Comment pensent les IA ? Le processus de détourage catégoriel au cœur de la genèse des concepts de la cognition neuronale synthétique

Michael Pichat[1,2,4], William Pogrund[1,5], Paloma Pichat[1,3], Armanouche Gasparian[1], Samuel Demarchi[1,4], Martin Corbet[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 21 janvier 2025.

Abstract

Cet article investigue, dans le champ de la neuropsychologie de l'intelligence artificielle, le processus de segmentation catégorielle effectué par des modèles de langage, qui consiste, au fil des différentes couches neuronales, à créer de nouvelles dimensions catégorielles fonctionnelles pour analyser les données textuelles introduites et réaliser les tâches requises. Chaque neurone d'un réseau MLP (multilayer perceptron) est associé à une catégorie spécifique, générée par trois facteurs portés par la fonction d'agrégation neuronale : l'amorçage catégoriel, l'attention catégorielle et le phasage catégoriel. A chaque nouvelle couche, ces facteurs président à la formation de nouvelles catégories à partir des catégories des neurones précurseurs. A travers un processus de détourage catégoriel, ces nouvelles catégories sont fabriquées par une extraction sélective de sous-dimensions spécifiques des catégories antécédentes, en construisant une distinction d'une forme d'un fond catégoriel. Nous évoquons de façon exploratoire plusieurs caractéristiques cognitives de ce détourage synthétique : la réduction catégorielle, la sélectivité catégorielle, la séparation des dimensions initiales d'embedding et la segmentation de zones catégorielles.

1 Introduction

La segmentation catégorielle synthétique est l'activité des neurones formels consistant à «découper», dans le monde des tokens auquel un modèle de langage est exposé, une dimension de caractéristiques utilisée pour analyser et catégoriser ces tokens, un concept-en-acte tenu pour pertinent étant donnée la nature de la tâche à réaliser par ce modèle dirait Vergnaud [133, 134]dans le champ de la psychologie cognitive du développement humain. Le résultat de cette segmentation se traduit par la création, par chaque neurone, d'une catégorie synthétique de pensée, d'un concept, ou pour le dire encore autrement, d'une dimension catégorielle portée par ce neurone [101, 102]. Cette catégorie conceptuelle synthétique est, entre autres, définie par son extension, c'est-à-dire l'ensemble des tokens pour lesquels le neurone associé à cette catégorie est (suffisamment) activé.

Dans un précédent travail [105], nous avons investigué des facteurs mathématico-cognitifs de la segmentation catégorielle opérée par les neurones synthétiques des modèles de langage. Dans cette étude exploratoire préalable, nous avons examiné, à la fois de manière quantitative et qualitative, les éléments génétiques influençant cette segmentation. En nous basant sur la fonction d'agrégation [1] de la forme $\Sigma(w_{i,j}x_{i,j}) + b$, qui pilote pour partie ce processus cognitif, nous avons identifié trois éléments-clés causaux d'ordre mathématique et cognitif impliqués dans ce processus de partition conceptuelle.

Premièrement, l'« effet x » ou amorçage catégoriel synthétique, se rapporte au fait que l'activation des catégories portées par les neurones précurseurs d'une couche n se répercute sur l'activation des catégories propres à leurs neurones d'arrivée associés en couche $n + 1$, impactant ainsi directement leur extension catégorielle. Autrement dit, plus un token appartient à l'extension d'une catégorie précurseure en couche n (i.e. plus ce token est activé dans le neurone impliqué), plus il est doté d'un potentiel d'appartenance à l'extension de sa catégorie superordonnée (i.e. d'un potentiel d'activation dans le neurone de couche $n + 1$). Ce phénomène d'amorçage catégoriel pilotant ainsi pour partie la segmentation catégorielle opérée en couche $n + 1$, c'est-à-dire la détermination du sous-ensemble de tokens constitutifs de l'extension catégorielle des concepts portés par les neurones en couche $n + 1$.

Deuxièmement, l'« effet w » ou l'attention catégorielle synthétique, a trait au fait que les poids des connexions entre un neurone d'arrivée (couche $n + 1$) et ses neurones précurseurs (couche n) pilotent le degré de pertinence accordé aux

1. Bills et al. (2023), dont nous allons exploiter les données dans le cadre de cette présente étude, au sein de leur compte github associé à leur article, indiquent une liste «of the upstream and downstream neurons with the most positive and negative connections» ; liste dont ils donnent la définition opérationnelle suivante : «Definition of connection weights : neuron-neuron : for two neurons (l1, n1) and (l2, n2) with l1 < l2, the connection strength is defined as $h\{l1\}.mlp.c\_proj.w[:, n1, :]$ @ diag($h\{l2\}.ln\_2.g$) @ $h\{l2\}.mlp.c\_fc.w[:, :, n2]$." Cette liste définit, dans le cadre des couches denses (i.e. couches pleinement connectées) de GPT2-XL, les poids à travers lesquels chaque neurone d'une couche d'arrivée $n + 1$ est relié à l'ensemble des neurones de la couche précédente n. C'est sur la base de ces poids que les fonctions linéaires d'agrégations neuronales, auxquelles nous nous référons dans ce présent article, opèrent.

catégories précurseures dans la construction du segment catégoriel des neurones d'arrivée correspondant ; cela se manifestant qualitativement par un processus de complémentation catégorielle consistant génétiquement à « apporter » à la constitution de l'extension (de tokens) d'une catégorie d'arrivée une sous-dimension catégorielle spécifique extraite de chaque catégorie précurseure, apport qui est fonction de l'intensité de la focalisation attentionnelle dont cette catégorie précurseure est l'objet. Le segment catégoriel propre à un neurone d'arrivée se retrouvant ainsi constitué (quant à son extension de tokens), sous-segment par sous-segment catégoriel antécédant sémantiquement complémentaire et extrait des extensions des catégories sous-ordonnées.

Enfin, l'«effet Σ "ou phasage catégoriel synthétique, se réfère au phénomène de cognition synthétique par lequel des sous-segments des segments catégoriels précurseurs (i.e. des tokens de neurones de la couche n) identiques qui se retrouvent activés simultanément au niveau des neurones d'arrivée (couche $n + 1$) rentrent alors en écho catégoriel, présidant ainsi pour partie à la détermination des tokens constitutifs de l'extension des catégories de ces neurones 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 des catégories d'arrivée, c'est-à-dire la segmentation catégorielle opérée les neurones de destination impliqués. L'extraction de sous-dimensions catégorielles ici réalisée dans les catégories précurseures étant une extraction de sous-dimensions 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.

Ces trois facteurs mathématico-cognitifs causaux de la segmentation catégorielle orchestrent, nous l'avons mentionné, au niveau d'un neurone d'arrivée (couche $n + 1$), un mécanisme 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 ainsi, partie par partie, le contenu de l'extension (en termes de tokens) de la catégorie de ce neurone superordonné, c'est-à-dire du segment catégoriel spécifiquement associé à ce neurone d'arrivée. Ce mécanisme conceptuel synthétique d'extraction de concept, largement étudié en psychologie cognitive via ses équivalents humains [15, 55, 42, 45, 9, 78, 150], est fascinant du point de vue épistémologique et participe à la construction de la « réalité » opérée par la cognition artificielle dans son interaction avec le monde des tokens qui lui fait face.

Notre étude préalable, que nous venons ici de résumer, étudiait le processus de l'extraction catégorielle synthétique en tentant de mettre en lumière comment il était le fruit d'une activité réalisée simultanément, au niveau d'un neurone d'arrivée, sur ces différents neurones précurseurs. Autrement dit, nous avons étudié comment l'extension (de tokens) de la catégorie d'un neurone d'arrivée était générée par l'interaction conjointe (en termes d'activation, d'intersection ou de complémentation catégorielle) des extensions des catégories de ses neurones prédécesseurs, chaque prédécesseur contribuant à une partie de cette extension.

Dans le cadre de la présente recherche, plus localisée, nous tentons maintenant de mieux appréhender par quoi cette phénoménologie d'extraction catégorielle se manifeste «émantiquement» en termes d'abstraction catégorielle au niveau local de la catégorie de chaque neurone précurseur donné ; c'est-à-dire en termes d'extraction de certains tokens (constitutifs d'une sous-dimension catégorielle précise) et non pas d'autres qui vont dès lors constituer le «fond catégoriel» duquel est détachée la sous-dimension catégorielle spécifiquement abstraite d'une catégorie précurseure donnée. Il s'agit dès lors de cerner comment un « détourage catégoriel » est réalisé sur la variabilité catégorielle relative des tokens constitutifs de l'extension de la catégorie de chaque neurone précurseur afin d'extraire, de chacune d'entre elles, un sous-ensemble de tokens catégoriellement homogènes (vis-à-vis d'une sous-dimension catégorielle donnée) et alignés avec la (nouvelle) catégorie spécifique que fabrique singulièrement leur neurone successeur corollaire.

2 Extraction catégorielle, conceptualisation et abstraction

Les capacités d'abstraction et de conceptualisation constituent une aptitude d'extraction catégorielle fondamentale de la cognition humaine. Elles jouent un rôle central dans l'intelligence [67] et rendent possible le raisonnement, la généralisation et la résolution de problèmes [34]. Elles apportent également une habileté-clé pour l'apprentissage, l'adaptation robuste de la connaissance à un nouveau domaine et l'analogie [82, 59].

Mais si les êtres humains font montre d'une forte aptitude à l'abstraction et à la conceptualisation, en composant de nouveaux concepts à partir de concepts préalables, la conduite de ces processus cognitifs est plus sensible pour les systèmes d'intelligence artificielle [125]. En effet, si les humains excellent à extraire des patterns abstraits à partir de différentes séquences, à filtrer les détails non pertinents et à transférer ces concepts généralisés à d'autres séquences, les réseaux neuronaux artificiels sont réputés éprouver plus de difficultés en la matière [145, 146].

Au sein de la théorie classique des concepts [80], la philosophie définissait l'abstraction comme l'extraction de définitions (i.e. de traits définitoires) constitutives de conditions nécessaires et suffisantes pour l'appartenance d'un élément à l'extension d'un concept. Les philosophes empiristes, par la suite, pensaient l'abstraction comme le fruit de l'extraction de traits communs à diverses expériences sensorielles ou à leurs stockages mnésiques, ces abstractions étant alors issues d'une «distillation» opérée à partir d'impressions concrètes incorporées dans la perception.

La question de l'abstraction est *de facto* épistémologiquement intimement liée à celle de la catégorisation [103, 104]. Dans le champ de la psychologie cognitive et développementale humaine, différentes approches de la nature du contenu et de l'organisation des catégories sont abordées, chacune se focalisant sur certains aspects (complémentaires ou divergeants) des entités conceptuelles :

les connaissances fondamentales [21, 121], les exemplaires et les prototypes comme entités nodales étalonnant les concepts [92, 115, 123, 138, 90]les réseaux sémantiques définissant la structuration des concepts entre eux [25, 26, 60], la finalisation des catégories dans une perspective de cognition située [10, 48, 73], la dimension métaphorisée des concepts [68], les concepts comme modèles mentaux ancrés dans l'activité perceptive [69], le rapport des concepts et d'un langage de la pensée au sein d'une approche probabiliste stochastique [52]. Ces différentes approches pointent chacune différentes modalités, fonctionnalités ou finalités de l'abstraction ; mais c'est sans doute Vergnaud [133, 134] qui les lie les plus directement en mettant l'accent sur le fait que la catégorisation conceptuelle est fondamentalement une activité d'abstraction et d'extraction de caractéristiques tenues pour pertinentes (concepts-en-acte) ou vraies (théorèmes-en-acte) dans le cadre d'une activité donnée.

3 La nature des contenus extraits par l'abstraction et la conceptualisation

Mais par quoi se traduit l'extraction catégorielle constitutive de l'abstraction et de la conceptualisation ? Autrement dit, quelle est la nature cognitive ou catégorielle de ce qui est extrait dans le processus d'abstraction ou de conceptualisation ?

Dans une épistémologie d'inspiration empiriste, le processus de l'abstraction est régulièrement positionné au sein d'une dichotomie concret / abstrait, l'abstraction catégorielle étant alors abordée en termes d'extraction d'éléments plus éthérés à partir d'éléments tangibles. Ainsi Cuccio [28], dans le domaine des neurosciences, indiquent que les catégories abstraites sont générées sur la base de concepts concrets qui constituent des abstractions préalablement construites. Dans cette lignée, Pulvermuller [110], et en pointant des limites de l'amodalité des approches purement liées aux traits définitoires et déconnectées du contexte des objets et des actions, met en lumière des mécanismes neurobiologiques de l'extraction de traits sémantiques communs en soulignant l'ancrage du symbolique dans les entités du monde «réel» ; cela, à travers une modélisation permettant d'identifier les processus de l'enracinement des symboles, des catégories et des concepts dans les entités du monde réel, en termes d'associations entre signes et éléments matériels de la référence aboutissant à la formation de représentations abstraites par extraction de traits sémantiques ; l'auteur pointe alors des représentations non symboliques sous la forme de circuits sémantiques émergeants dispersés dans différentes régions corticales spécifiques, en fonction de la nature sémantique du signe impliqué (mots d'action => aires motrices ; mots sensoriels => aires sensorielles). Dans le cadre spécifique du *machine learning*, les études ayant trait à la vision et au traitement d'images sont sources de nombreuses investigations relatives à la nature de l'extraction catégorielle. Elles font montre de l'extraction de traits de base (couleur, texture), complétée par des caractéristiques extraites plus complexes [126, 17, 75, 71, 54] ; ou encore de la

combinaison de catégories simples (dotées d'une référence simple et identifiable, perçue par les sens), lorsqu'elles sont en nombre suffisant, en catégories, plus abstraites, extraites [146] ; ou enfin de la construction et la généralisation de représentations abstraites à partir de séquences concrètes par extraction et segmentation de traits partagés d'items apparaissant dans le même contexte [145].

Diverses études, notamment dans le champ de la neurobiologie, décrivent le processus d'abstraction sous un angle d'extraction de cartes cognitives d'espaces mentaux relativement à un domaine de connaissance donné [122, 93, 41, 72, 94, 86] ; ces cartes étant par exemple produites par des neurones biologiques (de place et de grille) qui rendent possible une représentation des expériences et des souvenirs. Différentes techniques de simulation de l'émergence de ces cartes cognitives extraites des données peuvent être mises en place à partir de réseaux de neurones formels apprenant sur la base d'unités statistiques dont les caractéristiques sont encodées par des vecteurs de traits ; techniques au premier rang desquelles figurent les méthodologies de réduction de dimension et / ou de clusterisation opérant à partir de ces vecteurs.

Dans leur passionnante étude, Ponomarev & Agafonov [107] (en continuité de Sousa et al.[124]),modélisent l'extraction conceptuelle réalisée par les réseaux de neurones en termes d'ontologies spécifiquement localisées au niveau des activations des couches neuronales artificielles ; cela, pour aboutir, dans le domaine du traitement visuel, à des cartes de compositions d'arguments et de prédicats conceptuels abstraits de zones neuronales précises de ces couches. Fel et al.[44], quant à eux, montrent l'importance de l'étude locale des concepts à des fins de compréhension fine de la nature et de l'organisation des caractéristiques conceptuelles extraites par les réseaux artificiels ; en mobilisant, à partir de graphes de clusterisation opérés sur l'espace vectoriel des activations, des représentations visuelles des principaux traits utilisés par un modèle de vision pour extirper des concepts ; ainsi, expliquent les auteurs, le concept visuel d'«espresso» est-il le fruit du dégagement de traits visuels de type «bulles et mousse sur le café», «latte art», «tasses transparentes avec mousse et liquide noir», «anse de la tasse à café», « café dans la tasse» ; cela, par-delà de plus classiques approches de type «cartes de températures» se limitant plus à montrer le «où» que le «quoi» de l'extraction catégorielle [53] ; [13].

