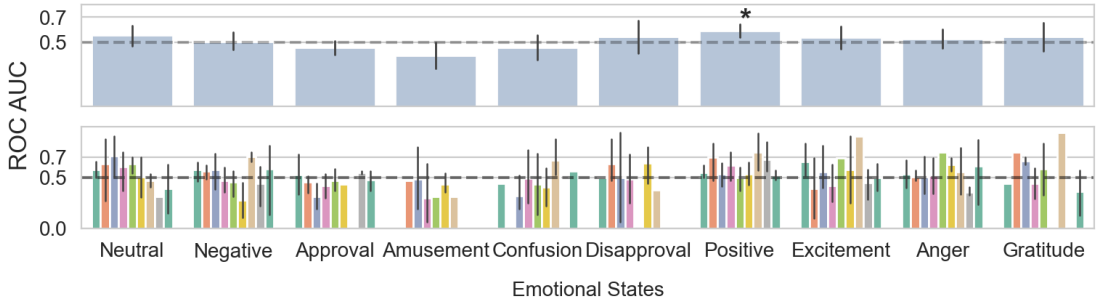


Figure 6.10: **Cross-modal Performance for One-vs-Rest Binary Decoders** The figure shows the ROC AUC scores for each of the 10 binary decoders when trained on story data and tested on data from the cognitive task blocks. Error bars indicate 95% confidence intervals. Only the decoder for the "Positive" state performed significantly above the chance level of 0.5.

cessfully generalize across modalities to the task data.

The decline in performance illustrates the generalizability problem as predicted by the constructionist view of emotion. Under this view, the neural representations of affective states are not innately coded or localized but are contextually constructed from the dynamic interaction of domain-general brain networks that support other cognitive functions [Barrett, 2017b, Lindquist et al., 2012, Pessoa, 2017]. Therefore, these neural representations are content- and modality-dependent: the brain activity corresponding to an emotion like sadness may differ substantially if it is elicited by a narrative versus a monetary loss [Siegel et al., 2018a].

The results are congruent with a broader literature demonstrating the context-dependency of neural emotion patterns. The failure of the decoders to generalize across modalities is particularly in alignment with fMRI studies showing that classifiers trained to decode emotions from one sensory modality struggle to generalize to another [Wallenwein et al., 2024]. While some evidence exists for supramodal emotion representations, a large portion of the neural response is specific to the sensory modality in which the emotion was elicited [Peelen et al., 2010]. Our results showed that the story driven emotional states were not recognizable in the task data. This is representing a primary obstacle on toward objective quantification of emotional state with neural data.

We did identify one successful generalization of the binary decoder to the task setting: the decoder trained on Positive labels in the stories showed significantly above-chance performance for task generalization. This might be seen as evidence for a generalizable

neural substrate for core affective valence. However, this interpretation must be treated with great caution, as it is directly contradicted by the failure of the 3-state multi-class and binary decoder even for the in-context settings. The 3-state configuration of decoders relied on distinguishing between positive, negative, and neutral states and performed at or below chance levels in all performance tests. This inconsistency suggests that the few successful findings are likely not robust enough. While statistically significant, such isolated, conflicting and close to chance level results have limited utility for any real-world application, which would demand far greater reliability. Therefore, the main outcome of this study is not that generalization is possible in a few select cases, but rather a broader demonstration of the graded decline in performance with increasing generalization. The primary contribution of this work is hence to shed light on the difficulty of finding a truly context-independent neural signature for emotion. The results reinforce the conclusion that the generalizability problem for emotional state decoders remains a fundamental, unsolved challenge.

Several limitations must be considered for this experimental work. First, we lack the "ground truth" during story listening. With this chapter, we opt to test the validity of our LLM-quantified labels (as described in chapter 3). While we provided empirical evidence on neural consistency within these LLM-quantified emotional states, they remained a proxy for true felt experience. Our results will be greatly benefited by validating the LLM-quantified emotional states against the subjective reports. Additionally, we can also compare the neural consistency between LLM-quantified and subjective reported emotional states. However, obtaining unbiased and uninterrupted subjective report during naturalistic story listening is methodologically challenging. While it is not impossible but we estimated that the resource demands for adding such measure exceeded the scope of this thesis. Future work can certainly improve upon our design by creatively include subjective reports during story listening such as having participants speak how they are feeling out loud. Second, our modest sample size (N=9) limits the statistical power and broader generalizability of our conclusions, which warrant replication in a larger cohort. Third, the outlier removal procedure for the out-of-sample analysis, while statistically justified, was a post-hoc correction that highlights the fragility of these generalizable signals. Finally, we conducted our analysis in the sensor space of MEG signal which used limited spatial resolution. Therefore, our findings reflect coarse cortical-level patterns, and it is entirely possible that there are more generalized brain pattern below the cortical surface that's consistent with our LLM-quantified emotional states. Therefore, future work with more advanced source localization methods or with more spatially oriented methods like fMRI might be needed for a more precise understanding of the neural representation of affective experiences.

In conclusion, this chapter provides an important test of an approach towards objective quantification of emotional states, and suggests that neural representations of emotional states might show substantial context-dependence. While decodable signatures of emotion were comparatively robust within training contexts, they were not easily transferable to new contexts. Overall, the findings are compatible with constructive accounts of emotion. As such, objective quantification of affective states may need to account for specifics of the induction modalities. Joy in a task is not quite the same as joy in a story. Tentatively, these findings might also be in keeping with our hypotheses that affective states are metareasoning states: joy in a task implies different behavior and attentional requirements than the same "emotion" while listening to a story. According to a metareasoning account the two should indeed differ meaningfully.

Chapter 7

Discussion and concluding remarks

This dissertation began with a fundamental challenge in affective science: the difficulty of objectively measuring affective states. These states are crucial, shaping our thoughts, decisions, and well-being, yet they are ephemeral, high-dimensional, and phenomenally complex. The gold standard, subjective self-report, is reactive and burdensome. Objective measures of physiology and behavior lack specificity, while neural measures have been mostly context-dependent.

Conceptually, we adopted the view of affective states as computational heuristics for metareasoning. Methodologically, we hypothesized that the computational structure of Large Language Models (LLMs), specifically their ability to model long-range contextual dependencies, could serve as a proxy for the temporal structure of affective states. The central goal was to test if this LLM-based framework could produce an objective, high-dimensional quantification of emotion that was fast, passive, non-interruptive, and, most importantly, generalizable across contexts.

7.1 Summary of Findings

The empirical work of this thesis progressed in a logical arc, from establishing a basic principle further linking computational cost in decision-making and affective state to developing and validating a novel quantification pipeline.

In Chapter 2, we demonstrated that cognitive effort, a form of computational cost, negatively impacts momentary self-reported happiness. This finding add to the growing evidences for the thesis's metareasoning framework, providing evidence that affective states track the computational cost of our actions. However, this study was limited to a single dimension of affective state (happiness or valence in general) and relied on

the very self-report methodology we sought to move beyond. We also further demonstrated the ineffectiveness of the single dimension measure of affective state, reiterating the need for a more granular, objective and passive method to quantify affective states.

In Chapter 3, we illustrated the first core methodological contribution of this thesis: a novel pipeline for objectively quantifying high-dimensional emotional states from text with LLM. This three-step process used a LLM to generate predictive continuations, a fine-tuned classifier to extract their emotional content, and a Hidden Markov Model to identify discrete, persistent states.

In chapter 4, the central finding was that these LLM-quantified states exhibited neural consistency. We found distinct patterns of MEG activity, in both the time and frequency domains, correspond to different emotional states quantified by the LLM pipeline. This showed that an LLM pipeline may potentially generate emotion quantifications that are both just semantically coherent and also neurally valid, at least within the context of story listening.

In Chapter 5, we built the necessary testbed for the second, more difficult part of our investigation: generalization. To test if the LLM-quantified emotional states captured an abstract, context-independent representation of emotional state, we needed a non-linguistic elicitation setting to test for generalization. This chapter detailed the design and validation of a battery of four cognitive tasks. Using high-dimensional self-report emotion selection, we confirmed that this battery successfully elicited a wide and granular range of affective states, from joy and pride to gratitude and anger, creating a rich dataset of self-reported emotions in a task-based context.

In Chapter 6, finally, we trained MEG decoders on the LLM-quantified emotional state labels from the story data (Chapter 3) and evaluated their performance on held-out stories and, most importantly, on the self-reported emotion data from the cognitive tasks (Chapter 5). The results revealed a clear gradient of generalization. The In-Context performance, where train and test data share contextual ground, was robust. These decoders performing above chance, confirming the neural consistency findings from Chapter 4. Out-of-Sample performance, testing on novel stories, showed a drop in performance, though decoders for some states remained weakly above chance. Cross-Modal performance, the most stringent cross-induction-modality generalization test, dropped to chance level for nearly all decoders.

7.2 Impacts on Existing Literature and Future Work in Affective Neuroscience

One central finding of this thesis is the failure of cross-modal generalization. This result, while a failure of our immediate goal to find a universal decoder of affective state, provides a significant piece of puzzle in the big picture questions on the nature of emotion.

This finding lends strong support to the Psychological Constructionist view of emotion [Barrett, 2017b, Lindquist et al., 2012] and to metareasoning theories of emotion [Huys and Renz, 2017]. This theory posits that emotions are not innate, discrete entities with dedicated neural "fingerprints". Instead, an instance of emotional states is constructed in the moment from the interaction of more basic, domain-general brain networks. Our results align with this prediction. A decoder trained on "story-induced-sadness" failed to recognize "task-induced-sadness" because, from a constructionist and metareasoning perspective, they should be neurally distinct. "Story-induced-sadness" is constructed using language and auditory networks, while "task-induced-sadness" is constructed using networks for reward processing and decision-making. Our decoders, trained on the neural patterns of one context, then had no basis to recognize the patterns of the other.

However, we must be cautious about over-interpreting a null result. We cannot definitively disprove the hypothesis that a context-independent neural signature for emotion does not exist. It is possible that a subtle, universal signature does exist, but that our study lacked the statistical power (with N=9) or spatial resolution to detect it (we used sensor-space MEG signal, but perhaps more nuanced spatial signal is needed). [Peelen et al., 2010]. The dominant signal, and thus the only one our decoders could learn, was the strong, context-specific one. The isolated, weak success of the "Positive" binary decoder in the cross-modal test is a tantalizing hint that a coarser, valence-level signal might generalize, but this finding was not robust and was not in keeping with the failure of the 3-state models.

Regardless of whether a subtle, universal signal exists, the findings suggest strong context dependence. This has practical implications for any attempt to objectively quantify emotion.

This work challenge some core tenets of standard approaches in affective neuroscience and computational sentiment analysis [Zhang et al., 2020, Torres et al., 2020, Zhang et al., 2024, Chen et al., 2022, Saarimäki, 2021, Zhou et al., 2021,

Ochsner et al., 2002]. Many studies build models that perform well within a single, isolated context: classifying affective faces, decoding emotion from movie clips, or analyzing the sentiment of tweets. Our results suggest this might insufficient to claim usefulness of these decoders. A model trained to decode ”fear” from static pictures might not to a classifier of fear but rather of fear-from-static-pictures.

We propose that cross-context and cross-modal generalization should become an important benchmark for any method claiming to quantify a general, abstract affective state. A neural decoder or computational model should not be considered a valid measure of ”anger” unless it can demonstrate an ability to identify anger across diverse elicitation methodss—from listening to a story, to losing a game, to receiving unfair social feedback. Our thesis provides a direct empirical framework for conducting such a test. This generalization test is necessary to move the field from building context-specific classifiers to identifying the core, abstract features of human affective experience.

7.3 Limitations and Future Directions

This work represents an initial step, and its limitations highlight a clear path for future research.

First, the LLM-quantification pipeline (Chapter 3) by definition lacked ”ground truth”. We demonstrated neural consistency for the LLM-proxy labels, but not their direct correspondence to the participants’ “felt” experience during story listening. Collection of such measure would have been difficult. One of the key drawbacks of self-report of emotional states is that we cannot obtain continuous measure without significantly disrupting the participants. Retroactive collection is also problematic, as recall accuracy is subject to memory biases, and the knowing the story’s progression can retroactively alter a participant’s appraisal of their earlier feelings. Lastly, it is resource intensive and beyond the scope of this thesis’s work. However, future studies may seek to overcome the methodological challenge and integrate moment-to-moment subjective reports, perhaps through thinking-out-loud procedure, to validate the LLM-generated labels against subjective human experience.

Second, we were also limited by the sample size (N=9) and the spatial resolution of sensor-space MEG. Although we collected rather lengthy amount of data per participant, we were only powered to detect larger effect sizes, limiting our ability to nullify hypotheses. A larger-scale study, perhaps using fMRI or source-localized MEG, could be better in detecting more generalizable signals with the additional signal source in deeper brain structures.

Third, the LLM tools themselves are evolving rapidly. Our pipeline, built with Llama 2, could be potentially improve with the use of newer, more powerful foundation models (e.g., Gemini, Llama 3, GPT-5) that may generate more sophisticated and more brain-like predictive representations, potentially leading to more generalizable emotional state labels.

The most significant bottleneck, however, is the lack of a large-scale, benchmark dataset for affective neuroscience. The fields of computer vision and natural language processing were revolutionized by massive, public datasets like ImageNet [Deng et al., 2009] and large text corpora. Affective science needs its own equivalent.

Binz and colleagues recently established a human cognition foundation model [Binz et al., 2025]. By fine-tuning a LLM with over 10,000,000 choices from 60,000 participants, their final model was able to capture human behaviors in held-out participants better than existing cognitive models such as RL. Further, their model also demonstrated the ability to generalize to behave like human in structural task modifications and new cognitive domains. This work demonstrated that through large dataset and mapping cognitive domains to the semantic space, the LLMs model structure can effectively navigate and represent human cognition.

The work by Binz and colleagues suggests that mapping cognition to a semantic space is a powerful method for achieving generalization. This approach suggests a promising path forward, even if the constructionist view is correct. If no single, universal neural signature for a felt affective state exists across contexts, then a generalizable signal can be found for the mapping of affective states onto the semantic space. This aligns with the notion that emotional states exist in a high-dimensional, structured space rather than a fixed set of modules. The specific categories we identify—whether 27 today, or 37 or 16 tomorrow—likely reflect the common challenges the emotional states were recruited to guide. Future work could therefore investigate if a robust, cross-contextual neural signature emerges when participants linguistically describe an emotional event, regardless of how it was first elicited. This would shift the target from a pure affective signal to a semantic-affective one, providing a new way to understand how emotion is represented in the brain's conceptual and linguistic systems.

It may also be beneficial to consider creating a large multi-modal multi-context dataset for human affective experiences. This would involve recording thousands of participants engaging in a wide variety of emotion-eliciting paradigms across different contexts, including passive consumption of naturalistic stimuli, engaging in social interactions, and participating in various cognitive tasks. Their affective states would be collected in conjunction with their affective history, such as whether they were in a de-

pressive or anxious episode. Multi-modal neuroimaging data would also be assessed. With such a dataset, one could train a "foundation model of human affective experience," one that is capable of predicting affective states regardless of eliciting context.

7.4 Overall conclusion

This dissertation sought to take an initial step toward an objective quantification of affective states by framing affective states as metareasoning heuristics. We developed a novel LLM-based pipeline that successfully generated neurally consistent emotion labels within a narrative context. However, a cross-induction-method test revealed that these neural representations failed to generalize to a non-linguistic task context.

We demonstrated that the neural representation of affective states was dominated by context-specific signals, which provides empirical support for a constructionist view: the neural basis of emotion is not a localized pattern but a dynamic, context-dependent construction. The primary contributions of this thesis are therefore twofold. First, it provides a novel, neurally-validated pipeline for quantifying affective states in a context-specific manner. Second, it provides an empirical demonstration of the generalizability gradient, setting a new, more rigorous benchmark for future research that aims to bridge the gap between objective brain data and subjective affective experience.

Bibliography

[Ablin et al., 2018] Ablin, P., Cardoso, J.-F., and Gramfort, A. (2018). Faster Independent Component Analysis by Preconditioning With Hessian Approximations. *IEEE Transactions on Signal Processing*, 66(15):4040–4049.

