How Do Artificial Intelligences Think? The Three Mathematico-Cognitive Factors of Categorical Segmentation Operated by Synthetic Neurons

Michael Pichat*
William Pogrund†
Armanush Gasparian‡
Paloma Pichat§
Samuel Demarchi¶
Michael Veillet-Guillem‖

Abstract

How do the synthetic neurons in language models create "thought categories" to segment and analyze their informational environment? What are the cognitive characteristics, at the very level of formal neurons, of this artificial categorical thought? Based on the mathematical nature of algebraic operations inherent to neuronal aggregation functions, we attempt to identify mathematico-cognitive factors that genetically shape the categorical reconstruction of the informational world faced by artificial cognition. This study explores these concepts through the notions of priming, attention, and categorical phasing.

1 Introduction

1.1 Synthetic Explainability and Cognitive Inference

Making an artificial neural network explainable means translating its operations into a language that is accessible and logical for humans [25, 56, 57, 58]. This involves examining the network's observable actions within an interpretative framework that assigns relevant meaning to its operations. In our approach, we

*Neocognition & Faculties of Philosophy and Psychology, Paris
†Neocognition and NP - Phelma, UGA
‡Neocognition
§Neocognition and Faculty of Medicine of Lyon East, University Lyon 1
¶Neocognition (Chrysippe R&D) and Department of Psychology, University of Paris 8
‖Neocognition (Chrysippe R&D) and Epitech Paris

utilize concepts derived from human cognitive psychology as heuristic or analogical bridges between human and artificial intelligence. This requires continuous consideration of potential pitfalls, such as anthropomorphizing algorithms [52], confusing behavior with cognition [12], or merging observer and observed system, a risk highlighted by cybernetics, systems theory, and enactive cognitive science [78, 79, 71, 73].

The practical utility of this cognitive explainability approach unfolds in two directions. First, it helps to prevent erroneous or potentially dangerous responses from the artificial neural system, such as cognitive biases [29], cultural biases [42], hallucinations [40, 49], or excessive emphasis on certain inputs [26]. Second, it improves the efficiency of language models [8] by further aligning them with human expectations [44].

In this study, we explore an approach to explainability focused on a fine cognitive granularity, referred to as mechanistic explainability. Rather than examining network outputs in relation to inputs on a global scale [86], this approach targets a microscopic analysis. Specifically, we delve into the fundamental cognitive units of formal neural networks—synthetic neurons, either individually or in groups within layers [21, 22, 32, 53]. Our objective is to infer the internal cognitive mechanism of artificial networks at a genetic level to understand how the categories and concepts vectorized by formal neurons are locally constituted.

2 Epistemological Status of Synthetic Thought Categories

2.1 Structural and Functional Construction of Synthetic Cognition

By structural (i.e., architectural) and functional (i.e., mathematical) design, the cognition of components within a synthetic neural network is inherently categorical [11, 32, 14, 43, 87, 57, 58]. In simplified terms, the functioning of each formal neuron can be described in three stages:

1. **Integration**: Each formal neuron receives inputs from its precursor neurons, where each input can be interpreted as the degree of membership of a current element (such as a token in language models) to the category associated with a precursor neuron.

2. **Weighted Combination**: Through an aggregation function [1], these inputs are combined to produce a resulting category. This combination is

[1] Bills et al. (2023) provide, on the GitHub repository associated with their article, a list "of the upstream and downstream neurons with the most positive and negative connections." They operationally define these connections as follows: "Definition of connection weights: neuron-neuron: for two neurons (l1, n1) and (l2, n2) with l1 < l2, the connection strength is defined as h{l1}.mlp.c_proj.w[:, n1, :] @ diag(h{l2}.ln_2.g) @ h{l2}.mlp.c_fc.w[:, :, n2]." This list specifies, within the dense layers (i.e., fully connected layers) of GPT2-XL, the weights through which each neuron in an arrival layer n+1 is connected to all neurons in the preceding

enhanced by a non-linear activation function to ensure sparsity [61, 88].

3. **Output Production**: This output will subsequently be used by successor neurons in further processing.

For each member (token) of a synthetic category, an associated activation value indicates the degree to which that element belongs to the artificial category, in alignment with fuzzy logic [84, 81]. The extension of each category can then be defined as the set of elements with a positive activation value, exceeding a specified threshold in the context of a fuzzy α-cut.

In our epistemological framework, synthetic categories, much like human categories [74], are immanent cognitive constructs. Each synthetic category is created during the training phase by the neural network itself. This artificial category acts as a segmentation tool within the vast, undetermined space of potential arguments and predicates [52]. These arguments and predicates may align with existing human-like categories or form entirely novel "alien-like" categories that could represent statistical constructs [11] or "polysemic concepts" [14, 53] not directly relatable to human cognitive categories.

In analyzing the unique categorical segmentation achieved by a synthetic neuron, the critical question is not its ontological alignment with a presumed pre-existing reality but rather its functional role (or, as Varela would say, its "coupling") within the goal-oriented task it is designed for [7]. Thus, in a constructivist perspective, a category is a pragmatic projection rather than the recognition of a pre-given property. Synthetic categories are therefore viewed as similar to "in-action concepts" as described by Vergnaud [74, 75], representing functional arguments and predicates pertinent to task performance without being verbalized, theorized, or consciously realized.

Synthetic categories can be inferred at various levels of neural network granularity: at the level of a single neuron (neural-localized category) [11], at the layer level, or across inter-layer connections (distributed category) [14, 53].

3 Problem Statement

How do synthetic neurons construct the categorical dimensions through which they segment and analyze their environment (e.g., tokens in language models)? What are the developmental characteristics of this artificial categorical thinking, and how are these categories vectorized by synthetic neurons? Specifically, what are the genetic factors that influence or govern these categorical constructions? More precisely, which factors determine the level of membership (i.e., activation level) of a token within a synthetic neural category, thereby shaping the extension and hence the "semantics" of this category? In other words, how do these factors quantitatively and qualitatively constitute the genetic variables of categorical segmentation (of the token world) performed by synthetic neurons?

layer n. These weights are the basis for the linear aggregation functions of neurons referred to in this article.

Investigating these questions requires recognizing that the cognitive and conceptual properties of artificial neural networks do not emerge by magic or chance. In an embodied cognition framework [71, 72, 2, 55], these properties directly result from the specific characteristics of the physical structure within which they emerge.

A central structural and functional component of a neural network is the aggregation function governing the linear combination and vector projection of input categorical dimensions into a resulting categorical dimension. This aggregation function, along with other elements (including the activation function), genetically and functionally shapes the dimensional categorical segmentation specific to each formal neuron.

Observing the nature and operators constitutive of this aggregation function, of the form $\sum(w_{i,j}x_{i,j})+a$, suggests that it mathematically generates and formats the categorical segmentation performed by synthetic neurons through at least three mathematico-cognitive factors. We will investigate these factors in this exploratory work: the first factor is associated with the variable $x_{i,j}$, representing the activation values of categorical outputs from precursor neurons, cognitively interpreted as categorical priming (or effect X). The second factor relates to the parameter $w_{i,j}$, the weighting assigned to these outputs, interpreted as categorical attention (or effect W). Finally, the third factor concerns the linear additive combination of the terms $w_{i,j}x_{i,j}$ within the aggregation function, cognitively denoted as categorical phasing (or effect $\sum$).

4 Methodology

4.1 Methodological Positioning

To better understand the positioning of our exploratory work, we provide a brief, non-exhaustive overview of various technical approaches that, with varying levels of cognitive granularity, seek to extract informational content or processes within formal neural networks, whether organized in layers, groups, or complete networks. These approaches are not mutually exclusive and may partially overlap.

As previously mentioned, studies at a macro-cognitive level focus on analyzing the differences between inputs and outputs to understand the relationship between initial data and outcomes in a language model. Among these methods, gradient-based approaches evaluate the role of each input by exploiting the derivatives relative to each input dimension [30]. Input characteristics can be evaluated based on elements such as features [23], token importance scores [30], or attention weights [5]. Concurrently, example-based approaches aim to observe how outputs vary with different inputs by examining the effect of slight input modifications (e.g., deletion, negation, mixing, or masking) [4, 80, 70]. Additionally, some studies focus on concept mapping of inputs to quantify their contributions to observed results [17].

Approaches with finer cognitive granularity focus on the intermediate states

of the language model rather than its final output, examining partial outputs or internal states of neurons or groups of neurons. In this context, certain approaches analyze and linearly decompose the activation score of a neuron in a given layer concerning its inputs (neurons, attention heads, or tokens) from the previous layer [76]. Other methods tend to simplify activation functions for easier interpretation [77]. Furthermore, some techniques, leveraging the model's vocabulary, focus on extracting encoded knowledge by projecting connections and intermediate representations through a matching matrix [24, 35]. Finally, certain methodologies use neural activation statistics in response to data sets [11, 50, 28, 77, 20]. Our exploratory study specifically fits within this last category.

4.2 Methodological Choices

In this exploratory research, we focus on the GPT model proposed by OpenAI, specifically its GPT-2XL version. This choice is due to GPT-2XL's sufficient complexity, allowing us to examine advanced synthetic cognitive phenomena without reaching the sophistication of GPT-4 or its multimodal version, GPT-4o. A practical consideration also guided our preference for GPT-2XL: in 2023, OpenAI shared, in the article by Bills et al. [11], parameter details as well as activation values for its neurons, which serve as the basis for our analysis.

For simplicity, this exploratory study is limited to the first two layers of GPT-2XL (layers 0 and 1), each comprising 6,400 neurons. Regarding tokens and their activation values among these 12,800 formal neurons (i.e., 2 x 6,400), we have decided to consider, for each neuron, the 100 tokens with the highest average activation values (referred to as "core-tokens").

4.3 Statistical Choices

Our descriptive and inferential statistical analyses were conducted using Python's SciPy library, following guidance from Howell [39] and Beaufils [10].