Citons enfin les études structurales et formelles, donnant à voir de façon fine la nature des extractions catégorielles réalisées par les neurones artificiels en termes de sous-espaces conceptuels, constitutifs de traits définitoires, extirpés à partir de données perceptives brutes [24] ; ou, dans une lignée analogue, les travaux modélisant l'extraction conceptuelle en termes d'espace de représentation doté de dimensions ou de sous-espaces séparables et constitutifs de traits conceptuels [57, 58].

4 Problématique

4.1 Le détourage catégoriel

Dans le domaine de l'édition d'images ou de vidéos et du graphisme, le terme de détourage désigne le processus de séparation d'un élément d'une scène visuelle (statique ou dynamique) de son arrière-plan. Cela, afin mettre en avant cet élément ainsi isolé ; ou afin de le manipuler par la suite, par exemple en l'intégrant au sein d'un autre champ visuel, en en faisant ressortir certains détails, en en changeant certaines nuances de couleurs ou encore en en améliorant la netteté. Le détourage impliquant fondamentalement une opération de délimitation du contour de ce qui est instancié comme étant l'objet pertinent à dissocier de ce qui est défini comme étant un fond, du bruit tenu pour non pertinent pour le dire dans un registre du traitement du signal.

Par analogie, nous définissons, dans le domaine de la cognition synthétique, le détourage catégoriel comme la création, l'instanciation, l'enaction (au sens de Varela)[130] d'une (sous-)dimension singulière au sein de l'espace vectoriel, de dimension infinie, des caractéristiques qu'il est possible d'attribuer à une entité donnée. Nous parlons bien ici de création originale et non pas d'identification d'une (sous-)dimension préexistante qui serait déjà disponible au sein d'un monde pré-donné de caractéristiques dotées d'une existence ontologique [131]. Précisons de plus que cette (sous-)dimension peut être analogue ou non à une catégorie de pensée actuellement existante dans la pensée humaine.

4.2 Détourage catégoriel et construction de la séparation forme / fond

Les études précédemment mentionnées, dans les champs de la psychologie humaine, des neurosciences et des réseaux de neurones, abordent avec intérêt la nature des contenus et de l'organisation des éléments extraits par l'abstraction et la conceptualisation en termes d'exemplaires et prototypes, de modèles mentaux, de traits de base issus d'éléments concrets, de circuits sémantiques émergeants, de cartes cognitives d'espaces mentaux, d'ontologies, de clusters représentationnels, etc. Ces diverses approches font montre de diverses modélisations possibles et intrigantes du résultat de ces contenus extraits. Mais à travers quels processus ces contenus sont-ils extraits par les neurones formels eux-mêmes ? Par quoi se manifeste cognitivement ou catégoriellement le processus de détourage catégoriel à l'endroit de ces contenus ? Par quelle phénoménologie se traduit, dans le cadre de son activité d'abstraction et de conceptualisation, la «décision» par la cognition synthétique de ce qui est la forme conceptuelle qui est à retenir et qui devra être dissociée d'un fond ?

Dans le champ des neurosciences humaines, Savioz et al. [117] décrivent des aspects de l'effet de la fonction d'activation sur la construction de la dissociation entre fond et forme en précisant comment les neuromodulateurs agissent sur la plasticité synaptique [77]. En effet, la dopamine et la noradrénaline provoquent un renforcement des synapses activées et, de façon inverse, un affaiblissement

des synapses non activées, générant dès lors un contraste entre un signal et un bruit de fond. Ce process pouvant être modélisé par une fonction sigmoïdale de transfert [119], dotée d'un paramètre de gain G . L'augmentation de G entraine alors des activations neuronales plus fortes suite à des inputs positifs et des inactivations plus fortes en réponse à des inputs inhibiteurs. Les inputs sont ainsi mieux discriminés et les stimuli pertinents sont mieux détectés sur un bruit de fond [61, 51]. Cela aboutissant à des représentations corticales plus distinctes et à une meilleure inhibition des informations non pertinentes. Zeki [149], montre par exemple l'effet de ce processus d'extraction d'une figure à partir d'un fond dans le traitement cortical des informations visuelles au niveau des aires occipitales V1 à V4.

Ces éléments manifestent que le processus de détourage dans la cognition humaine implique fondamentalement un mécanisme quantitatif : les inputs les plus forts sont amplifiés, les inputs les plus faibles sont inhibés. Il en est de même dans la cognition synthétique dans la mesure où elle est également portée par une fonction d'activation composée à une fonction d'agrégation ; fonction d'agrégation qui pilote de façon causale, nous l'avons mentionné précédemment à travers les facteurs mathématico-cognitifs d'amorçage, d'attention et de phasage catégoriels, l'intensité d'activation qui va spécifiquement être allouée à certains tokens à la différence d'autres tokens. Mais comment mieux comprendre de façon qualitative cette phénoménologie quantitative ? Par quelles propriétés qualitatives cognitives et catégorielles se traduit ce processus quantitatif de détermination, de construction de ce qui est la «forme» pertinente à extraire d'un «fond» tout aussi cognitivement construit ? De façon plus cernée, au niveau d'un neurone précurseur, quelles sont les propriétés (résultantes de l'agrégation opérée par un de ses neurones d'arrivée associé) du mécanisme d'abstraction d'une sous-dimension catégorielle (la forme catégorielle retenue) au sein de la dimension catégorielle portée par ce neurone précurseur ?

4.3 Abstraction réfléchissante et mode opératoire cognitivo-mathématique du détourage catégoriel

Jean Piaget est certainement l'un des chercheurs, dans le champ de la psychologie humaine, à avoir le plus profondément investigué conceptuellement et empiriquement la notion d'abstraction. Le référentiel théorique qu'il propose offre un précieux cadre heuristique pour penser la question du mode opératoire du processus de détourage catégoriel dans son activité de différenciation d'une forme (dimension catégorielle) d'un fond dans le domaine de la cognition synthétique. Piaget [19] distingue deux types d'abstraction (nous n'abordons pas ici le cas de l'abstraction réfléchie ni le sous cas de l'abstraction pseudo-empirique).

L'abstraction simple (dite également empirique) opère sur les dimensions «matérielles» et «immédiatement observables» d'un ensemble d'objets (ou d'actions). Elle porte sur des dimensions physiques «imposées» par la perception et «inhérentes» à l'objet (son poids, sa texture, sa couleur) ou à l'action (sa direction, sa force). Cette abstraction est pour partie liée à celle postulée par l'empirisme philosophique. Mais pour partie seulement, car le constructivisme de

Piaget précise qu'elle nécessite néanmoins des cadres de connaissance qui ont été antérieurement générés par une abstraction réfléchissante. Par exemple, la couleur n'est pas une donnée pleinement immédiate mais présuppose une catégorisation et une sériation des impressions en provenance de longueurs d'onde variées perçues, catégorisation non directement extraite de la «réalité» par abstraction empirique [85]Tout comme, même en physique, les grandeurs mesurées (la masse, la force, l'accélération, etc.) sont elles-mêmes construites et dès lors la résultante d'inférences issues de précédentes abstractions réfléchissantes [98].

L'abstraction réfléchissante est, quant à elle, l'activité consistant à identifier une dimension associée à un objet puis à mobiliser cette dimension comme un élément au sein d'une structure plus large et différente du seul cadre de la perception (à la différence donc de l'abstraction simple). Plus précisément, l'abstraction réfléchissante relève des activités et des coordinations mentales que l'individu opère sur des dimensions extraites, sans plus avoir besoin d'avoir comme support proximal de ces dernières les objets ou les actions qui leur sont associés [89] ; dès lors, ce qui est extrait ici ne sont pas des informations assez proches de «caractéristiques du monde », mais des modes de structuration que l'individu a lui-même introduit dans la «réalité». Dans ce registre Piaget mentionne le passage de l'arithmétique à l'algèbre qui implique une abstraction des opérations et des relations numériques, sans n'avoir plus besoin de nombres effectifs auxquels les appliquer.

Processus de formation de connaissances de nature plus endogène que l'abstraction empirique, l'abstraction réfléchissante piagétienne se déroule en trois temps [85] : (i) l'abstraction proprement dite, (ii) le réfléchissement, (iii) la réflexion. L'abstraction à proprement parler ne porte pas sur une propriété du monde physique (ex : la couleur), elle n'est donc pas une abstraction empirique : elle consiste à extraire une propriété de l'activité-même de l'individu, par exemple le fait de coordonner des éléments (ex : réunir, ordonner, mettre en correspondance). Ce premier moment de l'abstraction produit soit une connaissance nouvelle (par objectivation : un instrument de pensée devient lui-même un objet de pensée, un nouveau concept), soit un nouvel outil de la pensée (un schème). Deuxième moment du déroulement de l'abstraction réfléchissante, le réfléchissement se traduit par une projection sur un plan supérieur de connaissance (par sa nature ou sa complexité) de ce qui a été extrait du palier inférieur. Cette transposition sur un plan plus élaboré rendant possible une nouvelle utilisation, plus abstraite et décontingentée de la dimension ou de l'élément extrait. Dernier moment de l'abstraction réfléchissante, la réflexion relève d'une authentique reconstruction et réorganisation de la dimension ou de l'élément projeté sur le plan supérieur. Dimension ou élément qui va dès lors pouvoir être (i) traduit dans les termes (plus abstraits) du nouveau plan et (ii) manipulé mentalement comme un élément parmi d'autres et ainsi combiné à d'autres dimensions ou éléments eux-mêmes extraits.

4.4 Détourage catégoriel et abstraction réfléchissante synthétique

Les trois temps de l'abstraction réfléchissante, tels que définis par Piaget, éclairent les modalités qualitatives par lesquelles la fonction d'agrégation $\Sigma(w_{i,j}x_{i,j}) + b$, au sein des réseaux de neurones artificiels, «décollent» une forme d'un fond, une sous-dimension singulière à partir d'une infinité possible de sous-dimensions de segmentation catégorielle du monde (des tokens). En effet, l'activité de cette fonction d'agrégation peut être interprétée comme opérant ces trois temps comme suit. Premièrement, à travers une toute première abstraction («abstraction à proprement parler»), la fonction d'agrégation extrait une dimension, non pas intrinsèquement présente dans les tokens eux-mêmes mais fruit de l'activité-même du réseau de neurones, mais la dimension catégorielle $x_{i,j}$ construite dans l'espace d'activation de chaque neurone d'une couche n. Puis, dans un deuxième temps, la fonction d'agrégation, dans une activité de réfléchissement, projette cette dimension catégorielle extraite sur un nouveau plan vectoriel (représentationnel), plus abstrait, celui de la couche $n + 1$de neurones superordonnés. Enfin, en une dernière temporalité propre de réflexion, et dans ce nouveau plan vectoriel, les dimensions extraites sont l'objet elles-mêmes de nouvelles activités, de pondération $w_{i,j}$, de sommation Σ d'adjonction de biais b; activités qui, couplées à la fonction d'activation f_a, vont créer de toute pièce une dimension $f_a(\Sigma(w_{i,j}x_{i,j}) + b).$; cette nouvelle forme dimensionnelle ayant été ici (i.e. au niveau de cette étape spécifique de réflexion au sens de Piaget) singulièrement fabriquée à partir de la sous-dimension catégorielle spécifiquement détourée de chaque dimension catégorielle (le fond catégoriel) des neurones précurseurs de couche n.

Ainsi que l'écrit magnifiquement von Glaserfeld [49]en citant von Humboldt : « in order to reflect, the mind (...) must grap as a unit what was just presented, and thus posit it as a object again itself. The mind then compares the units, of which several can be created in that way, and separates and connects them according to its needs ». L'auteur complète (idem, pp.90-91) : « 'to grasp as a unit what was just presented' is to cut it out of the continuous experimental flow. In the literal sense of the term, this is a kind of abstraction (...). Focused attention picks a chunk of experience, isolates it from what came before and from what follows, and treats is as a closed entity ». Ainsi le détourage catégoriel opérant au niveau de la catégorie de chaque neurone précurseur non pas "arrache" mais fabrique à proprement parler une sous-dimension catégorielle originale sur la base d'une activité d'abstraction réflexive se traduisant par une singulière recombinaison pondérée de dimensions catégorielles d'entrée.

Mais, encore et toujours, par quoi cette abstraction réflexive, portant le détourage catégoriel opéré par chaque neurone d'une couche n à travers sa fonction d'agrégation propre, se traduit-elle catégoriellement, cognitivement ou encore sémantiquement ? Telle est la question à laquelle nous allons tenter d'apporter des éléments de réponse, ici choisis en termes de réduction sémantique, de séparation des embeddings des espaces vectoriels de GPT2-XL, de réorganisation de ces dimensions de ces embeddings et de nature sémantique des sous-dimensions

catégorielles extraites. Ainsi sont les propriétés et phénoménologies du détourage catégoriel vectorisé par l'abstraction réfléchissante synthétique que nous allons tenter de commencer à explorer dans cette présente investigation.

5 Méthodologie

5.1 Positionnement méthodologique

Pour situer méthodologiquement notre présente étude exploratoire, nous présentons ici un aperçu concis de diverses méthodes d'explicabilité visant, avec différents niveaux de détail cognitif, à extraire le contenu ou les processus informationnels des réseaux de neurones formels, qu'ils soient structurés en couches, en groupes ou en réseau complet.

Les recherches à large spectre cognitif se penchent sur l'analyse des écarts entre entrées et sorties, cherchant à élucider le lien entre les données initiales et les résultats dans un modèle de langage. Parmi ces approches, les méthodes basées sur les gradients évaluent le rôle de chaque donnée d'entrée en exploitant les dérivées de chaque dimension d'entrée [40]. Les caractéristiques des entrées peuvent être évaluées à travers des éléments tels que les traits [32], les scores d'importance des tokens [40], ou encore les poids d'attention[5]. . Par ailleurs, les approches par exemples examinent comment les sorties évoluent face à différents inputs, en observant les effets de légères modifications des entrées [140]ou d'altérations telles que la suppression, la négation, le mélange ou le masquage des entrées [4, 143, 129]. Certains travaux s'intéressent enfin au mapping conceptuel des entrées pour évaluer leur contribution aux outputs observés [22].

Les méthodes à granularité cognitive plus fine se concentrent sur les états intermédiaires du modèle de langage plutôt que sur sa sortie finale, analysant les sorties ou états internes partiels de neurones ou de groupes de neurones. Dans ce cadre, certaines approches analysent et décomposent linéairement le score d'activation d'un neurone d'une couche spécifique en relation avec ses entrées (neurones, têtes d'attention ou tokens) dans la couche précédente [139]D'autres cherchent à simplifier les fonctions d'activation pour en améliorer l'interprétation [140]. Certaines techniques, en s'appuyant sur le vocabulaire du modèle, s'orientent vers l'extraction des connaissances encodées en projetant les connexions et représentations intermédiaires via une matrice de correspondance [33, 47]. Enfin, certaines méthodologies s'appuient sur les statistiques d'activation neuronale en réponse à des ensembles de données [12, 84, 38, 140, 29]. Notre étude exploratoire actuelle s'insère spécifiquement dans ce dernier groupe de démarches.

5.2 Options méthodologiques

Dans notre recherche exploratoire, nous avons choisi d'explorer le modèle GPT développé par OpenAI, en nous concentrant spécifiquement sur la version GPT-2XL. Ce choix s'explique par le fait que GPT-2XL offre une complexité

suffisante pour analyser des phénomènes cognitifs synthétiques avancés, tout en restant moins complexe que GPT-4 ou sa version multimodale actuelle GPT-4o. Un aspect pratique a également influencé notre décision : en 2023, OpenAI a rendu disponible, grâce à l'article de Bills et al. (2023) [12],des informations détaillées sur les paramètres et valeurs d'activation des neurones du modèle, données essentielles pour notre étude.

