La restructuration catégorielle synthétique

Ou comment les IA découpent progressivement des régularités efficientes dans leur expérience du monde

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], Théo Dasilva[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

Publié sur arXiv le 25 février 2025.

Résumé

Comment les modèles de langage segmentent leur expérience interne du monde des mots afin d'apprendre progressivement à interagir de plus en plus efficacement avec lui ? Cette étude de neuropsychologie de l'intelligence artificielle s'intéresse au phénomène de restructuration catégorielle synthétique ; processus à travers lequel chaque nouvelle couche neuronale perceptron successive abstrait et combine des sous-dimensions catégorielles pertinentes des catégories de pensée de sa couche précédente ; cela, pour façonner de nouvelles catégories encore plus efficientes pour analyser et traiter l'expérience propre que réalise le système synthétique du monde extérieur langagier auquel il est exposé. Notre neurone viewer génétique, associé à cette étude, permet de visualiser le phénomène de restructuration catégorielle synthétique opérée lors du passage de la couche perceptron 0 à 1 de GPT2-XL.

1 Contexte de notre étude

1.1 Facteurs mathématico-cognitifs de la segmentation catégorielle synthétique

Dans une recherche antérieure [62], nous avons étudié les éléments mathématico-cognitifs influençant la segmentation catégorielle effectuée par les réseaux neuronaux artificiels des modèles de langage. Dans cette analyse exploratoire, nous avons investigué, de manière quantitative et qualitative, les facteurs qui influencent de manière génétique cette partition synthétique. En nous appuyant sur la fonction d'agrégation, prenant la forme $\sum(w_{i,j}\, x_{i,j}) + b$, et qui joue un rôle central dans ce processus cognitif, nous avons distingué trois facteurs, à la fois mathématiques et cognitifs, participant à cette compartimentation conceptuelle.

Premièrement, l'effet x, ou amorçage catégoriel synthétique, qui se rapporte au fait que l'activation des catégories synthétiques de pensée portées par les neurones de la couche n influe sur l'activation des catégories des neurones qui leur sont fortement connectés dans la couche $n+1$. En d'autres termes, plus un token appartient à une catégorie initiale dans la couche n (c'est-à-dire plus il active son neurone associé), plus il est susceptible d'appartenir aux catégories (fortement liées à cette catégorie précurseure) de la couche $n+1$. Ce phénomène d'amorçage des catégories préalables oriente dès lors, dans une certaine mesure, le façonnage de leurs catégories super-ordonnées (avec lesquelles elles ont un fort poids de connexion) en couche $n+1$; cela, en contribuant à la détermination des tokens qui deviennent partie intégrante de l'extension des catégories portées par les neurones de cette nouvelle couche.

Ensuite, l'effet w, ou attention catégorielle synthétique, qui concerne le fait que l'importance des poids de connexion entre un neurone cible (couche $n+1$) et ses neurones antérieurs (couche n) influe sur l'importance accordée à leurs catégories initiales associées lors de la formation de la catégorie de ce neurone cible. Cela se manifestant qualitativement par un processus de complémentation catégorielle consistant génétiquement à « apporter » à l'extension (de tokens) d'une catégorie d'arrivée une sous-dimension catégorielle spécifique et distincte extraite de chaque catégorie précurseure, apport qui est fonction de l'intensité de la focalisation attentionnelle dont chacune de cette catégorie précurseure est l'objet. La catégorie propre à un neurone d'arrivée se retrouvant ainsi constituée, sous-segment par sous-segment catégoriel antécédant sémantiquement complémentaire et extrait de l'extension de ses catégories sous-ordonnées.

Enfin, l'effet $\sum$, ou phasage catégoriel synthétique, par lequel des mêmes tokens qui se retrouvent simultanément activés au sein de différentes catégories précurseures significativement associées à un même neurone d'arrivée rentrent alors en écho catégoriel, présidant ainsi pour partie à la détermination des tokens constitutifs de l'extension de la catégorie de ce neurone d'arrivée ; processus, d'un point de vue qualitatif, se manifestant par un phénomène d'intersection catégorielle définissant génétiquement le contenu (en termes de tokens) de l'extension catégorielle de cette catégorie d'arrivée. L'extraction de

sous-dimensions catégorielles ici réalisée dans les catégories précurseures étant une extraction de sous-dimensions partiellement communes à ces catégories précurseures, et non pas une extraction de sous-dimensions différentes et complémentaires au sein de chacune de ces catégories précurseures, comme cela est le cas de l'effet w précédent.

Ces trois facteurs mathématico-cognitifs causaux de la segmentation catégorielle pilotent, au niveau d'un neurone d'arrivée (couche $n+1$), un processus d'extraction de sous-dimensions catégorielles spécifiques à partir des catégories portées par ses neurones précurseurs (couche n). Ces sous-dimensions catégorielles précurseures extraites, assemblées par la fonction d'agrégation au niveau du neurone d'arrivée, définissent génétiquement ainsi, le contenu de la catégorie de ce neurone superordonné. Ce processus synthétique d'extraction de concept, déjà abondamment examiné chez les êtres humains [9, 42, 34, 35, 6, 51, 101], constitue un sujet captivant sur le plan épistémologique et participe à l'élaboration de la « réalité » opérée par la cognition artificielle dans son interaction avec le monde des tokens qui lui sont présentés.

1.2 Le détourage catégoriel synthétique

Dans le cadre d'une autre étude préalable [63], nous avons pointé que ces trois facteurs mathématico-cognitifs de la segmentation catégorielle président à un processus de détourage catégoriel synthétique ; détourage consistant à fabriquer et à distinguer une forme d'un fond catégoriel. Il s'agissait dès lors de comprendre les propriétés de ce détourage catégoriel, réalisé sur la variabilité catégorielle relative des tokens constitutifs de l'extension de la catégorie de chaque neurone précurseur afin d'en extraire un sous-ensemble de tokens catégoriellement homogènes et alignés avec la (nouvelle) catégorie spécifique que crée et porte leur neurone d'arrivée correspondant.

Nous avons alors identifié, à titre exploratoire, plusieurs caractéristiques synthétiques de ce détourage catégoriel :

- **La réduction catégorielle**, traduisant le fait que la sous-dimension catégorielle extraite de la catégorie d'un neurone précurseur regroupe des tokens plus homogènes sémantiquement par rapport aux tokens constitutifs de l'extension de la catégorie initiale ; en tout cas à partir du référentiel sémantique d'observation constitué par les embeddings initiaux de GPT2-XL.
- **De façon corolaire, la sélectivité catégorielle**, faisant référence à l'extraction d'un ensemble significativement réduit de tokens à partir de l'ensemble des tokens constitutifs de l'extension de la catégorie de chaque neurone antérieur.
- **La séparation des dimensions initiales d'embedding**, liée au fait que le détourage opéré sur une catégorie de départ, lorsqu'observé dans le référentiel de l'espace vectoriel des embeddings de GPT2-XL, tend à se manifester par une compartimentation de ces embeddings, certains se retrouvant préférentiellement appariés à la forme (i.e. la sous-dimension)

catégorielle extraite, à la différence des autres plus associés au fond catégoriel restant dans cette catégorie de départ.

- **La segmentation de zones catégorielles des dimensions initiales d'embedding (réduites à deux dimensions)**, se donnant à voir par un éloignement relatif des centres de gravité catégorielle respectivement de la forme extraite et du fond non retenu, chacun de ces barycentres se retrouvant positionné dans des régions catégorielles différentes.

Ces éléments nous donnent à comprendre différentes propriétés cognitives synthétiques par lesquelles le détourage catégoriel crée des extractions de sous-dimensions catégorielles des neurones précurseurs, en regroupant en une forme des tokens convergeant au titre d'un segment catégoriel homogène fabriqué.

Les deux études préliminaires que nous venons de résumer proposent une série de notions explicatives permettant de nous représenter comment la cognition synthétique extrait, à partir d'une couche n, et durant le passage de cette couche n à une couche $n+1$, des sous-dimensions catégorielles singulières qui vont constituer la catégorie propre à chaque neurone d'arrivée. Mais comment comprendre et décrire plus avant les modalités cognitives synthétiques spécifiques à travers lesquelles les facteurs mathématico-cognitifs de la segmentation catégorielle (amorçage, attention et phasage) et le détourage catégoriel qui en découle façonnent la restructuration catégorielle qui est opérée lors du passage d'une couche neuronale à sa couche suivante ? Comment, plus en détail, opèrent et se combinent les processus cognitifs synthétiques mentionnés (amorçage, attention, phasage et détourage catégoriels) pour générer, à chaque nouvelle couche, une réorganisation du système de segmentation catégoriel porté par les neurones de cette nouvelle couche ? Quelles relations entretiennent les nouvelles catégories réorganisées super-ordonnées par rapport à leurs catégories génétiquement sous-ordonnées ? subordinate categories ?

2 Réflexions épistémologiques et conceptuelles autour de la notion de restructuration catégorielle synthétique

2.1 La structuration des catégories de pensée en psychologie cognitive humaine

Dans le champ de la cognition humaine, une première approche de la structuration des catégories de pensée est proposée à travers le paradigme des représentations en réseaux sémantiques [22, 80, 100]. Ici, les concepts sont représentés et stockés en mémoire à long terme sous forme de nœuds sémantiques, chaque nœud dénotant un concept. À travers des liens associatifs, chaque concept est relié à plusieurs autres concepts, chacun exprimant une caractéristique du concept impliqué. Les catégories sont alors organisées de façon hiérarchique, en fonction de leur ampleur de généralité : les concepts super-ordonnés sont spatialement situés au-dessus des sous-ordonnés. Cela, selon

une modalité d'économie cognitive : une information stockée, une fois pour toutes, à un niveau n, ne l'est pas à un autre niveau, supérieur ou inférieur ; cette conservation mnésique étant alors réalisée au degré de généralité le plus fort. Face à un objet donné à analyser, la récupération de l'information stockée en mémoire est alors réalisée via un processus d'activation diffusante [2] : l'activation d'un concept se propage aux concepts auxquels il est sémantiquement relié au sein de la structure catégorielle. Notons que cette modélisation de la structuration catégorielle a été l'objet de contradictions empiriques partielles, notamment en ce qui concerne les sujets de ressemblance sémantique [23] et de typicalité [76].

Les travaux relatifs à la structuration catégorielle par comparaison de concepts [79, 65, 64] proposent une alternative processuelle. Ils postulent que face à deux concepts à étudier quant à leur relation, une première phase est celle de l'encodage des entités qui leur sont respectivement associées (exemple : oiseau et animal). Puis une récupération de leurs traits sémantiques propres est alors engagée, concernant leurs traits définitionnels (indispensables, constitutifs de la catégorie) et leurs traits caractéristiques (traits fréquents, communs, mais non essentiels). Un examen de proximité est ensuite effectué, sur la base d'un calcul d'un indicateur de ressemblance à partir des deux séries de traits réactivés. Enfin, si jamais la valeur obtenue n'est pas suffisamment tranchée (proximité vs différence catégorielle), la comparaison est alors restreinte aux seuls traits définitionnels, permettant ainsi un arbitrage final.

De façon plus affinée mais plus complexe, d'autres recherches relatives à la structuration catégorielle en mémoire [88, 102, 38] postulent que la relation entre deux catégories est calculée en intégrant aussi bien les traits communs que différents entre ces deux catégories. Le process est alors le suivant : récupération mnésique des caractéristiques propres à chaque concept, détermination calculatoire du nombre de caractéristiques en commun, détermination du nombre de traits divergents, pondérations respectives des propriétés communes ou différentes et enfin, soustraction des deux valeurs pondérées afin de fixer un indice de proximité.

Dans une perspective assez contrastée, les études en termes de représentations propositionnelles [14, 39, 100, 52] posent que l'unité signifiante fondamentale est la proposition, qui agence entre elles les catégories. Ces propositions, stockées et récupérées en mémoire à long terme, contiennent les relations existantes entre les concepts ; cela, sous la forme d'une combinaison, plus ou moins élaborée, et pouvant avoir une valeur de vérité vraie ou fausse, de prédicats et d'arguments. Ce positionnement théorique étant assez proche épistémologiquement du paradigme des ontologies en NLP et plus largement en machine learning [66, 57, 83, 74, 56].

Mentionnons enfin les modélisations de la structuration catégorielle sous forme de représentations schématiques [12, 43]. Dans ce présent cadre théorique, des représentations à spectre large sont conservées en mémoire au sujet de concepts et d'événements du monde les concernant. Cela, sous la forme d'un schéma général épuré de synthèse, dont les scripts (organisés selon une dimension temporelle) sont les instances les plus fréquemment envisagées. La fonction de ces schémas étant, par exemple, de stocker au niveau mnésique les seules grandes étapes du déroulement relatif à une notion ou à une connaissance ; cette modalité

permettant une économie mnésique.

Les différentes perspectives que nous venons de synthétiser, dans le champ de la psychologie cognitive humaine, des modalités possibles de structuration des catégories nous donnent à prendre la mesure, si besoin il y avait, de la différence entre les cognitions catégorielles humaines et artificielles. En effet, ces différentes modélisations ne sont que peu adaptées aux spécificités de la cognition catégorielle synthétique. Cela, pour au moins les trois raisons suivantes :

- Ces modélisations sont **statiques** : elles ont pour finalité de représenter la structuration fixe des catégories de pensée ; alors que les systèmes catégoriels neuronaux synthétiques sont fondamentalement génétiques (bien que synchroniques et non diachroniques) : les couches neuronales sont des systèmes emboîtés de constitution progressive de catégories de plus en plus élaborées.
- Ces modélisations sont centrées, dans une logique aristotélicienne de tiers exclu, sur un **principe d'unicité sémantique** ; alors que les catégories synthétiques sont largement polysémiques [8, 15], en tout cas lorsque nous les étudions au sein d'un référentiel sémantique d'observation humain.
- Ces modélisations **ne prennent pas en compte l'activité d'extraction élective catégorielle** qui est celle opérée par les neurones : la catégorie de chaque neurone de couche $n + 1$ n'intègre pas, en l'état, chacune de ses catégories sous-ordonnées, mais en exfiltre certaines sous-dimensions particulières.