[Ackerman and Thompson, 2017] Ackerman, R. and Thompson, V. A. (2017). Meta-Reasoning: Monitoring and Control of Thinking and Reasoning. *Trends in Cognitive Sciences*, 21(8):607–617.

[Adolphs, 2008] Adolphs, R. (2008). Fear, faces, and the human amygdala. *Current Opinion in Neurobiology*, 18(2):166–172.

[Aguinis et al., 2021] Aguinis, H., Villamor, I., and Ramani, R. S. (2021). MTurk Research: Review and Recommendations. *Journal of Management*, 47(4):823–837.

[Amazon Inc, 2023] Amazon Inc (2023). Amazon Mechanical Turk.

[American Psychiatric Association, 2013] American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-V)*. American Psychiatric Association, fifth edition edition.

[American Psychiatric Association, 2022] American Psychiatric Association (2022). *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association Publishing, dsm-5-tr edition.

[Anderson et al., 2019] Anderson, E. C., Carleton, R. N., Diefenbach, M., and Han, P. K. J. (2019). The Relationship Between Uncertainty and Affect. *Frontiers in Psychology*, 10:2504.

[Appelhans and Luecken, 2006] Appelhans, B. M. and Luecken, L. J. (2006). Heart rate variability as an index of regulated emotional responding. *Review of General Psychology*, 10(3):229–240. Place: US Publisher: Educational Publishing Foundation.

[Arora et al., 2024] Arora, Y., Kumar, A., Subhash, D. A., Raj, R., Bajpai, S., and Sharma, N. (2024). Attention on Emotions: A Vision Transformer Approach to Advancing Facial Expression Recognition. In *2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS)*, pages 1–11.

[Asghari and Bialy, 2025] Asghari, H. and Bialy, F. (2025). Triangles, Justice, and AI: Testing Large Language Models' Comprehension of Political Ideologies. *P&D - Philosophy & Digitality*, 2(1):97–114.

[Ashcraft, 2002] Ashcraft, M. H. (2002). Math Anxiety: Personal, Educational, and Cognitive Consequences. *Current Directions in Psychological Science*, 11(5):181–185.

[Ashlock and Rogers, 2008] Ashlock, D. and Rogers, N. (2008). A model of emotion in the prisoner's dilemma. In *2008 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pages 272–279.

[Ashwani et al., 2024] Ashwani, S., Hegde, K., Mannuru, N. R., Sengar, D. S., Jindal, M., Kathala, K. C. R., Banga, D., Jain, V., and Chadha, A. (2024). Cause and Effect: Can Large Language Models Truly Understand Causality? *Proceedings of the AAAI Symposium Series*, 4(1):2–9.

[Assunção et al., 2022] Assunção, G., Patrão, B., Castelo-Branco, M., and Menezes, P. (2022). An Overview of Emotion in Artificial Intelligence. *IEEE Transactions on Artificial Intelligence*, 3(6):867–886.

[Aw et al., 2011] Aw, J. M., Vasconcelos, M., and Kacelnik, A. (2011). How costs affect preferences: experiments on state dependence, hedonic state and within-trial contrast in starlings. *Animal Behaviour*, 81(6):1117–1128.

[Bagnara et al., 2025] Bagnara, L., Moeck, E. K., Kuppens, P., Bianchi, V., and Koval, P. (2025). Why do feelings persist over time in daily life? Investigating the role of emotion-regulation strategies in the process underlying emotional inertia. *Emotion (Washington, D.C.)*.

[Baltodano et al., 2018] Baltodano, S., Garcia-Mancilla, J., and Ju, W. (2018). Eliciting Driver Stress Using Naturalistic Driving Scenarios on Real Roads. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '18, pages 298–309, New York, NY, USA. Association for Computing Machinery.

[Banks et al., 2012] Banks, S. J., Bellerose, J., Douglas, D., and Jones-Gotman, M. (2012). Bilateral skin conductance responses to emotional faces. *Applied Psychophysiology and Biofeedback*, 37(3):145–152.

[Barrett, 2004] Barrett, L. F. (2004). Feelings or Words? Understanding the Content in Self-Report Ratings of Experienced Emotion. *Journal of personality and social psychology*, 87(2):266–281.

[Barrett, 2017a] Barrett, L. F. (2017a). *How emotions are made: The secret life of the brain*. How emotions are made: The secret life of the brain. Houghton Mifflin Harcourt, Boston, MA. Pages: xv, 425.

[Barrett, 2017b] Barrett, L. F. (2017b). The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12(1):1–23.

[Barrett et al., 2019] Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A., and Pollak, S. D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological science in the public interest : a journal of the American Psychological Society*, 20(1):1–68.

[Bates et al., 2015] Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48.

[Batson, 1987] Batson, C. D. (1987). Prosocial Motivation: Is it ever Truly Altruistic? In Berkowitz, L., editor, *Advances in Experimental Social Psychology*, volume 20, pages 65–122. Academic Press.

[Baucom et al., 2012] Baucom, L. B., Wedell, D. H., Wang, J., Blitzer, D. N., and Shinkareva, S. V. (2012). Decoding the neural representation of affective states. *NeuroImage*, 59(1):718–727.

[Beck et al., 2011] Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., and Erbauch, J. (2011). Beck Depression Inventory. Institution: American Psychological Association.

[Becker and Leinenger, 2011] Becker, M. W. and Leinenger, M. (2011). Attentional selection is biased toward mood-congruent stimuli. *Emotion*, 11(5):1248–1254. Place: US Publisher: American Psychological Association.

[Bell et al., 2017] Bell, I. H., Lim, M. H., Rossell, S. L., and Thomas, N. (2017). Ecological Momentary Assessment and Intervention in the Treatment of Psychotic Disorders: A Systematic Review. *Psychiatric Services*, 68(11):1172–1181. Publisher: American Psychiatric Publishing.

[Bennett et al., 2021] Bennett, D., Radulescu, A., Zorowitz, S., Felso, V., and Niv, Y. (2021). Affect-congruent attention drives changes in reward expectations. Technical report, PsyArXiv.

[Bilal et al., 2025] Bilal, A., Ebert, D., and Lin, B. (2025). LLMs for Explainable AI: A Comprehensive Survey. arXiv:2504.00125 [cs].

[Bills et al., 2023] Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., Sutskever, I., Leike, J., Wu, J., and Saunders, W. (2023). Language models can explain neurons in language models.

[Binz et al., 2025] Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., Éltető, N., Griffiths, T. L., Haridi, S., Jagadish, A. K., Ji-An, L., Kipnis, A., Kumar, S., Ludwig, T., Mathony, M., Mattar, M., Modirshanechi, A., Nath, S. S., Peterson, J. C., Rmus, M., Russek, E. M., Saanum, T., Schubert, J. A., Schulze Buschoff, L. M., Singhi, N., Sui, X., Thalmann, M., Theis, F. J., Truong, V., Udandarao, V., Voudouris, K., Wilson, R., Witte, K., Wu, S., Wulff, D. U., Xiong, H., and Schulz, E. (2025). A foundation model to predict and capture human cognition. *Nature*, pages 1–8.

[Blain and Rutledge, 2020] Blain, B. and Rutledge, R. B. (2020). Momentary subjective well-being depends on learning and not reward. *eLife*, 9:e57977.

[Bonanno et al., 2004] Bonanno, G. A., Papa, A., Lalande, K., Westphal, M., and Coifman, K. (2004). The Importance of Being Flexible: The Ability to Both Enhance and Suppress Emotional Expression Predicts Long-Term Adjustment. *Psychological Science*, 15(7):482–487. Publisher: SAGE Publications Inc.

[Botunac et al., 2024] Botunac, I., Brkić Bakarić, M., and Matetić, M. (2024). Comparing Fine-Tuning and Prompt Engineering for Multi-Class Classification in Hospitality Review Analysis. *Applied Sciences*, 14(14):6254.

[Botvinick et al., 2009] Botvinick, M. M., Huffstetler, S., and McGuire, J. T. (2009). Effort discounting in human nucleus accumbens. *Cognitive, Affective, & Behavioral Neuroscience*, 9(1):16–27.

[Boucsein, 2012] Boucsein, W. (2012). *Electrodermal Activity*. Springer US, Boston, MA.

[Bower, 1981] Bower, G. H. (1981). Mood and memory. *American Psychologist*, 36(2):129–148. Place: US Publisher: American Psychological Association.

[Bradley et al., 2008] Bradley, M. M., Miccoli, L., Escrig, M. A., and Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, 45(4):602–607.

[Brennan, 2016] Brennan, J. (2016). Naturalistic Sentence Comprehension in the Brain. *Language and Linguistics Compass*, 10(7):299–313. _eprint: https://compass.onlinelibrary.wiley.com/doi/pdf/10.1111/lnc3.12198.

[Brinkmann and Gendolla, 2008] Brinkmann, K. and Gendolla, G. H. E. (2008). Does depression interfere with effort mobilization? Effects of dysphoria and task difficulty on cardiovascular response. *Journal of Personality and Social Psychology*, 94(1):146–157.

[Brosschot and Thayer, 2003] Brosschot, J. F. and Thayer, J. F. (2003). Heart rate response is longer after negative emotions than after positive emotions. *International Journal of Psychophysiology*, 50(3):181–187.

[Brown et al., 2020] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners.

[Burnell et al., 1999] Burnell, S. J., Evans, L., and Yao, S. (1999). The Ultimatum Game: Optimal Strategies without Fairness. *Games and Economic Behavior*, 26(2):221–252.

[Cacioppo et al., 1999] Cacioppo, J. T., Gardner, W. L., and Berntson, G. G. (1999). The affect system has parallel and integrative processing components: Form follows function. *Journal of Personality and Social Psychology*, 76(5):839–855.

[Cai et al., 2021] Cai, H., Jin, Y., Liu, S., Zhang, Q., Zhang, L., Cheung, T., Balbuena, L., and Xiang, Y.-T. (2021). Prevalence of suicidal ideation and planning in patients with major depressive disorder: A meta-analysis of observation studies. *Journal of Affective Disorders*, 293:148–158.

[Callaway et al., 2022] Callaway, F., van Opheusden, B., Gul, S., Das, P., Krueger, P. M., Griffiths, T. L., and Lieder, F. (2022). Rational use of cognitive resources in human planning. *Nature Human Behaviour*, 6(8):1112–1125.

[Caucheteux et al., 2023] Caucheteux, C., Gramfort, A., and King, J.-R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, 7(3):430–441.

[Chakriswaran et al., 2019] Chakriswaran, P., Vincent, D. R., Srinivasan, K., Sharma, V., Chang, C.-Y., Reina, D. G., Chakriswaran, P., Vincent, D. R., Srinivasan, K., Sharma, V., Chang, C.-Y., and Reina, D. G. (2019). Emotion AI-Driven Sentiment Analysis: A

Survey, Future Research Directions, and Open Issues. *Applied Sciences*, 9(24). Company: Multidisciplinary Digital Publishing Institute Distributor: Multidisciplinary Digital Publishing Institute Institution: Multidisciplinary Digital Publishing Institute Label: Multidisciplinary Digital Publishing Institute Publisher: publisher.

[Chen et al., 2022] Chen, J., Ro, T., and Zhu, Z. (2022). Emotion Recognition With Audio, Video, EEG, and EMG: A Dataset and Baseline Approaches. *IEEE Access*, 10:13229–13242.

[Christopoulos et al., 2016] Christopoulos, G. I., Uy, M. A., and Yap, W. J. (2016). The Body and the Brain: Measuring Skin Conductance Responses to Understand the Emotional Experience. *Organizational Research Methods*. Publisher: SAGE PublicationsSage CA: Los Angeles, CA.

[Clauss et al., 2022] Clauss, K., Gorday, J. Y., and Bardeen, J. R. (2022). Eye tracking evidence of threat-related attentional bias in anxiety- and fear-related disorders: A systematic review and meta-analysis. *Clinical Psychology Review*, 93:102142.

[Cohn et al., 2004] Cohn, M. A., Mehl, M. R., and Pennebaker, J. W. (2004). Linguistic Markers of Psychological Change Surrounding September 11, 2001. *Psychological Science*, 15(10):687–693.

[Colombatto and Fleming, 2024] Colombatto, C. and Fleming, S. M. (2024). Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness*, 2024(1):niae013.

[Cowen and Keltner, 2017] Cowen, A. S. and Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114(38):E7900–E7909. Publisher: Proceedings of the National Academy of Sciences.

[Cunningham and Brosch, 2012] Cunningham, W. A. and Brosch, T. (2012). Motivational Salience: Amygdala Tuning From Traits, Needs, Values, and Goals. *Current Directions in Psychological Science*, 21(1):54–59.

[Curby et al., 2012] Curby, K. M., Johnson, K. J., and Tyson, A. (2012). Face to face with emotion: Holistic face processing is modulated by emotional state. *Cognition and Emotion*, 26(1):93–102. Publisher: Routledge _eprint: https://doi.org/10.1080/02699931.2011.555752.

[Dan-Glauser and Scherer, 2011] Dan-Glauser, E. S. and Scherer, K. R. (2011). The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. *Behavior Research Methods*, 43(2):468–477.

[Darwin, 1872] Darwin, C. (1872). *The expression of the emotions in man and animals, 3rd ed*. The expression of the emotions in man and animals, 3rd ed. Oxford University Press, New York, NY, US. Pages: xxxvi, 472.

[Davidson, 1998] Davidson, R. J. (1998). Affective Style and Affective Disorders: Perspectives from Affective Neuroscience. *Cognition and Emotion*, 12(3):307–330. Publisher: Routledge _eprint: https://doi.org/10.1080/026999398379628.

[Davidson, 2010] Davidson, R. J. (2010). Empirical explorations of mindfulness: conceptual and methodological conundrums. *Emotion (Washington, D.C.)*, 10(1):8–11.

[Daw et al., 2011] Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., and Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215.

[Dawson et al., 2007] Dawson, M. E., Schell, A. M., and Filion, D. L. (2007). The electrodermal system. In *Handbook of psychophysiology, 3rd ed*, pages 159–181. Cambridge University Press, New York, NY, US.

[de Boer et al., 2017] de Boer, L., Axelsson, J., Riklund, K., Nyberg, L., Dayan, P., Bäckman, L., and Guitart-Masip, M. (2017). Attenuation of dopamine-modulated prefrontal value signals underlies probabilistic reward learning deficits in old age. *eLife*, 6:e26424.

[Dehon et al., 2010] Dehon, H., Larøi, F., and Van der Linden, M. (2010). Affective valence influences participant's susceptibility to false memories and illusory recollection. *Emotion*, 10(5):627–639.

[Dell'Acqua et al., 2010] Dell'Acqua, R., Sessa, P., Peressotti, F., Mulatti, C., Navarrete, E., and Grainger, J. (2010). ERP Evidence for Ultra-Fast Semantic Processing in the Picture–Word Interference Paradigm. *Frontiers in Psychology*, 1.

[Demidenko et al., 2021] Demidenko, M. I., Weigard, A. S., Ganesan, K., Jang, H., Jahn, A., Huntley, E. D., and Keating, D. P. (2021). Interactions between methodological and interindividual variability: How Monetary Incentive Delay (MID) task contrast maps vary and impact associations with behavior. *Brain and Behavior*, 11(5):e02093. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/brb3.2093.

[Demszky et al., 2020] Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). GoEmotions: A Dataset of Fine-Grained Emotions.

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. ISSN: 1063-6919.

[Diener and Larsen, 1984] Diener, E. and Larsen, R. J. (1984). Temporal stability and cross-situational consistency of affective, behavioral, and cognitive responses. *Journal of Personality and Social Psychology*, 47(4):871–883.

[Diener et al., 2009] Diener, E., Lucas, R. E., and Scollon, C. N. (2009). Beyond the Hedonic Treadmill: Revising the Adaptation Theory of Well-Being. In Diener, E., editor, *The Science of Well-Being: The Collected Works of Ed Diener*, Social Indicators Research Series, pages 103–118. Springer Netherlands, Dordrecht.

[Dillion et al., 2023] Dillion, D., Tandon, N., Gu, Y., and Gray, K. (2023). Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.