Afin de simplifier notre exploration, nous nous sommes concentrés sur l'examen des deux premières couches de GPT-2XL, comprenant chacune 6400 neurones, soit un total de 12800 neurones artificiels. En ce qui concerne les tokens et leurs valeurs d'activation au sein de ces neurones, nous avons choisi d'analyser, pour chaque neurone, les 100 tokens ayant les valeurs d'activation moyennes les plus élevées, que nous avons appelés « core-tokens ».

Afin d'étudier la proximité sémantique entre tokens dans le cadre des premiers résultats statiques que nous présenterons ci-après, nous avons fait le choix de mesurer le cosinus de similarité au sein de la base d'embeddings de GPT2-XL, et non pas dans la base par exemple de GPT-4 pourtant plus performante, afin de ne pas retomber dans la limite méthodologique mentionnée par Bills et al. [12] et Bricken (2023) [16]consistant à apparier des systèmes cognitifs synthétiques ne reposant pas sur le même système d'embeddings, c'est-à-dire pas sur le même référentiel de segmentation catégorielle ; même si, à des fins de comparaison et de vérification de la plausibilité de nos données, nous avons également eu recours à deux autres bases classiques d'embeddings librement disponibles : Alibaba-NLP/gte-large-en-v1.5 et BERT base model.

5.3 Choix statistiques

Pour nos analyses statistiques, nous avons utilisé les bibliothèques Python de la suite SciPy, en nous appuyant sur les recommandations de Howell (2024)[63] et Beaufils[11].

Pour évaluer la normalité de nos données, une exigence pour la réalisation de tests paramétriques, nous avons adopté une approche en deux volets. Premièrement, nous avons réalisé des tests inférentiels : le test de Shapiro-Wilk, adapté aux petits échantillons ; le test de Lilliefors, utile lorsque les paramètres de la distribution normale sont inconnus et estimés à partir des données ; le test de Kolmogorov-Smirnov, idéal pour les grands échantillons ; et enfin le test de Jarque-Bera, qui évalue la symétrie et l'aplatissement des données pour de grands échantillons. Deuxièmement, nous avons complété cette analyse par des indicateurs descriptifs tell que la symétrie (skewness) et l'aplatissement (kurtosis), et des méthodes graphiques telles que le QQ-plot pour confronter les distributions observées à une distribution normale théorique. En ce qui concerne l'évaluation de l'homoscédasticité (égalité des variances entre groupes), nous avons utilisé le test de Bartlett (sensible aux déviations de normalité) complété par le test de Levene (moins sensible à la non-normalité).

Les résultats, indiqués partiellement dans la suite de cet article, révèlent une normalité relative de nos données. En conséquence, nous avons principalement employé pour nos comparaisons de groupes et nos études de distribution :

- le test de Kruskall-Wallis, étudiant la relation entre une variable nominale définissant k groupes indépendants et une variable de classement ; cela, à travers un reclassement en rang de nos données numériques d'activation neuronale des tokens, et en respectant sa condition d'application d'effectifs de groupes strictement supérieurs à 5.
- le test du χ^2 univarié d'ajustement ; cela en respectant ses conditions d'application relatifs aux effectifs théoriques et aux effectifs bruts, ne nécessitant ainsi pas de variantes pour effectifs faibles (statistiques de Fisher ou Monte-Carlo).

Concernant notre étude statistique de l'orientation préférentielle des vecteurs de dimension d'embedding des tokens en fonction de la caractéristique de ces derniers d'être des « taken-tokens » ou des « left-tokens », nous avons mobilisé une approche de réduction dimensionnelle par analyse factorielle de type analyse en composantes principales (ACP). Une ACP a été réalisée pour les doublets neurone précurseur – neurone d'arrivée ; les neurones précurseurs étant ici des neurones de la couche 0 et les neurones d'arrivée les 10 neurones de la couche 1 avec lesquels chaque neurone précurseur présente le plus fort poids de connexion. Pour chaque ACP, les unités statistiques impliquées sont les 100 core-tokens du neurone précurseur, c'est-à-dire les 100 tokens en moyenne les plus activés au sein de ce neurone. Les variables ici utilisées sont les 1600 dimensions d'embeddings de GPT2-XL complétées de deux variables dichotomiques (0/1) antagonistes : « taken-token » et « left-token » ; un core-token d'un neurone précurseur étant un taken-token s'il est également un core-token de son neurone d'arrivée associé, ou un left-token si ce n'est pas le cas. Les conditions usuelles d'application de l'ACP sont : (i) un nombre d'unités statistiques supérieur à 100, (ii) un nombre d'unités statistiques supérieur à 10 fois le nombre de variables impliquées, (iii) une sphéricité vérifiée par la significativité (au seuil ($\alpha = 5\%$)) du test de Bartlett, (iv) une adéquation globale validée par un coefficient de Kaiser-Meyer-Olkin (KMO) supérieur à .5 voire .7. Ces conditions n'étant que partiellement respectées dans le cadre de notre présente étude, nos résultats en la matière seront à appréhender avec prudence et comme ayant une valeur heuristique pour penser des études complémentaires, dont l'étude t-STN supplémentaire qui sera directement présentée par la suite dans le cadre de cet article.

Nos options de paramétrage de l'ACP ont été les suivantes : (i) test de rotations orthogonales (de type varimax) et obliques, (ii) réduction préalable des données afin de les normaliser, (iii) vérification d'une qualité de représentation des variables sur les facteurs (communalité cos2) greater than 40%,supérieure à 40%, pour éviter des pertes de variance trop importantes durant la projection vectorielle sur les facteurs, (iv) examen des corrélations trop fortes ($|\rho| > .9$), afin de supprimer les variables potentiellement redondantes, (v) règle de Kaiser ($\lambda > 1$) associée à la règle du pourcentage de variance restituée ($\sigma^2 > .6$), afin de déterminer le nombre de facteurs à retenir, (vi) sur-pondération des deux variables «taken-token» et «left-token» à 1% (du nombre de dimensions d'embeddings), afin de formater l'ACP pour qu'elle produise un axe factoriel F2 recherché relatif au fait qu'un token soit un «taken-token» et «left-token», (vii)

sélection des 1671 neurones précurseurs de la couche 0 présentant un pourcentage de taken-tokens compris entre 15% et 85%, afin de garantir un nombre minimal de taken-tokens présents.

Afin d'explorer plus avant comment les 1600 dimensions des embeddings des tokens tendent à se distribuer préférentiellement dans l'espace vectoriel des embeddings de GPT2-XL en fonction du fait que ces tokens soient des taken-tokens ou des left-tokens, et compte-tenu des limites méthodologiques mentionnées ci-avant concernant notre étude ACP, nous avons mobilisé une autre approche de réduction dimensionnelle, la technique t-SNE. Une valeur ajoutée significative de cette dernière étant de ne pas présupposer une combinaison linéaire des variables d'origine dans la constitution des axes factoriels, ainsi que d'être significativement moins contraignante en termes de conditions d'application. Les unités statistiques et variables sont ici les mêmes que celles mentionnées concernant notre étude ACP, mais sans la présence des deux variables supplémentaires «taken-token» et «left-token». A l'instar de la technique précédente, une étude t-SNE a été réalisée sur les 1671 neurones précurseurs mentionnés ci-avant. Notre paramétrage de l'algorithme t-SNE a été le suivant : (i) standardisation préalable des données, (ii) perplexité fixée entre 5 et 50 et test d'une perplexité produisant les meilleurs résultats, (iii) nombre d'itérations compris entre 500 et 1000 (avec test du meilleur paramètre), (iv) taux d'apprentissage déterminé à 200.

5.4 Objectif et mise en œuvre de l'étude en termes d'observables statistiques

Une dimension catégorielle (au sens de Varela), à partir de laquelle va être détourée une sous-dimension catégorielle singulière, peut être *a minima* définie par trois types d'éléments :

- Sa fonction d'appartenance au sens de la logique floue (Zadeh, 1996 ; Wu et al., 2022)[148, 144], traduite comme suit. Soit X un ensemble et A un sous-ensemble flou de X caractérisé par une fonction d'appartenance $\mu_A : X \to [0, 1]$déterminant le niveau partiel (à la différence de la théorie classique des ensembles) d'appartenance d'un élément x de X à l'ensemble A. Cet ensemble flou étant doté d'une hauteur $h(A) = \max\{\mu_A(x) \mid x \in X\}$ et d'un noyau $\text{noy}(A) = \{x \in X \mid \mu_A(x) = 1\}$ (si A est normalisé, i.e., $h(A) = 1$)contenant les éléments x appartenant totalement à A. Cette fonction d'appartenance, dans le champ de la cognition synthétique catégorielle, étant vectorisée par la fonction d'agrégation (composée à la fonction d'activation) associée au neurone portant la dimension catégorielle impliquée.
- Son extension (Nadeau, 1999)[88]de tokens pour lesquels le neurone impliqué s'active, définie comme le support $\sup(A) = \{x \in X \mid \mu_A(x) > 0\}$. Et plus particulièrement une extension de core-tokens, i.e. de tokens fortement activés, déterminé par α-coupe $(A) = \{x \in X \mid \mu_A(x) \geq \alpha\}$, avec α suffisamment élevé.

- Sa compréhension, un «sens» attribué à cette dimension via une construction inférentielle interprétatie.

L'objectif de notre présente étude exploratoire est de mieux comprendre, dans le prolongement immédiat de nos récents travaux antérieurs (Pichat et al., 2024d), les propriétés du détourage catégoriel généré par les facteurs que sont l'amorçage, l'attention et le phasage catégoriels portés par la fonction d'agrégation neuronale. Propriétés du détourage catégoriel que nous allons empiriquement investiguer au niveau des core-tokens de chaque neurone précurseur d'une couche n (i.e. l'extension de la catégorie propre à chacun de ces précurseurs). Et détourage catégoriel consistant à extraire de ces core-tokens certains tokens spécifiques (les taken-tokens), et à en laisser d'autres (les left-tokens). Caractéristiques cognitives du processus synthétique de détourage catégoriel, que nous allons ici tenter d'investiguer :

- En termes de réduction catégorielle, opérationnalisée sur la base de distances sémantiques mesurées à partir de cosinus similarité (des dimensions d'embeddings de l'espace vectoriel d'entrée de GPT2-XL), dans notre premier lot d'études statistiques.
- Puis en termes de séparation comme de structuration de ces mêmes dimensions d'embeddings, en fonction de la nature «taken» ou «left» des tokens impliqués.

6 Résultats

6.1 Normalité des données

En regard de notre premier lot d'études statistiques ci-après, relatif à la notion de réduction catégorielle, et opérationnalisée en termes de distances sémantiques mesurées à partir de cosinus similarité, nous avons réalisé les vérifications de normalité et d'homogénéité des variances de nos données suivantes. Pour chaque neurone d'arrivée de la couche 1, les tests ont été initiés, parmi ses 10 neurones précurseurs (de la couche 0) à forts poids de connexion, uniquement sur les précurseurs pour lesquels le nombre taken-tokens comme le nombre de left-tokens est supérieur ou égal à 6 ; cela, afin d'avoir un nombre minimum de tokens de chaque classe. Soit un maximum théorique de 6.400 neurones d'arrivée (en couche 1) x 10 neurones précurseurs (en couche 0) = 64.000 études de clusters de taken-tokens (nommés «taken-clusters») ; et un nombre effectif de 9.007 taken-clusters étudiés (et de left-clusters pour le cas de l'homoscédasticité). Pour chaque étude, la variable impliquée est le cosinus similarité entre tous les tokens (différents) du cluster impliqué.

Le tableau n°1 rend compte d'une normalité contrastée des cosinus similarité entre les tokens des différents taken-clusters, avec des pourcentages de compatibilité des tests inférentiels avec une hypothèse de normalité variant de 52 à 93% à partir des embeddings de GPT2-XL, les plus fiables ; et des résultats plus faibles (de 21 à 73%) pour les autres embeddings. Le tableau n°2 fait de même quant à lui en montrant une homogénéité mitigée des variances des tokens

entre les taken et les left-clusters, avec 58% de tests inférentiels compatibles avec cette hypothèse pour les embeddings de GPT, et moins pour les autres (entre 29 et 43%). Ces résultats nous invitent à ne pas mobiliser par la suite de tests paramétriques (si ce n'est à titre d'indication) dont les conditions d'application ne sont dès lors que partiellement validées.

	GPT2-XL	Alibaba	BERT
% of $p_{sw} > .05$	52.52%	21.37%	29.57%
% of $p_{multiw} > .05$	64.39%	30.48%	32.71%
% of $p_{sx} > .05$	92.78%	65.36%	66.13%
% of $p_{ib} > .05$	79.52%	52.60%	72.58%
Mean	.537	.681	.907
Median	.524	.661	.913

Tableau n°1 : Statistiques des ratios de normalité des similarités cosinus entre taken-tokens de chaque taken-cluster (Couche 1 ; N=9007).

	GPT2-XL	Alibaba	BERT
% of $p_{varw} > .05$	58.61%	38.86%	42.61%
% of $p_{varbtw} > .05$	58.84%	29.12%	33.25%

Tableau n°2 : Statistiques des ratios d'homoscedasticité des similarités cosinus entre taken-tokens et left-tokens (Couche 1 ; N=9007).

6.2 Détourage et réduction catégorielle

Au sein des neurones artificiels des modèles de langage, ainsi que nous l'avons indiqué précédemment, le processus d'abstraction généré, entre autres, par la fonction d'agrégation$\Sigma(w_{i,j}x_{i,j}) + b$, «décolle» une forme catégorielle (construite) d'un fond, une sous-dimension catégorielle singulière à partir d'une infinité possible de sous-dimensions de segmentation catégorielle du monde (des tokens). Cela, étant généré par les trois mécanismes de la cognition synthétique, étayés mathématiquement par cette fonction d'agrégation, que sont l'amorçage, l'attention et le phasage catégoriels. Plus particulièrement, nous postulons que l'ultime étape de réflexion de l'abstraction réfléchissante piagétienne opérée par la fonction d'agrégation d'un neurone de couche $n + 1$, va provoquer l'extraction, le détourage d'une sous-dimension particulière de chacun de ses neurones précurseurs contributifs (i.e. sources effectives de taken-tokens) en couche n ; l'apposition successive, au niveau du neurone d'arrivée, de ces différents sous-segments catégoriels constituant ainsi «partie par partie», l'extension de la nouvelle catégorie propre à ce neurone d'arrivée, c'est-à-dire l'ensemble de ses core-tokens spécifiques.

Cette extraction, par un neurone d'arrivée, de sous-dimensions catégorielles spécifiques, à partir de la diversité catégorielle relative des core-tokens de chacun de ses neurones précurseurs, devrait se traduire par la segmentation de taken-clusters catégoriellement plus homogènes ; en tout cas, plus homogènes au titre d'un segment catégoriel, la sous-dimension catégorielle spécifiquement extraite de la dimension catégorielle de chaque précurseur. Autrement dit, nous postulons une première caractéristique du détourage catégoriel : la réduction catégorielle, se traduisant par le découpage, au sein de l'extension (de tokens) de la catégorie associée à chaque neurone précurseur, d'un cluster de tokens (i.e. un taken-cluster) présentant une moindre variabilité catégorielle par rapport à la dispersion de départ, cette dernière étant plus conceptuellement «resserrée» autour de la sous-dimension ainsi extraite.