Ce paradigme de la structuration des catégories en psychologie cognitive humaine, dans la mesure où il a une finalité statique et non pas dynamique, n'est donc a priori pas suffisamment pertinent pour nous aider à comprendre et décrire les modalités cognitives synthétiques à travers lesquelles les facteurs mathématico-cognitifs de la segmentation catégorielle et le détourage catégoriel génèrent la restructuration catégorielle qui est réalisée lors du passage d'une couche neuronale à sa couche super-ordonnée.

2.2 Extraction de régularités catégorielles et restructuration catégorielle

Quel référentiel conceptuel psychologique mobiliser pour penser, de façon heuristique et sous condition de transposition appropriée, la combinaison des processus cognitifs synthétiques que nous avons précédemment relatés (amorçage, attention, phasage et détourage catégoriels) dans le cadre de la génération, au niveau de chaque nouvelle couche synthétique, de la restructuration du système de segmentation catégorielle porté par cette nouvelle couche ?

Une réflexion épistémologique semble s'imposer à ce stade, et nous allons la positionner dans un référentiel constructiviste, qui nous semble plus pertinent pour penser la nature de l'activité cognitive réalisée par les systèmes neuronaux artificiels, dans ses relations avec la restructuration catégorielle opérée par ces réseaux de neurones. Cette réflexion épistémologique, partant d'un cadre de

psychologie humaine pour tenter de le transposer, de façon ajustée, au cadre de la cognition synthétique.

La pensée synthétique, tout comme la cognition humaine, ne fabrique pas des représentations internes qui sont des copies analogiques, des enregistrements passifs de propriétés du monde, propriétés qui seraient pré-données et préexistantes à l'activité de leur observation. En effet, la connaissance n'est pas un miroir de la nature [90]. Cela, à l'opposé de ce que postulerait une épistémologie empiriste ou réaliste, qui se caractérise par une « tendency to think of knowledge as the representation of a world outside [...] independent of the knower. The representation [is] supposed to reflect at least part of the world's structure and the principles according to which it works » [95, p.113].

Cette position interactionniste étant posée, comment ne pas tomber dans une impasse solipsiste ? Von Glaserfeld [95, p.113], précise en effet : « if we were to say that there [is] no such relation [between knowledge and objective world], we should find ourselves caught in solipsism, according to which the mind, and the mind alone, creates the world ». La réponse réside dans le fait que l'activité intelligente n'est jamais passive mais toujours au service d'une finalité [58]. Et cette finalité est d'ajuster ses contenus et ses modalités de fonctionnement à l'expérience, propre et singulière, que le système intelligent réalise du monde extérieur : « the cognizing subject has conceptually evolved in order to fit into the world as he or she experiences it » [95, p.114].

Mais comment s'opère cet ajustement ? De façon avant tout pragmatique et non pas épistémique : « a living system, due to its circular organization, is an inductive system and functions always in a predictive manner : what happened once will occur again. Its organization (genetic and otherwise) is conservative and repeats only that which works » [50, p.39]. Il s'agit donc pour un système intelligent, dans une logique opportuniste, d'identifier ou plutôt de construire cognitivement des régularités [89], des invariants opératoires dirait Piaget [58], au sein du monde tel qu'il est expérimenté dans le cadre de nos cadres sensoriels et cognitifs humains, ou au sein de leurs corollaires synthétiques, les « techno-umwelt » [32]. Ainsi que l'indique von Glaserfeld [95, p.118] : « what we ordinarily call reality is the domain of the relatively durable perceptual and conceptual structures which we manage to establish, use, and maintain in the flow of our actual experience » ; et l'auteur rajoute : « empirical facts, from the constructivist perspective, are constructs based on regularities in a subject's experience. They are viable if they maintain their usefulness and serve their purposes in the pursuit of goals » [95, p.128].

Dans le contexte spécifique de la cognition synthétique, que sont ces régularités et qu'est-ce qui les porte au niveau interne ? La réponse à cette question, en tout cas au niveau des couches de type perceptron, est assez immédiate : ces régularités sont essentiellement les poids de connexion entre les neurones des couches successives (nous faisons ici abstraction de la question des biais de la fonction d'agrégation neuronale). Ce sont en effet ces poids qui sont l'objet de l'apprentissage du réseau de neurones durant sa phase initiale de deep learning.

Et ce sont ces poids qui vont précisément piloter l'activité spécifique de restructuration catégorielle, durant la construction d'une nouvelle couche

catégorielle neuronale, à partir de la couche catégorielle précédente. En effet, ces poids vont régenter, ipso facto, l'activité attentionnelle synthétique, consistant à identifier, durant la constitution d'une nouvelle catégorie de pensée artificielle (en couche n), les catégories antécédentes (en couche $n-1$) sur lesquelles il convient de se focaliser particulièrement et dans quelle mesure (i.e. pondération) il convient de le faire. Cela, à l'instar de ce que von Glaserfeld [95, pp.90-91], à nouveau, mentionne dans le domaine de la pensée des systèmes vivants : « focused attention picks a chunk of experience, isolates it from what came before and from what follows, and treats it as a closed entity » ; et : « for this constructive activity, the role of attention is crucial [...]. Subjects can freely move their focus of attention in the perceptual field [...] ; [attention is an] originator of coordination and relation » [95, p.116].

La fonction des poids neuronaux est fondamentalement une fonction de comparaison et de pondération du niveau d'importance attentionnelle à accorder aux catégories synthétiques précurseures dans la coordination différenciée qui en est faite, par la fonction d'agrégation, dans le cadre de la genèse de leurs catégories synthétiques correspondantes d'arrivée. Cela, étant donné que le réseau de neurones a appris qu'il est pertinent de réaliser une telle restructuration catégorielle, afin de traiter efficacement les informations qui lui sont présentées. Ainsi que le précisait Humboldt [45, p.581], dans le registre humain : « in order to reflect, the mind must stand still for a moment in its progressive activity, must grasp as a unit what was just presented, and thus posit it as object against itself. The mind then compares the units [...] and separates and connects them according to its needs ».

À l'issue de cette présente section, la notion d'attention nous apparaît comme potentiellement pertinente comme référentiel heuristique d'analyse cognitive de l'activité synthétique de restructuration catégorielle opérée par les réseaux neuronaux.

2.3 Processus attentionnels et restructuration catégorielle

Physiologiquement, l'attention émane des capacités limitées du système nerveux à traiter l'information et se manifeste par des choix dans l'intégration, l'activation et l'utilisation des données sensorielles ou mémorisées (sémantiques, procédurales) [36, 4]. Cela se traduit par une réaction d'orientation, qui consiste à focaliser la recherche d'informations sur certaines caractéristiques particulières. Mentionnons, dans ce registre, les études proprement neurocognitives des processus attentionnels, dont l'approche issue de Posner [72, 73, 78] ; celle-ci pointant l'effet d'un système attentionnel frontal, au sein du lobe frontal (associé à la focalisation sémantique consciente et à la planification), et d'un système postérieur au sein du lobe pariétal (lié aux processus visuo-spatiaux et aux variations de localisations attentionnelles).

Dans le champ de la psychologie cognitive, l'attention est définie comme un ajustement d'une activité en vue de son objectif, entraînant une meilleure efficacité dans les processus de collecte d'informations (notamment la sélectivité) et de réalisation (en termes de précision et de rapidité) pour cette activité

particulière [68, 81, 67, 75, 85, 28, 82, 24, 99, 40]. En ce qui concerne l'exécution des tâches, l'attention renvoie au contrôle par le système nerveux central de l'activité, comme par exemple l'assignation d'un degré d'importance (priorité, ordre, fiabilité, etc.) à certaines informations internes (connaissances, représentations, schémas) ou la supervision de la qualité de l'exécution des tâches tout au long de leur déroulement dans le temps.

Deux fonctions cognitives sont habituellement assignées aux processus attentionnels [29] :

- **La détection du signal**, impliquant principalement le mécanisme de vigilance et celui d'exploration, à des fins d'identification de l'occurrence d'un stimulus donné.
- **L'attention sélective**, permettant la considération de certains stimuli spécifiques et non pas d'autres.

La vigilance se définit par l'aptitude d'un individu à se focaliser électivement sur une série d'informations, durant un laps de temps donné, afin d'identifier l'occurrence d'une information ciblée [49]. Stimulus dont le taux d'occurrence est faible mais pour lequel il convient d'être en capacité de réagir avec célérité [20, 41]. La vigilance est négativement impactée par le niveau d'incertitude relative aux éléments qui sont l'objet de la focalisation attentionnelle [16]. La vigilance peut être modélisée comme un projecteur attentionnel [69] et elle est fonction de l'attente relative à la présence de l'élément ciblé au sein d'une localisation donnée [70, 53, 55].

À la différence de la vigilance, l'exploration (ou inspection) visuelle [102, 99] est non pas une attente « passive » de l'émergence de données ciblées, mais la recherche active d'un stimulus [71]. Typiquement, l'exploration est conçue comme une stratégie de recherche de caractéristiques spécifiques (traits, attributs) par balayage, afin de les localiser au sein d'un environnement donné. Dans le cadre de la *théorie de l'intégration des attributs* [86, 77], une carte mentale est associée à chaque attribut, cette carte constituant une représentation de l'occurrence de la caractéristique visée dans le champ visuel. Les différentes cartes sont alors examinées de façon parallèle ; et une investigation simultanée d'une série de traits relevant d'une même entité est rendue possible par un processus d'attention conçu comme une « colle mentale » ayant pour fonction de réunir au sein d'une même zone représentationnelle les divers attributs impliqués ; cette dynamique pouvant être associée à un processus d'inhibition neuronale de caractéristiques non pertinentes via des neurones ad hoc. La *théorie de la similitude* [30], quant à elle, appréhende l'inspection attentionnelle en termes de proximité entre les stimuli cibles et les éléments non pertinents, ainsi qu'en termes de proximité entre ces éléments non pertinents eux-mêmes. Enfin, la *théorie de l'inspection guidée* [19, 1] conçoit l'exploration attentionnelle en deux moments : (i) une première phase d'activation d'une représentation conjointe de toutes les cibles potentielles, sur la base des caractéristiques spécifiques de la cible, (ii) une seconde phase, analysant de façon sérielle le niveau d'activation de toutes les cibles éventuelles afin de retenir celle dotée de la plus forte activation.

L'attention sélective, relevant d'une focalisation attentionnelle sur certains

éléments spécifiques, est notamment étudiée via le paradigme de l'*effet cocktail party* [21, 48, 10], relatif au suivi électif d'une conversation particulière parmi d'autres conversations. Trois paramètres de cette focalisation attentionnelle sur le discours du locuteur cible sont identifiés : (i) ses propriétés sensorielles particulières, (ii) son intensité sonore et (iii) son positionnement géographique. La *théorie du filtre attentionnel* [17, 103] postule que les informations en provenance d'une variété de vecteurs sensoriels parviennent à un filtre sélectionnant celles qui seront l'objet d'un traitement perceptif à proprement parler. Toutefois, la version initiale de cette approche tend à être supplantée par la *théorie de l'atténuation* [84, 47] postulant que toutes les informations sont diminuées quant à leur intensité perceptive, et que dès lors seules celles dont la densité est suffisamment importante peuvent demeurer ; à savoir celles qui sont les moins atténuées étant donnée leur proximité avec les caractéristiques spécifiquement ciblées. Citons enfin la *théorie des ressources attentionnelles limitées* [46], dans le cadre de la conduite parallèle d'activités, en lien avec une facilitation contrastée.

L'attention sélective est directement liée au process de conceptualisation [91, 92, 61]. En effet, lorsqu'il s'agit d'acquérir de façon élective des informations, l'attention est associée à la création de concepts, impliquant l'identification des seuls éléments pertinents (associés aux objets concernés par l'activité) nécessaires à la réussite de cette activité ; ces éléments permettant ainsi d'orienter l'action pour qu'elle soit correctement ajustée et donc efficace. Ici, l'attention relève d'un tri et d'une structuration d'une abondance d'informations perçues disponibles, et dès lors d'une exclusion (ou inhibition) de celles considérées comme peu pertinentes, afin de concentrer l'effort mental et la sélectivité sur certains objets et caractéristiques dont l'expérience antérieure a montré la pertinence pragmatique. Cette approche de l'attention, dans un registre de conceptualisation, semble particulièrement adaptée à la notion de poids attentionnels (perceptron) en deep learning, dans la mesure où il s'agit alors d'apprendre à fabriquer en couche neuronale perceptron de niveau $n + 1$ de nouvelles catégories de pensée, en combinant de façon pondérée et élective au niveau attentionnel les catégories portées par les neurones de la couche n précédente. Cette approche conceptuelle est dès lors celle que nous mobiliserons particulièrement dans le cadre de notre présente étude du processus de restructuration catégorielle synthétique ; cela, d'autant plus, que cette approche conceptuelle est directement épistémologiquement cohérente avec l'approche d'investigation catégorielle qui est la nôtre.