To assess the normality of our data, a necessary condition for performing parametric tests, we adopted a dual approach. First, we employed various inferential tests: the Shapiro-Wilk test (effective for small samples), the Lilliefors test (suitable for small samples when normal distribution parameters are unknown and estimated from the data), the Kolmogorov-Smirnov test (preferred for large samples), and the Jarque-Bera test (focusing on symmetry and kurtosis, valid for large samples). Second, we used a descriptive approach with indices such as skewness and kurtosis, and graphical methods like the QQ-plot to compare the observed distribution with a theoretical normal distribution.

The results, not reproduced here, indicate a relatively mixed normality in our data, leading us primarily towards Spearman's ordinal correlation studies in analyzing relationships between variables associated with our hypotheses. This approach allows us to avoid normality prerequisites and mitigate bias introduced by outliers. When necessary, we applied univariate goodness-of-fit tests to infer

the significance of observed phenomena (notably regarding the positivity and significance of ordinal correlations obtained for each neuron in layer 1).

In our statistical framework, the composite units include the 6,400 "destination" neurons in layer 1, their 100 respective core-tokens (tokens with the highest average activation levels), as well as the 10 precursor neurons (from layer 0) with the highest connection weights to each destination neuron. We focused on the 100 tokens most highly activated by each neuron, deeming it less relevant initially to examine tokens weakly or not activated by them, as they fall partially outside the extension of the category associated with each neuron.

4.4 Objective and Implementation of the Study in Terms of Statistical Observables

The objective of this exploratory study is to identify synthetic cognitive factors that partially drive the categorical segmentation performed by formal neurons. These factors are mathematically embedded in the neural aggregation function and influence the identification of core-tokens for a given neuron, that is, the determination of the content of its categorical extension.

More specifically, we aim to verify to what extent the membership of a core-token to the specific category of a destination neuron depends on three cognitive factors that we will define and propose: categorical priming, categorical attention, and categorical phasing. The level of membership of a core-token (in layer 1, the destination layer) to the category associated with a neuron will be measured by the activation value of this token within the relevant neuron. Priming will be evaluated based on the activation value of a token in its respective precursor neurons (in layer 0). Attention will be assessed through the connection weights linking destination neurons (layer 1) to their top 10 precursor neurons (those with the highest connection weights) in layer 0. Finally, categorical phasing will be quantified by analyzing the frequency with which a core-token within a destination neuron (layer 1) also appears as a core-token among the 10 associated precursor neurons (layer 0).

5 Definition of Synthetic Cognitive Concepts Studied and Results

5.1 Synthetic Categorical Priming

In human psychology, priming [3, 18, 83, 38] is a cognitive process in which an initial stimulus triggers a preliminary stage of cognitive processing, thus facilitating, accelerating, or preparing the reception of a second, related stimulus. Specifically, semantic priming is a process by which the meaning of one element (e.g., a word) becomes more accessible to an individual through prior exposure to another semantically related element. The priming effect is typically studied in terms of response delay in lexical decision or text comprehension tasks, where

response time can indicate the existence, structure, and strength of semantic relationships between words and concepts in long-term semantic memory.

The notion of priming is related to that of activation [47, 15, 46], postulating that cognitive contents or processes can exhibit variable intensity levels of activity. Prototypical examples involve biological neural structures whose activity levels can be physiologically "directly" measurable (even if this measurement is partly a methodological and statistical reconstruction). In the case of priming, activation is conceptualized as the propagation of activation: a cognitive characteristic (e.g., meaning) is "spread" from an entity A (which activates first) to an entity B (which activates as a causal result) (e.g., from one word to another) if A and B are structurally or temporarily linked.

We hypothesize a transposition of the concept of priming, as defined above in the fields of neuroscience and human cognitive psychology, into the domain of synthetic cognition. Mathematically, due to the construction of the aggregation function $\Sigma(w_{i,j}x_{i,j}) + a$, for a given element (e.g., a token or other), the activation value of the category carried by a destination neuron (on layer n) is directly a function (modulo the activation function) of the activation values $x_{i,j}$ of the categories associated with its precursor neurons (on the subordinate layer $n-1$). In other words, in epistemological alignment with the original notion of priming, the prior activation (when it exists for a given token) of the categories vectorized by precursor neurons should "mathematically propagate" the activation of the category associated with their corresponding destination neuron. We thus formulate, in these terms, a hypothesis of synthetic categorical priming within artificial neural networks.

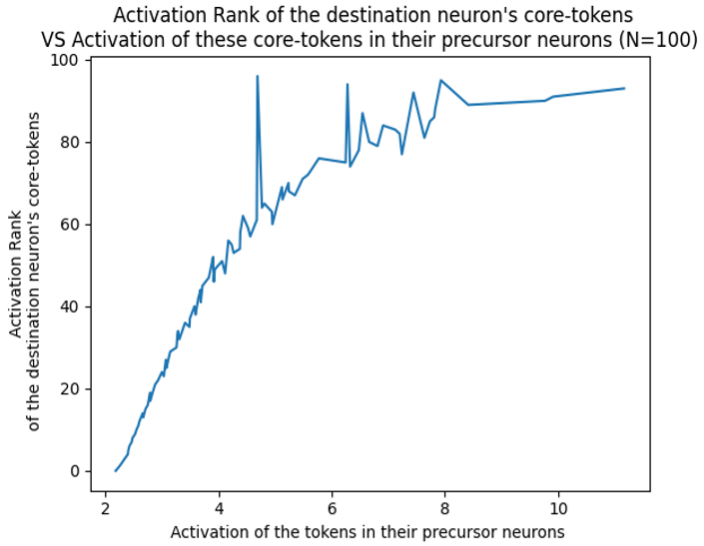
From a quantitative perspective, the empirical observable associated with our hypothesis of synthetic categorical priming is the activation value of destination neurons as a function of their precursor neurons. Specifically, data compatible with our hypothesis should show, for a given series of tokens, a relationship between the activation value of destination neurons on layer $n+1$ and that of their respective precursor neurons. We operationalize this approach on the 6,400 neurons in layer 1 of GPT-2XL, considering for each destination neuron its 10 precursor neurons with the highest connection weights and its 100 tokens associated with the highest average activation values (core-tokens) (only core-tokens activated in at least one precursor neuron are included).

Statistically, we test an ordinal relationship (Spearman's ρ) between the average activation rank (ranging from 1 to 100) of the 100 core-tokens of each of the 6,400 destination neurons in layer 1 and the mean cumulative activation values (i.e., summed) of these tokens within the 10 associated precursor neurons (each core-token of a destination neuron having a non-negative activation value for each of the 10 relevant precursors).

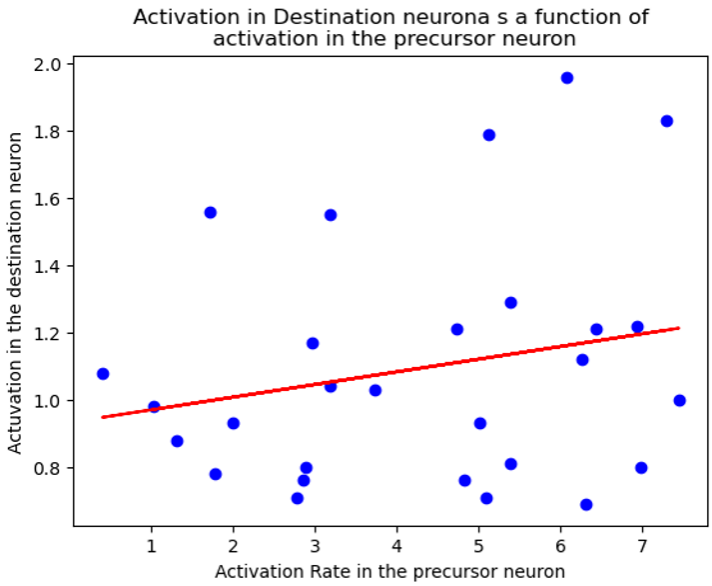
Table 1 shows a positive relationship with an extremely strong effect size ($\rho = .94$) and statistical significance ($p < .001$). Figure 1 illustrates this overall positive monotonic trend, though with occasional pronounced peaks in variability. Figure 2 provides a view for an example neuron, with a regression line again showing a positive relationship, although less pronounced in this case.

Strength of ρ	
(ρ)	.94
Significance of ρ	
p	.0000

*Table 1 : Spearman's correlation between activation rank of the destination neuron's core-tokens and mean activation of these tokens in their precursor neurons (Layer 1, Nmax=6400*100*10).*



Graph n°1 : Activation rank of the destination neuron's core-tokens as a function of the mean activation of these tokens in their precursor neurons (Layer 1).



Graph n°2 : Activation of the destination neuron's core-tokens as a function of the mean activation of these tokens in their precursor neurons (Layer 1, Control neuron 3000).

These results appear compatible with our hypothesis of synthetic categorical priming, which we term effect "X"—a mathematico-cognitive propagation of activation from precursor neural categories to their associated destination category in the superordinate layer.

5.2 Synthetic Categorical Attention

In human cognitive psychology, attention is defined as a specific calibration of activity according to its purpose, resulting in greater efficiency in information intake processes (including selectivity) and execution processes (including precision and speed) [60, 66, 59, 63, 69, 27, 68, 19, 82, 36]. In terms of external information intake, attention is related to conceptualization [74, 75], meaning the identification of only those parameters (objects relevant to the activity) whose consideration is crucial for successful task performance. Actions must thus be adjusted to these parameters to ensure efficiency. Here, attention involves filtering and structuring the excessively large amount of available perceived information, or inhibiting information deemed irrelevant, in order to focus mental effort and informational selectivity on specific objects and properties. Regarding task execution, attention is linked to the control, by the central system, of the activity, which may involve assigning varying degrees of weight (priority, order, reliability, etc.) to certain internal information (knowledge, representations, schemas) or verifying the quality of task performance within its temporal sequence.