[Drebitz et al., 2025] Drebitz, E., Rausch, L.-P., and Kreiter, A. K. (2025). Gamma-band synchronization between neurons in the visual cortex is causal for effective information processing and behavior. *Nature Communications*, 16(1):7380. Publisher: Nature Publishing Group.

[Du et al., 2023] Du, X., Deng, X., Qin, H., Shu, Y., Liu, F., Zhao, G., Lai, Y.-K., Ma, C., Liu, Y.-J., and Wang, H. (2023). MMPosE: Movie-Induced Multi-Label Positive Emotion Classification Through EEG Signals. *IEEE Transactions on Affective Computing*, 14(4):2925–2938.

[Duque and Vázquez, 2015] Duque, A. and Vázquez, C. (2015). Double attention bias for positive and negative emotional faces in clinical depression: Evidence from an eye-tracking study. *Journal of Behavior Therapy and Experimental Psychiatry*, 46:107–114.

[Ekman, 1999] Ekman, P. (1999). Basic emotions. In *Handbook of cognition and emotion*, pages 45–60. John Wiley & Sons Ltd, Hoboken, NJ, US.

[Ekman and Friesen, 1971] Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129. Place: US Publisher: American Psychological Association.

[Ekman et al., 1987] Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W. A., Pitcairn, T., Ricci-Bitti, P. E., Scherer, K., Tomita, M., and Tzavaras, A. (1987). Universals and cultural differences

in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, 53(4):712–717. Place: US Publisher: American Psychological Association.

[Eldar et al., 2018] Eldar, E., Roth, C., Dayan, P., and Dolan, R. J. (2018). Decodability of Reward Learning Signals Predicts Mood Fluctuations. *Current Biology*, 28(9):1433–1439.e7.

[Eldar et al., 2016] Eldar, E., Rutledge, R. B., Dolan, R. J., and Niv, Y. (2016). Mood as Representation of Momentum. *Trends in Cognitive Sciences*, 20(1):15–24.

[Ellis et al., 2022] Ellis, J. M., Midgette, E. P., Capiaghi, A. J., Schoemann, A. M., Sears, S. F., Kyle, B. N., Carels, R. A., and Whited, M. C. (2022). Mobile assessment of experiential avoidance, mood, stress, and attendance in cardiopulmonary rehabilitation. *Health Psychology*, 41(12):955–963.

[Emanuel and Eldar, 2023] Emanuel, A. and Eldar, E. (2023). Emotions as computations. *Neuroscience & Biobehavioral Reviews*, 144:104977.

[Erber and Tesser, 1992] Erber, R. and Tesser, A. (1992). Task effort and the regulation of mood: The absorption hypothesis. *Journal of Experimental Social Psychology*, 28(4):339–359.

[Ethofer et al., 2009] Ethofer, T., Van De Ville, D., Scherer, K., and Vuilleumier, P. (2009). Decoding of Emotional Information in Voice-Sensitive Cortices. *Current Biology*, 19(12):1028–1033.

[Farrell et al., 2022] Farrell, K., Lak, A., and Saleem, A. B. (2022). Midbrain dopamine neurons signal phasic and ramping reward prediction error during goal-directed navigation. *Cell Reports*, 41(2):111470.

[Farruque et al., 2024] Farruque, N., Goebel, R., Sivapalan, S., and Zaïane, O. R. (2024). Depression symptoms modelling from social media text: an LLM driven semi-supervised learning approach. *Language Resources and Evaluation*, 58(3):1013–1041.

[Fernández et al., 2012] Fernández, C., Pascual, J. C., Soler, J., Elices, M., Portella, M. J., and Fernández-Abascal, E. (2012). Physiological Responses Induced by Emotion-Eliciting Films. *Applied Psychophysiology and Biofeedback*, 37(2):73–79.

[Ferstl et al., 2022] Ferstl, M., Teckentrup, V., Lin, W. M., Kräutlein, F., Kühnel, A., Klaus, J., Walter, M., and Kroemer, N. B. (2022). Non-invasive vagus nerve stimulation boosts mood recovery after effort exertion. *Psychological Medicine*, 52(14):3029–3039.

[Fischer and Roseman, 2007] Fischer, A. H. and Roseman, I. J. (2007). Beat them or ban them: The characteristics and social functions of anger and contempt. *Journal of Personality and Social Psychology*, 93(1):103–115.

[Fridlund, 1994] Fridlund, A. J. (1994). *Human facial expression: An evolutionary view*. Human facial expression: An evolutionary view. Academic Press, San Diego, CA, US. Pages: xiv, 369.

[Frijda, 1986] Frijda, N. H. (1986). *The emotions*. The emotions. Editions de la Maison des Sciences de l'Homme, Paris, France. Pages: xii, 544.

[Frömer et al., 2021] Frömer, R., Lin, H., Dean Wolf, C. K., Inzlicht, M., and Shenhav, A. (2021). Expectations of reward and efficacy guide cognitive control allocation. *Nature Communications*, 12(1):1030.

[Gagne and Dayan, 2023] Gagne, C. and Dayan, P. (2023). The Inner Sentiments of a Thought. arXiv:2307.01784 [cs].

[Ganesan, 2020] Ganesan, K. (2020). Effort-Related Decision-Making and its Underlying Processes during Childhood. preprint, PsyArXiv.

[GBD 2019 Mental Disorders Collaborators, 2022] GBD 2019 Mental Disorders Collaborators (2022). Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet Psychiatry*, 9(2):137–150.

[Gendolla and Krüsken, 2002] Gendolla, G. H. E. and Krüsken, J. (2002). The joint effect of informational mood impact and performance-contingent consequences on effort-related cardiovascular response. *Journal of Personality and Social Psychology*, 83(2):271.

[Giachanou and Crestani, 2016] Giachanou, A. and Crestani, F. (2016). Like It or Not: A Survey of Twitter Sentiment Analysis Methods. *ACM Comput. Surv.*, 49(2):28:1–28:41.

[Gilam et al., 2019] Gilam, G., Abend, R., Shani, H., Ben-Zion, Z., and Hendler, T. (2019). The anger-infused Ultimatum Game: A reliable and valid paradigm to induce and assess anger. *Emotion*, 19(1):84–96.

[Gillan and Daw, 2016] Gillan, C. M. and Daw, N. D. (2016). Taking Psychiatry Research Online. *Neuron*, 91(1):19–23.

[Gläscher et al., 2010] Gläscher, J., Daw, N., Dayan, P., and O'Doherty, J. P. (2010). States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. *Neuron*, 66(4):585–595.

[Godman et al., 2014] Godman, M., Nagatsu, M., and Salmela, M. (2014). The Social Motivation Hypothesis for Prosocial Behavior. *Philosophy of the Social Sciences*, 44(5):563–587.

[Goyal et al., 2014] Goyal, M., Singh, S., Sibinga, E. M. S., Gould, N. F., Rowland-Seymour, A., Sharma, R., Berger, Z., Sleicher, D., Maron, D. D., Shihab, H. M., Ranasinghe, P. D., Linn, S., Saha, S., Bass, E. B., and Haythornthwaite, J. A. (2014). Meditation Programs for Psychological Stress and Well-being: A Systematic Review and Meta-analysis. *JAMA internal medicine*, 174(3):357–368.

[Greasley et al., 2000] Greasley, P., Sherrard, C., and Waterman, M. (2000). Emotion in Language and Speech: Methodological Issues in Naturalistic Approaches. *Language and Speech*, 43(4):355–375. Publisher: SAGE Publications Ltd.

[Gross, 1998] Gross, J. J. (1998). Antecedent- and response-focused emotion regulation: Divergent consequences for experience, expression, and physiology. *Journal of Personality and Social Psychology*, 74(1):224–237. Place: US Publisher: American Psychological Association.

[Grosscup and Lewinsohn, 1980] Grosscup, S. J. and Lewinsohn, P. M. (1980). Unpleasant and pleasant events, and mood. *Journal of Clinical Psychology*, 36(1):252–259.

[Grupe and Nitschke, 2013] Grupe, D. W. and Nitschke, J. B. (2013). Uncertainty and anticipation in anxiety: an integrated neurobiological and psychological perspective. *Nature Reviews Neuroscience*, 14(7):488–501. Publisher: Nature Publishing Group.

[Grèzes et al., 2021] Grèzes, J., Erblang, M., Vilarem, E., Quiquempoix, M., Van Beers, P., Guillard, M., Sauvet, F., Mennella, R., and Rabat, A. (2021). Impact of total sleep deprivation and related mood changes on approach-avoidance decisions to threat-related facial displays. *Sleep*, 44(12):zsab186.

[Gusnard et al., 2001] Gusnard, D. A., Akbudak, E., Shulman, G. L., and Raichle, M. E. (2001). Medial prefrontal cortex and self-referential mental activity: Relation to a default mode of brain function. *Proceedings of the National Academy of Sciences*, 98(7):4259–4264. Publisher: Proceedings of the National Academy of Sciences.

[Hagendorff et al., 2024] Hagendorff, T., Dasgupta, I., Binz, M., Chan, S. C. Y., Lampinen, A., Wang, J. X., Akata, Z., and Schulz, E. (2024). Machine Psychology.

[Hart et al., 2019] Hart, W., Breeden, C. J., and Richardson, K. (2019). Differentiating dark personalities on impression management. *Personality and Individual Differences*, 147:58–62.

[Hauk et al., 2004] Hauk, O., Johnsrude, I., and Pulvermüller, F. (2004). Somatotopic Representation of Action Words in Human Motor and Premotor Cortex. *Neuron*, 41(2):301–307.

[Haxby, 2012] Haxby, J. V. (2012). Multivariate pattern analysis of fMRI: The early beginnings. *Neuroimage*, 62(2):852–855.

[Heilbron et al., 2022] Heilbron, M., Armeni, K., Schoffelen, J.-M., Hagoort, P., and de Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, 119(32):e2201968119. Publisher: Proceedings of the National Academy of Sciences.

[Het and Wolf, 2007] Het, S. and Wolf, O. T. (2007). Mood changes in response to psychosocial stress in healthy young women: Effects of pretreatment with cortisol. *Behavioral Neuroscience*, 121:11–20.

[hmmlearn, 2025] hmmlearn (2025). hmmlearn — hmmlearn 0.3.3.post1+ge01a10e documentation.

[Ho et al., 2022] Ho, M. K., Abel, D., Correa, C. G., Littman, M. L., Cohen, J. D., and Griffiths, T. L. (2022). People construct simplified mental representations to plan. *Nature*, 606(7912):129–136.

[Huff and Tingley, 2015] Huff, C. and Tingley, D. (2015). "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics*, 2(3):2053168015604648.

[Huggingface, 2025] Huggingface (2025). Text classification.

[Hutto and Gilbert, 2014] Hutto, C. and Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

[Huys and Renz, 2017] Huys, Q. J. and Renz, D. (2017). A Formal Valuation Framework for Emotions and Their Control. *Biological Psychiatry*, 82(6):413–420.

[Huys et al., 2015] Huys, Q. J. M., Daw, N. D., and Dayan, P. (2015). Depression: A Decision-Theoretic Analysis. *Annual Review of Neuroscience*, 38(Volume 38, 2015):1–23. Publisher: Annual Reviews.

[Huys et al., 2012] Huys, Q. J. M., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., and Roiser, J. P. (2012). Bonsai Trees in Your Head: How the Pavlovian System Sculpts Goal-Directed Choices by Pruning Decision Trees. *PLOS Computational Biology*, 8(3):e1002410.

[Hämäläinen et al., 1993] Hämäläinen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J., and Lounasmaa, O. V. (1993). Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics*, 65(2):413–497.

[Inzlicht et al., 2018] Inzlicht, M., Shenhav, A., and Olivola, C. Y. (2018). The Effort Paradox: Effort Is Both Costly and Valued. *Trends in Cognitive Sciences*, 22(4):337–349.

[Isen, 2001] Isen, A. M. (2001). An Influence of Positive Affect on Decision Making in Complex Situations: Theoretical Issues With Practical Implications. *Journal of Consumer Psychology*, 11(2):75–85.

[Islam and Moushi, 2025] Islam, R. and Moushi, O. M. (2025). GPT-4o: The Cutting-Edge Advancement in Multimodal LLM. In Arai, K., editor, *Intelligent Computing*, pages 47–60, Cham. Springer Nature Switzerland.

[Jackson et al., 2022] Jackson, J. C., Watts, J., List, J.-M., Puryear, C., Drabble, R., and Lindquist, K. A. (2022). From Text to Thought: How Analyzing Language Can Advance Psychological Science. *Perspectives on Psychological Science*, 17(3):805–826.

[James, 1948] James, W. (1948). What is emotion? 1884. In *Readings in the history of psychology*, Century psychology series, pages 290–303. Appleton-Century-Crofts, East Norwalk, CT, US.

[Jangraw et al., 2023] Jangraw, D. C., Keren, H., Sun, H., Bedder, R. L., Rutledge, R. B., Pereira, F., Thomas, A. G., Pine, D. S., Zheng, C., Nielson, D. M., and Stringaris, A. (2023). A highly replicable decline in mood during rest and simple tasks. *Nature Human Behaviour*.

[Jansma et al., 2006] Jansma, J., Ramsey, N., de Zwart, J., van Gelderen, P., and Duyn, J. (2006). fMRI study of effort and information processing in a working memory task. *Human Brain Mapping*, 28(5):431–440.

[Joana, 2009] Joana, D. B. D. L. R. (2009). Mood and mental effort : informational mood impact on cardiovascular reactivity and the context-dependency of moods.

[Jääskeläinen et al., 2021] Jääskeläinen, I. P., Sams, M., Glerean, E., and Ahveninen, J. (2021). Movies and narratives as naturalistic stimuli in neuroimaging. *NeuroImage*, 224:117445.

[Kaindl, 1990] Kaindl, H. (1990). Tree Searching Algorithms. In Marsland, T. A. and Schaeffer, J., editors, *Computers, Chess, and Cognition*, pages 133–158, New York, NY. Springer.

[Kassam et al., 2013] Kassam, K. S., Markey, A. R., Cherkassky, V. L., Loewenstein, G., and Just, M. A. (2013). Identifying Emotions on the Basis of Neural Activation. *PloS One*, 8(6):e66032.

[Ke et al., 2025] Ke, J., Song, H., Bai, Z., Rosenberg, M. D., and Leong, Y. C. (2025). Dynamic brain connectivity predicts emotional arousal during naturalistic movie-watching. *PLOS Computational Biology*, 21(4):e1012994. Publisher: Public Library of Science.

[Keltner and Haidt, 1999] Keltner, D. and Haidt, J. (1999). Social functions of emotions at four levels of analysis. *Cognition and Emotion*, 13(5):505–521. Place: United Kingdom Publisher: Taylor & Francis.

[Keren et al., 2021] Keren, H., Zheng, C., Jangraw, D. C., Chang, K., Vitale, A., Rutledge, R. B., Pereira, F., Nielson, D. M., and Stringaris, A. (2021). The temporal representation of experience in subjective mood. *eLife*, 10:e62051.

[Ketelaar and Tung Au, 2003] Ketelaar, T. and Tung Au, W. (2003). The effects of feelings of guilt on the behaviour of uncooperative individuals in repeated social bargaining games: An affect-as-information interpretation of the role of emotion in social interaction. *Cognition and Emotion*, 17(3):429–453.

[Khalfa et al., 2002] Khalfa, S., Isabelle, P., Jean-Pierre, B., and Manon, R. (2002). Event-related skin conductance responses to musical emotions in humans. *Neuroscience Letters*, 328(2):145–149.

[Kim et al., 2018] Kim, H., Lu, X., Costa, M., Kandemir, B., Adams, R. B., Li, J., Wang, J. Z., and Newman, M. G. (2018). Development and validation of Image Stimuli for Emotion Elicitation (ISEE): A novel affective pictorial system with test-retest repeatability. *Psychiatry Research*, 261:414–420.

[King et al., 2018] King, B. M., Cespedes, V. M., Burden, G. K., Brady, S. K., Clement, L. R., Abbott, E. M., Baughman, K. S., Joyner, S. E., Clark, M. M., and Pury, C. L. S. (2018). Extreme under-reporting of body weight by young adults with obesity: relation to social desirability. *Obesity Science & Practice*, 4(2):129–133.

