Attention intra-neuronale au sein des modèles de langage

Rapports entre activation et sémantique

Michael Pichat[1,2,4], William Pogrund[1,5], Paloma Pichat[1,3], Armanouche Gasparian[1], Samuel Demarchi[1,4], Martin Corbet[1,2], Alois Georgeon[1,2], Michael Veillet-Guillem[1]

[1]Neocognition (Chrysippe R&D)
[2]Facultés Libres de Philosophie et de Psychologie de Paris (ER IPC)
[3]Faculté de Médecine de Lyon Est (Université Lyon 1)
[4]Université Paris 8
[5]INP-PHELMA, Université Grenoble Alpes

Résumé

Cette étude s'intéresse à la capacité des neurones de type perceptron des modèles de langage à opérer une attention intra-neuronale ; c'est-à-dire à repérer différents segments catégoriels homogènes au sein de la catégorie de pensée synthétique qu'ils portent, sur la base d'une segmentation de zones d'activation particulières des tokens pour lesquels ils s'activent particulièrement. L'objectif de ce travail est dès lors de déterminer dans quelle mesure les neurones formels peuvent établir une relation homomorphique entre segmentations activationnelles et catégorielles. Les résultats suggèrent l'existence d'une telle relation, ténue, au seul niveau des tokens dotés de niveaux d'activation très élevés. Cette attention intra-neuronale permet ensuite des processus de restructuration catégorielle au niveau des neurones de la couche suivante, contribuant ainsi à la formation progressive d'abstractions catégorielles de haut niveau.

1 Contexte théorique

1.1 Têtes attentionnelles

Avant l'essor des transformers, les modèles de traitement automatique du langage s'appuyaient principalement sur des architectures récurrentes (RNN, LSTM) et convolutionnelles (CNN). Les RNN et LSTM, bien que capables de capturer des dépendances à long terme grâce à leurs mécanismes de mémoire, souffrent du problème du gradient évanescent, rendant difficile l'apprentissage

de relations sur de longues séquences. De plus, leur nature séquentielle limite leur parallélisation et ralentit leur entraînement. Les CNN, en revanche, offrent une meilleure parallélisation mais sont mal adaptés aux dépendances globales, car leur champ réceptif est restreint et dépend de la profondeur du réseau. Ces approches souffraient ainsi de limitations dans la capture des dépendances à longue portée et de difficultés de parallélisation [124, 125]. es limitations ont motivé l'introduction des transformers et du mécanisme de self-attention, qui surmonte ces contraintes en permettant un traitement parallèle efficace tout en capturant des relations à longue portée. L'introduction du mécanisme de self-attention par Vaswani et al.[126] a permis de surmonter les contraintes mentionnant, marquant un tournant majeur dans le domaine du NLP.

Les transformers exploitent la self-attention, où chaque élément d'une séquence pondère l'importance des autres éléments, facilitant ainsi la modélisation des dépendances complexes. L'attention multi-tête enrichit cette approche en permettant à différentes têtes d'attention de se spécialiser dans divers aspects de la représentation. Devlin et al., [127] ont par exemple montré que certaines têtes de BERT capturent des relations syntaxiques, comme les liens entre sujets et verbes, tandis que d'autres se concentrent sur des relations sémantiques plus globales. Radford et al., [128] ont manifesté que l'utilisation de multiples têtes attentionnelles dans GPT permet de mieux modéliser le contexte des phrases en capturant des informations distribuées sur différentes positions de la séquence d'entrée. Cette capacité permet aux transformers d'améliorer la richesse et la hiérarchie des représentations, offrant une meilleure généralisation sur des tâches variées.

Au sein d'un transformer, chaque couche d'attention est composée de plusieurs têtes, chacune effectuant une opération d'attention sur une projection linéaire des entrées. La self-attention repose sur la projection des représentations en trois ensembles de vecteurs : les requêtes (Q), les clés (K) et les valeurs (V). Pour une séquence d'entrée $X \in \mathbb{R}^{T \times d}$, où T est la longueur de la séquence et d la dimension des vecteurs, ces matrices sont définies par :

$$Q = XW^T, \quad K = XW^T, \; V = XW^T$$

où $W^T \in \mathbb{R}^{d \times b_k}$ sont des matrices de poids apprises, etd_k est la dimension des clés et des requêtes. Le calcul de l'attention est donné par la scaled dot-product attention :

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

La normalisation par$\sqrt{d_k}$ stabilise l'apprentissage en évitant une explosion des valeurs des produits scalaires[126]. ML'attention multi-tête applique ces opérations en parallèle sur h projections différentes de Q, K, et V :

$$\text{head}_i = \text{Attention}(QW_i^G, KW_i^K, VW_i^V)$$

La concaténation des têtes suivie d'une projection linéaire produit la sortie finale :

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

Cette architecture permet d'extraire des informations contextuelles sous divers angles et d'améliorer la capture de relations complexes dans les données.

L'étude des mécanismes attentionnels a révélé que certaines têtes sont spécialisées tandis que d'autres sont redondantes. [129] ont montré que certaines têtes capturent des relations syntaxiques précises, tandis que d'autres se concentrent sur des relations sémantiques globales. Voita [130], et Mitchell [131] ont observé que la suppression de plusieurs têtes n'affecte pas significativement la performance du modèle, suggérant des mécanismes de compensation entre les têtes restantes.

Des approches issues de la mécanique statistique ont permis d'analyser l'interaction entre les chemins d'attention. Tiberi et al.[132] ont modélisé la contribution des têtes d'attention via une décomposition en noyaux :

$$K = \sum_{i=1}^{h} K_i$$

Chaque noyau K_i est associé à une tête spécifique, permettant d'évaluer leur rôle dans la représentation finale du modèle. Les résultats montrent que certaines têtes jouent un rôle structurant, tandis que d'autres peuvent être éliminées sans impact significatif. Cette observation ouvre la voie à des optimisations des architectures des transformer, par réduction du nombre de têtes redondantes et amélioration de l'interprétabilité du modèle.

L'efficacité computationnelle des transformers a fait l'objet de nombreuses recherches. Les Sparse Transformers [23] réduisent la complexité de l'attention à $O(n \log n)$ en introduisant une structure d'attention clairsemée. Reformer [133]optimise la gestion mémoire grâce à une factorisation des clés-valeurs et une attention locale. Performer [134] emplace l'attention classique par une approximation linéaire, réduisant la complexité à $O(n)$. Cette approche repose sur les projections aléatoires des clés et requêtes dans un espace de dimension réduite, où les produits scalaires sont calculés de manière approximative à l'aide des noyaux favorisant une factorisation efficace. Cela permet d'éviter le calcul coûteux des produits matriciels denses en maintenant une précision élevée, rendant l'attention scalable même pour de longues séquences. Longformer et BigBird [135] combinent des attentions locales et globales pour traiter efficacement de longues séquences.

D'autres travaux ont analysé la spécialisation des têtes d'attention dans des contextes spécifiques. Clark et al. [136] ont étudié les matrices d'attention de BERT et observé que certaines têtes apprennent des relations syntaxiques spécifiques, telles que la dépendance sujet-verbe ou les relations anaphoriques. Transformer-XL [137] a introduit un mécanisme de mémoire récurrente permettant de capturer des dépendances à plus long terme, améliorant la génération de texte et le dialogue.

Enfin, les mécanismes attentionnels se sont étendus à d'autres domaines, notamment la vision par ordinateur avec Vision Transformer (ViT) [138] et Swin

Transformer [139], ainsi qu'aux neurosciences et à la modélisation des processus cognitifs [140].

1.2 L'attention humaine

Biologiquement, l'attention humaine découle de l'aptitude restreinte du système neuronal à traiter l'information. Elle se manifeste par des sélections dans l'acquisition, l'activation et l'utilisation des données sensorielles ou mémorielles (telles les connaissances ou règles) [41, 4]. Ceci se traduit par une réponse orientée, focalisant la recherche d'informations sur quelques caractéristiques précises. On peut évoquer ici les recherches neurocognitives sur les mécanismes attentionnels, inspirées notamment du travail de Posner[81, 82, 87]. Cette approche met en avant l'existence d'un système frontal d'attention associé à l'attention consciente et la planification, et d'un système postérieur, localisé dans le lobe pariétal, impliqué quant à lui dans les processus visuo-spatiaux et le changement de focalisation attentionnelle.

En psychologie cognitive, l'attention est conceptualisée comme le calibrage d'une activité vers un but précis, augmentant ainsi l'efficacité dans la dynamique de recueil et d'exécution d'informations (sélectivité, précision, rapidité) pour une tâche donnée [77, 90, 76, 84, 148, 30, 91, 26, 108, 45]. Lors de l'exécution de tâches, l'attention est gérée par le système nerveux central, déterminant l'importance de certaines informations internes (comme des connaissances ou les schémas) et garantes de la qualité de l'exécution.

Généralement, deux fonctions cognitives sont associées à l'attention [31] :

- La détection de signaux, qui repose majoritairement sur la vigilance et l'exploration pour identifier l'apparition d'un stimulus particulier.
- L'attention sélective, focalisée sur des stimuli spécifiques en excluant d'autres.

La vigilance désigne la capacité à se concentrer sur un flux d'informations pour repérer un signal précis [56], lequel pourrait apparaître rarement mais nécessite une réaction rapide [21, 46]. Elle est négativement affectée par le niveau d'incertitude des éléments ciblés [17]. La vigilance peut être définie comme un faisceau d'attention ajustable, influencé par l'anticipation de l'apparition du signal dans un emplacement précis [78, 60, 62].

L'exploration visuelle [111, 108], quant à elle, est une quête active de stimuli, contrastant avec une attente passive de leur émergence [80]. Elle se caractérise par une stratégie de reconnaissance par balayage d'attributs pour les situer dans un environnement donné. Selon la théorie d'intégration des attributs *Feature Integration Theory* [95, 86], une carte mentale pour chaque attribut visuel permet une représentation des occurrences dans le champ de vision, régulièrement inspectées parallèlement. Les processus d'attention jouent ici un rôle de liaison mentale, rassemblant divers attributs d'un même objet et inhibant les caractéristiques non pertinentes. L'approche de similitude [32] a en ce qui la concerne, analyse l'exploration attentionnelle comme une évaluation de la proximité entre les stimuli cibles et les distractions. Tandis que l'approche

de l'inspection guidée [20, 1] divise l'attention exploratoire en deux phases : d'abord l'activation d'une représentation globale des cibles potentielles, suivie d'une analyse sérielle pour identifier la cible la plus activée.

L'attention sélective est souvent explorée à travers l'effet « cocktail party » [22, 55, 11], rqui se rapporte à la capacité de suivre une conversation parmi d'autres environnantes. Les caractéristiques de cette focalisation attentionnelle incluentsuggère qu'un filtre sélectionne parmi des flux sensoriels ceux qui recevront spécifiquement un traitement approfondi. Cependant, ce modèle a évolué en une approche atténuatrice [93, 53], où toutes les informations sont réduites en intensité perceptible, laissant résiduelles seulement celles qui sont proches de critères ciblés. La répartition des ressources attentionnelles limitées[52] e rapporte également à la gestion parallèle d'activités avec efficacité accrue.

Les aspects des mécanismes attentionnels qui sont pertinents pour nous, dans le cadre de notre présente étude, sont positionnés à l'intersection des activités de vigilance et d'attention sélective. En effet, ces activités, respectivement de détection d'un type ciblé d'information et de focalisation attentionnelle élective sur certaines caractéristiques de données, relèvent du phénomène qui nous intéresse ici : l'impact du niveau d'activation d'un neurone face à un token entrant sur la détection et la sélection attentionnelles intra-neuronale de tokens présentant certaines caractéristiques catégorielles spécifiques.

1.3 Processus attentionnels et conceptualisation

La notion de conceptualisation [1] est un des apports majeurs de la théorie des champs conceptuels proposée par Vergnaud [100, 101]. La conceptualisation, nous allons le voir dans ce qui suit, est un processus cognitif central à l'intersection des mécanismes de vigilance et d'attention sélective mentionnés précédemment, dans le cadre spécifique des questions de catégorisation des informations reçues par un système cognitif ; cette notion, développée dans le domaine de la pensée humaine, est dès lors particulièrement pertinente, de façon heuristique, en ce qui concerne notre présente investigation relative à l'identification et à la sélection attentionnelles de données particulières (tokens) au niveau du traitement interne réalisé par les neurones formels.

La conceptualisation est une activité représentationnelle attentionnelle dont la finalité est l'identification de caractéristiques opératoires des stimuli (pour nous des tokens) auxquels elle s'applique ; ceci, afin de fonder l'activité d'un système cognitif sur ces caractéristiques et dès lors de la rendre efficace. La conceptualisation a ainsi pour fonction cognitive d'extraire une forme opératoire

1. Classiquement, la recherche portant sur les concepts aborde ces derniers comme des unités cognitives rattachées en mémoire à un mot ou à une expression et auquel peut être associée une classe d'objets possédant des propriétés communes. Ces approches ont débouché, d'un point de vue expérimental, sur des études de jugement de familiarité ou de typicalité, de comparaison de catégories, de classification et plus généralement d'identification catégorielle ; études que nous avons synthétisées précédemment [67, 68]. Cependant, le terme de conceptualisation, tel qu'il est mobilisé par Vergnaud, ne renvoie pas à ces approches classiques du concept qui se centrent sur sa forme prédicative mais, à un positionnement développemental et pragmatique. Vergnaud met alors en avant la forme première du concept : sa forme opératoire

de connaissance, objet d'une focalisation attentionnelle spécifique ; cette dernière est, dans le champ de la cognition humaine, largement non consciente et non verbalisable et, pour cette raison, Vergnaud [100] la qualifie de connaissance-en-acte.

