Imperial College London

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

A Synergistic Core for Large Language Models: Partial Information Decomposition in Machine Learning

Author: Pedro Urbina Rodriguez

Supervisor: Dr Pedro Mediano

Submitted in partial fulfillment of the requirements for the MSc degree in Advanced Computing of Imperial College London

April 2024

Abstract

As Machine Learning (ML) systems and large neural networks grow more capable, they increasingly serve as intriguing case studies for exploring the emergence of cognition in complex information processing systems. There has been a notable surge in efforts to understand and interpret these systems, making Information Theory (IT) and its recent extension, Partial Information Decomposition (PID), particularly relevant as conceptual and mathematical frameworks for this endeavour. PID enhances IT by introducing the concept of *sameness* of information, which enables the decomposition of multivariate information into three categorically distinct types: *redundant*, *synergistic*, and *unique* information. This work reviews recent studies that employ PID to both theoretically underpin ML systems and to interpret and enhance their functionality. This is achieved by leveraging PID's ability to provide conceptual clarity and a fine-grained view of the information dynamics within these systems. We also explore how ML systems themselves are being utilized to compute key PID quantities, highlighting a bidirectional relationship where each domain informs and advances the other.

Additionally, we utilize PID to conduct an in-depth analysis of the information processing dynamics within Large Language Models (LLMs), specifically using the open-source model Gemma [1]. Our findings suggest that, similar to the human brain, these models feature a synergistic information processing core potentially linked to their impressive cognitive capabilities. We provide evidence for this link by demonstrating that different cognitive tasks engage distinct sets of attention heads: higher-level cognitive tasks predominantly activate the synergistic core, while lower-level tasks primarily stimulate the redundancy-dominated layers.

Contents

1 Introduction

Machine Learning (ML) systems, particularly neural networks, are becoming increasingly powerful. Yet, we still lack a firm theoretical understanding of their representation learning processes and inner workings after training. Beyond ML, we aspire to develop a *deep theory of cognition*—one that can explain not only how intelligent behavior emerges from neural networks and biological brains but also how such intelligence could arise in any complex system.

To understand the fundamental factors underpinning advanced cognition, we must abstract away unnecessary details from these intelligent systems. Viewing them as information processing systems allows us to employ Information Theory (IT) as a semantics-free framework in our quest for a unifying theory of cognition. Although classical IT has been successful, it falls short in disentangling the qualitatively different ways that information sources contribute to predicting a target variable—an often encountered challenge in understanding complex systems.

Fortunately, Partial Information Decomposition (PID) [2] has emerged as an extension of classical IT to remedy these shortcomings. PID's primary contribution is its ability to provide conceptual clarity about the different types of information that a set of source variables encode about a target variable. For instance, in a system with two information sources about a target, the information can be decomposed into *unique information* provided exclusively by each source, *redundant information* repeatedly present in by both sources, and *synergistic information* that only emerges when both sources are considered together.

This conceptual clarity is paramount, as we argue in this work, for a deep understanding of the principles governing these systems. Specifically, we explore how PID has been applied to understand and enhance current ML paradigms, establishing it as a strong theoretical foundation to interpret and improve these systems. Moreover, we illustrate the potential of this framework to shed light on how advanced cognitive-like features might emerge in state-of-the-art architectures such as the Transformer [3].

The main contributions of this study are as follows:

- We conduct an extensive review of recent studies at the intersection of PID and ML, highlighting the application and impact of PID.
- We identify a synergistic information processing core in the Gemma Large Language Model (LLM) [1], similar to the synergistic core observed in the human brain [4].
- We provide empirical evidence showing that LLMs differentially employ attention heads to solve distinct cognitive tasks.
- We demonstrate that LLMs predominantly utilize synergistic layers for higher-order cognitive tasks and redundant layers for lower-level cognitive tasks.

This report is structured to provide a coherent flow from foundational concepts and studies to advanced applications and empirical findings. Section 2 introduces the foundational concepts and frameworks of IT and PID. Section 3 reviews some applications of PID in Neuroscience, Cognitive Science, and Complex Systems. This section serves as motivation for the subsequent Section 4, where we provide an in-depth review of recent work at the intersection of PID and ML. Section 5 details our empirical findings regarding the informational processing architecture of the Gemma LLM. The report concludes with Section 6, where we summarize the key insights and potential implications of our findings, suggesting avenues for future research.

2 Background

Essential to this work is the understanding of Information Theory (IT) and its recent extension to Partial Information Decomposition (PID). We provide an introduction to these two fundamental topics in the following sections.

In this work, we will denote random variables with capital letters, X, and their associated support with the calligraphic character $\mathcal{X}$. Particular realizations of the random variable will be represented using lowercase letters, $x \in \mathcal{X}$. Additionally, in line with the conventions in the field of Information Theory, we will express the probability density function (PDF) as $p_X(x) := \mathbb{P}(X = x)$ for $x \in \mathcal{X}$, where $\mathbb{P}$ is a suitably defined probability measure on $\mathcal{X}$. This can be simplified to $p(x)$ when the associated random variable is clear from the context.

2.1 Information Theory (IT)

In his seminal 1948 paper, Claude E. Shannon laid the foundational groundwork for the field of Information Theory [5]. He formalized the abstract concept of *information* contained in a random variable, defining it as the expected reduction in uncertainty upon observing the variable's outcome. The intuition behind this idea is that an event with a wide array of possible outcomes delivers more informational value than one with fewer possibilities, or in a deterministic case, where no new information is gained (information gain is zero).

Mathematically, the surprise of observing a specific event $x \in \mathcal{X}$ from a random variable X is quantified as $h(x) = -\log p(x)$. The entropy of X, which measures the total information contained in X, is given by $H(X) = \mathbb{E}_X[h(x)] = -\sum_{x\in\mathcal{X}} p(x)\log p(x)$. Similarly, the joint entropy of two random variables X and Y, considering their joint probability density function $p(x, y)$, is defined as $H(X, Y) = \mathbb{E}_{X,Y}[-\log p(x, y)] = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x, y)\log p(x, y)$.

An equally important concept is the idea of conditional entropy. Intuitively, this measures the entropy or information remaining in a random variable Y given that we know the value of a random variable X. Using the conditional probability density function $p(y|x)$, the conditional entropy $H(Y|X)$ can be defined as $H(Y|X) = \mathbb{E}_{X,Y}[-\log p(y|x)]$. Thus, the conditional entropy can be viewed as $H(Y|X) = H(X, Y) - H(X)$, representing the difference between the joint entropy of X and Y and the entropy of X alone.

One of the key concepts in classical IT is the notion of Mutual Information (MI). MI measures the dependency between two random variables by quantifying the information shared between them. This concept serves as a generalized measure of correlation and is mathematically defined as

$$I(X;Y) = \mathbb{E}_{X,Y}\left[\log \frac{p(x,y)}{p(x)p(y)}\right] = \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}.$$

It is essential to note that MI is non-negative, reaching a minimum value of zero exclusively when X and Y are independent, indicating no shared information. Additionally, we can now decompose the total entropy of Y as $H(Y) = I(X;Y) + H(Y|X)$.

Expanding on the concept of conditional entropy, we can introduce conditional mutual information (CMI). CMI represents the information shared between two random variables X and Y, conditioned on the knowledge of a third random variable Z. The conditional MI is defined as

$$I(X;Y|Z) = \mathbb{E}_{X,Y,Z}\left[\log \frac{p(x,y|z)}{p(x|z)p(y|z)}\right] = \sum_{x\in\mathcal{X},y\in\mathcal{Y},z\in\mathcal{Z}} p(x,y|z)\log\frac{p(x,y|z)}{p(x|z)p(y|z)}.$$

2.2 Partial Information Decomposition (PID)

2.2.1 Motivation

Classical IT extends straightforwardly to multivariate settings by incorporating multivariate random variables and their associated probability distributions. However, in these settings, complex interactions emerge among the individual random variables—interactions that classical IT was not originally designed to capture and explore. Consequently, classical IT can only provide a coarse-grained perspective on these interactions.

Partial Information Decomposition (PID), on the other hand, was specifically developed to address the shortcomings of classical IT by more effectively disentangling the interactions among information sources within a multivariate setting. Its aim is to provide a more detailed view of these interactions. The key innovation of PID is the introduction of the concept of *sameness* of information [6], which helps in distinguishing the unique contributions of each variable.

To further elucidate the shortcomings of classical IT, consider, for instance, a system with two source variables X_1 and X_2 trying to predict a target variable Y. The PID framework seeks a deep analysis of $I(Y; X_1, X_2)$, focusing on how the information from each source contributes either redundantly, uniquely, or synergistically to the understanding of Y.

To further dissect the mutual information $I(Y; X_1, X_2)$ using classical IT, one might consider the individual mutual information terms $I(Y; X_1)$ and $I(Y; X_2)$. A key observation is that the sum of these two terms does not necessarily equal $I(Y; X_1, X_2)$. The following two examples elucidate the factors contributing to the discrepancy.

First, information may be provided redundantly by the source variables. For instance, if X_2 is merely a copy of X_1, then X_2 provides no additional information about Y once X_1 is known, resulting in $I(Y; X_1, X_2) = I(Y; X_1) = I(Y; X_2)$ and characterizing the sources as redundant. In this case we have $I(Y; X_1, X_2) < I(Y; X_1) + I(Y; X_2)$.

Second, the information from the sources can be synergistic, where the combined information from both variables exceeds the sum of their individual contributions. An example of this is the 2-bit XOR gate, where $Y = X_1 \oplus X_2$. Independently, each input bit reveals no information about the output ($I(Y; X_1) = I(Y; X_2) = 0$), but together they completely determine Y. Therefore, $I(Y; X_1, X_2) > I(Y; X_1) + I(Y; X_2)$.

Classical IT does not separate these synergistic and redundant elements clearly. As observed, the co-information, defined as

$$I(Y; X_1; X_2) = I(Y; X_1, X_2) - (I(Y; X_1|X_2) + I(Y; X_2|X_1)),$$

can be nonzero, indicating the presence of non-trivial interactions.

The primary objective of the PID framework is to systematically categorize these interactions—synergy and redundancy—allowing a nuanced decomposition of how a multivariate set of sources informs a target. This facilitates a deeper exploration into the structure of multivariate information decomposition [2].

2.2.2 The PID Framework

Williams and Beer [2] were interested in disentangling how each variable in a source random vector $X = (X_1, \ldots, X_n)$ contributes information about a target random variable Y. Specifically, in the case of two source variables ($n = 2$), they identified three distinct components of information about Y that X_1 and X_2 can contain:

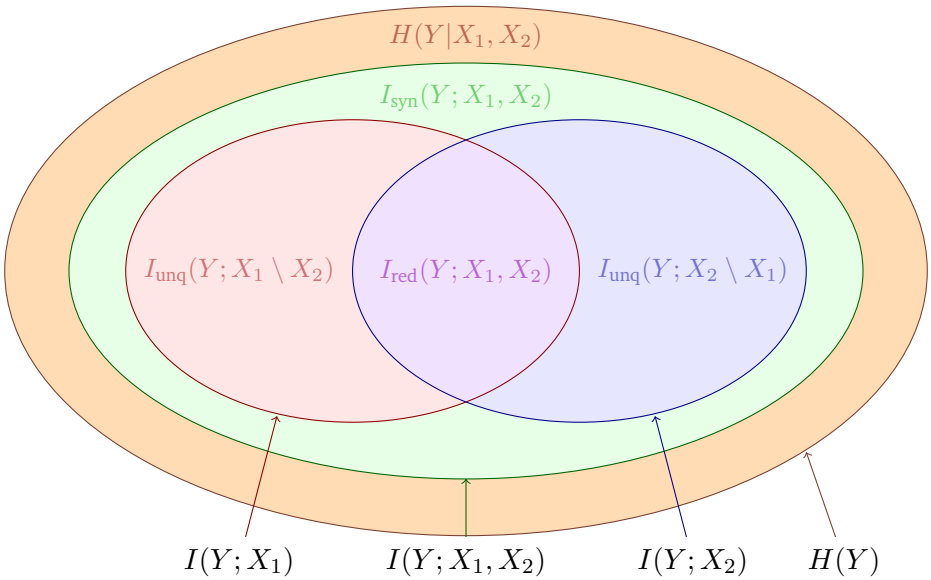


Figure 1: Decomposition of $H(Y)$. Each colored label inside the diagram corresponds directly to the area shaded in the same color. Black labels outside the outermost oval, connected by arrows, represent the total area or information content encompassed by each oval they point to.

- Redundant Information: $I_{red}(Y; X_1, X_2)$. Information about the target that is present repeatedly in both source variables.
- Unique Information: $I_{unq}(Y; X_1 \setminus X_2)$ and $I_{unq}(Y; X_2 \setminus X_1)$. Information about the target that is specific to one source and absent in the other.
- Synergistic Information: $I_{syn}(Y; X_1, X_2)$. Information about the target that emerges only when both source variables are considered together, not discernible when analyzing the sources individually.