3 Problématique

3.1 Extraction de régularités attentionnelles, facteurs du détourage et restructuration catégorielle

En relation immédiate avec la notion de conceptualisation, l'attention catégorielle synthétique [62], ou « effet w », est un facteur mathématico-cognitif décisif de la restructuration catégorielle, objet de notre présent travail. En effet,

les poids attentionnels neuronaux sont les régularités apprises par le système neuronal qui vont directement impacter sa focalisation attentionnelle élective, au niveau de la constitution d'une nouvelle catégorie de pensée en couche n, sur certaines catégories spécifiques en couche $n-1$. Et dès lors déterminer le détourage catégoriel [63] qui va consister à extraire une sous-dimension catégorielle [60] précise de chacune des catégories précurseures afin de constituer, via une combinaison pondérée, la nouvelle catégorie de pensée super-ordonnée réorganisée catégoriellement.

Mais, à nouveau, et cela est la finalité de l'étude que nous conduisons ici, quelles sont les modalités cognitives spécifiques par lesquelles l'attention catégorielle synthétique étaye la restructuration catégorielle opérée par les neurones d'une couche n à partir des catégories portées par les neurones de sa couche précédente ? Comment l'extraction de sous-dimensions catégorielles permise par l'attention catégorielle, constitutive de cette restructuration catégorielle, se manifeste-t-elle en termes d'activation et donc d'amorçage catégoriel ? Selon quels mécanismes l'attention catégorielle interagit-elle avec le phasage catégoriel pour générer cette restructuration ? Autrement dit, quelles sont les phénoménologies propres à travers lesquelles se manifeste la coactivité de ces trois facteurs mathématico-cognitifs dans leur façonnage du détourage et de la restructuration catégoriels synthétiques ?

Avant de présenter les choix que nous avons réalisés quant à notre investigation spécifique de ces questions à travers les opérationnalisations qui sont les nôtres, nous prenons un moment, ci-après, pour clarifier plus avant ce que nous entendons par restructuration catégorielle et définir une série d'éléments méthodologiques qui seront propres à ces opérationnalisations.

3.2 Précisions relatives au phénomène de restructuration catégorielle synthétique

Le graphe n°1 présente un cas de restructuration catégorielle synthétique, en prenant l'exemple du neurone 121 de la couche perceptron 1 de GPT2-XL, en relation avec trois de ses neurones prédécesseurs en couche 0, avec lesquels il entretient les plus forts poids de connexion (et donc de focalisation attentionnelle). Précisons que les clusters catégoriels (sous-groupes de tokens) que nous allons mentionner, concernant ce neurone d'arrivée et ses précurseurs, ne sont pas des réalités sémantiques intrinsèques, mais qu'ils sont le fruit d'une clusterisation sémantique opérée par GPT-4o.

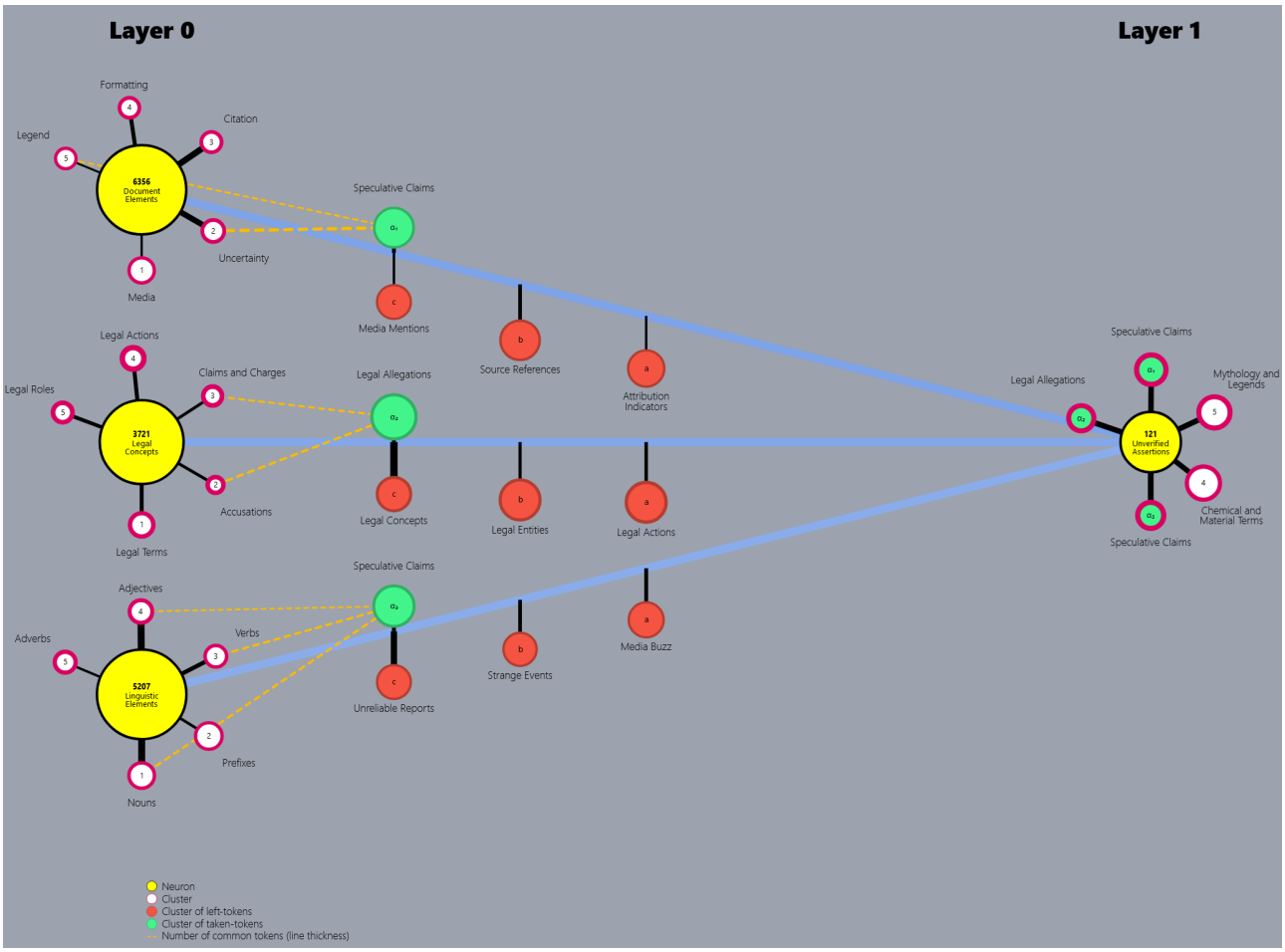
Commençons par le neurone précurseur (couche 0) n°372l. En nous centrant sur les 100 tokens qui activent le plus en moyenne ce neurone (que nous nommons les « core-tokens », dans la mesure où nous les définissons comme constituant l'extension catégorielle de la catégorie ici impliquée), nous pouvons associer ce neurone à la catégorie de pensée « legal concepts », catégorie que ce neurone est supposé détecter.

Cette catégorie peut être décomposée en 5 clusters catégoriels de tokens (parmi ces 100 core-tokens) :

- « **Legal terms** » : contenant les tokens `[patent]`, `[liability]`, `[arbitration]`, etc.
- « **Accusations** » : impliquant les tokens `[allegation]`, `[accusation]`, `[disputed]`, etc.
- « **Claims & charges** » : associés aux tokens `[billing]`, `[claim]`, `[charge]`, etc.
- « **Legal actions** » : relevant des tokens `[lawsuit]`, `[sue]`, `[embroiled]`, `[litigation]`, etc.
- « **Legal roles** » : contenant les tokens `[defendant]`, `[lawyers]`, `[plaintiff]`, `[claimant]`, etc.

Parmi ces 100 core-tokens de départ, certains activent également fortement en moyenne le neurone d'arrivée n°121 en couche 1 : `[charges]`, `[claims]`, `[accusations]`, `[allegations]`, `[alleging]`, etc. Ces tokens deviennent donc également des core-tokens constitutifs de l'extension de la catégorie de ce neurone d'arrivée. Ces tokens spécifiques, que nous nommons « taken-tokens », peuvent être interprétés comme ayant été extraits, détourés de la catégorie précurseure, afin de participer à la constitution de la catégorie d'arrivée. Ces taken-tokens forment un « taken-cluster » (en vert dans le graphe) qui a été ici interprété sémantiquement comme relevant d'une sous-dimension catégorielle « legal allegations », extraite de la catégorie de départ.

Notons que cette sous-dimension catégorielle « legal allegations » ne relève pas d'une simple sélection d'un des clusters catégoriels (« legal terms », « accusations », « claims & charges », « legal actions », « legal roles ») de la catégorie de départ « legal concepts ». Au contraire, cette sous-dimension catégorielle « legal allegations » est le fruit d'un authentique détourage catégoriel constructif consistant à extraire électivement et à regrouper certains tokens issus initialement des différents clusters catégoriels impliqués. Cette sous-dimension catégorielle singulière et originale illustre ainsi une reconstruction catégorielle, c'est-à-dire une nouvelle forme de segmentation catégorielle des objets du monde des tokens.



Graphe n° 1 : Exemple de restructuration catégorielle (cas du neurone 121 de la couche perceptron 1 de GPT2-XL, en relation avec trois de ses neurones précurseurs à poids maximum de connexion).

Ce mécanisme synthétique de construction et d'extraction d'une nouvelle sous-dimension catégorielle se produit pour chaque catégorie de départ. Chacune de ces sous-dimensions catégorielles devient à son tour un des clusters catégoriels du neurone d'arrivée, portant ici une catégorie pouvant être interprétée comme relevant de « unverified assertions ». Ainsi, lors du passage des catégories de départ à leur catégorie d'arrivée associée, via ces détourages catégoriels, nous observons un véritable processus de restructuration catégorielle opéré par ce neurone d'arrivée. Cette restructuration synthétique se traduisant par la construction de la distinction : (i) de formes catégorielles extraites (les sous-dimensions catégorielles détourées, portées par les taken-cluster, en vert dans le graphe) et (ii), d'un fond catégoriel non retenu (et se traduisant ici par les « left-clusters »,

en rouge).

Notre neurone viewer génétique, associé à cette étude, permet de visualiser le phénomène de restructuration catégorielle synthétique opérée lors du passage de la couche perceptron 0 à 1 de GPT2-XL.

3.3 La confluence catégorielle partielle synthétique

Un premier choix d'opérationnalisation conceptuelle et méthodologique de notre questionnement relatif à la restructuration catégorielle synthétique opérée durant le passage de catégories précurseures en couche n à leur catégorie d'arrivée associée en couche $n+1$ est le suivant.

Quel effet cognitif produit la coactivité de l'attention catégorielle et du phasage catégoriel synthétiques sur la restructuration catégorielle artificielle lors du passage des catégories précurseures à leur catégorie associée en couche ultérieure ? Dans quelle mesure cette interaction entre ces deux facteurs mathématico-cognitifs pilote le détourage catégoriel qui va générer cette restructuration ? Comment cela se manifeste-t-il au niveau des sous-dimensions catégorielles détourées de chacune des catégories précurseures ?

Nous postulons ici l'effet suivant : *la confluence catégorielle partielle*, pour une catégorie d'arrivée donnée en couche n, de ses sous-dimensions catégorielles génétiques extraites associées en couche $n-1$. Autrement dit, une convergence sémantique relative entre les clusters de tokens extraits (*taken-clusters*), c'est-à-dire entre les sous-dimensions catégorielles détourées de chaque catégorie de départ.

Ce phénomène de confluence catégorielle partielle repose sur la considération suivante : si l'attention catégorielle est à l'origine de l'extraction d'une sous-dimension catégorielle d'une catégorie de départ en couche $n-1$, la coordination de ce processus synthétique à celui du phasage catégoriel va mécaniquement tendre à provoquer, entre les différentes catégories de départ, un détourage de sous-dimensions sémantiquement liées. Cela, étant donné que, par construction mathématique de la fonction d'agrégation de la forme $\sum(w_{i,j}x_{i,j}) + b$, pour un token de départ donné, son activation résultante du neurone d'arrivée sera d'autant plus forte (et donc ce token sera extrait et deviendra un *taken-token*) s'il est simultanément activé au sein de plusieurs catégories de départ. Ainsi, le phasage catégoriel entraîne mécaniquement une convergence sémantique partielle.

Pour illustrer notre propos, reprenons l'exemple du graphe n°1 :

- La sous-dimension catégorielle α_1 *"spéculative claims"* contient, entre autres, les tokens : `[claim]`, `[allegedly]`, `[claimed]`, `[allege]`, `[infamous]`, `[notorious]`, `[rumor]`.
- La sous-dimension catégorielle α_2 *"legal allegations"* implique, entre autres, les tokens : `[claim]`, `[allegedly]`, `[claimed]`, `[allege]`, `[accusation]`, `[charges]`.

Ces deux sous-dimensions contiennent en commun les tokens `[claim]`, `[allegedly]`, `[claimed]`, `[allege]`, ce qui illustre une relative convergence sémantique.

3.4 La dispersion activationnelle catégorielle

Un deuxième choix d'orientation conceptuelle et méthodologique de notre investigation du phénomène de la restructuration catégorielle est celui qui suit.

Quel effet produit la coactivité de l'attention et de l'amorçage catégoriels synthétiques sur l'aspect activationnel de la restructuration catégorielle artificielle lors du passage des catégories précurseures à leur catégorie associée en couche ultérieure ? Nous postulons ici un effet de dispersion activationnelle catégorielle, à savoir qu'un cluster de taken-tokens, extrait d'une catégorie de départ, ne correspond pas à un segment continu de valeurs d'activation de ces tokens au sein du neurone de départ impliqué ; autrement dit, qu'une sous-dimension catégorielle détourée ne délimite pas un segment activationnel homogène (i.e. les taken-tokens d'un taken-cluster donné n'ont pas des valeurs d'activation proches au sein du neurone de départ concerné) ; autrement dit encore, que la restructuration catégorielle s'accompagne d'une restructuration activationnelle : la structure ou la topologie des activations des taken-tokens dans les neurones de départ n'est pas conservée.