From a physiological perspective, attention is driven by the limited information-processing capacity of the nervous system, leading to selective choices in the integration, activation, and utilization of sensory data or stored memory (semantic, procedural) [34, 6]. This process is achieved through an orientation response, which directs information-seeking activities toward a specific type of informational characteristics.

We hypothesize here a transposition of the concept of attention, as previously described in cognitive psychology and human neuroscience, into the field of artificial cognition. This is based on the mathematical construction of the aggregation function $\Sigma(w_{i,j}x_{i,j}) + a$, where, for a given element (a token), its activation value within the category associated with a destination neuron is inherently dependent (apart from the activation function) on the connection weights $w_{i,j}$ between this destination neuron and its precursor neurons. In other words, and in epistemological continuity with the original concept of attention, the connection weights with precursor neurons act as direct regulators of the level of information uptake (i.e., activation levels) derived from these precursor neurons—ranging from inhibition or filtering of data for negative, near-zero, or weakly positive weights, to strong mathematical-cognitive focus and integration for significant weights.Thus, in terms of execution, the neuronal connection weights govern the degree of information utilization that the artificial cognitive system deems relevant from preceding synthetic categories in performing the current task of a given successor neuron, which involves calculating the degree of membership of a given token in the category constitutive of this superordinate neuron. We define this hypothesis as synthetic categorical attention within artificial neurons, which we denote as effect "W."

5.2.1 Quantitative Approach to Synthetic Categorical Attention

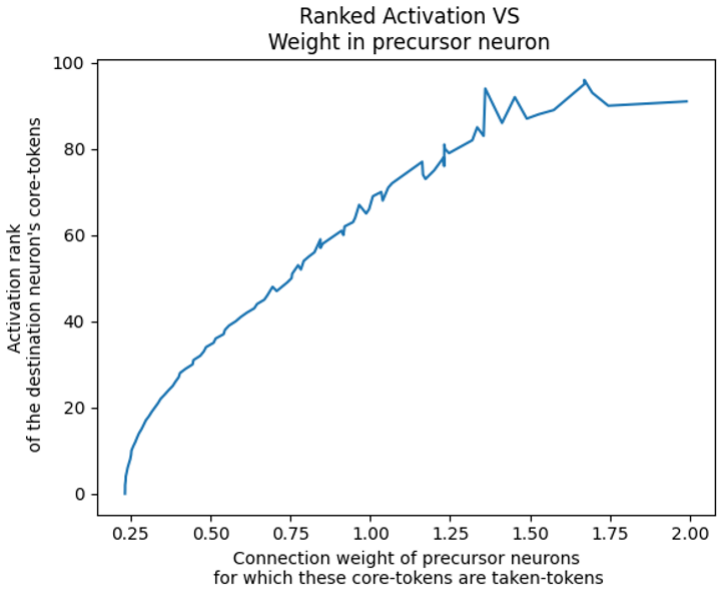
Quantitatively, the empirical observable associated with our hypothesis of synthetic categorical attention is, for a given token, the activation value of destination neurons as a function of their connection weights with respective precursor neurons. To test this hypothesis, we examine the relationship between the activation value of destination neurons and the connection weights with their precursor neurons. According to this hypothesis, activation should increase with higher values of these antecedent weights. We apply this approach to the 6,400 neurons in layer 1 of GPT-2XL, taking into account for each destination neuron its 10 precursor neurons with the highest connection weights and its 100 tokens with the highest average activation values (core-tokens) (note that only core-tokens activated in at least one precursor neuron are considered). From a statistical perspective, and in a more operationalized form, we test for an ordinal relationship (measured with Spearman's ρ) between (i) the average activation rank (ranging from 1 to 100) of the 100 core-tokens of each of the 6,400 destination neurons in layer 1, and (ii) the average cumulative connection weights (i.e., summed) with their respective (1 to 10) precursor neurons for which these tokens are also core-tokens.

In Table 2, we observe a positive ordinal relationship between the activation

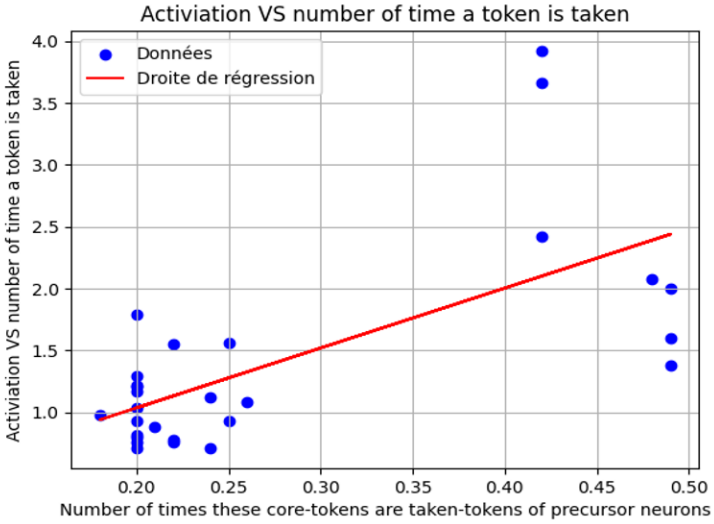
rank of destination neurons and the average cumulative connection weights with their respective precursor neurons. This relationship has an extremely strong effect size ($\rho = .999$) and high statistical significance ($p < .001$). Figure 3 illustrates this positive monotonic relationship across the entire dataset, while Figure 4 provides an example for a control neuron, with a regression line again showing a positive relationship, albeit less pronounced in this case.

N_{max}	6400*100*10=6400000
n(mean ranks)	100
(ρ)	.999
p	.0000

Table 2: Spearman's correlation between activation rank of the destination neuron's core-tokens and mean connection weights with their precursor neurons (Layer 1).



Graph n°3: Activation rank of the destination neuron's core-tokens as a function of the mean cumulative connection weights with their precursor neurons (Layer 1).

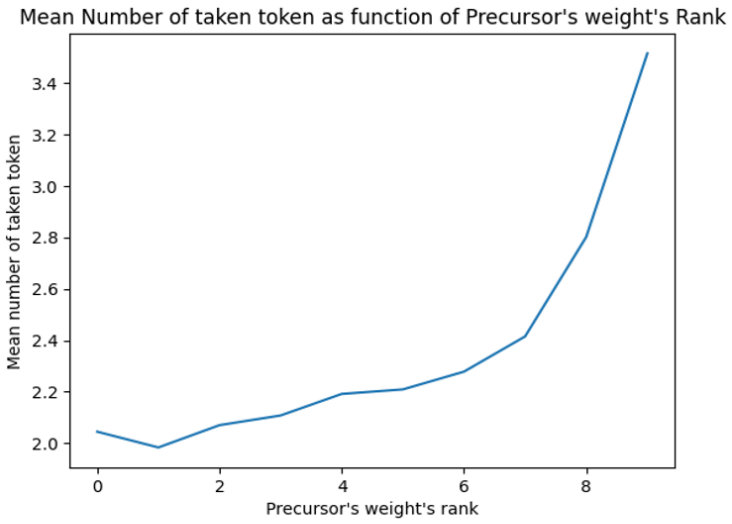


Graph n°4: Activation of the destination neuron's core-tokens as a function of the cumulative connection weights with their precursor neurons (Layer 1; control neuron 3000).

The exploratory data obtained here are compatible with our hypothesis of synthetic categorical attention, positing a positive monotonic ordinal relationship between the activation level of core-tokens in destination neurons and the connection weights of these destination neurons with precursor neurons that also contain these same tokens as core-tokens.

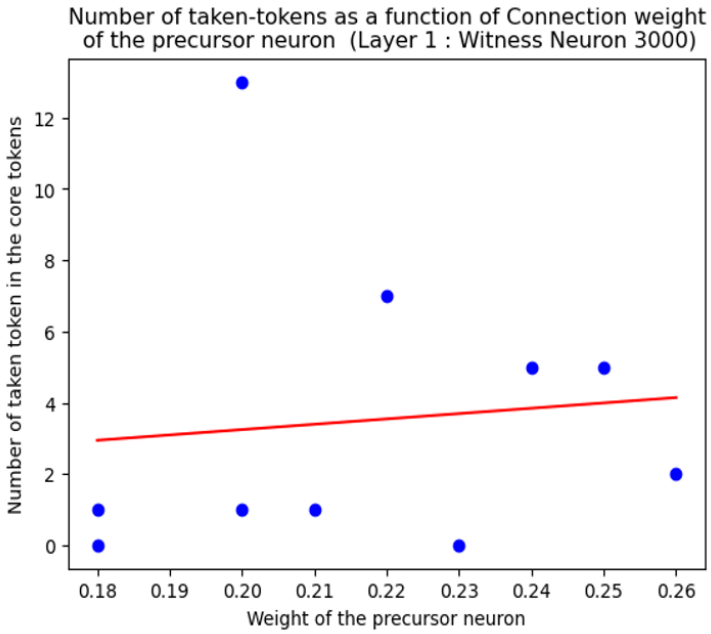
Still in our quantitative approach, we now seek to gain deeper cognitive insight into this synthetic categorical attention by examining its modus operandi in terms of information selection at the input of destination neurons. An intriguing question in this regard is the relationship, for a given destination neuron, between the intensity of its connection weights with its precursor neurons and the number of "shared" core-tokens between this destination neuron and its precursor neurons. This question can be reframed as follows: to what extent do precursor neurons with high connection weights contribute more core-tokens to their destination neurons? In other words, to what degree do strongly connected precursor neurons more actively influence the constitution of the categorical extension content of their destination neurons (i.e., the composition of their core-tokens)? Alternatively, to what extent does connection weight regulate the definition of the extension, and therefore the selection and categorical segmentation specifically operated by a given (destination) synthetic neuron? Our analysis reveals an extremely strong and significant positive ordinal correlation ($\rho = .989, p < .001$) between (i) the average rank of the connection weights of each destination neuron (in layer 1) with its precursor neurons, and (ii) the average number of core-tokens in the destination neuron that were also previously core-tokens of the involved precursor neurons (see Figure 6). This analysis includes $n = 6,400$ destination neurons in layer 1 and 10 precursor

neurons in layer 0, totaling 64,000 cases.



