Neuropsychologie et Explicabilité de l'IA: Approche Distributionnelle de la Relation Entre Activation & Similarité des Catégories Neuronales de la Cognition Synthétique

Michael Pichat[1,2], **Enola Campoli**[1,3], **William Pogrund**[1,4], **Jourdan Wilson**[1,5], **Michael Veillet-Guillem**[1,6], **Anton Melkozerov**[1,7], **Paloma Pichat**[1,8], **Armanush Gasparian**[1], **Samuel Demarchi**[1,9], and **Judicael Poumay**[1]

[1]Neocognition (Chrysippe R&D) contact@neocognition.ai
[2]Université de Paris & Facultés Libres de Philosophie et de Psychologie de Paris
[3]Département de Sciences Cognitives & Département de Neuropsychologie, Université Côte d'Azur
[4]Département de Sciences Cognitives, Université de Grenoble Alpes
[5]Département de Linguistique, Université Paris Cité & Université de Californie Los Angeles
[6]Epitech Paris
[7]Académie des Sciences de Russie, FRC CSC RAS
[8]Faculté de Médecine de Lyon Est, Université Lyon 1
[9]Département de Psychologie, Université Paris 8

Publié sur arXiv le 23 Octobre 2024

Nous proposons une approche neuropsychologique de l'explicabilité des réseaux neuronaux artificiels consistant à utiliser les concepts de la psychologie cognitive humaine comme repères heuristiques pertinents pour inventer des cadrages explicatifs synthétiques en phase avec les modalités humaines de pensée. Les concepts analogiques ici mobilisés, de nature à opérer un tel pont épistémologique, sont ceux de la catégorisation et de la similarité, ces notions étant particulièrement ajustées à la « nature » catégorielle du traitement reconstructif de l'information opéré par les réseaux de neurones artificiels. Notre étude tend à mettre à jour un processus de cognition synthétique singulier, celui de convergence catégorielle

des tokens à fortes activations. Processus que nous tentons d'expliquer par l'idée que le segment catégoriel créé par un neurone est en fait le fruit d'une superposition de sous-dimensions catégorielles au sein de l'espace vectoriel d'entrée qui est le sien.

1 Introduction

Au sein d'une démarche d'explicabilité, la neuropsychologie de l'intelligence artificielle se concentre sur l'étude des mécanismes cognitifs neuronaux synthétiques, envisageant ces derniers comme de nouveaux sujets d'étude de la psychologie cognitive. L'objectif est de rendre compréhensibles les réseaux de neurones artificiels utilisés dans les modèles de langage en adaptant les concepts de la psychologie de la cognition humaine à l'interprétation de la cognition neuronale artificielle. Dans ce contexte, la notion de catégorisation est particulièrement pertinente pour ce faire, dans la mesure où elle joue un rôle clé en tant que processus de segmentation et de reconstruction des données informationnelles par les vecteurs neuronaux de la cognition synthétique.

Il s'agira dès lors, dans le cadre de cette étude, de mobiliser la notion de catégorisation, telle qu'appréhendée par la psychologie cognitive humaine (notamment dans sa relation à la notion de similarité), afin de l'appliquer à l'analyse comportementale neuronale et à l'inférence de certains processus cognitifs synthétiques sous-jacents aux comportements observés.

2 Explicabilité catégorielle des réseaux de neurones artificiels

2.1 Épistémologie et utilité de l'explicabilité synthétique

L'explicabilité consiste à décrire l'activité d'un réseau de neurones artificiels dans des termes compréhensibles par l'être humain (Du et al., 2019 ; Pichat, 2023, 2024a, 2024b). Ou en tout cas de projeter le comportement observable d'un réseau de neurones dans un référentiel interprétatif permettant l'attribution d'un registre de signification à ce comportement qui soit pertinent pour un observateur, en fonction des finalités qui sont les siennes. En ce qui nous concerne, ce registre est celui de la psychologie cognitive, qui consistera dès lors à mobiliser les catégories de pensée de la cognition humaine (et, plus particulièrement pour nous, la notion de catégorisation) comme référents conceptuels afin de tisser des analogies heuristiques de conduites cognitives entre cognitions humaine et artificielle. En nous gardant de tomber dans les pièges épistémologiques de l'anthropomorphisme (Nadeau, 1999), du néo-comportementalisme (Bloch et al., 2011) ou de la confusion entre observateur et système observé contre laquelle nous met en garde la cybernétique, la systémie et les sciences cognitives de l'énaction (Watzlawick, 1977, 1984 ; Varela, 1984, 1996).

La fonction pragmatique de l'explicabilité est double. Premièrement, inhiber les réponses potentiellement fallacieuses voire dangereuses du système neuronal synthétique (Luo et al., 2024) : erreurs, biais cognitifs (Echterhoff, 2024) ou culturels (Kheya, 2024), hallucinations (Kandpal et al., 2023 ; McKenna et al., 2023), focalisation abusive sur certains éléments d'entrée (Du et al., 2023), etc. Deuxièmement, augmenter la performance des modèles de langage (Bastings et al., 2022) en améliorant leur cohérence avec un raisonnement humain (Ma et al., 2023). En un mot, il s'agit de développer un alignement cognitif (Pichat, 2023, 2024a ; Khamassi 2024) de la cognition artificielle sur l'humaine, même si cela peut parfois relever d'une paradoxale injonction contradictoire consistant à attendre des systèmes d'intelligence artificielle à la fois qu'ils dépassent et qu'ils respectent la pensée humaine.

Dans le cadre de ce travail, nous nous focalisons sur ce que l'on peut épistémologiquement qualifier d'une explicabilité à faible granularité cognitive, aussi nommée explicabilité mécaniste. C'est-à-dire, une explicabilité qui n'analyse pas cognitivement au niveau macroscopique les sorties finales d'un réseau de neurones en fonction de ses entrées ; par exemple ses biais cognitifs, ses biais culturels ou l'effet d'un étayage cognitif au sens de Bruner (1990) de la « chaîne de pensée » en prompt engineering (Zhen et al., 2024). Mais une explicabilité microscopique dont l'unité d'observation est le neurone formel (seul ou combiné à d'autres au sein de couches ou de clusters intra- ou inter-couches). Cette démarche explicative à faible granularité ayant pour finalité de rentrer dans le système cognitif interne du réseau de neurones artificiels en fabriquant des éléments de compréhension quant à la façon dont les catégories de pensée et concepts sont plus ou moins localement encodés et structurés au sein du modèle de langage (Dalvi et al., 2019, 2022). Il s'agit donc d'interpréter comment les connaissances catégorielles voire les processus cognitifs sont élaborés et mis en œuvre par les neurones formels (Fan et al, 2023).

2.2 Objets d'étude de l'explicabilité synthétique

Dans le registre de l'explicabilité à grain fin, les études portent schématiquement soit sur les neurones de type perceptron multicouche (Bricken 2023) soit sur les têtes attentionnelles des transformers (Clark et al., 2019). Concernant les têtes d'attention, il s'agira de comprendre le type de connaissances catégorielles ou de processus cognitifs encodés par les poids attentionnels. Ici, l'intérêt porte par exemple sur les effets de l'attention sur l'analyse de la fiabilité des données dans les couches initiales et médianes (Li et al., 2023), l'identification attentionnelle de caractéristiques catégorielles syntaxiques tels les compléments d'objet indirect (Wang et a., 2022) ou encore l'effet mnésique des poids attentionnels (Geva et al., 2021).

En termes de type d'entités cognitives étudiées au niveau neuronal, on peut distinguer deux classes de travaux d'explicabilité : ceux ayant trait à des contenus cognitifs (catégories, concepts) et ceux portant sur des processus cognitifs (circuits). Dans le premier registre, le sujet d'étude est principalement la relation entre neurones et catégories conceptuelles spécifiques (Sajjad et al, 2022 ;

Foote et al., 2023). Dans le second, les études s'intéressent par exemple à l'effet des poids de connexion entre neurones vis-à-vis de la réalisation d'opérations logiques élémentaires (ET, OU, etc.) dans certaines branches neuronales (Voss et al., 2021) ; ou encore à des sous-ensembles de neurones impliqués dans la prise de décision (Antverg & Belinkov, 2022).

En termes d'empan neuronal, deux types d'investigations peuvent être dissociés. Ceux se focalisant sur des encodages catégoriels propres à des neurones ou têtes attentionnelles isolés (Bills et al., 2023 ; Jaunet et al., 2021) et ceux, à spectre plus large, cherchant à tracer l'effet des connections neuronales sur la constitution de circuits spécifiques homogènes quant à leur activité cognitive (Anthropic, 2023 ; Olah et al, 2020 ; Bricken 2023).

Mentionnons enfin les études destinées à expliquer des phénomènes cognitifs statiques (la catégorie « capturée » ou le processus cognitif encapsulé dans un neurone ou un assemblage de neurones) par opposition à celle positionnées dans une perspective dynamique. Ces dernières vont par exemple analyser l'évolution de l'information attentionnelle à travers les têtes attentionnelles et les couches (Yeh et al., 2023), décomposer la représentation des tokens en n vecteurs intermédiaires au fil des couches (Modarressi et al., 2022 ; Yang et al., 2023) ; ou encore suivre comment, à l'issue de chaque couche, les nouveaux embeddings sont constitués par combinaison de l'embedding attentionnel et de l'embedding perceptron d'une couche, afin de tracker le traitement interne progressif des représentations vectorielles (Kobayashi et al., 2023).

2.3 Exemples d'explications catégorielles synthétiques à granularité faible

Diverses recherches dévoilent ou plutôt infèrent une variété de catégories (linguistiques, logiques, positionnelles, etc.) encodées au sein des neurones et des têtes d'attention. Nous en présentons certaines ici, en omettant donc les travaux relatifs au neurones porteurs de processus cognitifs.

Dans le cadre de l'expérimentation classique de Clark et al. (2019) sur BERT, les auteurs mettent en lumière les fonctions linguistiques convergentes des têtes d'attention issues de mêmes couches :

- Identification de catégories linguistiques syntaxiques ou morphologiques : objets directs des verbes, objets des déterminants du nom, objets des propositions, objets des pronoms possessifs, verbes modifiés par les verbes auxiliaires passifs.
- Identification de catégories linguistiques relatives à la coréférence : antécédents des mentions co-référentes (she/her, talks/negociations).
- Identification de catégories de type séparateur permettant la segmentation et la délimitation du texte : points, token séparateur “SEP”.
- Identification de catégories positionnelles : prochain token, token précédent.