[Kirschbaum et al., 1993] Kirschbaum, C., Pirke, K.-M., and Hellhammer, D. H. (1993). The 'Trier Social Stress Test' – A Tool for Investigating Psychobiological Stress Responses in a Laboratory Setting. *Neuropsychobiology*, 28(1-2):76–81.

[Kober et al., 2008] Kober, H., Barrett, L. F., Joseph, J., Bliss-Moreau, E., Lindquist, K., and Wager, T. D. (2008). Functional grouping and cortical-subcortical interactions in emotion: a meta-analysis of neuroimaging studies. *NeuroImage*, 42(2):998–1031.

[Kool et al., 2016] Kool, W., Cushman, F. A., and Gershman, S. J. (2016). When Does Model-Based Control Pay Off? *PLOS Computational Biology*, 12(8):e1005090.

[Kool et al., 2010] Kool, W., McGuire, J. T., Rosen, Z. B., and Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology. General*, 139(4):665–682.

[Kragel et al., 2022] Kragel, P. A., Hariri, A. R., and LaBar, K. S. (2022). The Temporal Dynamics of Spontaneous Emotional Brain States and Their Implications for Mental Health. *Journal of Cognitive Neuroscience*, 34(5):715–728.

[Kragel and LaBar, 2013] Kragel, P. A. and LaBar, K. S. (2013). Multivariate Pattern Classification Reveals Autonomic and Experiential Representations of Discrete Emotions. *Emotion (Washington, D.C.)*, 13(4):681–690.

[Kragel and LaBar, 2016] Kragel, P. A. and LaBar, K. S. (2016). Decoding the Nature of Emotion in the Brain. *Trends in Cognitive Sciences*, 20(6):444–455. Publisher: Elsevier.

[Kumar et al., 2018] Kumar, P., Goer, F., Murray, L., Dillon, D. G., Beltzer, M. L., Cohen, A. L., Brooks, N. H., and Pizzagalli, D. A. (2018). Impaired reward prediction error encoding and striatal-midbrain connectivity in depression. *Neuropsychopharmacology*, 43(7):1581–1588.

[Kumar et al., 2024] Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., Griffiths, T. L., Hawkins, R. D., and Nastase, S. A. (2024). Shared functional specialization in transformer-based language models and the human brain. *Nature Communications*, 15(1):5523. Publisher: Nature Publishing Group.

[Kurniawan et al., 2013] Kurniawan, I. T., Guitart-Masip, M., Dayan, P., and Dolan, R. J. (2013). Effort and Valuation in the Brain: The Effects of Anticipation and Execution. *Journal of Neuroscience*, 33(14):6160–6169.

[Kurzban, 2010] Kurzban, R. (2010). Does the Brain Consume Additional Glucose during Self-Control Tasks? *Evolutionary Psychology*, 8(2):147470491000800208.

[Kutas and Hillyard, 1980] Kutas, M. and Hillyard, S. A. (1980). Reading Senseless Sentences: Brain Potentials Reflect Semantic Incongruity. *Science*, 207(4427):203–205.

[LaBar et al., 1998] LaBar, K. S., Gatenby, J. C., Gore, J. C., LeDoux, J. E., and Phelps, E. A. (1998). Human amygdala activation during conditioned fear acquisition and extinction: a mixed-trial fMRI study. *Neuron*, 20(5):937–945.

[Lane et al., 2009] Lane, R. D., McRae, K., Reiman, E. M., Chen, K., Ahern, G. L., and Thayer, J. F. (2009). Neural correlates of heart rate variability during emotion. *NeuroImage*, 44(1):213–222.

[Lang et al., 1997] Lang, P. J., Bradley, M. M., and Cuthbert, B. N. (1997). International affective picture system (IAPS): Technical manual and affective ratings.

[Lang et al., 1993] Lang, P. J., Greenwald, M. K., Bradley, M. M., and Hamm, A. O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, 30(3):261–273. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8986.1993.tb03352.x.

[Lange and James, 1922] Lange, C. G. and James, W., editors (1922). *The emotions, Vol. 1.* Williams & Wilkins Co, Baltimore.

[Larsen et al., 1986] Larsen, R. J., Diener, E., and Emmons, R. A. (1986). Affect intensity and reactions to daily life events. *Journal of Personality and Social Psychology*, 51(4):803–814.

[Larsen and Berenbaum, 2014] Larsen, S. E. and Berenbaum, H. (2014). Substantial Symptom Changes in Naturalistic Recovery from Aversive Events. *Journal of Clinical Psychology*, 70(10):967–978.

[Larson et al., 2025] Larson, E., Gramfort, A., Engemann, D. A., Leppakangas, J., Brodbeck, C., Jas, M., Brooks, T. L., Sassenhagen, J., McCloy, D., Luessi, M., King, J.-R., Höchenberger, R., Brunner, C., Goj, R., Favelier, G., van Vliet, M., Wronkiewicz, M., Rockhill, A., Appelhoff, S., Holdgraf, C., Scheltienne, M., Massich, J., Bekhti, Y., Leggitt, A., Dykstra, A., Trachel, R., Luke, R., De Santis, L., Panda, A., Magnuski, M., Westner, B., Wakeman, D. G., Strohmeier, D., Bharadwaj, H., Linzen, T., Barachant, A., Ruzich, E., Bailey, C. J., Li, A., Moutard, C., Bloy, L., Raimondo, F., Huberty, S., Nurminen, J., Billinger, M., Montoya, J., Woodman, M., Lee, I., Schulz, M., Foti, N., Nangini, C., García Alanis, J. C., Orfanos, D. P., Hauk, O., Maddox, R., LaPlante, R., Drew, A., Dinh, C., Binns, T. S., Dumas, G., Martin, Benerradi, J., Hartmann, T., Ort, E., Pasler, P., Repplinger, S., Rudiuk, A., Radanovic, A., Buran, B., Woessner, J., Massias, M., Hämäläinen, M., Sripad, P., Chirkov, V., Mullins, C., Raimundo, F., Kaneda, M., Alday, P., Pari, R., Kornblith, S., Halchenko, Y., Luo, Y.-H., Kasper, J., Doelling, K., Jensen, M., Ruuskanen, S., Kern, S., Gahlot, T., Nunes, A., Gütlin, D., Heinila, E., Armeni, K., kjs, Weinstein, A., Lamus, C., Galván, C. M., Moënne-Loccoz, C., Altukhov, D., Peterson, E., Hanna, J., Houck, J., Klein, N., Roujansky, P., Luke, R., Rantala, A., Maess, B., Forster, C., O'Reilly, C., Welke, D., Kolkhorst, H., Banville, H., Zhang, J., Maksymenko, K., Clarke, M., Anelli, M., Chapochnikov, N., Bannier, P.-A.,

Choudhary, S., Férat, V., Kim, C., Klotzsche, F., Wong, F.-T., Kojcic, I., Nielsen, J. D., Lankinen, K., Tabavi, K., Thibault, L., Gerster, M., Alibou, N., Gayraud, N., Ward, N., Chu, Q., Herbst, S., Quinn, A., Gauthier, A., Pinsard, B., Stephen, E., Hornberger, E., Hathaway, E., Kalenkovich, E., Mamashli, F., Belonosov, G., O'Neill, G., Marinato, G., Anevar, H., Abdelhedi, H., Sosulski, J., Stout, J., Calder-Travis, J., Zhu, J. D., Eisenman, L., Esch, (2025). MNE-Python.

[Lazarus, 1991] Lazarus, R. S. (1991). Progress on a cognitive-motivational-relational theory of emotion. *American Psychologist*, 46(8):819–834.

[LeBel et al., 2023] LeBel, A., Wagner, L., Jain, S., Adhikari-Desai, A., Gupta, B., Morgenthal, A., Tang, J., Xu, L., and Huth, A. G. (2023). A natural language fMRI dataset for voxelwise encoding models. *Scientific Data*, 10:555.

[LeDoux, 2002] LeDoux, J. E. (2002). Emotion Circuits in the Brain. In *Fear and Anxiety*. Routledge.

[Lehne et al., 2015] Lehne, M., Engel, P., Rohrmeier, M., Menninghaus, W., Jacobs, A. M., and Koelsch, S. (2015). Reading a Suspenseful Literary Text Activates Brain Areas Related to Social Cognition and Predictive Inference. *PLOS ONE*, 10(5):e0124550. Publisher: Public Library of Science.

[Lerner et al., 2004] Lerner, J. S., Small, D. A., and Loewenstein, G. (2004). Heart Strings and Purse Strings: Carryover effects of emotions on economic decisions. *Psychological Science*, 15(5):337–341. Place: United Kingdom Publisher: Blackwell Publishing.

[Li and Lu, 2009] Li, M. and Lu, B.-L. (2009). Emotion classification based on gamma-band EEG. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2009:1323–1326.

[Lieberman et al., 2007] Lieberman, M. D., Eisenberger, N. I., Crockett, M. J., Tom, S. M., Pfeifer, J. H., and Way, B. M. (2007). Putting feelings into words: affect labeling disrupts amygdala activity in response to affective stimuli. *Psychological Science*, 18(5):421–428.

[Lindquist et al., 2012] Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., and Barrett, L. F. (2012). The brain basis of emotion: a meta-analytic review. *The Behavioral and Brain Sciences*, 35(3):121–143.

[Liu et al., 2019] Liu, Y., Dolan, R. J., Kurth-Nelson, Z., and Behrens, T. E. J. (2019). Human Replay Spontaneously Reorganizes Experience. *Cell*, 178(3):640–652.e14.

[Liu et al., 2018] Liu, Y.-J., Yu, M., Zhao, G., Song, J., Ge, Y., and Shi, Y. (2018). Real-Time Movie-Induced Discrete Emotion Recognition from EEG Signals. *IEEE Transactions on Affective Computing*, 9(4):550–562.

[Liuzzi et al., 2021] Liuzzi, L., Chang, K. K., Zheng, C., Keren, H., Saha, D., Nielson, D. M., and Stringaris, A. (2021). Magnetoencephalographic Correlates of Mood and Reward Dynamics in Human Adolescents. *Cerebral Cortex*, page bhab417.

[Lockwood et al., 2017] Lockwood, P. L., Hamonet, M., Zhang, S. H., Ratnavel, A., Salmony, F. U., Husain, M., and Apps, M. A. J. (2017). Prosocial apathy for helping others when effort is required \textbar Nature Human Behaviour. *Nature Human Behaviour*, 1(7):0131.

[Lopopolo et al., 2024] Lopopolo, A., Fedorenko, E., Levy, R., and Rabovsky, M. (2024). Cognitive Computational Neuroscience of Language: Using Computational Models to Investigate Language Processing in the Brain. *Neurobiology of Language*, 5(1):1–6.

[Low et al., 2020] Low, D. M., Rumker, L., Talkar, T., Torous, J., Cecchi, G., and Ghosh, S. S. (2020). Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study. *Journal of Medical Internet Research*, 22(10):e22635.

[Madsen et al., 1995] Madsen, P. L., Hasselbalch, S. G., Hagemann, L. P., Olsen, K. S., Bülow, J., Holm, S., Wildschiødtz, G., Paulson, O. B., and Lassen, N. A. (1995). Persistent Resetting of the Cerebral Oxygen/Glucose Uptake Ratio by Brain Activation: Evidence Obtained with the Kety—Schmidt Technique. *Journal of Cerebral Blood Flow & Metabolism*, 15(3):485–491.

[Mao et al., 2023] Mao, R., Liu, Q., He, K., Li, W., and Cambria, E. (2023). The Biases of Pre-Trained Language Models: An Empirical Study on Prompt-Based Sentiment Analysis and Emotion Detection. *IEEE Transactions on Affective Computing*, 14(3):1743–1753.

[Massar et al., 2018] Massar, S. A. A., Csatho, A., and Van der Linden, D. (2018). Quantifying the Motivational Effects of Cognitive Fatigue Through Effort-Based Decision Making. *Frontiers in Psychology*, 9.

[Matejka et al., 2013] Matejka, M., Kazzer, P., Seehausen, M., Bajbouj, M., Klann-Delius, G., Menninghaus, W., Jacobs, A. M., Heekeren, H. R., and Prehn, K. (2013). Talking about Emotion: Prosody and Skin Conductance Indicate Emotion Regulation. *Frontiers in Psychology*, 4. Publisher: Frontiers.

[Mayer et al., 2023] Mayer, C. W. F., Ludwig, S., and Brandt, S. (2023). Prompt text classifications with transformer models! An exemplary introduction to prompt-based learning with large language models. *Journal of Research on Technology in Education*, 55(1):125–141.

[McCredie and Morey, 2019] McCredie, M. N. and Morey, L. C. (2019). Who Are the Turkers? A Characterization of MTurk Workers Using the Personality Assessment Inventory. *Assessment*, 26(5):759–766.

[Mintz, 2010] Mintz, N. L. (2010). Effects of Esthetic Surroundings: II. Prolonged and Repeated Experience in a "Beautiful" and an "Ugly" Room. *The Journal of Psychology*.

[Mogg and Bradley, 1998] Mogg, K. and Bradley, B. P. (1998). A cognitive-motivational analysis of anxiety. *Behaviour Research and Therapy*, 36(9):809–848.

[Mortensen and Hughes, 2018] Mortensen, K. and Hughes, T. L. (2018). Comparing Amazon's Mechanical Turk Platform to Conventional Data Collection Methods in the Health and Medical Research Literature. *Journal of General Internal Medicine*, 33(4):533–538.

[Nava et al., 2016] Nava, E., Romano, D., Grassi, M., and Turati, C. (2016). Skin conductance reveals the early development of the unconscious processing of emotions. *Cortex*, 84:124–131.

[Nickolls et al., 2008] Nickolls, J., Buck, I., Garland, M., and Skadron, K. (2008). Scalable parallel programming with CUDA. In *ACM SIGGRAPH 2008 classes*, SIGGRAPH '08, pages 1–14, New York, NY, USA. Association for Computing Machinery.

[Niedenthal and Showers, 1991] Niedenthal, P. M. and Showers, C. (1991). The Perception and Processing of Affective Information and its Influences on Social Judgment. In *Emotion and Social Judgements*. Garland Science.

[Niv et al., 2007] Niv, Y., Daw, N. D., Joel, D., and Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology*, 191(3):507–520.

[Norman et al., 2006] Norman, K. A., Polyn, S. M., Detre, G. J., and Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10(9):424–430.

[Northoff et al., 2006] Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., and Panksepp, J. (2006). Self-referential processing in our brain–a meta-analysis of imaging studies on the self. *NeuroImage*, 31(1):440–457.

[Ochsner et al., 2002] Ochsner, K. N., Bunge, S. A., Gross, J. J., and Gabrieli, J. D. E. (2002). Rethinking feelings: an FMRI study of the cognitive regulation of emotion. *Journal of Cognitive Neuroscience*, 14(8):1215–1229.

[Ochsner and Gross, 2005] Ochsner, K. N. and Gross, J. J. (2005). The cognitive control of emotion. *Trends in Cognitive Sciences*, 9(5):242–249.

[Onysk and Huys, 2025] Onysk, J. and Huys, Q. (2025). Objective quantification of mood states using large language models.

[Osterhout and Holcomb, 1992] Osterhout, L. and Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, 31(6):785–806.

[Otto and Daw, 2019] Otto, A. R. and Daw, N. D. (2019). The opportunity cost of time modulates cognitive effort. *Neuropsychologia*, 123:92–105.

[Ousdal et al., 2008] Ousdal, O., Jensen, J., Server, A., Hariri, A., Nakstad, P., and Andreassen, O. (2008). The human amygdala is involved in general behavioral relevance detection: Evidence from an event-related functional magnetic resonance imaging Go-NoGo task. *Neuroscience*, 156(3):450–455.

[Ousdal et al., 2012] Ousdal, O. T., Reckless, G. E., Server, A., Andreassen, O. A., and Jensen, J. (2012). Effect of relevance on amygdala activation and association with the ventral striatum. *NeuroImage*, 62(1):95–101.

[O'Callaghan and Stringaris, 2019] O'Callaghan, G. and Stringaris, A. (2019). Reward Processing in Adolescent Depression Across Neuroimaging Modalities. *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie*, 47(6):535–541.