Vergnaud [101] définit un concept-en-acte, fruit de l'activité attentionnelle sélective de conceptualisation, comme une catégorie de pensée apprise comme étant pertinente relativement à une tâche (pour nous, le traitement finalisé de données textuelles). Relativement à cette notion de concept-en-acte, précisons trois points :

1. Les concepts-en-acte sont des catégories de pensée à travers lesquelles un système cognitif identifie, sélectionne et capture certaines informations présentes dans une situation à laquelle il est confronté (un ensemble de tokens dans le cas du traitement d'éléments langagiers). Autrement dit, les concepts-en-acte sont des filtres cognitifs attentionnels grâce auxquels une situation donnée est électivement « lue » ou « perçue ».
2. D'un point de vue épistémologique, il existe potentiellement une infinité de types formels de catégories de pensée. Les types les plus fréquemment rencontrés sont les suivants : objet, propriété, relation, transformation, condition, processus.
3. Les concepts-en-acte sont des vecteurs pragmatiques de la pensée, ici synthétique, qui organisent le traitement attentionnel de l'information en découpant le monde des tokens en fonction des seuls buts contingents de l'activité finalisée pour laquelle un modèle de langage a été entraîné. En effet, la fonctionnalité des concepts-en-acte réside dans le fait qu'ils permettent au système neuronal de focaliser son attention sur un nombre limité d'éléments sélectionnés et appris comme importants pour la réussite de l'activité de ce système synthétique. A ce titre, ils sous-tendent une représentation des seules variables de situation dont la prise en compte est centrale pour l'effectivité de l'activité.

The approach developed by Vergnaud [101] nstaure la conceptualisation comme une activité cognitive attentionnelle fondamentalement économique et pragmatique. Cette finalisation proprement pragmatique de la conceptualisation est à l'origine de la caractéristique des concepts qu'elle extrait d'être en acte, c'est-à-dire d'être encapsulés dans l'activité du système cognitif.

Un neurone artificiel peut être décrit comme un opérateur cognitif synthétique de conceptualisation, dont la finalité est de choisir, à partir d'un ensemble de tokens entrants, un (ou des) sous-groupe(s) particulier(s) de tokens ; sous-groupe(s) qui va (vont) être constitutif(s) de l'extension catégorielle du (des) concept(s)-en-acte « critique(s) » que ce neurone a pour fonction d'identifier électivement ; ce neurone réalisant dès lors une activité de focalisation attentionnelle sur certains types de tokens qu'il s'agit de sélectionner et de filtrer afin de rendre efficace l'activité de traitement langagier à laquelle participe ce neurone.

2 Problématique

2.1 Facteurs mathématico-cognitifs de la segmentation catégorielle et attention intra-neuronale

Dans une étude précédente [69], nous avons exploré les facteurs mathématico-cognitifs qui impactent la manière dont un réseau de neurones artificiels (de type perceptron) d'un modèle de langage effectue une segmentation catégorielle des tokens qui lui sont présentés. En nous fondant sur la fonction d'agrégation neuronale, de la forme $\sum(w_{i,j}\ x_{i,j})+b$, qui porte pour partie ce processus cognitif, nous avons identifié trois facteurs participant à ce découpage conceptuel.

- Le premier facteur est l'effet x l'amorçage catégoriel synthétique. Celui-ci se réfère à l'influence de l'activation des catégories de pensée synthétiques des neurones d'une couche n sur l'activation des catégories des neurones fortement connectés attentionnellement de la couche suivante. En d'autres termes, un token appartenant fortement à une catégorie initiale dans une couche n a plus de chances d'appartenir à une catégorie fortement associée de la couche $n+1$.
- Le deuxième facteur est l'effet w ou attention inter-catégorielle synthétique. Celui-ci influence l'importance qu'un neurone d'arrivée (en couche $n+1$) accorde aux catégories des neurones antérieurs (couche n), en fonction des poids de connexion. Ce processus se traduit par une complémentation catégorielle, où chaque catégorie précurseure focalisée attentionnellement apporte une sous-dimension catégorielle propre (constituée donc de tokens très spécifiques) à la constitution de la catégorie d'arrivée. Ainsi, la catégorie d'un neurone d'arrivée est construite par assemblage de sous-dimensions catégorielles complémentaires issues de ses catégories antérieures.
- Le troisième facteur, l'effet $\sum$ est le phasage catégoriel synthétique. Il relève du fait que plus des tokens sont conjointement activés au sein de catégories précurseures (couche n)et plus ils tendent à être constitutifs de l'extension de leur catégorie fortement associée attentionnellement en couche$n+1$. Cet effet se traduit par une intersection catégorielle.

Le phénomène qui nous intéresse ici est, pour rappel, la mesure dans laquelle le niveau d'activation spécifique d'un neurone face à un token est lié à une valeur catégorielle particulière de ce token au sein de la catégorie portée par ce neurone. Autrement dit, un segment donné de valeurs d'activations est-il lié à un segment catégoriel délimité, permettant ainsi à un neurone de focaliser son attention élective sur un segment catégoriel particulier sur la base de son empan d'activation associé spécifique ? Les trois facteurs mathématico-cognitifs que nous venons d'expliciter, et que l'on pourrait mettre en relation avec la genèse et la structuration des catégories de pensée humaines [10, 47, 38, 40, 7, 58, 110]sont particulièrement liés à la question qui est la nôtre. En effet, ce sont ces trois facteurs qui déterminent directement, avant l'application de la fonction d'activation non linéaire, la valeur d'activation que la fonction d'agrégation neuronale va assigner à un token donné. Autrement dit, ce sont ces trois facteurs

qui définissent la zone activationnelle, le segment activation au sein duquel va être positionné un token traité par le neurone impliqué. Dans quelle mesure ces trois facteurs, matrice des valeurs d'activation associées à des tokens par un neurone mais également porteurs d'effets catégoriels ainsi que nous l'avons précisé, délimitent-ils des segments activationnels particuliers appariés à des segments catégoriels particuliers qui pourraient dès lors être ainsi objets d'une focalisation attentionnelle spécifique ? Pour le dire de façon alternative, comment ces facteurs sont-ils associés, via les valeurs d'activation neuronale, à la conceptualisation de certains segments catégoriels sur lesquels il est pertinent et efficace pour le neurone de focaliser son attention interne ?

2.2 Détourage catégoriel synthétique et attention intra-neuronale

Dans le cadre d'une autre recherche antérieure [71],nous avons mis en lumière le fait que les trois facteurs mathématico-cognitifs de la segmentation catégorielle, que nous avons précisés ci-avant, pilotent un mécanisme de détourage catégoriel synthétique ; détourage se traduisant par l'élaboration et la séparation d'une forme d'un fond catégoriel. Plus précisément, le détourage catégoriel est le phénomène de la cognition synthétique par lequel une sous-dimension catégorielle particulière est extraite de la catégorie portée par un neurone précurseur (en couche n) afin de participer à la constitution d'une catégorie superordonnée (en couche $n + 1$).

Le détourage se manifeste via une série de caractéristiques synthétiques :

- La réduction catégorielle, à savoir le fait que la sous-dimension catégorielle extraite d'une catégorie précurseure contient des tokens sémantiquement plus homogènes par rapport à la catégorie de départ.
- De façon connexe, la sélectivité catégorielle, se traduisant par l'extraction d'un faible sous-groupe de tokens à partir du groupe de tokens qui caractérisaient la catégorie de départ.
- La séparation des dimensions initiales d'embedding, associée à une différenciation de ces embeddings, certains relevant plus de la sous-dimension catégorielle détourée.
- Le découpage de zones catégorielles des dimensions initiales d'embedding, certaines zones étant plus spécifiquement associées aux sous-dimensions extraites.

Le détourage catégoriel est une activité d'extraction d'une sous-dimension catégorielle de la catégorie associée à un neurone de couche n, qui est réalisée à l'« extérieur » de ce neurone de départ, à savoir au niveau d'un de ses neurones appariés en couche $n + 1$. Mais dans quelle mesure ce détourage, phénomène piloté *in fine* par des valeurs d'activation, est-il opéré sur la base d'une focalisation attentionnelle et d'une conceptualisation intra-neuronales de certains sous-segments catégoriels spécifiques au sein de la catégorie de départ impliquée, sous-segments qui seraient associés à des segments activationnels particuliers ?

2.3 Restructuration catégorielle et attention intra-neuronale

Dans le cadre d'une dernière étude antérieure [72], nous nous sommes intéressés au processus de restructuration catégorielle synthétique, à savoir à la genèse, à chaque couche neuronale $n+1$, de nouvelles catégories de pensée artificielles plus fonctionnelles pour découper le monde des tokens, au service de la finalité de l'activité du réseau de neurones. Ce processus relève d'une abstraction réfléchissante au sens de Piaget (Pichat et al., 2024e), opérée sur les catégories de couche n.

Cette restructuration catégorielle est directement fonction de la coactivité des trois facteurs de la segmentation catégorielle énumérés dans ce qui précède (l'amorçage catégoriel, l'attention inter-catégorielle, le phasage catégoriel). Nous avons avancé que le phénomène de restructuration est particulièrement lié à celui de l'attention inter-catégorielle étant donné que cette dernière, par construction mathématique de la fonction d'agrégation neuronale, est une condition nécessaire et amplificatrice des deux autres facteurs (l'amorçage et le phasage).

Nous avons manifesté le fait que l'action conjointe de l'attention inter-catégorielle et du phasage catégoriel est à l'origine d'une confluence catégorielle partielle : les sous-dimensions catégorielles détourées des catégories de couche$n$u niveau d'un neurone qui leur est fortement attentionnellement en couche $n+1$, tendent à converger sémantiquement dans une mesure relative. Nous avons également mis en lumière que l'impact de concert de l'attention inter-catégorielle et de l'amorçage catégoriel génère une dispersion activationnelle : une sous-dimension catégorielle extraite d'une catégorie en couche n ne relève pas d'un segment continu d'activations des tokens concernés au niveau du neurone de départ.

Comment cette confluence catégorielle partielle d'une part, et cette dispersion activationnelle d'autre part, au niveau du passage d'une couche n à une couche $n+1$, sont-elles réalisées à partir de la conceptualisation et de la focalisation attentionnelle intra-neuronale au niveau de la couche n ? Autrement dit, comment la confluence catégorielle partielle, pilotée par une coactivité des facteurs synthétiques de l'attention inter-catégorielle et du phasage catégoriel, est-elle rendue possible par l'identification, via l'activation, au niveau du neurone précurseur impliqué, de sous-segments catégoriels à prendre spécifiquement en compte ? Et dans quelle mesure la dispersion activationnelle serait-elle compatible avec une focalisation attentionnelle intra-neuronale à partir de segments activationnels particuliers, ce qui semblerait paradoxal ?

2.4 Conceptualisation et attention intra-neuronales

D'un point de vue mathématique, une unité de traitement neuronale synthétique est le fruit d'une composition de fonctions : une fonction d'activation non linéaire (RELU, SELU, GELU, ELU, etc.) appliquée à une fonction linéaire d'agrégation de la forme $\sum(w_{i,j}\ x_{i,j}) + b$. Au sein d'un modèle de langage, ce traitement matriciel conduit à associer à un token (ou plutôt à son embedding) en

entrée une valeur d'activation en sortie. Quel est, d'un point de vue sémantique, la signification épistémologique de cette valeur d'activation ? Cette valeur d'activation est-elle corolaire d'une valeur sémantique spécifique ? Différents segments activationnels, dans l'espace des activations, sont-ils associés à différents segments catégoriels distincts, dans l'espace sémantique associé au neurone impliqué ? Existe-t-il, pour un neurone donné, une relation d'homomorphisme entre son espace d'activation et son espace catégoriel ? Ces espaces sont-ils sécables, segmentables en zones activationnelles et catégorielles délimitables et appariables ? Autrement dit, est-il possible de localiser des segments d'activation intra-neuronaux qui seraient associés à des sous-sémantiques spécifiques au sein des catégories de pensée portées par les neurones ? Autrement dit encore, dans quelle mesure la valeur d'activation est-elle un quantificateur permettant une conceptualisation, un repérage attentionnel de certaines zones sémantiques spécifiques intra-neuronales ?

3 Méthodologie

3.1 Positionnement méthodologique

Afin de situer méthodologiquement notre présente recherche, nous la localisons parmi un ensemble de techniques d'investigation d'explicabilité destinées à rendre les réseaux neuronaux artificiels plus compréhensibles. Ces méthodes cherchent, avec des degrés de profondeur cognitive variés, à expliquer les mécanismes internes ou à interpréter le sens ou la fonction des flux d'informations de ces réseaux, qu'ils soient étudiés en couches neuronales individuelles, en groupes de couches ou dans leur intégralité.

Les études visant une explicabilité de type « macroscopique » se concentrent sur les fluctuations entre les données d'entrée et les résultats, pour clarifier le lien entre ce qui est donné au système et ce qu'il produit. Dans cette optique, les approches fondées sur les gradients établissent l'influence de chaque donnée en examinant les dérivées partielles de chaque dimension d'entrée [37]. Les attributs des entrées peuvent être évalués via divers éléments, y compris les particularités [28],l'importance des éléments (tokens) [37], ou les coefficients d'attention [6]. Par ailleurs, les méthodes exploitant des exemples scrutent les variations de sorties face à des modifications d'entrées, nécessaires pour observer les effets de légères adaptations des données [106], ou pour évaluer les implications d'altérations de données d'entrée comme la suppression, la négation, le mélange ou le camouflage [3, 107, 96]. D'autres approches se dédient également à cartographier conceptuellement les entrées afin d'évaluer leur influence sur les sorties constatées [19].