Nous opérationnalisons ce postulat à travers deux hypothèses, méthodologiquement proches, chacune constituant cependant un angle de vue doté de sa valeur ajoutée propre. Premièrement, pour chaque neurone précurseur (contributif) de chaque neurone d'arrivée, le taken-cluster impliqué devrait présenter un cosinus similarité moyen (mesurant la proximité moyenne entre tous les tokens constitutifs de ce cluster) supérieur au cosinus similarité moyen des core-tokens associés à ce neurone précurseur. Deuxièmement, pour chaque neurone précurseur (contributif) de chaque neurone d'arrivée, le taken-cluster impliqué devrait présenter un cosinus similarité moyen supérieur au cosinus similarité moyen des left-tokens associés à ce neurone précurseur. Nous réalisons 9007 tests statistiques de ces deux hypothèses, correspondant aux 9007 cas de taken-clusters et left-clusters contenant 6 ou plus tokens (conformément aux conditions d'application du test non-paramétrique de Kruskall-Wallis).

En ce qui concerne notre première hypothèse, le tableau n°3 (et sa visualisation au niveau du graphe n°1) manifestent une supériorité moyenne de l'homogénéité catégorielle des taken-tokens par rapport à celle des core-tokens de .14 (pour un calcul à partir des embeddings de GPT2-XL), ce qui est relativement important dans la mesure où le cosinus similarité varie de 0 à 1 en valeur absolue. Avec un pourcentage de 95% de cas de clusters où l'écart d (d = mean(COS(taken-tokens)) − mean(COS(core-tokens)))est positif et cela, de façon largement significative ($p(\chi^2) < .0001$). De même, le pourcentage de cas présentant une différence significative (p<.05) de moyennes de proximité cosinus est important (81% avec un test de Kruskall-Wallis et même 86% pour un t de student, même si ce dernier indicateur est moins pertinent comme évoqué précédemment). Les autres systèmes d'embeddings pointent des résultats analogues, bien que moins forts, a priori en raison de leur moindre capacité à opérer des analyses sémantiques adaptées au mode de segmentation catégorielle de GPT2. Ces résultats sont compatibles avec notre première hypothèse.

	GPT2-XL	Alibaba	BERT
Mean(Mean(COS(taken-tokens)))	.537	.681	.907
Mean(Mean(COS(core-tokens)))	.399	.585	.892
Mean(d)	.139	.097	.015
% of (p(t) < .05)	85.77%	84.13%	49.38%
% of (p(KW) < .05)	81.33%	78.04%	60.73%
% of d > 0	94.64%	93.07%	68.05%
p(χ^2) of d > 0	4.36E-19	7.03E-18	3.69E-04

Tableau n°3 : Comparaison inférentielle des distances moyennes de similarité cosinus entre taken-tokens et core-tokens, pour chaque neurone précurseur pertinent de chaque neurone de destination (Couche 1 ; N=9007).

Graphique n° 1 : Comparaison descriptive des distances moyennes de similarité cosinus entre taken-tokens et core-tokens, pour chaque neurone précurseur pertinent de chaque neurone d'arrivée (Couche 1 ; N = 9007).

Le tableau n°4 et son graphe n°2 associés illustrent ces résultats globaux avec le cas particulier d'un neurone précurseur du neurone n°3000 de la couche 1.

	GPT2-XL	Alibaba	BERT
Mean(COS(taken-tokens))	.604	.678	.855
Mean(COS(core-tokens))	.444	.593	.889
d	.161	.085	-.034
p(t)	5.41E-06	4.26E-06	9.87E-01
p(KW)	3.24E-06	2.69E-06	2.13E-01

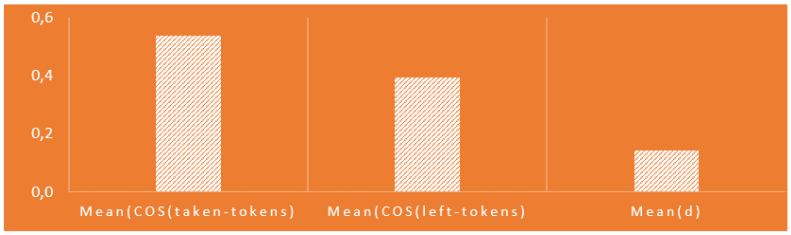
Tableau n°4 : Comparaison inférentielle des distances moyennes de similarité cosinus entre taken-tokens et core-tokens, pour un neurone précurseur du neurone 3000 (Couche 1).

Graphique n° 2 : Comparaison descriptive des distances moyennes de similarité cosinus entre taken-tokens et core-tokens, pour un neurone précurseur du neurone 3000 (Couche 1).

En ce qui concerne maintenant notre deuxième hypothèse, le tableau n°5 (et sa visualisation au sein du graphe n°3) montrent à nouveau une plus grande valeur moyenne de la proximité catégorielle des taken-tokens par rapport, cette fois-ci, à celle des left-tokens de 0.14 (embeddings GPT2-XL). Avec un pourcentage significatif ($p(\chi^2) < .0001$) de 94% de cas où d est positif. De plus, le pourcentage de cas associés à un écart significatif ($p < .05$) des moyennes de similarité cosinus est important (82% avec Kruskal-Wallis et 85% pour Student). Les autres systèmes d'embeddings pointent des résultats analogues, bien que moins marqués, a priori en raison de leur moindre capacité à opérer des analyses sémantiques adaptées au mode de segmentation catégorielle de GPT2-XL. Ces résultats, directement méthodologiquement complémentaires aux premiers, sont nécessairement également compatibles avec notre deuxième hypothèse.

	GPT2-XL	Alibaba	BERT
Mean(Mean(COS(taken-tokens)))	.537	.681	.907
Mean(Mean(COS(core-tokens)))	.394	.581	.892
Mean(d)	.143	.100	.015
% of (p(t)<.05)	85.48%	83.87%	50.11%
% of (p(KW)<.05)	81.65%	78.93%	63.44%
% of d>0	94.02%	92.54%	67.07%
$p(\chi^2)$ of d>0	1.33E-18	1.77E-147	6.00E-04

Tableau n° 5 : Comparaison inférentielle des distances moyennes de similarité cosinus entre taken-tokens et left-tokens, pour chaque neurone précurseur pertinent de chaque neurone destination (Couche 1 ; N=9007).

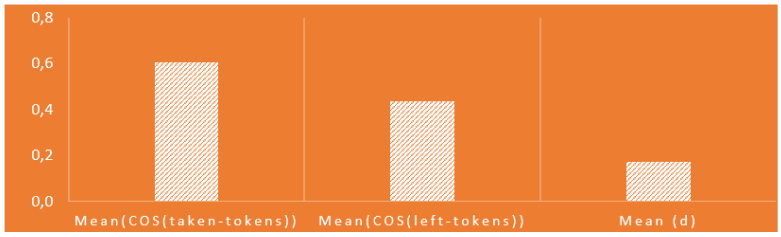


Graphe n° 3 : Comparaison descriptive des distances moyennes de similarité cosinus entre taken-tokens et left-tokens, pour chaque neurone précurseur pertinent de chaque neurone de destination (Couche 1 ; $N = 9007$).

Le tableau n°6 et son graphe n°4 corollaire illustrent ces résultats globaux avec de nouveau le cas d'un neurone précurseur du neurone n°3000 de la couche 1.

	GPT2-XL	Alibaba	BERT
Mean(COS(taken-tokens))	.604	.678	.855
Mean(COS(left-tokens))	.434	.590	.891
Mean(d)	.171	.088	-.036
$p(t)$	1.34E-06	2.10E-06	9.94E-01
$p(KW)$	9.55E-07	1.47E-06	1.81E-01

Tableau n° 6 : Comparaison inférentielle des distances moyennes de similarité cosinus entre taken-tokens et left-tokens, pour un neurone précurseur (Couche 1 ; Neurone de contrôle 3000).



Graphe n° 4 : Comparaison descriptive des distances moyennes de similarité cosinus entre taken-tokens et left-tokens, pour un neurone précurseur (Couche 1 ; Neurone de contrôle 3000).

L'ensemble des résultats ici obtenus sont ainsi compatibles avec notre postulat relatif au fait que le détourage catégoriel opéré par l'abstraction portée par la fonction d'agrégation se traduit par un processus de réduction catégorielle consistant pour chaque neurone d'arrivée en couche n+1 à extraire une sous-dimension catégorielle particulière de la dimension catégorielle associée à chacun de ses neurones précurseurs contributifs en couche n ; ce sous-segment

se caractérisant par un ensemble de tokens (des taken-tokens) plus homogènes entre eux, et constituant bien, dès lors, une sous-dimension catégorielle spécifique relativement unifiée et non pas un sous-groupe catégoriellement hétérogènes de tokens ou un sous-groupe de tokens catégoriellement équivalents à l'ensemble de leurs core-tokens de départ.

6.3 Détourage et sélectivité catégorielle

La caractéristique de réduction catégorielle pointée dans la section précédente nous conduit à formuler un autre postulat de propriété du détourage catégoriel porté par la fonction d'agrégation et son activité d'abstraction, celle d'une *sélectivité catégorielle*. Cette sélectivité catégorielle se traduisant par l'extraction, à partir de la catégorie propre à un neurone précurseur donné, d'une sous-dimension catégoriellement très restreinte quant à son extension propre (i.e. le nombre de tokens qu'elle contient), dans la mesure où cette catégorie précurseure présente *de facto* elle-même déjà une convergence catégorielle *a priori* conséquente et constitutive de la dimension catégorielle qu'elle vectorise.

Nous opérationnalisons ce postulat par l'hypothèse suivante : le cardinal des sous-ensembles de tokens que constituent les taken-clusters tend à être très faible par rapport au nombre (100) de core-tokens de départ de chaque neurone précurseur. Afin de tester cette hypothèse, nous fixons un seuil très bas, égal à 6 tokens, le nombre repère auquel nous allons comparer l'effectif de chaque taken-cluster possible associé à la couche 1 ; le nombre total de taken-clusters ici impliqué étant de 64 000 (6 400 neurones d'arrivée en couche 1 $\times$ 10 neurones précurseurs en couche 0).

Le tableau n°7 donne à voir une forte sur-représentation (86%) de taken-clusters constitués de moins de 6 taken-tokens, tendance largement significative ($p(\chi^2) < .0001$) ; notons que la taille très importante de l'effectif d'unités statistiques ici impliquées (64 000) est de nature à produire une significativité accrue biaisée, mais que ce biais n'est pas important dans la mesure où la taille d'effet importante ici en jeu ne peut *a priori* qu'être associée à une significativité élevée. Ce résultat est compatible avec notre hypothèse de sélectivité catégorielle associée au processus de détourage.

	< 6	≥ 6
Observed Frequencies	54993	9007
% of Observed Frequencies	.86	.14
Expected Frequencies	3200	60800
Weighted χ^2 Residuals	16.19	-0.85
χ^2	882406.20	
$p(\chi^2)$	.0000	

Tableau n° 7 : Distribution des taken-clusters de taille inférieure à 6 (Couche 1 ; $N = 64\ 000$).

6.4 Détourage et séparation des dimensions initiales d'embedding

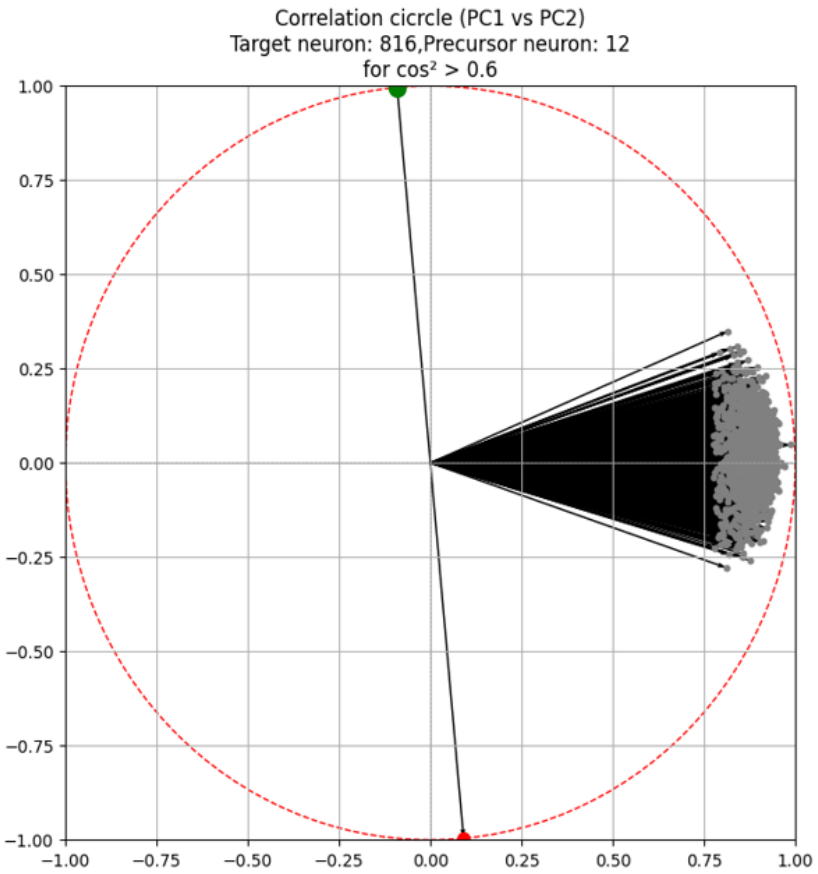
Continuons notre exploration des propriétés du détourage catégoriel généré par les facteurs que sont l'amorçage, l'attention et le phasage catégoriels portés par la fonction d'agrégation neuronale. Le processus de détourage catégoriel se traduit *ipso facto*, au niveau de l'extension (core-tokens) de la catégorie synthétique associée à chaque neurone précurseur de couche n, par la sélection, l'extraction de certains tokens spécifiques (les taken-tokens), qui vont devenir une partie de l'extension (core-tokens) de la nouvelle catégorie synthétique vectorisée par un neurone d'arrivée de couche $n+1$. Dans le cadre de la cognition synthétique, cette abstraction n'est certainement pas le fruit d'un traitement cognitif des tokens en tant que tels (i.e. les tokens comme une unité sémantique, à l'instar de ce qu'un être humain aurait le sentiment de traiter), mais bien de leurs coordonnées dans les dimensions d'embedding de l'espace vectoriel d'entrée de la couche impliquée. Ce qui est donc calculé et extrait de façon *princeps* et à proprement parler de la catégorie d'un neurone précurseur, n'est pas des tokens *per se* mais une sous-dimension catégorielle à l'intérieur de cette dimension catégorielle de départ.

Il est possible d'inventer une variété d'opérationnalisations de l'étude des caractéristiques propres à cette sous-dimension catégorielle extraite (la « forme » catégorielle), par rapport à celles du « reliquat » catégoriel non extrait (le « fond » catégoriel), dans le cadre du processus de détourage catégoriel. Le mode opératoire que nous mobilisons ici est constitué par le cadre analytique de l'espace vectoriel des embeddings d'entrée de GPT2-XL ; cela pouvant notamment être justifié par le fait que nous opérons au niveau de l'interaction génétique entre les couches 0 et 1, qui ne sont dès lors pas encore trop « éloignées » catégoriellement des dimensions catégorielles propres à cet espace vectoriel de départ. Et nous allons dès lors étudier, concernant les 100 core-tokens constitutifs de l'extension de la catégorie d'un neurone précurseur donnée en couche 0, quelles sont les caractéristiques respectives, en termes de dimensions de ces embeddings, de ces tokens « passant » ou « repris » à un neurone d'arrivée (avec un fort poids de connexion) en couche 1. Et plus précisément : quelles sont les caractéristiques respectives, vis-à-vis de cet espace vectoriel de départ, des tokens devenant des taken-tokens (i.e. ceux qui sont constitutifs de la sous-dimension catégorielle extraite de la catégorie de départ, et qui vont devenir à leur tour des core-tokens constitutifs de la catégorie d'arrivée), par opposition aux tokens qui ne sont pas retenus (i.e. les left-tokens, qui font partie du fond catégoriel non extrait).