Graph n°5 : Mean number of taken-tokens as a function of percursor's weight's rank (Layer 1).

We refer to such tokens as "taken-tokens"—tokens that are core-tokens in precursor neurons and are "reused" as core-tokens by their respective successor neurons. This result, consistent with the nature of the aggregation function, indicates that stronger attention weights lead to an overrepresentation of these taken-tokens. A high attention weight associated with a precursor neuron thus functions, in terms of information selection, as an "extractor" of a categorical sub-dimension (composed of the relevant taken-tokens) from the precursor neuron's categorical dimension. This sub-dimension, in turn, genetically "feeds" the extension (of core-tokens) of the category represented by the destination neuron, thereby contributing to its specific categorical segmentation. Figure 7 illustrates this trend with a sample neuron from layer 1.



Graph n°6 : Number of taken-tokens as a function of percursor's weight (Layer 1, control neuron 3000).

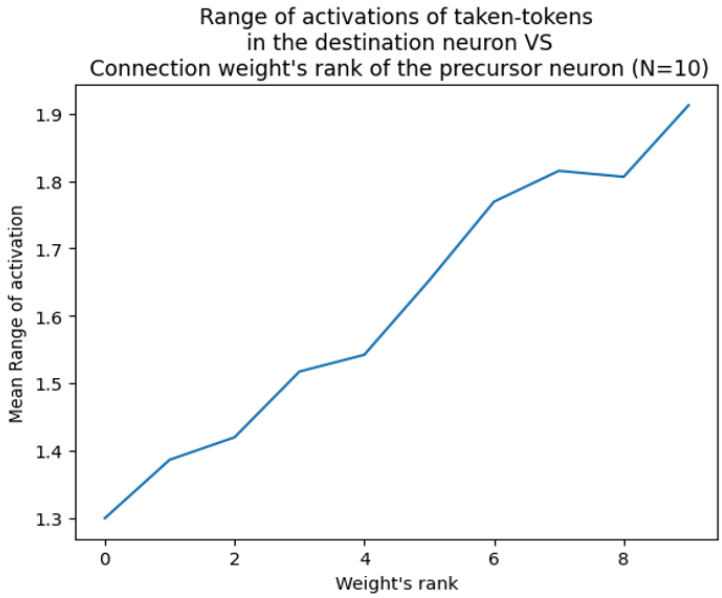
From a quantitative perspective, another highly interesting and informative result further illuminates the cognitive mechanism of synthetic categorical attention. Consistent with the mathematical nature of the aggregation function, we observe an extremely strong positive ordinal correlation ($\rho = .988, p < .001$) between (i) the rank of precursor-successor connection weights and (ii) the average activation range of associated taken-tokens in destination neurons. For this calculation, we only consider precursor neurons associated with at least two taken-tokens (Nmax = 6,400 destination neurons in layer 1 x 100 tokens x 10 precursor neurons). This trend is clearly illustrated in Figure 7.

In human categorization, Thibault (1997) and Roads et al. (2024) note, regarding Nosofsky's (1986) "generalized context model" of categorization, that the use of a weighted distance metric (specifically, Minkowski distance) to account for selective attention is associated with changes (contraction or expansion) in the metric of the categorical representation space: low attention weights "bring stimuli closer together" within the implicated dimension, whereas high attention weights (strong attention) "stretch" the representation space along that dimension, thereby increasing stimulus discrimination.

In the context of synthetic categorization, this is precisely what we observe here: a destination neuron with a high connection weight to a given precursor neuron displays greater variability in the activation range of its taken-tokens originating from this precursor neuron. In other words, taken-tokens

are more distinctly discriminated regarding their degree of membership to the category represented by this destination neuron. Put differently, strong connection weights increase the activation span of core-tokens in destination neurons, enabling better differentiation and sharper contrast in the degree of membership of a given token to the implicated category.

Thus, synthetic categorical attention may be associated with the discriminative power of a neuronal category and, consequently, its analytical precision within its specific token segmentation dimension. This aligns epistemologically with the conceptual characteristics of attention as defined in human psychology.



Graph n°7 : Mean range of activation of taken-tokens in the destination neuron as a function of percursor's weight's rank (Layer 1).

In a quantitative context, the empirical results of our current exploratory study appear to be compatible with a phenomenon of synthetic cognition—namely, artificial categorical attention, referred to as effect "W." This, in turn, points successively to three potential characteristics of this phenomenology associated with a significant attention-weighted connection: (i) the selection, by a destination neuron, of specific informational characteristics (i.e., certain types of core-tokens) from its precursor neurons (and not others), (ii) the associated extraction by a destination neuron of a particular sub-dimension (of core-tokens) within the categorical dimension carried by each of its precursor neurons, and (iii) the contrast enabling finer differentiation (reflected by activation level) of different types of elements constituting the extension (of core-tokens) of the category specific to a destination neuron.

5.2.2 Qualitative Approach to Synthetic Categorical Attention

Let us now delve deeper into these three converging characteristics (which are ultimately only alternative facets of one another) of synthetic categorical attention through a qualitative exploration of this phenomenology. For this purpose, we employ qualitative examples illustrating how the categories carried by precursor neurons with high attention-weighted neural connections selectively contribute to and generate the content (in terms of core-tokens) of the categorical extensions of their respective destination neurons. This occurs, as we will observe, through a process of "categorical complementation," which involves selectively focusing the computational attention of the aggregation function of the destination neuron on specific categorical sub-dimensions extracted from precursor neurons, thus constructing the unique categorical nature of this destination neuron—that is, the specific content of its categorical extension in terms of core-tokens.

Here, as a purely illustrative example (see Table 3) and without aiming for exhaustiveness, is a comparison of different categorical types of core-tokens selectively "contributed" through a process of categorical complementation by various precursor neurons with high attention-weighted connections. This process progressively builds, sub-category by sub-category, the categorical extension specific to their associated destination neuron. We qualitatively identify two main classes of categorical complementation: linguistic and non-linguistic.

Let us first examine linguistic categorical complementation. This can be semantic in nature, meaning it consists of categorical additions that can be interpreted in terms of operations analogous to human semantics:

- **Intra-lexical complementation**, consisting of adding tokens from the same root (tokenization variants). Example: a precursor neuron "contributes" the token "manager" to the destination neuron, another contributes "manag," and yet another provides "managerial." Intra-lexical complementation may also involve tokens from different roots; for instance, one precursor provides the token "manager" while another provides "director" (the lexical field remains consistent in this case).

- **Sub-lexical complementation**, consisting of adding tokens from a lexical sub-field. Example: a precursor neuron supplies the destination neuron with the token "manager," while another provides the tokens "Wenger," "Klopp," and "Mourinho" (these refer to football coaches and constitute a lexical sub-category of "manager").

- **Peri-lexical complementation**, involving the addition of tokens from a related lexical field. Example: one precursor provides "listen" while another provides "sound"; or one precursor neuron provides "order" and another "request"; or yet another provides "necessary" while another supplies "indispensable."

- **Para-lexical complementation**, through the addition of tokens from an antonymic lexical field. Example: one precursor neuron contributes "love"

and “adore” to the destination neuron, while another provides “hate,” “despise,” and “dislike.”

Linguistic categorical complementation can also be graphemic in nature. Example: a precursor neuron provides the token “Said” to the destination neuron, and another precursor provides the token “id” (both containing the same grapheme “id”).

Finally, linguistic categorical complementation can be phonological. Example: one precursor provides the tokens “be” and “bee,” while another provides “Eve” and “ea” (both containing the same sound /i/).

Turning to non-linguistic categorical complementation:

- **Quantitative complementation**: Example: a precursor neuron provides the tokens “er,” “cv,” and “ku,” while another provides “od,” “fx,” and “yw” (each token consistently contains exactly two graphemes).
- **Cultural complementation**: This type involves elements shared within a given human culture. Example: one neuron provides “ObamaCare,” while another supplies “Congress” (the U.S. Congress enacted this legislation in March 2010).
- **Other types**: These may not necessarily align with human thought categories but are based on statistical contingencies identified by the neural network during training. We term these “alien categories” or “non-human-like categories,” or even “polysemic categories” from our human cognitive perspective. Example: a precursor neuron provides the token “manager,” while another associates the token “ID,” without an observable (human) logic linking them.

Complémentation catégorielle linguistique	Sémantique	Intra-lexicale	(manager) VS (manag) ; (manager) VS (managerial)
			(manager) VS (director)
		Sub-lexicale	(manager) VS (wenger, Klopp, Mourinho)
		Péri-lexicale	(listen) VS (sound) ; (order) VS (request) ; (necessary) VS (indispensable)
		Para-lexicale	(love, adore) VS (hate, despise, dislike)
	Graphémique		(Said) VS (ID)
	Phonologique		(be, bee) VS (Eve, ea)
Complémentation catégorielle non linguistique	Quantitative		(er, cv, ku) VS (od, fx, yw)
	Culturelle		(ObamaCare) VS (Congres)
	Autre		(manager) VS (ID)

Table n°3 : Exemples de modalités qualitatives de complémentation catégorielle (Layer 1).

These illustrative examples, again without aiming for exhaustiveness or systematicity, help us understand how the process of incoming information selection can qualitatively operate through the mechanism of synthetic categorical attention. This occurs through an activity of categorical complementation, allowing the aggregation function of a destination neuron to extract from each of its precursor neurons with high attention-weighted connections a specific categorical sub-dimension that contrasts with others. The successive apposition of these sub-dimensions thus generates, sub-dimension by sub-dimension, the categorical content of the dimensional segment carried by this destination neuron.