[Packheiser et al., 2021] Packheiser, J., Donoso, J. R., Cheng, S., Güntürkün, O., and Pusch, R. (2021). Trial-by-trial dynamics of reward prediction error-associated signals during extinction learning and renewal. *Progress in Neurobiology*, 197:101901.

[Patil and Jadon, 2025] Patil, A. and Jadon, A. (2025). Advancing Reasoning in Large Language Models: Promising Methods and Approaches. arXiv:2502.03671 [cs].

[Peelen et al., 2010] Peelen, M. V., Atkinson, A. P., and Vuilleumier, P. (2010). Supramodal representations of perceived emotions in the human brain. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 30(30):10127–10134.

[Pekrun, 2006] Pekrun, R. (2006). The Control-Value Theory of Achievement Emotions: Assumptions, Corollaries, and Implications for Educational Research and Practice. *Educational Psychology Review*, 18(4):315–341.

[Pessoa, 2017] Pessoa, L. (2017). A Network Model of the Emotional Brain. *Trends in Cognitive Sciences*, 21(5):357–371.

[Phan et al., 2002] Phan, K. L., Wager, T., Taylor, S. F., and Liberzon, I. (2002). Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. *NeuroImage*, 16(2):331–348.

[Philip, 1971] Philip, B. (1971). Hedonic relativism and planning the good society. *Adaptation level theory*, pages 287 – 301.

[Phillips et al., 1997] Phillips, M. L., Young, A. W., Senior, C., Brammer, M., Andrew, C., Calder, A. J., Bullmore, E. T., Perrett, D. I., Rowland, D., Williams, S. C., Gray, J. A., and David, A. S. (1997). A specific neural substrate for perceiving facial expressions of disgust. *Nature*, 389(6650):495–498.

[Pike and Robinson, 2022] Pike, A. C. and Robinson, O. J. (2022). Reinforcement Learning in Patients With Mood and Anxiety Disorders vs Control Individuals: A Systematic Review and Meta-analysis. *JAMA psychiatry*, 79(4):313–322.

[Poldrack, 2010] Poldrack, R. A. (2010). Mapping Mental Function to Brain Structure: How Can Cognitive Neuroimaging Succeed? *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 5(6):753–761.

[Prolific Inc, 2023] Prolific Inc (2023). Prolific.

[Pólya and Csertő, 2023] Pólya, T. and Csertő, I. (2023). Emotion Recognition Based on the Structure of Narratives. *Electronics*, 12(4):919. Publisher: Multidisciplinary Digital Publishing Institute.

[R Core Team, 2020] R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Raffaelli et al., 2021] Raffaelli, Q., Mills, C., de Stefano, N.-A., Mehl, M. R., Chambers, K., Fitzgerald, S. A., Wilcox, R., Christoff, K., Andrews, E. S., Grilli, M. D., O'Connor, M.-F., and Andrews-Hanna, J. R. (2021). The think aloud paradigm reveals differences in the content, dynamics and conceptual scope of resting state thought in trait brooding. *Scientific Reports*, 11(1):19362.

[Raichle, 2015] Raichle, M. E. (2015). The Brain's Default Mode Network. *Annual Review of Neuroscience*, 38(Volume 38, 2015):433–447. Publisher: Annual Reviews.

[Redmiles et al., 2019] Redmiles, E. M., Kross, S., and Mazurek, M. L. (2019). How Well Do My Results Generalize? Comparing Security and Privacy Survey Results

from MTurk, Web, and Telephone Samples. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 1326–1343.

[Reimers and Gurevych, 2019] Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv:1908.10084 [cs].

[Richardson and Suinn, 1972] Richardson, F. C. and Suinn, R. M. (1972). The Mathematics Anxiety Rating Scale: Psychometric data. *Journal of Counseling Psychology*, 19(6):551–554. Place: US Publisher: American Psychological Association.

[Rilling et al., 2002] Rilling, J., Gutman, D., Zeh, T., Pagnoni, G., Berns, G., and Kilts, C. (2002). A neural basis for social cooperation. *Neuron*, 35(2):395–405.

[Robinson and Clore, 2002] Robinson, M. D. and Clore, G. L. (2002). Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin*, 128(6):934–960. Place: US Publisher: American Psychological Association.

[Robinson and Morsella, 2014] Robinson, M. M. and Morsella, E. (2014). The subjective effort of everyday mental tasks: Attending, assessing, and choosing. *Motivation and Emotion*, 38(6):832–843.

[Robles and Johnson, 2017] Robles, C. F. and Johnson, A. W. (2017). Disruptions in effort-based decision-making and consummatory behavior following antagonism of the dopamine D2 receptor. *Behavioural Brain Research*, 320:431–439.

[Rouhani and Niv, 2021] Rouhani, N. and Niv, Y. (2021). Signed and unsigned reward prediction errors dynamically enhance learning and memory. *eLife*, 10:e61077.

[Rude et al., 2004] Rude, S., Gortner, E.-M., and Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, 18(8):1121–1133. Publisher: Routledge _eprint: https://doi.org/10.1080/02699930441000030.

[Russell, 1994] Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin*, 115(1):102–141.

[Russell et al., 1989] Russell, J. A., Lewicka, M., and Niit, T. (1989). A cross-cultural study of a circumplex model of affect. *Journal of Personality and Social Psychology*, 57(5):848–856. Place: US Publisher: American Psychological Association.

[Russell and Wefald, 1991] Russell, S. and Wefald, E. (1991). Principles of metareasoning. *Artificial Intelligence*, 49(1):361–395.

[Rutledge et al., 2017] Rutledge, R. B., Moutoussis, M., Smittenaar, P., Zeidman, P., Taylor, T., Hrynkiewicz, L., Lam, J., Skandali, N., Siegel, J. Z., Ousdal, O. T., Prabhu, G., Dayan, P., Fonagy, P., and Dolan, R. J. (2017). Association of Neural and Emotional Impacts of Reward Prediction Errors With Major Depression. *JAMA Psychiatry*, 74(8):790.

[Rutledge et al., 2014] Rutledge, R. B., Skandali, N., Dayan, P., and Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences*, 111(33):12252–12257.

[Rutledge et al., 2015] Rutledge, R. B., Skandali, N., Dayan, P., and Dolan, R. J. (2015). Dopaminergic Modulation of Decision Making and Subjective Well-Being. *Journal of Neuroscience*, 35(27):9811–9822.

[Ryskin and Nieuwland, 2023] Ryskin, R. and Nieuwland, M. S. (2023). Prediction during language comprehension: what is next? *Trends in Cognitive Sciences*, 27(11):1032–1052. Publisher: Elsevier.

[Saarimäki, 2021] Saarimäki, H. (2021). Naturalistic Stimuli in Affective Neuroimaging: A Review. *Frontiers in Human Neuroscience*, 15. Publisher: Frontiers.

[Saarimäki et al., 2016] Saarimäki, H., Gotsopoulos, A., Jääskeläinen, I. P., Lampinen, J., Vuilleumier, P., Hari, R., Sams, M., and Nummenmaa, L. (2016). Discrete Neural Signatures of Basic Emotions. *Cerebral Cortex (New York, N.Y.: 1991)*, 26(6):2563–2573.

[Sander et al., 2003] Sander, D., Grafman, J., and Zalla, T. (2003). The human amygdala: an evolved system for relevance detection. *Reviews in the Neurosciences*, 14(4):303–316.

[Sandra and Otto, 2018] Sandra, D. A. and Otto, A. R. (2018). Cognitive capacity limitations and Need for Cognition differentially predict reward-induced cognitive effort expenditure. *Cognition*, 172:101–106.

[Sanh et al., 2020] Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2020). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108 [cs].

[Satpute et al., 2013] Satpute, A. B., Shu, J., Weber, J., Roy, M., and Ochsner, K. N. (2013). The functional neural architecture of self-reports of affective experience. *Biological Psychiatry*, 73(7):631–638.

[Scherer, 1984] Scherer, K. R. (1984). Emotion as a multicomponent process: A model and some cross-cultural data. *Review of Personality & Social Psychology*, 5:37–63. Place: US Publisher: Sage Publications, Inc.

[Schultz et al., 1997] Schultz, W., Dayan, P., and Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, 275(5306):1593–1599.

[Schupp et al., 2006] Schupp, H. T., Flaisch, T., Stockburger, J., and Junghöfer, M. (2006). Emotion and attention: event-related brain potential studies. *Progress in Brain Research*, 156:31–51.

[Seabrook et al., 2018] Seabrook, E. M., Kern, M. L., Fulcher, B. D., and Rickard, N. S. (2018). Predicting Depression From Language-Based Emotion Dynamics: Longitudinal Analysis of Facebook and Twitter Status Updates. *Journal of Medical Internet Research*, 20(5):e9267.

[Sergerie et al., 2010] Sergerie, K., Armony, J. L., Menear, M., Sutton, H., and Lepage, M. (2010). Influence of Emotional Expression on Memory Recognition Bias in Schizophrenia as Revealed by fMRI. *Schizophrenia Bulletin*, 36(4):800–810.

[Seymour and Dolan, 2008] Seymour, B. and Dolan, R. (2008). Emotion, Decision Making, and the Amygdala. *Neuron*, 58(5):662–671.

[Shapiro et al., 2013] Shapiro, D. N., Chandler, J., and Mueller, P. A. (2013). Using Mechanical Turk to Study Clinical Populations. *Clinical Psychological Science*, 1(2):213–220.

[Sheline et al., 2009] Sheline, Y. I., Barch, D. M., Price, J. L., Rundle, M. M., Vaishnavi, S. N., Snyder, A. Z., Mintun, M. A., Wang, S., Coalson, R. S., and Raichle, M. E. (2009). The default mode network and self-referential processes in depression. *Proceedings of the National Academy of Sciences*, 106(6):1942–1947. Publisher: Proceedings of the National Academy of Sciences.

[Shenhav et al., 2017] Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., and Botvinick, M. M. (2017). Toward a Rational and Mechanistic Account of Mental Effort. *Annual Review of Neuroscience*, 40(1):99–124.

[Sherman et al., 1975] Sherman, M., Trief, P., and Sprafkin, R. (1975). Impression management in the psychiatric interview: Quality, style, and individual differences. *Journal of Consulting and Clinical Psychology*, 43(6):867–871. Place: US Publisher: American Psychological Association.

[Shu et al., 2018] Shu, L., Xie, J., Yang, M., Li, Z., Li, Z., Liao, D., Xu, X., and Yang, X. (2018). A Review of Emotion Recognition Using Physiological Signals. *Sensors*, 18(7):2074. Publisher: Multidisciplinary Digital Publishing Institute.

[Siegel et al., 2018a] Siegel, E. H., Sands, M. K., Van den Noortgate, W., Condon, P., Chang, Y., Dy, J., Quigley, K. S., and Barrett, L. F. (2018a). Emotion fingerprints or emotion populations? A meta-analytic investigation of autonomic features of emotion categories. *Psychological Bulletin*, 144(4):343–393.

[Siegel et al., 2018b] Siegel, E. H., Wormwood, J. B., Quigley, K. S., and Barrett, L. F. (2018b). Seeing What You Feel: Affect Drives Visual Perception of Structurally Neutral Faces. *Psychological Science*, 29(4):496–503.

[Somarathna et al., 2023] Somarathna, R., Bednarz, T., and Mohammadi, G. (2023). Virtual Reality for Emotion Elicitation – A Review. *IEEE Transactions on Affective Computing*, 14(4):2626–2645.

[Song and Hakoda, 2018] Song, Y. and Hakoda, Y. (2018). Selective Impairment of Basic Emotion Recognition in People with Autism: Discrimination Thresholds for Recognition of Facial Expressions of Varying Intensities. *Journal of Autism and Developmental Disorders*, 48(6):1886–1894.

[Sonnentag and Grant, 2012] Sonnentag, S. and Grant, A. M. (2012). Doing Good at Work Feels Good at Home, but Not Right Away: When and Why Perceived Prosocial Impact Predicts Positive Affect. *Personnel Psychology*, 65(3):495–530.

[Stone and Neale, 1984] Stone, A. A. and Neale, J. M. (1984). Effects of severe daily events on mood. *Journal of Personality and Social Psychology*, 46(1):137–144.

[Talarico and Rubin, 2003] Talarico, J. M. and Rubin, D. C. (2003). Confidence, Not Consistency, Characterizes Flashbulb Memories. *Psychological Science*, 14(5):455–461.

[Taylor et al., 2024] Taylor, G. J., Porcelli, P., and Bagby, R. M. (2024). Alexithymia: A Defense of the Original Conceptualization of the Construct and a Critique of the Attention-Appraisal Model. *Clinical Neuropsychiatry*, 21(5):329–357.

[Thayer and Lane, 2000] Thayer, J. F. and Lane, R. D. (2000). A model of neurovisceral integration in emotion regulation and dysregulation. *Journal of Affective Disorders*, 61(3):201–216.

[Torre and Lieberman, 2018] Torre, J. B. and Lieberman, M. D. (2018). Putting Feelings Into Words: Affect Labeling as Implicit Emotion Regulation. *Emotion Review*, 10(2):116–124.

[Torres et al., 2020] Torres, E. P., Torres, E. A., Hernández-Álvarez, M., and Yoo, S. G. (2020). EEG-Based BCI Emotion Recognition: A Survey. *Sensors*, 20(18):5083. Publisher: Multidisciplinary Digital Publishing Institute.

[Touvron et al., 2023] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs].

[Treadway et al., 2012] Treadway, M. T., Bossaller, N. A., Shelton, R. C., and Zald, D. H. (2012). Effort-based decision-making in major depressive disorder: A translational model of motivational anhedonia. *Journal of Abnormal Psychology*, 121(3):553–558.

[Treadway et al., 2009] Treadway, M. T., Buckholtz, J. W., Schwartzman, A. N., Lambert, W. E., and Zald, D. H. (2009). Worth the 'EEfRT'? The Effort Expenditure for Rewards Task as an Objective Measure of Motivation and Anhedonia. *PLoS ONE*, 4(8).

[Vaish et al., 2016] Vaish, A., Carpenter, M., and Tomasello, M. (2016). The Early Emergence of Guilt-Motivated Prosocial Behavior. *Child Development*, 87(6):1772–1782.

[van Duijn et al., 2023] van Duijn, M., van Dijk, B., Kouwenhoven, T., de Valk, W., Spruit, M., and van der Putten, P. (2023). Theory of Mind in Large Language Models: Examining Performance of 11 State-of-the-Art models vs. Children Aged 7-10 on Advanced Tests. In Jiang, J., Reitter, D., and Deng, S., editors, *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 389–402, Singapore. Association for Computational Linguistics.

[Van Kleef et al., 2004] Van Kleef, G. A., De Dreu, C. K. W., and Manstead, A. S. R. (2004). The Interpersonal Effects of Emotions in Negotiations: A Motivated Information Processing Approach. *Journal of Personality and Social Psychology*, 87(4):510–528.

[Vaswani et al., 2023] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention Is All You Need.

[Vogel et al., 2020] Vogel, T. A., Savelson, Z. M., Otto, A. R., and Roy, M. (2020). Forced choices reveal a trade-off between cognitive effort and physical pain. *eLife*, 9:e59410.

[Wallenwein et al., 2024] Wallenwein, L. A., Schmidt, S. N., Hass, J., and Mier, D. (2024). Cross-modal decoding of emotional expressions in fMRI—Cross-session and cross-sample replication. *Imaging Neuroscience*, 2:imag–2–00289.

[Walsh, 1969] Walsh, W. B. (1969). Self-report under socially undesirable and distortion conditions. *Journal of Counseling Psychology*, 16(6):569–574. Place: US Publisher: American Psychological Association.

[Wang et al., 2018] Wang, C.-A., Baird, T., Huang, J., Coutinho, J. D., Brien, D. C., and Munoz, D. P. (2018). Arousal Effects on Pupil Size, Heart Rate, and Skin Conductance in an Emotional Face Task. *Frontiers in Neurology*, 9. Publisher: Frontiers.

[Wang et al., 2023] Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., and Anandkumar, A. (2023). Voyager: An Open-Ended Embodied Agent with Large Language Models.

[Wang et al., 2017] Wang, L., Zheng, J., and Meng, L. (2017). Effort provides its own reward: endeavors reinforce subjective expectation and evaluation of task performance. *Experimental Brain Research*, 235(4):1107–1118.