Les méthodes d'explicabilité de type « microscopique », quant à elles, scrutent les états internes intermédiaires des modèles langagiers plutôt que leur résultat global, en examinant les interconnexions et les activations des neurones individuels ou de groupes de neurones. Certains travaux examinent comment découper et rendre compte des activations neuronales d'une couche à partir des entrées de la

couche antérieure [103]. D'autres cherchent à ajuster les fonctions d'activation pour les rendre plus accessibles [106]. Certaines techniques se penchent sur les connaissances incorporées au sein des neurones, extrapolant des significations internes par des matrices de signification [29, 42]. Enfin, certaines pratiques évaluent les statistiques des réponses neuronales à partir de jeux de données spécifiques [9, 61, 33, 106, 27]. Notre travail actuel s'inscrit dans cette dernière catégorie méthodologique.

3.2 Modèle et unités statistiques mobilisés

Dans la continuité de nos travaux précédants [67, 68, 69, 70, 71], nous nous sommes intéressés au modèle transformer GPT d'OpenAI, particulièrement à la version GPT-2XL. Ce choix s'est imposé car GPT-2XL offre une complexité adéquate pour investiguer des processus cognitifs avancés mais sans nous confronter, pour une première exploration, à la complexité interprétative de GPT-4 ou GPT-4o. En 2023, OpenAI a partagé, comme exposé par Bills et al. [9], une documentation détaillée relative aux paramètres du modèle GPT-2XL, données que nous avons exploitées dans le cadre de notre présente investigation.

Dans le but de réduire l'ampleur de nos traitement statistiques, nous avons ciblé les deux premières couches perceptron de GPT-2XL, qui possèdent chacune 6400 neurones, cumulant 12800 neurones artificiels. Concernant les éléments langagiers (tokens) et leurs activations associées utilisés, notre étude s'est concentrée sur les 100 tokens exhibant les activations moyennes maximales par neurone, que nous avons nommé « core-tokens ». Lorsque nous avons analysé des relations entre neurones des couches 0 et 1, nous nous sommes limités pour chaque neurone de couche 1, à ses 10 neurones à plus forts poids de connexion attentionnelle associés en couche 0.

3.3 Traitements statistiques

Dans le cadre de nos analyses statistiques, nous avons utilisé la librairie SciPy, en accord avec les recommandations émises par Howell , Beaufils et Ellis. [50, 8, 34, 35].

Notre vérification de la normalité de nos données, afin d'évaluer la possibilité de mobiliser des tests paramétriques, a été opérée en deux étapes. Dans un premier temps, nous avons employé des tests d'inférence : le test de Shapiro-Wilk, pertinent pour les petits ensembles de données ; le test de Lilliefors, adapté lorsque les paramètres de distribution sont inconnus ; le test de Kolmogorov-Smirnov pour les larges échantillons ; et le test de Jarque-Bera quantifiant la symétrie et la distribution des données au sein des échantillons vastes. Complétant cela, dans un second temps, avec l'étude des coefficients de skewness et de kurtosis, et en visualisant nos distributions avec une démarche de type QQ-plot, afin de comparer les données enregistrées avec une distribution normale théorique. Pour contrôler l'homogénéité de variances entre les sous-groupes de données dont la relation était à étudier, nous avons appliqué le test de Bartlett, sensible face à la non-normalité, et le test de Levene, moins influencé par les écarts à la normalité.

De ces vérifications préliminaires est ressortie une normalité limitée de nos données. Par conséquent, nos études statistiques ont reposé sur des démarches non paramétriques, à savoir :

- Le test de Kruskal-Wallis, pour investiguer la relation d'une variable catégorielle définissant divers groupes indépendants et une variable ordinale ; test mobilisé à partir d'un ordonnancement des données numériques d'activation neuronale des tokens ; cela, dans le respect des conditions usuelles d'application de ce test, dont des groupes d'au minimum 6 observations. Les tailles d'effet associées au Kruskal-Wallis ont été mesurées à partir du paramètre *d*. de Cohen. Concernant nos comparaisons de groupes 2 à 2, k étant le nombre total de groupes, $k(k-1)/2$ comparaisons ont été effectuées, cela avec le coefficient de calcul de différence des rangs adapté à ce type de situations post hoc, mesuré sur un échelle de type $z$et en utilisant un seuil α de significativité divisé par$k(k-1)$.
- Le test χ^2 univarié d'ajustement, compte tenu de ses exigences concernant l'adéquation des effectifs théoriques et observés, s'abstenant ainsi de recourir aux alternatives pour de petits échantillons, telles les méthodes de Fisher ou Monte Carlo. Les tailles d'effets associées ont été mesurée à partir du risk ratio.

Pour certaines études, nous avons mobilisé une démarche d'analyse typologique par classification hiérarchique. Cette dernière a alors été paramétrée de la façon suivante : (i) choix d'une classification descendante, (ii) utilisation d'une distance euclidienne de mesure des distances entre unités statistiques, (iii) nombre de clusters définis à 5, (iv) sélection de la démarche de Ward comme méthode d'agrégation, (v) standardisation préalable des données.

3.4 Opérationnalisations méthodologiques

Afin de quantifier la proximité sémantique entre tokens, le cosinus de similarité de la base d'embeddings de GPT-2XL a été notre référence, évitant ainsi les limitations méthodiques repérées par Bills et al. [9]lorsqu'il s'agit de relier divers systèmes cognitifs artificiels sur des fondations non unifiées d'embeddings.

Dans le but d'étudier l'éventuelle relation, pour chaque neurone donné, entre ses segments catégoriels et ses segments d'activation, nous avons mobiliér deux types de démarches antagonistes :

- Premièrement, une démarche « descendante », consistant à partir de segments catégoriels définis, puis à étudier dans quelle mesure ils sont associés à des segments activationnels contrastés. Ces segments catégoriels étant définis dans un premier temps par classification hiérarchique descendante à partir des embeddings GP2-XL des 100 core-tokens définissant la catégorie associée à chaque neurone investigué ; puis, dans un second temps, à des fins de diversification méthodologique, par prompt engineering de segmentation avec le modèle GPT4o d'OpenAI. Cette première démarche a, enfin, été prolongée par une mesure des entrelacements et overlaps activationnels des segments catégoriels de tokens obtenus par prompt engineering.

- Deuxièmement, une approche « ascendante », se traduisant par le fait de partir de segments cette fois-ci activationnels définis, puis d'évaluer dans quelle mesure ces derniers sont l'objet d'une certaine proximité sémantique. Ces segments activationnels, à nouveau à des fins de diversification méthodologique, étant définis par quartiles d'activation puis par clusterisation des activations par classification hiérarchique descendante.

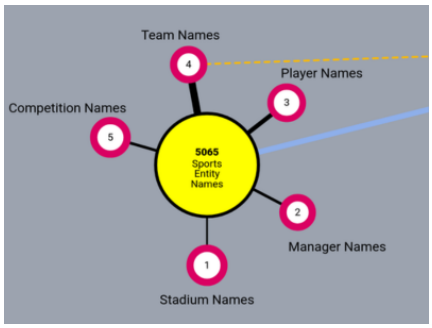
4 Résultats

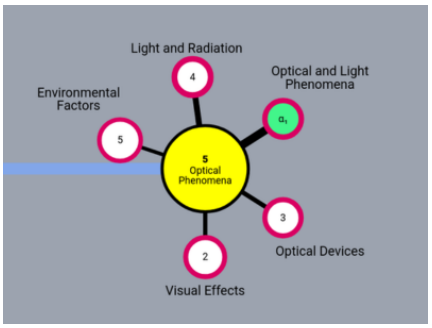
Pour rappel, notre questionnement est le suivant : existe-t-il un processus synthétique de conceptualisation et d'attention intra-neuronales, permettant à chaque neurone de repérer et d'isoler, au sein de la catégorie de pensée artificielle qu'il porte, certains segments catégoriels spécifiques à partir de segments activationnels déterminés. Nous avons opérationnalisé notre démarche empirique d'investigation de cette question à travers deux études méthodologiquement antagonistes, que nous allons présenter successivement :

- Une étude « top-down », au sein de laquelle nous sommes parti de segments catégoriels, pour examiner leurs valeurs activationnelles moyennes respectives.
- Puis, une étude « bottom-up », dans le cadre de laquelle nous avons en premier isolé des segments activationnels donnés, afin d'étudier ensuite leurs homogénéités catégorielles respectives.

4.1 Différentiation activationnelle des clusters catégoriels

Dans le cadre de notre étude « top-down », et à l'endroit de chaque neurone perceptron des couches 0 et 1 de GPT2-XL, nous avons décomposé la catégorie de pensée qu'il porte (via ses 100 core-tokens, c'est-à-dire ses tokens les plus activés en moyenne) en 5 clusters catégoriels. Ces derniers sont des sous-catégories, relativement homogènes au titre d'un critère sémantique donné, en lesquelles il est possible de partitionner leur catégorie de départ. Nous présentons à titre d'illustration ci-dessous, l'exemple des neurones n°5 (couche 1) et n°5065 (couche 0), à partir de notre *genetic neural viewer*.Cela, dans le but d'étudier la question de la relation entre segmentation (clusterisation) catégorielle et segmentation activationnelle ; et, de façon plus opérationnalisée, d'étudier dans quelle mesure il existait une différence significative de moyennes activationnelles entre les clusters catégoriels obtenus.





Une première opérationnalisation de notre démarche a consisté, pour chaque neurone, à générer 5 clusters catégoriels de tokens, à partir d'une classification hiérarchique descendante opérée sur les embeddings GPT2-XL des 100 core-tokens de ce neurone. D'un point de vue méthodologique, précisons les éléments suivants :

- Le recours aux embeddings GPT2-XL est, comme toute opérationnalisation, un choix singulier, qui n'exprime dès lors nullement la mesure d'une quelconque réalité sémantique intrinsèque, notion qui n'a aucun sens d'un point de vue épistémologique, mais uniquement une modalité particulière, parmi d'autres possibles, d'évaluation de phénomènes sémantiques au sein d'un espace sémantique contingent donné. Nos interprétations se devront dès lors d'être circonscrites à cette contingence sémantique spécifique.
- Au sein de nos statistiques globales, nous n'avons retenu que les neurones dont la clusterisation sémantique a abouti à des clusters contenants tous au moins 6 tokens, afin de respecter les conditions d'utilisation du test inférentiel de Kruskall-Wallis de comparaison de moyennes que nous avons mobilisé.
- Nous avons systématiquement, pour chaque neurone, nommé « K_1 » son cluster catégoriel associé à la plus faible valeur moyenne d'activation de ses tokens constitutifs, et ainsi de suite jusqu'à son cluster « K_5 ».

Le tableau n°1 synthétise nos résultats obtenus pour les 2194 neurones étudiés au sein de la couche 0. Nous obtenons de façon globale un faible pourcentage (21.46%) de neurones présentant une différence significative ($\alpha = 5\%$)de moyennes d'activation (notéesμ_{K_n})des tokens constitutifs de ses 5 clusters catégoriels associés. Ce qui est confirmé par les faibles pourcentages (notés $\pi(p_{K_n,K_m} < \alpha')$) de différences significatives de moyennes d'activation au niveau de tests post hoc comparant les clusters catégoriels 2 à 2 (avec un seuil de significativité ajusté $\alpha' = \alpha/20$).

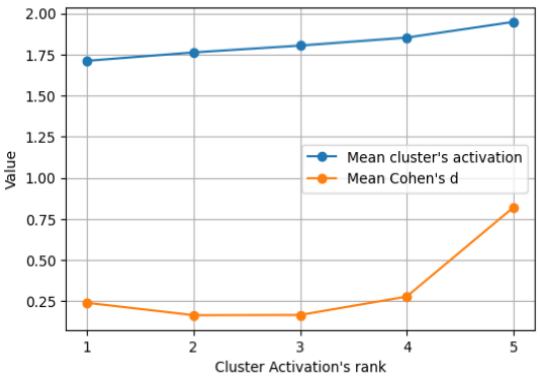
Cependant, une tendance intéressante est à noter : si les distances de moyennes activationnelles (notées$\mu(\delta_{K_n,K_m})$) entre les clusters catégoriels successifs sont faibles, la distance activationnelle moyenne entre les clusters catégoriels K_4 et K_5 ($\mu(\delta_{K_4,K_5}) = .0955$) est légèrement plus importante que la distance entre les clusters K_1 et K_2 ($\mu(\delta_{K_1,K_2}) = .0512$). Tendance en phase avec la

légère supériorité de taille d'effet (mesurée avec un$d$) e Cohen) de la distance activationnelle moyenne entre K_4 et K_5 ($\mu(d_{K_4,K_5}) = .2785$) par rapport à celle entre K_1 et K_2 ($\mu(d_{K_1,K_2}) = .2404$).

Mais cette tendance devient beaucoup plus manifeste et importante lorsque l'on regarde la très forte taille d'effet de la différence de moyenne d'activations entre les clusters K_1 et K_5 ($\mu(d_{K_1,K_5}) = .8202$) ; et l'on pourrait d'ailleurs formuler l'hypothèse que la faible significativité des tests inférentiels post hoc de comparaison de μ_{K_1} and μ_{K_5} ($\pi(p_{K_1,K_5} < \alpha') = 14.02\%$) est due à un biais issu du faible nombre de tokens impliqués. Le graphe n°1 synthétise visuellement nos principales données ici mentionnées.

