NEUROPSYCHOLOGIE DE L'IA : RAPPORT ENTRE PROXIMITE D'ACTIVATION ET PROXIMITE CATEGORIELLE AU SEIN DES CATEGORIES NEURONALES DE LA COGNITION SYNTHETIQUE

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 8 Octobre 2024

La neuropsychologie de l'intelligence artificielle s'intéresse à la cognition neuronale synthétique prise comme un nouveau type d'objet d'étude de la psychologie cognitive. Dans une visée d'explicabilité des réseaux de neurones artificiels constitutifs des modèles de langage, il s'agit de transposer des concepts de la psychologie cognitive au domaine de la construction interprétative de la cognition neuronale artificielle, en s'interrogeant dans une perspective épistémologique, sur les conditions et enjeux constructivistes d'une telle transposition. Le concept cognitif humain ici mobilité est celui de la catégorisation comme heuristique à la pensée du processus de segmentation et de construction du réel opéré par les vecteurs neuronaux de la cognition de synthèse.

1 Introduction

L'explicabilité vise à rendre l'activité d'un réseau de neurones artificiels compréhensible pour les humains (Du et al., 2019 ; Pichat, 2023, 2024a). Cela implique de traduire le comportement observable d'un réseau de neurones dans un cadre interprétatif, permettant d'attribuer une signification pertinente à ce comportement en fonction des objectifs de l'observateur. Dans notre contexte, ce cadre est celui de la psychologie cognitive. Il s'agit donc d'utiliser les catégories de pensée de la cognition humaine comme référents conceptuels pour établir des analogies entre les comportements cognitifs humains et artificiels. Plus spécifiquement, nous nous centrerons sur la notion de catégorisation dans la mesure où ce concept de la psychologie cognitive semble particulièrement pertinent pour analyser la cognition synthétique des modèles de langage qui relève *de facto* fortement d'une dynamique d'extraction d'invariants catégoriels linguistiques (Jawahar et al., 2019 ; Clark et al., 2019 ; Bills et al. 2023 ; Clark et al, 2023).

Dans ce travail, nous nous concentrons sur une explicabilité épistémologique à faible granularité cognitive (Pichat, 2024b). Autrement dit, nous examinons une explicabilité microscopique où l'unité d'observation est le neurone formel. Cette approche explicative à faible granularité vise à pénétrer directement le système "boîte noire" que représente 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).

2 La catégorisation humaine

2.1 Le Processus cognitif de la catégorisation humaine

La catégorisation joue un rôle central au sein d'une variété d'activités cognitives humaines de différentes ampleurs (Sternberg, 2007 ; Roads et al., 2024) : clas-

sification et sériation, identification et dénotation des objets, compréhension, raisonnement, résolution de problèmes, mémorisation, inférence et prédiction, transfert de propriétés, conceptualisation, etc.

D'un point de vue formel, une catégorie se définit par deux types d'éléments: sa compréhension et son extension (Nadeau, 1999). La compréhension d'une catégorie, également nommée intension, est l'ensemble des propriétés qui définissent de façon nécessaire et suffisante cette catégorie ; que ces propriétés soient physiques, structurelles, fonctionnelles, procédurales ou finalisées (c'est-à-dire liées au but de la tâche impliquée) (Tijus, 2004). Son extension est l'ensemble des membres de cette catégorie.

Historiquement, l'âge classique de la notion de catégorie est porté par Platon puis Aristote qui positionnent une catégorie comme étant définie par une série de propriétés nécessaires et suffisantes. Cet angle de vue a par exemple donné naissance aux catégories fondées sur les traits, conçues comme le fruit d'une ventilation d'une catégorie en une série de caractéritiques toutes nécessaires et collectivement suffisantes pour définir cette catégorie (Katz, 1972). Mais l'étude cognitive des processus humains effectifs de catégorisation a rapidement montré le caractère rigide de cette conception *princeps* des règles ou éléments théoriques définitoires de la catégorisation.

Rosch (1975) développe ainsi une approche de la catégorisation par attribution d'un niveau de ressemblance d'un objet candidat au prototype de la classe impliquée. Cela, sur la base de la constatation empirique que des individus, interrogés quant à ce qui définit une catégorie, ont plus tendance à en énoncer des traits de caractéristiques au détriment de propriétés déterminatives (Rosch et Mervis, 1975). Le prototype est alors défini comme l'exemple, réel ou extrapolé par construction mentale (tel le moyennage), le plus représentatif, le plus typique de la catégorie impliquée quant à ses traits caractéristiques (Singh et al., 2020 ; Vogel et al., 2021). Ce prototype pouvant être l'objet d'une forte variabilité pour différentes sous-catégories d'une même catégorie donnée (Malt et Smith, 1984). Dans l'approche du prototype, les traits de caractéristiques sont régulièrement présents dans les items constitutifs de la classe mais ce n'est pas toujours le cas. En effet, plus souple que la conception antérieure basée sur des définitions, cette préhension de la catégorisation permet de répondre à l'objection de Wittgenstein (1953) concernant « l'air de famille » : si le prototype se caractérise par les propriétés (1, 2, 3, 4), un élément A possédant les attributs (1, 2, 5, 6) tout comme un élément B possédant les attributs (7, 8, 3, 4) pourront tous deux être assignés à cette même catégorie quand bien même ils ne possèdent pas d'attributs communs. Approche du prototype que l'on retrouve également dans le champ des théorisations *princeps* de la perception (Posner et al, 1967 ; Bransford, 1971 ; Reed, 1972).

Connexe mais différente de l'approche du prototype, la théorie de la catégorisation par l'exemplaire (Medin et Schaffer, 1978 ; Nosofsky, 1992 ; Nosofsky et al., 2022) suggère que les objets sont catégorisés par comparaison à des exemples typiques de la catégorie, exemples stockés en mémoire. L'exemplaire le plus typique étant ici celui qui ressemble le plus à tous les exemplaires connus par l'individu. Cet exemplaire est celui qui exerce le plus fort pouvoir

d'attraction de par sa fréquence d'existence parmi les exemplaires conservés au niveau mnésique.

Notons rapidement, même si nous ne l'exploiterons pas ultérieurement, la définition de la catégorisation par réseaux sémantiques (Collins et Quillian, 1969; Hornsby et al., 2020). Dans le cadre de cette conception, les catégories sont structurées au sein d'un réseau de nœuds (les concepts) et de liens entre ces nœuds (les relations entre concepts) ; cela, des catégories les plus spécifiques à celles qui possèdent le plus important niveau de généralité.

Mentionnons enfin l'approche de la catégorisation contextuelle, circonstancielle ou encore finalisée par le but de l'action et de la tâche, à l'instar des catégories *ad hoc* (Barsalou, 1983 ; Glaser et al., 2020). Dans le registre de ces approches fonctionnelles (Rips 1989 ; Keil, 1989 ; Wisniewski et Medin, 1994; Barsalou, 1995 ; Bove et al., 2022), la finalité devient le centre de l'activité de définition d'une catégorie, au détriment d'une logique ou d'une sémantique générale, ou encore de l'apparence. C'est ici la situation, par sa fin et ses éléments de contexte propres, qui guide la catégorisation et les termes de cette catégorisation n'ont pas d'existence en dehors de cette situation *hic et nunc*.

2.2 La Catégorisation humaine par jugement de similarité

Les approches de la catégorisation par similarité postulent qu'un objet est assimilé à une classe par estimation de la proximité de celui-ci avec ce qui représente cette classe (Thibault, 1997 ; Jacob et al., 2021 ; Kaniuth et al., 2022 ; Roads et al., 2021, 2024) ; cela, sur la base (i) d'un espace de traits ou de dimensions retenus comme pertinents pour effectuer la comparaison, (ii) d'une modalité de calcul de distance entre les instances comparées.

L'implication de l'évaluation de la similarité comme base de la catégorisation semble à spectre large (Thibault, 1997), en tout cas pour les classes n'impliquant pas de définition explicite et présentant une organisation hiérarchisée rendant possible le fait que certains items appartiennent clairement à une catégorie (Hampton, 1997).

Les théories de la catégorisation par le prototype (Posner et Keele, 1968; Reed, 1972 ; Rosch & Mervis, 1975 ; Medin et Schaffer, 1978) mentionnées précédemment mettent de facto en avant la fonction de la similarité dans le processus de catégorisation (Sanborn et al., 2021) : est attribué à une catégorie ce qui est jugé proche de la représentation centrale qu'est le prototype. Il en va de même concernant l'approche de la catégorisation par l'exemplaire (Medin et Schaffer, 1978 ; Brooks, 1987 ; Nosofsky, 1992) : est assigné à une classe catégorielle ce qui est estimé le plus à proximité des éléments significatifs composant cette classe. Dans les deux cas, la catégorisation est la résultante de la distance estimée entre l'item impliqué et ce qui représente la catégorie (Ayeldeen et al., 2015 ; Roads et al., 2024).

2.3 Les arguments opposés à la possibilité d'une catégorisation humaine par similarité

Les arguments s'opposant à une fondation effective ou possible de la catégorisation sur un raisonnement de similarité sont divers (Love, 2002) :

• Un élément est assigné à la catégorie qui présente la meilleure capacité explicative de cet élement (Murphy et Medin, 1985), par-delà l'éventuelle classification initiale par similarité (Keil, 1989).