D'un point de vue statistique, une démarche pertinente ici est d'avoir recours à une analyse factorielle de type analyse en composantes principales (ACP), avec les (1600) dimensions d'embeddings de GPT2-XL prises comme variables et les tokens impliqués comme unités statistiques ; cela, en complétant avec deux variables dichotomiques : une relative au fait qu'un token soit un taken-token, l'autre au fait qu'un token soit un left-token. Nous surpondérons ces deux dernières variables (à hauteur de 1% du nombre de 1600 variables d'embeddings), afin de conduire l'ACP à produire un axe factoriel (doté d'une valeur propre

suffisante) relatif au fait qu'un token soit un taken-token versus un left-token. Et nous allons dès lors étudier comment se distribuent les variables d'embedding de départ par rapport à ce facteur différenciant les taken-tokens des left-tokens ; cela, en étant prudent dans nos interprétations, dans la mesure où nous avons indiqué précédemment en section méthodologique que les conditions usuelles d'application de l'ACP ne sont pas toutes pleinement respectées dans le cadre de notre présente étude (nous ne retiendrons néanmoins dans ce cadre que des vecteurs d'embedding ayant une qualité de représentation $\cos^2 > .6$).

Une première application de notre démarche (neurone précurseur n°12 de la couche 0, dans son interaction génétique avec le neurone d'arrivée n°816 de la couche 1) aboutit au graphe n°5. Ce cercle de corrélation est très instructif. Nous y voyons une différenciation nette entre un premier facteur horizontal (avec 63% de variance restituée) regroupant les vecteurs d'embeddings (dotés d'une bonne qualité de représentation) et un deuxième facteur vertical (associé à 17% de variance extraite) opposant les taken-tokens (en vert) aux left-tokens (en rouge). Les vecteurs d'embedding tendent à se distribuer préférentiellement en fonction du caractère « taken » ou « left » des tokens impliqués, et nous obtenons un graphe d'extraction de dimensions de l'espace vectoriel des embeddings : les taken-tokens sont plus corrélés à certaines dimensions d'embedding alors que les left-tokens sont plus corrélés à d'autres dimensions d'embedding. Précisons qu'il est normal que ces corrélations ne soient pas très fortes, dans la mesure où (i) l'espace vectoriel de dimensions catégorielles fabriqué par le layer 1 commence à s'éloigner catégoriellement de l'espace vectoriel des embeddings de départ (sans quoi le premier n'aurait pas de valeur ajoutée par rapport au second), et (ii) les fonctions non linéaires également impliquées (GELU) déforment nécessairement la topologie vectorielle d'origine.



Graphe n° 5 : Cercle de corrélation ACP sur les variables d'embedding et la variable d'appartenance au groupe (left/taken) associée aux left-tokens et aux taken-tokens (Pondération = 1% ; Neurone cible 816 ; Neurone précurseur 12 ; sur les 78,91% de variables avec $cos^2 > .6$).

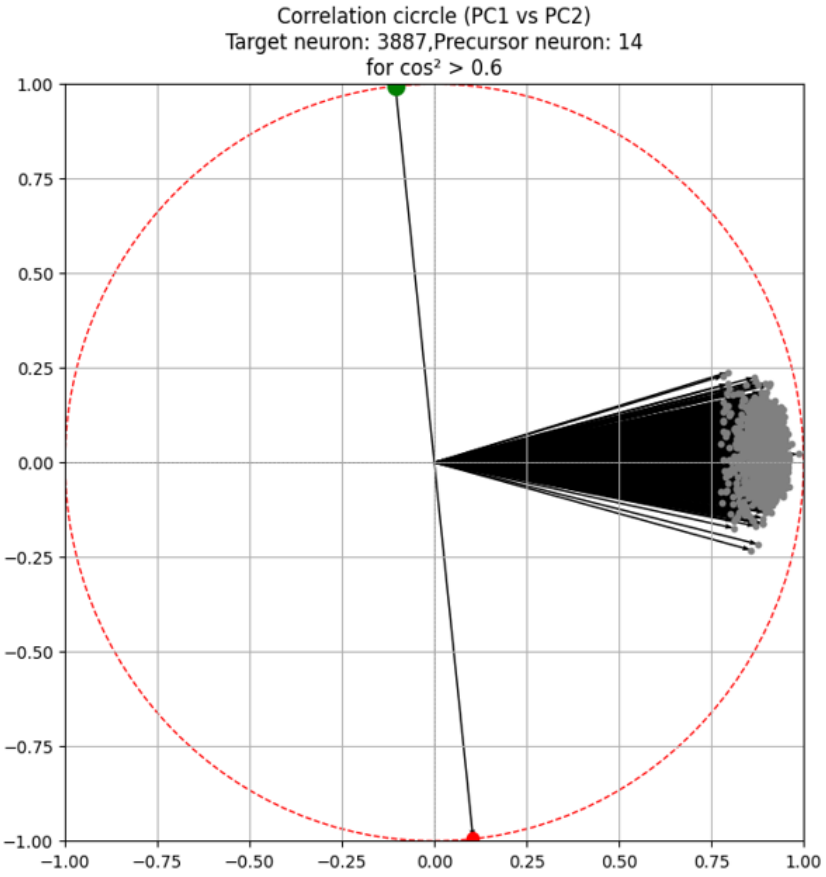
Le tableau n°8, associé au graphe précédent, manifeste une minorité de variables d'embedding associées aux taken-tokens (26%) par opposition aux left-tokens (74%), ce qui est compatible, bien que dans un registre autre, avec notre précédent postulat de sélectivité catégorielle du détourage. Dans la continuité du graphe, les moyennes de produits scalaires font montre d'une projection certes faible (pour les raisons indiquées ci-dessus) mais non négligeable des variables d'embedding sur la variable d'opposition taken vs left-tokens ; précisons, à nouveau de façon compatible avec notre postulat de sélectivité catégorielle, que les produits scalaires moyens sont plus faibles pour les taken-tokens que pour les left-tokens.

	Taken-tokens	**Left-tokens**
% of projected variables	26.13	73.87
Mean(cos)	.0622	.1015
Mean(scalar product)	.0628	.1025

Tableau n° 8 : Statistiques de projection des coordonnées ACP des variables d'embedding sur la variable d'appartenance au groupe (left/taken) (Pondération = 1% ; Neurone cible 816 ; Neurone précurseur 12 ; sur les 78,91% de variables avec $cos^2 > .6$*).*

Purement descriptives et relatives à un seul doublet de neurones, les données que nous obtenons ci-avant tendent à être compatibles avec l'idée suivante : si l'on observe le processus de détourage catégoriel à partir du référentiel constitué par l'espace vectoriel des embeddings de départ de GPT2-XL, ce processus semble se traduire par le fait que l'abstraction, par un neurone d'arrivée, d'une sous-dimension catégorielle (associée aux taken-tokens) à partir de la dimension catégorielle portée par un neurone précurseur donné a pour corollaire l'extraction sélective de certaines dimensions (minoritaires) d'embedding qui sont dès lors catégoriellement spécifiquement liées à cette sous-dimension abstraite.

Un autre exemple (neurone précurseur n°14 de la couche 0, dans son interaction génétique avec le neurone d'arrivée n°3887 de la couche 1) produit le même type de données et d'interprétation (*cf.* graphe n°6 et tableau n°9).



Graphe n° 6 : Cercle de corrélation ACP sur les variables d'embedding et la variable d'appartenance au groupe (left/taken) associée aux left- tokens et aux taken-tokens (Pondération = 1% ; Neurone cible 3887 ; Neurone précurseur 14 ; sur les 86,61% de variables avec $cos^2 > .6$).

	Taken-tokens	Left-tokens
% of projected variables	14.74	85.26
Mean(cos)	.0338	.0888
Mean(scalar product)	.0342	.0896

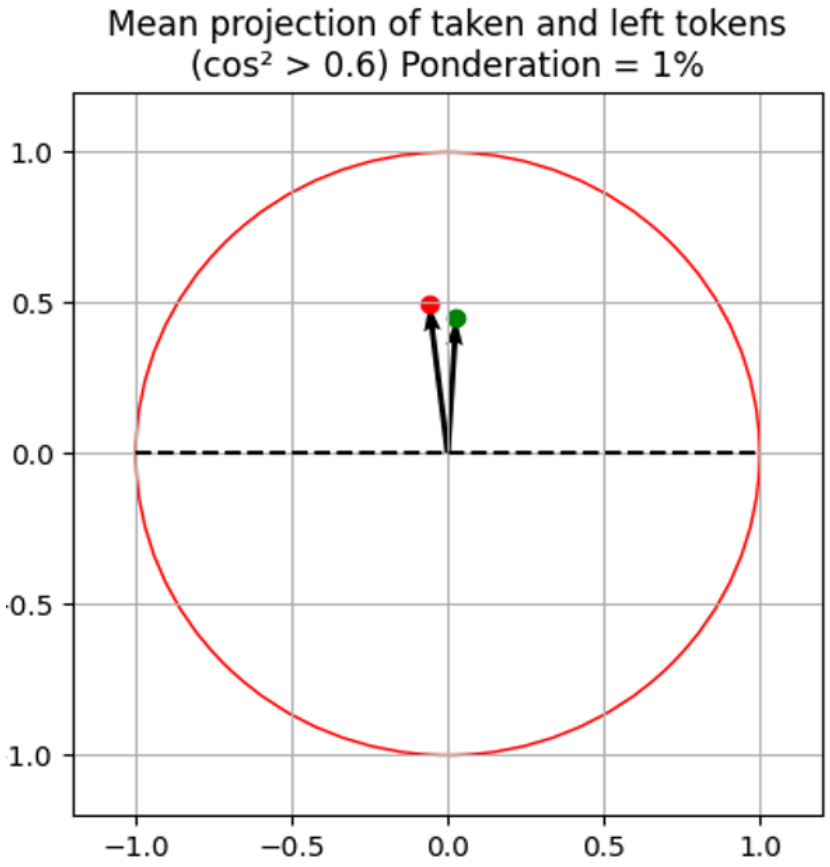
Tableau n° 9 : Statistiques de projection des coordonnées ACP des variables d'embedding sur l'axe d'appartenance au groupe (left/taken) (Pondération = 1% ; Neurone cible 3887 ; Neurone précurseur 14 ; sur les 86,61% de variables avec $cos^2 > .6$).

Examinons maintenant si ces tendances locales, au niveau de certains neurones pris comme exemples, semblent perdurer à un niveau plus global. Pour ce faire,

nous avons appliqué notre démarche d'ACP à l'ensemble, lorsque pertinent, des doublets génétiques (core-tokens de la catégorie d'un neurone de départ de la couche 0, caractère taken ou left de ces tokens vis-à-vis d'un neurone d'arrivée (à fort poids de connexion) associé de la couche 1). Pour des raisons statistiques, tous les croisements ne sont pas possibles et nous n'avons conservé que les cas (i) où 15 à 85% des core-tokens de départ devenaient des taken-tokens ($N = 1671$), puis (ii) où l'indice KMO était supérieur à .5 ($N = 950$). Notons que sur ces 950 cas, 22% seulement des core-tokens impliqués en layer 0 deviennent des taken-tokens (et donc des core-tokens) en layer 1, ce qui est à nouveau pleinement en phase avec notre précédent postulat de sélectivité catégorielle.

Le graphe n°7, après compilation de l'ensemble des données obtenues, rotation (l'axe des abscisses devenant celui opposant taken et left-tokens) et reconstruction d'un vecteur moyen de dimensions d'embedding associées aux takens-tokens (doté cependant d'une norme assez faible de .45) et d'un vecteur moyen de dimensions d'embedding associées aux left-tokens (doté d'une norme de .5), confirme notre tendance *princeps*. Sur la globalité mentionnée, nous voyons à nouveau que la partition taken-tokens vs left-tokens est associée à une segmentation des vecteurs d'embeddings de départ de GPT2-XL.

Autrement dit, le détourage catégoriel opéré par la fonction d'agrégation et consistant à extraire une sous-dimension catégorielle (dont l'extension est le cluster de taken-tokens spécifiquement impliqués pour chaque cas) se traduit par le fait que cette sous-dimension singulièrement extraite est associée à des dimensions particulières d'embedding, abstraites électivement de l'ensemble des dimensions possibles d'embedding de départ. Même si, en phase avec l'ultime étage de «réflexion » de l'abstraction réfléchissant de Piaget, nous gardons en mémoire que les sous-dimensions catégorielles détourées ne se réduisent pas à de simples extractions électives mais, plus avant, bien à d'authentiques recombinaisons originales de ces dernières via la fonction d'agrégation.



Graphe n° 7 : Reconstruction des coordonnées ACP du vecteur moyen d'embedding associé aux taken-tokens et du vecteur moyen d'embedding associé aux left-tokens (Pondération = 1% ; Couche 1 ; tous $cos^2 > .6$).

De façon plus analytique, le tableau n°10 met en lumière à nouveau une minorité de variables d'embeddings associées aux taken-tokens ; cela, toujours en phase avec notre postulat précédent de sélectivité catégorielle, bien que cette minorité moyenne (36%) soit plus forte que dans nos illustrations précédentes de cas. Et nous retrouvons le même type de résultats : des produits scalaires, certes faibles pour les raisons indiquées, mais non négligeables, de projection des vecteurs moyens d'embedding sur l'axe opposant taken et left-tokens.

	Taken-tokens	Left-tokens
Mean (% projected variables)	35.95	64.05
Mean (mean(cos))	.0639	.166
Mean (mean(scalar product))	.0646	.1177

Tableau n° 10 : Statistiques moyennes de projection des coordonnées ACP des variables d'embedding sur l'axe d'appartenance au groupe (Pondération = 1% ; Couche 1 ; $N = 950$; sur les 79,62% de variables avec cos > .6 ; Où KMO > 0.5 ; Moyenne(KMO) = .501).