N_{neuron}		2194	
$\pi(p_{KW} < .05)$		21.4612	
μ_{K_1}	1.7117	$\mu(\delta_{K_1,K_2})$	.0512
μ_{K_2}	1.7629	$\mu(\delta_{K_2,K_3})$	.0416
μ_{K_3}	1.8045	$\mu(\delta_{K_3,K_4})$	.0484
μ_{K_4}	1.8530	$\mu(\delta_{K_4,K_5})$	.0955
μ_{K_5}	1.9485	$\mu(\delta_{K_1,K_5})$	.2368
$\mu(d_{K_1,K_2})$	.2404	$\pi(p_{K_1,K_2} < \alpha')$	.1370
$\mu(d_{K_2,K_3})$	.1650	$\pi(p_{K_2,K_3} < \alpha')$	.0913
$\mu(d_{K_3,K_4})$	.1664	$\pi(p_{K_3,K_4} < \alpha')$	.2283
$\mu(d_{K_4,K_5})$	.2785	$\pi(p_{K_4,K_5} < \alpha')$	1.3242
$\mu(d_{K_1,K_5})$	.8202	$\pi(p_{K_1,K_5} < \alpha')$	14.0183

Tableau n°1 : Comparaison des activations moyennes entre les clusters catégoriels issus d'une classification hiérarchique sur les embeddings des tokens (couche 0).



Graph n°1 : Comparison of mean activations between categorical clusters from hierarchical classification on tokens' embeddings (layer 0).

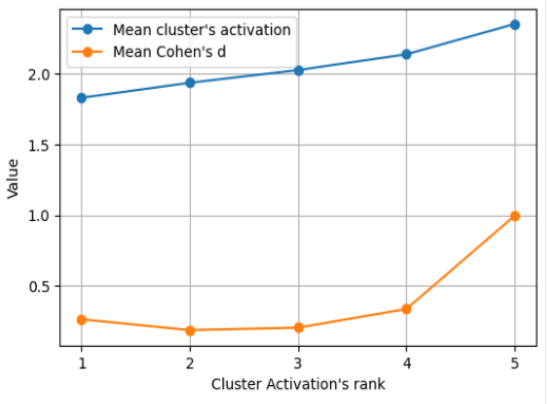
Il semble ressortir de ces premiers résultats la double tendance suivante :

- Une indifférenciation activationnelle (i.e. une faible différence de valeurs moyennes d'activation) entre les clusters catégoriels associés à de plus faibles valeurs moyennes d'activation.
- Une différentiation activationnelle relative (i.e. une plus forte différence de valeurs moyennes d'activation) entre les clusters catégoriels impliquant une ou plusieurs plus forte(s) valeur(s) moyenne(s) d'activation.

Le tableau n°2 et son graphe n°2 de synthèse associé font montre de résultats analogues, mais avec des tendances plus marquées encore ; dont une différence de taille d'effet plus contrastée entre les clusters K_4et K_5 ($\mu(d_{K_4,K_5}) = .3363$) et les clusters K_1 et K_2 ($\mu(d_{K_1,K_2}) = .2631$) ; et une taille d'effet extrêmement forte lorsque l'on compare les clusters K_1 et K_5($\mu(d_{K_1,K_5}) = .9962$),associée à une significativité accrue ($\pi(p_{K_1,K_5} < \alpha') = 22.08\%$).

N_{neuron}		2192	
$\pi(p_{KW} < .05)$		31.4325	
μ_{K_1}	1.8300	$\mu(\delta_{K_1,K_2})$	.1062
μ_{K_2}	1.9362	$\mu(\delta_{K_2,K_3})$	.0896
μ_{K_3}	2.0258	$\mu(\delta_{K_3,K_4})$	.1119
μ_{K_4}	2.1377	$\mu(\delta_{K_4,K_5})$	.2143
μ_{K_5}	2.3519	$\mu(\delta_{K_1,K_5})$	.5219
$\mu(d_{K_1,K_2})$	.2631	$\pi(p_{K_1,K_2} < \alpha')$	.0456
$\mu(d_{K_2,K_3})$	.1869	$\pi(p_{K_2,K_3} < \alpha')$	.3193
$\mu(d_{K_3,K_4})$	.2042	$\pi(p_{K_3,K_4} < \alpha')$	.5931
$\mu(d_{K_4,K_5})$	.3363	$\pi(p_{K_4,K_5} < \alpha')$	2.1898
$\mu(d_{K_1,K_5})$	.9962	$\pi(p_{K_1,K_5} < \alpha')$	22.0803

Tableau n°2 : Comparaison des activations moyennes entre les clusters catégoriels issus d'une classification hiérarchique sur les embeddings des tokens (couche 1).



Graphe n°2 : Comparaison des activations moyennes entre les clusters catégoriels issus d'une classification hiérarchique sur les embeddings des tokens (couche 1).

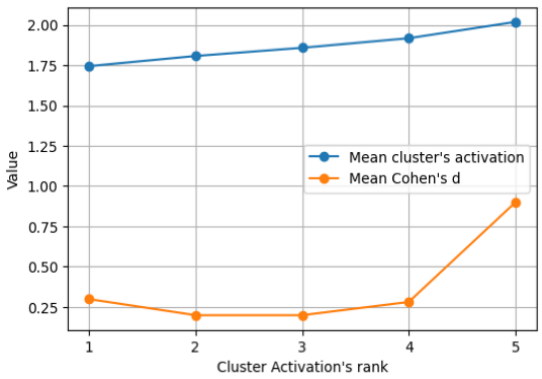
Toujours dans le cadre de notre étude « top-down », une deuxième opérationnalisation de notre démarche a consisté, pour chaque neurone, à générer 5 clusters catégoriels de tokens (toujours à partir des 100 core-tokens de ce neurone), sur la base cette fois-ci d'un système de prompts opérés sur GPT4 au sein d'une structure de codage Python ; cela, afin d'investiguer notre sujet d'étude à partir d'un autre référentiel d'observation sémantique couplé à une autre méthodologie de système de clusterisation.

Le tableau n°3 et son graphe n°3 associé, portant ici sur 2316 neurones de

la couche 0, manifestent de façon invariante les mêmes types de résultats que précédemment.

N_{neuron}		2316	
$\pi(p_{KW} < .05)$		18.3938	
μ_{K_1}	1.7443	$\mu(\delta_{K_1,K_2})$	.0629
μ_{K_2}	1.8072	$\mu(\delta_{K_2,K_3})$	.0512
μ_{K_3}	1.8584	$\mu(\delta_{K_3,K_4})$	.0592
μ_{K_4}	1.9176	$\mu(\delta_{K_4,K_5})$	.1015
μ_{K_5}	2.0190	$\mu(\delta_{K_1,K_5})$	.2747
$\mu(d_{K_1,K_2})$	.2981	$\pi(p_{K_1,K_2} < \alpha')$	.0864
$\mu(d_{K_2,K_3})$	.1982	$\pi(p_{K_2,K_3} < \alpha')$	.0432
$\mu(d_{K_3,K_4})$	.1982	$\pi(p_{K_3,K_4} < \alpha')$	.0000
$\mu(d_{K_4,K_5})$	.2803	$\pi(p_{K_4,K_5} < \alpha')$	.2591
$\mu(d_{K_1,K_5})$	.8985	$\pi(p_{K_1,K_5} < \alpha')$	11.9603

Tableau n°3 : Comparaison des activations moyennes entre les clusters catégoriels issus d'un prompt de clustering GPT4 sur les embeddings des tokens (couche 0).

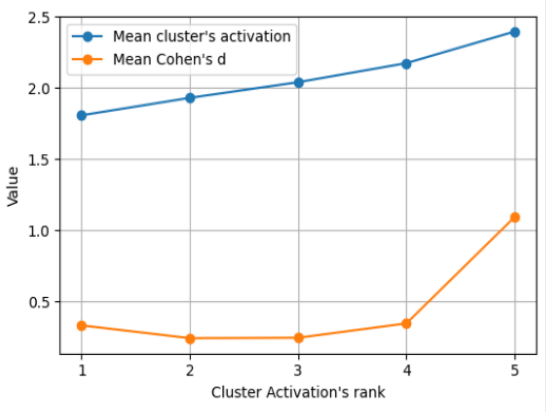


Graph n°3 : Comparison of mean activations between categorical clusters from GPT4 clustering prompt (layer 0).

Et il en va à nouveau de même au niveau du tableau n°4 et son graphe n°4 respectif, ayant trait à 1942 neurones de la couche 1. Ce qui pourrait au demeurant appuyer une hypothèse postulant que les phénomènes synthétiques que nous pointons ont tendance à augmenter en couches plus profondes.

N_{neuron}		1942	
$\pi(p_{KW} < .05)$		29.8661	
μ_{K_1}	1.8086	$\mu(\delta_{K_1,K_2})$	.1229
μ_{K_2}	1.9316	$\mu(\delta_{K_2,K_3})$	.1099
μ_{K_3}	2.0415	$\mu(\delta_{K_3,K_4})$	.1341
μ_{K_4}	2.1756	$\mu(\delta_{K_4,K_5})$	.2223
μ_{K_5}	2.3979	$\mu(\delta_{K_1,K_5})$	.5892
$\mu(d_{K_1,K_2})$	.3321	$\pi(p_{K_1,K_2} < \alpha')$	.1030
$\mu(d_{K_2,K_3})$	.2424	$\pi(p_{K_2,K_3} < \alpha')$	.1030
$\mu(d_{K_3,K_4})$	.2451	$\pi(p_{K_3,K_4} < \alpha')$	.0515
$\mu(d_{K_4,K_5})$	.3461	$\pi(p_{K_4,K_5} < \alpha')$	.8754
$\mu(d_{K_1,K_5})$	1.0948	$\pi(p_{K_1,K_5} < \alpha')$	23.0690

Tableau n°4 : Comparaison des activations moyennes entre les clusters catégoriels issus d'un prompt de clustering GPT4 sur les embeddings des tokens (couche 1).



Graphe n°4 : Comparaison des activations moyennes entre les clusters catégoriels issus d'un prompt de clustering GPT4 (couche 1).

De ces différentes premières investigations « top-down », il semble pointer de façon invariante une double tendance quant à la question de l'existence d'un éventuel mécanisme synthétique d'attention intra-neuronale, rendant possible pour un neurone d'identifier et de localiser, à l'intérieur de la catégorie de pensée artificielle qu'il vectorise, certains segments catégoriels déterminés sur la base de segments activationnels donnés :

- Une indifférenciation activationnelle entre les clusters catégoriels relevant de plus faibles valeurs moyennes d'activation.

- Une différentiation activationnelle relative entre les clusters catégoriels ayant trait à une ou plusieurs plus forte(s) valeur(s) moyenne(s) d'activation.

4.2 Entrelacement activationnel des clusters catégoriels

Nous sommes, à nouveau, dans le cadre de notre investigation « top-down » relative à l'existence d'une dynamique synthétique de conceptualisation et d'attention interne à chaque neurone, permettant à celui-ci de déterminer et de différencier, à l'intérieur de sa catégorie de pensée associée, certains segments catégoriels précis à partir de zones activationnelles particulières. Mais, ici, nous adoptons maintenant un angle méthodologique significativement différent. Cela, en investiguant dans quelle mesure les clusters catégoriels que nous obtenons (avec la dernière technique précédente impliquant un prompt appliqué à GPT4) forment des segments distincts quant à leur niveau d'activation. Autrement dit, dans quelle ampleur les segments activationnels propres aux clusters catégoriels se chevauchent (i.e. contiennent des tokens de différents clusters catégoriels) ou non (i.e. le segment activationnel de chaque cluster catégoriel ne contient que des tokens issus de ce seul cluster).

Cela, via l'opérationnalisation suivante :

- Soient $x_{\min}(K_i)$ l'activation la plus faible des tokens du cluster catégorielK_i, et $x_{\max}(K_i)$ l'activation la plus forte des tokens de ce même cluster.
- Soient $n(K_i)$ le nombre de tokens (issus de l'ensemble des 5 clusters K_1, K_2, K_3, K_4, K_5) confondus) dont l'activation appartient à$[x_{\min}(K_i), x_{\max}(K_i)]$, et $m(K_i)$ le nombre de tokens appartenant au cluster K_i.
- Soient Nle nombre total de tokens clustérisés (100, ou moins lorsque GPT4 n'a pas été en mesure de clustériser tous les tokens).
- Pour un cluster catégoriel donnéK_i, il n'y a pas chevauchement si $n(K_i) = m(K_i)$.

Pour chaque neurone de chacune des couches 0 et 1 de GPT2-XL, nous avons étudié cette opérationnalisation à partir de tableaux de contingence, de la forme présentée ci-dessous et d'un calcul inférentiel de χ^2 en ne retenant que les cas des clusters présentant un effectif théorique strictement supérieur à 5, dans le respect des conditions d'application du χ^2) et un effectif observé également strictement supérieur à 5.