Pourquoi postuler une telle dispersion activationnelle catégorielle ? Elle semble en effet contre-intuitive. Cela étant donné que le phénomène d'amorçage catégoriel tendrait à nous donner à penser que des tokens activant fortement un neurone précurseur devraient fortement activer son neurone d'arrivée associé ; et dès lors être *de facto* présents dans le taken-cluster extrait de la catégorie de départ. La raison est que penser les choses ainsi reviendrait à oublier l'impact significatif du processus de phasage catégoriel ; autrement dit, l'amorçage catégoriel ne suffit pas à nécessairement produire des activations fortes des neurones d'arrivée et donc à extraire des tokens. En effet, pour qu'un token devienne un taken-token, il convient qu'il active suffisamment fortement son neurone de départ (amorçage catégoriel) mais également, statistiquement en tout cas, qu'il soit l'objet d'un phasage catégoriel (avec un autre neurone de départ). Or ces deux phénomènes de la cognition synthétique sont *a priori* indépendants. Et, dès lors, deux tokens proches en termes d'activation (même forte) au sein d'un neurone de départ, ne sont pas forcément conjointement l'objet d'un processus de phasage catégoriel : statistiquement, l'un en sera l'objet et pas l'autre. Et dès lors le premier deviendra un taken-token et pas l'autre, *in fine* pas assez activé au sein du neurone d'arrivée.

De plus, pour le même type de raisons, nous postulons également une dispersion activationnelle des taken-tokens au niveau du neurone d'arrivée. Autrement dit, que des tokens d'un même taken-cluster ne sont pas forcément proches quant aux activations qu'ils provoquent au sein de ce neurone d'arrivée. Et donc qu'une sous-dimension catégorielle donnée, associée à une catégorie d'arrivée, ne délimite pas une zone activationnelle déterminée dans ce neurone d'arrivée. Cela, entre autres, étant donné que deux sous-dimensions catégorielles vont avoir tendance à contenir certains tokens en commun (cf phénomène de confluence catégorielle partielle synthétique dont nous faisons l'hypothèse), mais que cette tendance est partielle. Ainsi, deux taken-clusters (associés à un même neurone d'arrivée) auront tendance à avoir certains tokens en commun, et d'autres en différence ; les premiers tokens seront alors l'objet du processus de phasage

catégoriel, à la différence des seconds. Ce qui provoquera, par construction mathématique de la fonction d'agrégation, des discontinuités de valeurs d'activation du neurone d'arrivée. Phénomène de dispersion activationnelle des taken-tokens au niveau du neurone d'arrivée que nous postulons d'autant plus, qu'il est directement compatible avec les caractéristiques de discontinuité catégorielle des core-tokens successifs quant à leur niveau d'activation (postulant qu'il existe des cosinus similarité particulièrement faibles entre core-tokens successifs) et d'inhomogénéité catégorielle mono-activationnelle des core-tokens successifs (posant que les core-tokens ayant les mêmes niveaux d'activation ne sont pas catégoriellement les plus proches) que nous avons pointé dans des travaux précédents (Pichat et al., 2024a). Et, enfin, phénomène de dispersion activationnelle des taken-tokens au niveau du neurone d'arrivée impliquant que la restructuration catégorielle opérée au niveau d'un neurone d'arrivée n'est pas corollaire d'une structuration isomorphe de l'espace d'activation de ce neurone d'arrivée.

3.5 La distanciation catégorielle

Notre troisième et dernier choix d'opérationnalisation conceptuelle et méthodologique de la restructuration catégorielle est le suivant :

Si la coactivité des trois facteurs mathématico-cognitifs (amorçage, attention et phasage catégoriels) impliqués dans le détourage produit une restructuration catégorielle, alors cela signifie que le système de segmentation catégorielle d'une couche neuronale $n+1$ diffère de celui de la couche précédente n. Autrement dit, la segmentation catégorielle d'un neurone d'arrivée diffère de celle de chacun de ses neurones précurseurs.

Nous postulons ainsi la propriété de *distanciation catégorielle synthétique*, à savoir que la catégorie portée par un neurone d'une couche $n+1$ est distante des catégories vectorisées par ses neurones prédécesseurs en couche n. Nous étudierons empiriquement cette caractéristique en mesurant la distance sémantique entre les clusters catégoriels de *core-tokens* d'un neurone d'arrivée et ceux de ses neurones précurseurs ayant un fort poids de connexion. Cette distance sémantique sera évaluée dans le référentiel d'observation constitué par les embeddings de GPT2-XL.

4 Méthodologie

4.1 Cadre méthodologique

Afin de situer méthodologiquement notre présente étude exploratoire, nous présentons ici un bref aperçu de différentes techniques d'explicabilité. Celles-ci tentent, avec des niveaux variés d'acuité cognitive, d'élucider le contenu ou les processus informationnels des réseaux de neurones artificiels, qu'ils soient structurés en termes de couches neuronales, de groupes de couches ou de réseaux entiers.

Les recherches d'explicabilité adoptant une perspective cognitive large se concentrent sur l'examen des variations entre les entrées et les sorties afin de clarifier la relation entre les données d'entrée et les résultats d'un modèle de langage. Parmi ces méthodes, les approches basées sur les gradients évaluent l'impact de chaque caractéristique d'entrée en étudiant les dérivées partielles associées à chaque dimension [33]. Les caractéristiques des entrées peuvent être analysées à travers divers aspects tels que les traits [26], l'importance des tokens [33] ou les poids d'attention [5]. De plus, les approches basées sur les exemples explorent les changements de sortie en réponse à des variations d'entrée, notamment par la suppression, la négation, le mélange ou le masquage de tokens [3, 98, 87]. D'autres travaux se concentrent sur la cartographie conceptuelle des entrées pour estimer leur contribution aux résultats observés [18].

Les méthodes d'explicabilité ayant une plus grande acuité cognitive s'attachent à analyser les états internes du modèle, en examinant les sorties partielles ou les activations neuronales intermédiaires. Certaines recherches décomposent ainsi linéairement l'activation d'un neurone donné en fonction de ses entrées dans la couche précédente [94]. D'autres techniques visent à simplifier les fonctions d'activation pour en faciliter l'interprétation [97]. Certaines approches exploitent le lexique du modèle en projetant les connexions et représentations intermédiaires dans une matrice de correspondance [27, 37]. Enfin, certaines méthodes s'appuient sur l'analyse statistique des activations neuronales en réponse à un ensemble de données [8, 54, 31, 97, 25]. Notre étude s'inscrit précisément dans cette dernière approche.

4.2 Approche méthodologique

nous avons décidé d'examiner le modèle transformer GPT développé par OpenAI, en nous focalisant spécifiquement sur la version GPT-2XL. Cette sélection s'explique par le fait que GPT-2XL présente une complexité suffisante pour analyser des phénomènes cognitifs artificiels avancés, tout en étant moins complexe que GPT-4 ou sa version multimodale actuelle, GPT-4o. Un autre facteur ayant motivé notre décision est que, en 2023, OpenAI a fourni, via la publication de Bills et al. [8]des informations complètes sur les paramètres et valeurs d'activation neuronale du modèle, essentielles pour notre recherche.

Pour simplifier notre investigation, nous avons concentré notre analyse sur les deux premières couches perceptron de GPT-2XL, chacune comptant 6400 neurones, soit un total de 12800 neurones artificiels. Concernant les tokens et leurs valeurs d'activation dans ces neurones, nous avons choisi d'examiner, pour chaque neurone, les 100 tokens ayant les valeurs d'activation moyennes les plus élevées, que nous avons appelés « core-tokens ». Enfin, dans le cadre de cette étude du processus de restructuration catégorielle, et toujours à des fins de simplification, nous nous sommes focalisé, pour chaque neurone d'arrivée en couche 1, sur ses seuls neurones précurseurs en couche 0 à plus fort poids positif de connexion (10 précurseurs au maximum par neurone d'arrivée).

Pour évaluer la similarité sémantique entre les tokens, nous avons choisi de mesurer le cosinus de similarité au sein de la base d'embeddings de GPT-2XL.

Nous avons évité la base du GPT-4, bien qu'elle soit plus performante, afin de contourner la limitation méthodologique mentionnée par Bills et al. [8] et Bricken [15], qui est d'associer des systèmes cognitifs artificiels ne reposant pas sur le même système d'embeddings, c'est-à-dire sur une segmentation catégorielle différente.

4.3 Options statistiques

Pour évaluer la normalité de nos données, caractéristique essentielle pour réaliser des tests paramétriques, nous avons suivi une méthode en deux phases. Tout d'abord, nous avons effectué des tests statistiques inférentiels, incluant : le test de Shapiro-Wilk, qui est efficace pour les petits ensembles de données ; le test de Lilliefors, pertinent lorsque les paramètres de la distribution normale ne sont pas connus et sont estimés à partir des données ; le test de Kolmogorov-Smirnov, adapté aux grands ensembles de données ; et le test de Jarque-Bera, qui évalue la symétrie et la platitude des données dans les grands échantillons. Ensuite, nous avons complété cette approche par des statistiques descriptives telles que le *skewness* (pour l'asymétrie) et le *kurtosis* (pour le degré d'aplatissement), accompagnées de techniques visuelles comme le QQ-plot pour comparer les données observées à une distribution normale théorique. Pour vérifier l'homogénéité des variances entre groupes, nous avons appliqué le test de Bartlett (sensible aux déviations de normalité), ainsi que le test complémentaire de Levene (moins affecté par ces écarts).

Les résultats, partiellement présentés dans cet article, montrent une normalité approximative de nos données. Par conséquent, nos analyses statistiques reposent ici principalement sur des démarches non paramétriques ; cela, via :

- **Le test de Kruskal-Wallis**, qui explore la relation entre une variable catégorielle définissant plusieurs groupes indépendants et une variable ordinale ; test appliqué en ordonnant nos données numériques sur l'activation neuronale des tokens ; et, conformément aux conditions d'application de cet outil, pour des groupes d'au moins 5 observations.
- **Le test du χ^2 univarié d'adéquation**, en tenant compte de ses exigences en ce qui concerne les effectifs théoriques et observés, pour éviter d'avoir recours aux alternatives pour des petits effectifs, tels les tests de Fisher ou Monte Carlo.
- **Le test binomial**, en s'assurant de la binarité des variables impliquées, ainsi que de l'indépendance des épreuves et de l'égalité des probabilités des résultats des modalités.

5 Présentation de nos résultats

5.1 Confluence catégorielle partielle

Comme indiqué précédemment, notre premier angle d'étude du phénomène synthétique de restructuration catégorielle porte sur l'effet que produit la

coactivité de l'attention catégorielle et du phasage catégoriel synthétiques sur cette restructuration, à l'occasion du passage des catégories précurseures à leur catégorie associée en couche ultérieure. Plus particulièrement, nous cherchons à mieux comprendre comment cette restructuration se manifeste au niveau des sous-dimensions catégorielles détourées respectivement de chacune des catégories précurseures.

Dans ce cadre, nous avons postulé l'existence d'un phénomène de la cognition synthétique : la confluence catégorielle partielle. A savoir, pour un neurone d'arrivée (en couche 1 au niveau de notre étude), une convergence sémantique relative entre les clusters de tokens extraits (taken-clusters) de leurs neurones précurseurs (à plus forts poids de connexion) ; c'est-à-dire, entre les sous-dimensions catégorielles détourées de chaque catégorie de départ.

D'un point de vue méthodologique, nous avons procédé ainsi. Pour chaque neurone d'arrivée en couche perceptron 1, nous nous sommes centrés sur ses 10 neurones précurseurs en couche 0, avec lesquels il possède un plus fort poids (attentionnel) de connexion ; et plus particulièrement, sur les 10 taken-clusters détourés de ces précurseurs. Puis nous avons comparé chacun de ces taken-clusters aux autres, en ne retenant que les cas où chacun des deux taken-clusters comparés contenait au moins 6 taken-tokens (cardinal minimal des taken-clusters qui a été retenu pour toutes les analyses de la présente étude, notamment à des fins d'application ultérieure du test inférentiel de Kruskal-wallis). Pour chacun de ces croisements dits effectifs, la distance sémantique entre les tokens impliqués a été évaluée en termes de cosinus similarité à partir du référentiel d'observation constitué par les embeddings de GPT2-XL, selon la formule (sans doublons) :

$$m = \text{moyenne}(\cos(\text{tokens-cluster}_x, \text{tokens-cluster}_y))$$

Au sein du tableau n°1, nous pouvons observer une distance sémantique moyenne de 0,45 entre les taken-clusters associés, qui témoigne d'une convergence sémantique relative entre ces clusters. Sur les 9463 croisements effectifs, 72 % sont inférieurs à 0,5 (le cosinus similarité variant, pour rappel, de 0 à 1) ; cela, de façon largement significative ($p(\chi^2) < 0,0001$) (avec un χ^2 d'ajustement basé sur une hypothèse d'équi-distribution). Ces données sont compatibles avec notre hypothèse de confluence catégorielle partielle, de nature à montrer l'effet d'intersection catégorielle généré par le facteur de phasage catégoriel, sans pour autant empêcher un impact seul (i.e. sans trop d'effet de phasage catégoriel) de l'effet de complémentation catégorielle produit par le facteur d'attention catégorielle. Autrement dit, les taken-clusters extraits des catégories précurseures ont bien tendance à converger sémantiquement, sans pour autant que cette convergence soit synonyme d'identité catégorielle (dernière caractéristique qui se produirait si les taken-clusters avaient tendance à être identiques, c'est-à-dire uniquement produits par phasage catégoriel sans effet de complémentation). En tout cas, lorsque nous étudions la question de la distance sémantique à partir du référentiel d'observation sémantique constitué par les embeddings de GPT2-XL.