5.3 Synthetic Categorical Phasing

Through the construction of the aggregation function $\Sigma(w_{i,j}x_{i,j}) + a$, we postulate a third mathematico-cognitive factor influencing the level of token attribution to a neuronal categorical dimension. We term this factor "synthetic categorical phasing," or effect "Σ," as the aggregation function of a destination neuron sums, for a given token, the weighted values of its activations $w_{i,j}x_{i,j}$ within its respective precursor neurons. Several studies in human cognitive psychology and neuroscience involving the notion of phasing could potentially serve as partial analogies for synthetic cognition in this area; for example, studies on perceptual modality topics [48, 41] or brain synchronizations [1, 16, 62, 67, 65].

In the realm of synthetic cognition, we define synthetic categorical phasing by the notion that a token, previously highly activated for different precursor neurons (i.e., a core-token of these precursor neurons), must, due to the mathematical construction of the aggregation function, be associated with a high activation level within the related destination neuron. This is because the token is co-activated within the various terms constituting the aggregation function; this co-activation leads to an additive concatenation, resulting in a significant activation level for this token at the output of the destination neuron. Such a token is therefore theoretically subject to the phasing of the neural categories of the involved precursors: these precursor categorical segments, though conceptually potentially distinct, are jointly activated, generating a categorical "echo" or "resonance" for this specific token. This occurs through a categorical intersection traced across these dimensions, thereby strengthening the output activation level of the destination dimension.

5.3.1 Quantitative Approach to Synthetic Categorical Phasing

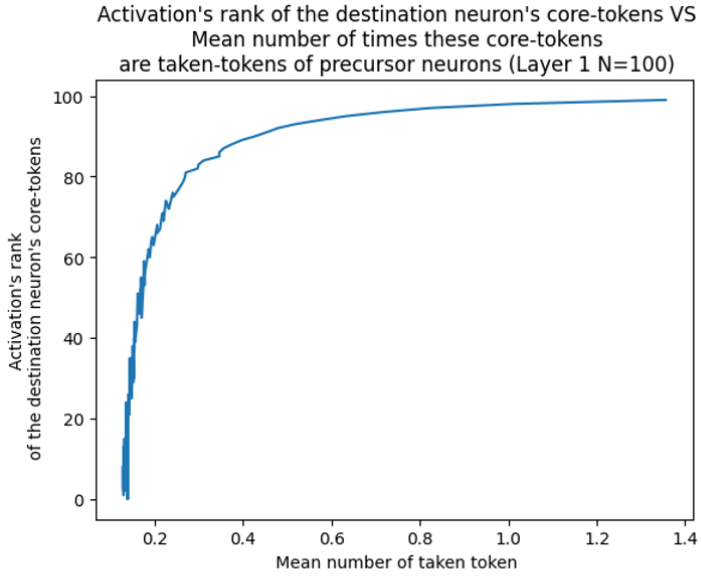
Quantitatively, we operationalize our hypothesis of categorical phasing as follows: the more a destination neuron's core-token is also a core-token for a greater number of its precursor neurons (with high connection weights), the higher its activation level in this destination neuron. This hypothesis thus posits a positive monotonic relationship between these two variables. Table 4 presents the compiled results from local-level analytic testing of this hypothesis, i.e., for each of the 6,400 individual destination neurons in layer 1. We observe a strong ordinal correlation between the two variables, with a large effect size (Mean (ρ) = .976), high significance (% of ($p(\rho) < .05$) = 99.40%; $p(\chi^2) < .0001$), and overwhelmingly positive directionality (% of ($\rho > 0$) = 99.45%, $p(\chi^2) < .0001$). Table 5 displays the results of global-level testing of this hypothesis across all data as a whole (Nmax = 6,400 neurons in layer 1 x 10 precursors in layer 0 x 100 core-tokens). We again find a strong, positive, and significant ordinal correlation between the two variables ($\rho = .989$, $p(\rho) < .001$). Figure 9 graphically illustrates this trend, showing a pronounced logarithmic distribution leading to an asymptotic plateau, while Figure 10 provides an example for a control neuron with a distinctly positive regression line.

Strength of ρ	
Mean (ρ)	.976
Significance of ρ	
% of (p(ρ)<.05)	99.40%
p(χ2) of (p(ρ)<.05)	.0000
Positivity of ρ	
% of (ρ>0)	99.45%
p(χ2) of (ρ>0)	.0000

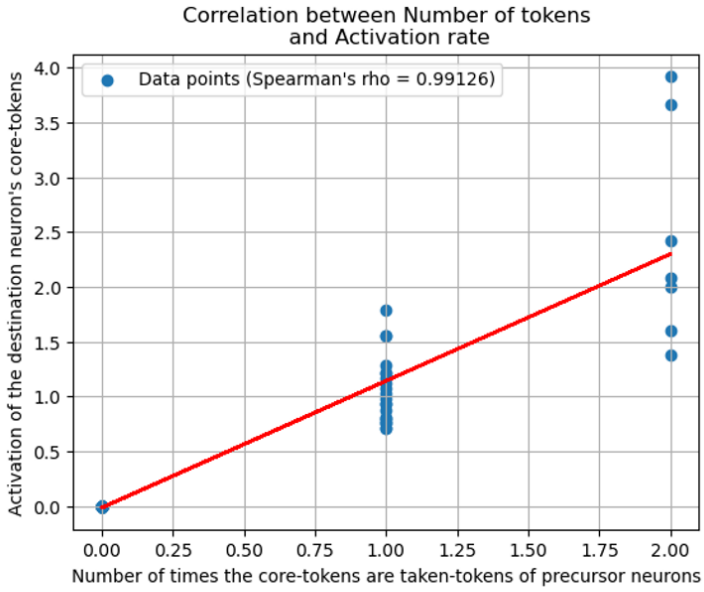
Table n°4 : Spearman's ρ correlation: Core-tokens' activation in the destination neuron VS Number of times these core-tokens are core-tokens of related precursor neurons (N=6400 ; Layer 1).

Strength of ρ	
(ρ)	.989
Significance of ρ	
p	.0000

Table n°5 : Spearman's ρ correlation: Core-tokens' activation's rank in the destination neurons VS Mean number of times these core-tokens are core-tokens of related precursor neurons (Layer 1).



Graph n°8 : Core-tokens' activation's rank in the destination neurons as a function of mean number of times these core-tokens are core-tokens of related precursor neurons (Layer 1).



Graph n°9 : Core-tokens' activation's rank in the destination neuron as a function of number of times these core-tokens are core-tokens of related precursor neurons (Layer 1, Control neuron 3000).

Quantitatively, the results obtained support our hypothesis of categorical phasing, termed effect Σ, positing that the more a token is strongly activated (core-token) in multiple precursor neurons, the more likely it is to be strongly activated at the output level of the associated destination neuron. This token thereby appears at the categorical intersection of the categories represented by these precursors, which are thus locally phased.

5.3.2 Qualitative Approach to Synthetic Categorical Phasing

Let us now, within a qualitative framework, establish reference points to understand the cognitive modalities through which categories—initially distinct or, at the very least, non-isomorphic—associated with precursor neurons can become locally phased categorically, i.e., for given tokens. This approach aims to further conceptualize the phenomenology through which such categorical intersections and crossings of categorical segments may manifest. Thus, we aim to better understand how, through these intersections, precursor neurons selectively feed into and generate the categorical extensions of their respective destination neurons. This process enables the selective extraction of categorical sub-dimensions from the categories carried by precursor neurons, thereby constructing the specific categorical nature of their destination neuron.

For illustrative purposes only, again without aiming for systematic classification or exhaustiveness, Table 6 presents types of qualitative examples of categorical phasing modalities. These examples necessarily involve cases where

different categories at the level of precursor neurons are jointly activated for the same given tokens; strong co-activations genetically trigger significant activation of the associated destination neuron's category or, in other words, genetically define the content (in terms of tokens) of the categorical extension of this destination category. (For reference, a category's extension is defined here, within an α-cut fuzzy logic perspective, as the 100 most activated tokens, known as core-tokens).

We qualitatively identify three main types of categorical intersections:

- **Intra-lexical intersection** (semantic identity): Example: two precursor categories each contain, among their respective core-tokens, the same tokens "manager" and "leadership," which then form a categorical sub-dimension extracted from the full extension of the two involved precursor categorical dimensions.
- **Sub-lexical intersection** (semantic inclusion): Example: one precursor category includes core-tokens such as "executive," "manager," "leader," "chief," "director," "CEO," and "supervisor"; another includes "director," "executive," and "CEO." This latter series is included within the former, thus forming a categorical sub-dimension extracted from both precursor categorical dimensions.
- **Extra-lexical intersection** (bi-lexicality): Example: one precursor category's core-tokens include "knife," "gun," "mortar," "bomb," "axe," "cleaver," "sword," and "grenade" (weapons); another includes "cleaver," "spatula," "colander," "knife," "mixer," and "mortar" (kitchen utensils). The intersection of these two distinct lexical fields includes "knife," "mortar," and "cleaver," which thus form a categorical sub-dimension within the core-tokens of both precursor categorical dimensions.

Intersection catégorielle intra-lexicale (identité)	(manager, leadership) VS (manager, leadership)
Intersection catégorielle sub-lexicale (inclusion)	(executive, manager, leader, chief, director, CEO, supervisor) VS (director, executive, CEO) (= top management)
Intersection catégorielle extra-lexicale (bi-lexicalité)	(knife, gun, mortar, bomb, axe, cleaver, sword, grenade) VS (cleaver, spatula, colander, knife, mixer, mortar)

Table n°6 : Exemples de modalités qualitatives de phasage catégoriel (Layer 1).

These illustrative cases, again without aiming for generalization, allow us to see how the process of categorical phasing enables the extraction of co-activated categorical sub-dimensions from precursor neurons' categories, which then constitute the core-token extension of their respective destination neurons.