Dans leur passionnante étude sur GPT2-XL, Bills et al. (2023) manifestent une série de neurones singuliers, en pointant pour certains leur forte prise en compte des éléments de contexte :

- Des neurones catégoriels associés à tout token relevant d'un champ lexical spécifique. Par exemple un neurone s'activant pour des mots décrivant un mouvement impliquant les pieds (ran, walked, danced, kicked, hopped, stepping, kicked, tiptoed) ; ou un neurone lié aux choses réalisées correctement.

- Des neurones associés à la catégorie des phrases possédant une certaine valence sémantique, tel le neurone "simile" ayant trait aux phrases impliquant la certitude ou la confiance.

- Des neurones détecteurs d'anomalies, par exemple la catégorie des mots tronqués ou étranges.

- Des neurones réagissant à un token précis mais uniquement dans un contexte linguistique donné, formant ainsi une catégorie contextuelle relative à un token. A l'instar du neurone "hypothetical had" s'activant pour le token "had" dans un contexte hypothétique ou dans lequel les choses auraient pu être autrement; ou un neurone actif pour "together" mais seulement lorsque précédé par "get"; ou encore des neurones devenant opérationnels pour des mots spécifiques en début de texte.

- Des neurones détectant des séquences logiques comme la catégorie de la répétition de tokens identiques ou celle de la rupture de la logique d'une séquence (1, 2, 3, 5).

- Des neurones d'anticipation catégorielle (clairement liés à la finalité pour laquelle le modèle a été entraîné) devenant fonctionnels dans un contexte correspondant à un probable prochain token, par exemple lorsque le prochain token est probablement "from".

Les études pointant également une distribution géographique du type spécifique d'activité catégorielle neuronale réalisé en fonction du niveau de profondeur des couches qui les portent. Ainsi les têtes d'attention des premières couches ont une attention plus large par rapport aux autres plus centrées sur les tokens (Clark et al, 2023). Ou le fait que les premières couches réagissent plus à des catégories d'éléments morphologiques au niveau des mots, alors que les couches plus tardives sont plus sensibles aux caractéristiques catégorielles syntaxiques relatives aux phrases (voix passive/active, temps) ainsi qu'aux informations catégorielles sémantiques (Jawahar et al., 2019).

3 Catégorisation synthétique et catégorisation humaine

3.1 Principales caractéristiques catégorielles d'un réseau de neurones artificiels
neurones artificiels

Un réseau de neurones artificiels, notamment un modèle de langage, est organisé en trois composantes : des couches d'entrée et de sortie, ayant respectivement une fonction perceptive et effectrice (Savioz et al., 2010), et des couches intermédiaires dites cachées. Dans le cas des transformers, ces couches intermédiaires regroupent des couches de type perceptron multicouches, qui sont les moins étudiées (Garde et al., 2023) et sur lesquelles va porter notre présent travail, et des couches attentionnelles.

Les couches de type perceptron multicouches se matérialisent par une fonction d'agrégation et une fonction d'activation. La fonction d'agrégation, de la forme $\sum(w_{i,j}x_{i,j})+a$ et dont les paramètres spécifiques sont propres à chaque neurone, réalise en sortie la segmentation catégorielle (i.e. crée une nouvelle catégorie) constitutive de ce neurone, sur la base de la représentation vectorielle qui lui arrive en entrée. Dans un premier temps, chaque poids $w_{i,j}$ du vecteur d'agrégation indique dans quelle mesure sa dimension catégorielle j associée, de l'espace vectoriel sémantique d'entrée, est sélectivement à prendre en compte afin de générer la nouvelle dimension catégorielle de sortie. Autrement dit, chaque poids est un sélecteur épistémologique (Pichat, 2024b) qui réalise, d'un point de vue cognitif, une activité d'attention sélective catégorielle, en régissant le niveau d'importance à accorder à une caractéristique dimensionnelle catégorielle d'entrée donnée. Dans un second temps, l'ensemble des produits (poids attentionnel x dimension catégorielle) sont additionnés. Cette concaténation additive réalise cognitivement une activité de fusion épistémologique pondérée (Pichat, 2024) des dimensions catégorielles d'entrée impliquées. Il résulte, *in fine*, de cette combinaison linéaire pondérée catégorielle, la création d'un nouveau segment catégoriel (i.e. une nouvelle dimension catégorielle j', une nouvelle catégorie), plus abstraite, pertinente pour l'activité finalisée du réseau de neurones. Au niveau neurobiologique, par analogie comparative, les poids sont instanciés par l'efficacité synaptique, c'est-à-dire la quantité dans laquelle la synapse libère son neurotransmetteur dans la fente synaptique (Savioz et al., 2010).

La fonction d'activation (Sigmoïde, Tanh, GeLU, ReLU, Leaky ReLU, Softmax, etc.), outre une normalisation éventuelle des sorties des neurones, introduit de la non-linéarité dans le système neuronal. D'un point de vue cognitif, cette non-linéarité augmente le contraste catégoriel (meilleure distinction signal / bruit) et facilite alors la convergence de l'activation des neurones pour certaines catégories (i.e. facilite la construction de catégories stables et différenciantes). Elle permet une hiérarchisation de la distribution des catégories construites par le réseau (catégories plus élémentaires sur les premières puis augmentation

progressive de leur complexité dans les couches ultérieures) et une sparsité relative du réseau. Cette sparsité, c'est-à-dire l'activation limitée des neurones en fonction du type d'entrée, se traduit par une prévention du sur-apprentissage (catégories trop spécifiques rendant impossibles la généralisation et l'adaptation à la diversité) et par une réduction du coût computationel du réseau. Le corollaire biologique de la fonction d'activation est la fonction de transfert (Savioz et al., 2010), de type 1/(1+exp(-(G*activation net+b))), à l'échelle de la génération du potentiel d'action au niveau du hile de l'axone ainsi qu'à l'échelle de la neuromodulation par la dopamine et la noradrénaline au niveau de la synapse (Servan-Schreiber, 1990).

Les têtes attentionnelles (Vaswani et al, 2017) reposent quant à elles sur le mécanisme fondamental de l'auto-attention permettant de calculer des pondérations contextuelles attentionnelles de chaque partie catégorielle des données d'entrée ; cela, à travers trois matrices apprises de poids (queries Q, keys K, values V). Les scores d'attention ainsi déterminés, par comparaison de chaque élément avec tous les autres éléments d'une séquence informationnelle donnée, rendent possibles une sélectivité informationnelle (i.e. une focalisation de l'attention sur des catégories d'éléments pertinents au détriment des non pertinents) et une flexibilité contextuelle (i.e. un ajustement du poids attentionnel accoré à certains éléments catégoriels en fonction du contexte global et des relations entre éléments). A vocation fondamentalement organisationnelle, chaque tête attentionnelle permet de capter ou plutôt de construire un type de relation structurelle catégorielle entre les différents éléments d'une séquence d'entrée puis d'injecter ces « méta-informations catégorielles » dans les représentations des éléments à traiter, ces représentations étant ainsi informationnellement enrichies des relations catégorielles qu'entretiennent ces éléments avec les autres éléments importants de la séquence en jeu.

3.2 La catégorisation humaine

Dans le registre de la pensée humaine, la catégorisation est essentielle dans diverses activités cognitives, telles que la classification, l'identification des objets, la compréhension, le raisonnement, la résolution de problèmes, la mémorisation, l'inférence, la prédiction et la conceptualisation (Sternberg, 2007 ; Roads et al., 2024).

Rosch (1975) propose une approche de la catégorisation basée sur la ressemblance d'un objet à un prototype de la catégorie, notant que les individus tendent à énoncer des traits caractéristiques plutôt que des propriétés déterminatives (Rosch et Mervis, 1975) ; le prototype étant l'exemple le plus représentatif de la catégorie (Singh et al., 2020 ; Vogel et al., 2021). La théorie de la catégorisation par l'exemplaire (Medin et Schaffer, 1978 ; Nosofsky, 1992 ; Nosofsky et al., 2022), quant à elle, suggère quant à elle que les objets sont comparés à des exemples typiques stockés en mémoire, l'exemplaire le plus typique étant celui qui ressemble le plus aux exemplaires connus. Enfin, l'approche contextuelle ou finalisée par le but de l'action (Barsalou, 1983 ; Glaser et al., 2020) met l'accent sur la finalité de la situation pour définir une catégorie, plutôt que sur

une logique ou une sémantique générale.

Les approches de la catégorisation par similarité, qui nous intéressent plus dans le cadre du présent travail, postulent qu'un objet est assimilé à une classe en fonction de sa proximité avec une représentation de cette classe (Thibault, 1997 ; Jacob et al., 2021 ; Kaniuth et al., 2022 ; Roads et al., 2021, 2024). Les théories du prototype et de l'exemplaire mentionnées à l'instant mettent en avant la similarité comme base de la catégorisation (Sanborn et al., 2021 ; Ayeldeen et al., 2015 ; Roads et al., 2024). Les critiques de la catégorisation par similarité soulignent cependant que le choix des critères de jugement de similarité est arbitraire et peut ne pas correspondre aux fondements de l'attribution catégorielle (Love, 2002 ; Kalyan et al., 2012 ; Reppa et al., 2013 ; Poth, 2023). Le raisonnement par similarité est alors jugé trop équivoque pour être fonctionnel (Wixted, 2018). Les défenseurs de la similarité dans la catégorisation (Bobadilla et al., 2020 ; Hebart et al., 2020) contre-argumentent dès lors que ces critiques reposent sur deux erreurs épistémologiques : (i) une erreur réaliste qui suppose que la catégorisation doit saisir un réel prédéfini, et (ii) une erreur rationaliste qui impose une logique normative à la catégorisation, logique que tout individu devrait comprendre.

4 Problématique

Ainsi que nous l'avons mentionné, l'objectif de l'explicabilité synthétique est de rendre les opérations d'un réseau de neurones artificiels accessibles à la compréhension humaine (Du et al., 2019). Cela nécessite de convertir le comportement observable des réseaux neuronaux en un cadrage interprétatif contenant des éléments explicatifs compatibles avec les référentiels cognitifs de la pensée humaine. Dans ce contexte, la psychologie cognitive humaine apparaît comme un cadre heuristique pertinent pour fabriquer des analogies explicatives de la cognition synthétique. Et plus particulièrement la psychologie cognitive de la catégorisation dans la mesure où le traitement de l'information synthétique relève de façon significative d'un comportement de segmentation et d'analyse catégorielles (Pichat et al., 2024) ; en effet, nombre de travaux (Jawahar et al., 2019 ; Clark et al., 2019 ; Bills et al. 2023 ; Clark et al, 2023) mettent en lumière la cognition artificielle des modèles de langage comme reposant en grande partie sur une dynamique d'extraction d'invariants catégoriels linguistiques au sens large.