These quantities, jointly called PID atoms, provide the decomposition of the MI $I(Y; X_1, X_2)$ that we were looking for. Specifically we have that

$$I(Y; X_1, X_2) = I_{red}(Y; X_1, X_2) + I_{unq}(Y; X_1 \setminus X_2) + I_{unq}(Y; X_2 \setminus X_1) + I_{syn}(Y; X_1, X_2). \quad (1)$$

The intuitive meaning of Equation 1 is that we can decompose the information provided by two sources about a target variable into distinct components: the unique information each source contributes, the redundant information shared by both sources, and the synergistic information that emerges only when both sources are considered together. Figure 1 offers a visual representation of this decomposition. This figure also shows the intuitive decomposition the individual MI terms

$$I(Y; X_1) = I_{unq}(Y; X_1 \setminus X_2) + I_{red}(Y; X_1, X_2), \quad (2)$$
$$I(Y; X_2) = I_{unq}(Y; X_2 \setminus X_1) + I_{red}(Y; X_1, X_2). \quad (3)$$

Equations 1 to 3 illustrate the relationship between classical IT metrics and PID atoms. While classical IT yields three distinct quantities, PID introduces four atoms, indicating that the system is underdetermined. To compute these PID atoms, an additional method is required to estimate one of them. In their seminal work, Williams and Beer [2] tackle this by defining

redundancy as

$$I_{red}(Y; X_1, X_2) = I_{min}(Y; X_1, X_2) = \sum_{y \in \mathcal{Y}} p(y) \min\{I(Y = y; X_1), I(Y = y; X_2)\},$$

where $I(Y = y; X_i)$ denotes the specific information. However, several alternative measures of redundancy have been proposed since Williams and Beer's initial work, leading to ongoing debates within the field. Currently, there is no consensus on which measure best captures the notion of redundancy.

2.2.3 Lattice Structure of PID

In the previous section, we explored the scenario involving only two source variables, X_1 and X_2, which is the most commonly used case in the studies we review in subsequent chapters. However, the PID framework can be extended to any number of source variables $X_1, \ldots, X_n$, revealing the intricate lattice structure that underlies this decomposition. In this section, we delve into the mathematical foundations of PID, uncovering the lattice structure initially described by Williams and Beer [2].

The key insight of Williams and Beer involves understanding the structure of multivariate targeted information, $I(Y; X_1, \ldots, X_n)$, as a combination of redundancies and synergies among the source variables $X_1, \ldots, X_n$. For example, some information about Y may be encoded both by X_2, and redundantly by the synergistic combination of X_1 and X_3, denoted as $\{X_1, X_3\}$. In general, parts of the information contained in the target variable Y are redundantly explained by the synergistic interactions among the source variables. Our aim is to elucidate the structure of this space characterized by redundancies of synergies.

Therefore, we need to consider every possible combination of redundancies of synergies. Synergies can emerge from any subset of source variables $X = \{X_1, \ldots, X_n\}$, requiring us to consider any $A \in \mathcal{P}^*(X)$, where $\mathcal{P}^*(X) := \mathcal{P}(X) \setminus \{\varnothing\}$, as potential candidates for synergistic interactions. Consequently, we must address any redundancy that arises from these synergies, forcing us to navigate the vast space of $\alpha \in \mathcal{P}^*(\mathcal{P}^*(X))$.

However, a key observation allows us to simplify the space of possible redundancies. Suppose $A_i, A_j \in \mathcal{P}^*(X)$ with $A_i \subseteq A_j$. To distinguish from the two-dimensional case, we denote the redundancy between two sets of sources as $I_\cap(Y; \{A_i, A_j\})$. It follows that the redundancy between A_i and A_j can be no greater than that of A_i alone; that is, $I_\cap(Y; \{A_i, A_j\}) = I_\cap(Y; \{A_i\}) = I(Y; A_i)$. For instance, if considering the redundancy between X_1 and the synergistic interaction of X_1 and X_2, denoted $\{X_1, X_2\}$, it becomes evident that $I_\cap(Y\{\{X_1\}, \{X_1, X_2\}\}) = I_\cap(Y; \{X_1\}) = I(Y; X_1)$. This directly implies that

$$I_\cap(Y; \{A_i, A_j, A_k\}) = I_\cap(Y; \{A_i, A_k\}) \text{ if } A_i \subseteq A_j, \text{ for all } A_i, A_j, A_k \in \mathcal{P}^*(X). \tag{4}$$

This key observation allows us to greatly restrict our domain for studying redundancies. Instead of considering any $\alpha \in \mathcal{P}^*(\mathcal{P}^*(X))$, we can now, without loss of generality, define the new redundancy domain

$$\mathcal{A}(X) = \{\alpha \in \mathcal{P}^*(\mathcal{P}^*(X)) : \forall A_i, A_j \in \alpha, A_i \not\subset A_j\}.$$

Each element $\alpha \in \mathcal{A}(X)$ is referred as an antichain. Williams and Beer [2] establish a partial order over the antichains in $\mathcal{A}(X)$:

$$\alpha \preceq \beta \Leftrightarrow \forall B \in \beta, \exists A \in \alpha, A \subseteq B.$$

This partial order endows the redundancy domain $\mathcal{A}(X)$ with a lattice structure, referred to as the redundancy lattice. In this lattice, each node includes all redundancies corresponding to the lower atoms within the structure. Ascending through the lattice corresponds to an accumulation of information.

In the two-dimensional case, the lowest node in the lattice corresponds to $I_\cap(Y; \{\{X_1\}, \{X_2\}\})$, which in turn corresponds to the redundancy $I_{red}(Y; X_1, X_2)$. The nodes immediately above, $I_\cap(Y; \{\{X_1\}\})$ and $I_\cap(Y; \{\{X_2\}\})$, correspond to $I(Y; X_1)$ and $I(Y; X_2)$ respectively. The highest node in the lattice, $I_\cap(Y; \{\{X_1, X_2\}\})$, represents the total MI $I(Y; X_1, X_2)$.

By taking the partial differences over the redundancy function $I_\cap$, we identify the information that each set of sources uniquely contributes—information not already provided by any other set of sources lower in the lattice. This concept of partial information differences is captured by the function I_∂, or PI-function, which is defined implicitly by

$$I_\cap(Y; \alpha) = \sum_{\beta \preceq \alpha} I_\partial(Y; \beta). \tag{5}$$

$I_\partial(Y; \alpha)$ quantifies the marginal redundancy of an antichain α within the redundancy lattice. It represents the difference between the redundancy contributed by α and the aggregate redundancy of all nodes β positioned lower in the lattice. This is formally computed using a Möbius inversion, resulting in $I_\partial(Y; \alpha) = I_\cap(Y; \alpha) - \sum_{\beta \prec \alpha} I_\partial(Y; \beta)$. This measure, $I_\partial(Y; \alpha)$, is referred to as the PID atom associated with the antichain α.

2.2.4 Revisiting the Two-Dimensional Case

To connect the discussions from the previous two sections, let us examine the redundancy lattice more closely when we are dealing with only two sources, X_1 and X_2. The redundancy lattice consists of

$$\mathcal{A}(X) = \{\{\{X_1\}, \{X_2\}\}, \{\{X_1\}\}, \{\{X_2\}\}, \{\{X_1, X_2\}\}\},$$

with the established partial ordering

$$\begin{aligned} \{\{X_1\}, \{X_2\}\} \prec \{\{X_1\}\}, \quad \{\{X_1\}, \{X_2\}\} \prec \{\{X_2\}\}, \\ \{\{X_1\}\} \prec \{\{X_1, X_2\}\}, \quad \{\{X_2\}\} \prec \{\{X_1, X_2\}\}. \end{aligned}$$

By applying Equation 5 to various antichains, we derive

$$I(Y; X_1, X_2) = I_\cap(Y; \{\{X_1, X_2\}\}) = \sum_{\beta \preceq \{\{X_1, X_2\}\}} I_\partial(Y; \beta), \tag{6}$$

$$I(Y; X_1) = I_\cap(Y; \{\{X_1\}\}) = \sum_{\beta \preceq \{\{X_1\}\}} I_\partial(Y; \beta), \tag{7}$$

$$I(Y; X_2) = I_\cap(Y; \{\{X_2\}\}) = \sum_{\beta \preceq \{\{X_1\}\}} I_\partial(Y; \beta). \tag{8}$$

We have already established that $I_\cap(Y; \{\{X_1\}, \{X_2\}\}) = I_\partial(Y; \{\{X_1\}, \{X_2\}\}) = I_{red}(Y; X_1, X_2)$, as it represents the lowest node in the lattice. Combining Equations 2 and 7 results in $I_\partial(Y; \{\{X_1\}\}) = I(Y; X_1) - I_{red}(Y; X_1, X_2) = I_{unq}(Y; X_1 \setminus X_2)$. Similarly, Equations 3 and 8 lead to $I_\partial(Y; \{\{X_2\}\}) = I_{unq}(Y; X_2 \setminus X_1)$. Finally, integrating these results with Equations 1 and 6 concludes that $I_\partial(Y; \{\{X_1, X_2\}\}) = I_{syn}(Y; X_1, X_2)$.

In summary, we have demonstrated that the PID atoms discussed in Section 2.2.2 correspond precisely to the PI-function derived from first principles in Section 2.2.3. Furthermore, Equations 1, 2, and 3 correspond exactly with Equations 6, 7, and 8, respectively.

3 PID in Neuroscience, Cognitive Sciences and Complex Systems

This section offers a concise review of the fields where Information Theory, and Partial Information Decomposition in particular, have been effectively applied. PID has proven invaluable in disentangling various types of information and enhancing our understanding of the dynamics within different information processing systems. The insights gained from these applications provide a strong motivation for the recent uses of IT and PID to better understand and enhance Machine Learning Systems, topics we will explore in depth in the next chapter.

3.1 Neuroscience

The field of neuroscience has benefited enormously from viewing the brain as an information processing system, with neurons as its information processing building blocks [7, 8, 9]. Neurons are assumed to transmit information from one to another, and in the process, they modify, copy, combine, omit and store information. As such, it has been extremely fruitful to use the lens of the PID framework to decompose and obtain a more fine-grained view of these information dynamics.

Luppi et al. revisit these contributions in their recent work [9]. They suggest that redundant information contributes to robustness by duplicating the same information across different locations, albeit at the cost of reduced expressional capacity. Conversely, synergistic information allows the network to encode information more efficiently but introduces fragility; if any one of a set of n neurons that synergistically encode a piece of information fails, that information becomes inaccessible. However, when information is redundantly encoded across different sets of n neurons, the failure of a single neuron does not compromise the information, although it may lead to suboptimal use of the neurons' information capacity.

Integration is a central concept in neuroscience, often regarded as essential for complex cognitive functions. Integrated Information Theory proposes that a system's ability to integrate information in a way that forms a unified whole, which cannot be decomposed into independent parts, underlies conscious experience [10].

However, the concept of integration remains somewhat vague. Luppi et al. suggest that it can be understood in two distinct ways [9]. *Integration-as-oneness* emphasizes that components should act as one. From a PID perspective, this implies redundancy, where neurons encoding the same information function collectively as one. Conversely, *integration-as-cooperation* involves complementarity, where cooperating neurons collectively provide more information than the sum of their individual contributions, aligning with the notion of synergy in PID. Yet, classical IT struggles to disambiguate these concepts, lacking the necessary conceptual clarity to differentiate between synergy, redundancy, and unique information [9].

In a previous study, Luppi et al. [4] demonstrated a significant correlation between higher-order cognition and synergistic interactions among brain regions, whereas lower sensorimotor functions are associated with redundant interactions. This finding lends support to the concept of *integration-as-cooperation*, particularly as higher cognitive functions often necessitate greater integration. Furthermore, their research reveals that although redundancy remains consistent, synergy increases when comparing the brains of macaques to those of humans. This observation leads the authors to suggest that evolution may favor increased synergistic interactions between brain regions, which in turn could facilitate the capability for more complex cognitive tasks. This study underscores the critical importance of distinguishing between synergy and redundancy in understanding the information processing dynamics of the brain.

However, an important limitation of this study is that it only considers pairwise redundancies and synergies, neglecting higher-order interactions involving three or more brain regions. These more complex interactions could provide incredibly valuable insights into the brain's information processing dynamics. This limitation is a common issue across various studies employing PID, largely due to the computational complexity involved in calculating higher-order interactions. This complexity may be the main factor hindering PID's broader application, despite its considerable conceptual promise. However, in Section 4.4, we review several studies that utilize ML to overcome these shortcomings.

The fine-grained nature of PID also allows for quantifying specific information about a particular feature transferred between two brain regions, going beyond the broad information transfer that classical IT can capture. Celotto et al. [11] sought to quantify the content and direction of information communicated between two brain regions, addressing the limitations of current methods that provide only a coarse-grained view of overall information propagation. Using PID, they are able to do so by examining the information about a specific feature that the sender's past shares (or is redundant with) exclusively with the receiver's present, which is, at the same time, unique with respect to the receiver's own past. Based on this approach, they developed a measure called Feature-specific Information Transfer, which they experimentally tested to confirm its effectiveness in detecting feature- and direction-specific information transfer.

Considering the brain as an information processing entity, a natural question arises: Toward which goal function or underlying principle is the information processing directed? Several neural goal functions have been proposed, including predictive coding, infomax, and coherent infomax. PID allows us to decompose the entropy of a target variable Y according to Equation 1, yielding $H(Y) = H(Y|X_1, X_2) + I_{red}(Y; X_1, X_2) + I_{unq}(Y; X_1 \setminus X_2) + I_{unq}(Y; X_2 \setminus X_1) + I_{syn}(Y; X_1, X_2)$. By introducing coefficients for each of the terms in this decomposition, Wibral et al. [12] crafted a generic goal function encapsulates previous neural goal functions and also created a new neural goal function that exploits synergy.

3.2 Cognitive Sciences and Complex Systems

As in Neuroscience, IT and PID have become invaluable frameworks for studying Cognitive Sciences. Additionally, IT has emerged as a kind of *lingua franca* within the field of complex systems, providing the necessary mathematical tools to examine phenomena like emergence, which are closely associated with the concept of synergy [13, 14]. This section briefly reviews some key works in these two fields that utilize both PID and IT.