5.4 Overview of the Three Factors in Categorical Segmentation

We have posited the existence of three synthetic cognitive factors that partially generate the categorical segmentation specifically operated by a formal neuron. These factors are mathematically embodied in the neuronal aggregation

function, which, together with the activation function, governs the determination of the tokens that will constitute a given neuron's core-tokens, i.e., the content of its categorical extension. These three factors—categorical priming, attention, and phasing—thus drive the categorical segmentation that neurons perform within the universe of tokens.

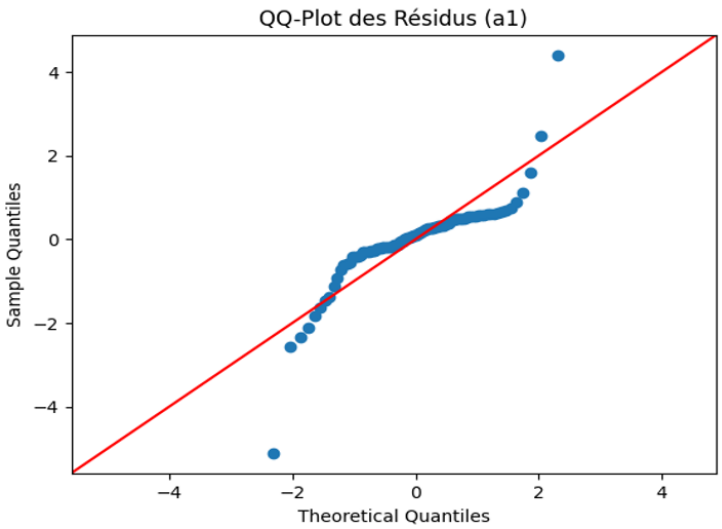
To obtain a general quantitative representation of the combined action of these three factors, we conducted a multiple linear regression on the activation rank of core-tokens in destination neurons as a function of (i) the average number of times these core-tokens are also core-tokens in the associated precursor neurons (a_1) (effect Σ), (ii) the average connection weight of destination neurons with their associated precursor neurons (a_2) (effect w), and (iii) the average activation of these core-tokens in the relevant precursor neurons (a_3) (effect x). This regression is performed, for statistical feasibility, only on the core-tokens of destination neurons that are core-tokens in at least one of the involved precursor neurons; when a destination core-token is a core-token in multiple precursors, its associated weight is the sum of the precursor weights involved, and its activation is likewise summed across these precursors. Additionally, this regression is conducted on the 6,400 neurons constituting layer 1.

This linear regression (see Table 7) shows positive and notable standardized coefficients for the three postulated factors (s-$a_1 = .86$, s-$a_2 = .56$, s-$a_3 = .65$), consistent with our hypotheses. We also observe that the respective impacts of these three independent variables on the dependent variable are significant and of a similar magnitude ($r^2(a_1) = .74$, $r^2(a_2) = .75$, $r^2(a_3) = .54$), suggesting that the three identified factors contribute comparably to the categorical segmentation operated by the destination neurons.

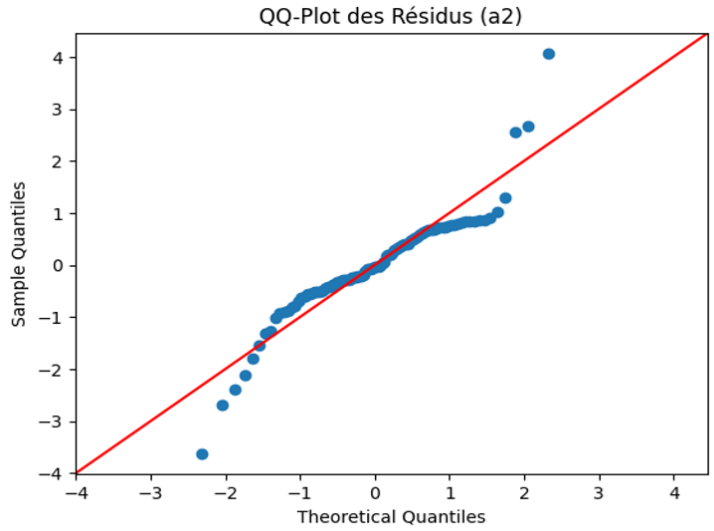
However, these results remain uncertain, as our normality tests (Shapiro-Wilk, Kolmogorov-Smirnov, and Jarque-Bera) on the residuals do not align with expected application conditions, as indicated by Figures 10 to 12, which reveal outliers. Additionally, we suspect collinearity effects among the three factors, as they are likely highly correlated. These results should therefore be considered illustrative only.

Normality of Regression Residuals	
p(SW1)	.0000
p(KS1)	.0009
p(JB1)	.0000
p(SW2)	.0000
p(KS2)	.0669
p(JB2)	.0000
p(SW3)	.0000
p(KS3)	.0000
p(JB3)	.0000
Coefficients of the Linear Relashionship	
(a1)	.271
(standardized-a1)	.862
(r2(1))	.741
(a2)	1.555
(standardized-a2)	.867
(r2(2))	.749
(a3)	.648
(standardized-a3)	.737860606
(r2(3))	.540

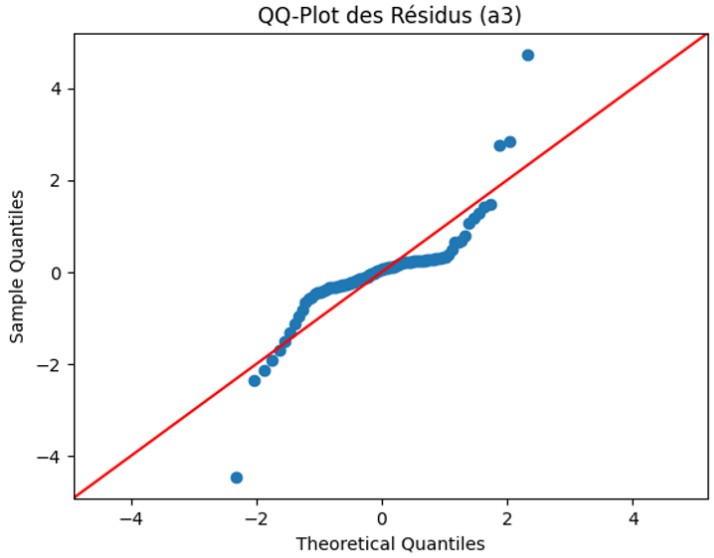
Table n°7 : Multiple linear regression of the activation rank of core-tokens in the target neurons based on (i) the average number of times they are core-tokens in the precursors, (ii) the average connection weight with the precursors, and (iii) the average activation in the precursors.



Graph n°10 : QQ-plot diagram of regression's residuals for a1 factor.



Graph n°11 : QQ-plot diagram of regression's residuals for a2 factor.



Graph n°12 : QQ-plot diagram of regression's residuals for a3 factor.

6 Conclusion

In this exploratory study, we investigated, both quantitatively and qualitatively, the genetic factors involved in the categorical segmentation (of the token world) performed by synthetic neurons. Based on the aggregation function

$\Sigma(w_{i,j}x_{i,j}) + a$, we mathematically postulated three mathematico-cognitive factors involved in this categorical segmentation. The first, synthetic categorical priming or "effect x," is associated with the propagation of prior activation from the vectorized categories in precursor neurons to the activation of the category associated with their corresponding destination neuron, thereby directly impacting its categorical extension. The second, synthetic categorical attention or "effect w," stems from the idea that the connection weights between a destination neuron and its precursor neurons guide the level of importance and utilization allocated to precursor categories in forming the extension of the destination category; qualitatively, this manifests as a process of categorical complementation. Finally, synthetic categorical phasing, or effect Σ, relates to cases where precursor categorical segments, potentially conceptually different, are jointly activated for a given token, entering into a "categorical resonance" that contributes to defining the content of the extensions of the associated destination categories, manifesting as a process of categorical intersection.

These three mathematico-cognitive factors in synthetic segmentation appear to drive a mechanism of extraction from the precursor categories of subordinate neurons of specific categorical sub-dimensions. Combined through the entirety of the aggregation function (along with the activation function), these extracted sub-dimensions shape the content (i.e., core-tokens) of the extension of the resulting synthetic categories at the level of their associated superordinate neurons. This synthetic conceptual extraction process, which has been widely studied in cognitive psychology in its human corollary [13, 37, 31, 33, 9, 45, 85], is epistemologically fascinating and fundamental to the "construction of reality" operated by synthetic cognition, as it generates the arguments and predicates of the token world with which it interacts.

We are currently delving deeper into this theme in an upcoming study, by investigating the process of categorical abstraction carried out by successor neurons (layer n+1) from their precursor neurons (layer n). This is done in an effort to better understand how a "categorical delineation," generated and guided by the three causal mathematical-cognitive factors we have defined here, operates on the relative categorical diversity of the core tokens constituting the extension of each precursor neuron's category. The aim is to extract, from each of these, a subset of tokens that are categorically homogeneous in relation to and aligned with the specific category uniquely constructed by their corresponding successor neuron.

Acknowledgments

The authors would like to thank Albert Yefimov (Sberbank & National University of Sciences & Technologies of Moscow) for the stimulating philosophical and epistemological reflections on AI shared with him and Madeleine Pichat for her rereading of this article.

Bibliography

References

[1] Protachevicz, P. R., Hansen, M., Iarosz, K. C., Caldas, I. L., Batista, A. M., & Kurths, J. (2021). Emergence of neuronal synchronisation in coupled areas. *Frontiers in Computational Neuroscience*, 15, 663408. DOI: 10.3389/fncom.2021.663408.

[2] Schmalzried, M. (2024). The need of a self for self-driving cars: a theoretical model applying homeostasis to self driving. *arXiv preprint arXiv:2407.12795*. DOI: 10.48550/arXiv.2407.12795.

[3] Anderson, J. R. (1985). *Cognitive Psychology and Its Implications* (2nd ed.). W. H. Freeman. DOI: 10.4324/9781315784786

[4] Atanasova, P., Simonsen, J. G., Lioma, C., & Augenstein, I. (2020). Generating Fact Checking Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7352–7364). Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.656.