N (croisements effectifs)	9 463
Moyenne (m)	0,453
% de ($m < 0{,}5$)	72,241
$p(\chi^2)$ de ($m < 0{,}5$)	6,78E-213

Tableau n° 1 : Statistiques de similarités cosinus relatives aux taken-clusters entre les couches 0 et 1 de GPT2-XL.

5.2 Dispersion activationnelle catégorielle

Comme évoqué dans la présentation de notre problématique, notre deuxième axe d'investigation du phénomène synthétique de restructuration catégorielle a trait à l'impact que produit la coactivité de l'attention et de l'amorçage catégoriels sur l'aspect activationnel de cette restructuration. Et nous avons postulé un effet de dispersion activationnelle catégorielle, posant qu'un cluster de taken-tokens extrait d'une catégorie de départ, ne correspond pas à un segment continu de valeurs d'activation de ces tokens au sein du neurone de départ impliqué ; autrement dit, qu'une sous-dimension catégorielle extraite ne compartimente pas une zone activationnelle homogène (i.e. un segment de valeurs d'activation proches).

Notre démarche méthodologique a été la suivante. Pour chaque taken-cluster, nous avons comparé la distance moyenne d'activation (sur le neurone de départ) de ses tokens constitutifs au premier quartile des distances d'activation des 100 core-tokens (sur ce même neurone de départ) ; le premier quartile exprimant la fourchette basse de ces distances. Cela, à travers la formule :

$$\begin{aligned} d = \text{moyenne}\left(\left|\text{activation}_{\text{taken-token}_x} - \text{activation}_{\text{taken-token}_y}\right|\right) \\ - Q1\left(\left|\text{activation}_{\text{core-token}_n} - \text{activation}_{\text{core-token}_m}\right|\right) \end{aligned}$$

Des valeurs positives de d indiquent une non-proximité des valeurs d'activation des *taken-tokens*. Seuls les *taken-clusters* contenant au moins six tokens ont été inclus dans cette analyse.

Le tableau n°2 fait montre, à partir des 9007 distances calculées, d'une valeur moyenne de la distance d égale à 0,33. Et un pourcentage de cas où cette distance est positive égal à 88,45 %, tendance largement significative ($p(\chi^2) < 0,0001$). Ces données sont compatibles avec notre hypothèse de dispersion activationnelle et tendent ainsi à manifester le fait que la restructuration catégorielle, durant le passage d'une couche neuronale à sa couche suivante, va de pair avec une restructuration activationnelle : des segments activationnels de tokens, en neurone de départ, ne définissent pas les segments catégoriels de taken-clusters qui vont être détourés, dans le cadre de cette restructuration, des catégories associées à ces neurones de départ.

N	9 007
Moyenne (d)	0,331
% de ($d > 0$)	88,448
$p(\chi^2)$ de ($d > 0$)	1,47E-14

Tableau n° 2 : Comparaisons des distances d'activation entre taken-tokens et core-tokens (couche 0 de GPT2-XL).

Le tableau n°3, établi selon la même méthodologie, indique qu'il en va de même concernant notre postulat relatif à une dispersion activationnelle des taken-tokens au niveau, cette fois-ci, du neurone d'arrivée. Avec une distance moyenne de 0,43 et un pourcentage, fortement significatif ($p(\chi^2) < 0,0001$), de cas où d est positif égal à 88,4%. Ces données sont donc compatibles elles aussi avec notre hypothèse de dispersion activationnelle au niveau des neurones d'arrivée : les sous-dimensions catégorielles détourées des catégories des neurones précurseurs ne délimitent pas de zones activationnelles spécifiques au sein de leurs neurones d'arrivée associés.

N	9 007
Moyenne (d)	0,432
% de ($d > 0$)	88,385
$p(\chi^2)$ de ($d > 0$)	1,63E-14

Tableau n° 3 : Comparaisons des distances d'activation entre taken-tokens et core-tokens (couche 1 de GPT2-XL).

5.3 Distanciation catégorielle

Notre dernier angle d'étude du phénomène synthétique de restructuration catégorielle a trait à l'existence d'une distanciation catégorielle entre la segmentation catégorielle portée par un neurone d'arrivée et celles portées par ses neurones précurseurs (à forts poids de connexion). Et, plus spécifiquement, à l'existence d'une différence de segmentation catégorielle entre les clusters catégoriels de la catégorie d'un neurone d'arrivée et les clusters catégoriels des catégories de ses neurones précurseurs. Postulat que nous étudions à travers l'évaluation de la distance sémantique entre les clusters catégoriels de core-tokens d'un neurone d'arrivée et les clusters catégoriels de core-tokens de chacun de ses neurones précurseurs avec lesquels il a un fort poids de connexion ; cette distance sémantique étant ici, à nouveau, mesurée ici dans le référentiel d'observation constitué par les embeddings de GPT2-XL.

Dans un registre méthodologique, rappelons que les clusters catégoriels étudiés ici n'ont rien de segmentations absolues mais sont bien entendu relatifs à l'opérationnalisation qui a été choisie pour les générer. Dans le cadre de notre présente étude, nous avons mobilisé, par prompt engineering, GPT4o pour produire ces clusters, que nous avons arbitrairement fixé au nombre invariant de 5 pour chaque neurone. Les clusters catégoriels de la catégorie portée par chaque neurone d'arrivée ont été comparé, pour chacune des catégories de ses

neurones précurseurs, à leurs clusters catégoriels propres. Cela, encore une fois, en nous centrant sur (i) les seuls 10 neurones précurseurs (en couche 0) à plus forts poids de connexion avec chaque neurone d'arrivée associé (en couche 1) et, (ii) sur les seuls clusters catégoriels contenant au moins 6 tokens. Pour chacun de ces croisements, d'un cluster catégoriel de départ x en couche 0 et d'un cluster d'arrivée y en couche 1, nous avons comparé la proximité sémantique entre les tokens x et y d'une part, avec la proximité sémantique entre les tokens x d'autre part. Cela, avec la formule suivante :

$$d = \text{mean}(\cos(\text{tokens}_x, \text{tokens}_y)) - \text{mean}(\cos(\text{tokens}_x))$$

. Une valeur négative de d traduisant ainsi une distance sémantique *a minima* conséquente entre les clusters catégoriels d'arrivée et ceux de départ, distance constitutive de l'effet de distanciation catégorielle que nous investiguons.

Le tableau n°4 présente les résultats obtenus, sur les 4692 neurones d'arrivée en couche 1, ayant au moins un neurone précurseur (en couche 0) doté d'au moins un cluster catégoriel comprenant au moins 6 tokens. 138367 distances d ont été calculées, avec une valeur moyenne de -0,14, et un pourcentage extrêmement élevé (99,83%) de cas où d est négatif ; cela, de façon largement significative ($p(\chi^2) < 0,0001$). Avec un pourcentage élevé (82,27%) de cas où la distance d est significative (au titre du test de Kruskal-Wallis), à nouveau de façon largement significative ($p(\chi^2) < 0,0001$).

N_d	138367
Moyenne (d)	-0,142
% de ($d < 0$)	99,829
$p(\chi^2)$ de ($d < 0$)	2,15E-23
% de ($p_{\text{KW}} < 0,05$)	82,274
$p(\chi^2)$ de ($p_{\text{KW}} < 0,05$)	1,08E-10

Tableau n°4 : Statistiques sur la distance sémantique entre les clusters catégoriels des neurones d'arrivée (couche 1) et ceux de leurs neurones précurseurs (couche 0).

De façon complémentaire, nous avons méthodologiquement procédé à une autre évaluation de la distance sémantique entre les clusters catégoriels d'arrivée et de départ. Cela, en dénombrant, pour chaque croisement, le nombre n de tokens en commun. Puis en calculant l'indice $d' = n - \frac{nx}{10}$, où nx est le nombre de tokens contenus dans un cluster catégoriel de départ x. Une valeur négative de d' indiquant dès lors un faible nombre relatif de tokens en commun.

Le tableau n°5 permet d'observer un faible nombre moyen (0,35) de tokens communs entre un cluster catégoriel d'un neurone d'arrivée et chacun des clusters catégoriels de ses neurones de départ associés. Et un nombre relatif moyen de tokens en commun négatif (-1,3), avec un pourcentage très fort (97,81%) et significatif ($p(\chi^2) < 0,0001$) de cas où d' est négatif ; et, de façon convergente, un pourcentage de cas important (94,59%) dans lesquels un test binomial relatif à la négativité de d' est significatif (ce pourcentage étant lui-même largement significatif avec $p(\chi^2) < 0,0001$).

N_d	138367
Moyenne ($\overline{n}$)	0,349
Moyenne ($\overline{d'}$)	-1,298
% de ($\overline{d'} < 0$)	97,811
$p(\chi^2)$ de ($\overline{d'} < 0$)	1,15E-21
% de ($p_b(d' < 0) < 0,05$)	94,589
$p(\chi^2)$ de ($p_b(d' < 0) < 0,05$)	4,76E-19

Tableau n° 5 : Statistiques sur le nombre de tokens communs entre les clusters catégoriels d'arrivée (couche 1) et leurs clusters de départ correspondants (couche 0).

Ces différents résultats sont compatibles avec notre postulat de distanciation catégorielle relatif au fait que la restructuration catégorielle, opérée lors du passage de la segmentation catégorielle des neurones précurseurs (couche 0) à la segmentation catégorielle de leurs neurones d'arrivée correspondant (à fort poids de connexion), se manifeste par un écart sémantique entre les catégories synthétiques portées par ces neurones corollaires respectivement de départ et d'arrivée ; en tout cas lorsque l'on mobilise comme référentiel d'observation sémantique, celui des embeddings de GPT2-XL, et que l'on compare les sous-groupes catégoriels (clusters catégoriels) segmentés par les neurones formels. La restructuration catégorielle, durant le passage d'une couche neuronale à une autre, définit bien un nouveau système de découpage catégoriel du monde des tokens, ce qui est bien entendu la fonction princeps des couches neuronales synthétiques successives, afin de pallier les limites catégorielles d'origine des embeddings de départ, embeddings non assez efficients pour réaliser les tâches impliquées, et pour lequel le réseau de neurones a justement été entraîné.

6 Illustration qualitative de nos résultats

En complément de nos analyses quantitatives précédentes, nous illustrons maintenant les principaux processus cognitifs synthétiques explicatifs que nous avons mobilisés dans le cadre de notre présente investigation, afin de comprendre qualitativement plus avant comment opère empiriquement le processus de restructuration catégorielle artificielle.

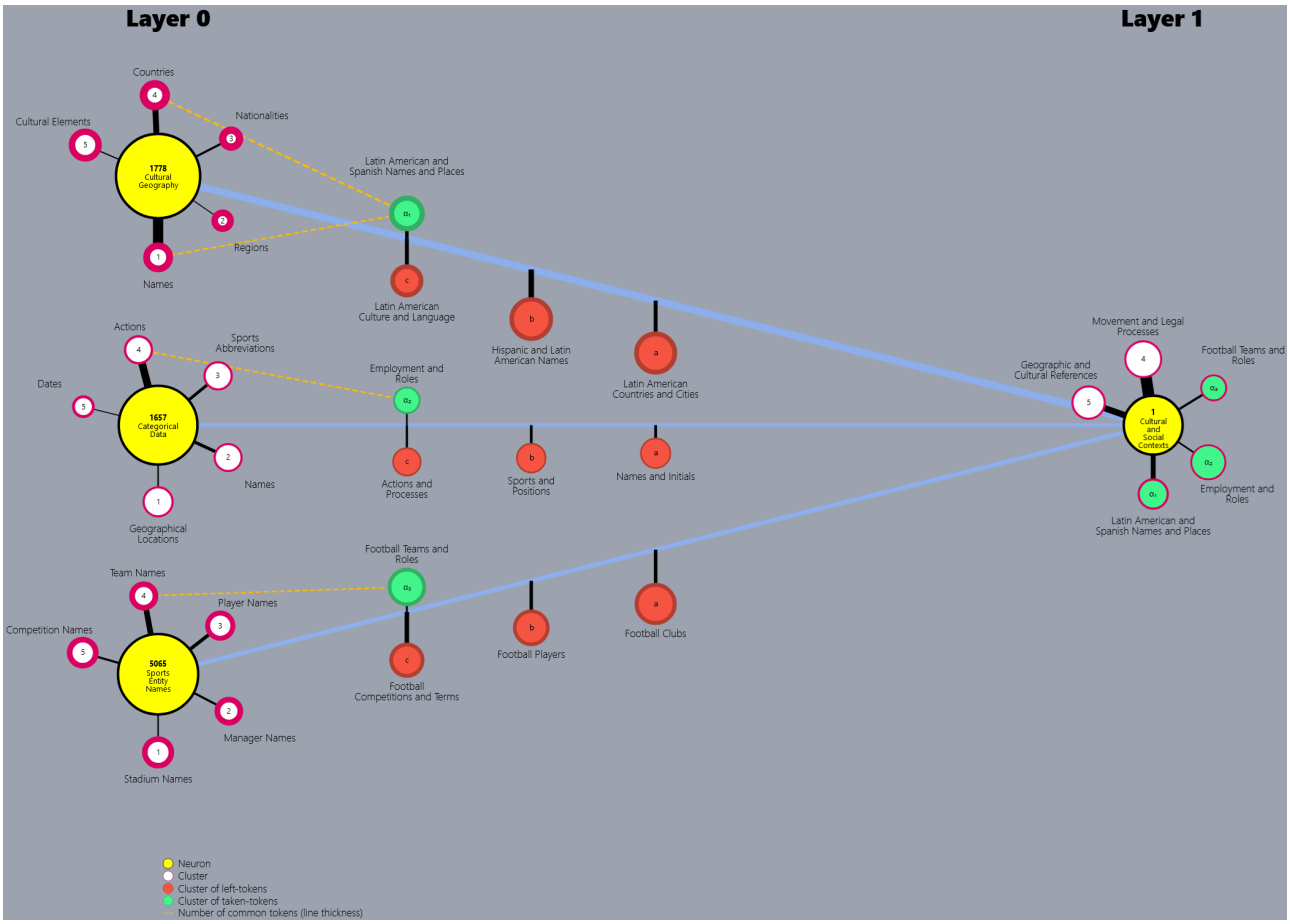
6.1 La complémentation catégorielle synthétique

Pour rappel, l'attention catégorielle synthétique, ou effet w, relève du fait que, lors de son activité de restructuration catégorielle d'une couche n à une couche $n+1$, un neurone d'arrivée, de par sa fonction d'agrégation propre, va porter plus d'attention sur certaines catégories sous-ordonnées spécifiques afin de constituer sa catégorie spécifique. Cela se traduisant par un processus de complémentation catégorielle consistant génétiquement à « apporter » à l'extension (de tokens) de cette catégorie d'arrivée une sous-dimension catégorielle singulière extraite d'une catégorie précurseure, sous-dimension sémantiquement différente de et

complémentaire à d'autres sous-dimensions catégorielles extraites du reste des catégories précurseures.