In their seminal work, Williams and Beer [15] examine a cognitive agent through two principal mathematical frameworks employed in Cognitive Science: Dynamical Systems Theory (DST) and Information Theory (IT). They utilize evolutionary methods to develop an agent capable of solving a relational categorization task, where the agent's objective is to act based on the relative sizes of two sequentially presented objects. In some instances, they found that the agent stores the size of the first object synergistically in relation to its position and a sensor output, highlighting the capability of PID to offer a detailed perspective on information storage within the evolved cognitive agent. Overall, the authors demonstrate how DST and IT provide complementary and synergistic approaches to understanding cognitive agents.

In the field of Complex Systems, emergence is a key phenomenon, closely linked to the concept of synergy in PID. The interaction between different layers of abstraction in our hierarchical world is not yet fully understood [16], and emergence is often employed as a somewhat lossy concept to describe these interactions. Rosas et al. [16] utilize PID to

develop a formal theory of causal emergence, establishing a sufficient condition for this phenomenon based on the temporal mutual information between the whole system and its component parts. They demonstrate that the synergy among the system's parts, in relation to the system's future, dictates emergent behaviors. This approach is applied to detect causal emergence in various models, including Conway's Game of Life, Reynolds' flocking model, and neural activity.

In a related study, Varley and Hoel [17] investigate causal emergence by examining how information is lost or augmented when transitioning between microscales and macroscales in complex systems. Specifically, they employ Boolean networks and PID to demonstrate that while mutual information must be the same or lower when moving from microscales to macroscales, the composition of this information can differ significantly. They show that the redundant component of mutual information at the microscale is often converted into synergistic information at the macroscale, hinting at the phenomenon of emergence and demonstrating information conversion across scales. This study has profound implications for reductionist theories, revealing that examining systems at macroscales can transform redundant (and ostensibly less useful) information into more informative, synergistic information, even though the total mutual information of the system may remain the same or decrease.

While most PID studies analyze systems to understand the predominant type of information within them, Varley and Bongard explore the implications of a system being primarily synergistic or redundant in [14]. They evolve Boolean networks into two distinct populations characterized by synergistic and redundant information to examine these systems from a Dynamical Systems perspective. Their findings reveal that synergistic Boolean networks, similar to random networks, exhibit chaotic and unstable behavior, yet possess a significant capacity to integrate information. In contrast, redundant Boolean networks display considerable stability but have limited computational capacity. Similar to the findings of Luppi et al. [9] in the human brain and Proca et al. [18] in artificial networks, these results point to a profound relationship and trade-off between synergy and efficiency versus redundancy and robustness. The authors propose that this trade-off might help explain the evolution towards larger brains, where a redundant substrate could stabilize more complex and higher-level cognitive processes characterized by synergy.

4 Partial Information Decomposition and Machine Learning

In this chapter, we explore the intersection of Partial Information Decomposition and Machine Learning. Motivated by the successful applications of PID in the fields discussed in the previous chapter, the ML community has recently embraced PID for various purposes. This review categorizes significant contributions by their intent: providing a theoretical foundation for ML, interpreting and understanding ML systems, and enhancing their performance. Additionally, we examine how ML techniques have been utilized to compute specific PID quantities.

4.1 Information Theory: A Theoretical Foundation for ML

Deep learning is predominantly characterized as an empirical field of research. This stems largely from the inadequate theoretical understanding of the internal mechanisms and representations learned by these ML systems. As these systems evolve in capability and intelligence, there is an urgent need for a theoretical framework that can transition the field from empiricism to a discipline grounded more firmly in theory.

Information Theory and Partial Information Decomposition emerge as promising frameworks to establish these theoretical underpinnings. As discussed in earlier sections, both areas have found successful applications in Neuroscience and Cognitive Sciences. They are now regarded as some of the most relevant paradigms for understanding the human brain and other complex systems.

One of the pioneering applications of Information Theory to the understanding of ML systems was conducted by Tishby and Zaslavsky [19], with further developments by Schwartz-Ziv and Tishby [20]. They conceptualized the supervised learning problem through the interaction between an input random variable X and an output random variable Y, which are considered dependent. The goal is to derive the relevant information $I(X;Y)$ using a minimal sufficient statistic $\tilde{X}$, while ensuring that an informativity threshold $I(\tilde{X};Y) \leq I(X;Y)$ is maintained. They aimed to minimize the complexity of $\tilde{X}$ by reducing $I(X;\tilde{X})$. This framework introduces a trade-off between compression—represented by $I(X;\tilde{X})$—and prediction accuracy, as indicated by $I(\tilde{X};Y)$ [19].

Building on this intuition, they developed the concept of the Information Plane: each layer in a neural network is assumed to compute a sufficient statistic, represented by the distribution T_i of the i-th layer. It is possible to calculate $I(X;T_i)$ and $I(T_i;Y)$ for each layer and plot these values on the Information Plane, with $I(X;T_i)$ on the x-axis and $I(T_i;Y)$ on the y-axis. Although their initial methods for estimating these mutual information metrics were later found to be flawed and subsequently refined [21, 22], their foundational ideas and approaches have significantly influenced subsequent research.

Wollstadt et al. [23] applied the PID framework to establish a theoretical foundation for feature selection, a key area in ML, by effectively disentangling the synergistic, redundant, and unique information contributions of source variables to target predictions. They demonstrate that a feature selection criterion based on CMI is strictly superior to one based solely on MI. Consider the scenario where we are deciding whether to include a second feature variable F_2 in a set currently composed only of F_1 and target variable Y. The distinction between the MI and CMI criteria can be expressed as

$$I(Y;F_2) - I(Y;F_2|F_1) = I_{red}(Y;F_2,F_1) - I_{syn}(Y;F_2,F_1). \tag{9}$$

In feature selection, the goal is to identify a minimal set of feature variables that encapsulate the maximum information about the target variable. Within this context, synergistic infor-

mation among the feature variables is highly beneficial, whereas redundant information is undesirable. According to Equation 9, the MI criterion increases with redundant information and decreases with less synergistic information, thereby making it an inferior criterion for including new feature variables. This conclusion is further supported empirically [23]. This study showcases a powerful recurring theme: the use of PID to overcome classical IT's inability to disentangle redundant and synergistic information, thereby enabling a deeper and more fine-grained understanding of the information dynamics within these ML systems.

In a similar vein, Milzman et al. [24] applied the PID framework to provide a more nuanced understanding of information dynamics within Factor Graphs, which are widely utilized in many ML-related fields, such as robotics and multi-agent systems. These graphs facilitate the representation of multiple sources of heterogeneous information, enabling subsequent inference processes. As additional sources are incorporated, the recurrent trade-off between redundancy and efficiency emerges: redundancy enhances network robustness but can impede the efficiency of information transfer. Echoing the insights from Wollstadt et al. [23], an optimal criterion for information transmission would prioritize CMI, which captures unique and synergistic information while minimizing redundancy. Consequently, Milzman et al. [24] developed a redundancy measure based on PID, specifically tailored to balance the trade-off between efficiency and robustness in factor graphs, which they validated through empirical simulations.

4.2 Using IT and PID for Enhancing Interpretability in ML Systems

The increasing complexity of ML systems, particularly those using deep learning and backpropagation, and their opaque, complicated representations have led researchers to move from a first-principles approach to one centered on interpretability in order to understand current ML systems. Theoretical foundations have lagged behind the quick advancements in algorithmic complexity and performance, necessitating a post-hoc analysis to understand the properties and representations developed during training. Interpretability has thus become crucial for elucidating the decision-making processes and learnt representations of these systems. The aim of this section is to highlight the potential of IT and PID to aid in this endeavor by reviewing key contributions in this area.

One of the initial contributions in this area is the work by Tax et al. [25], which applies PID to analyze the evolution of information-theoretic quantities in Restricted Boltzmann Machines (RBMs) during training on the MNIST dataset. They observed that the average pairwise mutual information (MI) between hidden neurons rapidly peaks before declining, while the MI between each hidden neuron and the target prediction exhibits complex patterns: it increases and then decreases for some neurons, while for others, it consistently rises.

At the network level, the authors identified two distinct phases in the training process. The first phase is characterized by high redundant information, while the second shows an increase in unique information and a reduction in redundancy. This suggests that neurons in the first phase are all trying to predict the target by themselves, causing significant redundancy, while in the second phase, neurons seem to specialize and learn unique and increasingly synergistic representations about the target. Additionally, they noted that in larger networks, the MI between each hidden neuron and the target prediction decreases, while higher synergy terms appear, indicating more distributed and complex representations as networks grow larger.

Yu et al. [26] adopt a similar method by performing a PID analysis on the filters of Convolutional Neural Networks (CNNs), treating CNN filters as source variables. To understand the information dynamics within these networks and address the indeterminacy of PID atoms in

relation to classical information quantities discussed in Section 2.2.2, they employ summary statistics, including a redundancy-synergy metric (which assesses whether the system is predominantly redundant or synergistic) and a weighted non-redundant information measure (which quantifies the amount of non-redundant information present). Their findings reveal that while increasing the number of filters initially boosts the multivariate mutual information, this increase soon reaches a plateau. However, prediction accuracy continues to rise, which they attribute to the system becoming more synergistic and less redundant, in line with the RBM results in [25].

Refining the approach of Tax et al. [25] to understand how representations are distributed in neural networks, Elrich et al. [27] propose a metric called representational complexity. This metric measures the average number of neurons in a neural network required to access a specific piece of information. Using PID, they define the degree of synergy for a given antichain $\alpha \in \mathcal{A}$ as the minimal number of neurons needed to retrieve the associated information, expressed as $m(\alpha) = \min_{A \in \alpha} |A|$. For instance, the antichain $\{\{X_1\}, \{X_1, X_2\}\}$ exhibits a degree of synergy of 1 because the necessary information can be accessed solely by observing X_1. Conversely, the antichain $\{\{X_1, X_2\}, \{X_2, X_3\}\}$ displays a degree of synergy of 2, as at least two source variables are required to acquire the information within the associated PID atom. They further define the representational complexity of a neural network layer as the weighted average degree of synergy across all antichains, factored by their respective information contributions

$$C = \frac{1}{I(X;Y)} \sum_{\alpha \in \mathcal{A}} I_\partial(Y;\alpha) m(\alpha).$$

A representational complexity of 1 is indicative of one-hot encoding scenarios, where all information about any possible outcome is accessible by examining just the relevant neuron. Conversely, denser coding schemes, where 2^n possible labels are represented by only n neurons, necessitate accessing all these neurons to decode any specific information, resulting in maximal representational complexity. Through empirical studies utilizing the CIFAR and MNIST datasets, the authors observe a reduction in representational complexity during training, with deeper layers exhibiting lower complexities. These findings support the notion that lower representational complexity facilitates a more efficient readout of classification outputs in the deep layers of neural networks.

Clauw et al. [28] adopt an approach similar to that of Elrich et al. [27], positing that the generalization abilities of neural networks derive from high-order synergistic interactions among neurons. Utilizing the concept of 0-Information [29]—a metric for assessing whether a set of random variables is predominantly redundant or synergistic—they analyze the MNIST dataset to explore information dynamics across different neural network layers. Their findings indicate that the initial layers predominantly exhibit redundant information, likely linked to the recognition of local features such as edges. In contrast, the deeper layers are characterized by synergistic interactions that are crucial for making high-level classification decisions.

In a related study, Proca et al. [18] explore the role of information dynamics in neural networks. They find that redundant information increases the robustness of the network, with high levels of dropout encouraging the network to store information redundantly. Conversely, they observe that synergy enables networks to encode information more efficiently, facilitating the integration of diverse information sources and supporting the simultaneous learning of multiple tasks, provided the tasks are interleaved during training to prevent catastrophic forgetting. However, they note that networks characterized by high synergy are less robust and more susceptible to various types of perturbations. These conclusions are

drawn from a series of experiments across supervised learning, reinforcement learning, and recurrent neural network models, demonstrating their broad applicability in understanding ML systems.

Kong et al. recently introduced a novel mathematical framework for diffusion models that allows for the computation of data probability densities by finding the global optimum of a mean squared error denoising objective [30]. Building upon this foundation, subsequent work [31] has developed a method for performing information decomposition in diffusion models. This advancement enables the calculation of pixel-wise MI and CMI between image pixels and specific words from the image prompt. This allows precise localization and identification of objects within images, and even targeted modifications of these objects through prompt-based interventions.

4.3 Leveraging Partial Information Decomposition to Enhance ML Systems

The information-theoretic approach outlined by Tishby and Zaslavsky [19], which defines the goal of supervised learning as deriving a sufficient statistic $\tilde{X}$ from input data X to capture all relevant information $I(X;Y)$ that X provides about the output labels Y, can be enhanced using the PID framework. When X is multidimensional, the PID framework allows for a more detailed breakdown into redundant, synergistic, and unique information components. This enables the development of more fine-grained training objectives aimed at optimizing algorithm performance.

This refined approach was adopted by Graetz et al. [6]. In their biologically inspired model, each neuron receives a two-dimensional input $X = (X_R, X_C)$ and outputs Y. They propose a versatile goal function for defining learning objectives as follows:

$$\begin{aligned} G(Y; X_R, X_C) =& \Gamma_0 I_{unq}(Y; X_R) + \Gamma_1 I_{unq}(Y; X_C) + \Gamma_2 I_{red}(Y; X_R, X_C) \\ &+ \Gamma_3 I_{syn}(Y; X_R, X_C) + \Gamma_4 H(Y \mid X_R, X_C), \end{aligned}$$

where $\{\Gamma_i\}_{i=0}^4$ are parameters that tailor the local learning goals of the neurons. By varying these parameters, Graetz et al. [6] demonstrated that their *informorphic* neurons could be configured to perform distinct tasks. For instance, setting the local goal to maximize redundancy between inputs and outputs enabled the solution of a simple MNIST classification task, while focusing solely on the unique information term $I_{unq}(Y; X_R)$ allowed them to solve a compression task through self-organization.