• Les catégories qui sont l'objet d'une définition explicite ne peuvent être directement fondées sur un jugement de similarité (Kalyan et al., 2012).

Mais les principaux arguments reposent sur l'idée que le choix singulier des critères du jugement de similarité, critères qui ne sont qu'un possible parmi d'autres dans l'espace des traits ou dimensions, n'est pas nécessairement en adéquation avec ce qui fonde ou devrait fonder l'attribution catégorielle (Reppa et al., 2013 ; Poth, 2023) :

• La catégorisation est impactée par des informations externes aux objets à classer : des théories générales sur le monde ou des éléments spécifiquement liés à la catégorie impliquée (Rips, 1989).

• L'assignation d'un objet à une catégorie est également fonction des relations entre les autres objets qui constituent cette catégorie (Medin et al, 1993).

• Là où la catégorisation est fonction de la finalité de tâches, notamment dans le cas des catégories ad hoc (Barsalou, 1991), la similarité va mobiliser des critères de comparaison qui ne seront pas adaptés à cette activité finalisée.

Ces arguments convergent vers l'idée que le raisonnement par similarité est trop équivoque pour pouvoir fonder de façon fonctionnelle l'attribution catégorielle (Wixted, 2018). Rips (1989) évoque dans ce registre une relation non-monotone entre similarité et appartenance catégorielle. Le critère utilisé par un jugement de similarité varie en effet en fonction des contextes (Murphy et Medin, 1985), par exemple culturels Whorf (1941). Autrement dit, les critères mobilisés par le jugement de similarité ne sont pas assez contraints et donc trop fonction du choix singulier de segmentation opéré *hic et nunc* (Goodman, 1972).

Dès lors, il ressort que les jugements de similarité et d'appartenance catégorielle ne sont pas en phase (Rips, 1989 ; Medin, 1993) et que la similarité ne peut pas, ou ne devrait pas, impliquer la catégorisation.

2.4 Les contre-argumentations en faveur d'une fondation de la catégorisation humaine sur la similarité

Face aux arguments s'opposant à l'implication ou à la pertinence de la similarité dans le processus de catégorisation, diverses réponses sont apportées (Bobadilla et al., 2020 ; Hebart et al., 2020).

Goldstone (1994) propose la quadruple contre-argumentation suivante : (i) l'invocation de la trop forte labilité de la similarité pour fonder la catégorisation ne tient pas dans la mesure où elle présuppose que la catégorisation elle-même ne serait pas également flexible ; (ii) même superficielle dans certains cas, la similarité est fonctionnelle dans la mesure où elle permettra génétiquement la

découverte d'indicateurs plus « profonds » de catégorisation et dès lors la création de nouvelles catégories plus « fondamentales » ; (iii) l'expérimentation montre que la similarité n'est pas si instable qu'il est régulièrement argué ; (iv) les catégories non fondées par la similarité se montrent réfractaires à la généralisation.

Thibault (1997), quant à lui, face aux critiques de la relativité subjective de la similarité, postule que la catégorisation est en fait un sous-type de similarité. L'auteur reconnaît en effet que si la similarité est bien fonction de la contingence du choix des critères comparés, la catégorisation procède de même, à ceci près que cette dernière choisit ses critères propres parmi un ensemble de traits définissant la catégorie en cause. Dénonçant un essentialisme psychologique, Thibault (idem) affirme également que l'argument de la faiblesse de la similarité (face aux éléments de contexte par exemple) ne tient pas car cette position postule que les critères de segmentation catégorielle ont, pourraient avoir ou devraient avoir une valeur *per se*, intrinsèque, ontologique, indépendante de l'individu.

Enfin, Hampton (1997) invoque le fait que la catégorisation elle-même peut être impactée par des éléments de similarité non pertinents (souvent perceptifs et notamment visuels), ou en tout cas décrétés comme tels par une analyse logique *a priori*affirmant de façon péremptoire que la pensée devrait fonctionner en phase avec les canons de ce qui est instancié comme étant la logique. Plus encore, l'auteur montre, en se basant sur une approche de logique floue, que si les catégories ont parfois du mal à être définies par les sujets, ces derniers n'ont pour autant pas de difficulté à indiquer dans quelle mesure deux items de ces catégories diffèrent ou quels en sont les membres typiques, c'est-à-dire à mobiliser des activités cognitives de similarité à leur endroit. Dans les deux cas, à nouveau, Hampton met en avant le fait que les faiblesses cognitives attribuées à la similarité sont fondées sur un postulat erroné d'assujettissement de la catégorisation à ce qui est instancié de façon normative et *a priori* comme étant La logique (classique).

Il ressort un invariant de ces contre-arguments : affirmer les limites de la similarité pour fonder la catégorisation relève d'une double faille épistémologique : (i) faille réaliste consistant à artificiellement décréter la catégorisation comme étant un processus se saisissant ou devant se saisir d'un réel ontologiquement pré-défini, (ii) faille rationaliste assignant à la catégorisation un (devoir d') assujettissement à une logique invoquée comme étant une évidence, évidence dont tout individu devrait de surcroit être en mesure de se saisir.

3 Problématique

3.1 Contexte

De nombreuses études révèlent ou déduisent une diversité de catégories (linguistiques, logiques, positionnelles, etc.) codées dans les neurones et les têtes d'attention. Dans l'expérimentation classique de Clark et al. (2019) sur BERT, les auteurs mettent en évidence les fonctions linguistiques convergentes des têtes

d'attention provenant des mêmes couches. Dans leur étude fascinante sur GPT2-XL, Bills et al. (2023) identifient une série de neurones spécifiques, soulignant pour certains leur forte sensibilité au contexte. Les recherches montrent également une distribution géographique du type d'activité neuronale en fonction de la profondeur des couches. Ainsi, les premières couches réagissent davantage à des catégories morphologiques au niveau des mots, tandis que les couches plus profondes sont plus sensibles aux caractéristiques catégorielles syntaxiques des phrases (voix passive/active, temps) et aux catégories sémantiques (Jawahar et al., 2019).

Dans le cadre de notre présent travail, nous repartons de certains aspects de l'étude menée par Bills et al. (2023), en les réorientant vers d'autres sujets d'interrogation plus cognitifs et épistémologiques (Pichat, 2024), ainsi que nous allons le préciser dans la rubrique suivante. Avant cela, expliquons rapidement la démarche de Bills et al. (idem). Sur la base de l'hypothèse qu'un neurone s'active spécifiquement pour une propriété donnée, les chercheurs ont entrepris une vaste analyse de la sémantique catégorielle de l'ensemble des neurones de GPT-2XL. Sur le plan méthodologique, ils ont soumis GPT-2XL à une série étendue de séquences de tokens, sélectionnées aléatoirement parmi les données internet utilisées pour l'entraînement du modèle. Pour chaque token, les valeurs d'activation de tous les neurones à travers toutes les couches ont été enregistrées. GPT-4 a ensuite été utilisé pour identifier automatiquement les éléments auxquels chaque neurone réagit (c'est-à-dire pour générer «l'explication» catégorielle), en se basant sur un prompt d'instruction et d'exemple appliqué uniquement aux cinq séquences avec les activations maximales.

3.2 De la catégorisation humaine à la catégorisation synthétique

En phase avec le questionnement que nous avons présenté, dans le champ de la cognition humaine, concernant la relation entre catégorisation et similarité, notre transposition de cette question heuristique dans le champ de la cognition synthétique est : le niveau d'appartenance catégorielle des tokens (parvenant à un neurone donné, sous forme d'embedding) à la catégorie associée à ce neurone est-il lié à leur niveau de similarité ? Autrement dit, l'intensité d'appartenance catégorielle et l'intensité de similarité de tokens, analysés par un neurone donné, sont-il deux aspects liés du même phénomène ? Autrement dit encore, l'espace neuronal de l'appartenance catégorielle est-il segmenté en fonction de la segmentation de l'espace de similarité ? Problématique largement inexplorée à ce jour dans le champ de l'explicabilité synthétique (Fan et al., 2023 ; Luo et al., 2024 ; Zhao et al., 2024) et qu'il nous semble dès lors particulièrement pertinente d'investiguer.

Notons d'un point de vue épistémologique que nous avons transposé la notion d'appartenance catégorielle (mesurée sur une échelle nominale dichotomique oui / non), dans le registre de la cognition humaine, en la notion de niveau d'appartenance catégorielle (mesurée donc sur une échelle ordonnée), dans le registre de la cognition synthétique ; cela, dans la mesure où dans le champ

neuronal artificiel, l'appartenance catégorielle (i.e. l'activation, ainsi que nous allons le développer plus tard) est une valeur numérique et non pas booléenne.