[Wang et al., 2013] Wang, X., Zhang, C., Ji, Y., Sun, L., Wu, L., and Bao, Z. (2013). A Depression Detection Model Based on Sentiment Analysis in Micro-blog Social Network. In Li, J., Cao, L., Wang, C., Tan, K. C., Liu, B., Pei, J., and Tseng, V. S., editors, *Trends and Applications in Knowledge Discovery and Data Mining*, pages 201–213, Berlin, Heidelberg. Springer.

[Watson et al., 1988] Watson, D., Clark, L. A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54(6):1063–1070. Place: US Publisher: American Psychological Association.

[Weierich et al., 2019] Weierich, M. R., Kleshchova, O., Rieder, J. K., and Reilly, D. M. (2019). The Complex Affective Scene Set (COMPASS): Solving the Social Content Problem in Affective Visual Stimulus Sets. *Collabra: Psychology*, 5(1):53.

[Weiner, 1985] Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review*, 92(4):548–573.

[Weninger et al., 2015] Weninger, F., Wöllmer, M., and Schuller, B. (2015). Emotion Recognition in Naturalistic Speech and Language—A Survey. In *Emotion Recognition*, pages 237–267. John Wiley & Sons, Ltd. Section: 10 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118910566.ch10.

[Westbrook et al., 2013] Westbrook, A., Kester, D., and Braver, T. S. (2013). What is the subjective cost of cognitive effort? Load, trait, and aging effects revealed by economic preference. *PloS One*, 8(7):e68210.

[World Health Organization, 2022] World Health Organization (2022). *World Mental Health Report: Transforming Mental Health for All*. World Health Organization, Geneva, 1st ed edition.

[Xu et al., 2024] Xu, Y., Zhou, Y., Cai, Y., Xie, J., Ye, R., and Wu, Z. (2024). Multimodal Emotion Captioning Using Large Language Model with Prompt Engineering. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, pages 104–109, Melbourne VIC Australia. ACM.

[Yang et al., 2014] Yang, X.-h., Huang, J., Zhu, C.-y., Wang, Y.-f., Cheung, E. F., Chan, R. C., and Xie, G.-r. (2014). Motivational deficits in effort-based decision making in individuals with subsyndromal depression, first-episode and remitted depression patients. *Psychiatry Research*, 220(3):874–882.

[Yang et al., 2025] Yang, Z., Su, Q., Xie, J., Su, H., Huang, T., Han, C., Zhang, S., Zhang, K., and Xu, G. (2025). Music tempo modulates emotional states as revealed through EEG insights. *Scientific Reports*, 15(1):8276. Publisher: Nature Publishing Group.

[Ye et al., 2024] Ye, A., Moore, J., Novick, R., and Zhang, A. X. (2024). Language Models as Critical Thinking Tools: A Case Study of Philosophers.

[Youssofzadeh et al., 2023] Youssofzadeh, V., Roy, S., Chowdhury, A., Izadysadr, A., Parkkonen, L., Raghavan, M., and Prasad, G. (2023). Mapping and decoding cortical engagement during motor imagery, mental arithmetic, and silent word generation using MEG. *Human Brain Mapping*, 44(8):3324–3342.

[Youssofzadeh et al., 2020] Youssofzadeh, V., Stout, J., Ustine, C., Gross, W. L., Conant, L. L., Humphries, C. J., Binder, J. R., and Raghavan, M. (2020). Mapping language from MEG beta power modulations during auditory and visual naming. *NeuroImage*, 220:117090.

[Zadra and Clore, 2011] Zadra, J. R. and Clore, G. L. (2011). Emotion and perception: the role of affective information. *WIREs Cognitive Science*, 2(6):676–685.

[Zhang et al., 2020] Zhang, J., Yin, Z., Chen, P., and Nichele, S. (2020). Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion*, 59:103–126.

[Zhang et al., 2013] Zhang, W.-N., Chang, S.-H., Guo, L.-Y., Zhang, K.-L., and Wang, J. (2013). The neural correlates of reward-related processing in major depressive disorder: A meta-analysis of functional magnetic resonance imaging studies. *Journal of Affective Disorders*, 151(2):531–539.

[Zhang et al., 2024] Zhang, Z., Fort, J. M., and Giménez Mateu, L. (2024). Mini review: Challenges in EEG emotion recognition. *Frontiers in Psychology*, 14. Publisher: Frontiers.

[Zhao et al., 2022] Zhao, G., Li, Y., and Xu, Q. (2022). From emotion AI to cognitive AI. Accepted: 2023-11-29T13:02:08Z Journal Abbreviation: From Emotion AI to Cognitive AI Publisher: Scilight Press.

[Zhou et al., 2021] Zhou, F., Zhao, W., Qi, Z., Geng, Y., Yao, S., Kendrick, K. M., Wager, T. D., and Becker, B. (2021). A distributed fMRI-based signature for the subjective experience of fear. *Nature Communications*, 12(1):6643.

[Zhu et al., 2019] Zhu, J., Ji, L., and Liu, C. (2019). Heart rate variability monitoring for emotion and disorders of emotion. *Physiological Measurement*, 40(6):064004. Publisher: IOP Publishing.

[Zupan and Babbage, 2017] Zupan, B. and Babbage, D. R. (2017). Film clips and narrative text as subjective emotion elicitation techniques. *The Journal of Social Psychology*, 157(2):194–210. Publisher: Routledge _eprint: https://doi.org/10.1080/00224545.2016.1208138.

Appendices

Appendix A

Supplemental information for chapter 2

A.1 Reaction Time Analysis

To affirm that difficulty manipulation does lead to higher cognitive effort expenditure, we analyzed the effect of difficulty on another metric that is linked to cognitive effort: reaction time. Because in the Multi-Attempt Letter Task, the participants were allowed to repeat a given trial, so their reaction times were not suitable for this analysis. We then only applied this analysis to the reaction time data from the Single-Attempt Letter Task.

We first log-transformed all the reaction time data and constructed a mixed effect linear regression model as following:

$$log(RT_{t,i}) = (\alpha_0 + \beta_{0,i}) + (\alpha_n + \beta_{n,i})n_{t,i} + (\alpha_\tau + \beta_{\tau,i})t + \epsilon_{t,i} \tag{A.1}$$

where $RT_{t,i}$ is the reaction time for a given participant (i) for a given trial (t). The rest of the notations are similar to our mixed linear effect model for momentary mood ratings, as illustrated in the Methods section.

We found that indeed increased number of letters (n) leads to higher reaction time (est. = 0.061, std. error = 0.006, df = 209, t = 10.157, $p < 0.001$).

We conclude that indeed higher difficulty leads to higher reaction time, which indicates increased exertion of cognitive effort.

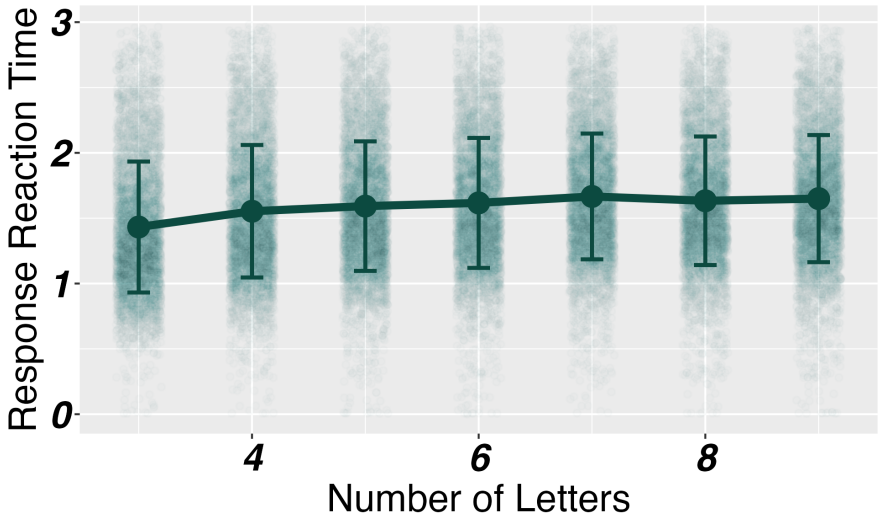


Figure A.1: Reaction Time data from the Single-Attempt Letter Task.

A.2 Time-in-task Analysis

To further investigate how participants' performance and mood changed over the course of task, we provided additional figures (Figure A.2) to illustrate that the main effects we have reported is stable and do not change over the course of the task.

We also conducted a more precise examination of the time-in-task effect by replacing trial number (t) with time-in-task (τ). We implemented this change to the full models of both Multi-Attempt Letters Task and Single-Attempt Letters Task. We did not find that such change influence the results that we have reported. More, we found that using τ instead of t resulted in a worse mood predicting model for data in both the Multi-Attempt Letters Task ($AIC_t = 53786, AIC_\tau = 54017$, Chi square = 230.99) and the Single-Attempt Letters Task ($AIC_t = 111809, AIC_\tau = 111902$, Chi square = 92.917).

A.3 Exclusion Summary

To illustrate that our cut-offs for each exclusion criteria is reasonable for our sample, the aggregated performance data from both studies are illustrated in Fig. A.3.

Using base R, We generated the chance level performance via simulation following procedure:

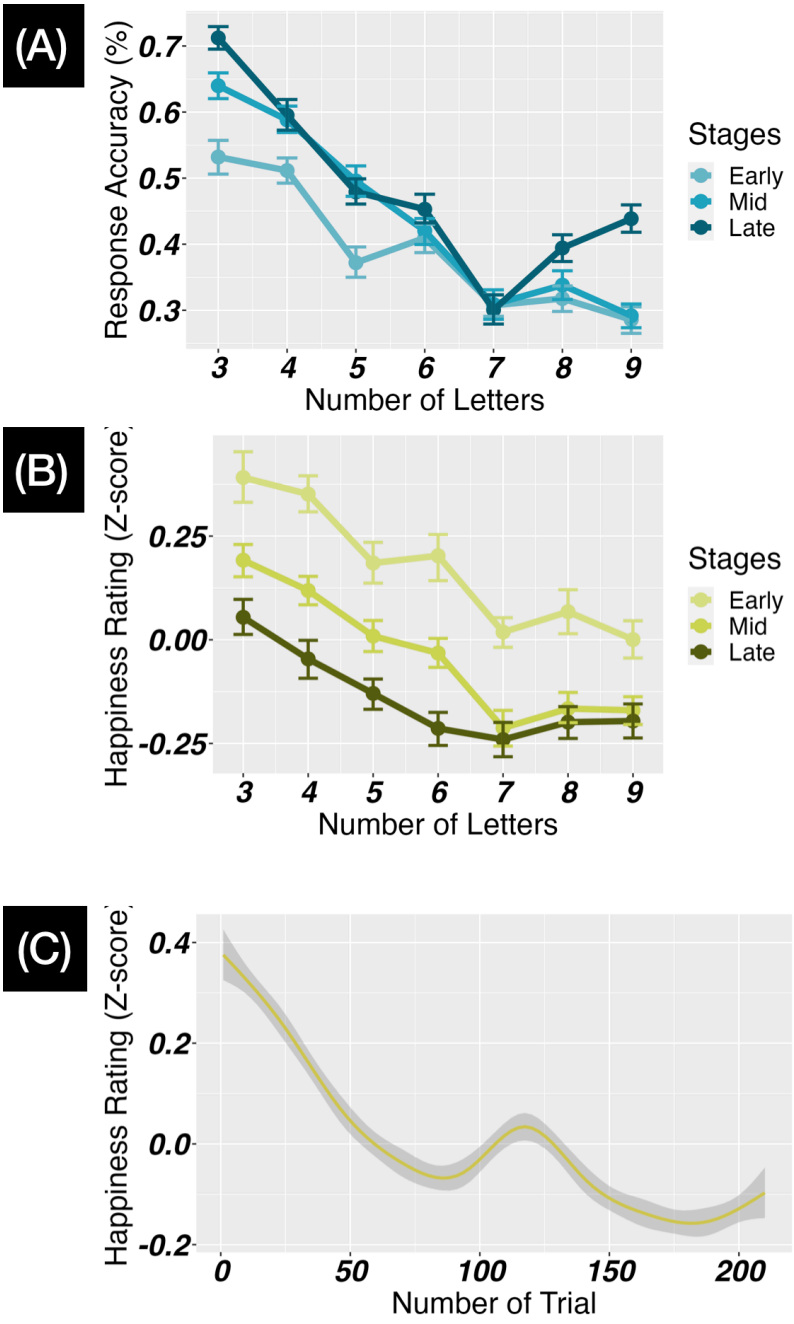


Figure A.2: The effects of number of letters on accuracy and mood over the course of the task. Stage (early, mid late) indicates a factorized form of the time elapsed since the beginning of the task (evenly split into 3 sections: early/mid/late). A) Accuracy over the course of the task showed no significant change; B) Difficulty's impact on mood is stable over the course of the task; and C) Mood showed downward drifting over time.

1. We constructed one simple agent that randomly selects one of the available options. For example, for a trial that had 6 letters, the agent will randomly select from 1 to 6.

2. We simulated behaviors from this agent in both the two task environments for 10,000 times

3. Because of central limit theorem, we can then calculate the probability density function of performance benchmarks (average number of error per trial for Multi-Attempt Letter Task and total accuracy count for Single-Attempt Letter Task).

4. We then calculate the performance benchmark at the 95th percentile, which we interpret as the minimal performance required to be considered as above chance performance.

A.4 Performance Feedback in Single-Attempt Letters Task

To demonstrate that content of the sparse feedbacks in the Single-Attempt Letters Task does not influence the results we have reported in the Results section, we conducted two additional analysis, we first included shown accuracy as additional predicting variable in our full model. We then tested whether trial-by-trial factual error (whenever participants made an error) has an impact on our results.

In the first analysis, we replace the variable f that indicates whether performance feedback was displayed to a, the accuracy percentage, in the Full mood predicting model that we have reported. We further allow this variable to interact with number of trial since last feedback (λ). With this model, we can estimate the effect of accuracy feedback on the trial it was shown as well as it is lasting effects. This Full + Accuracy (FA) model statistics were reported in Table. A.1

In the second analysis, we further include the factual correctness (c) in the model reported above. Doing so, we aim to examine whether participants' mood were influenced by objective correctness even if the outcomes were not explicitly showed. This Full + Accuracy + Correctness (FAC) model statistics were reported in Table. A.1 as well.

We conducted a ordinary likelihood ratio tests to see which model best explained the data and found that the FAC model held superior fit (Chi square = 2165.1, d.f. = 8, $p < 0.001$).

Model	$R^2_{marginal}$	$R^2_{conditional}$	Effect	Estimate	t-value	p-value	Effect size
FA	0.009	0.897	n	-0.174	-7.976	<0.001	0.181
			a	0.989	3.379	<0.001	1.029
			λ	0.180	3.911	<0.001	0.187
			t	-0.210	-4.851	<0.001	0.218
			$a * \lambda$	-0.334	-3.270	0.001	0.347
FAC	0.011	0.904	n	-0.143	-7.422	<0.001	0.153
			a	1.015	3.473	<0.001	1.092
			c	0.248	7.192	<0.001	0.267
			λ	0.174	3.806	<0.001	0.187
			t	-0.221	-5.111	<0.001	0.238
			$a * \lambda$	-0.334	-3.315	0.001	0.360

Table A.1: Linear mixed effect model predicting trial-by-trial mood ratings. n denotes the difficulty or number of letters; a indicates the accuracy percentage displayed in the most recent performance feedback; c indicates the objective not-displayed correctness; λ represents the number of trials elapsed since last feedback was displayed; t is the trial number, which we use to infer the mood drift over time effect and time in task effect. The degree of freedom for all the independent variables in the models is 206.