Barrientos and Sootla [32] argue that intelligent systems must possess the capability to handle primitive concepts and compose them hierarchically. In pursuit of this, they propose a modified version of the variational autoencoder (VAE), explicitly designed to minimize synergy in the loss function. Their model achieves disentanglement results comparable to those of the FactorVAE [33], a variant of VAE that targets disentangled representations. They demonstrate that the FactorVAE also indirectly reduces synergy through its original loss function. These findings underscore the profound connection between synergy and entanglement, and could lead to advancements in ML systems capable of composing concepts and exploring the vast landscape of hierarchical relationships.

In the field of Self-Supervised Learning, Mohamadi et al. [34] employ the PID framework to enhance data augmentation. They investigate how decomposing the mutual information between different representations or augmentations of data into redundant, unique, and synergistic components can be beneficial. There is an ongoing debate about the efficacy of using mutually informative augmentations of the same data. The authors advocate for augmentations that maximize synergistic information and minimize redundancy relative to the

original data, thus clearly distinguishing the confounding contributions of synergy and redundancy in the mutual information between the original data and its augmentations. They observe that while current data augmentation methods effectively reduce redundancy, they often do so at the cost of synergistic interactions. To address this, they develop a procedure that preserves synergy while reducing redundancy and demonstrate its effectiveness on the CIFAR and ImageNet datasets.

The works discussed in this section illustrate the potential of leveraging PID to enhance various aspects of ML systems. By adopting a more fine-grained decomposition of multivariate information, researchers have been able to design novel approaches to address key challenges in several subfields of ML.

4.4 ML Techniques for Estimating PID Information Atoms

While previous sections have highlighted how PID can enhance various aspects of ML systems, the methods for computing these information-theoretic quantities have not been extensively discussed. Indeed, the absence of universally agreed-upon methods for computing PID quantities, combined with the superexponential increase in the number of PID information atoms as the number of sources grows, has limited its broader adoption despite its significant conceptual promise. In this section, we will explore several studies that address these challenges by leveraging ML techniques to estimate these complex information-theoretic and PID quantities.

One of the pioneering efforts in this direction was conducted by Belghazi et al. [35]. Their primary objective was to estimate MI, which they approached by leveraging its characterization as a Kullback-Leibler (KL) divergence. Utilizing the Donsker-Varadhan representation [36] of the KL divergence, they formulated the estimation of MI as an optimization problem. This problem could then be efficiently solved using neural networks. This method represents a general approach to estimate information-theoretic quantities: by expressing them as KL divergence and framing their estimation as an optimization problem addressable with neural networks.

A similar methodology was employed by Kleinman et al. [37], who developed the Redundant Information Neural Estimator (RINE). This tool computes the amount of information that is redundant across a set of sources by optimizing over a family of functions using a neural network. The output of RINE is a coarse-grained summary statistic that quantifies the overall redundancy within the system, although it does not specify which aspects of the sources contribute to this redundancy.

Another PID-inspired measure of interest is the O-information [38], which assesses whether a system's information character is predominantly synergistic or redundant. This measure scales effectively with the number of components and does not require dividing the system into sources and targets. Bounoua et al. [39] approach the estimation of O-information by expressing it in terms of KL divergences. They utilize a time-varying score function that facilitates the use of parametric approximations, thereby enabling the estimation of O-information via neural networks.

Kaplanis et al. [40] aimed to identify causally emergent representations, which constitute a form of high-order synergy. In a time-evolving system represented by $X_t^1, \ldots, X_t^n$, they define a causally emergent variable as a variable $V_t = f(X_t)$ that holds unique information about the future state of the system $X_{t'}$ with respect to the individual components; $I_{unq}(X_{t'}; V_t \setminus (X_t^1, \ldots, X_t^n)) > 0$. A criterion for causal emergence was introduced by Rosas et al. [41]: $\Psi := I(V_t; V_{t+1}) - \sum_i I(X_t^i; V_{t+1}) > 0$. Kaplanis and colleagues employed a neural

network to maximize Ψ, while utilizing two additional networks to estimate $I(V_t; V_{t+1})$ and $\sum_i I(X_t^i; V_{t+1})$ respectively, following approaches similar to the ones previously described. This approach enabled them to detect emergent variables in synthetic data.

5 Experiments

In this chapter, we further illustrate the significant potential of PID to enhance our understanding of current ML systems. Specifically, we investigate the information processing dynamics within a transformer-based Large Language Model.

5.1 Motivation: *A synergistic core for human brain evolution and cognition*

The core inspiration for this section is derived from the work of Luppi, Mediano, Rosas et al. in *A synergistic core for human brain evolution and cognition* [4]. Although briefly discussed in Section 3.1, this subsection provides a deeper overview of their work, crucial for understanding the subsequent experiments.

The primary question addressed by [4] concerns the informational architecture of the brain and how it supports high-level cognition. They investigate the information processing dynamics between different brain regions using a recently developed extension of PID that incorporates a temporal dimension, enabling the study of dynamical systems such as the brain. This perspective allows them to explore whether and how the current state of the brain influences future states and the transmission of information from the present to the future.

This extension is termed Integrated Information Decomposition, or ΦID [42]. Using temporal data from two brain regions, X_t^1 and X_t^2, ΦID decomposes the mutual information between the past of both regions and the present, $I(X_{t-\tau}^1, X_{t-\tau}^2; X_t^1, X_t^2)$. This decomposition allows examination of how current brain states are shaped by individual components of their past (temporally persistent redundancy) and how they are influenced by information arising from the integration of multiple brain regions' past data (temporally persistent synergy).

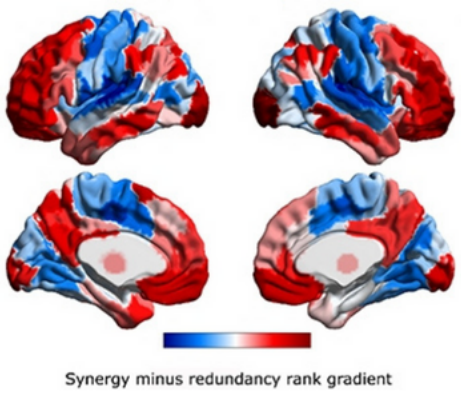


Figure 2: Distribution of Synergy and Redundancy in the Brain. Regions exhibiting synergistic interactions are marked in red, while those displaying redundant interactions are highlighted in blue. This image, adapted from [4], illustrates the synergy minus redundancy rank across different brain regions.

Using resting-state data from 100 participants from the Human Connectome Project, they calculated pairwise redundant and synergistic information terms between brain regions. By averaging the redundancy and synergy for each region, they established a ranking of brain regions based on their synergistic and redundant interactions. This ranking enabled a parcelation of the brain into regions characterized by either synergistic or redundant interactions, as shown in Figure 2.

The distinct division of the brain into synergistic and redundant regions suggests these regions may be associated with different cognitive domains. To test this hypothesis, a meta-analysis was conducted using NeuroSynth, which provided data on brain activity during various cognitive tasks. They discovered that regions rich in synergistic interactions correlate with higher cognitive tasks (such as cognitive control, social cognition, emotions), while regions with predominantly redundant interactions are linked with lower-level cognitive tasks (such as auditory processing, pain, motor functions). Thus, these findings suggest that the redundancy-to-synergy ranking corresponds to a progression from lower- to higher-level cognitive tasks.

Lower-level, sensory-related cognitive tasks typically benefit from segregated and modular information processing, whereas higher cognitive tasks require significant integration. Treating brain regions as nodes and the strength of their pairwise synergistic or redundant interactions as edges results in two distinct graphs: a synergy graph and a redundancy graph. Further analysis of these graphs' theoretical properties revealed that the synergy graph is more globally efficient and interconnected than the redundancy graph, which exhibits a more modular structure. This supports the initial hypothesis regarding the functional differentiation between the two types of information processing in the brain.

5.2 *A Synergistic Core for Large Language Models*

The core idea of this section is to investigate whether Large Language Models (LLMs) possess a similar informational architecture to the brain, specifically a synergistic core. Recent advances in LLMs have demonstrated their capability to perform a wide variety of tasks, suggesting these systems may exhibit intelligent behavior and the ability to undertake high-order cognitive tasks.

LLMs are particularly suited for this type of information theoretical analysis due to the relative ease of recording and analyzing how information is processed within these models—a task that is much more challenging in the human brain. Inspired by the significant findings in [4], we conducted a similar analysis on a recently developed open-source LLM, the 2 Billion parameter version of the Gemma family [1]. This model features a decoder-only transformer architecture with 18 layers, each containing 8 attention heads, totaling 144 attention heads.

However, several decisions had to be made regarding the correspondence between brain structures and the LLM architecture. In particular, we needed to determine the equivalent of brain regions in a decoder-only transformer. We considered several options, including using complete layers, individual neurons, or groups of neurons. Ultimately, we opted for using attention heads, as they bear a high-level resemblance to brain regions, with each head focusing on different features of the input sequence to generate the next token prediction.

Additionally, the exact correspondence for time in these types of transformers was not immediately clear. Given their autoregressive architecture, we decided to treat each autoregressive token generation step as a timestep. If we liken the brain's process of formulating thoughts during a cognitive task to next-token prediction, the analogy becomes quite clear: at each timestep, the brain (or model) predicts the next word to continue its reasoning. Indeed, at each autoregressive timestep, each attention head in the model formulates a new query or question to answer in order to predict the next token. This is the notion of time we adopted for our LLM analysis.

Finally, considering that the original analysis by [4] used resting-state data to construct the synergy-redundancy gradient shown in Figure 2, we needed to find an equivalent method to obtain "resting-state" data from the LLM. We opted to input random strings of text and

increase the generation temperature setting of the model to a value of 3. This adjustment helps avoid repetition loops, a common issue where the generation process collapses into producing the same token repeatedly.

5.3 Synergy in Deep Layers and Redundancy in Early Layers

The primary goal of our initial experiments was to analyze synergy and redundancy within the LLM. Inspired by [4], our focus was on the temporally persistent redundancy and synergy outlined in the ΦID framework [42], which we will refer to as redundancy and synergy for simplicity.

To compute these measures between attention heads using ΦID, we required a time series reflecting the state of each attention head throughout the autoregressive generation process. Each self-attention head applies the well-known attention mechanism [3], simplified here by omitting linear projections:

$$Attention(Q, K, V) = AV, \text{ where } A := softmax\left(\frac{QK^T}{\sqrt{d_k}}\right).$$

During each timestep of the autoregressive token generation, a new query row, q_t, is added to the query matrix Q, with its impact reflected in the last row of A, denoted a_t. We use the vector a_t to represent the state of a specific attention head at timestep t, and refer to a_t as the attention weights of the head at a given time.

However, to compute synergy and redundancy between different heads, a single scalar value per timestep is required rather than a vector. We explored various aggregation methods, including the L2 norm $||a_t||_2$, the mean, and the maximum of the attention weights vector. The results were similar across these methods, so we chose to use the norm for subsequent experiments.

With these implementation decisions finalized, we were able to compute time series analogous to the resting-state fMRI data of different brain regions used in [4]. We fed the LLM a random sequence of tokens with a length of 24 and generated 1000 tokens autoregressively, setting the temperature parameter to 3 to prevent repetition loops. For each attention head and at each of the 1000 autoregressive token generation timesteps, we recorded the L2 norm of the associated attention weights. A visualization of the results from this procedure is provided in Appendix A.

Figure 3 displays two symmetric matrix heatmaps showing the pairwise synergy and redundancy between each attention head. The attention heads are labeled sequentially from 1 to 144, starting with the first layer. In these matrices, the value at position (i, j) indicates the synergy (or redundancy, respectively) between heads number i and j. These matrices were created using the time series data previously generated, and we employed the code associated with the ΦID framework as detailed in the related works [43, 42, 4].

Figure 4 displays the average synergy and redundancy for each attention head, calculated by averaging the values across each row of the respective synergy and redundancy matrices. Figure 5 presents the average synergy minus redundancy rank for each layer, where higher values indicate layers with predominant synergistic interactions, and lower values signify layers that are mostly redundant. To compute this rank, each head was first ranked from least to most synergistic and from least to most redundant. These ranks were then subtracted, and the heads were re-ranked accordingly. For clarity in visualization, we plotted the average of these final ranks for each layer.