Dans le registre de cette étude, comme indiqué déjà, nous nous focalisons sur une explicabilité épistémologique à faible granularité cognitive (Pichat, 2024). En d'autres termes, nous examinons une explicabilité microscopique où l'unité d'observation est le neurone formel. Cette approche interprétative à faible granularité vise à pénétrer directement le système "boîte noire" que constitue un réseau de neurones artificiels, en créant des éléments de compréhension sur la manière dont les catégories de pensée et les concepts sont encodés et structurés localement au sein d'un modèle de langage (Dalvi et al., 2019, 2022). L'objectif est donc d'interpréter comment les connaissances catégorielles sont construites

et mobilisées par les éléments fondamentaux des réseaux, à savoir les neurones eux-mêmes (Fan et al., 2023). En ce qui concerne la question spécifique de la mobilisation, nous nous centrons ici sur la question particulière de la relation entre activation et similarité catégorielle.

En effet, en lien avec la problématique que nous avons soulevée, dans le domaine de la cognition humaine, concernant la relation entre catégorisation et similarité, nous avons, dans une étude précédente (Pichat et al., 2024), transposé cette question relationnelle heuristique au domaine de la cognition artificielle en nous soumettant à l'interrogation suivante : le degré d'appartenance (opérationnalisé en termes de niveau d'activation) des tokens (reçus par un neurone sous forme d'embedding) à la catégorie associée à ce neurone est-il lié à leur niveau de similarité (opérationnalisé en termes de cosinus similarité) ? En d'autres termes, l'intensité de l'appartenance catégorielle et l'intensité de similarité des tokens, telles qu'évaluées par un neurone, sont-elles deux facettes d'un même phénomène ? Cette question, largement inexplorée à ce jour dans le domaine de l'explicabilité des systèmes artificiels (Fan et al., 2023 ; Luo et al., 2024 ; Zhao et al., 2024), nous semblait particulièrement pertinente à examiner.

Dans le cadre de notre étude antérieure, nous avons montré la compatibilité de nos résultats avec deux hypothèses qui ont été formulées : (i) une discontinuité catégorielle des core-tokens (i.e. à fort niveau moyen d'activation) successifs en termes de niveau d'activation (suggérant des cosinus de similarité particulièrement bas entre ces core-tokens) et, (ii) une hétérogénéité catégorielle des core-tokens ayant des niveaux d'activation similaires (ces core-tokens ne sont pas les plus proches en termes de cosinus similarité). Ces deux hypothèses se complètent mutuellement et concernent toutes deux le sujet de la relation globale entre proximité d'activation et similarité cosinus. Mais elles se positionnent au sein d'une approche statique, visant à investiguer la relation entre proximité d'activation et similarité cosinus indépendamment des niveaux de valeurs des activations des tokens impliqués. Or d'autres résultats, non encore exploités durant cette recherche initiale, nous invitent à poursuivre cette investigation sous un angle distributionnel, s'interrogeant dès lors sur une possible évolution de la relation entre proximité d'appartenance catégorielle (i.e. proximité d'activation) et proximité catégorielle (i.e. proximité ou similarité cosinus) en fonction des niveaux d'activation des tokens impliqués. C'est le travail auquel nous nous livrons dans le cadre de cette présente recherche.

5 Methodologie

5.1 Méthodologies de l'explicabilité synthétique

A des fins de contextualisation méthodologique de notre étude, nous présentons ici un bref panorama non exhaustif d'approches techniques, à faible ou à forte granularité, visant à inférer les contenus ou les processus cognitifs encapsulés dans les neurones formels comme dans leurs assemblages (en couches, en clusters ou en réseau global) ; ces méthodes n'étant pas exclusives les unes des autres,

elles présentent dès lors une relative porosité.

Les techniques à forte granularité, comme évoqué déjà, sont ancrées sur le contraste entrée / sortie et se donnent pour objectif d'étudier la relation entre informations initiales et informations finales d'un modèle de langage. Dans ce registre, les méthodes basées sur les gradients visent à mesurer l'importance de chaque entrée en analysant les dérivées partielles de la sortie par rapport à chacune des dimensions d'entrée (Enguehard. 2023). Que les caractéristiques des entrées soient par exemple mesurées en termes de trait (Danilevsky et al., 2020), de score d'importance de tokens (Enguehard, 2023) ou de poids attentionnel (Barkan et al., 2021). De façon connexe, les approches fondées sur les exemples visent à comprendre dans quelle mesure l'output change avec différents inputs. Cela, en montrant comment les sorties des réseaux sont impactées par de petits changements en entrée (Wang et al., 2022) ou par des altérations (suppression, négation, mélange, masquage) des entrées (Atanasova et al., 2020 ; Wu et al., 2020 ; Treviso et al., 2023). Citons enfin ici les travaux effectuant un mapping conceptuel des entrées puis mesurant la contribution de ces concepts aux sorties constatées (Captum, 2022).

Les méthodologies à granularité fine, nous l'avons déjà abordé également, prennent comme sorties non par l'output final du modèle de langage étudié mais ses outputs ou états intermédiaires au niveau de neurones ou de clusters ou couches de neurones. Dans ce cadre, certaines méthodes visent à décomposer linéairement le score de pertinence d'un neurone d'une couche donnée en fonction de ses inputs (neurones, têtes attentionnelles ou tokens) dans la couche précédente (Voita et al., 2021). D'autres méthodes s'attachent à linéariser les fonctions d'activation afin de rendre plus aisée l'activité d'interprétation neuronale (Wang et al., 2022). D'autres méthodes encore, basées sur le vocabulaire du modèle, cherchent à repérer les connaissances encapsulées en projetant les poids de connexion comme les représentations intermédiaires dans l'espace de vocabulaire de ce modèle via une matrice de désembedding (Dar et al., 2023 ; Geva et al., 2023). Enfin, mentionnons les démarches fondées sur les statistiques d'activation neuronale en réponse à des corpus (Bills et al., 2023 ; Mousi et al., 2023 ; Durrani et al., 2022 ; Wang et al., 2022 ; Dai et al., 2022). C'est dans le cade spécifique de ces dernières démarches que s'inscrit notre présente étude.

5.2 L'étude d'explicabilité source de nos données

Nos données sont issues de la passionnante étude Bills et al. (2023). Partant de l'hypothèse qu'un neurone s'active spécifiquement pour une propriété à déterminer, cette propriété pouvant inclure le contexte, les auteurs mettent en place une vaste étude d'interprétation de la sémantique catégorielle de la totalité des neurones de GPT-2XL.

Méthodologiquement, les chercheurs d'OpenAI procèdent de la façon suivante. Ils soumettent GPT-2XL (le modèle "sujet") à une large série de séquences de 64 tokens, extraites au hasard des données issues d'internet avec lesquelles le modèle a été entraîné. Pour chaque token, ses valeurs d'activation pour l'ensemble des neurones de la totalité des couches sont enregistrées. Un mod-

èle plus élaboré, GPT-4, le modèle "explicateur", est ensuite mobilisé pour identifier de façon automatisée les éléments auxquels réagissent chaque neurone (i.e. pour générer "l'explication"), sur la base d'un prompt d'instruction et d'exemple (few shot learning) opérant uniquement sur les cinq séquences textuelles à activation maximale (i.e. contenant au moins un token à activation maximale), déterminées par le quantile des activations maximales. GPT-4 est par la suite utilisé comme modèle "simulateur" : sur la base d'un prompt lui indiquant "l'explication" de chaque neurone, le simulateur doit prédire le niveau d'activation de chaque token pour les mêmes séquences de 64 tokens. Enfin, toujours pour chaque neurone, ses activations réelles et prédites pour chaque token sont comparées pour calculer un score d'explication supposé mesurer la qualité de l'interprétation générée.

Les principaux résultats de l'étude sont les suivants. Concernant les tokens à mobiliser pour l'explication neuronale : la focalisation sur les 5 top activations des neurones est interprétée par les auteurs comme la plus efficace en termes de score de prédiction et l'augmentation du nombre de tokens n'augmente pas significativement le score de prédiction ; l'injection de tokens à plus faibles activations, quant à elle, diminue le score de prédiction. Concernant la dimension quantitative des scores d'explication : le score moyen d'explication est faible à 0.15 (seulement 1000 neurones sur 307200 ont un score supérieur à 0.8), ce score diminue avec la profondeur des couches, les scores d'explication croissent avec l'augmentation de la complexité du modèle explicateur comme avec sa sparsité. Concernant la dimension qualitative des scores d'explication : les explications de GPT-4 comme les humaines présentent de faibles scores de prédiction, les explications automatiques générées sont trop larges (hyperonymes descriptifs trop vastes là où des hyponymes seraient plus spécifiques aux données précises impliquées).

Dans le cadre de notre présent travail, nous repartons des données d'activation, neurone par neurone, obtenues pour le vaste empan de tokens mobilisés dans l'étude menée par Bills et al. (2023). Données que nous réutilisons à d'autres fins, dans le but d'étudier, comme évoqué, la relation entre activation et similarité au niveau de la cognition catégorielle neuronale synthétique.

5.3 Sélection et interprétation des données

Nous présentons ici de façon synthétique les choix méthodologiques que nous avons opérés dans le cadre de cette présente étude, en continuité de ceux réalisés durant notre étude préalable (Pichat et al., 2024) relative à la relation entre appartenance catégorielle (activation) et proximité catégorielle (similarité), dont ce travail actuel est la poursuite.

Pour simplifier notre étude, nous avons limité notre analyse aux deux premières couches de GPT-2XL (couches 0 et 1) et aux 6400 neurones de chaque couche. Pour chaque neurone parmi ces 12800 neurones au total, nous avons choisi de considérer ses 100 tokens les plus activés en moyenne (les core-tokens), ainsi que leurs valeurs d'activation respectives. Cette approche diffère de celle de Bills et al. (2023), qui se concentre uniquement sur les tokens à hyperac-

tivation. Nous estimons en effet cette méthode en partie limitée (même si elle permet de précieux effets comme indiqué par les auteurs) car elle ne capture pas la variabilité des tokens pour lesquels un neurone s'active, et préférons ainsi une vue plus complète de la catégorie de tokens à laquelle un neurone réagit.

Nous considérons que le niveau d'activation moyen d'un token dans un neurone est une bonne mesure de l'appartenance catégorielle de ce token à la catégorie neuronale impliquée. En effet, l'activation moyenne des tokens les plus activés semble bien représenter la mesure dans laquelle ces tokens font partie de l'extension d'une catégorie. Cela est en phase avec l'hypothèse de Bills et al. (2023) selon laquelle un neurone s'active pour une propriété spécifique.

Le cosinus de similarité entre deux tokens semble également une bonne mesure de la similarité catégorielle entre items. Cela, en phase avec Thibault (1997) définissant la similarité comme basée sur une modalité de calcul de distance entre dimensions catégorielles comparées. Le cosinus de similarité, couramment utilisé en NLP pour mesurer la proximité sémantique (Ham, 2023), s'aligne bien avec cette définition.