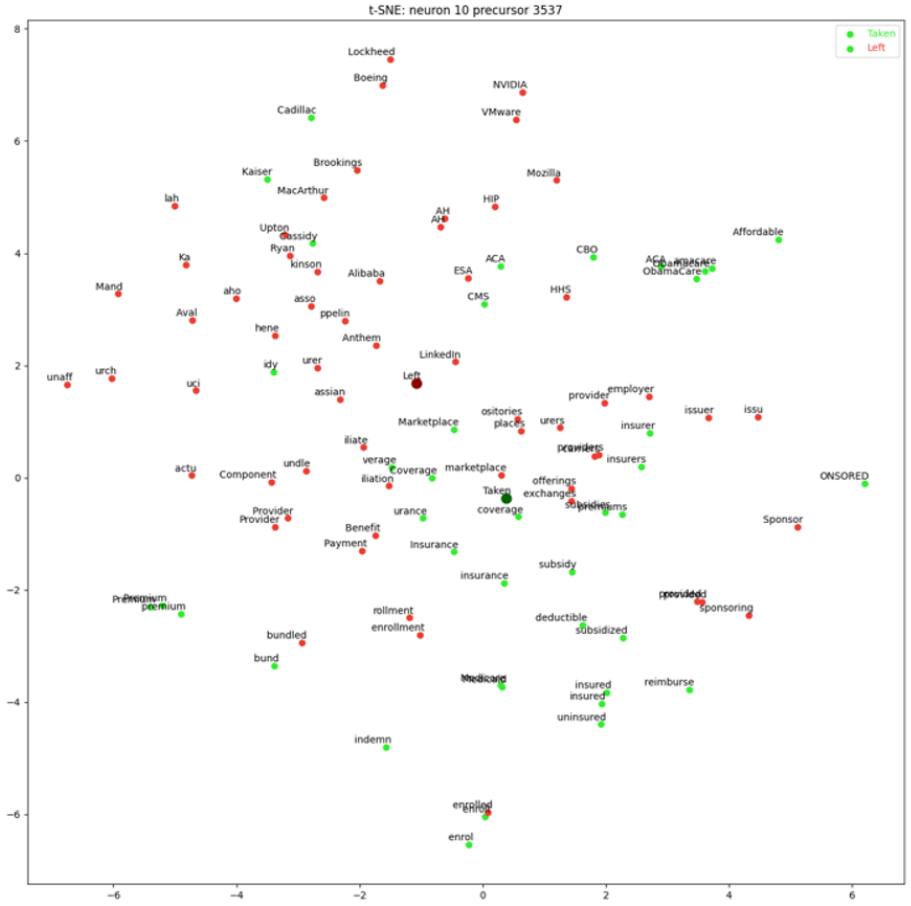
Le détourage opéré génétiquement par les catégories des neurones d'arrivée afin de se constituer consiste à abstraire, des catégories de leurs neurones précurseurs, des sous-dimensions catégorielles. Ces sous-dimensions ne sont pas catégoriellement aléatoires, mais catégoriellement homogènes ou tout du moins convergentes. La projection de ces sous-dimensions singulières dans un référentiel d'observation (l'espace vectoriel des embeddings de départ de GPT2-XL) catégoriellement pas trop éloigné, ce que nous avons réalisé dans la présente section, permet de mettre en lumière cette convergence catégorielle partielle ; cela, en faisant montre d'une segmentation, d'une compartimentation de ces dimensions d'embedding : certaines, et uniquement certaines, de ces dimensions (et donc de leurs sémantiques associées) tendent à confluer et ainsi à constituer des formes catégorielles (les sous-dimensions extraites) séparées d'un fond catégoriel distinct et non retenu (les autres dimensions d'embedding restantes). Mais le fait que les corrélations (dimensions d'embedding / sous-dimensions catégorielles) obtenues soient relativement faibles est un phénomène particulièrement remarquable : ces sous-dimensions ne sont pas à proprement «extraites », selon une abstraction simple au sens de Piaget et de l'empirisme philosophique : leurs combinaisons extraites ne sont pas pré-données et préexistantes dans les dimensions catégorielles respectives de départ. Au contraire, et en phase avec l'ultime étape de «réflexion » de l'abstraction réfléchissante piagétienne synthétique portée par la fonction d'agrégation, le détourage est le fruit d'une recombinaison originale et créatrice des dimensions d'embedding de départ. Le détourage n'est ainsi pas la séparation d'une forme d'un fond catégoriel déjà existants mais l'authentique et singulière construction d'une forme séparée d'un fond tout autant fabriqué.

6.5 Détourage et segmentation de zones catégorielles

Dans la rubrique précédente, nous avons tenté de mettre en lumière le fait que, dans le cadre du détourage catégoriel, les sous-dimensions extraites ou plutôt fabriquées sont chacune catégoriellement convergentes (et non pas aléatoires ou sémantiquement totalement chaotiques) au niveau interne : chaque sous-dimension tend à être associée à des dimensions d'embeddings de départ spécifiques et pas à d'autres. Tentons maintenant de faire un pas de plus dans la compréhension de cette convergence catégorielle. Cela, toujours au sein d'une approche de réduction des dimensions de l'espace vectoriel des embeddings de

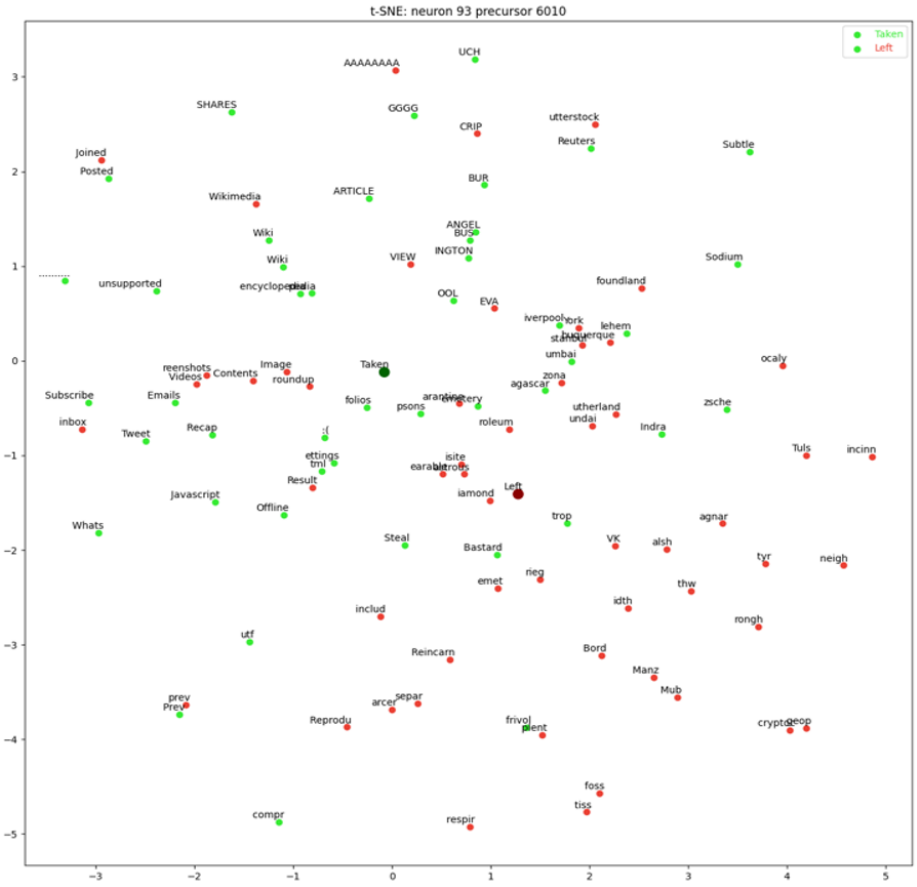
départ de GPT2-XL mais cette fois-ci (i) en s'affranchissant de la contrainte d'une construction linéaire des axes catégoriels réduits (et de prendre en compte ainsi l'effet de la fonction non linéaire d'activation couplée à la fonction d'agrégation), (ii) en se libérant de certaines conditions statistiques partiellement non respectées dans le cadre de nos ACP précédentes et, (iii) en couplant notre approche à une démarche de clusterisation de tokens (respectivement taken et left). Et toujours sur la base des (100) core-tokens des catégories des neurones précurseurs du layer 0, dont 15 à 85% deviennent des taken-tokens dans le cadre de l'interaction génétique avec un neurone d'arrivée en couche 1.

Le graphe n°8, réalisé sur les core-tokens du neurone précurseur n°3537 du layer 0 et leur caractère taken (points verts) ou left (points rouges) dans le cadre de l'interaction génétique avec le neurone d'arrivée n°10 en layer 1, est largement instructif. Intéressons-nous aux barycentres respectifs des taken-tokens (*cf.* point vert foncé noté « Taken ») et des left-tokens (*cf.* point rouge foncé noté « Left »), et prenons comme mesure de distance le fait que ces barycentres soient ou non présents dans les mêmes quadrants définis par le croisement des deux axes dimensionnels réduits obtenus. Nous pouvons observer que ces barycentres, même s'ils ne sont pas extrêmement éloignés l'un de l'autre, ne sont pas présents dans les mêmes quadrants, c'est-à-dire ne sont pas positionnés dans les mêmes zones catégorielles d'embeddings. Ainsi, la sous-dimension catégorielle extraite (ou plutôt fabriquée) ici par le neurone d'arrivée à partir de l'ensemble de la dimension catégorielle du neurone de départ est bien catégoriellement définie : cette sous-dimension créée (en son barycentre « Taken ») relève d'une zone catégorielle distincte de celle du fond catégoriel non extrait (en son barycentre « Left »). Il y a bien ici, à l'instar de nos résultats précédents relatifs à l'ACP, abstraction d'une sous-dimension catégorielle relativement homogène et spécifique vis-à-vis d'un référentiel d'observation opérationnalisé en termes d'embeddings de départ de GPT2-XL (sinon les barycentres seraient positionnés au même endroit et il n'y aurait pas de segmentation catégorielle). Mais, de façon cette fois plus qualitative, rendue possible par l'approche t-SNE préservant relativement bien les structures locales de proximité entre tokens, nous pouvons voir que les taken-tokens impliqués (« reimburse », « bund », « enrolled », « subsidy », « insured », « deductible », « insurance », « coverance », « obamacare », etc.) tendent à renvoyer à un champ lexical bien défini, ici le champ lexical de l'assurance santé, forme catégorielle extraite bien distincte du fond catégoriel constitué par les left-tokens.



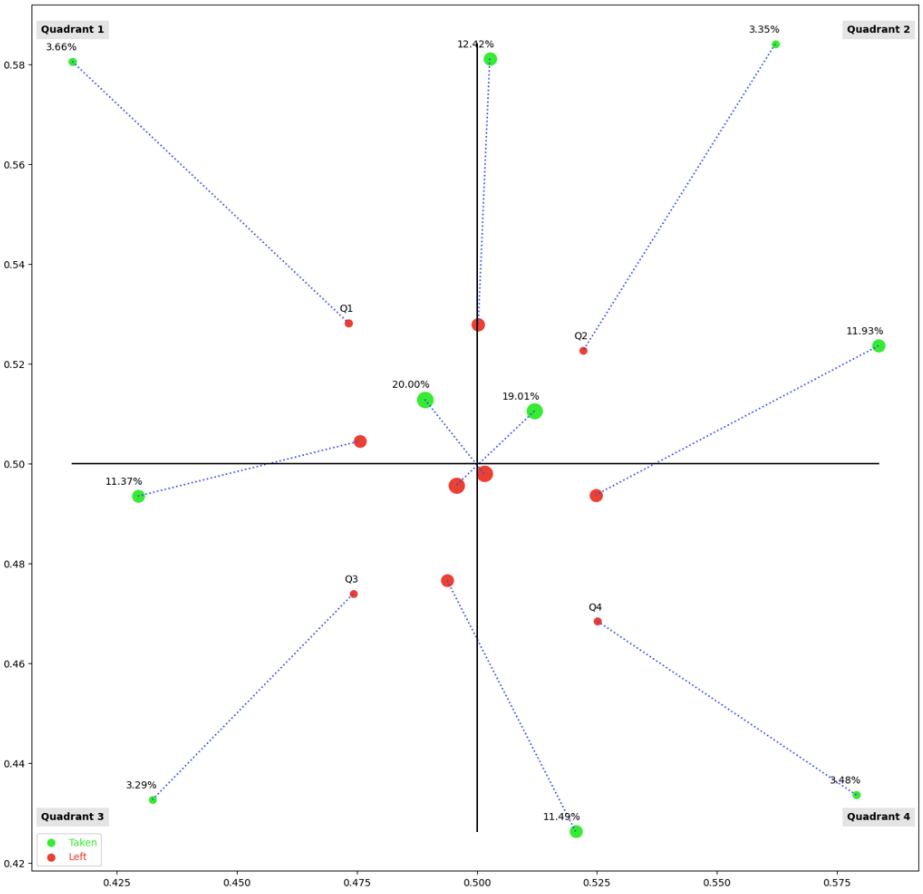
Graphe n° 8 : Distribution des taken-tokens et left-tokens dans l'espace d'embedding réduit par t-SNE (Neurone précurseur 3537 ; Neurone d'arrivée 10).

Il en va de même pour le graphe n°9, ayant trait au doublet neurone précurseur 6010 / neurone d'arrivée 93. Les barycentres sont à nouveau positionnés au sein de zones catégorielles (quadrants) contrastées. Les taken-tokens (« wiki », « shares », « posted », « recap », « offline », « encyclopedia », « article », « emails », « subscribe », « folios », « tweet », « javascript », « whats », etc.) constitutifs de la sous-dimension catégorielle ici fabriquée tendent à converger cette fois-ci autour du champ lexical relatif au domaine des communications numériques et des plateformes en ligne ; champ lexical relativement bien distinct de la sémantique associée aux left-tokens impliqués.



Graphe n° 9 : Distribution des taken-tokens et left-tokens dans l'espace d'embedding réduit par t-SNE (Neurone précurseur 6010 ; Neurone d'arrivée 93).

De façon maintenant plus systématique, nous avons appliqué notre démarche t-SNE aux 1610 doublets (neurone précurseur du layer 0, neurone d'arrivée associé et à fort poids de connexion du layer 1) pour lesquels notre approche t-SNE était possible et pour lesquels le pourcentage de taken-tokens extraits des core-tokens de départ était à nouveau compris entre 15 et 85%. Le graphe n°10 présente une reconstruction synthétique des données obtenues, dans laquelle figurent les barycentres moyens des taken-tokens et des left-tokens selon les différents cas de figure obtenus de croisement entre leurs positions respectives au sein des quatre quadrants (zones catégorielles d'embeddings). La même tendance que celle précédemment identifiée au niveau des exemples présentés se manifeste, et de façon largement significative ($\chi^2 = 217.75$, $p < .0001$, $ddl = 9$) : les barycentres moyens des taken-tokens extraits tendent (à hauteur de 87% ici) à être positionnés au sein de zones catégorielles (quadrants) différentes de celles des barycentres moyens des left-tokens qui leur sont associés.



Graphe n° 10 : Distribution graphique, au sein des quadrants, des barycentres moyens des paires de clusters taken et left dans l'espace d'embedding réduit par les axes moyens du t-SNE (Couches 0/1 ; $N = 1610$).

Dans la lignée de ce que nous venons d'indiquer, le tableau n°11 manifeste une sous-représentation systématique des barycentres taken/left associés positionnés dans les mêmes quadrants, par opposition à une sur-représentation forte (39% des effectifs concernés) des barycentres associés dispersés entre les quadrants Q1 et Q4, ou entre les quadrants Q2 et Q3. Ce dernier point montre de façon intéressante que les sous-dimensions catégorielles extraites tendent à être catégoriellement distinctes vis-à-vis des deux axes factoriels t-SNE conjointement, c'est-à-dire d'autant plus distinctes catégoriellement.

	LEFT-CLUSTERS			
	Quadrant 1	**Quadrant 2**	**Quadrant 3**	**Quadrant 4**
Quadrant 1	-0.44	0.04	-0.20	0.60
Quadrant 2	-0.08	-0.46	0.48	0.01
Quadrant 3	0.05	0.66	-0.44	-0.20
Quadrant 4	0.50	-0.13	0.07	-0.45

Tableau n° 11 : Sous- et sur-représentations de la distribution, au sein des quadrants, des ***barycentres moyens*** *des paires de clusters taken et left dans l'espace d'embedding réduit par les axes moyens du t-SNE (Couches 0/1 ; $N = 1610$).*

Il ressort de cette présente étude t-SNE, une convergence avec les résultats de l'étude précédente ACP : les sous-dimensions catégorielles extraites des catégories des neurones précurseurs (couche 0) par les fonctions d'agrégation des neurones d'arrivée associés et à fort poids de connexion (couche 1) ne sont pas catégoriellement aléatoires mais tendent à être catégoriellement homogènes au niveau interne, ici du point de vue du référentiel d'observation constitué par l'espace vectoriel des embeddings de GPT2-XL. Plus précisément, le détourage catégoriel opéré à partir de la catégorie portée par un neurone précurseur tend à créer une sous-dimension catégorielle dont le centre de gravité est catégoriellement distinct et (relativement) distant du centre de gravité catégoriel du reste de cette catégorie de départ. Autrement dit, le détourage catégoriel façonne, dans des zones catégoriellement différentes, une forme catégorielle (sous-dimension) activement distinguée d'un fond catégoriel. Et une démarche de clusterisation des (taken versus left) tokens impliqués donne à voir le résultat qualitatif sémantique de ce processus de disjonction, de différenciation catégorielle opéré par le phénomène de détourage catégoriel (en tout cas, lorsque la sémantique impliquée correspond à des éléments de la sémantique humaine, ce que l'on peut supposer être plus fortement le cas pour les premières couches).