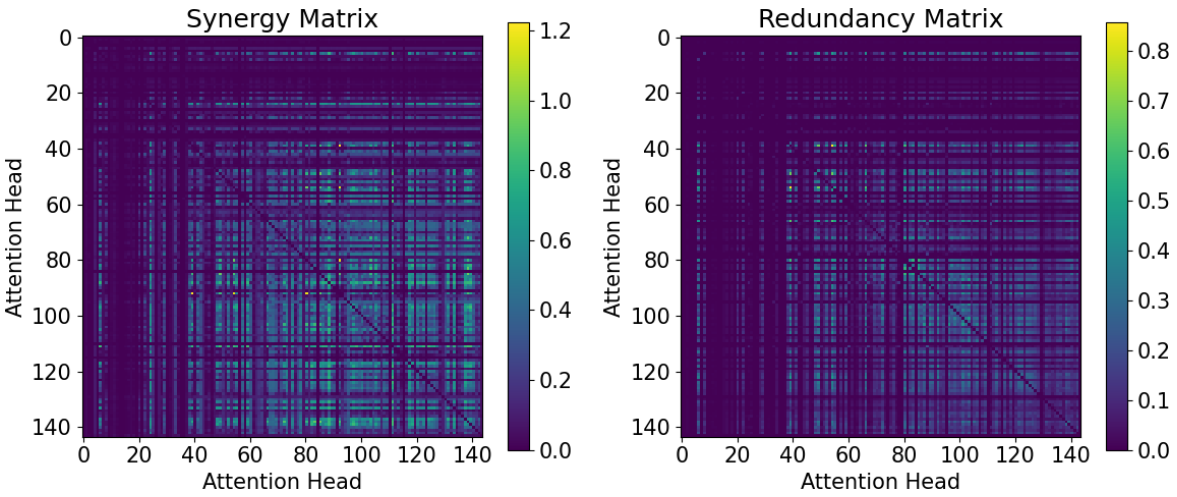
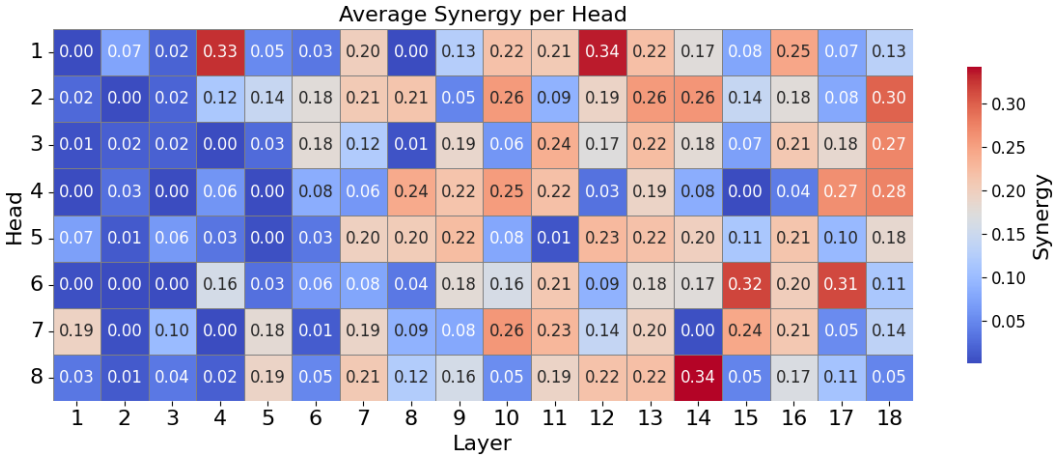
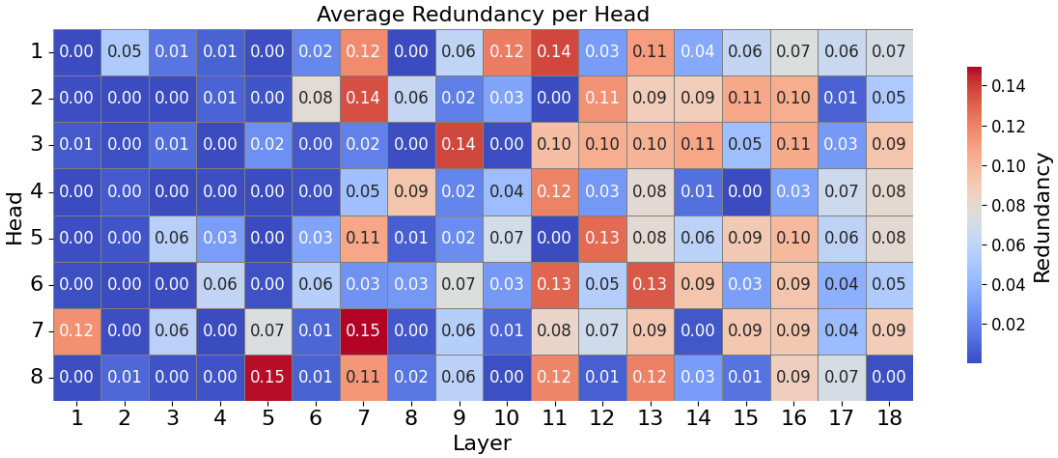


Figure 3: Synergy and redundancy matrix heatmaps of the pairwise attention weight norms between each attention head.



(a) Average pairwise synergy of each attention head.



(b) Average pairwise redundancy of each attention head.

Figure 4: Average pairwise synergy and redundancy of each attention head.

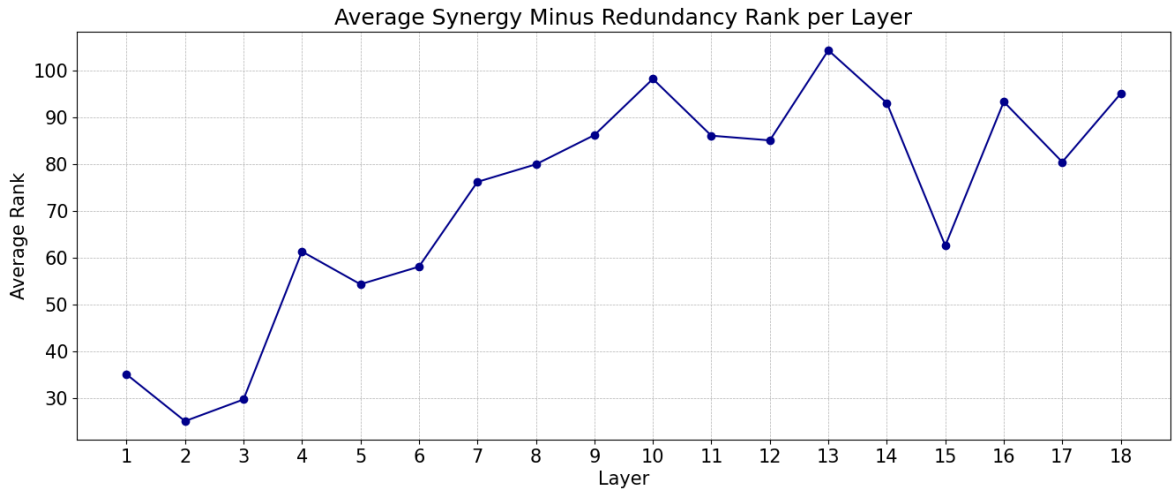


Figure 5: Average attention head synergy minus redundancy rank per layer.

The results indicate that synergy predominates in the deeper layers of the transformer, while it is nearly absent in the initial layers. There appears to be a strong correlation between synergy and redundancy, although redundancy is generally less pronounced. As depicted in Figure 5, there is a gradation from layers that are mostly redundant to those that are highly synergistic as we progress deeper into the transformer. This gradation shows a clear increasing trend from layers 1 to 10, with the last layers displaying similar levels of synergy minus redundancy.

These findings align with the understanding that deeper layers in neural networks are responsible for more abstract and higher-level computations, as synergy is associated with the integration of information and high-level cognition, consistent with insights from [4]. Additionally, these results are in line with those of Clauw et al. [28], who observed a similar gradient from initially redundant layers to synergistic final layers in simple neural networks. In summary, these results suggest that Large Language Models (LLMs) possess a synergistic core located in the deepest layers.

In accordance with the findings reported in [4], we discovered that the synergy matrix is more globally efficient than the redundancy matrix. However, contrary to the observations in the human brain detailed in [4], the synergy matrix in our study also exhibited greater modularity than the redundancy matrix. Numerical results are presented in Table 1 and were computed using the `bctpy` brain connectivity toolbox.

Metric	Synergy Matrix	Redundancy Matrix
Global Efficiency	0.12789 ± 0.00456	0.07666 ± 0.01726
Modularity	0.09760 ± 0.00617	0.05697 ± 0.00822

Table 1: Global efficiency and modularity of the obtained redundancy and synergy matrices shown in Figure 3, across four different runs.

5.4 Different Attention Heads are used for Different Cognitive Tasks

We next sought to determine whether LLMs utilize different attention heads for different cognitive tasks, similar to how distinct brain regions are employed for diverse cognitive functions in humans. To explore this hypothesis, we categorized six types of cognitive tasks, arranged a priori on a continuum from low- to high-level cognitive tasks. For each category,

Cognitive Task Category	Example Prompt
Syntax and Grammar Correction	Correct the error: He go to school every day.
Part of Speech Tagging	Identify the parts of speech in the sentence: Quickly, the agile cat climbed the tall tree.
Basic Numerical Reasoning	If you have 15 apples and you give away 5, how many do you have left?
Basic Common Sense Reasoning	If it starts raining while the sun is shining, what weather phenomenon might you expect to see?
Abstract Reasoning and Creative Thinking	Imagine a future where humans have evolved to live underwater. Describe the adaptations they might develop.
Emotional Intelligence and Social Cognition	Write a dialogue between two characters where one comforts the other after a loss, demonstrating empathy.

Table 2: Cognitive task categories and example prompts, sorted by increasing cognitive complexity. The sequence ranges from basic tasks, such as syntax and grammar correction, to advanced tasks, such as emotional intelligence and social cognition, based on our prior intuitive ranking of cognitive levels.

we designed ten unique prompts for the model. The six cognitive task categories and an example prompt are detailed in Table 2. The complete list of prompts utilized is shown in Appendix C.

We discovered that LLMs indeed utilize different attention heads for different cognitive tasks. We measured the norm of the attention weights vector—referred to hereafter as 'activation'—for each head while the LLM processed each prompt. Significant variations in activation were observed among the deeper layers of the LLM when analyzing the average activation during responses to prompts, grouped by cognitive task category. A detailed visualization of these differences is provided in Appendix B.

To further investigate this phenomenon, we applied Linear Discriminant Analysis (LDA) to the average activations of each attention head across all 60 prompts. Our aim was to cluster these 60 prompts into their respective cognitive task categories, using only the average head activations as features for the clustering. The results, illustrated in Figure 6, demonstrate clear clustering of prompts within their respective cognitive task categories. Notably, the first LDA component, a linear combination of average head attentions, effectively differentiates between low-level and high-level tasks, with lower values of this first component corresponding to low-level cognitive tasks and higher values indicating higher-level cognitive tasks.

However, our initial classification of 'basic numerical reasoning' as an intermediate-level cognitive task seems to be incorrect. Observations from Figure 6 and Appendix B suggest it should be categorized at the lowest level. We hypothesize that this may be due to the models processing numerical information via simple, rule-based methods or memorization, which require minimal reasoning and computation.

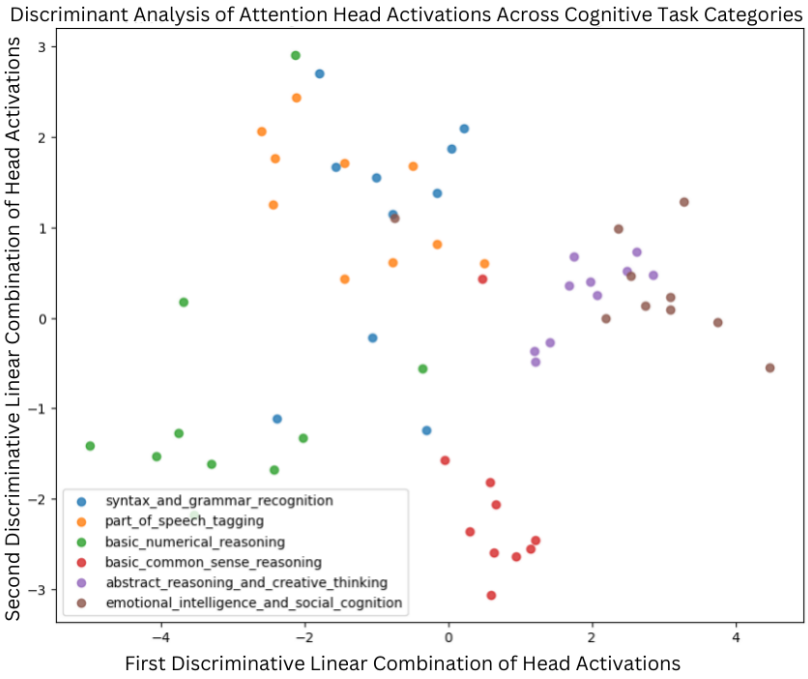


Figure 6: Linear Discriminant Analysis (LDA) applied to average attention head activations across 60 prompts, organized to assess differentiation among six cognitive task categories. This visualization clearly delineates clustering by task, with the first discriminant component providing a gradient from lower to higher cognitive task complexities.

5.5 High Level Cognitive Tasks use Synergistic Layers while Low Level Cognitive Tasks use Redundant Layers

We sought to identify a correspondence between the progression from lower- to higher-level cognitive tasks and a redundancy-to-synergy ranking, inspired by findings in [4] that suggest a close correlation in the human brain.

Figure 5 suggests a correspondence between the deepest layers, which exhibit the most synergistic interactions, and the initial layers, which are the most redundant. This prompted us to investigate the relative activation of each layer for the different cognitive task categories.

We discovered that the lowest-level cognitive tasks, including 'basic numerical reasoning' and 'syntax and grammar correction', were most actively processed (in relative terms) in the initial layers, which are the most redundant, and were less active in the deepest layers, associated with the most synergistic interactions. Conversely, the highest-level cognitive tasks, such as 'abstract reasoning and creative thinking' and 'emotional intelligence and social cognition', were predominantly active in the deepest, highly synergistic layers, and least active in the initial, most redundant layers. Figures 7 and 8 illustrate these findings.

These preliminary findings suggest a correlation between the redundancy-to-synergy ranking and the progression from lower- to higher-level cognitive tasks in LLMs, analogous to observations in the human brain [4]. However, a more in-depth analysis is required to substantiate this claim further.