4 Méthodologie

4.1 Le choix de GPT-2XL

Dans le cadre de notre étude, nous nous sommes intéressés à GPT d'open AI car cette suite de modèles, inaugurale d'une partie significative de l'IA générative contemporaine, présente le paradoxe d'être à la fois la plus populaire en termes de médiatisation et de nombre d'utilisateurs et en même temps la moins étudiée directement en interne, c'est-à-dire en termes d'explicabilité à faible granularité. Le modèle retenu est GPT-2XL. En effet, ce modèle présente l'intérêt d'être suffisamment élaboré pour pouvoir étudier les phénomènes cognitifs synthétiques de haut niveau qui nous intéressent sans pour autant atteindre la complexité de GPT-4, et *a fortiori* du multimodal GPT-4o, complexité par laquelle il ne nous semble pas pertinent de commencer dans le cadre du premier questionnement cognitif qui est le nôtre ; autrement dit, GPT-2XL nous apparaît présenter un bon niveau de compromis. Par-delà cette raison épistémique, une raison pragmatique préside également à notre choix de GPT-2XL : pour la première fois, en 2023, Open AI rompt sa tradition « black box » (certainement logique d'un point de vue commercial) concernant ces produits, en livrant dans le cadre de l'article de Bills et al. (2023) les informations relatives aux paramètres et aux valeurs d'activation des neurones constitutifs de GPT-2XL, paramètres et valeurs d'activation qui vont dès lors pouvoir constituer nos données de départ dans le cadre de cette étude.

A toutes fins utiles, précisons que GPT-2 XL est la variante à spectre large de GPT (Generative Pre-trained Transformer) 2, développé par OpenAI et sorti durant l'année 2019. Comme indiqué par son nom, GPT est un transformer, combinant des couches de têtes attentionnelles et des couches de type perceptron multicouche à propagation avant. Sa fonction d'activation est GeLU. Fruit d'un entrainement non supervisé (en tout cas directement) sur un dataset de 8 millions de pages internet, le modèle possède environ 1,5 milliard de paramètres répartis sur 48 couches. Ces dernières sont constituées de 6400 neurones chacune et opèrent sur des embeddings à 1600 dimensions ; chaque couche (ou bloc de transformer) est composée d'une sous-couche attentionnelle de 25 têtes attentionnelles et de deux sous-couches de type réseau à propagation avant. La finalité de l'entrainement du modèle est la complétion et la génération de texte, ce qui le rend capable, dans le registre de performance qui est le sien, d'une variété de tâches.

4.2 Nos choix spécifiques concernant les données

A des fins de simplification, nous nous sommes cantonnés dans le cadre de cette étude exploratoire aux deux premières couches de GPT-2XL (layer 0 et 1) et

aux 6400 neurones de chacune de ces couches.

Concernant les tokens et leurs valeurs d'activation au sein de ces 2 x 6400 = 12800 neurones formels, nous avons fait le choix, pour chacun de ces neurones, de considérer comme données pertinentes ses 100 tokens les plus activés en moyenne avec leurs valeurs d'activation respectives. En effet, la sélection des seuls tokens à hyperactivation, telle qu'opérée par Bills et al. (2023), nous semble trop restrictive dans la mesure où elle n'est pas représentative de la variabilité des tokens pour lesquels un neurone s'active, nous donnant ainsi potentiellement une vue trop restrictive de la catégorie de tokens à laquelle un neurone donné réagit ; autrement dit, nous pensons que Bills et al. (idem) n'interprètent pas *in extenso* les neurones mais identifient une sous-catégorie très limitée de la catégorie encodée par chacun de ces neurones. Un autre argument ayant généré notre choix est que les tokens à fortes valeurs moyennes d'activation retenus sont *de facto* moins sensibles aux effets de contextes qui, bien que cruciaux (et cela constitue en un sens une limite de notre approche), peuvent eux-mêmes également limiter l'extension des tokens appartenant relativement à une catégorie neuronale et donc la sémantique catégorielle des neurones impliqués.

4.3 La construction interprétative de nos observables

Comme nous l'avons mentionné, le niveau d'activation moyen dont un token est l'objet dans un neurone nous apparaît comme une bonne opérationnalisation de l'équivalent, dans la cognition synthétique, du niveau d'appartenance catégorielle dans la cognition humaine. En effet, l'activation moyenne des 100 tokens les plus activés nous semble être de facto bien représentative de la mesure dans laquelle un token fait partie de l'extension d'une catégorie. Cela, dans le champ de la cognition synthétique, en phase avec l'hypothèse de Bills et al. (2023) qu'un neurone s'active spécifiquement pour une propriété et les fondements des études d'explicabilité à faible granularité. Illustrons cette conception avec un cas pointé par Bills et al. (idem) : concernant les neurones s'activant après une occurrence répétée de tokens, plus la répétition est forte (ie plus la séquence de tokens impliquée satisfait la condition définitoire de cette catégorie, à savoir la répétition de tokens) et plus l'activation est forte. Cette transposition niveau d'appartenance catégorielle / niveau d'activation nous semble également justifiée, dans le champ neurobiologique de la cognition humaine cette fois-ci, par le fait que la fonction d'activation est le corollaire analogique de la fonction de transfert (Savioz et al., 2010) dont la finalité est précisément de clarifier les inputs appartenant à la catégorie à laquelle un neurone biologique doit réagir, en augmentant le contraste signal/bruit, forme/fond (Servan-Schreiber, 1990), c'est-à-dire le contraste entre ce qui appartient à la catégorie d'éléments pour lesquels le neurone doit s'activer par apposition aux éléments résiduels.

Le cosinus de similarité entre deux tokens nous apparaît de même être une bonne mesure de la notion de similarité entre deux items transposée du registre de la cognition humaine à celui de la cognition artificielle. En effet, dans le domaine de la pensée humaine, nous avons vu que (Thibault, 1997) définit la similarité sur la base (i) d'un espace de traits ou de dimensions retenus comme

pertinents pour effectuer la comparaison, (ii) d'une modalité de calcul de distance entre les instances comparées. En adéquation avec cette définition, le cosinus de similarité, est une mesure couramment utilisée en NLP pour mesurer la proximité sémantique entre deux éléments (Ham, 2023) ; cela sur la base du produit scalaire entre les deux vecteurs multidimensionnels impliqués, consistant à mesurer la distance, dimension par dimension de l'espace vectoriel sémantique en jeu, entre les deux items à comparer ; la similarité cosinus étant normée de -1 à 1, où -1 indique des vecteurs opposés (similarité opposée), 0 des vecteurs orthogonaux (pas de similarité) et +1 des vecteurs identiques (similarité totale). Notons que nous avons fait le choix central de mesurer le cosinus de similarité au sein de la base d'embeddings de GPT2-XL, et non pas dans la base par exemple de GPT-4 pourtant plus performante, afin de ne pas retomber dans la limite méthodologique mentionnée par Bills et al. (2023) et Bricken (2023) consistant à apparier des systèmes cognitifs synthétiques ne reposant pas sur le même système d'embeddings, c'est-à-dire pas sur le même système de segmentation catégorielle ; même si, à des fins de comparaison et de vérification de la plausibilité de nos données, nous avons également eu systématiquement recours à trois autres bases classiques d'embeddings librement disponibles : Alibaba-NLP/gte-large-en-v1.5, Mixedbread-ai/mxbai-embed-large-v1 et WhereIsAI/UAE-Large-V1.

4.4 Précisions statistiques

Nos calculs statistiques descriptifs comme inférentiels ont été réalisés via des librairies Python issues de la suite SciPy, sur la base de la consultation de Howell (2008) et Beaufils (1996).

L'étude préalable de la normalité des données, à des fins exploratoires mais également de vérification des conditions de réalisation de certains tests paramétriques (ANOVA, regression, test de Grubbs), a été double. Premièrement via différents tests inférentiels ayant leurs avantages respectifs : tests de Shapiro-Wilk (valable pour de petits échantillons), Lilliefors (valable pour les petits échantillons et les cas où paramètres de la distribution normale sont inconnus et estimés à partir des données), Kolmogorov-Smirnov (plus adapté pour les grands échantillons) et Jarque-Bera (centré sur la symétrie et l'aplatissement, valable pour de grands échantillons) ; notons que, concernant les mesures de similarité cosinus, ces tests ont systématiquement été réalisés sur l'ensemble des quatre embeddings mentionnés, à des fins de vérification de la convergence des résultats. Deuxièmement via une approche descriptive numérique (indices skweness de symétrie et kurtosis d'aplatissement, écart moyenne-médiane) et graphique (QQ-plot de comparaison distribution effective / distribution théorique normale). Cette variabilité d'approches nous dotant d'une vue à spectre large.

Deux types d'unités statistiques ont été dégagés. Concernant nos investigations « micro », neurone par neurone, les unités statistiques instanciées sont des tokens, les 100 tokens les plus activés en moyenne pour chaque neurone, tokens que nous nommons « core-tokens ». Sur ces unités statistiques de premier ont été réalisés les tests paramétriques suivants : le test de Fisher pour la régression

linaire d'une distribution à résidus normaux (comparant variance expliquée par le modèle de régression et variance non-expliquée), le test de Grubbs permettant d'identifier les outliers d'une distribution normale, le t de Student univarié pour comparaison d'une moyenne à une norme sur une variable à distribution normale (avec correction pour petits effectifs) ; et les tests non paramétriques suivants : ρ de spearman (pour les corrélations sur échelle ordinale), Wilcoxon-Mann-Whitney (pour la comparaison de moyennes de groupes indépendants sur échelle ordinale). Relativement aux mesures de similarité cosinus, l'ensemble de ces tests paramétriques et non paramétriques a été réalisé sur les quatre embeddings indiqués, toujours à des fins de vérification de la convergence de nos résultats.