Le graphe n°2 présente un exemple de complémentation catégorielle générée par l'attention catégorielle synthétique (cas du neurone d'arrivée n°1 en couche 1 de GPT2-XL). Le neurone d'arrivée (en couche 1), associé à la catégorie *"Cultural & social contexts"* extrait ici, à partir de ses neurones précurseurs en couche 0, par complémentation, 3 sous-dimensions catégorielles sémantiquement différentes et complémentaires les unes des autres :

- La sous-dimension catégorielle α_1 *"Latin american & spanish names & places"*, extraite de la catégorie précurseure n°1778 *"Cultural geography"*, et contenant entre autres les tokens : `[Argentine]`, `[Luis]`, `[Juan]`, `[Puerto]`, `[Nicarag]`.
- La sous-dimension catégorielle α_2 *"Employment & roles"*, détourée de la catégorie précurseure n°1657 *"Categorical data"*, et impliquant notamment les tokens : `[transfer]`, `[incub]`, `[stint]`.
- La sous-dimension catégorielle α_3 *"Football teams and roles"*, abstraite de la catégorie précurseure n°5065 *"Sports entity names"*, et relevant de divers tokens dont : `[goalkeeper]`, `[relegation]`, `[Cardiff]`, `[Southampton]`.



Graphe n°2 : Cas de complémentation catégorielle générée par l'attention catégorielle synthétique (neurone d'arrivée n°1 en couche 1 de GPT2-XL).

6.2 Le phasage catégoriel synthétique

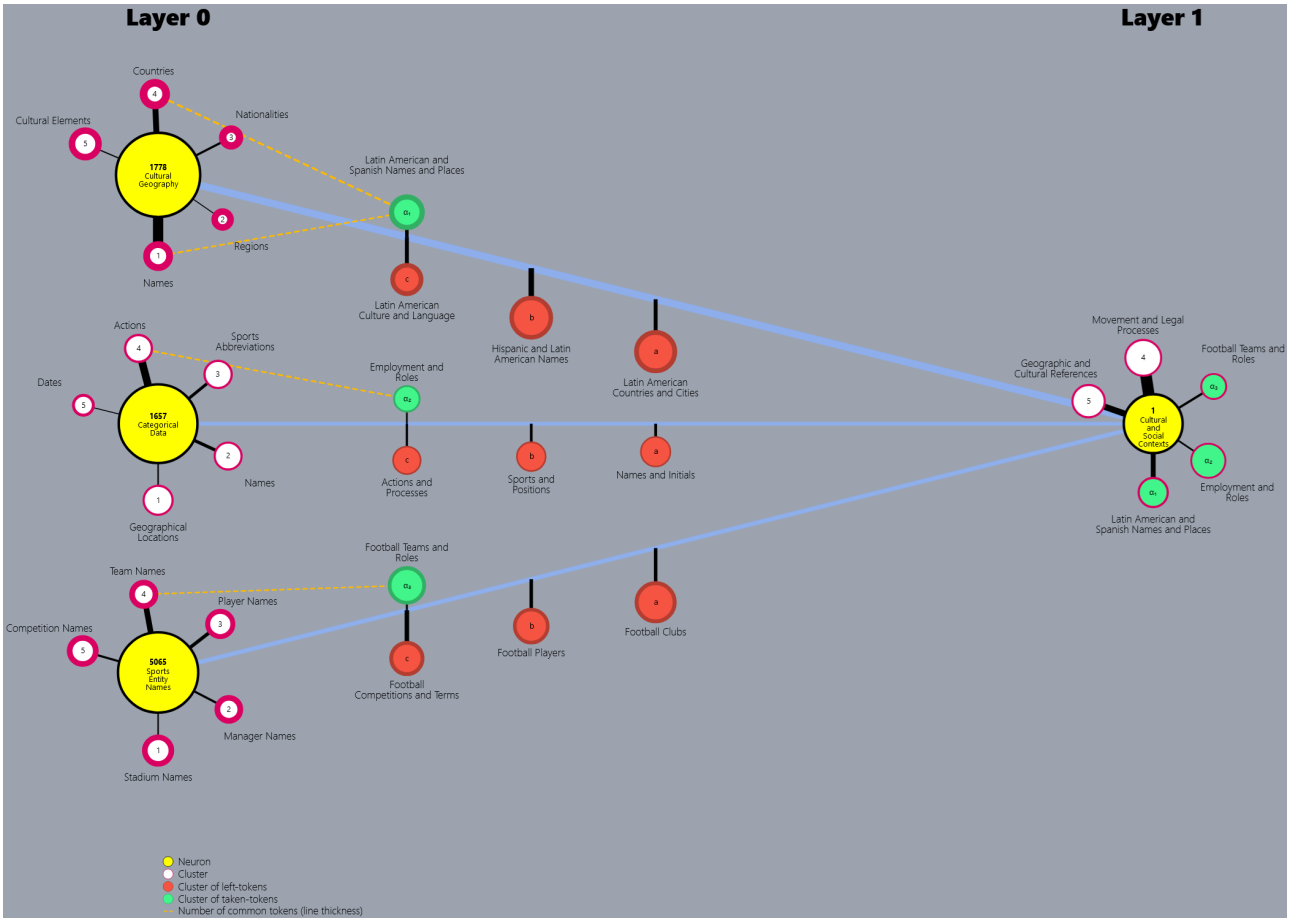
Comme mentionné précédemment, le phasage catégoriel synthétique, ou effet Σ, a trait au fait que, lors de son activité de restructuration catégorielle d'une couche n à une couche $n+1$, un neurone d'arrivée, étant donnée sa fonction d'agrégation particulière, va extraire de plusieurs de ses catégories précurseures des sous-dimensions catégorielles corrélées sémantiquement entre elles. Cela se manifestant par un processus d'intersection catégorielle, au sein duquel des mêmes tokens, rentrant alors en écho catégoriel, sont conjointement extraits des catégories de départ et dès lors simultanément présents dans l'extension des sous-dimensions extraites.

Le graphe n°3 présente un exemple d'intersection catégorielle produite par phasage catégoriel synthétique (cas du neurone d'arrivée n°121 en couche 1 de

GPT2-XL). Le neurone d'arrivée (en couche 1), associé à la catégorie *"Unverified assertions"* exfiltre ici, à partir de ses neurones précurseurs en couche 0, par intersection, 3 sous-dimensions catégorielles sémantiquement liées de façon conséquente les unes aux autres :

- La sous-dimension catégorielle α_1 *"Speculative claims"*, extraite de la catégorie précurseure n°6356 *"Document elements"*.
- La sous-dimension catégorielle α_2 *"Legal allegations"*, détourée de la catégorie précurseure n°3721 *"Legal concepts"*.
- La sous-dimension catégorielle α_3 *"Speculative claims"*, issue de la catégorie précurseure n°5207 *"Linguistic elements"*.

Ces sous-dimensions catégorielles, ainsi que le montre en quelques exemples non exhaustifs le tableau n°6, relèvent de taken-clusters contenant une série de tokens en commun.



Graphe n°3 : Cas d'intersection catégorielle générée par le phasage catégoriel synthétique (neurone d'arrivée n°121 en couche 1 de GPT2-XL).

	α_1	α_2	α_3
[claim]	X	X	X
[allegedly]	X	X	X
[purported]	X		X
[reportedly]	X		X
[alleges]		X	X
[alleged]		X	X

Tableau n°6 : tokens identiques entre les taken-clusters α_1, α_2, α_3 extraits respectivement des catégories prédécesseures n°6356, 3721 et 5207 en couche 0, par le neurone d'arrivée n°121.

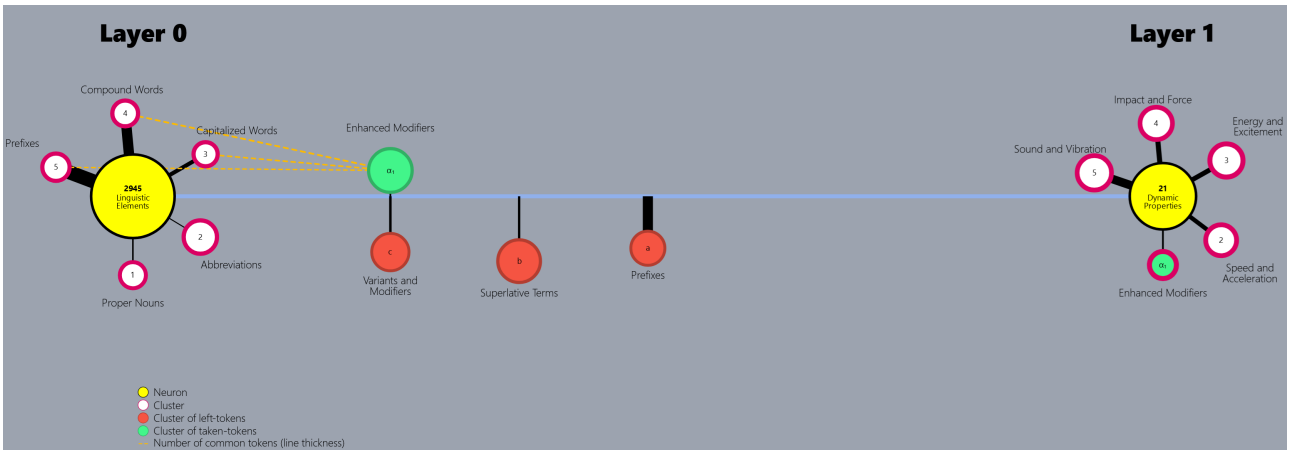
6.3 Détourage Catégoriel Synthétique

L'attention et le phasage catégoriels, avec l'amorçage catégoriel, sont les processus synthétiques (portés par les fonctions d'agrégation neuronales) qui façonnent le détourage catégoriel lors de la restructuration catégorielle. Le détourage, nous l'avons évoqué, est le mécanisme de la cognition artificielle consistant à construire activement une forme catégorielle qui va être distinguée d'un fond catégoriel (tout aussi construit). Ce détourage fabrique la sous-dimension catégorielle (matérialisée par un token-cluster) qui est singulièrement abstraite de chaque catégorie de départ.

Le Graphe 4 présente un cas de détourage catégoriel de type sémantique : les token-tokens constitutifs de la sous-dimension catégorielle α_1, extraite comme *"Enhanced modifiers"* de la catégorie *"Linguistic elements"* du neurone précurseur n° 2945 (couche 0), sont homogènes d'un point de vue sémantique (en tout cas lorsque nous prenons comme référentiel d'observation celui de la sémantique humaine). Ces tokens sont en effet (pour partie) les suivants : `[Hyper]`, `[Super]`, `[turbo]`. La sous-dimension catégorielle α_1 est bien fabriquée et non pas simplement passivement identifiée à partir d'un sous-groupe sémantique intrinsèquement préexistant ; cette sous-dimension est en effet sémantiquement distincte des autres clusters catégoriels possibles de ce neurone, tels que : *"Proper nouns"*, *"Abbreviations"*, *"Capitalized words"*, *"Compound words"*, *"Prefixes"*.

Cette sous-dimension—forme catégorielle détourée—est également sémantiquement distincte du fond catégoriel non extrait de la catégorie de départ, fond qu'il est possible de segmenter à travers les trois left-clusters suivants :

- *Left-cluster a "Prefixes"*, constitué entre autres des tokens : `[altern]`, `[aux]`, `[counter]`, `[subst]`.
- *Left-cluster b "Superlative terms"*, comprenant notamment les tokens : `[superflu]`, `[superhuman]`, `[superpower]`, `[superior]`, `[supers]`.
- *Left-cluster c "Variants & modifiers"*, constitué par exemple des tokens : `[mini]`, `[doub]`, `[dual]`, `[extra]`, `[triple]`, `[sup]`.