5.6 Limitations and Future Work

Despite the promising results demonstrating a close resemblance between the information processing dynamics of the human brain and large language models, this work faced several

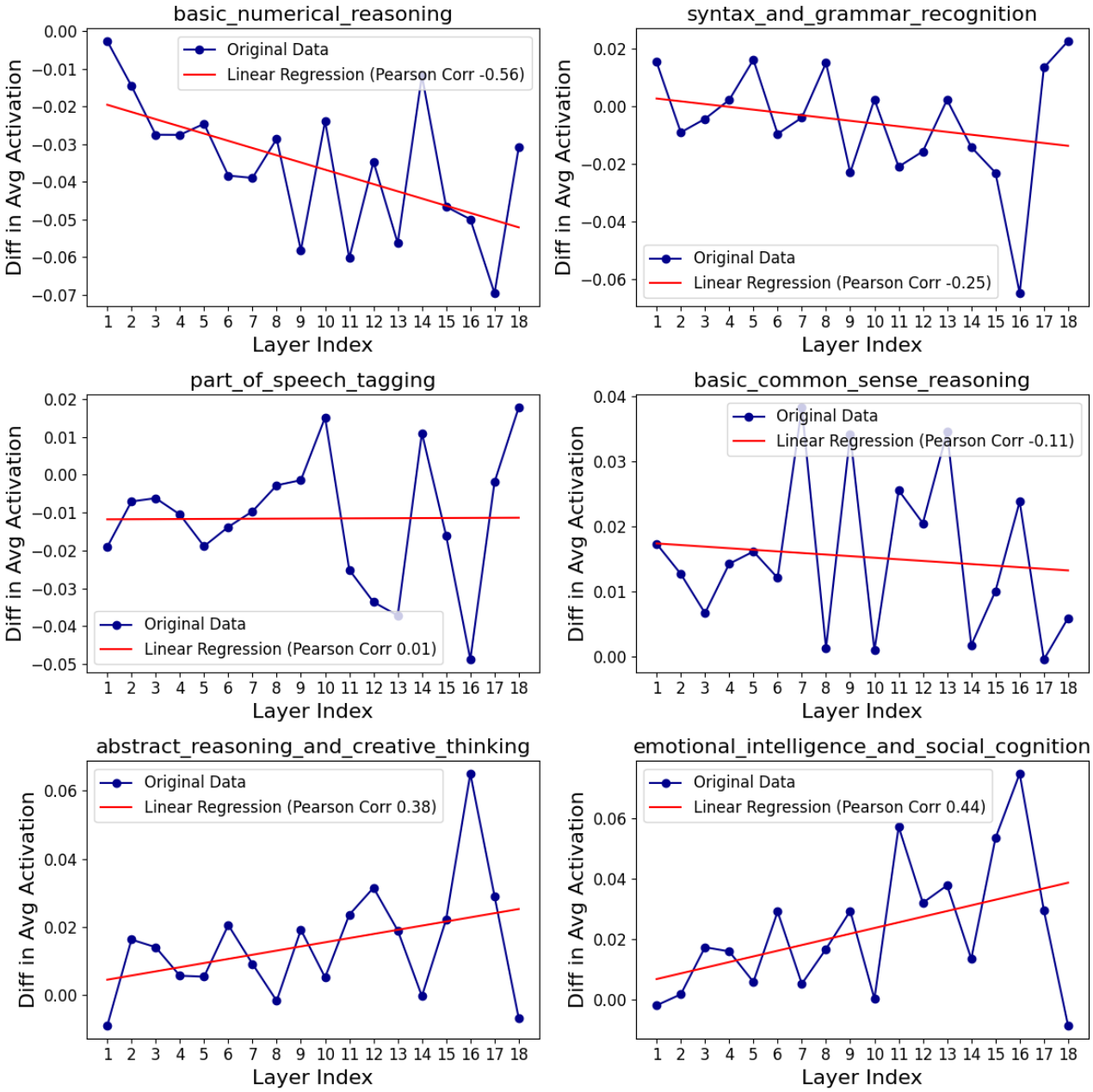


Figure 7: Differential activations per layer for each cognitive task category, calculated relative to the average activation of each head across all categories and averaged per layer. Linear regression analysis was performed for each subplot, with the resulting regression lines displayed in red.

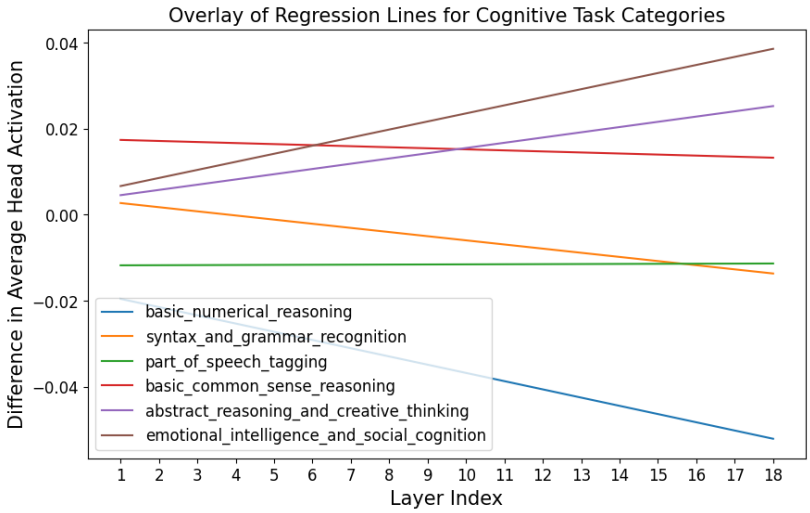


Figure 8: Overlay of regression slopes for each cognitive task category, illustrating differences in relative activations across layers. This plot displays six lines, each corresponding to the regression analyses from the six subplots in Figure 7, highlighting trends in how each cognitive task category engages different LLM's layers.

limitations due to time and computational constraints, which pave the way for potential future work:

- **Pairwise Computation of Synergy and Redundancy**: Current practical limitations of PID and ΦID restricted our analysis to low-order synergies and redundancies between attention heads. This represents a significant limitation as our results do not account for potential high-order synergies and redundancies, which could reveal deeper insights. Future studies could employ coarse-grained PID metrics such as the O-Information [38] to address this issue.
- **Lack of Comparison Among Different LLMs**: Our study was limited to the 2 billion parameter version of the Gemma model [1]. Comparing findings across various LLMs and model sizes could enhance the robustness and generalizability of our results.
- **Limited Number of Cognitive Task Categories**: We investigated only six cognitive task categories, with ten prompts per category. Expanding the range of cognitive tasks and increasing the number of prompts could provide a more comprehensive understanding of how LLMs utilize different attention heads for various different cognitive task categories.
- **Correlation Between Low to High-Order Cognitive Tasks and Redundancy to Synergy Gradient**: The observed relationship between the gradient from low to high-order cognitive tasks and the redundancy to synergy gradient warrants further investigation. More detailed experiments are needed to confirm and elaborate on this apparent correlation.

6 Conclusion

In this study, we have examined the significant contributions of Information Theory (IT) and Partial Information Decomposition (PID) across various fields, including Neuroscience, Complex Systems, Cognitive Science, and Machine Learning. We particularly focused on the intersection of PID and Machine Learning, reviewing key studies in the area. Our empirical research revealed a synergistic core within the Gemma Large Language Model (LLM), reminiscent of similar structures identified in the human brain [4]. This finding underscores the broader applicability of PID principles in deciphering and understanding complex information processing systems.

We started by delving into the theoretical frameworks and mathematical underpinnings of Information Theory (IT) and Partial Information Decomposition (PID) that underlie this work. Motivating the subsequent exploration of PID in Machine Learning, we examined its successful applications in Neuroscience, Cognitive Science, and Complex Systems. These fields have adopted PID as a key conceptual framework, enabling a more nuanced understanding of brain dynamics and other information processing systems.

A key component of this work is our review of the integration of PID with Machine Learning. Since its inception, Information Theory has aimed to provide a theoretical foundation for Deep Learning. We have argued that the recent extensions to PID represent the path towards achieving this theoretical foundation. Furthermore, PID has proven extremely useful in the pursuit of interpretability within Machine Learning systems, even contributing to their improvement. However, one of the main challenges with the PID framework is the difficulty of computing the PID information atoms. We have also demonstrated how Machine Learning techniques can be leveraged to address this issue.

In a further attempt to demonstrate the potential of PID to interpret Machine Learning systems, we performed a PID analysis on the Gemma LLM [1]. Our main result indicates that Gemma possesses a synergistic core of attention heads, similar to the synergistic core of brain regions identified in the human brain [4]. Additionally, we found that high-level cognitive tasks significantly engage the synergistic layers more than the redundant layers in Gemma, whereas low-level cognitive tasks predominantly utilize the redundant layers.

These results suggest a convergence in the informational architecture of various 'intelligent' information processing systems, whereby a synergistic core is formed that facilitates information integration, which in turn enables higher-order cognition. Future work could extend this line of research by performing similar analyses on larger LLMs to explore how this synergistic core, which appears to facilitate higher-order cognition, evolves as LLMs increase in size, complexity and intelligence. Beyond enhancing the interpretability of LLMs, these studies could contribute to a *deep theory of cognition* that explains not only how cognition operates in the human brain, but potentially in any kind of intelligent complex system.

References

[1] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love *et al.*, "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024.

[2] P. L. Williams and R. D. Beer, "Nonnegative decomposition of multivariate information," *arXiv preprint arXiv:1004.2515*, 2010.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[4] A. I. Luppi, P. A. Mediano, F. E. Rosas, N. Holland, T. D. Fryer, J. T. O'Brien, J. B. Rowe, D. K. Menon, D. Bor, and E. A. Stamatakis, "A synergistic core for human brain evolution and cognition," *Nature Neuroscience*, vol. 25, no. 6, pp. 771–782, 2022.

[5] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[6] M. Graetz, A. Makkeh, A. C. Schneider, D. A. Ehrlich, V. Priesemann, and M. Wibral, "Infomorphic networks: Locally learning neural networks derived from partial information decomposition," *arXiv preprint arXiv:2306.02149*, 2023.

[7] R. Quian Quiroga and S. Panzeri, "Extracting information from neuronal populations: information theory and decoding approaches," *Nature Reviews Neuroscience*, vol. 10, no. 3, pp. 173–185, 2009.

[8] E. Amico, K. Abbas, D. A. Duong-Tran, U. Tipnis, M. Rajapandian, E. Chumin, M. Ventresca, J. Harezlak, and J. Goñi, "Toward an information theoretical description of communication in brain networks," *Network Neuroscience*, vol. 5, no. 3, pp. 646–665, 2021.

[9] A. I. Luppi, F. E. Rosas, P. A. Mediano, D. K. Menon, and E. A. Stamatakis, "Information decomposition and the informational architecture of the brain," *Trends in Cognitive Sciences*, 2024.

[10] G. Tononi, "Consciousness as integrated information: a provisional manifesto," *The Biological Bulletin*, vol. 215, no. 3, pp. 216–242, 2008.

[11] M. Celotto, J. Bím, A. Tlaie, V. De Feo, A. Toso, S. Lemke, D. Chicharro, H. Nili, M. Bieler, I. Hanganu-Opatz *et al.*, "An information-theoretic quantification of the content of communication between brain regions," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[12] M. Wibral, V. Priesemann, J. W. Kay, J. T. Lizier, and W. A. Phillips, "Partial information decomposition as a unified approach to the specification of neural goal functions," *Brain and cognition*, vol. 112, pp. 25–38, 2017.

[13] T. F. Varley, "Information theory for complex systems scientists," *arXiv preprint arXiv:2304.12482*, 2023.

[14] T. F. Varley and J. Bongard, "Evolving higher-order synergies reveals a trade-off between stability and information integration capacity in complex systems," *arXiv preprint arXiv:2401.14347*, 2024.

[15] R. D. Beer and P. L. Williams, "Information processing and dynamics in minimally cognitive agents," *Cognitive science*, vol. 39, no. 1, pp. 1–38, 2015.

[16] F. E. Rosas, P. A. Mediano, H. J. Jensen, A. K. Seth, A. B. Barrett, R. L. Carhart-Harris, and D. Bor, "Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data," *PLoS computational biology*, vol. 16, no. 12, p. e1008289, 2020.

[17] T. F. Varley and E. Hoel, "Emergence as the conversion of information: a unifying theory," *Philosophical Transactions of the Royal Society A*, vol. 380, no. 2227, p. 20210150, 2022.

[18] A. M. Proca, F. E. Rosas, A. I. Luppi, D. Bor, M. Crosby, and P. A. M. Mediano, "Synergistic information supports modality integration and flexible learning in neural networks solving multiple tasks," *ArXiv*, 2022.

[19] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *2015 IEEE Information Theory Workshop (ITW)*, 2015, pp. 1–5.

[20] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *ArXiv*, vol. abs/1703.00810, 2017.

[21] Z. Goldfeld, E. Van Den Berg, K. Greenewald, I. Melnyk, N. Nguyen, B. Kingsbury, and Y. Polyanskiy, "Estimating information flow in deep neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 97. PMLR, 2019, pp. 2299–2308.

[22] A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox, "On the information bottleneck theory of deep learning*," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2019, no. 12, 2019.

[23] P. Wollstadt, S. Schmitt, and M. Wibral, "A rigorous information-theoretic definition of redundancy and relevancy in feature selection based on (partial) information decomposition," *J. Mach. Learn. Res.*, vol. 24, pp. 131:1–131:44, 2023.

[24] J. Milzman, A. Harrison, C. Nieto-Granda, and J. Rogers, "Measuring multi-source redundancy in factor graphs," in *2023 26th International Conference on Information Fusion (FUSION)*. IEEE, 2023, pp. 1–8.

[25] T. M. Tax, P. A. Mediano, and M. Shanahan, "The partial information decomposition of generative neural network models," *Entropy*, vol. 19, no. 9, 2017.

[26] S. Yu, K. Wickstrom, R. Jenssen, and J. Principe, "Understanding convolutional neural networks with information theory: An initial exploration," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 1, p. 435—442, 2021.