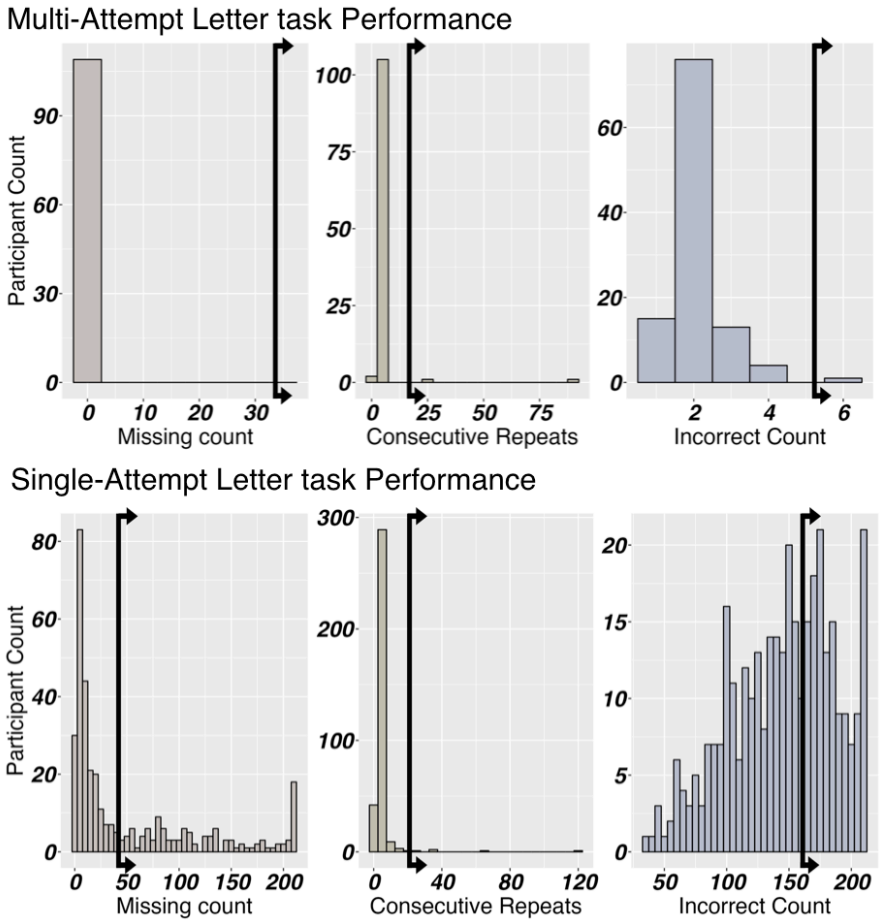


Figure A.3: Participant Performance in both studies. Top panel demonstrate participants recruited for the Multi-Attempt Letters Task and the bottom panel showed the ones in Single-Attempt Letters Task. For both studies, from left to right, the graphs illustrate the missing trial count, highest number of consecutive repeats (across trials) and inaccurate counts. The vertical black bar in each graphs showed the cut-off of the exclusion criteria, with arrows pointing toward excluded data points.

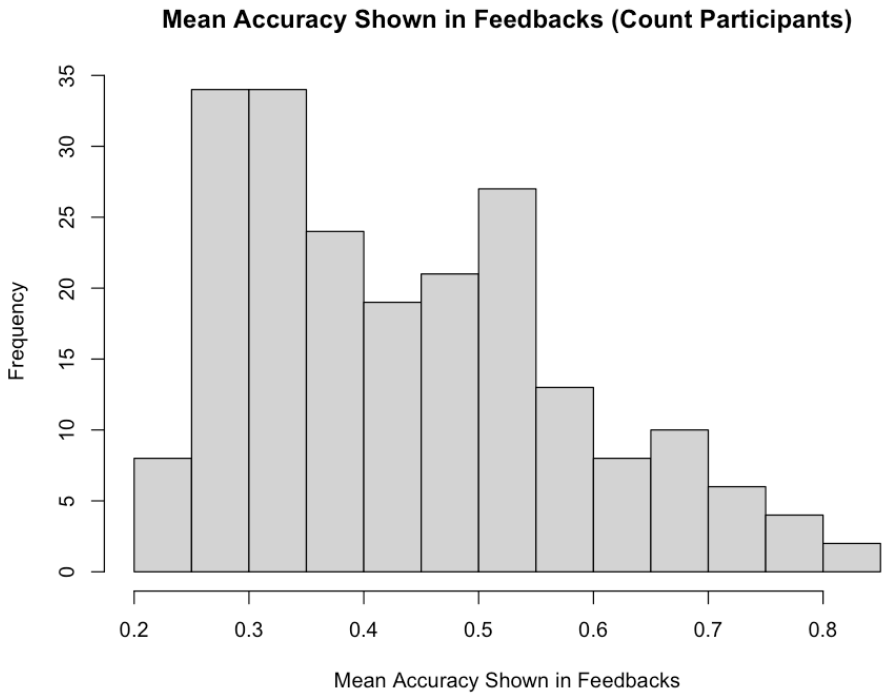


Figure A.4: The majority of the participants seen negative performance feedbacks. Each count represent one participant's overall mean accuracy as displayed in the in the Single-Attempt Letter Task.

Appendix B

Supplemental information for chapter 6

B.1 Stop-word related analysis

With identical decoding models using PSD, we exclude words that are considered ”meaningless” such as transitional words, ”the”, ”a” etc., and observe difference in model performance. We identified that the pattern of generalization across time has changed. In the no stop-word models, we found that the models are peaking ROC AUC more on the edge.

B.1.1 Significant greater similarity for epoch pairs that are within a story

To illustrate consistency, we calculated cosine similarity between any two epochs for all participant, this should indicate topographic consistency between two epochs. A reminder that one epochs now consists of 2 seconds of data that may include multiple words. The label for each epoch is decided by the highest duration. First, we concatenated all power bands and showed that across emotional states, and calculated the mean difference between pairs that come from one story versus ones that come from different stories. While the difference is extremely small (95% confidence interval: 0.00061, 0.00097), it is statistically significant (two independent sample t-test, t=8.703, df=77906401, $p < 0.001$).

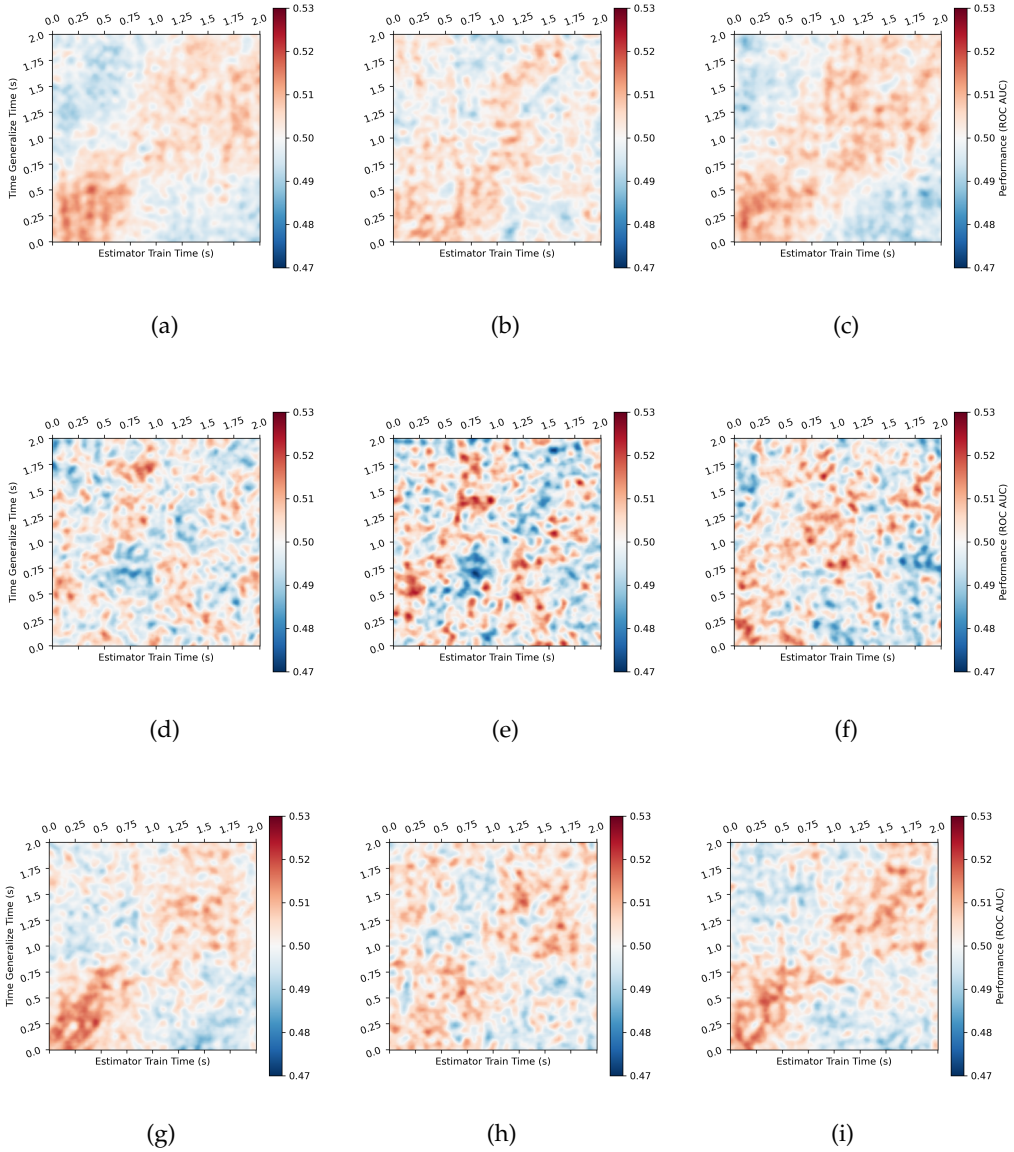
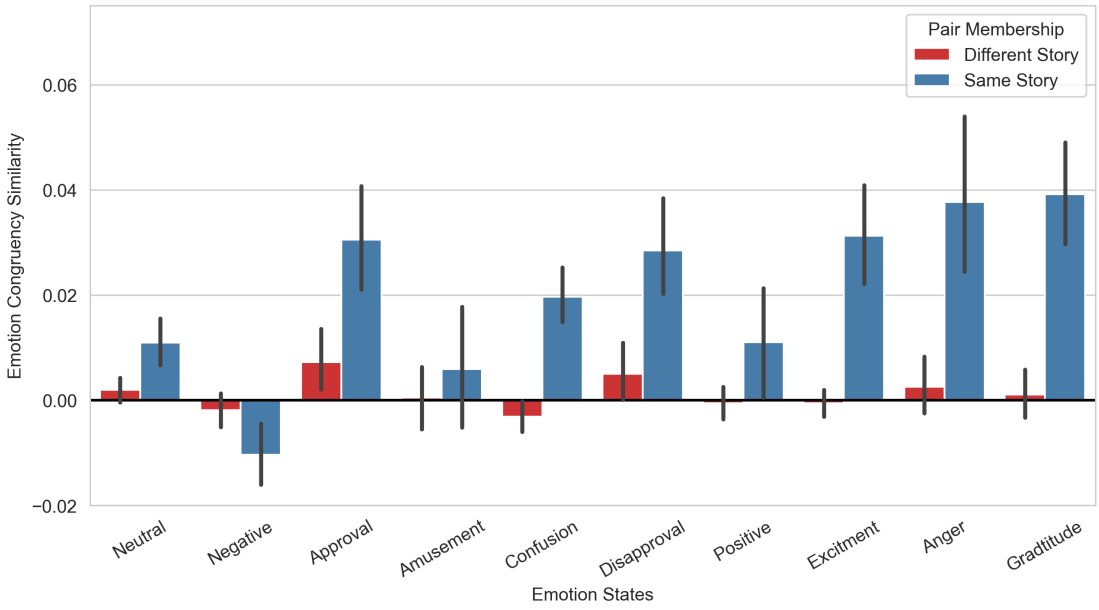


Figure B.1: Illustrations of the logistic regression on classifying HMM states's out-of-sample performance. Performance is characterized by the Area under the curve of the Receiver Operating Characteristic curve (plotting the true positive rate going over false positive rate). A perfect classifier will have 1 as the ROC AUC. Here, each cell in the heat-map corresponds to the ROC AUC of an estimator trained on specific time since onset (y) test on out-of-sample dataset's time since onset (x). Therefore, a peak in the diagonal line indicate a classifier has good specificity. For all 3 components, the ROC AUC heat maps indicate that the hmm raw model is not very specific, nor is it well performing. Top row shows the average ROC AUC across participants and the bottom rows showed the max across participants; middle row illustrate the ROC AUC but with stop-words filtered out; and the bottom row showed ones with only stop words.

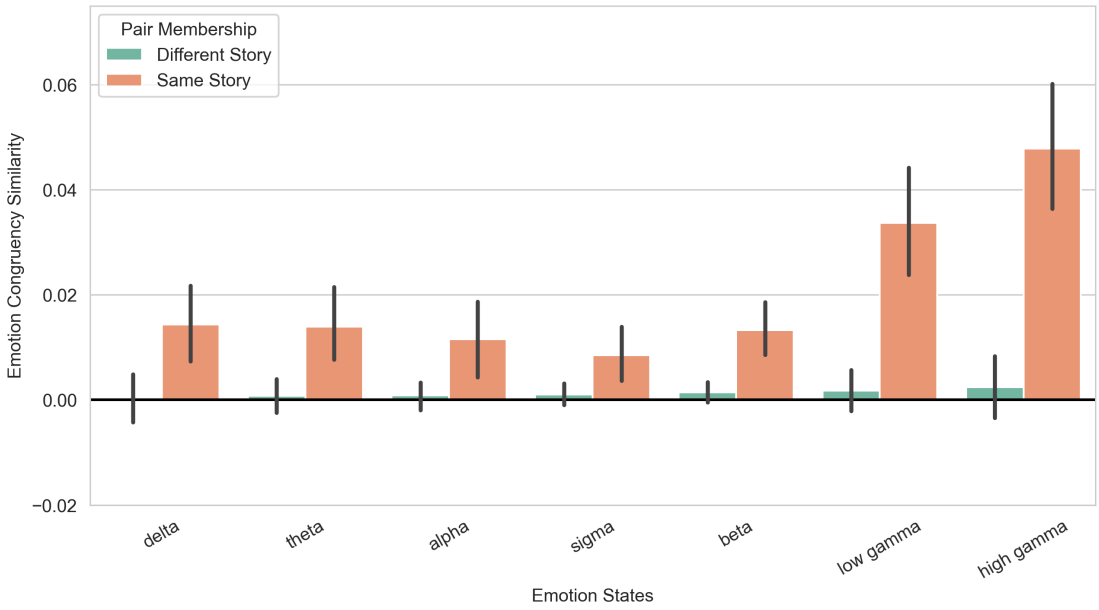


(a)

Figure B.2: Emotion congruency similarity (defined as the cosine similarity difference between emotion congruent pairs of epochs and the incongruent ones) by emotions (x-axis) and story membership (color). Error bar indicate 95% confidence interval.

B.1.2 Emotionally congruent pairs showed greater within story similarity

However, we are more interested in how could this difference meaningfully change our decoder performance. Therefore, we then separately calculated the within story similarity (mean difference of within minus between story pairs) for each emotion and emotion congruency (whether the two epochs in the pair share the same emotion label or not). We showed that in Figure B.2 that nearly across all emotions, when only looking at pairs that are emotionally congruent, the within story similarities are significantly greater than between pairs (except for negative emotional state). The results for emotionally incongruent pairs are inconsistent. However, this is expected as the incongruent pairs consists many different comparisons. One surprising result is that only for incongruent amusement pairs, the within story similarity is significantly positive. That is, regardless of what other amusement paired against, the similarity within a story is greater than between ones.