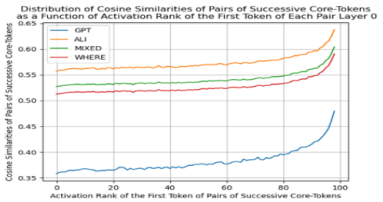
Nous avons choisi de mesurer le cosinus de similarité au sein de la base d'embeddings de GPT-2XL pour éviter les limites méthodologiques mentionnées par Bills et al. (2023) et Bricken (2023), qui consistent à apparier des systèmes cognitifs synthétiques basés sur des systèmes d'embeddings différents. Pour comparaison et vérification, nous avons également utilisé trois autres bases d'embeddings librement disponibles : Alibaba-NLP/gte-large-en-v1.5, Mixedbread-ai/mxbai-embed-large-v1, et WhereIsAI/UAE-Large-V1.

6 Résultats

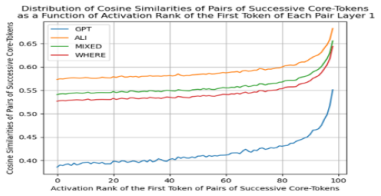
6.1 Évolution de la relation entre proximité cosinus et activation

Nous étudions, pour rappel, concernant les core-tokens successifs de chaque neurone, la relation entre proximité d'activation et proximité cosinus, c'est-à-dire la relation entre la proximité de niveau d'appartenance catégorielle entre deux tokens et la proximité catégorielle entre ces deux tokens. Cela, en nous focalisant plus spécifiquement sur la dynamique de cette distribution en fonction des niveaux d'activation. Autrement dit, nous nous posons la question de l'éventuelle évolution, au fil des segments activationnels, de la relation entre proximité cosinus et niveau d'activation des core-tokens successifs. Les graphes n°1 (couche 0) et n°2 (couche 1) présentent la distribution moyenne des cosinus similarité des paires de core-tokens successifs en fonction des rangs du niveau d'activation du premier token de chaque paire. Nous pouvons assez limpidement y voir une évolution, d'abord très lente (rangs 0 à 40), puis plus rapide (rangs 40 à 80) et enfin d'allure exponentielle (à partir du rang 80), de la proximité catégorielle en fonction de la proximité activationnelle.

Plusieurs points peuvent être notés : (i) cette évolution semble invariante pour les 4 bases d'embeddings (même si la dynamique de croissance est plus



Graphe 1 : distribution des cosinus similarité entre paires de core-tokens successifs en fonction du rang d'activation du premier token de chaque paire (Couche 0, n=6400).



Graphe 2 : distribution des cosinus similarité entre paires de core-tokens successifs en fonction du rang d'activation du premier token de chaque paire (Couche 0, n=6400).

	GPT2	Alibaba	Mixedbread	WhereIsAI
Min	0.06	0.42	0,41	0,4
Max	0.82	0.88	0,85	0,84
Mean	0.38	0.57	0,54	0,53
s	0.16	0.08	0,07	0,07
CV	0.42	0,14	0,13	0,13
Range	0.76	0,46	0,44	0,45
Q_1	0.27	0,52	0,5	0,48
Q_2	0.37	0,56	0,53	0,52
Q_3	0.48	0,61	0,57	0,56

Tableau 1 : moyennes statistiques des indices descriptifs de position et de dispersion des cosinus similarité des paires de core-tokens successifs (Couche 0, n=6400).

	GPT2	Alibaba	Mixedbread	WhereIsAI
Min	0,1	0,43	0,42	0,41
Max	0,87	0,94	0,93	0,92
Mean	0,42	0,59	0,56	0,54
s	0,16	0,09	0,09	0,09
CV	0,38	0,15	0,15	0,16
Range	0,77	0,51	0,51	0,52
Q_1	0,31	0,53	0,51	0,49
Q_2	0,4	0,58	0,54	0,53
Q_3	0,51	0,63	0,59	0,57

Tableau 2 : moyennes statistiques des indices descriptifs de position et de dispersion des cosinus similarité des paires de core-tokens successifs (Couche 1, n=6400).

marquée pour les embeddings de GPT-2XL, dans la mesure où ils sont plus discriminants comme évoqué dans Pichat et al., 2024), (ii) le segment de croissance exponentielle semble encore plus marqué pour la couche 1 par rapport à la couche 0 (une accélération serait-elle à noter au fil de la profondeur des couches ?), (iii) à partir du rang 80 pour la couche 0 et du rang 40 pour la couche 1, les moyennes de cosinus similarité par rang deviennent supérieures aux moyennes globales (respectivement 0.38 et 0.42 selon les embeddings de GPT-2XL, cf tableaux n°1 et n°2), (iv) ce phénomène de croissance de la proximité cosinus en fonction de la proximité d'activation est malgré tout borné par des *maxima* pas très élevés de cosinus similarité (respectivement 0.48 et 0.55 avec les embeddings de GPT-2XL).

Cette première vue, exploratoire et descriptive, de l'évolution monotone positive de la distribution des cosinus similarité en fonction du niveau d'activation nous donne à interroger de façon plus formelle l'hypothèse du phénomène de cognition synthétique suivant : celui de la convergence catégorielle des paires de core-tokens successifs (quant à leur niveau d'activation) en fonction de l'augmentation de ces niveaux d'activation. Autrement dit, hypothèse postulant le fait que plus les niveaux d'activation des core-tokens successifs (i.e. proches au niveau activationnel) augmentent et plus la variabilité catégorielle de ces core-tokens diminue (i.e. plus la proximité catégorielle augmente). Nous allons tester dans ce qui suit cette hypothèse à partir de différentes opérationnalisations.

	Min(x)-Q1(x)	Q1(x)-Q2(x)	Q2(x)-Q3(x)	Q3(x)-Max(x)
Observed Freq.	357	316	240	144
% of Observed Freq.	33,77	29,90	22,71	13,62
Expected Freq.	264,25	264,25	264,25	264,25
Weighted χ^2 Residuals	0,35	0,20	-0,09	-0,46
% of χ^2 Contribution	32,67	10,17	2,23	54,92
p(χ^2)	0,0000			

Tableau 3 : distribution des outliers inférieurs des cosinus similarité des paires de core-tokens successifs en fonction des quartiles d'activation du premier token de ces paires (Couche 0, N = 6400, outliers issus des embeddings de GPT-2XL avec étendue interquartile).

	Min(x)-Q1(x)	Q1(x)-Q2(x)	Q2(x)-Q3(x)	Q3(x)-Max(x)
Observed Freq.	367	314	242	177
% of Observed Freq.	33,36	28,55	22,00	16,09
Expected Freq.	275,00	275,00	275,00	275,00
Weighted χ^2 Residuals	0,33	0,14	-0,12	-0,36
% of χ^2 Contribution	40,93	7,36	5,27	46,45
p(χ^2)	0,0000			

Tableau 4 : distribution des outliers inférieurs des cosinus similarité des paires de core-tokens successifs en fonction des quartiles d'activation du premier token de ces paires (Couche 1, N = 6400, outliers issus des embeddings de GPT-2XL avec étendue interquartile).

6.2 Convergence Catégorielle et Valeurs Extrêmes de Cosinus Similarité

Une première opérationnalisation de nature à tester notre hypothèse de convergence catégorielle des paires de core-tokens successifs en fonction de l'augmentation des niveaux d'activation peut se fonder sur l'étude de la distribution des valeurs inférieures extrêmes des cosinus similarité. En effet, conformément à cette hypothèse, le nombre de *minima* de cosinus similarité devrait décroître avec l'augmentation des valeurs d'activations.

Les tableaux n°3 (couche 0) et n°4 (couche 1) présentent la distribution du nombre d'outliers inférieurs des cosinus similarité des paires de core-tokens successifs en fonction des quartiles d'activation du premier token de ces paires (outliers issus des embeddings de GPT-2XL avec étendue interquartile). On constate, pour la couche 0, une sur-représentation (+0.35) des écarts pondérés relatifs au premier segment de quartilage des activations, décroissant progressivement jusqu'à une sous-représentation (-0.46) des écarts pondérés relatifs au dernier segment. De même, concernant la couche 1, nous trouvons une sur-représentation de +0.33 décroissant progressivement jusqu'à une sous-représentation de -0.36. Dans les deux cas, l'analyse inférentielle de χ^2 d'ajustement, avec une équi-distribution théorique de 25% conforme à notre segmentation en quartile, souligne cette diminution $p(\chi^2) < 0.05$.

Dans la même logique, intéressons-nous maintenant aux faibles valeurs de cosinus, que nous définissons, neurone par neurone, comme inférieures au seuil du minimum du cosinus du neurone augmenté de 10% de son étendue. Les tableaux 5 (couche 0) et 6 (couche 1) manifestent la même tendance de diminution progressive, au fil des quartiles d'activation, des effectifs de faibles cosinus ; tendance toujours significative $p(\chi^2) < 0.05$.

Ces deux lots de résultats mettent en lumière le fait que l'effectif de valeurs extrêmes faibles de coninus similarité diminue en fur et à mesure de l'accroissement des valeurs d'activation. Ils sont ainsi compatibles avec notre hypothèse de convergence catégorielle postulant le fait que plus les niveaux d'activation des core-tokens successifs augmentent et plus la variabilité catégorielle de ces core-tokens diminue.

	Min(x) -Q1(x)	Q1(x)- Q2(x)	Q2(x)- Q3(x)	Q3(x)- Max(x)
Observed Freq.	9040	9035	8193	6152
% of Observed Freq.	27,88	27,87	25,27	18,98
Expected Freq.	8105	8105	8105	8105
Weighted χ^2 Residuals	0,12	0,11	0,01	-0,24
% of χ^2 Contribution	15,72	15,55	0,14	68,59
$p(\chi^2)$	0,0000			

Tableau 5 : distribution des faibles valeurs des cosinus similarité des paires de core-tokens successifs en fonction des quartiles d'activation du premier token de ces paires (Couche 0, N = 6400, calcul avec embeddings de GPT-2XL).

	Min(x)- Q1(x)	Q1(x)- Q2(x)	Q2(x)- Q3(x)	Q3(x)- Max(x)
Observed Freq.	9650	8997	7727	6455
% of Observed Freq.	29,39	27,41	23,54	19,66
Expected Freq.	8207,25	8207,25	8207,25	8207,25
Weighted χ^2 Residuals	0,18	0,10	-0,06	-0,21
% of χ^2 Contribution	34,66	10,38	3,84	51,12
$p(\chi^2)$	0,0000			

Tableau 6 : distribution des faibles valeurs des cosinus similarité des paires de core-tokens successifs en fonction des quartiles d'activation du premier token de ces paires (Couche 1, N = 6400, calcul avec embeddings de GPT-2XL).