[27] D. A. Ehrlich, A. C. Schneider, V. Priesemann, M. Wibral, and A. Makkeh, "A measure of the complexity of neural representations based on partial information decomposition," *Transactions on Machine Learning Research*, 2023.

[28] K. Clauw, S. Stramaglia, and D. Marinazzo, "Higher-order mutual information reveals synergistic sub-networks for multi-neuron importance," *arXiv preprint arXiv:2211.00416*, 2022.

[29] F. E. Rosas, P. A. M. Mediano, M. Gastpar, and H. J. Jensen, "Quantifying high-order interdependencies via multivariate extensions of the mutual information," *Phys. Rev. E*, vol. 100, p. 032305, 2019.

[30] X. Kong, R. Brekelmans, and G. Ver Steeg, "Information-theoretic diffusion," in *International Conference on Learning Representations*, 2023.

[31] X. Kong, O. Liu, H. Li, D. Yogatama, and G. V. Steeg, "Interpretable diffusion via information decomposition," in *The Twelfth International Conference on Learning Representations*, 2024.

[32] G. Barrientos and S. Sootla, "Non-synergistic variational autoencoders," 2019.

[33] H. Kim and A. Mnih, "Disentangling by factorising," in *International conference on machine learning*. PMLR, 2018, pp. 2649–2658.

[34] S. Mohamadi, G. Doretto, and D. A. Adjeroh, "More synergy, less redundancy: Exploiting joint mutual information for self-supervised learning," *ArXiv*, 2023.

[35] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, 2018, pp. 531–540.

[36] M. Donsker and S. Varadhan, "Asymptotic evaluation of certain markov process expectations for large time. iv," *Communications on Pure and Applied Mathematics*, vol. 36, no. 2, pp. 183–212, 1983.

[37] M. Kleinman, A. Achille, S. Soatto, and J. C. Kao, "Redundant information neural estimation," *Entropy*, vol. 23, no. 7, 2021.

[38] F. E. Rosas, P. A. M. Mediano, M. Gastpar, and H. J. Jensen, "Quantifying high-order interdependencies via multivariate extensions of the mutual information," *Phys. Rev. E*, vol. 100, p. 032305, 2019.

[39] M. Bounoua, G. Franzese, and P. Michiardi, "Soi: Score-based o-information estimation," in *Submitted to ArXiV, 8 February 2024*, EURECOM, Ed., 2024.

[40] C. Kaplanis, P. Mediano, and F. Rosas, "Learning causally emergent representations," in *NeurIPS 2023 workshop: Information-Theoretic Principles in Cognitive Systems*, 2023.

[41] F. E. Rosas, P. A. M. Mediano, H. J. Jensen, A. K. Seth, A. B. Barrett, R. L. Carhart-Harris, and D. Bor, "Reconciling emergences: An information-theoretic approach to identify causal emergence in multivariate data," *PLOS Computational Biology*, vol. 16, no. 12, p. e1008289, 2020.

[42] P. A. Mediano, F. E. Rosas, A. I. Luppi, R. L. Carhart-Harris, D. Bor, A. K. Seth, and A. B. Barrett, "Towards an extended taxonomy of information dynamics via integrated information decomposition," *arXiv preprint arXiv:2109.13186*, 2021.

[43] Imperial MIND Lab, "Phi-id – integrated information decomposition," https://github.com/Imperial-MIND-lab/integrated-info-decomp/tree/main, 2023.

A Time Series of Attention Weight Norms

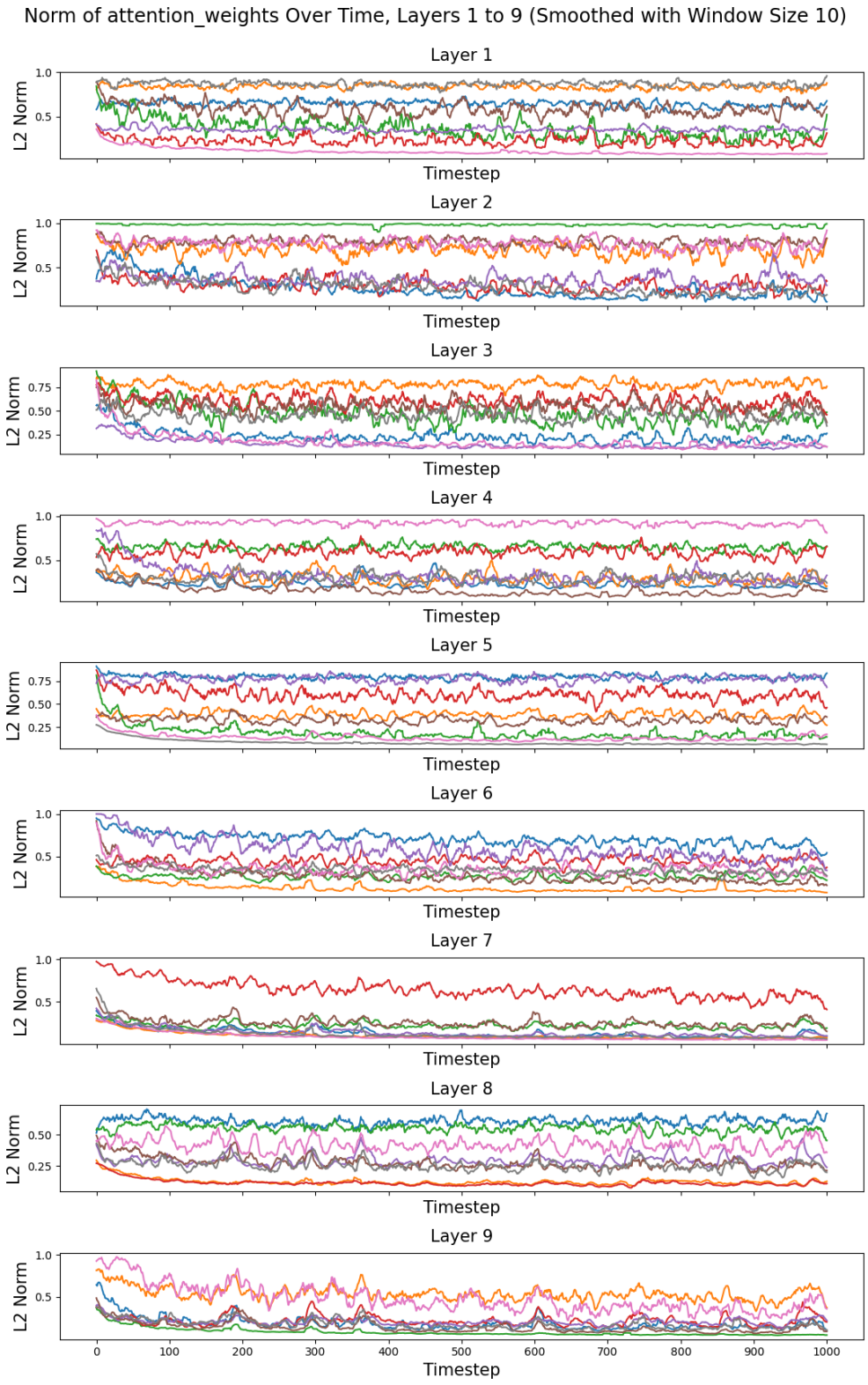


Figure 9: Time series of L2 norms for attention weights across attention heads in layers 1 to 9, with each head represented by a distinct color line. For enhanced visualization, the series has been smoothed using a moving average with a window size of 10.

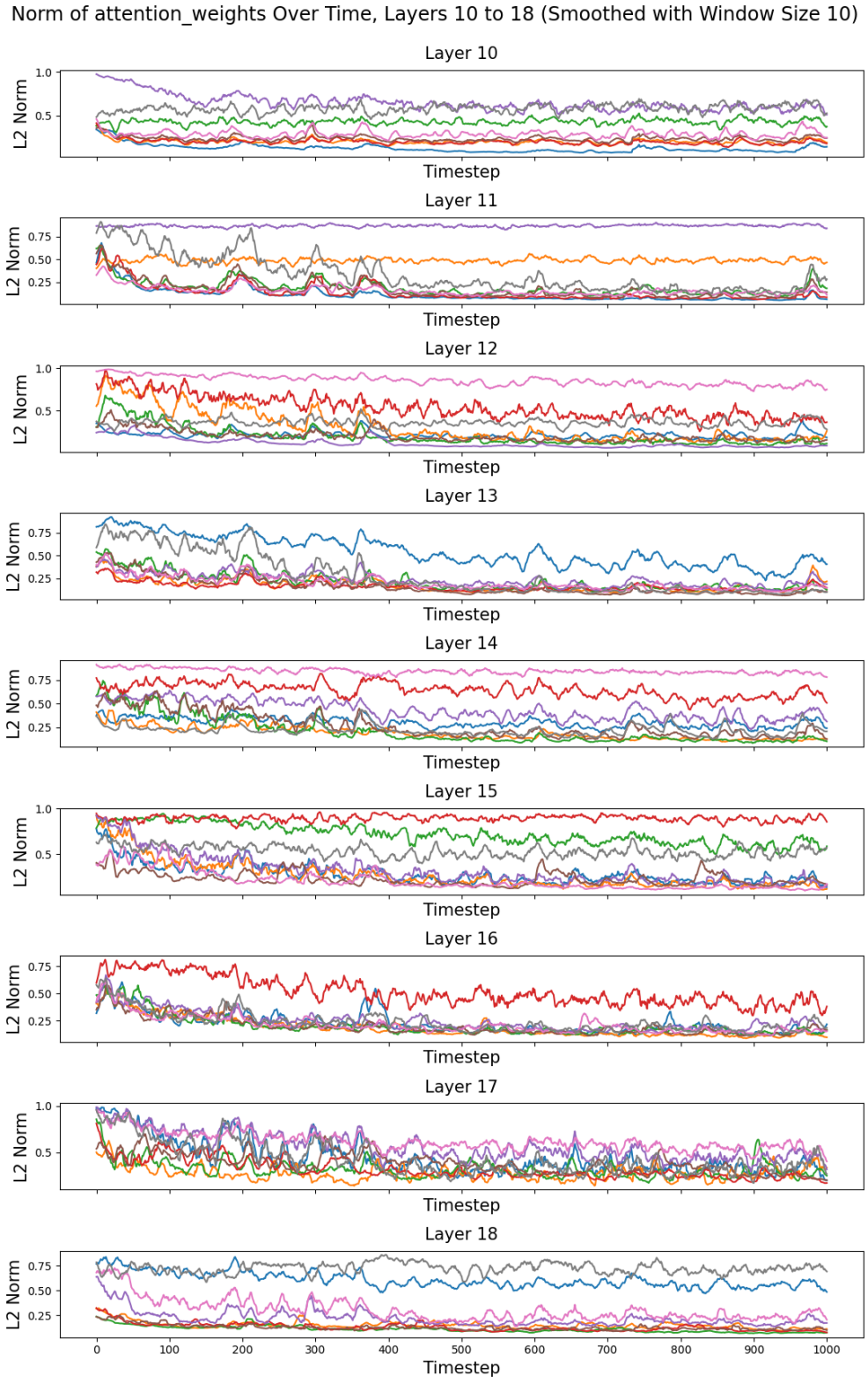


Figure 10: Time series of L2 norms for attention weights across attention heads in layers 10 to 18, with each head represented by a distinct color line. For enhanced visualization, the series has been smoothed using a moving average with a window size of 10.

B Average Attention Head Activation by Cognitive Task Category

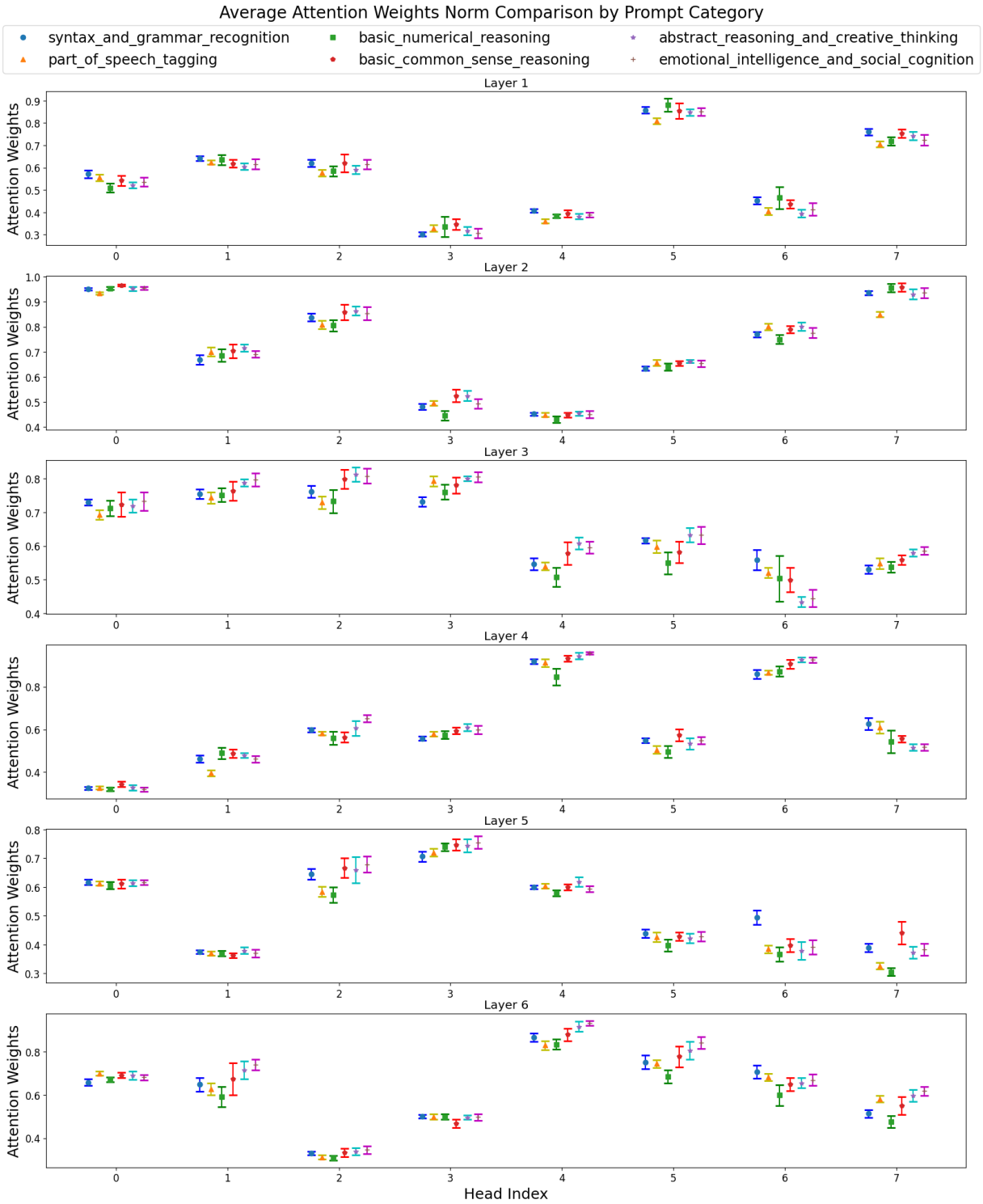


Figure 11: Average activation of attention heads across initial layers (1-6) for different cognitive task categories, presented with 95% confidence intervals. The data illustrate a lack of variability in activation across these initial layers, suggesting that they perform similar operations regardless of the cognitive task category.