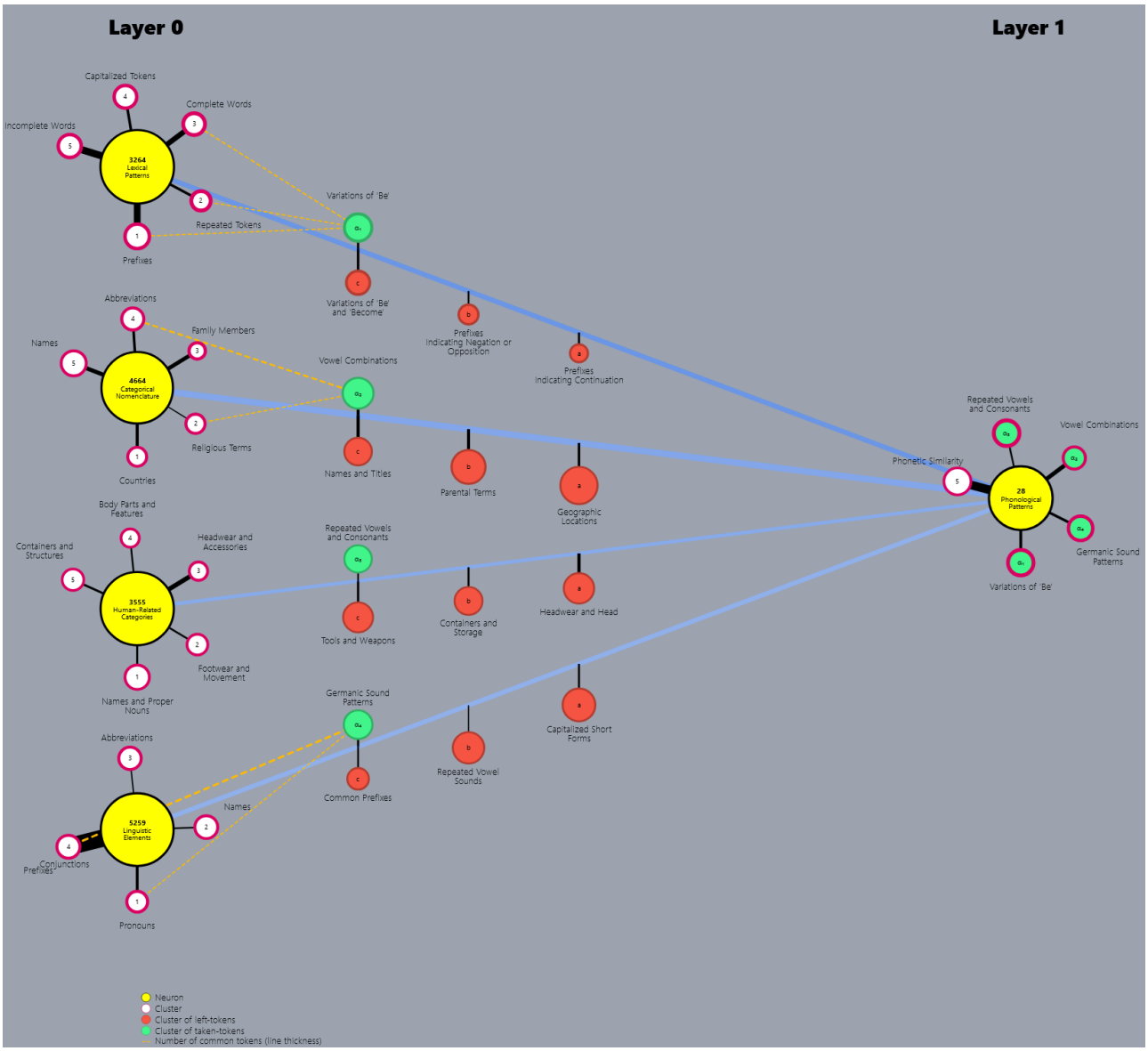


Graphe n°4 : Cas de détourage catégoriel synthétique de type sémantique (neurone d'arrivée n°21 en couche 1 de GPT2-XL).

Le Graphe 5, quant à lui, nous permet d'observer une situation de détourage catégoriel de type graphémique-phonologique :

- Les token-tokens constitutifs de la sous-dimension catégorielle α_1, extraite comme *"Variations of be"* de la catégorie *"Lexical patterns"* du neurone précurseur n° 3264 (couche 0), sont homogènes d'un point de vue graphémique et/ou phonologique. Ces tokens impliqués sont par exemple : `[be]`, `[beau]`, `[bes]`, `[bei]`, `[beaver]`. La sous-dimension catégorielle unique α_1 est distincte d'autres clusters catégoriels possibles de ce neurone, tels que : *"Prefixes"*, *"Repeated tokens"*, *"Complete words"*, *"Capitalized tokens"*, *"Incomplete words"*.
- Les token-tokens formant la sous-dimension catégorielle α_2, *"Vowel combinations"*, détourée de la catégorie *"Categorical nomenclature"* du neurone précurseur n° 4664 (couche 0), sont homogènes d'un point de vue graphémique et/ou phonologique. Ces tokens impliqués sont notamment : `[AE]`, `[EA]`, `[EE]`, `[IE]`, `[EEE]`. Cette sous-dimension catégorielle α_2 est également distincte des autres clusters catégoriels possibles de ce neurone, tels que : *"Countries"*, *"Religious terms"*, *"Family members"*, *"Abbreviations"*, *"Names"*.
- Les token-tokens générant la sous-dimension catégorielle α_3, *"Repeated vowels & consonants"*, détachée de la catégorie *"Human-related categories"* du neurone précurseur n° 3555 (couche 0), convergent au niveau graphémique et/ou phonologique. Ces tokens incluent, entre autres : `[EE]`, `[oo]`. Cette sous-dimension catégorielle α_3 est différenciée des autres clusters catégoriels possibles de ce neurone, tels que : *"Names & proper nouns"*, *"Footwear & movement"*, *"Headwear & accessories"*, *"Body parts & features"*, *"Containers & structures"*.
- Enfin, les token-tokens définissant la sous-dimension catégorielle α_4,

"Germanic sound patterns", détourée de la catégorie *"Linguistic elements"* du neurone précurseur n° 5259 (couche 0), sont cohérents au niveau graphémique et/ou phonologique. Cette sous-dimension contient les tokens : `[Die]`, `[Bei]`, et `[Lie]`. Cette sous-dimension catégorielle α_4 est également distincte des autres clusters catégoriels possibles de ce neurone, tels que : *"Pronouns"*, *"Names"*, *"Abbreviations"*, *"Prefixes"*, *"Conjunctions"*.



Graphe n° 5 : Cas de détourage catégoriel synthétique de type graphémique-phonologique (neurone d'arrivée n°28 en couche 1 de GPT2-XL).

6.4 Restructuration Catégorielle Synthétique

Les facteurs mathématico-cognitifs synthétiques que sont l'amorçage, l'attention et le phasage catégoriels formatent le processus de détourage catégoriel consistant à nouveau, pour la catégorie d'un neurone d'arrivée en couche $n+1$,

à se constituer génétiquement en extrayant (et en combinant) de chacune de ses catégories précurseures (en couche n) une sous-dimension catégorielle singulière. Ces sous-dimensions catégorielles détourées vont devenir des sous-dimensions (possibles) constitutives de la nouvelle catégorie super-ordonnée ainsi fabriquée et de facto produire un nouveau système de segmentation catégorielle du monde des tokens ; système original et distinct de celui de chacune des catégories précurseures impliquées. Ce que nous dénotons comme le phénomène synthétique de la restructuration catégorielle.

Le graphe n°6 illustre de façon éclairante le processus de restructuration catégorielle. D'un premier neurone précurseur en couche 0, n°6255, associé à la catégorie de pensée *"Current affairs"*, la sous-dimension catégorielle α_1 *"Disasters & threats"* est détourée. Elle contient, parmi d'autres, les tokens : `[qaida]`, `[nightmare]`, `[malware]`, `[ransomware]`.

On observe assez aisément comment ces tokens ont pu être très sélectivement extraits, d'au moins 4 des 5 clusters catégoriels avec lesquels il était initialement possible de décrire la catégorie de départ (ces tokens étant respectivement sémantiquement liés à certains aspects de ces clusters catégoriels) : *"Health"*, *"Political figures / entities"*, *"Technology"*, *"Criminal activities"*. Cette sous-dimension extraite, premier élément de la restructuration catégorielle ici observée, est sémantiquement une authentique construction catégorielle bien distincte du fond catégoriel qui a été ici identifié comme non pertinent à être retenu. Fond catégoriel que nous pouvons classer ici à travers les 3 left-clusters :

- *Left-cluster a "Political figures & entities"*, constitué entre autres des tokens : `[Maduro]`, `[Hezbollah]`, `[Chavez]`, `[Boko]`, `[Fidel]`, `[scientology]`.
- *Left-cluster b "Natural & environmental events"*, comprenant les tokens : `[armageddon]`, `[tempest]`, `[typhoon]`, `[tornado]`, `[wildfire]`, `[irma]`.
- *Left-cluster c "Health & medical conditions"*, formé notamment des tokens : `[leukemia]`, `[herpes]`, `[pox]`, `[allergies]`, `[tumor]`, `[diabetes]`.

Toujours au niveau du graphe n°6, nous voyons qu'une autre sous-dimension catégorielle, α_2 *"Magical & mystical elements"* est extraite du neurone précurseur n°6040, relevant de la catégorie synthétique *"Symbolic elements"*. L'extension de cette sous-dimension comprend, entre autres, les tokens : `[spiral]`, `[scourge]`, `[ginny]` (de Ginny Weasley, personnage de *"Harry Potter"*). Ces tokens ont potentiellement été électivement extraits d'au moins 4 des 5 clusters catégoriels avec lesquels il était à l'origine possible de segmenter la catégorie de départ : *"Spirituality"*, *"Names"*, *"Jewelry & gems"*, *"Weapons"*. Cette deuxième sous-dimension détourée, part supplémentaire de la restructuration catégorielle ici décrite, est à nouveau une élaboration catégorielle originale bien différente du fond catégoriel laissé pour compte. Fond catégoriel que nous pouvons décomposer via les 3 left-clusters :

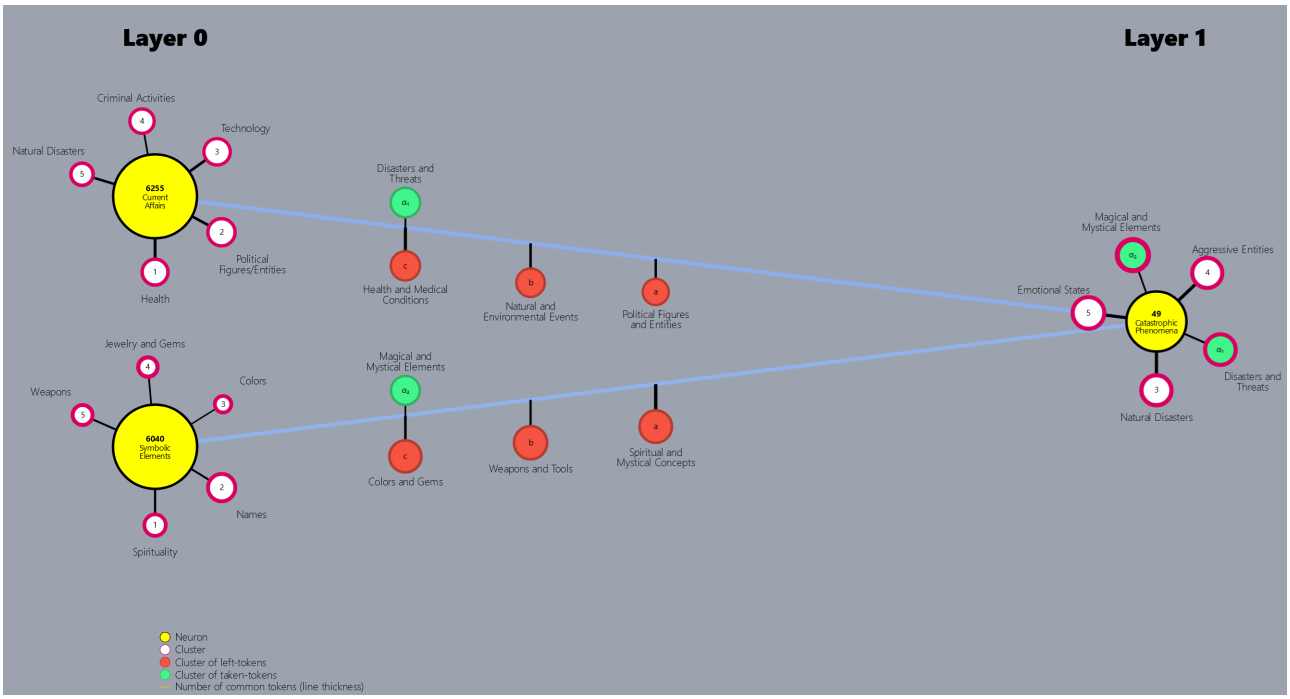
- *Left-cluster a "Spiritual & mystical concepts"*, formé pour partie des tokens : `[soul]`, `[virtue]`, `[grace]`, `[sacrament]`, `[gift]`, `[spirit]`, `[mystery]`.

- *Left-cluster b "Weapons & tools"*, impliquant entre autres les tokens : `[spear]`, `[sword]`, `[dagger]`, `[lightsaber]`, `[scythe]`, `[baton]`.
- *Left-cluster c, enfin, "Colors & gems"*, formé notamment des tokens : `[sparkle]`, `[jewel]`, `[gem]`, `[pearl]`, `[ruby]`.

Dans le cadre de la restructuration catégorielle que nous sommes en train de détailler au niveau du graphe n°6, nous voyons que les sous-dimensions catégorielles détourées que nous venons de mentionner, α_1 *"Disasters & threats"* et α_2 *"Magical & mystical elements"*, parmi d'autres sous-dimensions catégorielles potentielles détourées d'autres catégories précurseures en couche 0, peuvent constituer des clusters catégoriels de la nouvelle catégorie qui leur est associée en couche 1. Cette nouvelle catégorie synthétique super-ordonnée, nommée ici *"Catastrophic phenomena"*, pouvant ici être décrite avec les 3 autres clusters catégoriels :

- *Cluster 3 "Natural disaster"*, constitué entre autres des tokens : `[calam]`, `[catastrophic]`, `[pandemonium]`, `[katrina]`, `[turmoil]`, `[havoc]`.
- *Cluster 4 "Aggressive entities"*, comprenant notamment les tokens : `[malicious]`, `[ferocious]`, `[insurgents]`, `[demonic]`, `[taliban]`, `[devils]`, `[jihad]`, `[rebel]`.
- *Cluster 5 "Emotional states"*, mobilisant par exemple les tokens : `[stirred]`, `[messed]`, `[raging]`, `[mad]`.