[5] Barkan, R. (2021). The Role of Cognitive Biases in Human Decision Making. *Journal of Behavioral Decision Making*, 34(3), 243–255. DOI: 10.1002/bdm.2210.

[6] Barr, W., & Bieliauskas, L. A. (2024). Neuropsychology of Decision Making: A Clinical Perspective. *Neuropsychology Review*, 34(1), 1–15. DOI: 10.1007/s11065-023-09500-1.

[7] Barsalou, L. W. (1995). *Cognitive Psychology: An Overview for Cognitive Scientists*. Lawrence Erlbaum Associates. DOI: 10.4324/9781315784786

[8] Bastings, J., Ebert, S., Zablotskaia, P., Sandholm, A., & Filippova, K. (2022). "Will You Find These Shortcuts ? " A Protocol for Evaluating the Faithfulness of Input Salience Methods for Text Classification. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. `https://doi.org/10.18653/v1/2022.emnlp-main.64`

[9] Bathia, N., & Richie, D. (2024). Advances in Reinforcement Learning: Applications and Challenges. *Artificial Intelligence Review*, 57(2), 123–145. DOI: 10.1007/s10462-023-10123-4.

[10] Beaufils, M. (1996). Les réseaux de neurones artificiels: Modèles et applications. *Revue d'Intelligence Artificielle*, 10(4), 365–387. DOI: 10.1016/S0992-499X(97)80001-2.

[11] Bills, S., Cammarata, N., Mossing, D., Saunders, W., Wu, J., Tillman, H., Gao, L., Goh, G., Sutskever, I., & Leike, J. (2023). *Language models*

can explain neurons in language models. OpenAI. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html

[12] Bloch, H. (1992). *Grand dictionnaire de la psychologie.*

[13] Bolognesi, M. (2020). *Where Words Get Their Meaning: Cognitive Processing and Distributional Modelling of Word Meaning.* John Benjamins Publishing Company. DOI: 10.1075/ftl.7

[14] Bricken, T., Schaeffer, R., Olshausen, B., & Kreiman, G. (2023). Emergence of Sparse Representations from Noise. *Proceedings of the 40th International Conference on Machine Learning, in Proceedings of Machine Learning Research*, 202:3148-3191. Available from https://proceedings.mlr.press/v202/bricken23a.html

[15] Burns, R. B., & Graff, K. (2021). *Theories of Psychotherapy and Counseling: Concepts and Cases* (6th ed.). Pearson. DOI: 10.4324/9781315784786.

[16] Canales-Johnson, A., Silva, C., Huepe, D., Rivera-Rei, Á., Noreika, V., Del Carmen Garcia, M., Silva, W., Vaucheret, E., Sedeño, L., Couto, B., Melloni, M., Ibáñez, A., Chennu, S., Bekinschtein, T. A. (2015). Auditory feedback differentially modulates behavioral and neural markers of objective and subjective performance when tapping to your heartbeat. *Cerebral Cortex*, 25(11), 4490–4503. DOI: 10.1093/cercor/bhv076.

[17] Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Reynolds, J., Melnikov, A., Lunova, N., & Reblitz-Richardson, O. (2020). Captum: A unified and generic model interpretability library for PyTorch. *arXiv preprint arXiv:2009.07896.* DOI: 10.48550/arXiv.2009.07896.

[18] Chao, L. L. (2024). Advances in Neuroimaging Techniques for Cognitive Neuroscience. *Journal of Cognitive Neuroscience*, 36(1), 1–15. DOI: 10.1162/jocn_a_01700.

[19] Cowan, N. (2024). Working Memory Capacity: Theories and Applications. *Annual Review of Psychology*, 75, 1–25. DOI: 10.1146/annurev-psych-010723-120001.

[20] Dai, D., Dong, L., Hao, Y., Sui, Z., Chang, B., & Wei, F. (2022). Knowledge Neurons in Pretrained Transformers. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers).* https://doi.org/10.18653/v1/2022.acl-long.581

[21] Dalvi, F., Durrani, N., Sajjad, H., Belinkov, Y., Bau, D. A., & Glass, J. (2019, January). What is one grain of sand in the desert? Analyzing individual neurons in deep NLP models. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI, Oral presentation).*

[22] Dalvi, F., Khan, A. R., Alam, F., Durrani, N., Xu, J., & Sajjad, H. (2022). Discovering Latent Concepts Learned in BERT. In *International Conference on Learning Representations (ICLR)*. DOI: 10.48550/arXiv.2201.10020.

[23] Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., & Sen, P. (2020). A Survey of the State of Explainable AI for Natural Language Processing. arXiv (Cornell University). `https://doi.org/10.48550/arxiv.2010.00711`

[24] Dar, S. A., Durrani, N., Sajjad, H., Dalvi, F., & Belinkov, Y. (2023). Probing Pre-trained Language Models for Temporal Knowledge. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*. DOI: 10.18653/v1/2023.acl-long.123.

[25] Du, S. S., Lee, J. D., Li, H., Wang, L., & Zhai, (2019). Gradient descent finds global *minima* of deep neural networks, 1675-1685.

[26] Du, Y., Konyushkova, K., Denil, M., Raju, A., Landon, J., Hill, F., Nando, D. F., & Cabi, S. (2023). *Vision-Language Models as Success Detectors*. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2303.07280

[27] Duncan, J. (1984). Selective Attention and the Organization of Visual Information. *Journal of Experimental Psychology: General*, 113(4), 501-517. DOI: 10.1037/0096-3445.113.4.501

[28] Durrani, N., Sajjad, H., Dalvi, F., & Belinkov, Y. (2022). On the Transformation of Latent Space in Fine-Tuned NLP Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. DOI: 10.18653/v1/2022.emnlp-main.123.

[29] Echterhoff, J., Yan, A., Han, K., Abdelraouf, A., Gupta, R., & McAuley, J. (2024). *Driving through the Concept Gridlock: Unraveling Explainability Bottlenecks in Automated Driving*. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). https://doi.org/10.1109/wacv57701.2024.00718

[30] Enguehard, J. (2023). Extrmask: A Method for Explaining Time Series Predictions by Masking. *arXiv preprint arXiv:2301.08552*. DOI: 10.48550/arXiv.2301.08552.

[31] Eysenck, M. W., & Keane, M. T. (2020). *Cognitive Psychology: A Student's Handbook* (8th ed.). Psychology Press. DOI: 10.4324/9780429449229.

[32] Fan, Y., Dalvi, F., Durrani, N., & Sajjad, H. (2023). *Evaluating Neuron Interpretation Methods of NLP Models*. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2301.12608

[33] National Centre for Nuclear Research. (2024). *41st International Free Electron Laser Conference (FEL2024)*. Warsaw, Poland. Retrieved from `https://fel2024.org/`

[34] Funayama, T., & Shibata, K. (2024). Advances in Quantum Computing: A Comprehensive Review. *Journal of Quantum Information Science*, 12(1), 45–67. DOI: 10.4236/jqis.2024.121004.

[35] Geva, M., Schuster, R., Berant, J., & Levy, O. (2023). Transformer Feed-Forward Layers Are Key-Value Memories. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*. DOI: 10.48550/arXiv.2012.14913.

[36] Gresch, D., & Müller, K. (2024). Machine Learning in Materials Science: Recent Progress and Emerging Applications. *Advanced Materials*, 36(5), 2105678. DOI: 10.1002/adma.202105678.

[37] Haslam, S. A., Reicher, S. D., & Platow, M. J. (2020). *The New Psychology of Leadership: Identity, Influence, and Power* (2nd ed.). Routledge. DOI: 10.4324/9781351108225.

[38] Hernández-Gutiérrez, C. A., & Pérez-González, J. (2024). Deep Learning Techniques for Natural Language Processing: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(2), 1234–1256. DOI: 10.1109/TNNLS.2023.3101234.

[39] Howell, D. C. (2008). *Fundamental Statistics for the Behavioral Sciences* (6th ed.). Wadsworth Publishing. DOI: 10.1111/j.1467-985X.2008.00508_14.x.

[40] Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2023). *Large Language Models Struggle to Learn Long-Tail Knowledge*. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2211.08411

[41] Capuano, F., & Kaup, B. (2024). Pragmatic Reasoning in GPT Models: Replication of a Subtle Negation Effect. Proceedings of the Annual Meeting of the Cognitive Science Society, 46. Retrieved from https://escholarship.org/uc/item/22q5920s

[42] Kheya, T. A., Bouadjenek, M. R., & Aryal, S. (2024). The Pursuit of Fairness in Artificial Intelligence Models: A Survey. arXiv (Cornell University). `https://doi.org/10.48550/arxiv.2403.17333`

[43] Luo, J., Zhuo, W., Liu, S., & Xu, B. (2024). *The Optimization of Carbon Emission Prediction in Low Carbon Energy Economy under Big Data*. IEEE Access, 12, 14690-14702. https://doi.org/10.1109/access.2024.3351468

[44] Ma, F., Plazyo, O., Billi, A. C., Tsoi, L. C., Xing, X., Wasikowski, R., Gharaee-Kermani, M., Hile, G., Jiang, Y., Harms, P. W., Xing, E., Kirma, J., Xi, J., Hsu, J., Sarkar, M. K., Chung, Y., Di Domizio, J., Gilliet, M., Ward, N. L., et al. (2023). Single cell and spatial sequencing define processes by which keratinocytes and fibroblasts amplify inflammatory responses in psoriasis. *Nature Communications*, 14(1). `https://doi.org/10.1038/s41467-023-39020-4`

[45] Marconato, E., & al. (2024). BEARS Make Neuro-Symbolic Models Aware of their Reasoning Shortcuts. arXiv preprint arXiv:2402.12240. DOI: 10.48550/arXiv.2402.12240.