6.6 Phénoménologies catégorielles du détourage

Terminons notre présente investigation des caractéristiques du détourage catégoriel en réalisant un pas supplémentaire dans son étude qualitative. Cela, en donnant à voir, purement à titre d'exemple, une typologie partielle manifestant une façon possible de catégoriser différents cas catégoriels possibles de détourage (*cf.* tableau n°12).

Ces exemples sont tous issus d'une analyse qualitative des caractéristiques de clusters de taken-tokens (la forme catégorielle) extraits au niveau de la couche 1 par rapport à leurs clusters correspondant de left-tokens (le fond catégoriel). Et ils relèvent tous, dans le cas présent, d'une approche «linguistique humaine ».

Une première classe catégorielle de détourage peut être qualifiée de «sémantique ». Elle peut être subdivisée en deux sous-classes :

- Premièrement, le **détourage hétéro-lexical**, caractérisé par la segmentation de tokens appartenant à un même champ lexical et distinct de celui du fond catégoriel :
  a) par fractionnement de tokens de même racine. Exemple : les tokens *manager* et *managerial* sont différenciés d'un fond catégoriel contenant les tokens *Pharma*, *Middles*, *villa*.
  b) par fragmentation de tokens issus de racines différentes. Exemple : les tokens *manager*, *Wenger* (célèbre entraîneur de l'équipe de foot d'Arsenal) et *logistics* sont dissociés d'un fond catégoriel comprenant les tokens *Nurse*, *Motorist*, *frustrated*.
- Deuxièmement, le **détourage sub-lexical**, déterminé par la scission de tokens appariés à un sous-champ lexical du champ lexical de la catégorie impliquée :
  a) par dispersion de tokens de même racine. Exemple : les tokens *order*, *ordered*, *ordering*, *Ord* sont séparés de *want*, *aspir*, *needing*, *desirable*, *desire*.

     Ou encore : les tokens *Psych*, *psychopath*, *psychiatrie* sont distingués d'une catégorie contenant *McGill*, *Graduation*, *NYU*, *lecturer*, *academ*, *Prof* (champ lexical de l'enseignement supérieur) et *pediatric*, *mindfulness*, *PTSD*, *hypnot*, *clinicians* (champ lexical de la médecine).
  b) par désunion de tokens provenant de racines distinctes. Exemple : les tokens *listen*, *Audio*, *Speakers* (champ lexical audio non technique) sont discriminés des tokens *kHz*, *dB*, *Codec*, *frequencies*, *spectrum* (champ lexical audio technique).

     Ou encore : les tokens *audio*, *listen*, *ear*, *sound*, *headphone* (champ lexical de la perception sonore) sont coupés des tokens *soundtrack*, *Melody*, *orchestra*, *Musical*, *Bach*, *singers* (champ lexical de la production sonore artistique), des tokens *dB*, *MIDI*, *kHz*, *Frequency*, *wav*, *reverber* (champ lexical du son dans un aspect technique), et encore des tokens *Yamaha*, *drums*, *violin*, *bells*, *guitar* (champ lexical des instruments de musique).

Une deuxième classe catégorielle, plus restreinte, de détourage peut être qualifiée de «graphémique ». Elle peut être caractérisée par une démarcation de tokens dotés d'un groupe spécifique de lettres.

Exemple : les tokens *ID*, *id*, *pid*, *IDA* sont isolés de *mom*, *phone*, *CAR*, *but*, *Ratio*.

Semantic clipping	Hetero-lexical	By homo-root tokens	(manager, managerial) vs (Pharma, Middles, villa)
		By hetero-root tokens	(manager, Wenger, logistics) vs (Nurse, Motorist, frustrated)
	Sub-lexical	By homo-root tokens	(order, ordered, ordering, Ord) vs (want, aspir, needing, desirable, desire) (Psych, psychopath, psychiatrie) vs (McGill, Graduation, NYU, lecturer, academ, Prof) + (pediatric, mindfulness, PTSD, hypnot, clinicians)
		By hetero-root tokens	(listen, Audio, Speakers) vs (kHz, dB, Codec, frequencies, spectrum) (audio, listen, ear, sound, headphone) vs (soundtrack, Melody, orchestra, Musical, Bach, singers) + (dB, MIDI, kHz, Frequency, wav, reverber) + (Yamaha, drums, violin, bells, guitar)
Graphemic clipping			(ID, id, pid, IDA) vs (mom, phone, CAR, but, Ratio)

Tableau n°12 : Exemples de types de détourage linguistique catégoriel : différences entre les left-tokens et les taken-tokens issus des core-tokens des mêmes neurones (Couche 1).

Ces présents exemples mentionnés, à titre purement illustratif et sans volonté de systématicité, nous donnent à comprendre qualitativement comment le détourage catégoriel peut créer des extractions de sous-dimensions catégorielles, en regroupant des tokens convergeant au titre d'un segment catégoriel homogène et défini. Même si ces présents exemples, relatifs à des couches précoces et donc encore potentiellement proches de logiques humaines, sont aisément interprétables dans le cadre des catégories de pensée humaines qui sont les nôtres, ils ne doivent pas nous donner à penser qu'il en est forcément ainsi : les sous-dimensions extraites sont régulièrement de l'ordre d'*alien concepts*, non isomorphes à des *primes* humains de pensée ; il y a bien dans leur cas une rationalité de convergence catégorielle, mais *synthétique*, c'est-à-dire propre à des construits statistiques non immédiatement ou aisément explicables dans les catégories de la sémantique humaine.

7 Discussion

7.1 Synthèse de notre démarche d'exploration génétique de la segmentation catégorielle synthétique

Nous nous sommes intéressés, de façon exploratoire, au processus de segmentation catégorielle opéré par la cognition synthétique des modèles de langage, consistant à découper et fabriquer, dans le monde des tokens, de nouvelles dimensions catégorielles ; chaque neurone formel MLP pouvant être associé à une dimension catégorielle spécifique, traçable par son extension propre de tokens afférents s'activant particulièrement pour cette catégorie (ses core-tokens).

D'un point de vue structurel causal, cette segmentation est entre autres pilotée par la fonction d'agrégation inhérente à chaque neurone [105], fonction incarnant trois facteurs présidant à la genèse et à l'activation des catégories portées par les neurones des couches $n+1$ à partir des catégories des neurones précurseurs en couche n :

1. l'*effet x* ou amorçage catégoriel synthétique (l'activation des catégories précurseures se répercute par propagation sur la fabrication des catégories d'arrivée),
2. l'*effet w* ou l'attention catégorielle synthétique (les poids de connexions entre neurones d'arrivée et précurseurs pilotent le degré de pertinence accordé aux catégories précurseures dans la construction des nouvelles catégories d'arrivée),
3. l'*effet S* ou phasage catégoriel synthétique (des sous-groupes de core-tokens identiques de différents précurseurs simultanément activés rentrent en écho catégoriel dans la genèse des nouvelles catégories d'arrivée).

D'un point de vue fonctionnel, ces trois facteurs mathématico-cognitifs de la segmentation catégorielle président, au niveau d'un neurone d'arrivée (couche n), à un mécanisme d'extraction d'une sous-dimension catégorielle spécifique de la catégorie portée par chacun de ses neurones précurseurs (couche $n-1$) ; l'union de ces sous-dimensions générant la nouvelle dimension catégorielle vectorisée par ce neurone d'arrivée. Cette abstraction se traduit par un processus de *détourage catégoriel synthétique* consistant à fabriquer et à distinguer une forme d'un fond catégoriel.

Il s'agit 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.

Plusieurs caractéristiques fonctionnelles de ce détourage catégoriel ont été ici proposées de façon exploratoire :

- **Réduction catégorielle** : la sous-dimension catégorielle extraite est associée à un ensemble de (taken-)tokens catégoriellement plus homogènes entre eux, par rapport à l'ensemble des (core-)tokens de départ de la catégorie précurseure impliquée.
- **Sélectivité catégorielle** : extraction, à partir de la catégorie d'un neurone précurseur, d'une extension significativement quantitativement plus restreinte en termes de nombre de (taken-)tokens.
- **Séparation des dimensions initiales d'embedding** : le détourage catégoriel, lorsqu'observé dans le référentiel de l'espace vectoriel des embeddings de départ, tend à se manifester par une compartimentation élective dichotomique 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 la catégorie de départ.
- **Segmentation de zones catégorielles** : 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 contrastées, attestables, lorsque possible, par des

interprétations sémantiques humaines différenciées des clusters de tokens impliqués.

Ces différentes propriétés nous donnent à comprendre qualitativement différentes caractéristiques cognitives synthétiques par lesquelles le processus de 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.

7.2 Quel statut épistémologique accorder aux sous-dimensions extraites par le détourage catégoriel ?

Ainsi que nous avons pu déjà l'aborder dans des travaux précédents [102, 106], il nous appartient de ne pas tomber dans le piège épistémologique de l'anthropomorphisme consistant à ne chercher à analyser la cognition synthétique qu'à travers le seul filtre de nos concepts cognitifs et catégoriels humains propres. Et comprendre que les sous-dimensions catégorielles extraites par le détourage catégoriel ne sont pas nécessairement phasées avec des catégories humaines habituelles de pensée [43, 12, 96, 95, 16].

De même, il nous appartient de sortir d'une forme de naïveté épistémologique réaliste et empiriste consistant à «naturaliser » les sous-dimensions instanciées par le processus de détourage catégoriel, en croyant qu'elles seraient des représentations correspondant à des sous-catégories déjà présentes dans le monde matériel et corollaires d'une forme de réalité intrinsèque, *per se*, pré-donnée et ontologique.

Ainsi que nous le dit magnifiquement von Glaserfeld [49] :

> « In order to judge the goodness of a representation that is supposed to depict something else, one would have to compare it to what it is supposed to represent. In the case of 'knowledge' that would be impossible, because we have no access to the 'real' world except through experience and yet another act of knowing, and this, by definition, would simply yield another representation (...). It is logically impossible, however, to compare a representation with something it is supposed to depict, if that something is supposed to exist in a real world that lies beyond our experimental interface. »[49, p.93].

Et alors que l'auteur constructiviste indique que l'idée même de « 'representation' (...) implies a reproduction, copy, or other structure that is in some way isomorphic with an original » [49, p.94], il nous invite dès lors à préférer la formulation «re-presentation » plus conforme à l'idée qu'elle dénote « a re-play of my own experiences, not a piece of some independent, objective world » [49, p.95].

7.3 Quel statut cognitif accorder aux sous-dimensions extraites par le détourage catégoriel ?

La théorie de la conceptualisation développée par Vergnaud [135], dans le champ du développement de la cognition humaine, nous semble particulièrement fructueuse pour nous doter d'un cadre de pensée des phénomènes synthétiques de construction de segmentation catégorielle et de détourage catégoriel.

Au sein de la théorie de Vergnaud, la conceptualisation est une activité représentationnelle dont la finalité est la construction cognitive de caractéristiques opératoires ; ceci, afin de fonder la conduite sur ces caractéristiques et dès lors de la rendre efficace [136]. Autrement dit, la fonction de la conceptualisation est d'établir des homomorphismes entre le plan des objets du monde sur lesquels il convient d'agir et le plan des opérations et des contenus de la pensée. Ainsi, pour Vergnaud [137], le processus de conceptualisation est une activité cognitive économique fondamentalement pragmatique : « on conceptualise pour agir efficacement ». Plus encore, dans une large mesure, pour l'auteur, la conceptualisation n'est pas cognitivement positionnée dans le registre de la théorisation explicite, conscientisée et verbalisée, mais dans celui de l'action. Cette finalisation proprement opérationnelle est à l'origine de la caractéristique de la conceptualisation d'être «en acte », c'est-à-dire d'être encapsulée dans l'action. Ainsi, les concepts et théorèmes mobilisés par l'individu sont nommés «concepts-en-acte » et «théorèmes-en-acte » en tant qu'ils ne sont activés que dans la seule action et n'ont pour unique finalité que de rendre efficace cette action.

Vergnaud [133] définit un concept-en-acte comme une catégorie de pensée tenue pour pertinente par l'individu relativement à une classe de situations d'action. Les concepts-en-acte sont des catégories de pensée à travers lesquelles le sujet crée et intègre des informations liées au type de situations auquel il est confronté. Autrement dit, les concepts-en-acte sont des filtres cognitifs grâce auxquels une situation donnée est « lue » ou construite sur le plan cognitif. D'un point de vue épistémologique, il existe potentiellement une infinité de types formels de catégories de pensée ; les types les plus fréquemment rencontrés sont les suivants : objet, propriété, relation, transformation, condition, processus. À nouveau, les concepts-en-acte sont des vecteurs pragmatiques de la pensée qui organisent le traitement de l'information en découpant les objets du monde en fonction des seuls buts contingents de l'activité finalisée [99]. En effet, la fonctionnalité des concepts-en-acte réside dans le fait qu'ils permettent au sujet de focaliser son attention sur un nombre limité d'éléments sélectionnés et expérimentés comme importants pour la réussite de l'action. À ce titre, ils sous-tendent une représentation des seules variables de situation dont la prise en compte est estimée centrale pour l'effectivité de l'action.

Sur la base de ces éléments de définition, nous pouvons indiquer que la catégorie portée par un neurone formel présente, d'un point de vue épistémologique, les traits définitoires d'un concept-en-acte. En effet, une catégorie synthétique n'est pas une entité conscientisée, démontrée, justifiée ou explicitée par la cognition synthétique. De plus, une catégorie artificielle a

fondamentalement une raison d'être pragmatique : sa genèse et son existence sont finalisées par l'efficience du traitement de l'information et de l'activité cognitive qu'elle rend possible : réaliser les tâches pour lesquelles le réseau de neurones a été conçu et entraîné, encoder le contexte (positionnel et « sémantique ») des tokens. Nous pouvons dès lors à ce titre qualifier les catégories associées aux neurones formels de concepts-en-acte synthétiques.

Vergnaud [134] définit un théorème-en-acte comme une proposition de pensée tenue pour vraie par le sujet relativement à une classe de situations d'action. Les théorèmes-en-acte fondent l'efficacité de l'action en la faisant reposer sur les « théories pratiques » implicites qu'ils constituent et que le sujet élabore à partir de l'emboîtement des propriétés des objets qui lui font face. Autrement dit, les théorèmes-en-acte permettent à l'individu de construire une représentation pratique d'un mode de fonctionnement et d'organisation des caractéristiques fonctionnelles des situations d'actions sur lesquelles il tend à agir efficacement [100]. Plus précisément, ces théorèmes-en-acte sous-tendent une représentation de la manière dont il convient de combiner ces variables situationnelles critiques.

D'un point de vue formel, une proposition résulte de la composition de prédicats et d'arguments ; un théorème-en-acte est ainsi une imbrication de concepts-en-acte. Ainsi que nous l'avons indiqué précédemment, l'activité de segmentation catégorielle opérée par la cognition synthétique est régulée de façon centrale (entre autres) par la fonction d'agrégation définitoire de chaque neurone formel.

De par sa forme, de type $S(w_{i,j}x_{i,j}) + b$, cette fonction d'agrégation fabrique la catégorie associée à un neurone de couche n, en combinant de façon spécifique (pondérée) les catégories portées par ces neurones précurseurs en couche $n-1$.