Concernant les investigations « macro », sources d'inférence globale sur l'ensemble des neurones d'une couche donnée, les unités statistiques mobilisées sont des neurones, les 6400 neurones de chaque couche. Sur ces unités statistiques de second ordre un seul test, non paramétrique, a été réalisé (majoritairement sur l'embedding GPT-2XL) : le χ^2, univarié, d'ajustement, nous permettant d'inférer la significativité des phénomènes « micro » sur l'ensemble de l'échelle « macro » des neurones d'une couche donnée. Notons que si notre étude avait été plus large, en portant sur l'ensemble des 48 couches de GPT-2XL, un troisième type pertinent d'unités statistique aurait été dégagé, celui des couches, permettant une généralisation des phénomènes pointés à l'ensemble du modèle.

4.5 La question investiguée et son opérationnalisation

En opérant le choix, mentionné précédemment, d'opérationnaliser l'appartenance catégorielle par l'activation et la similarité par la mesure du cosinus de similarité, notre question, opérationnalisée, devient : le niveau d'activation des tokens est-il lié à leur niveau de similarité cosinus ? Exprimée de façon fonctionnelle (au sens mathématique du terme), cette question se traduit comme suit : existe-t-il une relation entre l'activation (dans l'espace d'activation) et la distance de similarité cosinus (dans l'espace des similarité cosinus) ?

Afin d'opérationnaliser plus avant cette question, nous faisons le choix de l'étudier sous un angle de proximité d'intensité d'activation entre tokens. Cela, dans la mesure où l'on pourrait inférer que si le niveau de similarité (ie de proximité catégorielle) était lié au niveau d'appartenance catégorielle, alors cela signifierait vraisemblablement une relation entre proximité de l'intensité de similarité catégorielle et proximité de l'intensité de l'appartenance catégorielle. La question devient alors opérationnalisée comme suit : existe-t-il une relation entre proximité d'activation et similarité (proximité) cosinus entre les tokens ?

En termes d'unités statistiques, comme indiqué déjà, nous faisons le choix de nous focaliser, pour chaque neurone, sur les 100 tokens qui sont l'objet de ses 100 activations moyennes les plus fortes. Cela, en raison du fait qu'il ne nous semble pas pertinent de nous intéresser à des tokens n'appartenant pas, ou de façon trop labile, à la catégorie associée à chaque neurone (i.e. peu activés) ; en effet, il nous apparait peu probable, hormis hasard statistique,

de trouver de tels tokens qui seraient l'objet d'un lien systématique en termes de similarité. Dès lors, notre questionnement, tel qu'instancié relativement à la proximité d'activation, nous guide d'un point de vue méthodologique vers le choix final suivant d'unités statistiques premières : la suite des paires de core-token successifs quant à leur niveau d'activation.

In fine, notre questionnement devient le suivant : existe-t-il, au niveau des paires de core-tokens successifs de chaque neurone, une relation entre proximité d'activation et similarité (proximité) cosinus ?

5 Résultats

5.1 Explorations statistiques préliminaires à fins méthodologiques

Pour les deux couches (cf tableaux n°1 & n°2 et graphes analogues n°1 & n°2), la comparaison des valeurs moyennes min, mean, range, Q1 et CV (coefficient de variation) pour le cosinus similarité entre paires de core-tokens successifs (quant à leur niveau d'activation) semble pointer une carence relative des trois modèles d'embeddings Alibaba, Mixedbred et WhereIsAI par rapport à GPT-2XL : un pouvoir discriminant plus important de ce dernier. Phénomène qui pourrait éventuellement être expliqué partiellement par le biais méthodologique suivant, justement voulu comme indiqué préalablement dans notre rubrique méthodologique : les core-tokens ici impliqués sont *de facto* conformes au système de tokenisation de GPT-2XL et pas nécessairement en phase avec les modalités de segmentation qui ont présidé aux trois autres bases d'embedding. Il en résulte, dans tous les cas, que les valeurs de cosinus issues des embeddings GPT-2XL seront estimées plus fiables dans la suite de cette étude, ce qui n'implique pas pour autant de rejeter les indications en provenance des autres modèles d'embedding (i) à des fins de vérification de la convergence inter-embeddings de nos résultats, (ii) et cela d'autant plus que les présents résultats sont très convergents entre ces trois systèmes d'embedding ce qui serait malgré tout de nature à étayer les concernant une certaine fiabilité.

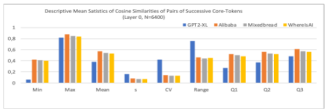
Pour les deux couches (cf tableaux n°3 & n°4), nous constatons que les indicateurs inférentiels (Shapiro-Wilk, Lilliefors, Kolmogorov-Smirnov et Jarque-Bera ; pour α=.05) comme descriptifs (comparaison moyenne/mode, symétrie et aplatissement) relatifs au cosinus similarité entre paires de core-tokens successifs sont compatibles avec (ce qui ne signifie pas pour autant qu'ils prouvent)

	GPT2	Alibaba	Mixedbred	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

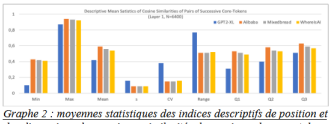
Tableau 1 : moyennes statistiques des indices descriptifs de position et de dispersion des cosinus similarité des paires de core-tokens successifs (Couche 0, n=6400).

	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).



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



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

une hypothèse de normalité dans 2/3 des cas pour une mesure effectuée à partir des embeddings de GPT-2XL. Mais que ces indicateurs chutent de façon assez convergente pour les mesures réalisées sur la base des trois autres systèmes d'embedding (cf. annexes pour les résultats relatifs aux « neurones témoins », dont les graphes QQ-plots / droite de Henry basés sur les embeddings de GPT-2XL uniquement). Cette divergence peut peut-être à nouveau être partiellement expliquée par la variabilité des systèmes de tokenisation. Dans les deux cas, la normalité semble moindre pour la couche 1 comparativement à la couche 0. Mais, quoi qu'il en soit, ces résultats semblent indiquer une plus grande pertinence, dans nos dispositifs statistiques à venir, de l'utilisation de tests non paramétriques (i.e. ne supposant pas une normalité de la distribution de la variable impliquée), ou en tout cas une prudence toute particulière dans l'interprétation des quelques-uns de nos résultats partiels qui seront fondés sur des tests paramétriques.

5.2 Proximité d'activation et proximité cosinus

Concernant notre investigation, au niveau des core-tokens successifs de chaque neurone, de la relation entre proximité d'activation et proximité (similarité) cosinus, les tableaux n°1 (neurones de la couche 0) et n°2 (neurones de la couche 1) semblent éclairants. Nous y constatons en effet des moyennes faibles, par rapport à un empan théorique de 0 à 1, des valeurs cosinus similarité, avec respectivement des valeurs de 0.38 et 0.42 pour une mesure réalisée sur la base des embeddings de GPT-2XL (les plus fiables comme indiqué). Une mesure inférentielle de χ^2 d'adéquation opérée sur le pourcentage de neurones ayant des valeurs moyennes cosinus similarité inférieures à 0.5 (c'est-à-dire plutôt faibles) est amplement compatible avec cette première constatation descriptive ($p(\chi^2)<0.05$ dans le cas des deux couches, N=6400) si l'on prend une hypothèse

	GPT2	Alibaba	Mixedbread	WhereIsAI
% of (p(SW)>.05)	55,94	22,84	20,53	18,16
% of (p(Lil)>.05)	73,22	42,09	34,2	31,31
% of (p(KS)>.05)	97,91	83,03	70,75	66,62
% of (p(JB)>.05)	68,41	24,05	20,33	18
% of ((m-Q_2)<(.05*m))	100	100	100	100
% of (|Skewness|<1)	94,64	47,97	39,88	36,88
% of (|Kurtosis|<1)	87,41	36,31	28,28	25,91

Tableau 3 : pourcentages des statistiques inférentielles (α=.05) et descriptives de normalité des cosinus similarité des paires de core-tokens successifs (Couche 0, n=6400).

	GPT2	Alibaba	Mixedbread	WhereIsAI
% of (p(SW)>.05)	45,86	9,97	8,14	7,03
% of (p(Lil)>.05)	65,69	22,25	15,45	13,45
% of (p(KS)>.05)	96	61,7	43,28	38,8
% of (p(JB)>.05)	56,77	10,61	8,28	7,19
% of ((m-Q_2)<(.05*m))	100	100	100	100
% of (|Skewness|<1)	89,93	26,34	17,33	15,88
% of (|Kurtosis|<1)	81,42	20,28	12,48	11,47

Tableau 4 : pourcentages des statistiques inférentielles (α=.05) et descriptives de normalité des cosinus similarité des paires de core-tokens successifs (Couche 0, n=6400).

d'équi-distribution théorique. Ce résultat est de même cohérent avec les valeurs moyennes également faibles de Q3 (respectivement 0.48 et 0.51), et les valeurs moyennes très faibles des *minima* de cosinus (respectivement 0.06 et 0.1). Cette première vue, exploratoire, va dans le sens du fait, au niveau de la globalité de la distribution des cosinus prise comme un tout, que proximité d'activation (i.e. proximité de niveau d'appartenance catégorielle entre deux tokens) ne va pas de pair avec proximité cosinus (i.e. proximité catégorielle entre ces deux tokens).