	K_i	
Observé	$n(K_i)$	$N - n(K_i)$
Attendu	$m(K_i)$	$N - m(K_i)$
Risk ratio	$\frac{n(K_i)}{m(K_i)}$	

Les tableaux n°5 et n°6, respectivement réalisés sur 2316 neurones éligibles de la couche 0 et 1942 neurones éligibles de la couche 1, font montre d'un chevauchement, d'un entrelacement activationnel fort. Cela, avec des moyennes (μ_ρ)de risk ratio très élevées, oscillant entre 4.5 et 5, manifestant une importante taille d'effet de chevauchement : en moyenne, un segment activationnel associé à un cluster catégoriel donné, contient 4 à 5 fois plus de tokens que le nombre de tokens effectivement associé à ce token. Cette phénoménologie étant très largement significative au niveau inférentiel, avec des pourcentages ($\pi(p(\chi^2) < .05)$)très importants, pour chaque cluster, de cas où distributions observées et théoriques de tokens sont fortement contrastées.

N_{neuron}	2316	
K_1	μ_ρ	5.1103
	$\pi(p(\chi^2) < .05)$	100
K_2	μ_ρ	4.9323
	$\pi(p(\chi^2) < .05)$	100
K_3	μ_ρ	4.8721
	$\pi(p(\chi^2) < .05)$	100
K_4	μ_ρ	4.9705
	$\pi(p(\chi^2) < .05)$	100
K_5	μ_ρ	5.0677
	$\pi(p(\chi^2) < .05)$	99.8705

Tableau n°5 : Pourcentages d'entrelacement des activations des clusters catégoriels de core-tokens (couche 0).

N_{neuron}	1942	
K_1	μ_ρ	4.6122
	$\pi(p(\chi^2) < .05)$	99.9485
K_2	μ_ρ	4.6212
	$\pi(p(\chi^2) < .05)$	100
K_3	μ_ρ	4.6998
	$\pi(p(\chi^2) < .05)$	99.9485
K_4	μ_ρ	4.6624
	$\pi(p(\chi^2) < .05)$	99.9485
K_5	μ_ρ	4.5496
	$\pi(p(\chi^2) < .05)$	99.6395

Tableau n°6 : Pourcentages d'entrelacement des activations des clusters catégoriels de core-tokens (couche 1).

L'entrelacement des segments activationnels définis par les clusters catégoriels que nous pointons ici tend à montrer que des segments catégoriels déterminés ne sont pas associés à des segments activationnels délimités. Ce résultat est cohérent avec l'indifférenciation activationnelle (entre les clusters catégoriels relevant de plus faibles valeurs moyennes d'activation) que nous avons mis en lumière au sein de la section précédente. Et il n'est a priori pas incohérent, en l'état, avec la différentiation activationnelle (relative entre les clusters catégoriels ayant trait à de plus fortes valeurs moyennes d'activation) mentionnée précédemment ; en effet, ce résultat ne nous permet pas de savoir dans quelle mesure les tokens « surnuméraires » (i.e. entrelacés quant à leur niveau d'activation) des clusters K_5 ont plutôt tendance à venir de tous les clusters catégoriels (ce qui pourrait alors contredire le phénomène de différenciation activationnelle) ou des clusters K_4 et peu des clusters K_1 (ce qui ne serait alors pas incohérent avec le phénomène de différenciation activationnelle).

4.3 Homogénéité catégorielle des clusters activationnels

Examinons maintenant les résultats de notre seconde étude « bottom-up », dans le cadre de laquelle nous avons en premier isolé des segments activationnels donnés afin d'investiguer ensuite leurs niveaux homogénéité catégorielle respectifs ; autrement dit, afin de voir dans quelle mesure des tokens issus d'un segment activationnel donné tendent à être proches d'un point de vue catégoriel. Cela, toujours au service de notre questionnement central dans cet article : existe-t-il un processus synthétique de conceptualisation et d'attention intra-neuronales, permettant à chaque neurone de repérer et d'isoler, au sein de la catégorie de pensée artificielle qu'il porte, certains segments catégoriels spécifiques à partir de segments activationnels déterminés.

D'un point de vue méthodologique, la proximité catégorielle a été ici décidée d'être étudiée à travers un calcul de similarité cosinus sur la base des embeddings de GPT2-XL ; précisons d'emblée, à l'instar de ce que nous avons déjà pointé, que l'utilisation des embeddings GPT2-XL est, comme n'importe quelle opérationnalisation, un choix spécifique, qui ne relate dès lors pas l'évaluation d'une quelconque ontologie sémantique absolue, mais uniquement une façon particulière, parmi d'autres, de mesurer une proximité catégorielle au sein d'un espace sémantique particulier donné ; et que nos interprétations devront ainsi être contextualisées au sein de cette contingence sémantique donnée.

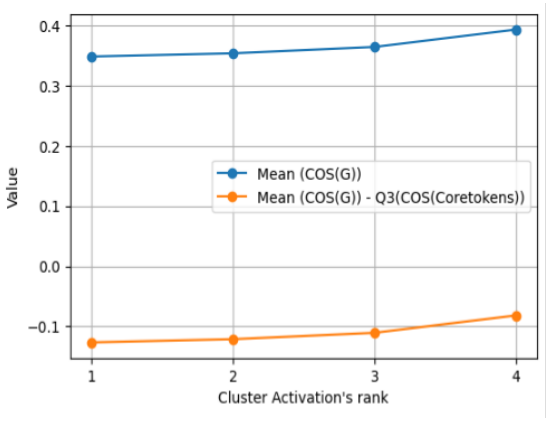
Plus précisément, au niveau méthodologique, nous avons procédé comme suit. Pour chacun des 6400 neurones de chacune des deux premières couches de GPT2-XL, nous avons scindé l'étendue de l'espace d'activation en 4 segments activationnels ; et ainsi obtenu 4 groupes (G_1, G_2, G_3, G_4) de tokens, ordonnés quant à leurs niveaux moyens d'activation. Pour chacun de ces 4 groupes, nous avons déterminé son homogénéité sémantique interne ($\cos_{Gi}$) en calculant la proximité cosinus moyenne de tous ses tokens 2 à 2. Puis nous avons calculé un indice : $d = \cos_{Gi} - Q_3(\cos_{100})$, $Q_3(\cos_{100})$ étant le 3ème quartile de l'ensemble des proximités 2 à 2 des 100 core-tokens du neurone impliqué. Cet indice d exprime donc dans quelle mesure les tokens du groupe G_i sont parmi les plus proches sémantiquement (par rapport à l'ensemble des proximités sémantiques de tokens du neurone) ; plus précisément, plus d est négatif et moins les tokens concernés sont proches sémantiquement (toujours relativement à l'ensemble des tokens concernés par un neurone donné).

Dans un premier temps, nous avons appliqué cette démarche en opérationnalisant la segmentation de l'étendue de l'espace d'activation en 4 segments activationnels en termes de quartiles, produisant ainsi 4 groupes comprenant chacun 25 % des 100 core-tokens de départ, pour un neurone donné. Cette démarche ayant l'avantage d'étudier des clusters activationnels homogènes quant à leurs effectifs respectifs de tokens. Le tableau n°7 indique les résultats obtenus sur les 6400 neurones de la couche 0. Nous voyons premièrement que les cosinus similarité moyens($\mu(\cos_{Gi})$) obtenus au sein des 4 groupes sont assez faibles, allant de .34 à .39 ; cela, avec des pourcentages ($\pi(\mu(\cos_{Gi} - Q_3(\cos_{100}) < 0)$) très importants (tous supérieurs à 97%) de cas où les proximités cosinus moyennes sont inférieures au 3ème quartile de l'ensemble des proximités, c'est-à-dire de cas où les proximités cosinus moyennes de groupes G_i ne sont pas parmi les plus fortes de chaque neurone concerné ; pourcentages associés à des probabilités inférentielles ($p(\chi^2)$),extrêmement faibles et donc significatives (avec hypothèse d'équi-distribution théorique). Ces éléments traduisent une faible homogénéité catégorielle des clusters activationnels. Seconde classe de résultats à noter, nous pouvons observer avec intérêt que les cosinus similarité moyens ($\mu(\cos_{Gi})$ augmentent, de façon certes relative mais résolument systématique, des groupes G_1 (les moins activés quant à leurs tokens propres) (.3485) aux groupes G_4 (les plus activés quant à leurs tokens) (.3933) ; et que, de façon corolaire, les distances $\mu(\cos_{Gi}) - Q_3(\cos_{100})$ tendent à légèrement diminuer en valeur absolue, tout comme leurs pourcentages de négativité associés ($\pi(\mu(\cos_{Gi}) - Q_3(\cos_{100}) < 0)$) ; cela exprimant une tendance, légère mais

régulière, d'augmentation de l'homogénéité catégorielle des clusters activationnels pour les valeurs activationnelles les plus fortes. Le graphe n°5 synthétise visuellement les résultats que nous venons de détailler.

N_{neuron}	6400
$\mu(\cos_{G_1})$	.3485
$\mu(\cos_{G_1}) - Q_3(\cos_{100})$	-.1267
$\pi(\mu(\cos_{G_1}) - Q_3(\cos_{100}) < 0)$	99.9531
$p(\chi^2)$	.0000
$\mu(\cos_{G_2})$	.3538
$\mu(\cos_{G_2}) - Q_3(\cos_{100})$	-.1213
$\pi(\mu(\cos_{G_2}) - Q_3(\cos_{100}) < 0)$	99.9844
$p(\chi^2)$	.0000
$\mu(\cos_{G_3})$	.3644
$\mu(\cos_{G_3}) - Q_3(\cos_{100})$	-.1108
$\pi(\mu(\cos_{G_3}) - Q_3(\cos_{100}) < 0)$	99.9688
$p(\chi^2)$	.0000
$\mu(\cos_{G_4})$	.3933
$\mu(\cos_{G_4}) - Q_3(\cos_{100})$	-.0819
$\pi(\mu(\cos_{G_4}) - Q_3(\cos_{100}) < 0)$	97.6250
$p(\chi^2)$	.0000

Tableau n°7 : Similarités cosinus moyennes des clusters d'activation par quartiles (couche 0).

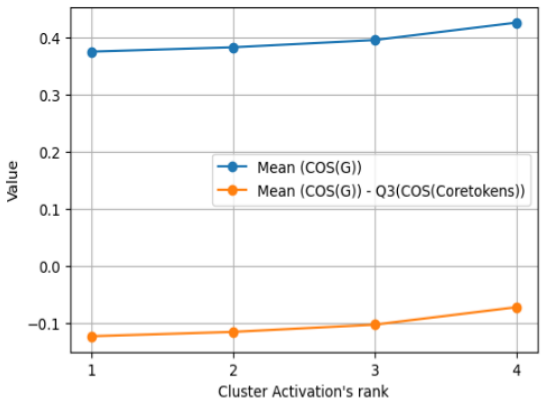


Graphe n°5 : Similarités cosinus moyennes des clusters d'activation par quartiles (couche 0).

Le tableau n°8 et son graphe de synthèse associé n°6 relatent exactement le même type de résultats pour les 6400 neurones de la couche 1. Nous y notons à nouveau de faibles valeurs $\mu(\cos_{G_1})$ mais une relative augmentation progressive de ces dernières au fils de la croissance des niveaux d'activation, ainsi qu'une diminution progressive relative de l'écart de $\mu(\cos_{G_1})$ à $Q_3(\cos_{100})$.

N_{neuron}	6400
$\mu(\cos_{G_1})$	.3753
$\mu(\cos_{G_1}) - Q_3(\cos_{100})$	-.1235
$\pi(\mu(\cos_{G_1}) - Q_3(\cos_{100}) < 0)$	99.9219
$p(\chi^2)$	.0000
$\mu(\cos_{G_2})$	.3829
$\mu(\cos_{G_2}) - Q_3(\cos_{100})$	-.1159
$\pi(\mu(\cos_{G_2}) - Q_3(\cos_{100}) < 0)$	99.9375
$p(\chi^2)$	.0000
$\mu(\cos_{G_3})$	.3956
$\mu(\cos_{G_3}) - Q_3(\cos_{100})$	-.1033
$\pi(\mu(\cos_{G_3}) - Q_3(\cos_{100}) < 0)$	99.9219
$p(\chi^2)$	.0000
$\mu(\cos_{G_4})$	.4261
$\mu(\cos_{G_4}) - Q_3(\cos_{100})$	-.0728
$\pi(\mu(\cos_{G_4}) - Q_3(\cos_{100}) < 0)$	96.6563
$p(\chi^2)$	.0000

Tableau n°8 : Similarités cosinus moyennes des clusters d'activation par quartiles (couche 1).



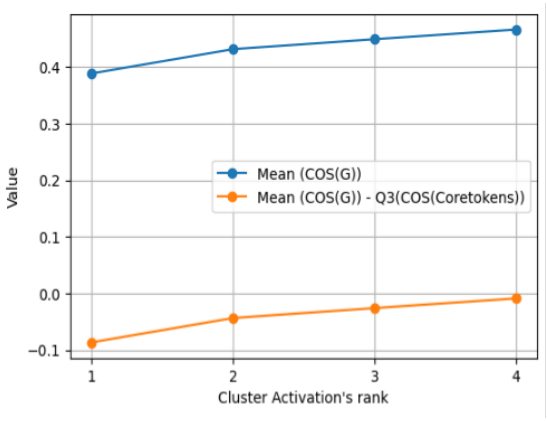
Graphe n°6 : Average cosine similarities of activation clusters by quartiles (Layer 1).