À ce titre, en phase avec les définitions précédentes, une fonction d'agrégation peut être assimilée à un **théorème-en-acte synthétique** présidant à la coordination singulière (propre à chaque neurone) d'une série de concepts-en-acte synthétiques antécédents.

Un tel théorème-en-acte artificiel réalisant de facto une activité de réflexion, au sens de l'abstraction réfléchissante de Piaget, dans la mesure où il opère une authentique reconstruction et réorganisation de catégories précurseures (en couche $n-1$), projetées sur le plan supérieur de la nouvelle catégorie neuronale générée en couche n.

Réalisons un dernier pas supplémentaire. Les sous-dimensions catégorielles extraites des catégories des neurones précurseurs sont elles-mêmes des concepts-en-acte synthétiques ; en effet, d'un point de vue épistémologique, une sous-dimension est une catégorie, ou plus précisément une sous-catégorie fabriquée au sein de sa catégorie d'origine ; à ce titre, une sous-dimension catégorielle peut être qualifiée de **sous-concept-en-acte**.

Mais il est intéressant de comprendre dans quel contexte mathématico-cognitif singulier ces sous-concepts-en-acte (i.e. ces sous-dimensions catégorielles) synthétiques émergent : dans le contexte «paroxystique » spécifique où le théorème-en-acte porté par la fonction d'agrégation est associé à une valeur d'activation résultante forte. C'est ainsi que l'on peut comprendre comment, au sein d'une forme de transition de phase catégorielle, le théorème-en-acte

définitoire de la catégorie d'un neurone d'arrivée détoure des sous-concepts-en-acte (i.e. des sous-dimensions) des catégories précurseures et formate cette nouvelle forme catégorielle ainsi construite et extraite de son fond catégoriel d'origine.

8 Conclusion

Nous avons investigué de façon exploratoire, dans le domaine de la neuropsychologie de l'intelligence artificielle, les modalités de segmentation catégorielle réalisées par un modèle de langage, celui de GPT2-XL en l'espèce. Ce processus implique, à travers différentes couches neuronales, la création de nouvelles dimensions catégorielles pour analyser les données textuelles et accomplir les tâches requises du modèle. Dans un réseau de perceptrons multicouches (MLP), chaque neurone est associé à une catégorie spécifique, déterminée par trois facteurs issus de la fonction d'agrégation neuronale : l'amorçage, l'attention et le phasage catégoriels. À chaque nouvelle couche, ces facteurs pilotent l'émergence de nouvelles catégories dérivées des catégories des neurones précédents. Par un processus de détourage catégoriel, ces nouvelles catégories sont formées par une abstraction sélective de sous-dimensions spécifiques de leurs catégories antécédentes, en distinguant une forme d'un fond catégoriel. Plusieurs caractéristiques cognitives synthétiques de ce détourage ont été ici distinguées : la réduction catégorielle, la sélectivité catégorielle, la séparation des dimensions initiales d'embedding et la segmentation des zones catégorielles.

Ces propriétés du détourage catégoriel ont été interprétées comme la manifestation de théorèmes-en-acte synthétiques, associés aux fonctions d'agrégation neuronales, qui, dans une phase paroxystique correspondant aux maxima de ces fonctions, génèrent une abstraction réfléchissante de sous-concepts-en-acte synthétiques singuliers, dont la recombinaison formate la fabrication de nouvelles catégories synthétiques encore plus fonctionnelles vis-à-vis des finalités d'activités qui sont celles du réseau neuronal impliqué.

Dans le cadre d'une nouvelle étude, en cours de réalisation, nous poursuivons notre exploration de la segmentation catégorielle synthétique en tentant de mieux comprendre comment se manifeste la restructuration catégorielle opérée d'une couche catégorielle neuronale n à sa couche catégorielle neuronale $n+1$. Cela, en investiguant la phénoménologie cognitive synthétique à travers laquelle les sous-concepts-en-acte (i.e. les sous-dimensions catégorielles), détourés de différents neurones précurseurs, se trouvent être ou non sémantiquement et activationnellement convergents entre eux-mêmes et ainsi générer, dans leurs neurones d'arrivée associés, de nouvelles structures catégorielles originales et singulières de la cognition synthétique.

Acknowledgments

Les auteurs remercient Madeleine Pichat pour sa relecture attentive de cet article, l'équipe de recherche ER IPC (Facultés Libres de Philosophie et de Psychologie de Paris) dans le cadre de laquelle la sous-équipe «Neocognition » a réalisé cette présente étude, Andrew Ponomarev (Petersburg Federal Research Center of the Russian Academy of Sciences) pour les idées stimulantes partagées avec lui, Judicael Poumay (Neocognition) pour ses éclairages pertinents, Martin Courbet et Théo DaSilva (Neocognition & Facultés Libres de Philosophie et de Psychologie) pour leur précieuse participation à la mise en forme de cet article, et Jeanne-Théoline Reybier (Chrysippe) pour ses activités propres rendant possible les conditions de réalisation de cette présente étude.

Bibliography

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

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

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

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

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

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

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

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

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

[10] L. W. Barsalou. Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 364(1521), 1281–1289, 2009.

[11] 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.

[12] 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`

[13] Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and aggregating feature-based model explanations. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2020.

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

[15] 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

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

[17] Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46, 904–911.

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

[19] Campbell, R. L., & Piaget, J. (2014). *Studies in Reflecting Abstraction.* Psychology Press.

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

[21] S. Carey. Precis of *The Origin of Concepts. Behavioral and Brain Sciences*, 34(3) :113, 2011.

[22] 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.

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

[24] Clark, S., Lerchner, A., von Glehn, T., Tieleman, O., Tanburn, R., Dashevskiy, M., & Bosnjak, M. (2021). Formalising Concepts as Grounded Abstractions. *arXiv preprint* arXiv :2101.05125.

[25] 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`.

[26] Collins, A. M., & Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological Review*, 82(6), 407–428.

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

[28] Cuccio, V., & Gallese, V. (2018). A Peircean account of concepts : grounding abstraction in phylogeny through a comparative neuroscientific perspective. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 373(1752), 20170128. https ://doi.org/10.1098/rstb.2017.0128

[29] 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`

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

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

[32] 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`

[33] 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.

[34] Stanislas Dehaene, Fosca Al Roumi, Yair Lakretz, Samuel Planton, and Mathias Sablé-Meyer. Symbols and mental programs : a hypothesis about human singularity. *Trends in Cognitive Sciences*, 26(9) :751–766, 2022. ISSN 1364-6613. doi : `https://doi.org/10.1016/j.`

tics.2022.06.010. URL https://www.sciencedirect.com/science/article/pii/S1364661322001413.

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

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

[37] 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

[38] 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.

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

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

[41] Russell A. Epstein, Eva Zita Patai, Joshua B. Julian, and Hugo J. Spiers. The cognitive map in humans : spatial navigation and beyond. *Nature Neuroscience*, 20(11) :1504–1513, 2017.

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

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

[44] 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)*, 2023.

[45] 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.

[46] 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.

[47] 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.

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

[49] Von Glaserfeld, E. (2002). *Radical Constructivism : A Way of Knowing and Learning*. London : RoutledgeFalmer.

[50] 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.

[51] Green, Isobel, Ryunosuke Amo, and Mitsuko Watabe-Uchida. Shifting attention to orient or avoid : a unifying account of the tail of the striatum and its dopaminergic inputs. *Current Opinion in Behavioral Sciences*, 59, 101441, 2024.

[52] N. Goodman, J. B. Tenenbaum, and T. Gerstenberg. Concepts in a probabilistic language of thought. In E. Margolis and S. Laurence, editors, *The Conceptual Mind : New Directions in the Study of Concepts*, pages 623–654. The MIT Press, 2015.

[53] Amirata Ghorbani, James Wexler, James Y. Zou, and Been Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, pages 9273–9282, 2019.

[54] Harrison, S., Gualdoni, E., & Boleda, G. (2023). Run like a girl! Sports-related gender bias in language and vision. *arXiv preprint* arXiv :2305.14468.

[55] 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.

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

[57] Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M., & Lerchner, A. (2017). β-VAE : Learning basic visual concepts with a constrained variational framework. In *Proceedings of ICLR 2017*.

[58] Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., & Lerchner, A. (2018). Towards a definition of disentangled representations. *arXiv preprint* arXiv :1812.02230.

[59] D. Hofstadter and E. Sander. *Surfaces and Essences*. Basic Books, 2013.

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

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

[62] Howell, D. C. (2008). *Fundamental Statistics for the Behavioral Sciences* (6th ed.). Wadsworth Publishing. DOI : 10.1111/j.1467-985X.2008.00508_-14.x.

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

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

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

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

[67] George Konidaris. On the necessity of abstraction. *Current Opinion in Behavioral Sciences*, 29 :1–7, October 2019. ISSN 2352-1546. doi : 10.1016/j.cobeha.2018.11.005. URL `https://www.sciencedirect.com/science/article/pii/S2352154618302080`.

[68] G. Lakoff. The contemporary theory of metaphor. In *Metaphor and Thought*, pages 202–251. Cambridge University Press, 2008.

[69] G. Fauconnier. *Mappings in Thought and Language*. Cambridge University Press, 1997.

[70] M. Johnson. *Embodied Mind, Meaning, and Reason : How Our Bodies Give Rise to Understanding*. University of Chicago Press, Chicago, IL, 2017.

[71] Kastner, M. A., Ide, I., Nack, F., Kawanishi, Y., Hirayama, T., Deguchi, D., & Murase, H. (2020). Estimating the imageability of words by mining visual characteristics from crawled image data. *Multimedia Tools and Applications*, 79(25), 18167–18199.

[72] Nathaniel J. Killian and Elizabeth A. Buffalo. Grid cells map the visual world. *Nature Neuroscience*, 21(2), 2018.

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

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

[75] Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster Sensorimotor Norms : Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, 52, 1271–1291.

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

[77] Magee, J. C., & Grienberger, C. (2020). Synaptic plasticity forms and functions. *Annual Review of Neuroscience*, 43(1), 95–117.

[78] 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.

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

[80] Margolis, E., & Laurence, S. (2019). Concepts. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2019 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/sum2019/entries/concepts/`.

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

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

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

[84] 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.

[85] Montangero, J., & Maurice-Naville, D. (1994). *Piaget ou l'intelligence en marche : aperçu chronologique et vocabulaire*. Editions Mardaga.

[86] Edvard I. Moser, May-Britt Moser, and Bruce L. McNaughton. Spatial representation in the hippocampal formation : a history. *Nature Neuroscience*, 20(11) :1448–1464, 2017.

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

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

[89] Nisa, et al. (2020). Advances in Social Science, Education and Humanities Research, volume 574.

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

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

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

[93] J. O'Keefe and J. Dostrovsky. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, 34(1) :171–175, 1971.

[94] John O'Keefe and Lynn Nadel. *The Hippocampus as a Cognitive Map*. Oxford University Press, 1978.

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

[96] Olah, C. (2023). Distributed Representations : Composition & Superposition. Retrieved from `https://transformer-circuits.pub/2023/superposition-composition/index.html`

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

[98] Piaget, J. (1974). *Adaptation vitale et psychologie de l'intelligence*. Paris : Hermann.

[99] Pichat, M. (2002). *Composantes intra-individuelle et contractuelle de la conceptualisation mathématique en situation didactique de traitement de tâches algébriques* (thèse de doctorat). Saint-Denis : Université Paris 8.

[100] Pichat, M., & Merri, M. (2007). *Psychologie de l'éducation*. Paris : Bréal.

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

[102] Pichat, M. (2024a). Psychologie de l'IA et alignement cognitif. Actes du colloque *Intelligence artificielle collaborative, management et développement des organisations* du 24/05/2024 coorganisé par l'Uni-

versité Paris Dauphine-PSL et le Cabinet Chrysippe R&D. Available online : https://www.youtube.com/watch?v=9TMmgbELaxQ&list=PLD25p-Bh6_sz6Sr7ms643GpCWW2LlIqeQ&index=6

[103] 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

[104] 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.

[105] Pichat, M., Campoli, E., Pogrund, W., Wilson, J., Veillet-Guillem, M., Melkozerov, A., Gasparian, A., Pichat, P., Poumay, J. (2024). *Neuropsychology and Explainability of AI : A Distributional Approach to the Relationship Between Activation Similarity of Neural Categories in Synthetic Cognition*. arXiv preprint arXiv :2411.07243.

[106] 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

[107] A. Ponomarev and A. Agafonov, "Ontology Concept Extraction Algorithm for Deep Neural Networks," in *Proceedings of the 32nd Conference of Open Innovations Association (FRUCT)*, IEEE, 2022, pp. 221–226. doi : https://doi.org/10.23919/FRUCT56874.2022.9953838.

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

[109] 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

[110] Pulvermuller, F. (2018). Neurobiological Mechanisms for Semantic Feature Extraction and Conceptual Flexibility. *Topics in Cognitive Science*, 10, 590–620.

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

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

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

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

[115] E. Rosch. Principles of categorization. In E. Margolis and S. Laurence, editors, *Concepts : Core Readings*, pages 189–206. MIT Press, 1999.

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

[117] Savioz, A., Leuba, G., Vallet Philippe, G., & Walzer, C. (2010). *Introduction aux réseaux neuronaux : de la synapse à la psyché*. De Boeck.

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

[119] Servan-Schreiber, D., Printz, H., & Cohen, J. D. (1990). A network model of catecholamine effects : gain, signal-to-noise ratio, and behavior. *Science*, 249(4971), 892–895.

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

[121] E. S. Spelke and K. D. Kinzler. Core knowledge. *Developmental Science*, 10(1) :89–96, 2007.

[122] Stoewer, P., Schilling, A., Maier, A., & Krauss, P. (2022). Neural network based formation of cognitive maps of semantic spaces and the emergence of abstract concepts. *arXiv preprint* arXiv :2210.16062.

[123] Singh, V., Gupta, I., & Jana, P. K. (2020). An energy efficient algorithm for workflow scheduling in IaaS cloud. *Journal of Grid Computing*, 18(3), 357–376. `https://doi.org/10.1007/s10723-019-09490-2`.

[124] M. de Sousa Ribeiro and J. Leite, "Aligning Artificial Neural Networks and Ontologies towards Explainable AI," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 6, pp. 4932–4940, 2021.

[125] Tang, Y., Bi, J., Xu, S., Song, L., Liang, S., Wang, T., ... & Xu, C. (2023). Video understanding with large language models : A survey. *arXiv preprint arXiv :2312.17432*.

[126] Tater, T., Walde, S. S. I., & Frassinelli, D. (2024). Unveiling the mystery of visual attributes of concrete and abstract concepts : Variability, nearest neighbors, and challenging categories. *arXiv preprint* arXiv :2410.11657.

[127] 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

[128] 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

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

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

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

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

[133] 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.

[134] 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.

[135] Vergnaud, G. (2020a). Héritages (pp. 27–37). In M. Merri (Ed.), *Activité humaines et conceptualization*. Toulouse : Presses universitaires du Midi.

[136] Vergnaud, G. (2020b). Réponses de Gérard Vergnaud (pp. 341–357). In M. Merri (Ed.), *Activité humaines et conceptualization*. Toulouse : Presses universitaires du Midi.

[137] Vergnaud, G. (2020c). 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.

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

[139] 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.

[140] 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.

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

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

[143] 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

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

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

[146] Xie, Y., Goyal, A., Zheng, W., Kan, M. Y., Lillicrap, T. P., Kawaguchi, K., & Shieh, M. (2024). Monte Carlo Tree Search Boosts Reasoning via Iterative Preference Learning. *arXiv preprint* arXiv :2405.00451.

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

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

[149] Zeki, S. (2002). *Inner Vision : An Exploration of Art and the Brain*. Oxford : Oxford University Press.

[150] 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.

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

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

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