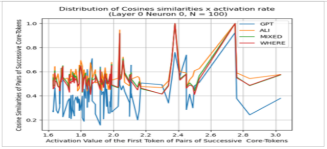
Les graphes n°3 (neurone témoin 0 de la couche 0) et n°4 (neurone témoin 0 de la couche 1) font montre d'exemples représentatifs du type de distribution des cosinus similarité en fonction de la valeur d'activation du premier token de chaque paire (pour les 100 core-tokens retenus) (cf annexes pour le cas des autres neurones témoins). Sur ces exemples ponctuels, nous pouvons globalement constater : (i) à nouveau des valeurs relativement faibles des cosinus similarité (notamment pour le cas d'une mesure avec les embedding de GPT-2XL), (ii) une variabilité qui semble qualitativement non négligeable des cosinus similarité ; cela, de façon stable pour les 4 modèles d'embeddings (même si la variabilité semble plus importante lorsque le cosinus est calculé à partir des embeddings de GPT-2XL, ce qui est normal étant donné le pouvoir sémantique plus discriminant de ce système d'embedding). Cette vue, qualitative dans la mesure où elle ne porte que sur des exemples, illustre le potentiel fait que, au niveau de la globalité de la distribution des cosinus, proximité d'activation n'irait pas de pair avec proximité cosinus ; en effet, nous n'obtenons pas ici des graphes stables (i.e. linéaires de type y=a) avec une valeur relativement élevée et constante des cosinus similarité des core-tokens successifs quant à leur niveau d'activation.

[h]

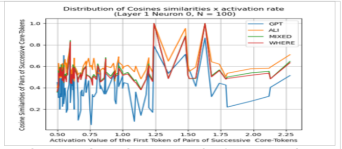
Nous allons dans ce qui suit, étudier quantitativement plus spécifiquement cette première tendance globale de non équivalence métrique de la proximité d'activation et de la proximité cosinus ; à travers deux phénomènes de cognition synthétique, de nature à explorer une version extrémisée de cette tendance, afin de montrer la force qui peut éventuellement être la sienne.

5.3 Discontinuité catégorielle des core-tokens successifs

Afin d'investiguer la tendance évoquée ci-dessus, à savoir que, au niveau de la globalité de la distribution des cosinus, la proximité d'activation (i.e. la proximité de niveau d'appartenance catégorielle entre deux tokens) n'est pas synonyme d'une proximité cosinus (i.e. d'une proximité catégorielle entre ces deux tokens),



Graphe 3 : Distribution des cosinus similarité entre paires de core-tokens successifs en fonction de la valeur d'activation du premier token de chaque paire (Couche 0, neurone témoin 0).



Graphe 4 : Distribution des cosinus similarité entre paires de core-tokens successifs en fonction de la valeur d'activation du premier token de chaque paire (Couche 1, neurone témoin 0).

nous testons la première hypothèse suivante, qui est un premier angle de vue porté sur cette tendance, angle de vue extrémisé comme déjà indiqué afin de montrer l'intensité potentielle que pourrait parfois avoir cette tendance : il existe une discontinuité catégorielle des core-tokens successifs quant à leur niveau d'activation. Autrement dit : il existe des points de rupture catégorielle (i.e. de rupture sémantique) entre les core-tokens successifs. Autrement dit encore : il existe des cosinus similarité particulièrement faibles entre core-tokens successifs relativement à leur niveau d'activation.

Afin de la tester, nous réalisons une première opérationnalisation de cette hypothèse en termes d'outliers de la distribution de la variable cosinus similarité, et plus précisément d'outliers inférieurs, la notion d'outlier inférieur incarnant de jure parfaitement bien l'esprit de notre hypothèse. Les tableaux n°5 (couche 0) et n°6 (couche 1) indiquent les nombres moyens d'outliers inférieurs significatifs ($p<0.05$) par neurone obtenus de façon inférentielle avec le test de Grubbs. Ces nombres moyens (respectivement 0.007 et 0.005 avec les embeddings de GPT-2XL) sont très faibles et ne vont pas dans le sens de notre hypothèse. Mais notons que l'utilisation de ce test n'est pas fiable ici dans la mesure où sa condition d'application, la normalité de la distribution de la variable cosinus similarité, n'est pas bien vérifiée ainsi que nous l'avons indiqué précédemment. Une approche non paramétrique, ici celle de l'écart inter-quartile, est dès lors plus sécurisée ; et elle fait montre de plus d'outliers inférieurs : en moyenne, respectivement 0.151 et 0.149 outliers inférieurs par neurone (toujours avec GPT-2XL), avec des moyennes de moyennes de cosinus similarité extrêmement faibles (respectivement 0.057 et 0.082), manifestant bien l'intensité forte que peut parfois prendre la discontinuité catégorielle des core-tokens successifs, même si ce phénomène demeure ici marginal (mais non inexistant) et relativement statistiquement normal lorsqu'il est opérationnalisé avec une approche d'outliers. Le tableau n°7 illustre qualitativement cette discontinuité catégorielle en faisant montre de core-tokens successifs nettement sémantiquement éloignés.

		GPT2	Alibaba	Mixedbr	WhereIs
According to Grubbs' Test (p<.05)	n	0,007	0,003	0,009	0,005
	m (inf-out)	0,152	0,353	0,343	0,330
	m (non-inf-out)	0,382	0,573	0,541	0,527
	s	0,019	0,000	0,000	0,000
	CV	0,145	0,000	0,000	0,001
	Range	0,043	0,000	0,000	0,000
According to Interquartile Range	n	0,151	0,188	0,292	0,255
	m (inf-out)	0,057	0,396	0,384	0,374
	m (non-inf-out)	0,383	0,573	0,542	0,527
	s	0,008	0,002	0,002	0,002
	CV	0,043	0,004	0,006	0,005
	Range	0,019	0,004	0,005	0,004

Tableau 5 : Statistiques moyennes des outliers inférieurs des cosinus similarité des paires de core-tokens successifs (Couche 0, n=6400).

		GPT2	Alibaba	Mixedbr	WhereIs
According to Grubbs' Test (p<.05)	n	0,005	0,002	0,005	0,004
	m (inf-out)	0,164	0,364	0,353	0,345
	m (non-inf-out)	0,416	0,590	0,558	0,544
	s	0,027	0,000	0,000	0,000
	CV	0,121	0,000	0,000	0,000
	Range	0,072	0,000	0,000	0,000
According to Interquartile Range	n	0,149	0,161	0,231	0,202
	m (inf-out)	0,082	0,404	0,393	0,383
	m (non-inf-out)	0,416	0,591	0,559	0,544
	s	0,008	0,002	0,002	0,002
	CV	0,136	0,005	0,005	0,005
	Range	0,017	0,004	0,005	0,004

Tableau 6 : Statistiques moyennes des outliers inférieurs des cosinus similarité des paires de core-tokens successifs (Couche 0, n=6400).

Neurone	Activation du premier core-token de la paire	Paires de core-tokens
163	2.09, 2.11	(' Library', ' kinderg'), (' kinderg', 'menu')
1095	1.11	(' cogn', ' Eleanor')

Tableau 7 : exemples d'outliers inférieurs de cosinus similarité de paires de core-tokens successifs (couche 1, calcul avec étendue interquartile sur embeddings GPT-2XL).

Poursuivons l'étude de notre hypothèse qu'il existe, au niveau de la globalité de la distribution des cosinus, des cosinus similarité particulièrement faibles

entre core-tokens successifs relativement à leur niveau d'activation ; mais, cette fois-ci, avec une opérationnalisation moins extrémisée, de nature à la rendre plus manifeste. Cela, en prenant comme indicateur les cosinus faibles, que nous définissons, neurone par neurone, comme inférieurs au seuil du minimum du cosinus du neurone augmenté de 10 % de son étendue ; ce qui correspond, en valeurs moyennes, à des seuils de 0.14 pour le neurone 0 et 0.18 pour le neurone 1 (cf tableaux n°1 & n°2). Nous constatons alors mécaniquement (cf tableaux n°8 & n°9) des pourcentages moyens nettement plus importants de fréquences de valeurs faibles par neurone (mesurées avec les embeddings de GPT-2XL), respectivement 5,06% et 5.17% ; cela, avec des moyennes très basses de cosinus (0.089 pour le neurone 0 et 0.129 pour le neurone 1), notamment comparativement aux restes des cosinus moyens (respectivement 0.397 et 0.431). Ces importants pourcentages moyens de cosinus faibles apparaissent significatifs ($p(\chi^2)<0.05$; $\chi^2{}_1=1064$; $\chi^2{}_2=1125$) lorsqu'on les évalue de façon inférentielle à l'aide d'un χ^2 d'adéquation en prenant comme distribution théorique une répartition 1%/99% correspondant à une situation où ces cosinus faibles seraient quasiment inexistants (ce qui devrait être le cas si il y avait relation entre proximité catégorielle et proximité d'activation). Ces éléments sont compatibles avec notre hypothèse de discontinuité catégorielle des core-tokens successifs postulant l'existence de cosinus similarité particulièrement faibles ; hypothèse, à nouveau, de nature à mettre en lumière de façon relativement extrémisée le fait que proximité activationnelle (i.e. proximité de niveau d'appartenance catégorielle entre deux tokens) et proximité de cosinus (i.e. proximité catégorielle entre ces deux tokens) ne vont pas de pair (au niveau de la globalité de la distribution).