Dans un deuxième temps, et toujours afin de diversifier notre méthodologie, nous avons appliqué notre démarche « bottom-up » en opérationnalisant la segmentation de l'étendue de l'espace d'activation toujours en 4 segments

activationnels à partir, cette fois-ci, d'une classification hiérarchique opérée sur les valeurs d'activation des 100 core-tokens de chaque neurone. Cette démarche ayant l'avantage d'étudier des clusters activationnels plus homogènes quant à leurs valeurs respectives moyennes d'activation ; valeur ajoutée d'autant plus pertinente que notre analyse est ici centrée sur la différenciation des proximités catégorielles en fonction des zones activationnelles, zones activationnelles a priori identifiées de façon plus pertinente par une approche de classification hiérarchique justement de nature à différencier des segments activationnels plus contrastés entre eux (maximalisation de la variance inter) et plus homogènes en leur seins propres (minimalisation de la variance intra). Le tableau n°9 et son graphe associé n°7, relatifs aux 6400 neurones de la couche 0, ainsi que le tableau n°10 et son graphe corrolaire n°8, relatifs aux 6400 neurones de la couche 0, font montre d'une constance des phénomènes synthétiques que nous venons de pointer. Mais, ici, avec une accentuation de l'effet d'augmentation de l'homogénéité catégorielle des clusters activationnels pour les valeurs activationnelles les plus fortes.

N_{neuron}	6400
$\mu(\cos_{G_1})$	.3888
$\mu(\cos_{G_1}) - Q_3(\cos_{100})$	-.0863
$\pi(\mu(\cos_{G_1}) - Q_3(\cos_{100}) < 0)$	73.5938
$p(\chi^2)$	.0000
$\mu(\cos_{G_2})$	.4318
$\mu(\cos_{G_2}) - Q_3(\cos_{100})$	-.0433
$\pi(\mu(\cos_{G_2}) - Q_3(\cos_{100}) < 0)$	65.4219
$p(\chi^2)$	.002
$\mu(\cos_{G_3})$	.4494
$\mu(\cos_{G_3}) - Q_3(\cos_{100})$	-.0258
$\pi(\mu(\cos_{G_3}) - Q_3(\cos_{100}) < 0)$	59.4844
$p(\chi^2)$	.0578
$\mu(\cos_{G_4})$	.4665
$\mu(\cos_{G_4}) - Q_3(\cos_{100})$	-.0087
$\pi(\mu(\cos_{G_4}) - Q_3(\cos_{100}) < 0)$	50.5625
$p(\chi^2)$	.9104

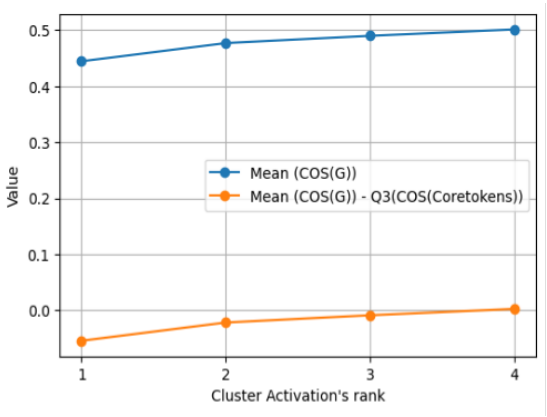
Tableau n°9 : Similarités cosinus moyennes des clusters d'activation par classification hiérarchique (couche 0).



Graphe n°7 : Similarités cosinus moyennes des clusters d'activation par classification hiérarchique (couche 0).

N_{neuron}	6400
$\mu(\cos_{G_1})$	.4444
$\mu(\cos_{G_1}) - Q_3(\cos_{100})$	-.0544
$\pi(\mu(\cos_{G_1}) - Q_3(\cos_{100}) < 0)$	66.4688
$p(\chi^2)$	.001
$\mu(\cos_{G_2})$	.4769
$\mu(\cos_{G_2}) - Q_3(\cos_{100})$	-.0219
$\pi(\mu(\cos_{G_2}) - Q_3(\cos_{100}) < 0)$	60.5625
$p(\chi^2)$	.0346
$\mu(\cos_{G_3})$	.4899
$\mu(\cos_{G_3}) - Q_3(\cos_{100})$	-.0090
$\pi(\mu(\cos_{G_3}) - Q_3(\cos_{100}) < 0)$	55.1875
$p(\chi^2)$	.2995
$\mu(\cos_{G_4})$	.5013
$\mu(\cos_{G_4}) - Q_3(\cos_{100})$	.0024
$\pi(\mu(\cos_{G_4}) - Q_3(\cos_{100}) < 0)$	47.0625
$p(\chi^2)$	.5569

Tableau n°10 : Similarités cosinus moyennes des clusters d'activation par classification hiérarchique (couche 1).



Graphe n°8 : Similarités cosinus moyennes des clusters d'activation par classification hiérarchique (couche 1).

Dans le cadre de notre dernière exploration de la question de l'existence d'un processus synthétique d'attention intra-neuronale, l'ensemble de nos résultats relatifs à l'homogénéité catégorielle des clusters activationnels tendent à indiquer une faible homogénéité catégorielle des clusters activationnels, assortie d'une progressive évolution positive de l'homogénéité catégorielle de ces clusters au fur et à mesure de la croissance des valeurs activationnelles.

5 Discussion des résultats et conclusion

5.1 Synthèse & interprétation des résultats obtenus

La question qui nous a intéressée dans le cadre de la présente étude a consisté en la mesure dans laquelle le niveau d'activation particulier d'un neurone exposé à un token est lié à une valeur catégorielle spécifique de ce token au sein de la catégorie propre à ce neurone. Pour le dire dans une logique topologique et plus détaillée, un segment donné de valeurs d'activations d'un neurone est-il corrélé à un segment catégoriel délimité, permettant ainsi de façon opérationnelle à ce neurone de focaliser électivement son attention intra-neuronale sur un segment catégoriel déterminé sur la base de la zone d'activation spécifique à ce segment ? Autrement dit encore, d'un point de vue épistémologique et fonctionnel, l'espace d'activation d'un neurone est-il compartimenté en zones activationnelle dont la signification et la fonction sont de rendre possible la détection attentionnelle de zones catégorielles spécifiques ?

Notre démarche méthodologique a été double :

- Une approche « top-down », dans laquelle nous sommes parti de segments catégoriels afin d'examiner leurs valeurs activationnelles moyennes respectives.

- Une approche « bottom-up », pour laquelle nous avons dans un premier temps isolé des segments activationnels afin d'investiguer par la suite leurs homogénéités catégorielles propres. Étant entendu, à nouveau, que nous avons ici réalisé le choix singulier d'opérationnaliser la mesure de cette homogénéité catégorielle à partir du seul référentiel d'observation constitué par le système d'embeddings de GPT2-XL, et que d'autres choix auraient pu être possibles et pertinents.

Dans une logique « top-down », une première étude a investigué à quel point il existait une différence significative de moyennes activationnelles entre clusters catégoriels. Il en est ressorti une double tendance quant au sujet de l'existence d'un potentiel processus synthétique d'attention intra-neuronale, rendant possible pour un neurone le fait d'identifier et de localiser, à l'intérieur de la catégorie de pensée artificielle qu'il porte, certains segments catégoriels spécifiques à partir de leurs empans activationnels respectifs :

- Une indifférenciation activationnelle (i.e. une faible différence de valeurs moyennes d'activation) entre les clusters catégoriels associés à de plus faibles valeurs moyennes d'activation.
- Une différenciation activationnelle relative (i.e. une plus forte différence de valeurs moyennes d'activation) entre les clusters catégoriels impliquant une ou plusieurs plus forte(s) valeur(s) moyenne(s) d'activation.

Toujours dans le cadre de notre investigation « top-down » du potentiel phénomène de l'attention intra-neuronale, nous avons mis à jour un entrelacement activationnel des clusters catégoriels ; à savoir le fait que des segments catégoriels donnés ne sont pas compartimentés selon des segments activationnels distincts, mais au contraire se chevauchant. Et nous avons indiqué que ce résultat est a priori compatible avec l'indifférenciation activationnelle (entre clusters catégoriels à plus faibles valeurs d'activation moyennes) ; et a priori non incompatible, en l'état, avec la différentiation activationnelle (entre les clusters catégoriels à plus fortes valeurs moyennes d'activation).

Enfin, au niveau de notre étude « bottom-up » de la question de l'existence d'un processus synthétique d'attention intra-neuronale, nos résultats tendent à manifester une faible homogénéité catégorielle des clusters activationnels, associée à une relative progressive évolution positive de cette homogénéité en fonction de l'augmentation de la valeur moyenne d'activation de ces segments activationnels.

L'ensemble de ces résultats tendent à être compatibles avec l'existence d'un phénomène d'attention intra-neuronale, semblant pouvoir être caractérisée comme suit : (i) une absence de relation entre segmentation activationnelle d'une part et catégorielle d'autre part pour des tokens à forts niveaux d'activation (car les tokens impliqués ici sont des core-tokens, c'est-à-dire des tokens fortement activés en moyenne), (ii) une relation ténue mais systématique entre segmentations activationnelle et catégorielle pour des tokens à très forts niveaux d'activation. Autrement dit, l'activation jouerait un rôle de vecteur attentionnel intra-neuronal : elle permettrait à un neurone de repérer et de délimiter, parmi l'ensemble des tokens constitutifs de son extension catégorielle,

certains tokens, ceux qui sont spécifiquement associés aux plus fortes valeurs activationnelles ; tokens relativement plus homogènes entre eux d'un point de vue catégoriel et constituant de facto un segment catégoriel particulier et présentant un intérêt spécifique. Cette attention intra-neuronale, en rendant possible une focalisation attentionnelle sur les seules très fortes activations, opérerait ainsi une dichotomisation attentionnelle, c'est-à-dire, à partir d'un continuum quantitatif d'activations, un seuil qualitatif d'activation (i.e. les très fortes activations) à partir duquel l'attention sélective intra-neuronale doit opérer. Cette attention intra-neuronale étant ainsi directement liée à des activités de vigilance [60, 21, 46, 62] and selective attention processes [55, 11, 112, 53].

5.2 Articulation de nos résultats actuels et antérieurs

La non présence d'une relation entre segmentation activationelle et segmentation catégorielle pour les tokens à forts niveaux d'activation est largement compatible avec le processus synthétique de divergence catégorielle postulé dans le cadre d'une de nos investigations précédentes [67]., divergence catégorielle renvoyant aux deux phénomènes cognitifs synthétiques suivants :

1. La discontinuité catégorielle des core-tokens successifs quant à leur niveau d'activation, à savoir d'existence de cosinus similarité particulièrement faibles entre core-tokens successifs.
2. L'inhomogénéité catégorielle des core-tokens à mêmes valeurs d'activation, c'est-à-dire le fait que les core-tokens ayant les mêmes niveaux d'activation ne sont pas catégoriellement les plus proches.

Phénomène de divergence catégorielle pouvant être associé aux aspects polysémantiques des concepts neuronaux [64, 39, 9, 16]à la différence des approches catégorielles classiques humaines [49, 149, 150, 151, 152, 153, 154, 155].

Inversement, la corrélation relative obtenue entre segmentations activationelle et catégorielle pour les tokens à très hautes valeurs d'acticonvergence catégorielle postulant le fait que plus les niveaux d'activation des core-tokens successifs augmentent et plus la variabilité catégorielle de ces mêmes core-tokens diminue. Cela, en phase avec la caractéristique centrale des catégories de pensée d'être ad hoc [5, 156, 153], c'est-à-dire de remplir une finalité dont on peut imaginer qu'elle puisse être, pour les catégories synthétiques, de se phaser a minima avec des catégories de pensée humaines et donc de partiellement converger vers des éléments de sens humains. Précisons que les tokens spécifiquement impliqués par cette activation de très haut niveau et cette convergence catégorielle pourraient être interprétés comme relevant de phénomènes relatifs à des prototypes catégoriels [149, 150].

Dans une autre étude précédente [69] nous avons identifié trois facteurs mathématico-cognitifs de la segmentation catégorielle opérée par chaque neurone synthétique :

- L'amorçage catégoriel (effet x), relevant du fait qu'un token activant fortement un neurone de couche n a une probabilité plus élevée d'activer

un neurone fortement relié de couche $n+1$. L'amorçage catégoriel étant ici défini dans le prolongement de son corolaire humain [113, 114, 115, 116].

- L'attention inter-catégorielle (effet$w$), i.e. le fait que plus un neurone de couche $n+1$ est fortement relié à un neurone de couche n,et plus un token fortement activé dans le premier a de probabilités de l'être également dans le neurone de couche $n+1$.
- Le phasage catégoriel (effet Σ), associé au fait que des tokens activant simultanément différents neurones de couche n ont une probabilité plus élevée d'activer un neurone fortement relié de couche $n+1$. La notion de phasage catégoriel étant ici pensée en lien avec ses notions homologues dans les domaines de la psychologie cognitive et des neurosciences humaines [54, 117, 118, 119, 120, 121, 122, 123].

Ces trois facteurs de la compartimentation catégorielle neuronale sont ceux qui président à la détermination de la valeur d'activation associée à un token donné. Par construction mathématique de la fonction d'agrégation, un effet fort et conjoint de ces trois facteurs, relativement à un token donné, apparie une très forte valeur d'activation à ce token. Et dès lors, un positionnement de ce dernier au sein du segment, pour un neurone donné, des tokens à très fortes valeurs activationnelles. Ces facteurs que sont l'amorçage, l'attention et le phasage catégoriels sont dès lors ceux qui rendent possible et pilotent l'attention intra-neuronale, celle-ci se traduisant par une focalisation attentionnelle élective sur les tokens à très forte activation, ces derniers définissant un segment catégoriel spécifique et particulièrement pertinent pour le neurone impliqué.