6.3 Convergence catégorielle et positivité de la monotonie relationnelle

Une deuxième opérationnalisation pertinente pour tester notre hypothèse de convergence catégorielle des paires de core-tokens successifs en fonction de l'accroissement des niveaux d'activation implique un angle plus fonctionnel (au sens mathématique du terme), en étudiant la monotonie et le sens de la relation liant cosinus similarité et valeur d'activation catégorielle. En phase avec notre hypothèse, nous devrions observer une relation monotone et positive (fonction croissante) entre ces deux variables.

Une première approche dans ce registre consiste à réaliser une étude de régression linéaire. Les tableaux 7 (couche 0) et 8 (couche 1) présentent les principaux termes de la régression linéaire du cosinus similarité des paires de core-tokens successifs en fonction de l'activation du premier token de chaque paire. Notons d'emblée une compatibilité variable et modérée de la condition d'application de la régression qu'est la normalité des résidus (les pourcentages de cas de tests inférentiels mobilisés associés à un $p > .05$ variant de 62% to 99% pour la couche 0 et de 53% to 98% pour la couche 1) : ces données de régression sont dès lors peu fiables (cf également les graphes 3 et 4 semblant aller dans le même sens de réserve). Une modélisation linéaire se prête peu à nos données, avec seulement 34% de tests de Fisher significatifs pour la couche 0, et 45 % pour la couche 1. Néanmoins, en changeant d'amplitude d'unités statistiques (i.e. en passant des tokens aux neurones), plusieurs résultats intéressants se manifestent. Premièrement, la moyenne du coefficient directeur a est positive, bien que faible (respectivement 0.06 pour la couche 0 et 0.04 pour la couche 1 avec les embeddings de GPT-2XL), cela étant relativement stable pour les 4 systèmes d'embeddings. Deuxièmement, le pourcentage de neurones assortis d'un tel coefficient directeur positif est extrêmement élevé (respectivement 80% pour la couche 0 et 85% pour la couche 1 via les embeddings de GPT-2XL), cela étant pleinement constant pour les 3 autres systèmes d'embeddings ; tendance largement significative au niveau inférentiel sur la base d'un χ^2 d'adéquation basé sur une équi-distribution dichotomique théorique $p(\chi^2) < 0.05$ pour les deux couches). *Modulo* notre réserve d'applicabilité, ce dernier résultat serait de nature à être compatible avec notre hypothèse, même si à nouveau, le choix d'une approche linéaire ne s'avère pas bien adapté ici. A titre d'illustration, le

	GPT	Ali	Mixed	WhereIs
% p(SW)>.05	61,97	27,16	23,53	20,44
% p(KS)>.05	98,86	90,17	80,08	75,72
% p(JB)>.05	72,98	28,16	22,75	20,13
Linear Relashionship Significance				
% p(F)<.05	34,45	39,97	40,89	41,41
Positivity of the Linear Relashionship				
Mean(a)	0,06	0,05	0,04	0,04
% a>0	79,59	82,02	82,38	82,69
p(χ^2) a>0	0,000	0,000	0,000	0,000

Tableau 7 : paramètres de la régression linéaire du cosinus similarité des paires de core-tokens successifs en fonction de l'activation du premier token de chaque paire (Couche 0, N = 6400).

	GPT	Ali	Mixed	WhereIs
% p(SW)>.05	52,80	12,75	9,33	8,02
% p(KS)>.05	97,97	75,27	55,50	50,83
% p(JB)>.05	63,39	13,25	9,38	8,09
Linear Relashionship Significance				
% p(F)<.05	45,31	46,73	50,45	50,80
Positivity of the Linear Relashionship				
Mean(a)	0,04	0,04	0,03	0,02
% a>0	85,23	83,28	85,05	85,20
p(χ^2) a>0	0,000	0,000	0,000	0,000

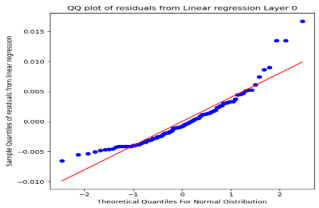
Tableau 8 : paramètres de la régression linéaire du cosinus similarité des paires de core-tokens successifs en fonction de l'activation du premier token de chaque paire (Couche 1, N = 6400).

graphe 5 nous produit un exemple de régression linéaire neuronale, illustrant la légère pente positive reliant les deux variables conformément à notre hypothèse (cf annexes pour une illustration par des neurones témoins des couches 0 et 1).

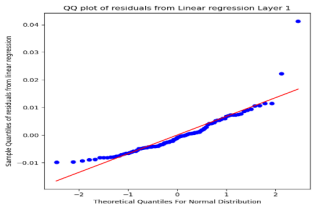
Nous complétons notre première approche méthodologique ci-dessus par une seconde démarche, non paramétrique cette fois-ci (en nous dégageant donc des conditions de normalité) et peut être plus adaptée dans la mesure où elle est ordinale, à travers un de spearman. Nous retrouvons le même type de résultats, que précédemment, dans les tableaux n°9 et n°10 relatifs aux paramètres de la relation ordinale du cosinus similarité des paires de core-tokens successifs en fonction de l'activation du premier token de chaque paire. Un pourcentage assez faible de cas où l'effet de corrélation ordinale est significatif (33% pour la couche 0 et 40% pour la couche 1), montrant à nouveau une insuffisante pertinence de ce nouveau mode de modélisation choisi pour rendre compte de notre effet postulé. Mais des moyens plus importants que leurs corollaires de coefficients directeurs linéaires précédents (0.10 pour la couche 0 et 0.13 pour la couche 1, avec les embeddings de GPT-2XL, valeurs stables pour les autres embeddings), faisant montre d'une légère meilleure adéquation de notre modélisation ordinale ; et toujours donc d'une positivité de la relation entre nos deux variables. Derechef, un pourcentage important de neurones associés à un coefficient positif de corrélation ordinale (75% pour la couche 0 et 81% pour la couche 1), résultat à nouveau significatif ($p(\chi^2) < 0.05$ for pour les deux couches).

Les résultats des deux approches méthodologiques complémentaires que nous venons de présenter ci-dessus sont pour partie compatibles avec notre hypothèse de convergence catégorielle des paires de core-tokens successifs en fonction de l'augmentation des niveaux d'activation.

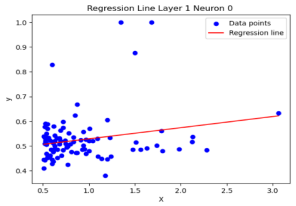
Précisons qu'une nouvelle vue sur les graphes n°1 et n°2, présentant la distribution des cosinus similarité entre paires de core-tokens successifs en fonction du rang d'activation du premier token de chaque paire, nous montre qu'une modélisation plus adaptée est en fait de la forme $y = a + \exp\left(\frac{x}{b} + c\right)$ (y représentant le



Graphe 3 : diagramme QQ-plot de comparaison distribution théorique normale / distribution effective relatif aux résidus de la régression du cosinus similarité sur l'activation (Couche 0, N=6400).



Graphe 4 : diagramme QQ-plot de comparaison distribution théorique normale / distribution effective relatif aux résidus de la régression du cosinus similarité sur l'activation (Couche 1, N=6400).



Graphe 5 : exemple de régression linéaire du cosinus similarité en fonction de l'activation (Couche 1, neurone 0).

	GPT2	Ali	Mixed	WhereIs
% p(ρ)<.05	32,52	35,56	35,38	35,75
Positivity of the Ordinal Relashionship				
Mean(ρ)	0,10	0,11	0,11	0,11
% ρ>0	74,45	77,27	77,66	78,11
p(χ^2) ρ>0	0,000	0,000	0,000	0,000

Tableau 9 : paramètres de la relation ordinale du cosinus similarité des paires de core-tokens successifs en fonction de l'activation du premier token de chaque paire (Couche 0, N=6400).

	GPT2	Ali	Mixed	WhereIs
% p(ρ)<.05	39,80	36,64	40,33	40,64
Positivity of the Ordinal Relashionship				
Mean(ρ)	0,13	0,12	0,13	0,13
% ρ>0	80,58	76,69	78,80	79,08
p(χ^2) ρ>0	0,000	0,000	0,000	0,000

Tableau 10 : paramètres de la relation ordinale du cosinus similarité des paires de core-tokens successifs en fonction de l'activation du premier token de chaque paire (Couche 1, N=6400).

cosinus similarité et x la valeur d'activation) ; le paramètre b, positif, exprimant le taux de croissance exponentielle du cosinus similarité (plus b est faible et plus ce taux de croissance est rapide) et le paramètre c, négatif, exprimant la zone d'activation correspondant à l'apparition de la cassure de croissance exponentielle du cosinus similarité ; le paramètre a dénotant, quant à lui, l'ordonnée à l'origine, c'est-à-dire le cosinus similarité minimal associé à l'activation minimale. Dès lors, une analyse de régression de type exponentiel est, bien mieux que nos approches linéaires ou ordinale, de nature à rendre compte d'une monotonie positive de la relation cosinus / activation, dans les conditions que nous avons décrites concernant les graphes 1 et 2, à savoir une évolution, d'abord très lente du cosinus en fonction de l'activation, puis plus rapide et enfin d'allure exponentielle.

6.4 Convergence Catégorielle et Moyennes Activationnelles Contrastées

Une troisième et dernière opérationnalisation pertinente pour tester notre hypothèse de convergence catégorielle des core-tokens successifs en fonction de l'activation consiste en une approche par contraste, consistant à comparer la proximité catégorielle de paires de core-tokens sucessifs entre groupes de tokens extrémisés quant à leur niveau d'activation. Selon notre hypothèse, pour chaque neurone, la moyenne des cosinus similarité des paires de core-tokens successifs devrait être supérieure pour les paires à fortes activations par rapport à celles à faibles activations.