	GPT2	Alibaba	Mixedbre	WhereIs
n	5,057	6,527	7,381	8,198
m(cos_{min})	0,089	0,443	0,429	0,420
m(non-cos_{min})	0,397	0,582	0,549	0,535
s	0,022	0,013	0,013	0,013
CV	0,034	0,030	0,030	0,031
Range	0,058	0,037	0,036	0,038

*Tableau 8 : Statistiques moyennes des faibles cosinus similarité (cosine < min+0,1*Range) des paires de core-tokens successifs (Couche 0, n=6400).*

	GPT2	Alibaba	Mixedbre	WhereIs
n	5,172	7,633	9,242	10,332
m(cos_{min})	0,129	0,456	0,445	0,436
m(non-cos_{min})	0,431	0,601	0,569	0,556
s	0,023	0,015	0,015	0,015
CV	0,225	0,034	0,034	0,035
Range	0,059	0,044	0,045	0,046

*Tableau 9 : Statistiques moyennes des faibles cosinus similarité (cosine < min+0,1*Range) des paires de core-tokens successifs (Couche 1, n=6400).*

5.4 Inhomogénéité catégorielle mono-activationnelle des core-tokens successifs

Toujours afin d'explorer plus avant la tendance mentionnée initialement au niveau de la globalité de la distribution des cosinus, à savoir que la proximité d'activation n'est pas équivalente à une proximité cosinus, nous testons maintenant la deuxième hypothèse suivante, qui est un second angle de vue porté sur cette tendance, angle à nouveau extrémisé afin de montrer l'intensité que peut éventuellement avoir cette tendance : il existe une inhomogénéité catégorielle des core-tokens successifs à même niveau d'activation. Autrement dit, les core-tokens ayant les mêmes niveaux d'activation ne sont pas catégoriellement les

plus proches. Ce nouveau point de vue extrémisé consiste cette fois-ci non plus à nous focaliser d'emblée sur les cosinus similarité les plus faibles, comme cela a été le cas avant, mais à nous intéresser aux cas où les activations sont proches au point d'être (quasi) identiques et devraient alors fortement être associées à des cosinus similarité forts si proximité d'activation et proximité cosinus étaient des phénomènes allant de concert.

Dans le but d'opérationnaliser le test de cette hypothèse, nous définissons les core-tokens successifs à activations (quasi) identiques comme ceux dont les niveaux d'activation sont égaux à 2 décimales près. Et nous définissons, pour chaque neurone, un indicateur de distance «d» qui est égal à l'écart entre le cosinus similarité maximal de ce neurone et le cosinus similarité des core-tokens successifs à même activation; distance dont nous vérifions ensuite la supériorité au seuil que constitue le premier quartile (Q1) de la distribution des cosinus pour ce neurone, afin de montrer que statistiquement les core-tokens à même activation ne sont pas catégoriellement les plus proches. Les tableaux n°10 (couche 0) et n°11 (couche 1) indiquent tout d'abord que les tokens à activations (quasi)identiques sont très nombreux (respectivement 47,25 et 35,31 tokens pour 100 tokens par neurone), ce qui va nous permettre d'étudier de façon consistante le phénomène qui nous intéresse ici. Nous pouvons voir que les distances moyennes sont très fortes (0.44 et 0.46, lorsque mesurées avec les embeddings de GPT-2XL, mais aussi assez importantes avec les autres embeddings pourtant plus enclins à sur-représenter les forts cosinus similarité). Les pourcentages de neurones témoignant d'une distance d supérieure à Q1 sont extrêmement élevés (respectivement 80.72% et 78.94%); et significatifs ($p(\chi^2)<0.05$) lorsqu'évalués de façon inférentielle avec un χ^2 d'ajustement univarié sur la base d'une hypothèse de distribution théorique 25%/75% cohérente avec notre recours à Q1. Ces éléments semblent compatibles avec notre hypothèse d'inhomogénéité catégorielle mono-activationnelle des core-tokens successifs, au niveau de la globalité de la distribution des cosinus.

	GPT2	Alibab	Mixedbrea	WhereIs
n	47,25			
m(d)	0,444	0,310	0,312	0,320
m(d/Range)	0,580	0,664	0,684	0,694
% of d>Q_1	80,72	3,13	8,69	15,36
P(χ) of d>Q_1	0,000	-	-	-

Tableau 10 ; statistiques moyennes de la distance d (d = Max(COS(neurone)) - (COS(core-token(n),core-token(n')))) & comparaison d-Q_1(cosinus) pour les paires de core-tokens successifs à même activation (Couche 0, n=6400).

	GPT2	Alibab	Mixedbrea	WhereIs
n	35,31			
m(d)	0,462	0,359	0,376	0,387
m(d/Range)	0,596	0,693	0,726	0,737
% of d>Q_1	78,94	4,91	12,42	21,95
P(χ) of d>Q_1	0,000	-	-	-

Tableau 11 ; statistiques moyennes de la distance d (d = Max(COS(neurone)) - (COS(core-token(n),core-token(n')))) & comparaison d-Q_1(cosinus) pour les paires de core-tokens successifs à même activation (Couche 1, n=6400).

A des fins de dotation d'un autre angle d'étude de notre hypothèse d'inhomogénéité catégorielle, nous mettons en place l'opérationnalisation complémentaire suivante, consistant pour chaque neurone à comparer son cosinus similarité moyen relatif aux core-tokens successifs à même activation au seuil de son 3ème quartile (Q3) de cosinus similarité, afin de montrer, en cohérence avec notre hypothèse, l'infériorité du premier au second. Les tableaux n°12 et n°13 manifestent que cela est très massivement le cas pour tous les systèmes d'embedding (dont 100% avec les embeddings de GPT-2XL). Ce qui

est confirmé au niveau inférentiel avec des pourcentages extrêmement élevés de cas où $p(t)<0.05$ dans le cadre d'une comparaison de Student univariée de moyenne (des cosinus) à une norme (Q3); cela à nouveau pour tous les modèles d'embeddings disponibles; dont 99,67% pour la couche 0 et 95.66% pour la couche 1 avec les embeddings de GPT-2XL. Notons également des cosinus similarité moyens assez faibles avec une mesure à base d'embeddings de GPT-2XL (respectivement 0.375 et 0.406), tout comme avec les autres embeddings (en gardant à l'esprit leur tendance à surestimer les valeurs fortes de cosinus). Ce second angle de vue est à nouveau compatible, au niveau de la globalité de la distribution des cosinus, avec notre hypothèse postulant que les core-tokens ayant les mêmes niveaux d'activation ne sont pas catégoriellement les plus proches, de nature à montrer de façon exacerbée à quel point proximité activationnelle et proximité de cosinus ne seraient pas des éléments isomorphes; c'est-à-dire dans quelle mesure la proximité de niveau d'appartenance catégorielle entre deux tokens est un phénomène qui serait dissocié de la proximité catégorielle entre ces deux tokens. Le tableau n°14 exemplifie qualitativement de façon illustrante cette hypothèse d'inhomogénéité catégorielle mono-activationnelle en faisant montre de paires de tokens à mêmes activations largement disparates d'un point de vue sémantique.

	GPT2	Aliba	Mixed	Where
n	47,25			
m(COS(n,n'))	0,375	0,568	0,537	0,522
m(COS(n,n')/Range)	0,503	1,329	1,334	1,299
% of m(COS(n,n'))<Q_3	100	99,97	99,98	99,91
% of (p(t)<.05)	99,67	96,83	93,67	92,52

Tableau 12 ; statistiques moyennes des cosinus des paires de core-tokens successifs à même activation & comparaison à Q_3(cosinus) (Couche 0, n=6400).

	GPT2	Aliba	Mixedbre	WhereIs
n	47,25			
m(COS(n,n'))	0,406	0,584	0,551	0,537
m(COS(n,n')/Range)	0,537	1,186	1,146	1,105
% of m(COS(n,n'))<Q_3	100	99,422	98,641	98,063
% of (p(t)<.05)	95,66	83,86	76,77	74,70

Tableau 13 ; statistiques moyennes des cosinus des paires de core-tokens successifs à même activation & comparaison à Q_3(cosinus) (Couche 1, n=6400).

n	Mono-activations des paires de tokens	Paires de tokens	d moyen / Range
19	1.4, 1.41, 1.41, 1.44, 1.5, 1.55, 1.66, 1.68, 1.7, 1.72, 1.76, 1.81, 1.81, 2, 2.01, 2.15, 2.25, 2.25, 2.62	(' compr', 'ecake'), (' Cinderella', ' drinkers'), (' drinkers', ' Vi'), (' rye', ' Gentleman'), (' pubs', ' gluten'), (' Rye', ' cocoa'), (' alcohol', ' sweets'), (' alcoholism', ' nicotine'), (' Hungary', 'isky'), (' Uzbek', ' mildly'), (' caramel', ' narc'), (' Presbyter', ' Vanilla'), (' Vanilla', ' sweetness'), (' tobacco', ' Croatia'), (' Croatia', 'itzerland'), (' whiskey', ' Tobacco'), (' Denmark', ' vanilla'), (' vanilla', ' Switzerland'), (' Finland', ' Sard')	0,553

Tableau 14 ; exemples de paires de core-tokens à même niveau d'activation (Couche 1, neurone 113).