Dans le cadre d'une autre recherche préalable [70], nous avons mis en lumière un mécanisme de détourage catégoriel synthétique, consistant, au niveau neuronal, en la séparation d'une forme d'un fond catégoriel [141, 142, 143]. Plus en détail, le détourage catégoriel est le processus à travers lequel une sous-dimension catégorielle spécifique est extraite [144, 145, 146, 147] de la catégorie vectorisée par un neurone précurseur (en couche n) afin de contribuer à la détermination d'une catégorie superordonnée (en couche $n+1$).Ce détourage est, nous l'avons vu, directement piloté par les facteurs mathématico-cognitifs que nous venons de mentionner. Et la sous-dimension catégorielle ainsi détourée est de façon immédiate déterminée par le processus d'attention intra-neuronale ; c'est en effet cette dernière qui, par construction mathématique de la fonction d'agrégation neuronale, va permettre une isolation attentionnelle sur un segment de tokens, ceux à très fortes valeurs d'activation, desquels vont être statistiquement extraits les tokens qui constitueront spécifiquement les sous-dimensions catégorielles détourées.

Nous avons également montré [71] le fait que la coactivité de l'attention catégorielle et du phasage catégoriel génère un phénomène synthétique de confluence catégorielle partielle ; à savoir que les sous-dimensions catégorielles extraites des catégories de couche n en lien avec un neurone qui leur est fortement lié au niveau attentionnel en couche $n+1$, font montre d'une tendance à converger sémantiquement pour partie. La confluence catégorielle est directement impactée par l'attention intra-neuronale dans la mesure où cette attention permet une

identification et une sélection des tokens singuliers, ceux à très fortes activations, à partir desquels va être opéré le phénomène de phasage catégoriel et dès lors de confluence catégorielle.

5.3 Attention intra-neuronale et conceptualisation

Ainsi que nous l'avons abordé, la conceptualisation [100, 101] est l'identification de caractéristiques opératoires spécifiques au sein d'une classe d'objets (dans notre présent cas, des tokens) sur lesquels un système cognitif doit agir, afin que cette action soit adaptée et dès lors efficace. Le processus de conceptualisation est résolument un phénomène attentionnel dans le sens où il permet une focalisation sélective de l'attention sur un sous-ensemble limité d'objets ou de caractéristiques particulières de ces objets, afin de calibrer l'activité en phase avec les propriétés particulières de ces objets et caractéristiques ainsi identifiées.

Un neurone artificiel peut être interprété comme un agent cognitif synthétique de conceptualisation, dont la finalité est de sélectionner attentionnellement, à partir d'un ensemble de tokens auxquels il réagit fortement, un sous-ensemble spécifique de tokens ; sous-groupe qui va être constitutif de l'extension catégorielle du concept-en-acte « critique » que ce neurone a pour fonction de repérer de façon élective ; ce neurone réalisant dès lors une activité de focalisation attentionnelle sur certains types de tokens qu'il s'agit de sélectionner et de filtrer afin de rendre efficace l'activité de traitement langagier à laquelle participe ce neurone. Le concept-en-acte paroxystique ainsi déterminé est constitué de tokens à très forts niveaux d'activation. Ces tokens particuliers, objets d'une différenciation activationnelle, tendent à converger sémantiquement (homogénéité catégorielle) vers une caractéristique catégorielle pertinente pour le neurone impliqué. La différenciation activationnelle du segment catégoriel repéré par ce concept-en-acte synthétique critique est intimement lié à l'effet d'amorçage catégoriel (effet x), qui, combiné au phasage catégoriel et à l'attention inter-catégorielle, vont rendre possible le détourage catégoriel ainsi que la confluence catégorielle. La conceptualisation et l'attention intra-neuronales, opérés au niveau d'un neurone de couche n, sont dès lors les processus synthétiques fondamentaux qui vont ensuite rendre possible, au niveau ultérieur de la couche$n + 1$, la constitution de nouvelles catégories de pensée, encore plus fonctionnelles pour réaliser les tâches assignées à un réseau de neurones, tâches pour lesquelles il a été entraîné.

6 Conclusion

Dans le cadre de cette présente étude nous avons investigué dans quelle mesure il existe, au niveau des neurones de type perceptron des modèles de langage, un processus synthétique de conceptualisation et d'attention intra-neuronales, permettant à chaque neurone de repérer et d'isoler, au sein de la catégorie de pensée artificielle qu'il porte, un segment catégoriel spécifique à partir de son espace d'activation. Cela, en lien avec la question de savoir si les neurones formels sont les porteurs, en leur sein, d'une relative relation homomorphique

entre segmentation activationnelle et segmentation catégorielle ; et ainsi penser la fonction et le sens catégoriel qu'il est pragmatiquement et épistémologiquement possible d'accorder à la notion d'activation. Notre étude tend à répondre que cette relation semble exister, de façon ténue mais systématique, au niveau des seuls tokens à très forts niveaux d'activation. Cette attention intra-neuronale segmente, à l'intérieur même d'un neurone donné d'une couche n, une zone activationnelle associée à un concept-en-acte neuronal particulier ; concept-en-acte synthétique à partir duquel les processus de restructuration catégorielle (dont le détourage catégoriel et la confluence catégorielle) vont alors pouvoir être conduits au niveau des neurones de couche $n + 1$, et ainsi piloter la genèse d'abstractions catégorielles de plus haut niveau, constitutives des catégories de pensée de ces neurones superordonnés.

Notons à nouveau, pour terminer, que le phénomène de plus grande homogénéité catégorielle des tokens à très forts niveaux d'activation, phénomène qui est central dans notre présente mise en lumière du processus d'attention et de conceptualisation intra-neuronales, a été ici opérationnalisé à partir du système d'embeddings d'entrée de GPT2-XL. Et que le processus de différentiation activationnelle que nous avons pointé a été déterminé à partir du modèle GPT4o. L'avantage de cette méthodologie est d'étudier le processus d'attention intra-neuronale à partir de référentiels d'observation sémantiques analogues ou en tout cas relativement proches de catégories de pensée humaines, ce qui semble pertinent dans la mesure où il s'agit d'investiguer dans quelle mesure cette attention intra-neuronale permet à un modèle de langage d'être ajusté à des tâches humaines, à l'endroit desquelles il doit dès lors réaliser en partie une activité de conceptualisation ainsi qu'une mobilisation de catégories de pensée en phase avec des catégories humaines de pensée. Mais, d'un autre point de vue épistémologique, il peut néanmoins sembler biaisé, autocentré et anthropomorphique de raisonner ainsi. En effet, il pourrait être pertinent de nous intéresser à la question de la relation attentionnelle, au sein d'un neurone donné, entre segmentation activationnelle et catégorielle en utilisant comme référentiel d'observation sémantique les embeddings propres à l'entrée de chaque couche neuronale respectivement impliquée. La valeur ajoutée de cette nouvelle approche, plus proche des catégories de pensée singulières et propres à chaque couche neuronale, serait d'être plus ajustée au monde catégoriel et aux « aliens concepts » spécifiques à chaque neurone formel. Cette alternative méthodologique pourrait éventuellement, dès lors, mettre en lumière une autre phénoménologie de l'attention et de la conceptualisation intra-neuronales, pouvant par exemple se traduire par une homogénéité catégorielle plus forte au niveau des tokens à très fortes activation ; voire par une relation homomorphique entre segmentations activationnelle et catégorielle plus étendue que celle que nous avons ici manifesté au seul niveau des très fortes activations neuronales.

Remerciements

Les auteurs remercient Madeleine Pichat pour sa relecture attentive du présent article. Ainsi que Chantal Colle pour les projets stimulant portés avec elle en matière d'intelligence artificielle.

Bibliographie

[1] Alahmari, S. S., Gardner, M. R., & Salem, T. (2024). Attention guided approach for food type and state recognition. *Food and Bioproducts Processing*.

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

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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`

[10] 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

[11] Bosker, H. R., & Frost, R. L. A. (2024). Statistical learning at a virtual cocktail party. *Psychonomic Bulletin & Review*, 31, 849-861.

[12] Bosker, H. R., & Frost, R. L. A. (2024). Statistical learning at a virtual cocktail party. *Psychonomic Bulletin & Review*, 31, 849-861.

[13] Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology*, 11(2), 177-220.

[14] Brewer, W. F., & Hay, A. E. (1984). Reconstructive recall of linguistic style. *Journal of Verbal Learning and Verbal Behavior*, 23(2), 237-249.

[15] Brewer, W. F., & Hay, A. E. (1984). Reconstructive recall of linguistic style. *Journal of Verbal Learning and Verbal Behavior*, 23(2), 237-249.

[16] 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`

[17] Broadbent, D. E., & Gregory, M. (1965). Effects of noise and of signal rate upon vigilance analysed by means of decision theory. *Human Factors*, 7(2), 155-162.

[18] Deutsch, J. A. (1958). Perception and Communication. *Nature*, 182(4649), 1572-1572.

[19] 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.

[20] Cave, K. R., & Wolfe, J. M. (1990). Modeling the role of parallel processing in visual search. *Cognitive Psychology*, 22(2), 225-271.

[21] Chen, T., Zhang, Y., Wang, H., Liu, J., & Li, X. (2024). Electrophysiological correlation between executive vigilance and attention network based on cognitive resource control theory. *International Journal of Psychophysiology*, 203, 112393.

[22] Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, 25, 975-979.

[23] Child, R., Gray, S., Radford, A., & Sutskever, I. (2019). Generating long sequences with sparse transformers. *arXiv preprint arXiv :1904.10509*. `https://arxiv.org/abs/1904.10509`

[24] Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8(2), 240–247. `https://doi.org/10.1016/s0022-5371(69)80069-1`.

[25] Conrad, C. (1972). Cognitive economy in semantic memory. *Journal of Experimental Psychology*, 92(2), 149-154.

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

[27] 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

[28] 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

[29] 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.

[30] 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

[31] Duncan, J. (1999). Attention. In R. A. Wilson & F. C. Keil (Eds.), *The MIT Encyclopedia of Cognitive Sciences*. Cambridge, MA : MIT Press.

[32] Duncan, J., & Humphreys, G. (1992). Beyond the search surface : Visual search and attentional engagement. *Journal of Experimental Psychology : Human Perception and Performance*, 18(2), 578-588.

[33] 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.

[34] Ellis, P. (2010). *The essential guide to effect sizes*. Cambridge University Press.

[35] Ellis, P. (2020). *Effect Size Matters : How Reporting and Interpreting Effect Sizes Can Improve Your Publication Prospects and Make the World a Better Place !* London : MadMethods.co.

[36] Efimov, A., Dubrovsky, D., & Matveev, F. (2023). What's stopping us achieving artificial general intelligence ? *Philosophy Now*, April/May.

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

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

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

[40] Fel, J., Smith, A., & Wang, T., "A Holistic Approach to Unifying Automatic Concept Extraction and Concept Importance Estimation," in *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2024.

[41] 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.

[42] 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.

[43] Giallanza, T., & Campbell, D. I. (2024, March). Context-Sensitive Semantic Reasoning in Large Language Models. In *ICLR 2024 Workshop on Representational Alignment*.

[44] Gernsbacher, M. A. (1985). Surface information loss in comprehension. *Cognitive Psychology*, 17(3), 324-363.

[45] 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.

[46] Hanzal, S., Müller, C., Schwarz, J., Binder, L., & Schröder, P. (2024). EEG markers of vigilance, task-induced fatigue and motivation during sustained attention : Evidence for decoupled alpha-and beta-signatures. *bioRxiv*, 2024-10.

[47] 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.

[48] Hastie, R. (2022). Schematic principles in human memory. *Social Cognition*, 39-88.

[49] Hornsby, A. N., & Love, B. C. (2020). How decisions and the desire for coherency shape subjective preferences over time. *Cognition*, 200, 104244. `https://doi.org/10.1016/j.cognition.2020.104244`

[50] Howell, D. C. (2024). *Méthodes statistiques en sciences humaines*. De Boeck Supérieur.

[51] von Humboldt, W. (1907). *Werke*, vol. 7, part 2. Berlin : Leitmann.

[52] Kahneman, D. (1973). *Attention and Effort*. Englewood Cliffs, NJ : Prentice-Hall.

[53] Lin, Z. (2024). Attenuation Theory. In *The ECPH Encyclopedia of Psychology* (pp. 1-2). Singapore : Springer Nature Singapore.

[54] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2020). Swin Transformer : Hierarchical Vision Transformer using Shifted Windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. `https://arxiv.org/abs/2103.14030`

[55] Liu, H., Zhang, Y., Wang, F., Li, J., & Chen, T. (2024). Electrophysiological correlation of auditory selective spatial attention in the “cocktail party” situation. *Human Brain Mapping*, 45(11), e26793.

[56] Mackworth, N. H. (1948). The breakdown of vigilance during prolonged visual search. *Quarterly Journal of Experimental Psychology*, 1(1), 6-21.

[57] Maturana, H. (1970). *Biology of cognition* (Vol. 9). Urbana : Biological Computer Laboratory, Department of Electrical Engineering, University of Illinois.

[58] 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.

[59] Moreira, G., Hauptmann, A., Marques, M., & Costeira, J. P. (2024). Learning Visual-Semantic Subspace Representations for Propositional Reasoning. *arXiv preprint*, arXiv :2405.16213.

[60] Motter, B. C. (1999). Attention in the animal brain. In R. A. Wilson & F. C. Keil (Eds.), *The MIT Encyclopedia of the Cognitive Sciences* (pp. 39–41). Cambridge, MA : MIT Press.

[61] 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.

[62] Murray, S. (2024). The Nature and Norms of Vigilance. *American Philosophical Quarterly*, 61(3), 265-278.