Nous utilisons une approche non paramétrique, celle de Wilcoxon-Mann-Whitney, étant donnée la relative normalité de nos données et les groupes à faibles effectifs que nous allons mobiliser ici. Les tableaux n°11 (couche 0) et n°12 (couche 1) contiennent les principaux résultats ayant trait aux différences moyennes des moyennes de cosinus similarité des 21 paires de core-tokens successifs ayant les plus faibles / forts rangs d'activation du premier token de chaque paire. Nous pouvons observer un écart de moyennes globales systématique pour les 4 embeddings, même si les écarts obtenus avec les embeddings de GPT-2XL sont plus contrastés dans la mesure où ce modèle d'embeddings est plus différenciant ($\cos_{\min}$=.364 / $\cos_{\max}$=.417 en moyenne pour la couche 0; $\cos_{\min}$=.393 / $\cos_{\max}$=.459 pour la couche 1); avec cependant une significativité de cette différence (p(MW) < .05) pour seulement 30% des neurones de la couche 0 (associée à une taille moyenne d'effet effectivement négative et non négligeable de -0.17) et 37% de ceux de la couche 1 (associée à une taille d'effet effectivement négative et non négligeable de -0.21) ; notons également un contraste plus fort pour la couche 1 par rapport à la 0, évoquant la question d'une éventuelle accentuation de ce clivage au fil de la profondeur des couches. Si ce contraste n'est pas significatif notons cependant que, de façon largement majoritaire, pour les 4 embeddings, nous obtenons une supériorité de la moyenne de similarité cosinus dans les groupes de core-tokens à fortes activations par rapport aux faibles activations (76% des neurones de la couche 0 et 82% de ceux de la couche 1), avec dans tous les cas (les deux couches et les 4 embeddings) un effet significatif de ces

	GPT	Ali	Mixed	Where
$m(cos_{min})$	0,364	0,561	0,531	0,516
$m(cos_{max})$	0,417	0,596	0,562	0,548
% p(MW)<.05	30,34	33,22	33,23	33,23
m(Cliff's Δ)	-0,17	-0,19	-0,19	-0,19
% ($m(cos_{min}) < m(cos_{max})$)	76,17	80,52	80,58	81,02
$p(\chi^2)$ ($m(cos_{min}) < m(cos_{max})$)	0,000	0,000	0,000	0,000

Tableau 11 ; différences moyennes des moyennes de cosinus similarité des 21 paires de core-tokens successifs ayant les plus faibles / forts rangs d'activation du premier token de chaque paire (Couche 0, N=6400).

	GPT	Ali	Mixed	Where
$m(cos_{min})$	0,393	0,576	0,544	0,530
$m(cos_{max})$	0,459	0,619	0,588	0,574
% p(MW)<.05	36,52	34,75	37,05	37,00
m(Cliff's Δ)	-0,21	-0,20	-0,21	-0,21
% ($m(cos_{min}) < m(cos_{max})$)	82,31	82,50	83,86	84,36
$p(\chi^2)$ ($m(cos_{min}) < m(cos_{max})$)	0,000	0,000	0,000	0,000

Tableau 12 ; différences moyennes des moyennes de cosinus similarité des 21 paires de core-tokens successifs ayant les plus faibles / forts rangs d'activation du premier token de chaque paire (Couche 1, N=6400).

pourcentages sur la base d'un χ^2 d'ajustement fondé sur une équi-distribution dichotomique théorique $p(\chi^2) < 0.05$. Ces résultats sont à nouveau compatibles avec notre hypothèse de convergence catégorielle des paires de core-tokens successifs en fonction de l'augmentation de ces niveaux d'activation. Précisons qu'un contraste certainement nettement plus fort aurait été trouvé neurone par neurone si nous avions mobilisé des groupes de tokens plus contrastés quant à leur activation, en prenant par exemple comme classe de tokens à faible activation des tokens non issus des core-tokens, c'est-à-dire des 100 tokens les plus activés en moyenne par neurone.

Au terme de nos trois opérationnalisations complémentaires du test de notre hypothèse de convergence catégorielle postulant le fait que plus les niveaux d'activation des core-tokens successifs augmentent et plus la variabilité catégorielle de ces core-tokens diminue (i.e. plus la proximité catégorielle augmente), une partie majeure de nos exploitations statistiques obtenues semble compatible avec cette hypothèse.

7 Discussion

7.1 La Convergence Catégorielle pour les Fortes Activations comme Révélatrice de la Co-activation des Sous-dimensions Catégorielles des Catégories Synthétiques

Nous nous sommes questionnés, concernant les core-tokens successifs, à propos de l'éventuelle existence d'une évolution de relation entre proximité cosinus et activation. Nos résultats dans ce registre tendent vers une notion de convergence catégorielle des paires de core-tokens successifs en fonction de l'activation ; i.e. vers le fait que plus les niveaux d'activation des core-tokens successifs (i.e. proches au niveau activationnel) augmentent et plus la variabilité catégorielle de ces core-tokens diminue (i.e. plus la proximité catégorielle augmente). Comment penser ce résultat ? Nous allons mobiliser dans ce qui suit différentes clés explicatives partielles avant de tenter de les coordonner à des fins de réponse à cette question.

Dans leur travail relatif au lien entre catégorisation et similarité en psychologie cognitive humaine, Roads et al. (2024) mettent l'accent sur les représentations mentales de type géométrique, c'est-à-dire celles pour lesquelles la représentation d'un objet (à catégoriser ou dont il faut évaluer la similarité avec un autre) se traduit par des coordonnées multidimensionnelles dans un espace à plusieurs axes. Mode de représentation en phase avec les sous-dimensions catégorielles relatives à un neurone donné que nous allons évoquer dans ce qui suit. Dans ce type d'approche représentationnelle, Thibault (1997) indique que les travaux liant catégorisation et similarité partent du principe qu'un objet est assimilé à une classe par estimation de la proximité de celui-ci avec ce qui représente cette classe ; cela, sur la base (i) d'un espace de dimensions retenues comme pertinentes pour effectuer la comparaison, (ii) d'une modalité de calcul de distance entre les éléments comparés.

Ainsi que nous l'avons évoqué, les tenants d'une dissociation de la catégorisation et de la similarité (Murphy et Medin, 1985 ; Rips, 1989 ; Barsalou, 1991 ; Medin et al, 1993 ; Love, 2002 ; Kalyan et al., 2012 ; Reppa et al., 2013 ; Poth, 2023) opposent la subjectivité, la labilité et la versatilité des critères choisis hic et nunc pour fonder le jugement de similarité, critères qui ne sont qu'un choix singulier et subjectif parmi d'autres possibles dans l'espace des dimensions impliquées. Ces objections pointent que le raisonnement par similarité est équivoque, c'est-à-dire non assez contraint (Goodman, 1972 ; Wixted, 2018).

Mais, nous l'avons évoqué également, Goldstone (1994) contre-argumente que la similarité n'est pas toujours instable, et Thibault (1997) dénonce un essentialisme psychologique de ces contradicteurs en leur reprochant que l'argument de la faiblesse de la similarité prend pour parti pris que les critères de la segmentation catégorielle devraient avoir une valeur intrinsèque et stable, relevant d'une conception d'un réel ontologiquement défini à copier. Plus encore, Hampton (1997), dans un positionnement issu de la logique floue, leur rétorque que les critères de la catégorisation ne peuvent pas être pensés de façon univoque et étanche, c'est-à-dire comme relevant du mode de la logique classique. Cela, en phase avec les positions de divers défenseurs de la similarité dans la catégorisation (Bobadilla et al., 2020 ; Hebart et al., 2020).

Ces différentes contradictions nous semblent *in fine* en partie en phase avec notre observation de convergence de la similarité catégorielle des paires de core-tokens successifs en fonction de l'activation. En effet, cette convergence à partir d'un certain seuil activationnel exprime la discordance suivante : il n'y pas ou peu de relation entre similarité catégorielle et proximité catégorielle pour les « faibles » niveaux d'activation, alors que pour les plus fortes activations apparaît un lien entre ces deux phénomènes. Et cela, précisément en raison du type d'arguments et de contre-arguments résumés ci-avant, que nous allons transposer à la cognition synthétique dans ce qui suit, en conservant à dessein leur antagonisme.

Comme déjà évoqué dans notre précédente étude (Pichat et al., 2024), les tokens successifs à faibles niveaux d'activation sont associés à des faibles valeurs de cosinus similarité entre eux pour un neurone donné car ils correspondent en fait à une variabilité, une diversité de sous-dimensions catégorielles associées à ce neu-

rone ; un token donné relèvera d'une certaine sous-dimension catégorielle, alors que son voisin en terme d'activation relèvera d'une autre sous-dimension catégorielle, le tout formant un faible cosinus similarité traduisant une hétérogénéité, une disparité catégorielle (en tout cas, à partir de notre point de vue humain). Cela, en phase avec les arguments ci-avant en défaveur d'un lien entre catégorisation et similarité invoquant la versatilité des dimensions utilisées pour évaluer la similarité. Autrement dit, un neurone encode une catégorie complexe, multidimentionnelle ; catégorie qui n'est pas homogène mais polysémique (Fan et al, 2023, Bills et al., 2023). Pour un neurone donné, cette polysémie nous fait percevoir cette catégorie comme un alien concept (Bills et al., 2023) (ce qu'elle est effectivement pour notre cognition humaine) car elle résulte d'une superposition de différentes sous-dimensions catégorielles générées par la base vectorielle intermédiaire de ce neurone (base que nous ne concevons pas de la même manière que Bricken et al. (2023) mais plutôt, pour un neurone donné dans une couche n, en termes de dimensions catégorielles de sortie de ses neurones précurseurs dans la couche $n-1$).

Alors que, en ce qui les concerne, les tokens successifs à forts niveaux d'activation sont associés à de plus fortes valeurs de cosinus similarité entre eux. Cela, étant donné que, par construction mathématique de la fonction d'agrégation, ils ont un niveau élevé d'activation précisément parce qu'ils relèvent conjointement, simultanément de plusieurs sous-dimensions catégorielles (issues de leurs neurones précurseurs) co-activées. Co-activations de différentes sous-dimensions catégorielles qui ne peuvent se produire que dans la mesure où, dans les cas spécifiques des tokens fortement activés ici impliqués, ces différentes sous-dimensions se trouvent être catégoriellement, sémantiquement convergeantes pour les tokens impliqués ; autrement dit, que dans les situations particulières où les tokens en jeu sont à l'intersection sémantique de ces diverses sous-dimensions catégorielles. Ces intersections catégorielles produisent alors une réduction des degrés de liberté sémantiques et provoquent ainsi, à l'instar d'un puits de potentiel en physique, des *minima* locaux sémantiques ne pouvant que converger vers des éléments stables de sens, que vont exprimer alors les plus forts cosinus similarité qui sont les leurs. Autrement dit, à ces intersections catégorielles précises, ne peuvent exister qu'un nombre plus réduit de possibles sémantiques. Et cela va correspondre aux arguments cognitifs humains des tenants de critères plus fixes, invariables et unifiés (ou en tout cas nous apparaissant comme tels) de la catégorisation ou de la similarité évoqués ci-avant.