5.5 Synthèse

Résumons maintenant nos traitements statistiques visant à étudier, au niveau des core-tokens successifs de chaque neurone, des caractéristiques d'une éventuelle existence d'une relation entre proximité d'activation et similarité (proximité) cosinus. Nous avons obtenu des résultats statistiques de nature à être cohérents avec les deux hypothèses qui ont été formulées concernant des phénomènes de la cognition synthétique : 1. Une hypothèse de discontinuité catégorielle des core-tokens successifs quant à leur niveau d'activation, postulant qu'il existe des cosinus similarité particulièrement faibles entre core-tokens successifs. 2. Une hypothèse d'inhomogénéité catégorielle mono-activationnelle des core-tokens successifs, posant que les core-tokens ayant les mêmes niveaux d'activation ne sont pas catégoriellement les plus proches.

6 Discussion de nos résultats

Notre série d'études s'est interrogée, concernant les core-tokens successifs, sur l'éventuelle existence d'une relation entre proximité de niveau d'appartenance catégorielle et proximité de similarité catégorielle ; proximités opérationnalisées en termes de niveau d'activation pour la première et de niveau de similarité cosinus pour la seconde. Nos résultats en la matière tendent vers l'idée d'une indépendance, au niveau global, entre proximité d'activation et proximité cosinus ; i.e. vers une non équivalence, pour des tokens donnés, entre leur proximité de niveau d'appartenance à une catégorie neuronale et l'intensité de leur similarité. Autrement dit, ce n'est pas parce que deux tokens sont proches en termes d'activation qu'ils sont proches au niveau catégoriel. Dans le cadre de notre présent travail, cette phénoménologie se manifeste à travers deux observables de la cognition synthétique que nous avons tenté de mettre à jour : la discontinuité catégorielle des core-tokens successifs et leur inhomogénéité catégorielle mono-activationnelle.

6.1 Tendances actuelles en explicabilité neuronale artificielle

Plusieurs tendances interprétatives actuelles, dans le champ de l'investigation de l'explicabilité neuronale synthétique, nous semblent constituer des clés élémentaires, à combiner, de la compréhension de cette phénoménologie de divergence entre activation et similarité. Nous les présentons sommairement dans ce qui suit avant de nous livrer à une possible élaboration de leur articulation.

La polysémie neuronale synthétique en est une première. Elle renvoie à l'idée que les neurones synthétiques sont conceptuellement distributifs (Fan et al, 2023), c'est-à-dire qu'ils peuvent conjointement correspondre à plusieurs concepts sémantiques (Bills et al, 2023). Et c'est ainsi que ces derniers indiquent que les neurones (i) ne peuvent peut-être pas avoir d'explications simples mais seulement des interprétations longues et disjonctives, (ii) ne doivent peut-être pas être pensés comme des unités de calcul homogènes au niveau sémantique. Ce que Bricken et al. (2023) dénotent en indiquant que les neurones formels répondent à des traits non reliés entre eux.

Les auteurs tendent à lier la polysémie synthétique à la notion de superposition qui exprime l'idée que des propriétés cognitives et des traits sémantiques peuvent être ventilés au sein de nombreux neurones polysémiques (Olah et al, 2020). Et qu'un même concept est ainsi distribué à travers différents neurones (Bills et al, 2023).

Une autre notion introduite par Bills et al (2023) qui nous semble ici pertinente, et qui renvoie à une potentielle illusion d'interprétabilité (Pichat, 2024b), est celle d'alien concept. A savoir que les concepts neuronaux formels peuvent être des concepts pour lesquels les êtres humains n'ont pas de mot (pas de signifiant au sens de Saussure) voire même correspondre à des « abstractions naturelles » non encore découvertes par les humains (absence de signifié hu-

main). Cela dans la mesure où, indiquent les auteurs, les modèles de langage s'occupent de choses différentes de nous, par exemple des construits statistiques utiles pour la tâche pour laquelle ils ont été entrainés, à l'instar de la prédiction du token suivant.

Enfin, Bricken et al. (2023) avancent que le neurone ne constitue pas une bonne unité d'interprétation sémantique en introduisant l'idée de l'existence d'espaces vectoriels synthétiques sémantiques intermédiaires. Un réseau de neurones créerait un espace vectoriel intermédiaire virtuel dont chaque vecteur de la base serait un trait sémantique *a priori* indépendant, fondamental, unique et mono-sémantique. Chacun de ces traits est obtenu par combinaison linéaire de neurones, i.e. chaque trait est un vecteur sur ces neurones. Chaque trait constitue ainsi une direction linéaire interprétable, une direction sémantique élémentaire. Et, dès lors, le vecteur d'activation à la sortie d'une couche neuronale pourrait être décomposé dans cet espace intermédiaire dont les vecteurs unitaires sont les traits élémentaires. Chacun de ces traits serait par définition invisible au niveau d'un seul neurone, raison pour laquelle le neurone ne serait pas forcément la bonne unité d'analyse selon les auteurs ; ces derniers indiquant par exemple, dans le cadre de leur étude, que seulement 512 neurones peuvent représenter des dizaines de milliers de traits. A partir de ces directions sémantiques fondamentales, des directions plus complexes seraient créées, celles que constituent les neurones, qui nous apparaissent alors *de facto* polysémiques dans la mesure où ils sont conceptuellement une projection compressée, i.e. à faible dimension, de ces espaces vectoriels intermédiaires beaucoup plus vastes.

Catégorielle et de la Similarité Catégorielle comme Signe de la Singularité Catégorielle de la Cognition Synthétique]La dissociation de la proximité d'appartenance

Catégorielle et de la Similarité Catégorielle comme Signe de la Singularité Catégorielle de la Cognition Synthétique Suite à cette présentation de tendances explicatives neuronales actuelles, tentons une réponse à notre observation empirique centrale, en les remaniant et en les transposant au sein d'un tout explicatif adapté. Pourquoi la proximité d'activation n'est pas un corollaire de la similarité ? Parce qu'un neurone code une catégorie synthétique, qu'il a créé dans le cadre de son activité finalisée, qui n'est pas unifiée, c'est-à-dire qui est polysémique. Cette polysémie nous fait apparaître cette catégorie comme un alien concept (ce qu'elle est effectivement pour notre cognition humaine) dans la mesure où elle est le fruit d'une superposition de sous-dimensions catégorielles générées par sa base vectorielle catégorielle intermédiaire (base que nous ne pensons pas dans les mêmes termes que Bricken et al (2023) mais plutôt, pour un neurone donné en couche n, en termes dimensions catégorielles de sortie de ses neurones précurseurs en couche n-1). Deux activations proches (cf notion de discontinuité catégorielle) et même identiques (cf notion d'inhomogénéité catégorielle mono-activationnelle) peuvent ainsi correspondre à des cristallisations, à des matérialisations, à des instanciations locales (par analogique quantique, nous pourrions parler d'effondrements de fonctions d'onde) de sous-dimensions catégorielles différentes. Autrement dit, des activations proches peuvent ainsi relever d'actualisations de sous-dimensions distinctes ; qui vont dès lors mé-

caniquement se traduire par des mesures de cosinus similarité faisant montre d'une discontinuité ou d'une inhomogénéité catégorielle, concepts spécifiques à la cognition synthétique, en tout cas concepts nous apparaissant lorsque nous étudions cette cognition synthétique à partir de notre propre référentiel humain de pensée nous donnant à postuler un *a priori* de logique cohérence sémantique entre activation et similarité.

La polysémie neuronale n'est ainsi pas segmentée catégoriellement dans les segments activationnels : segments catégoriels et segments activationnels sont deux registres dissociés dans la cognition synthétique neuronale. Car cette cognition catégorielle neuronale synthétique, à la différence de notre pensée humaine, n'est pas unifiée catégoriellement, en tout cas n'est pas unifiée au sein de concepts analogues aux nôtres. Dès lors, rechercher une convergence entre proximité catégorielle et proximité cosinus relève en partie *ipso facto* d'une démarche anthropocentrée qui ne pouvait qu'aboutir à la divergence corrélationnelle empirique dont les notions synthétiques de discontinuité et d'inhomogénéité catégorielle font montre. Cela, *a fortiori* dans le cadre de notre démarche méthodologique d'utilisation du cosinus similarité comme instrument de mesure de la proximité catégorielle : le cosinus similarité étant fondé sur un espace catégoriel vectoriel, celui des embeddings de départ de GPT-2XL (qui est par construction plus en phase avec la sémantique humaine), qui n'est pas celui des espaces vectoriels recombinés (par les valeurs de leurs fonctions d'agrégation respectives) des neurones des couches investiguées.

Parler de polysémie neuronale relèverait donc épistémologiquement d'un anthropocentrisme cognitif (Pichat, 2024). En effet, nous mobilisons le terme de polysémie car les catégories synthétiques nous apparaissent sémantiquement inhomogènes étant donné que nous n'avons pas de catégories de pensée humaines auxquelles les apparier. Mais n'est-ce pas le propre de l'abstraction catégorielle que de réunir des segments catégoriels initialement séparés ? La polysémie invoquée n'est qu'en fait le fruit d'un regroupement catégoriel auquel nous ne sommes pas (à ce stade de notre évolution conceptuelle) habitués, et nous ressemblons ainsi en mobilisant ce terme aux habitants du « flatland » de Watzlawick (1977) cantonnés à voir un monde multi-dimentionnel à travers le seul prisme sous-dimensionnel propre à leur monde.