[63] Ortiz-Rodriguez, F., Tiwari, S., Panchal, R., Medina-Quintero, J. M., & Barrera, R. (2022, June). MEXIN : multidialectal ontology supporting NLP approach to improve government electronic communication with the Mexican Ethnic Groups. In *DG.O 2022 : The 23rd Annual International Conference on Digital Government Research* (pp. 461-463).

[64] Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom In : An Introduction to Circuits. Retrieved from `https://distill.pub/2020/circuits/zoom-in/`. Accessed 24-11-2023.

[65] Patel, A. S., Merlino, G., Puliafito, A., Vyas, R., Vyas, O. P., Ojha, M., & Tiwari, V. (2023). An NLP-guided ontology development and refinement approach to represent and query visual information. *Expert Systems with Applications*, 213, 118998.

[66] Piaget, J. (1974). *La prise de conscience*. Paris : Presses Universitaires de France.

[67] Pichat, M. (2024). 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`

[68] 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`

[69] Pichat, M., Campoli, E., Pogrund, W., Wilson, J., Veillet-Guillem, M., Melkozerov, A., Gasparian, A., Pichat, P., Poumay, J. (2024). *Neuropsychology of AI : Relationship Between Activation Proximity and Categorical Proximity Within Neural Categories of Synthetic Cognition*. arXiv preprint arXiv :2410.11868.

[70] Pichat, M., Pogrund, W., Gasparian, A., Pichat, P., Demarchi, S., & Veillet-Guillem, M. (2024). How Do Artificial Intelligences Think ? The Three Mathematico-Cognitive Factors of Categorical Segmentation Operated by Synthetic Neurons. *arXiv preprint*, arXiv :2501.06196.

[71] Pichat, M., Pogrund, W., Gasparian, A., Pichat, P., Demarchi, S., Veillet-Guillem, M., Corbet, M., & Dasilva, T. (2025). The Process of Categorical Clipping at the Core of the Genesis of Concepts in Synthetic Neural Cognition. arXiv e-prints, arXiv-2502.

[72] Pichat, M., Pogrund, W., Pichat, P., Gasparian, A., Demarchi, S., Corbet, M., Georgeon, A., Dasilva, T., & Veillet-Guillem, M. (2025). *Synthetic Categorical Restructuring Or How AIs Gradually Extract Efficient Regularities from Their Experience of the World*. arXiv :submit/6232664 [cs.AI], 25 Feb 2025.

[73] Planchuelo, C., Hinojosa, J. A., & Duñabeitia, J. A. (2024). The nature of lexical associations in a foreign language : valence, arousal and concreteness. *Bilingualism : Language and Cognition*, 1-10.

[74] Polyn, S. M. (2024). 15 Attribute Theories of Memory. In M. J. Kahana & A. D. Wagner (Eds.), *The Oxford Handbook of Human Memory, Two Volume Pack : Foundations and Applications* (p. 417). Oxford University Press.

[75] Ponomarev, A., & Agafonov, A. (2022, November). Ontology concept extraction algorithm for deep neural networks. In *2022 32nd Conference of Open Innovations Association (FRUCT)* (pp. 221-226). IEEE.

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

[77] 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

[78] Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32(1), 3-25.

[79] Posner, M. I., Snyder, C. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology : General*, 109(2), 160.

[80] Posner, M. I., & DiGirolamo, G. J. (1998). Executive Attention : Conflict, Target Detection, and Cognitive Control. In R. Parasuraman (Ed.), *The Attentive Brain* (pp. 401-423). Cambridge, MA : MIT Press.

[81] Posner, M. I., & Rafal, R. D. (1995). Inhibition of return : Neural basis and function. Cognitive Neuropsychology, 12(3), 505–524.

[82] Posner, M. I. (2024). Orienting of attention and spatial cognition. *Cognitive Processing*, 25(Suppl 1), 55-59.

[83] Qiu, Q., Huang, Z., Xu, D., Ma, K., Tao, L., Wang, R., ... & Pan, Y. (2023). Integrating NLP and Ontology Matching into a Unified System for Automated Information Extraction from Geological Hazard Reports. *Journal of Earth Science*, 34(5), 1433-1446.

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

[85] Rosch, E. (1978). Cognition and categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and Categorization*. Lawrence Erlbaum Associates.

[86] Rosenholtz, R. (2024). Visual Attention in Crisis. *Behavioral and Brain Sciences*, 1-32.

[87] Rueda, M. R. (2024). Developing the attentive brain : Contribution of cognitive neuroscience to a theory of attentional development. *Human Development*, 1-16.

[88] Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory. *Psychological Review*, 81(3), 214-241.

[89] Sartori, G., Coltheart, M., Miozzo, M., & Job, R. (2024). Category Specificity and Informational Memory, *Memory*, 1, 604.

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

[91] 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

[92] Thukral, A., Dhiman, S., Meher, R., & Bedi, P. (2023). Knowledge graph enrichment from clinical narratives using NLP, NER, and biomedical ontologies for healthcare applications. *International Journal of Information Technology*, 15(1), 53-65.

[93] Treisman, A. (1964). Monitoring and storage of irrelevant messages in selective attention. *Journal of Verbal Learning and Verbal Behavior*, 3(6), 449-459.

[94] 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

[95] Treisman, A. (1986). Features and objects in visual processing. *Scientific American*, 255(5), 114B-125.

[96] Treviso, M., Lee, J. U., Ji, T., Van Aken, B., Cao, Q., Ciosici, M. R., & Schwartz, R. (2023). Efficient methods for natural language processing : A survey. *Transactions of the Association for Computational Linguistics*, 11, 826-860.

[97] Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4), 327-352.

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

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

[100] 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.

[101] 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`.

[102] Vergnaud, G. (2020). A Classification of Cognitive Tasks and Operations of Thought Involved in Addition and Subtraction Problems. In P. Carpenter, M. Moser & A. Romberg (Eds.), *Addition and Subtraction : A Cognitive Perspective*. London : Routledge.

[103] 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.

[104] von Glaserfeld, E. (2002). *Radical Constructivism*. London : Routledge Falmer.

[105] von Humboldt, W. (1907). *Werke* (Vol. 7, Part 2). Berlin : Leitmann.

[106] 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.

[107] 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

[108] Wu, D., & Zhang, S. (2024). Does visual attention help ? Towards better understanding and predicting users' good abandonment behavior in mobile search. *Library Hi Tech*, 42(3), 867-884.

[109] Yang, Y., Li, L., de Deyne, S., Li, B., Wang, J., & Cai, Q. (2024). Unraveling lexical semantics in the brain : Comparing internal, external, and hybrid language models. *Human Brain Mapping*, 45(1), e26546.

[110] 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.

[111] Zhang, C., Yin, Z., & Qin, R. (2024). Attention-Enhanced Co-Interactive Fusion Network (AECIF-Net) for automated structural condition assessment in visual inspection. *Automation in Construction*, 159, 105292.

[112] Zhao, M., Xu, D., & Gao, T. (2024). From Cognition to Computation : A Comparative Review of Human Attention and Transformer Architectures. *arXiv preprint arXiv :2407.01548*.

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

[114] 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.

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

[116] 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.

[117] 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

[118] 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

[119] 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.

[120] 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.

[121] 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.

[122] 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.

[123] Rzechorzek, A. (2024). Understanding Cognitive Processes : Insights from Recent Research. *Journal of Cognitive Neuroscience*. DOI : 10.1162/jocn_-a_01678.

[124] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv :1409.0473*. https://arxiv.org/abs/1409.0473

[125] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv :1406.1078*. https://arxiv.org/abs/1406.1078

[126] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*. https://arxiv.org/abs/1706.03762

[127] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, 1, 4171-4186. https://aclanthology.org/N19-1423/

[128] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI blog*. https://openai.com/research/language-unsupervised

[129] Luong, M.-T., Pham, H., & Manning, C. D. (2015). Effective Approaches to Attention-based Neural Machine Translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1412–1421. https://doi.org/10.18653/v1/D15-1166

[130] Voita, E., Talbot, D., Moiseev, F., Sennrich, R., & Titov, I. (2019). Analyzing multi-head self-attention : Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv :1905.09418*. https://arxiv.org/abs/1905.09418

[131] Michel, P., Levy, O., & Neubig, G. (2019). Are sixteen heads really better than one ? An empirical study of attention in Transformers. *arXiv preprint arXiv :1905.10650*. https://arxiv.org/pdf/1905.10650

[132] Tiberi, L., Mignacco, F., Irie, K., & Sompolinsky, H. (2024). Dissecting the Interplay of Attention Paths in a Statistical Mechanics Theory of Transformers. *arXiv preprint arXiv :2405.15926*. https://arxiv.org/abs/2405.15926

[133] Kitaev, N., Kaiser, Ł., & Levskaya, A. (2020). Reformer : The Efficient Transformer. *International Conference on Learning Representations*. https://openreview.net/forum?id=rkgNKkHtvB

[134] Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontañón, S., Pham, P., Ravula, A., Wang, Q., Yang, L., & Ahmed, A. (2020). Big Bird : Transformers for Longer Sequences. *Advances in Neural Information Processing Systems*, 33, 17283–17297. https://proceedings.neurips.cc/paper/2020/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf

[135] Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2022). Efficient Transformers : A Survey. *ACM Computing Surveys*, 55(6), 1–28. https://doi.org/10.1145/3530811

[136] Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What Does BERT Look at ? An Analysis of BERT's Attention. *Proceedings of the 2019 ACL Workshop BlackboxNLP : Analyzing and Interpreting Neural Networks for NLP*, 276-286. https://aclanthology.org/W19-4828/

[137] Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-XL : Attentive language models beyond a fixed-length context. *arXiv preprint arXiv :1901.02860*. https://arxiv.org/abs/1901.02860

[138] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An Image is Worth 16x16 Words : Transformers for Image Recognition at Scale. *International Conference on Learning Representations*. https://arxiv.org/abs/2010.11929

[139] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer : Hierarchical Vision Transformer using Shifted Windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10012–10022. https://arxiv.org/pdf/2103.14030

[140] Traylor, A., Merullo, J., Frank, M. J., & Pavlick, E. (2024). Transformer Mechanisms Mimic Frontostriatal Gating Operations When Trained on Human Working Memory Tasks. *arXiv preprint arXiv :2402.08211*. https://arxiv.org/abs/2402.08211

[141] Zeki, S. (2002). *Inner vision : An exploration of art and the brain*. Oxford University Press.

[142] Hock, R. M., et al. (2024). Effects of manipulating prefrontal activity and dopamine D1 receptor signaling in an appetitive feature-negative discrimination learning task. *Behavioral Neuroscience*.

[143] Green, I., Amo, R., & Watabe-Uchida, M. (2024). Shifting attention to orient or avoid : a unifying account of the tail of the striatum and its dopaminergic inputs. *Current Opinion in Behavioral Sciences*, 59, 101441.

[144] Ghorbani, A., Wexler, J., Zou, J. Y., & Kim, B. (2019). Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 9273–9282.

[145] Bhatt, U., Weller, A., & Moura, J. M. F. (2020). Evaluating and aggregating feature-based model explanations. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.

[146] Clark, P. (2021). Formalising concepts. *arXiv preprint arXiv :2101.05125v1*. https://arxiv.org/abs/2101.05125

[147] Ponomarev, D., & Agafonov, A. (2022). Ontology Concept Extraction Algorithm for Deep Neural Networks. https://www.researchgate.net/

publication/365833644_Ontology_Concept_Extraction_Algorithm_for_Deep_Neural_Networks

[148] 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

[149] Singh, V., Gupta, I., & Jana, P. K. (2020). An Energy Efficient Algorithm for Workflow Scheduling in IaaS Cloud. *Journal of Grid Computing*, 18(3), 357–376. https://doi.org/10.1007/s10723-019-09490-2

[150] Vogel, T., Ingendahl, M., & Winkielman, P. (2021). The architecture of prototype preferences : Typicality, fluency, and valence. *Journal of Experimental Psychology : General*, 150(1), 187–194. https://doi.org/10.1037/xge0000798

[151] Sanborn, A. N., Heller, K., Austerweil, J. L., & Chater, N. (2021). REFRESH : A new approach to modeling dimensional biases in perceptual similarity and categorization. *Psychological Review*, 128(6), 1145–1186. https://doi.org/10.1037/rev0000310

[152] Nosofsky, R. M., Meagher, B. J., & Kumar, P. (2022). Contrasting exemplar and prototype models in a natural-science category domain. *Journal of Experimental Psychology : Learning, Memory, and Cognition*, 48(12), 1970–1994. https://doi.org/10.1037/xlm0001069

[153] Love, A. H., Zdon, A., Fraga, N. S., Cohen, B., Mejia, M. P., Maxwell, R., & Parker, S. S. (2022). Statistical evaluation of the similarity of characteristics in springs of the California Desert, United States. *Frontiers in Environmental Science*, 10. https://doi.org/10.3389/fenvs.2022.1020243

[154] Poth, N., & Dolega, K. (2023). Bayesian belief protection : A study of belief in conspiracy theories. *Philosophical Psychology*, 36(6), 1182–1207. https://doi.org/10.1080/09515089.2023.2168881

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

[156] Glaser, J. I., Benjamin, A. S., Chowdhury, R. H., Perich, M. G., Miller, L. E., & Kording, K. P. (2020). Machine Learning for Neural Decoding. *eNeuro*, 7(4), ENEURO.0506-19.2020. https://doi.org/10.1523/eneuro.0506-19.2020