En phase avec notre observation empirique, cette explication tend à rendre compte de ce que Bricken et al. (2023), dans le registre de la cognition synthétique, indiquent : beaucoup de neurones apparaissent mono-sémantiques si on les regarde avec des tokens top activés mais se relèvent être poly-sémantiques si on les étudie sur la base de tokens à activations plus basses. Ou du fait que Bill et al. (2023), se cantonnent, à des fins de simplification, à l'interprétation des neurones formels uniquement à partir de leurs quelques tokens suractivés, en indiquant que la prise en compte des plus faibles activations baisse le pouvoir explicatif de leurs explications à visée mono-sémantique.

7.2 La convergence catégorielle « human like » comme manifestation des catégories synthétiques à l'interface des cognitions humaine et synthétique

Nous avons postulé que les cas de co-activation de différentes sous-dimensions catégorielles associées à un neurone génèrent une convergence catégorielle. Mais pourquoi cette convergence semble-t-elle orientée vers des tokens relevant d'une relative homogénéité sémantique isomorphe à des catégories humaines de pensée, attestée par l'augmentation des cosinus similarité, alors que l'on pourrait également imaginer qu'elle puisse être réalisée en direction de tokens dont l'unité relèverait d'alien concepts ? Nous évoquons ci-après le postulat explicatif suivant.

Ainsi que l'évoquent Thibault (1997) et Roads et al. (2024), en commentant des modélisations mathématisées de la catégorisation humaine, dont le « generalized context model » de Nosofsky (1986) et Nosofsky et al. (2022), la fonction de similarité peut être implémentée dans ces modèles en utilisant par exemple une distance pondérée de Minkowski afin de formaliser la notion d'attention sélective : l'attention spécifiquement accordée à une dimension utilisée pour le jugement de similarité est ainsi modélisée par le poids spécifiquement associé à cette dimension. Les auteurs indiquent que cette pondération sélective rend compte d'un processus de contraction ou de dilatation de l'espace cognitif apparié à cette dimension. Un poids élevé d'attention relatif à une dimension étire alors l'espace le long de cette dimension, ce qui a pour effet d'éloigner les *stimuli* (dans notre cas les tokens) dont la proximité catégorielle est évaluée sur cette dimension et permet ainsi une discrimination plus forte de ces *stimuli* sur cette dimension.

Par construction mathématique de la fonction d'agrégation, ce que nous venons d'indiquer est précisément le cas des tokens à (très) forte activation : leur activation, sur un neurone donné, ne peut être forte que dans la mesure où ils tendent à être co-activés sur leurs sous-dimensions catégorielles (issues de leur espace vectoriel d'entrée) générées par ceux de leurs neurones précurseurs qui ont les plus forts poids de connexion avec le neurone impliqué. Il s'en suit dès lors, pour ces tokens à forte activation, une plus grande capacité du neurone à les discriminer sur les sous-dimensions catégorielles impliquées.

Or cette capacité discriminante mentionnée nous semble être fondamentalement la fonction des catégories synthétiques neuronales, qui sont à ce titre des catégories ad hoc (Barsalou, 1995 ; Glaser et al., 2020 ; Bove et al., 2022) ; cela, étant donné qu'elles sont finalisées vers le but d'opérer des distinctions fines entre tokens (distinctions fines progressivement construites par des catégories synthétiques de plus en plus abstraites et subtiles) afin de parvenir à réaliser ce pour quoi le modèle de langage a été entraîné, à savoir prédire avec une très forte acuité différenciante le token suivant précisément pertinent (dans le cas de GPT). Mais cette précision prédictive ne pourra être adaptée que dans la mesure où elle est partiellement en phase avec la sémantique humaine, car c'est bien de générer des phrases ajustées à cette sémantique dont il s'agit. Il faut donc que la discrimination évoquée soit, en partie (mais pas en totalité car

l'efficience des catégories synthétiques provient aussi de leur dimension d'alien concept), convergente au vu de la sémantique humaine ; ce qui se traduira dès lors par des cosinus similarité plus importants et faisant montre d'une plus forte homogénéité catégorielle (selon le point de vue de l'univers sémantique de la cognition humaine) pour les tokens à (très) fortes activations.

Pour le dire autrement et de façon plus synthétisée, au niveau d'un neurone donné, sa discrimination accrue sur certaines sous-dimensions catégorielles (issues de des précurseurs neuronaux auxquels il est fortement relié) génère une meilleure segmentation et séparation des tokens sur les différentes sous-dimensions impliquées, ce qui rend alors possible des intersections catégorielles fines de ses sous-dimensions, abstractions affinées au service de la finalité de prédiction différenciante de prochains tokens adaptés aux modalités sémantiques humaines (auxquelles le modèle a été entrainée).

Terminons ce questionnement en indiquant que les tokens à (très) fortes activations peuvent être pensés comme étant ou tendant vers les prototypes, au sens de Rosh (1975) (mais aussi Singh et al., 2020 ; Vogel et al., 2021), de leurs neurones respectifs (le vrai prototype ultime étant, pour un neurone donné, le token fictif qui maximalise la fonction d'agrégation associée à ce neurone). Cela, dans la mesure où, à nouveau par construction mathématique de la fonction d'agrégation, les tokens fortement activés satisfont conjointement le plus les différentes sous-dimensions catégorielles (de l'espace vectoriel d'entrée) de leurs neurones associés. Interpréter les choses ainsi abonde dans le sens d'une relation forte entre catégorisation et similarité (en phase avec les supporters d'une indexation de la catégorisation sur la similarité en psychologie cognitive humaine) dans la mesure où nous observons que les tokens les plus fortement activés sont associés à une proximité de cosinus similarité accrue. Mais, plus fondamentalement, interpréter les core-tokens à très fortes activations comme les prototypes des catégories synthétiques dont ils relèvent nous permet de penser ces catégories synthétiques, en leurs prototypes, comme des états cognitifs limites, des points de transition de phase décrirait une analogie issue de la physique, à l'interface de deux registres de contraintes qu'il appartient à un réseau de neurones de concilier : (i) faire montre en sortie d'un respect des contraintes qui sont celles de la sémantique humaine (but), (ii) créer des segmentations catégorielles artificielles relevant d'aliens concepts distincts de la sémantique humaine mais efficaces pour parvenir à la finalité mentionnée (moyen). Les prototypes dénoteraient alors une sous-zone catégorielle interface à la limite (car si la similarité cosinus des tokens qui en relève est plus forte, les données montrent qu'elle demeure relative) de la clôture informationnelle (au sens de Varela) des aliens concepts que sont les concepts synthétiques des neurones formels.

8 Conclusion

Face à nos résultats empiriques majoritairement compatibles avec notre hypothèse de convergence catégorielle, postulant le fait que plus les niveaux d'activation des core-tokens successifs augmentent et plus la variabilité catégorielle de ces

core-tokens diminue, l'élément central explicatif que nous avons invoqué est le suivant : le segment catégoriel construit par un neurone d'une couche n (plus précisément par sa fonction d'agrégation entre autres) peut être décomposé en un espace vectoriel de sous-dimensions catégorielles. Ces sous-dimensions résultent d'une projection de l'espace vectoriel d'entrée de ce neurone, lequel est constitué, par construction mathématique de sa fonction d'agrégation, des dimensions catégorielles de sortie de chaque neurone de la couche précédente (n-1). En d'autres termes, pour comprendre un neurone, il faut le concevoir comme étant multidimensionnel, composé de sous-dimensions catégorielles. Les tokens à faible activation (mono-déclenchements catégoriels) relèvent de ces sous-dimensions de manière distincte, tandis que les tokens à forte activation (co-déclenchements catégoriels) les impliquent conjointement, ce que traduit notre observation de convergence catégorielle. Nous explorons actuellement ce postulat de sous-dimensions catégorielles dans le cadre d'une étude « génétique » visant à expliquer l'abstraction catégorielle produite par les neurones artificiels en termes de reconstruction des segmentations catégorielles de leurs neurones précurseurs les plus influents (c'est-à-dire ceux avec les connexions neuronales les plus fortes).

Remerciements

Michael Pichat remercie Christian Ganem (Chrysippe-R&D) pour ses précieux conseils de lectures innovantes en matière d'IA, Pierre Laniray (Université Paris Dauphine-PSL) pour son impulsion à étudier les problématiques d'explicabilité en matière d'IA, Stéphane Fadda (Université Sorbonne) pour ses conseils stimulants dans le registre de l'IA et Alexander Krainov (Yandex) pour les discussions challengeantes que nous avons pu avoir concernant la pertinence d'une étude de la psychologie de l'IA.

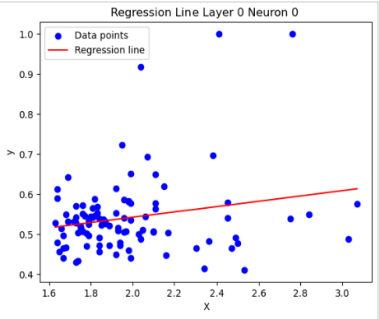
Contributions des auteurs

Michael Pichat a réalisé le design conceptuel et méthodologique de l'étude et en est le responsable scientifique. Enola Campoli a participé à divers aspects opérationnels de l'étude. William Pogrund a réalisé le formatage des données et leurs traitements statistiques. Michael Veillet-Guillem a géré la partie SysAdmin de l'étude. Jourdan Wilson a participé à des activités de prompt engineering, a réalisé la traduction anglaise et a formaté le texte publié. Anton Melkoezrov a participé à des activités de prompt engineering et au formatage des tableaux et schémas. Samuel Demarchi a conseillé les études statistiques. Armanouche Gasparian et Paloma Pichat ont rendu possible et étayé la réalisation de cette étude.