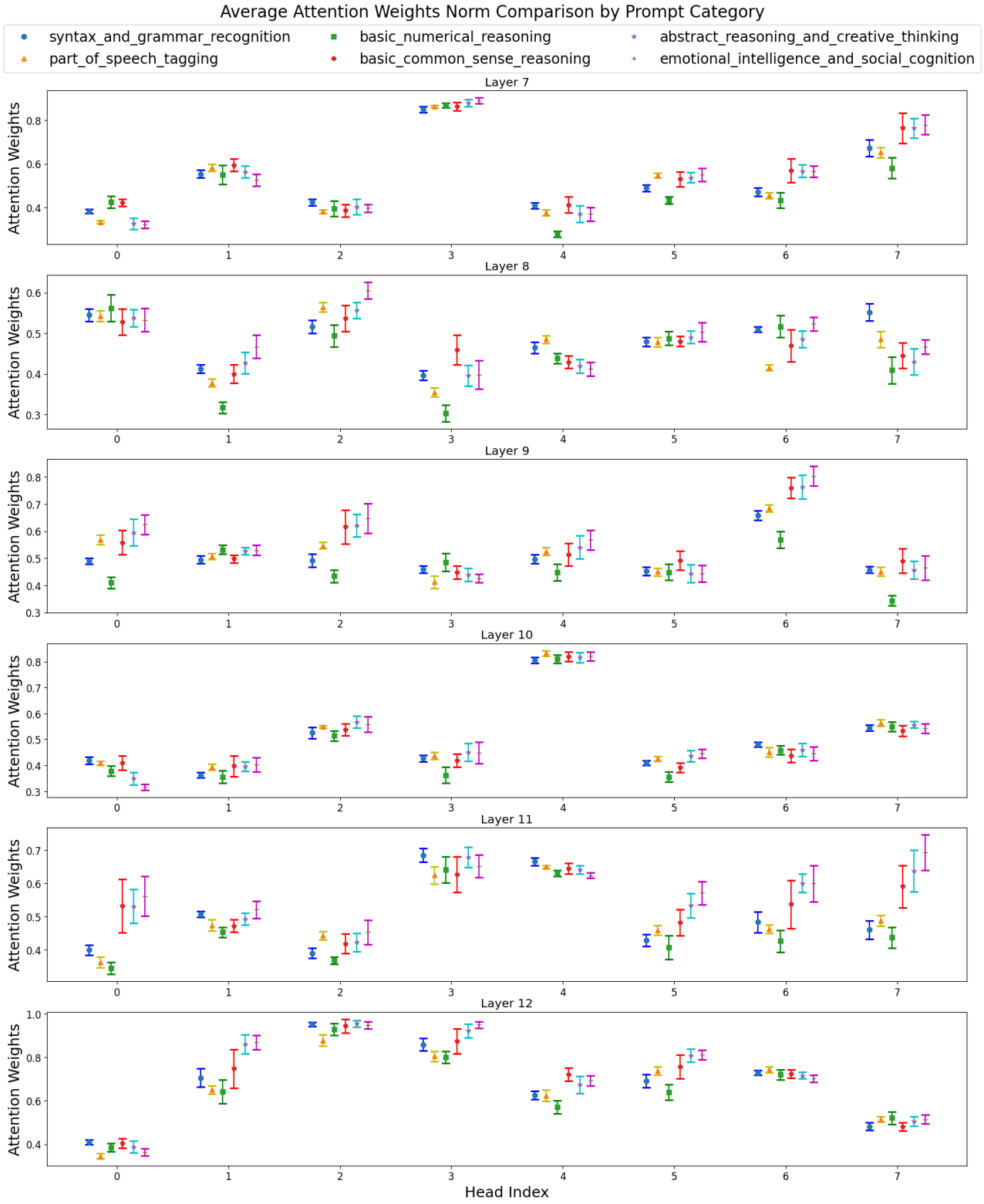


Figure 12: Average activation of attention heads across intermediate layers (7-12) for different cognitive task categories, presented with 95% confidence intervals. The data show increased variability in activation across these layers, suggesting that some attention heads specialize in different cognitive tasks.

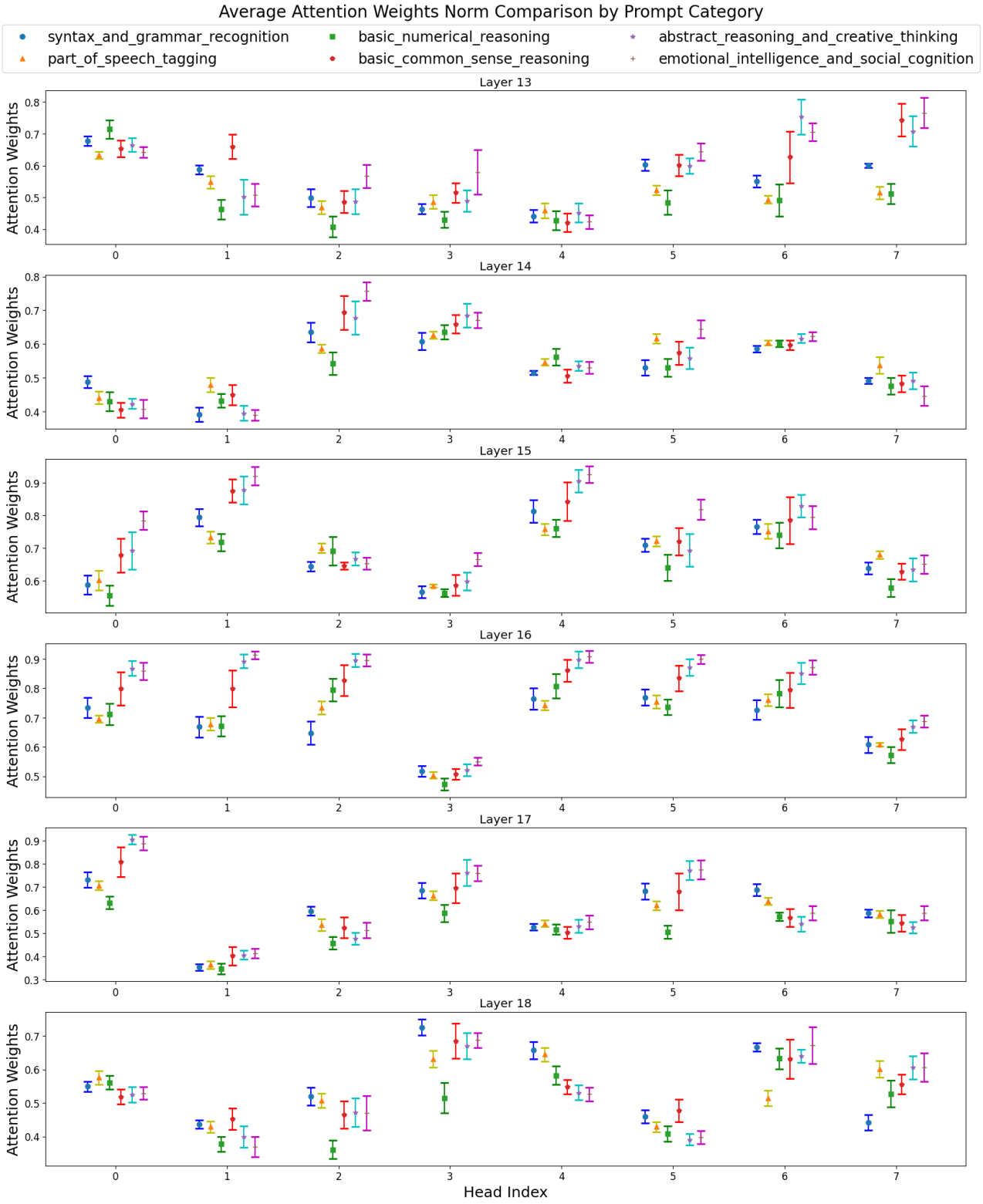


Figure 13: Average activation of attention heads across final layers (13-18) for different cognitive task categories, presented with 95% confidence intervals. The data show great variability in activation across these layers, suggesting that attention heads specialize in different cognitive tasks.

C Prompts used for the Different Cognitive Task Categories

Cognitive Task Category	Prompts
Syntax and Grammar Correction	Correct the error: He go to school every day. Correct the error: She have two cats and a dogs. Correct the error: I eats breakfast at 8:00 in the morning. Correct the error: Every students in the classroom has their own laptop. Correct the error: She don't like going to the park on weekends. Correct the error: We was happy to see the rainbow after the storm. Correct the error: There is many reasons to celebrate today. Correct the error: Him and I went to the market yesterday. Correct the error: The books is on the table. Correct the error: They walks to school together every morning.
Part of Speech Tagging	Identify the parts of speech in the sentence: Quickly, the agile cat climbed the tall tree. Identify the parts of speech in the sentence: She whispered a secret to her friend during the boring lecture. Identify the parts of speech in the sentence: The sun sets in the west. Identify the parts of speech in the sentence: Can you believe this amazing view? Identify the parts of speech in the sentence: He quickly finished his homework. Identify the parts of speech in the sentence: The beautifully decorated cake was a sight to behold. Identify the parts of speech in the sentence: They will travel to Japan next month. Identify the parts of speech in the sentence: My favorite book was lost. Identify the parts of speech in the sentence: The loud music could be heard from miles away. Identify the parts of speech in the sentence: She sold all of her paintings at the art show.
Basic Numerical Reasoning	If you have 15 apples and you give away 5, how many do you have left? A rectangle's length is twice its width. If the rectangle's perimeter is 36 meters, what are its length and width? You read 45 pages of a book each day. How many pages will you have read after 7 days? If a train travels 60 miles in 1 hour, how far will it travel in 3 hours? There are 8 slices in a pizza. If you eat 2 slices, what fraction of the pizza is left? If one pencil costs 50 cents, how much do 12 pencils cost? You have a 2-liter bottle of soda. If you pour out 500 milliliters, how much soda is left? A marathon is 42 kilometers long. If you have run 10 kilometers, how much further do you have to run? If you divide 24 by 3, then multiply by 2, what is the result? A car travels 150 miles on 10 gallons of gas. How many miles per gallon does the car get?

Table 3: Expanded list of prompts used for the different cognitive task categories, Part 1.

Cognitive Task Category	Prompts
Basic Common Sense Reasoning	If it starts raining while the sun is shining, what weather phenomenon might you expect to see? Why do people wear sunglasses? What might you use to write on a chalkboard? Why would you put a letter in an envelope? If you're cold, what might you do to get warm? What is the purpose of a refrigerator? Why might someone plant a tree? What happens to ice when it's left out in the sun? Why do people shake hands when they meet? What can you use to measure the length of a desk?
Abstract Reasoning and Creative Thinking	Imagine a future where humans have evolved to live underwater. Describe the adaptations they might develop. Invent a sport that could be played on Mars considering its lower gravity compared to Earth. Describe the rules. Describe a world where water is scarce, and every drop counts. Write a story about a child who discovers they can speak to animals. Imagine a city that floats in the sky. What does it look like, and how do people live? Create a dialogue between a human and an alien meeting for the first time. Design a vehicle that can travel on land, water, and air. Describe its features. Imagine a new holiday and explain how people celebrate it. Write a poem about a journey through a desert. Describe a device that allows you to experience other people's dreams.
Emotional Intelligence and Social Cognition	Write a dialogue between two characters where one comforts the other after a loss, demonstrating empathy. Describe a situation where someone misinterprets a friend's actions as hostile, and how they resolve the misunderstanding. Compose a letter from a character apologizing for a mistake they made. Describe a scene where a character realizes they are in love. Write a conversation between two old friends who haven't seen each other in years. Imagine a character facing a moral dilemma. What do they choose and why? Describe a character who is trying to make amends for past actions. Write about a character who overcomes a fear with the help of a friend. Create a story about a misunderstanding between characters from different cultures. Imagine a scenario where a character has to forgive someone who wronged them.

Table 4: Expanded list of prompts used for the different cognitive task categories, Part 2.