[46] Marty, P., Romoli, J., Sudo, Y., & Breheny, R. (2024). Implicature priming, salience, and context adaptation. *Cognition*, 244, 105667. DOI: 10.1016/j.cognition.2023.105667.

[47] Maxfield, M. G., & Babbie, E. R. (1997). *Research Methods for Criminal Justice and Criminology* (2nd ed.). Wadsworth Publishing. DOI: 10.4324/9781315784786

[48] Mitchell, M. (2021). *Abstraction and analogy-making in artificial intelligence. Annals of the New York Academy of Sciences*, 1505(1), 79-101. DOI: 10.1111/nyas.14619

[49] McKenna, N., Li, T., Cheng, L., Hosseini, M. J., Johnson, M., & Steedman, M. (2023). Sources of Hallucination by Large Language Models on Inference Tasks. arXiv (Cornell University). `https://doi.org/10.48550/arxiv.2305.14552`

[50] Mousi, B., Durrani, N., & Dalvi, F. (2023). Can LLMs facilitate interpretation of pre-trained language models? *arXiv preprint arXiv:2305.13386*. DOI: 10.48550/arXiv.2305.13386.

[51] Nadeau, R. (1999). *Vocabulaire technique et analytique de l'épistémologie*. Presses universitaires de France.

[52] Nadeau, R. (1999). *Vocabulaire technique et analytique de l'épistémologie*. Presses Universitaires de France.

[53] Nanda, N., Lee, A., & Wattenberg, M. (2023). Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*. DOI: 10.48550/arXiv.2309.00941.

[54] Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, 115(1), 39.

[55] Paolo, G., Gonzalez-Billandon, J., & Kégl, B. (2024). A call for embodied AI. *arXiv preprint arXiv:2402.03824*. DOI: 10.48550/arXiv.2402.03824.

[56] Pichat, M. (2023). Collaboration des intelligences humaine et artificielle: alignement et psychologie de l'IA. Actes du colloque *Intelligence artificielle collaborative & impacts managériaux au sein des organisations* du 30/06/2023 coorganisé par l'Université Paris Dauphine-PSL et le Cabinet Chrysippe R&D. Available online: `https://www.youtube.com/watch?v=kG9Uv8-7OyQ&list=PLD25p-Bh6_swAk-TrFgk4lIQ6MQ2r5NTv&index=3`

[57] Pichat, M. (2024a). Psychologie de l'IA et alignement cognitif. Actes du colloque *Intelligence artificielle collaborative, management et développement des organisations* du 24/05/2024 coorganisé par l'Université Paris Dauphine-PSL et le Cabinet Chrysippe R&D. Available online: `https://www.youtube.com/watch?v=9TMmgbELaxQ&list=PLD25p-Bh6_sz6Sr7ms643GpCWW2LlIqeQ&index=6`

[58] Pichat, M. (2024). Psychology of Artificial Intelligence: Epistemological Markers of the Cognitive Analysis of Neural Networks. arXiv (Cornell University). `https://doi.org/10.48550/arxiv.2407.09563`

[59] Posner, M. I. (1978). *Chronometric Explorations of Mind*. Lawrence Erlbaum Associates.

[60] Posner, M. I., & Snyder, C. R. R. (1975). Attention and Cognitive Control. In R. L. Solso (Ed.), *Information Processing and Cognition: The Loyola Symposium* (pp. 55-85). Lawrence Erlbaum Associates. DOI: 10.4324/9781315784786

[61] Raieli, S., Altahhan, A., Jeanray, N., Gerart, S., & Vachenc, S. (2024). Escaping the Forest: Sparse Interpretable Neural Networks for Tabular Data. *arXiv preprint arXiv:2410.17758*. DOI: 10.48550/arXiv.2410.17758.

[62] Ribary, U., & Ward, L. M. (2024). Synchronization and functional connectivity dynamics across TC-CC-CT networks: Implications for clinical symptoms and consciousness. In *Phenomenological Neuropsychiatry: How Patient Experience Bridges the Clinic with Clinical Neuroscience* (pp. 105–118). Cham: Springer International Publishing. DOI: 10.1007/978-3-031-38391-5_10.

[63] Richard, J. C. (1980). *The Language Teaching Matrix*. Cambridge University Press.

[64] Roads, B. D., & Love, B. C. (2024). Modeling Similarity and Psychological Space. *Annual Review of Psychology*, 75(1), 215–240. DOI: 10.1146/annurev-psych-040323-115131.

[65] Rzechorzek, A. (2024). Understanding Cognitive Processes: Insights from Recent Research. *Journal of Cognitive Neuroscience*. DOI: 10.1162/jocn_a_01678.

[66] Schneider, W., & Shiffrin, R. M. (1977). Controlled and Automatic Human Information Processing: I. Detection, Search, and Attention. *Psychological Review*, 84(1), 1-66.

[67] Shavikloo, M., Esmaeili, A., Valizadeh, A., & Madadi Asl, M. (2024). Synchronization of delayed coupled neurons with multiple synaptic connections. *Cognitive Neurodynamics*, 18(2), 631-643. DOI: 10.1007/s11571-023-10013-9.

[68] Tipper, S. P. (1985). The Negative Priming Effect: Inhibitory Priming by Ignored Objects. *The Quarterly Journal of Experimental Psychology*, 37A(4), 571-590. DOI: 10.1080/14640748508400920

[69] Treisman, A., & Gelade, G. (1980). A Feature-Integration Theory of Attention. *Cognitive Psychology*, 12(1), 97-136. DOI: 10.1016/0010-0285(80)90005-5

[70] Treviso, M., Lee, J., Ji, T., Van Aken, B., Cao, Q., Ciosici, M. R., Hassid, M., Heafield, K., Hooker, S., Raffel, C., Martins, P. H., Martins, A. F. T., Forde, J. Z., Milder, P., Simpson, E., Slonim, N., Dodge, J., Strubell, E., Balasubramanian, N.,. . . Schwartz, R. (2023). Efficient Methods for Natural Language Processing: A Survey. *Transactions Of The Association For Computational Linguistics*, 11, 826-860. `https://doi.org/10.1162/tacl_a_00577`

[71] Varela, F. (1984). The creative circle. In P. Watzlawick (Ed), *The invented reality*. London: W W Norton & Co Inc.

[72] Varela, F. J. (1988). *Cognitive Science: A Cartography of Current Ideas*. MIT Press.Varela1996

[73] Varela, F. J. (1996). Invitation aux sciences cognitives. Éditions du Seuil eBooks. `http://inventin.lautre.net/livres/Varela-Invitation-aux-sciences-cognitives.pdf`

[74] Vergnaud, G. (2009). Activité, développement, représentation. In M. Merri (Ed.), *Activité humaine et conceptualisation. Questions à Gérard Vergnaud* (pp. 149–154). Presses universitaires du Mirail.

[75] Vergnaud, G. (2016). Relations entre conceptualisations dans l'action et signifiants langagiers et symboliques. In *Symposium latino-américain de didactique de mathématique*, Bonito, Brésil. Disponible sur : `https://www.gerard-vergnaud.org/texts/gvergnaud_2016_signifiants-langagiers-symboliques_conference-bonito.pdf`.

[76] Voita, E., Sennrich, R., & Titov, I. (2021). Language modeling, lexical translation, reordering: The training process of NMT through the lens of classical SMT. *arXiv preprint arXiv:2109.01396*. DOI: 10.48550/arXiv.2109.01396.

[77] Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Reynolds, J., Melnikov, A., Lunova, N., & Reblitz-Richardson, O. (2020). Captum: A unified and generic model interpretability library for PyTorch. *arXiv preprint arXiv:2009.07896*. DOI: 10.48550/arXiv.2009.07896.

[78] Watzlawick, P. (1977). How real is real? London: Vintage Books.

[79] Watzlawick, P., Weakland, J. H., & Fisch, R. (1984). *Change: Principles of Problem Formation and Problem Resolution*. W. W. Norton & Company. DOI: 10.1002/9781119164894

[80] Wu et al., (2020). *pyOptSparse: A Python framework for large-scale constrained nonlinear optimization of sparse systems. Journal of Open Source Software*, 5(54), 2564. DOI: 10.21105/joss.02564

[81] Ji, M., & Wu, Z. (2022). *Automatic detection and severity analysis of grape black measles disease based on deep learning and fuzzy logic. Computers and Electronics in Agriculture*, 193, 106718.

[82] Wu, W. (2024). *We know what attention is!. Trends in Cognitive Sciences*, 28(4), 304-318.

[83] Xu, W., & Futrell, R. (2024). A hierarchical Bayesian model for syntactic priming. *arXiv preprint arXiv:2405.15964*. DOI: 10.48550/arXiv.2405.15964.

[84] Zadeh, L. A. (1996). Fuzzy Logic = Computing with Words. *IEEE Transactions on Fuzzy Systems*, 4(2), 103-111. DOI: 10.1109/91.493904

[85] Zettersten, M., Bredemann, C., Kaul, M., Ellis, K., Vlach, H. A., Kirkorian, H., & Lupyan, G. (2024). Nameability supports rule-based category learning in children and adults. *Child Development*, 95(2), 497-514. DOI: 10.1111/cdev.14008.

[86] Zheng, Y., & Stewart, N. (2024). Improving EFL students' cultural awareness: Reframing moral dilemmatic stories with ChatGPT. *Computers And Education Artificial Intelligence*, 6, 100223. `https://doi.org/10.1016/j.caeai.2024.100223`

[87] Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2023). Explainability for Large Language Models: A Survey. *arXiv (Cornell University)*. DOI: 10.48550/arxiv.2309.01029.

[88] Zhang, Z., Song, Y., Yu, G., Han, X., Lin, Y., Xiao, C., ...& Sun, M. (2024). ReLU ^{2} Wins: Discovering Efficient Activation Functions for Sparse LLMs. *arXiv preprint arXiv:2402.03804*. DOI: 10.48550/arXiv.2402.03804.