Ainsi vient de s'opérer, sous nos yeux si nous pouvons le dire ainsi, une restructuration catégorielle, par le processus de détourage catégoriel généré par les facteurs synthétiques que sont l'amorçage, l'attention et le phasage catégoriels. D'une segmentation catégorielle initiale en couche 0, opérée notamment autour des catégories *"Current affairs"* et *"Symbolic elements"*, émerge, en couche 1, une nouvelle catégorie synthétique originale *"Catastrophic phenomena"*, dont nous venons partiellement de suivre la genèse progressive.



Graphe n°6 : Cas de restructuration catégorielle synthétique (neurone d'arrivée n°49 en couche 1 de GPT2-XL).

Notre neurone viewer génétique, associé à cette étude, permet de visualiser le phénomène de restructuration catégorielle synthétique opérée lors du passage de la couche perceptron 0 à 1 de GPT2-XL.

7 Discussion

7.1 Synthèse de nos résultats empiriques

Nous avons étudié le phénomène de restructuration catégorielle synthétique, c'est-à-dire la construction, à chaque nouvelle couche neuronale $n+1$, de nouvelles catégories de pensée synthétiques plus efficientes pour segmenter le monde des tokens, en vue de réaliser les tâches attendues du modèle de langage ; restructuration catégorielle consistant à combiner, via un processus analogue à l'abstraction réfléchissante définie par Piaget [63], les catégories antécédentes pour en fabriquer de nouvelles.

Nous avons cherché à comprendre comment cette restructuration catégorielle était réalisée par le processus de détourage catégoriel [63] consistant en une construction singulière, au niveau de chaque neurone d'une couche de niveau $n+1$, d'une distinction :

- de formes catégorielles extraites (i.e. sous-dimensions catégorielles détourées) des catégories portées par ses neurones précurseurs en couche n et,
- d'un fond catégoriel non retenu, non pertinent.

Et nous avons également tenté de comprendre comment cette restructuration et ce détourage catégoriels sont façonnés par l'activité et la coactivité de trois facteurs de la segmentation catégorielle [62], portés par la fonction d'agrégation neuronale, que sont :

- L'amorçage catégoriel, ou effet x, à savoir que plus des tokens sont activés au sein d'une catégorie de couche n, et plus ils ont des chances d'en être extraits pour devenir constitutifs de l'extension des catégories de couche $n+1$ qui leur sont particulièrement liées.
- L'attention catégorielle, ou effet w, exprimant le fait que plus une catégorie de couche $n+1$ a un fort poids de connexion avec une catégorie de couche n, et plus il est probable que les tokens de la catégorie prédécesseure en soient extraits afin de devenir membre de l'extension de la catégorie d'arrivée.
- Le phasage catégoriel, ou effet Σ, associé au fait que plus des tokens sont simultanément activés au sein de différentes catégories de couche n et plus il est probable que ces tokens appartiennent à l'extension de la catégorie de couche $n+1$ avec laquelle ces catégories antécédentes sont fortement reliées.

Nous nous sommes particulièrement intéressés au phénomène synthétique d'attention catégorielle dans notre étude de la restructuration catégorielle, dans la mesure où :

- celui-ci rend possible et catalyse les deux autres phénomènes synthétiques que sont l'amorçage et le phasage catégoriels, ces derniers ne pouvant significativement opérer que dans les cas où les neurones de couches successives sont fortement reliés,
- les poids de connexion attentionnels neuronaux sont les régularités construites et apprises à partir du monde de tokens auquel on soumet le modèle de langage durant sa phase d'entrainement ; régularités fabriquées afin de permettre à ce modèle de langage de construire un ordre interprétatif durant son interaction avec ce monde.

Dans un premier temps, nous avons mis en lumière le fait que la coactivité de l'attention et du phasage catégoriels générait le phénomène synthétique de *confluence catégorielle partielle*, une première propriété de la restructuration catégorielle ; à savoir que les sous-dimensions catégorielles détourées des catégories portées par des neurones de couche n, au niveau d'un neurone qui leur est fortement lié en couche $n+1$, tendent à converger dans un espace sémantique, afin de participer à la constitution de la catégorie associée à ce neurone d'arrivée. Cela, étant donné que pour que des tokens soient détourés d'une catégorie de couche n (i.e. deviennent des taken-tokens constitutifs d'une sous-dimension extraite), par un neurone d'arrivée, il convient que le résultat de leur fonction d'activation

$(\Sigma(w_{i,j}x_{i,j}) + b)$ au niveau de ce neurone d'arrivée soit fort ; donc que l'effet Σ de phasage catégoriel opère de façon importante ; et dès lors que ces tokens soient simultanément présents dans les sous-dimensions catégorielles (i.e. les taken-clusters) extraites des différentes catégories de couche n. Ce qui provoque ainsi, mécaniquement, une convergence sémantique entre les différentes sous-dimensions catégorielles impliquées, dans le cadre du processus de restructuration catégorielle.

Dans une deuxième démarche, nous avons manifesté le fait que l'interaction de l'attention et de l'amorçage catégoriels provoquait un processus synthétique de dispersion activationnelle catégorielle. C'est-à-dire qu'une sous-dimension catégorielle extraite d'une catégorie portée par un neurone en couche n ne correspond pas à un segment continu de valeurs d'activation (i.e. des valeurs d'activation proches) des tokens afférents (i.e. taken-tokens) au sein de l'espace d'activation de ce neurone de départ ; autrement dit, que des segments activationnels de tokens, dans un neurone de départ, ne définissent pas les segments catégoriels de taken-clusters qui vont être détourés, dans le cadre de la restructuration catégorielle, de la catégorie associée à ce neurone prédécesseur. Phénomène de dispersion catégorielle que nous avons également pointé au niveau de l'espace d'activation du neurone d'arrivée. Ce phénomène de dispersion étant dû au fait de l'impact significatif du processus de phasage catégoriel durant le détourage catégoriel : l'amorçage catégoriel ne suffit pas à nécessairement produire des activations suffisantes des neurones d'arrivée et donc à extraire des tokens constitutifs d'une sous-dimension détourée : devenir un taken-token implique, pour un token de départ, qu'il active suffisamment intensément son neurone de départ (amorçage catégoriel) mais aussi qu'il soit intégré au sein d'un phasage catégoriel (avec un ou plusieurs autre(s) neurone(s) de départ). Or ces deux processus sont a priori non directement liés. Et, dès lors, deux tokens proches en termes d'activation (même forte) au sein d'un neurone de départ, ne sont pas forcément tous deux l'objet d'un processus de phasage catégoriel : statistiquement, l'un peut l'être et non l'autre.

Dans une troisième et dernière phase de notre étude, nous avons manifesté que le processus synthétique de restructuration catégorielle relève d'un phénomène distanciation catégorielle : la modalité de segmentation catégorielle des objets du monde (des tokens) des catégories portées par les neurones de couche n sont sémantiquement différentes de celles des catégories vectorisées par leurs neurones antécédants (avec lesquels ils sont fortement reliés) ; en tout cas si cette distance sémantique est mesurée dans le référentiel d'observation constitué par les embeddings de départ du modèle de langage impliqué. Distanciation catégorielle qui est, bien entendu, la fonction princeps des couches neuronales synthétiques successives, afin de pallier les limites catégorielles d'origine des embeddings de départ, embeddings non assez efficients pour réaliser les tâches impliquées.

7.2 Interprétation fonctionnelle de nos résultats

Ainsi que le pointent von Glaserfeld [95] et Varela [89, 90], l'activité d'un système cognitif vise à extraire des régularités fonctionnelles au sein du flux informationnel de son expérience d'interaction avec le monde extérieur auquel il est exposé, le monde des tokens dans le cadre des modèles de langage.

Cette extraction est finalisée : elle est au service de l'efficience de ce système intelligent (efficience renseignée par les feedbacks administrés à ce dispositif d'IA lors de sa phase d'apprentissage profond). Cette extraction n'est pas passive mais active : elle n'identifie pas des propriétés intrinsèques et pré-données du monde extérieur, mais fabrique, en fonction des paramètres mathématiques et architecturaux propres au système synthétique (ainsi qu'en fonction de ses données d'entraînement et des feedbacks dont il est l'objet), des invariants fonctionnels dans la singularité de l'expérience que ce système a du monde extérieur de tokens avec lequel il interagit. Cette extraction de régularités, en tout cas au niveau des neurones de type perceptron, se traduit par l'apprentissage de poids attentionnels, constitutifs de la clé de cryptage (un théorème-en-acte au sens de Vergnaud [91, 92, 93]) déterminant comment combiner, dans une dynamique d'abstraction réfléchissante analogue à celle de Piaget [58], les catégories de pensée (concepts-en-acte au sens de Vergnaud à nouveau [91]) synthétiques d'une couche de niveau n afin de former celles de la couche suivante ; afin de rendre cette dernière encore plus discriminante et dès lors fonctionnelle.

La restructuration catégorielle est le marqueur de cette extraction finalisée, active et attentionnelle. Elle est le fruit du détourage catégoriel synthétique, produit par les facteurs cognitivo-mathématiques que sont l'amorçage, l'attention et le phasage catégoriels. Elle se traduit par une confluence catégorielle partielle, une dispersion activationnelle et une distanciation catégorielle, elles-mêmes également déterminées par ces facteurs cognitivo-mathématiques.

Cette restructuration catégorielle synthétique est l'activité qui réalise, de façon équivalente à ce que nous indique Varela [89, 90] à propos des systèmes intelligents vivants, le couplage structurel constitutif de la fabrication d'un système intelligent. Couplage structurel, s'effectuant dans l'histoire de l'apprentissage profond de ce système synthétique, et consistant à progressivement recueillir et choisir (au sens étymologique latin de l'intelligence) les bons poids attentionnels à travers lesquels il est pertinent de combiner et de façonner son système d'interconnexions catégorielles d'analyse de son expérience fonctionnelle d'interaction avec le monde des tokens. Et c'est cette restructuration catégorielle qui va dès lors permettre à ce système intelligent de « s'adjoindre à un monde de signification préexistant » ainsi que le dirait Varela, celui des tokens et des significations humaines auquel il convient de se coordonner afin de conceptualiser (au sens de Vergnaud) et d'agir efficacement.

8 Conclusion

Les modèles de langage organisent leur appropriation interne du monde en segmentant et en restructurant progressivement leurs modalités propres d'analyse des objets issus de ce monde, les tokens et leurs relations. La phénoménologie de restructuration catégorielle, que nous avons tenté de mettre en lumière dans cette présente étude, détoure et combine, à chaque nouvelle couche neuronale d'un réseau perceptron, des sous-dimensions clés des catégories de pensée de la couche précédente, pour former de nouvelles catégories synthétiques toujours plus efficaces. Ces nouvelles catégories aident le système artificiel à mieux segmenter et traiter son expérience des mots et concepts du monde extérieur, afin de se coupler de façon fonctionnelle à ces éléments. Dans le cadre d'une prochaine étude, en cours de réalisation, nous tenterons de mieux comprendre dans quelle mesure les segmentations catégorielles générées par la restructuration synthétique sont associées, ou non, à une segmentation spécifique des espaces d'activation des neurones formels portant ces segmentations catégorielles.

Remerciements

Les auteurs remercient Madeleine Pichat pour sa relecture attentive de cet article et Stéphane Fadda (Sorbonne Center for Artificial Intelligence) pour les précieuses impulsions opérationnelles qu'il apporte à l'équipe Neocognition.

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] Barkan, R. (2021). The Role of Cognitive Biases in Human Decision Making. *Journal of Behavioral Decision Making*, 34(3), 243–255. DOI : 10,1002/bdm.2210.

[6] 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.

[7] 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.

[8] 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`

[9] 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

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

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

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

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

[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] 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`

[16] 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.

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

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

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

[24] Cowan, N. (2024). Working Memory Capacity : Theories and Applications. *Annual Review of Psychology*, 75, 1–25. DOI : 10,1146/annurev-psych-010723-120001.

[25] 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

[26] 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

[27] 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.

[28] 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

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

[30] 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.

[31] 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.

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

[33] Enguehard, J. (2023). Extrmask : A Method for Explaining Time Series Predictions by Masking. *arXiv preprint arXiv :2301.08552*. DOI : 10,48550/arXiv.2301.08552.

[34] Eysenck, M. W., & Keane, M. T. (2020). *Cognitive Psychology : A Student's Handbook* (8th ed.). Psychology Press. DOI : 10,4324/9780429449229.

[35] 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.

[36] 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.

[37] 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.

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

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

[40] 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.

[41] 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,

[42] 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.

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

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

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

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

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

[48] 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.

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

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

[51] 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.

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

[53] 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.

[54] 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.

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

[56] 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).

[57] 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.

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

[59] 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`

[60] 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`

[61] 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.

[62] 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.

[63] 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 preprint*, arXiv :submit/6145488 [cs.AI].

[64] 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,

[65] 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.

[66] 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.

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

[68] 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

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

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

[71] 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.

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

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

[74] 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.

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

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

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

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

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

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

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

[82] 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

[83] 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.

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

[85] 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

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

[87] 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,

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

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

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

[91] 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.

[92] 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`.

[93] 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.

[94] 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.

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

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

[97] 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.

[98] 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

[99] 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.

[100] 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.

[101] 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.

[102] 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.

[103] 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*.