(a)

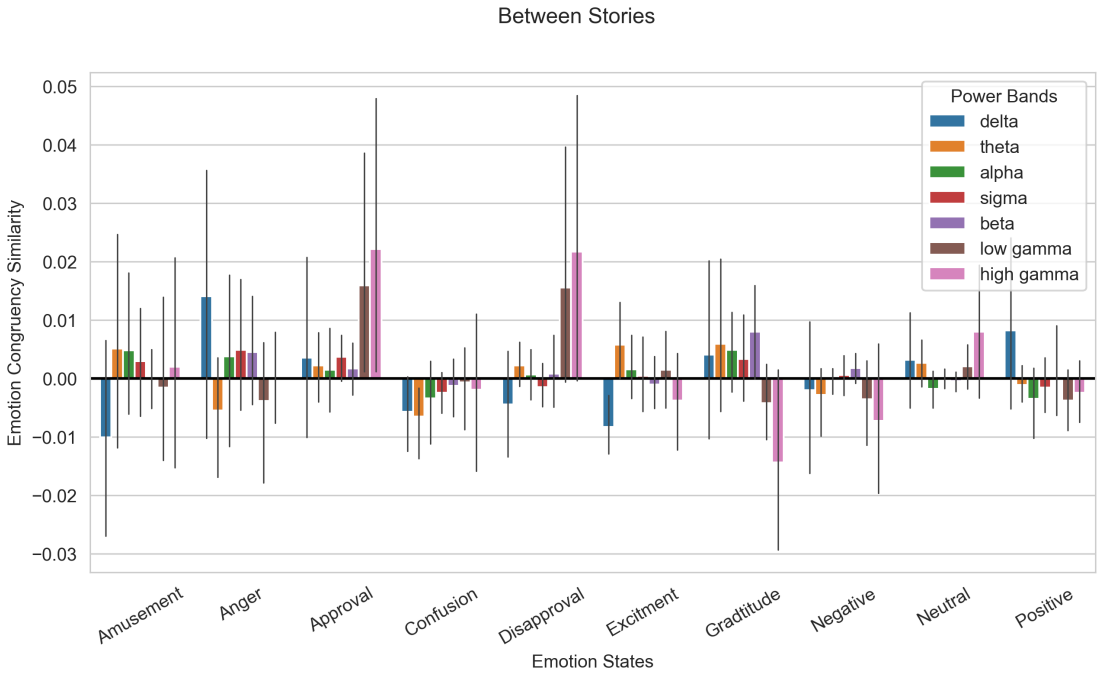
Figure B.3: Emotion congruency similarity by power band (x-axis) and story membership (color). Error bar indicate 95% confidence interval.

B.1.3 Within story pairs showed greater consistency across emotion

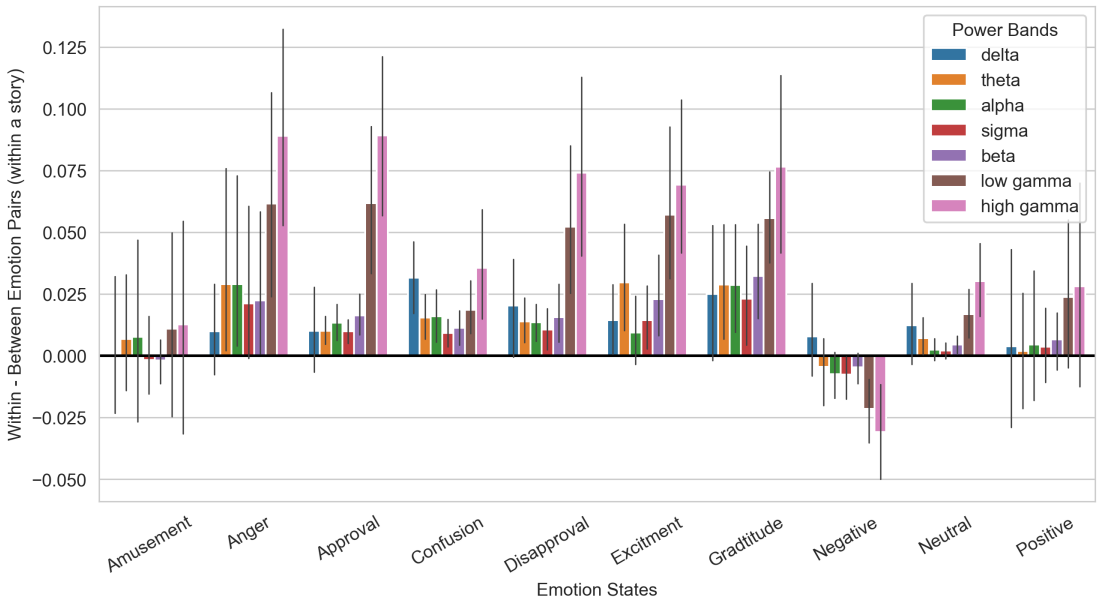
Similarly, if we calculate the emotion congruency similarity (mean difference between within emotion pairs against between ones), we also show that there's greater emotion congruency similarity for pairs within a story (Figure B.2).

B.1.4 Between story emotional congruent pairs show limited consistency

If we further examine the emotional congruency similarity only in the between story pairs, we found that only high / low gamma bands for approval, theta bands for confusion, and delta band for excitement showed significant greater similarity for emotional congruent pairs than incongruent ones (Figure B.4a). Comparing to Figure B.4b, which shows the within story similarity, and almost all band/emotion combinations are significantly different than 0, showing good indication that emotional states within a story is much easier to differentiate. One surprise finding is that for the negative state, emotion incongruent pairs are more similar than the congruent one.

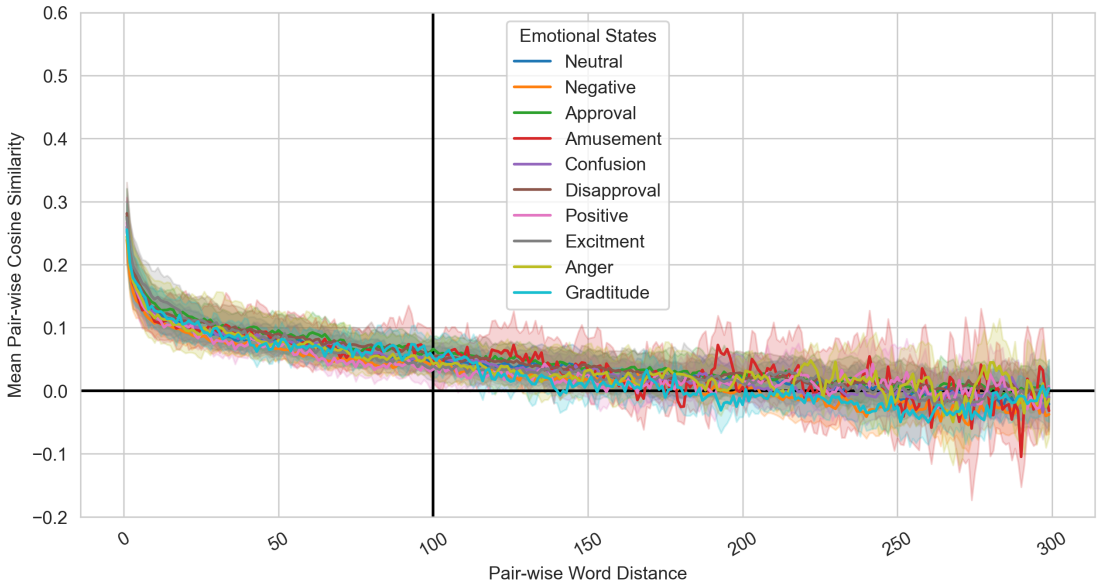


(a)



(b)

Figure B.4: Both graphs showed emotion congruency similarity by emotion and power bands. (a) showed only the cross story pairs, e.g. delta band topography similarity between two epochs that are both amusement from different stories. (b) includes pairs from within a story. Comparing (a) and (b) showed that within story pairs tend to have much greater similarity.



(a)

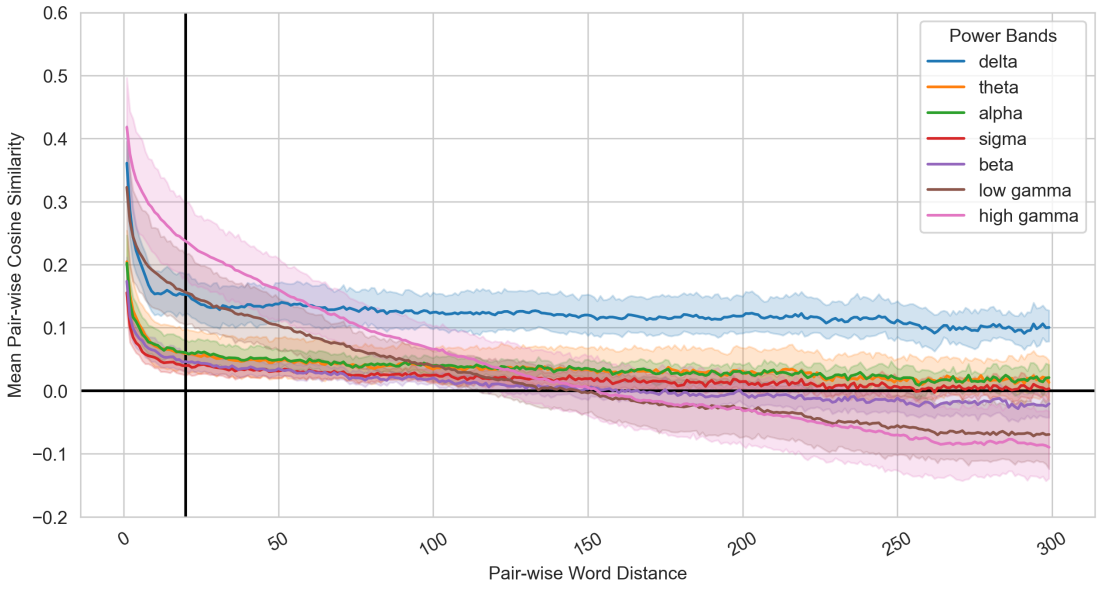
Figure B.5: Cosine similarity (y-axis) by word distance (x-axis) and split by emotions. This showed that autocorrelation for each emotion tend to persist at least 100 words.

B.1.5 Autocorrelation lead to greater emotion congruency similarity

Autocorrelation in a story might contribute to why within story similarity is much more differentiable than between story. In Figure B.5 we showed that the cosine similarity in topography only start to significantly de-correlate after 100 words, indicating that likely the source of high similarity within story is driven by autocorrelation.

B.1.6 Different power bands de-correlate at different timescale

Intuitively, different power bands de-correlate at different time-scale as the effective sample taken into each band's power calculation is time dependent (Figure B.6). We noticed an interesting issue where the gamma bands did not simply de-correlate but also goes further down to have significant negative correlation. That is, pairs including an early and late epoch for the gamma band showed significant anti-correlation. There are a two main possible reasons for this: 1) measurement noise: high frequency bands like the gammas are more sensitive to muscle tension, could be that participants tend to be more relaxed toward the end of the stories. Similarly, slow drift in participants' head position might also contribute to this anti-correlation 2) semantic processing related: prior research suggest that gamma synchrony is related to perceptual encoding,



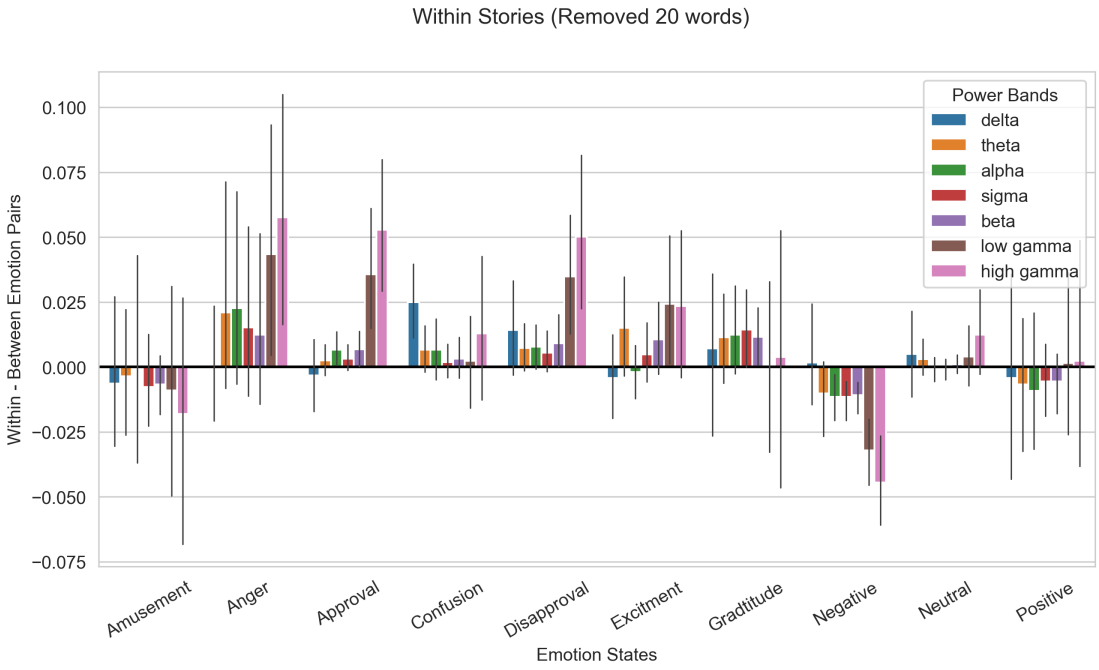
(a)

Figure B.6: Cosine similarity (y-axis) by word distance (x-axis) and split by power bands. This showed that autocorrelation for each power band tend to persist at least 20 words, regardless of emotion states.

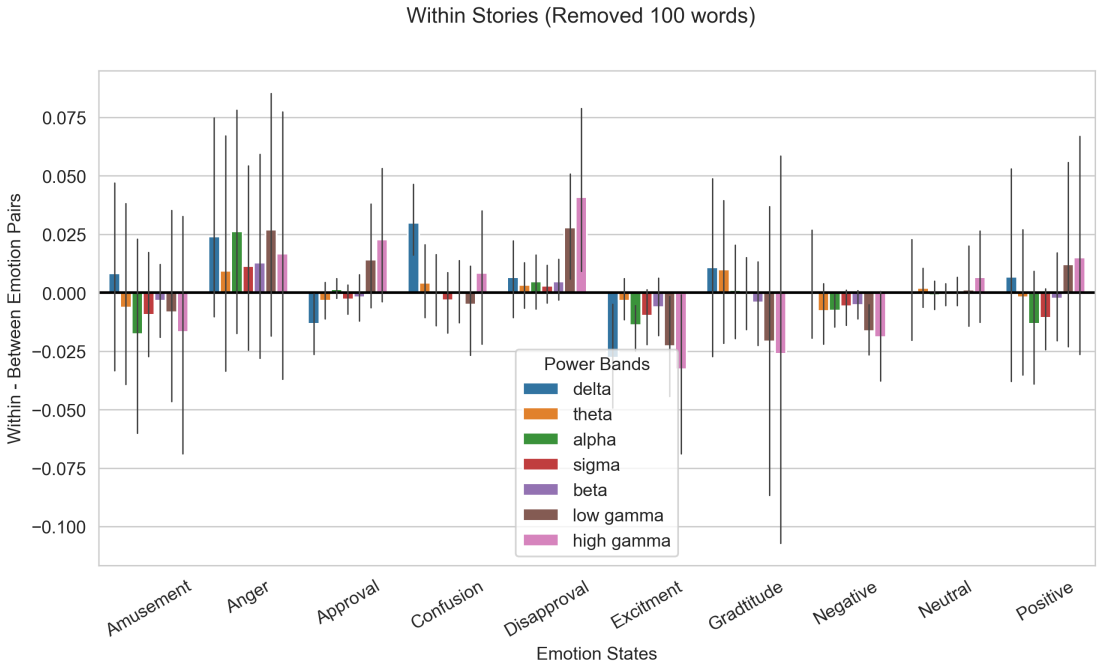
selective attention, salience, and working memory, it is possible that as the story ”dies down”, the need for these cognitive functions decrease.

B.1.7 Removal of neighboring epochs reduced emotion congruency similarity within a story

Because we now have a threshold where we expect the autocorrelation to drop off (100 if we mainly want to reduce autocorrelation related to emotion or 20 for power bands), we can then remove pairs that are less than these threshold and examine whether the emotion congruency similarity persist. As showed in both panels in Figure B.7, we see that the within story similarity now are more similar to the between story ones.



(a)



(b)

Figure B.7: Both graphs showed emotion congruency similarity by emotion and power bands. (a) showed removal of 20 neighboring words while (b) showed removal of 100. We see most band-emotion pairs did not survive the removal, with exception of delta band predicting confusion, and the two gamma bands predicting disapproval.