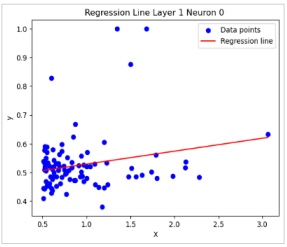
Annexes

Normality of Regression Residuals of Cosine Similarities Values of Pairs of Successive Core-Tokens as a Function of Activation of the First Token of Each Pair (Witness Neurons)

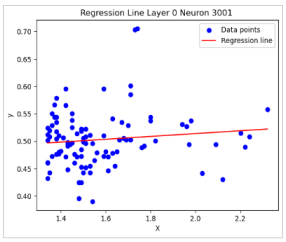
Layer 0 Neuron 0				
N (tokens) = 100	GPT2-XL	Alibaba	Mixedbread	WhereIsAI
Linear Relashionship Significance				
F	4,58	8,61	3,59	4,54
p(F)	0,03	0,00	0,06	0,04
Parameters of the Linear Relashionship				
a	0,09	0,09	0,06	0,07
b	0,22	0,39	0,44	0,41



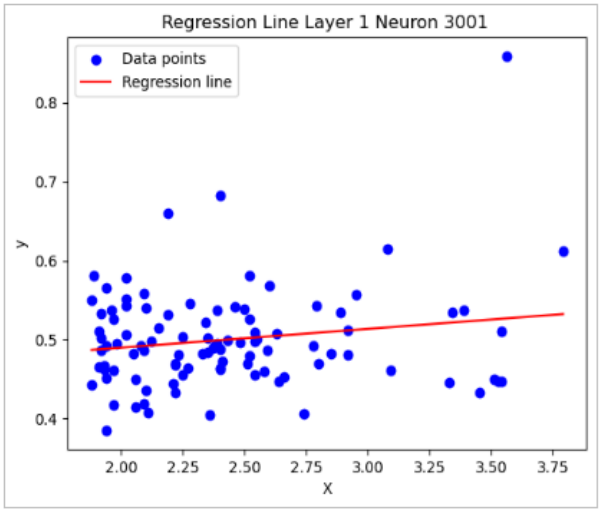
Layer 0 Neuron 0				
N (tokens) = 100	GPT2-XL	Alibaba	Mixedbread	WhereIsAI
Linear Relashionship Significance				
F	1,13	11,05	4,36	4,81
p(F)	0,29	0,00	0,04	0,03
Parameters of the Linear Relashionship				
a	0,04	0,06	0,04	0,05
b	0,37	0,53	0,50	0,48



Layer 0 Neuron 3001				
N (tokens) = 100	GPT2-XL	Alibaba	Mixedbread	WhereIsAI
Linear Relashionship Significance				
F	0,026	3,761	1,526	1,333
p(F)	0,873	0,055	0,220	0,251
Parameters of the Linear Relashionship				
a	0,01	0,05	0,03	0,03
b	0,28	0,47	0,47	0,46



Layer 1 Neuron 3001				
N (tokens) = 100	GPT2-XL	Alibaba	Mixedbread	WhereIsAI
Linear Relashionship Significance				
F	0,55	2,20	2,90	3,06
p(F)	0,46	0,14	0,09	0,08
Parameters of the Linear Relashionship				
a	0,02	0,02	0,02	0,02
b	0,30	0,49	0,46	0,44



Bibligraphie

[1] Antverg, O., & Belinkov, Y. (2021). On the Pitfalls of Analyzing Individual Neurons in Language Models. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2110.07483

[2] Ayeldeen, H., Hegazy, O., & Hassanien, A. E. (2015). Case Selection Strategy Based on K-Means Clustering. In *Advances in intelligent systems and computing* (p. 385-394). https://doi.org/10.1007/978-81-322-2250-7_39

[3] Barkan, O., Hauon, E., Caciularu, A., Katz, O., Malkiel, I., Armstrong, O., & Koenigstein, N. (2021). Grad-SAM. *CIKM '21 : Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. https://doi.org/10.1145/3459637.3482126

[4] Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition*, 11(3), 211-227. https://doi.org/10.3758/BF03196968

[5] 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

[6] 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

[7] Bobadilla-Suarez, S., Ahlheim, C., Mehrotra, A., Panos, A., & Love, B. C. (2020). Measures of Neural Similarity. *Computational Brain & Behavior*, 3(4), 369-383. https://doi.org/10.1007/s42113-019-00068-5

[8] 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

[9] Bruner, J. (1990). *Acts of meaning*. Harvard University Press.

[10] Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What Does BERT Look At ? An Analysis of BERT's Attention. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1906.04341

[11] 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

[12] 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)*.

[13] 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

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

[15] Durrani, U. K., Naymat, G. A., Ayoubi, R. M., Kamal, M. M., & Hussain, H. (2022). Gamified flipped classroom versus traditional classroom learning: Which approach is more efficient in business education? *The International Journal of Management Education*, 20(1), 100595. https://doi.org/10.1016/j.ijme.2021.100595

[16] 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

[17] Enguehard, J. (2023). Sequential Integrated Gradients: a simple but effective method for explaining language models. *Association For Computational Linguistics*. https://doi.org/10.18653/v1/2023.findings-acl.477

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

[19] Garde, A., Kran, E., & Barez, F. (2023). DeepDecipher: Accessing and Investigating Neuron Activation in Large Language Models. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2310.01870

[20] Geva, M., Bastings, J., Filippova, K., & Globerson, A. (2023). Dissecting Recall of Factual Associations in Auto-Regressive Language Models. *Association For Computational Linguistics*. https://doi.org/10.18653/v1/2023.emnlp-main.751

[21] 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

[22] Goodman, L. A. (1972). A Modified Multiple Regression Approach to the Analysis of Dichotomous Variables. *American Sociological Review*, 37(1), 28. https://doi.org/10.2307/2093491

[23] Ham, G., Kim, S., Lee, S., Lee, J., & Kim, D. (2023). Cosine Similarity Knowledge Distillation for Individual Class Information Transfer. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2311.14307

[24] Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, 4(11), 1173-1185. https://doi.org/10.1038/s41562-020-00951-3

[25] Jaunet, T., Kervadec, C., Vuillemot, R., Antipov, G., Baccouche, M., & Wolf, C. (2022). VisQA: X-raying Vision and Language Reasoning in Transformers. *IEEE Transactions On Visualization And Computer Graphics*, 28(1), 976-986. https://doi.org/10.1109/tvcg.2021.3114683

[26] Jawahar, G., Sagot, B., & Seddah, D. (2019b). What Does BERT Learn about the Structure of Language? *Proceedings Of The 57th Annual Meeting Of The Association For Computational Linguistics*. https://doi.org/10.18653/v1/p19-1356

[27] Kalyan, S. (2012). Similarity in linguistic categorization: The importance of necessary properties. *Cognitive Linguistics*, 23(3), 539-554. https://doi.org/10.1515/cog-2012-0016

[28] Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2022). Large Language Models Struggle to Learn Long-Tail Knowledge. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2211.08411

[29] Kaniuth, P., & Hebart, M. N. (2022). Feature-reweighted representational similarity analysis: A method for improving the fit between computational models, brains, and behavior. *NeuroImage*, 257, 119294. https://doi.org/10.1016/j.neuroimage.2022.119294

[30] Khamassi, M., Nahon, M., & Chatila, R. (2024). Strong and weak alignment of large language models with human values. *Scientific Reports*, 14(1). https://doi.org/10.1038/s41598-024-70031-3

[31] 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

[32] Kobayashi, G., Kuribayashi, T., Yokoi, S., & Inui, K. (2023). Transformer Language Models Handle Word Frequency in Prediction Head. *Findings Of The Association For Computational Linguistics: ACL 2023*. https://doi.org/10.18653/v1/2023.findings-acl.276

[33] 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

[34] Modarressi, A., Mohebbi, H., & Pilehvar, M. T. (2022). AdapLeR: Speeding up Inference by Adaptive Length Reduction. *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.1

[35] 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

[36] Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85(3), 207-238. https://doi.org/10.1037/0033-295X.85.3.207

[37] Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92(3), 289-316. https://doi.org/10.1037/0033-295x.92.3.289

[38] Mousi, B., Durrani, N., & Dalvi, F. (2023). Can LLMs facilitate interpretation of pre-trained language models? arXiv (Cornell University). https://doi.org/10.48550/arxiv.2305.13386

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

[40] Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In A. F. Healy, S. M.

[41] 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

[42] Olah, C., Cammarata, N., Voss, C., Schubert, L., & Goh, G. (2020). Naturally Occurring Equivariance in Neural Networks. *Distill*, 5(12). https://doi.org/10.23915/distill.00024.004

[43] 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

[44] Pichat, M., Campoli, E., Pogrund, W., Wilson, J., Veillet-Guillem, M., Melkozerov, A., Pichat, P., Gasparian, A., & Demarchi, S. (2024). Neuropsychology of AI: Relationship Between Activation Proximity and Categorical Proximity Within Neural Categories of Synthetic Cognition. arXiv. Available online: https://arxiv.org/abs/2410.11868

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

[46] 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

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

[48] Reppa, V., & Polycarpou, M. M. (2014). Adaptive Approximation for Multiple Sensor Fault Detection and Isolation of Nonlinear Uncertain Systems. *IEEE Transactions On Neural Networks And Learning Systems*, 25(1), 137-153. https://doi.org/10.1109/tnnls.2013.2250301

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

[50] Rosch, E. (1975). Cognitive representations of semantic categories. *Journal Of Experimental Psychology. General*, 104(3), 192-233. https://doi.org/10.1037/0096-3445.104.3.192

[51] Rosch, E., & Mervis, C. B. (1975). Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4), 573-605. https://doi.org/10.1016/0010-0285(75)90024-9

[52] Sajjad, H., Durrani, N., Dalvi, F., Alam, F., Khan, A. R., & Xu, J. (2022). Analyzing Encoded Concepts in Transformer Language Models. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2206.13289

[53] Savioz, A., Leuba, G., Vallet, P. G., & Walzer, C. (2010). Introduction aux réseaux neuronaux: De la synapse à la psyché. De Boeck Supérieur.

[54] Servan-Schreiber, E., & Anderson, J. R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(4), 592–608. https://doi.org/10.1037/0278-7393.16.4.592

[55] 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

[56] Sternberg, R. J. (2007). Manuel de psychologie cognitive: Du laboratoire à la vie quotidienne. De Boeck Supérieur.

[57] 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

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

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

[60] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1706.03762

[61] 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

[62] Voita, E., Sennrich, R., & Titov, I. (2021b). Language Modeling, Lexical Translation, Reordering: The Training Process of NMT through the Lens of Classical SMT. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2109.01396

[63] Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., & Dai, J. (2023). VisionLLM: Large Language Model is also an Open-Ended Decoder for Vision-Centric Tasks. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2305.11175

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

[65] Wixted, J. T., Vul, E., Mickes, L., & Wilson, B. M. (2018). Models of lineup memory. *Cognitive Psychology*, 105, 81–114. https://doi.org/10.1016/j.cogpsych.2018.06.001

[66] Wu, N., Kenway, G., Mader, C., Jasa, J., & Martins, J. (2020). pyOptSparse: A Python framework for large-scale constrained nonlinear optimization of sparse systems. *The Journal Of Open Source Software*, 5(54), 2564. https://doi.org/10.21105/joss.02564

[67] Yang, S., Saïd, M., Peyre, H., Ramus, F., Taine, M., Law, E. C., Dufourg, M., Heude, B., Charles, M., & Bernard, J. Y. (2023). Associations of screen use with cognitive development in early childhood: the ELFE birth cohort. *Journal Of Child Psychology And Psychiatry*, 65(5), 680–693. https://doi.org/10.1111/jcpp.13887

[68] Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., Ermon, S., & Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications*, 11(1). https://doi.org/10.1038/s41467-020-16185-w

[69] 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). https://doi.org/10.48550/arxiv.2309.01029

[70] 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