Conclusion

Notre hypothèse de discontinuité catégorielle des core-tokens successifs quant à leur niveau d'activation (postulant qu'il existe des cosinus similarité particulièrement faibles entre core-tokens successifs) ainsi notre hypothèse d'inhomogénéité catégorielle mono-activationnelle des core-tokens successifs (posant que les core-tokens ayant les mêmes niveaux d'activation ne sont pas catégoriellement les plus proches) sont complémentaires; cela, dans la mesure où elles ont trait à la question de l'existence globale d'une relation entre proximité d'activation et proximité (similarité) cosinus. Elles devraient être complétées par une troisième hypothèse, portant quant à elle, sur l'éventuelle évolution de la dynamique distri-

butionnelle de cette relation en fonction du niveau de valeur des segments activationnels; hypothèse visant à étudier une éventuelle convergence catégorielle des paires de core-tokens successifs en fonction de l'activation, en proposant que plus les niveaux d'activation des core-tokens successifs (ie proches au niveau activationnel) augmentent et plus la variabilité catégorielle de ces core-tokens diminue (i.e. plus la proximité catégorielle augmente). Nous allons très prochainement publier nos résultats en la matière (Pichat et al., in press a), d'autres résultats non mentionnés ici nous poussant dans la direction de cette nouvelle hypothèse formulée.

L'élément clé postulé, dans notre discussion des résultats obtenus afin de tenter de les doter d'un cadre explicatif cohérent, est que le segment catégoriel créé par un neurone donné d'une couche n (plus exactement par sa fonction d'agrégation entre autres) est *de facto* décomposable en un espace vectoriel de sous-dimensions catégorielles; ces dernières étant le fruit d'une projection de l'espace vectoriel d'entrée de ce neurone, espace vectoriel d'entrée étant (par construction mathématique de sa fonction d'agrégation) composé des dimensions catégorielles de sortie de chacun des neurones précurseurs (sur la couche n-1) de ce neurone. Autrement dit, qu'en matière d'explicabilité, un neurone doit d'emblée être pensé comme étant multi-dimensionnel, c'est-à-dire composé de sous-dimensions catégorielles dont vont tendre à séparément relever les tokens; en tout cas les tokens à «faibles activation» (mono-déclenchements catégoriels) à la différences des tokens à «fortes activations» qui vont plus tendre à conjointement en relever (co-déclenchements catégoriels) ainsi que nous sommes en train de le mettre à jour (Pichat et al., in press a). Nous allons prochainement explorer ce postulat, dans le cadre d'une étude «génétique» qui visera à expliquer l'abstraction catégorielle opérée en sortie par les neurones artificiels sur la base d'une recomposition reconstructrice des segmentations catégorielles de leurs neurones précurseurs (en couche immédiatement sous-ordonnée) les plus contributifs (i.e. à poids de connexion neuronale les plus importants).

Remerciements

Michael Pichat remercie Sébastien Duizabo et David Abonneau (Université Paris Dauphine PSL) pour les stimulants colloques et expérimentations pédagogiques en matière d'IA appliquée qu'ils rendent possibles avec nous, Emmanuel Brochier (IPC-Facultés Libres de Philosophie et de Psychologie de Paris) pour les passionnants projets académiques que nous avons avec lui en matière d'IA, Igor Zatsman (Académie des Sciences de Russie) et Nadia Buntman (Université d'Etat de Moscou Lomonossov) pour les captivants échanges IA et linguistique, Madeleine Pichat pour sa relecture attentive de cet article.

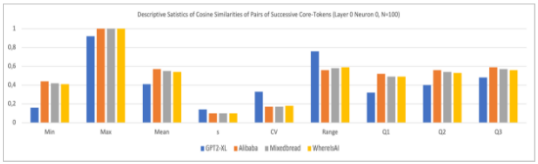
Contributions des auteurs

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.

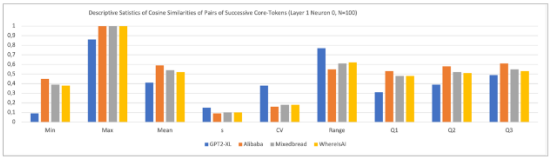
Annexes

A.1 Descriptive Mean Satistics of Cosine Similarities of Pairs of Successive Core-Tokens (Witness Neurons)

Layer 0 Neuron 0				
N (tokens) = 100	GPT2-XL	Alibaba	Mixedbread	WhereIsAI
Min	0,16	0,44	0,42	0,41
Max	0,92	1	1	1
Mean	0,41	0,57	0,55	0,54
s	0,14	0,1	0,1	0,1
CV	0,33	0,17	0,17	0,18
Range	0,76	0,56	0,58	0,59
Q1	0,32	0,52	0,49	0,49
Q2	0,4	0,56	0,54	0,53
Q3	0,48	0,59	0,57	0,56

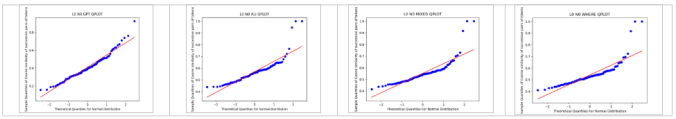


Layer 1 Neuron 0				
N (tokens) = 100	GPT2-XL	Alibaba	Mixedbread	WhereIsAI
Min	0,09	0,45	0,39	0,38
Max	0,86	1	1	1
Mean	0,41	0,59	0,54	0,52
s	0,15	0,09	0,1	0,1
CV	0,38	0,16	0,18	0,18
Range	0,77	0,55	0,61	0,62
Q1	0,31	0,53	0,48	0,48
Q2	0,39	0,58	0,52	0,51
Q3	0,49	0,61	0,55	0,53

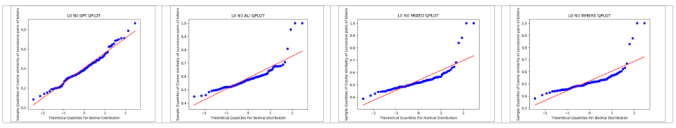


A.2 Normality Ratio Statistics of Cosine Similarities of Pairs of Successive Core-Tokens (Witness Neurons)

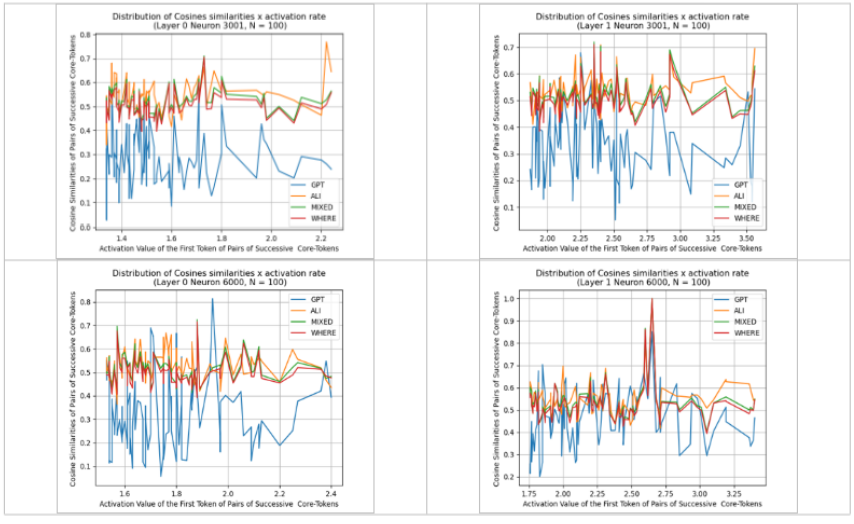
Layer 0 Neuron 0				
N (tokens) = 100	GPT2-XL	Alibaba	Mixedbread	WhereIsAI
SW p	0,0036	0,0000	0,0000	0,0000
Lilliefors p	0,1402	0,001	0,001	0,001
KS p	0,5342	0,0174	0,001	0,0018
JB p	0,0001	0,0000	0,0000	0,0000
(Mean-Q2)<(.05*Mean)	YES	YES	YES	YES
Skewness	0,85	2,3	2,69	2,8
Kurtosis	1,33	7,74	9,68	10,25



Layer 1 Neuron 0				
N (tokens) = 100	GPT2-XL	Alibaba	Mixedbread	WhereIsAI
SW p	0,1336	0,0000	0,0000	0,0000
Lilliefors p	0,5759	0,001	0,001	0,001
KS p	0,875	0,0058	0,0001	0,0001
JB p	0,1436	0,0000	0,0000	0,0000
(Mean-Q2)<(.05*Mean)	YES	YES	YES	YES
Skewness	0,48	2,46	2,98	3,11
Kurtosis	0,17	8,3	10,91	11,7



A.3 Distribution of Cosine Similarities of Pairs of Successive Core-Tokens as a Function of Activation Rank of the First Token of Pairs (Witness Neurons)



A.4 Sample of Inferior Outliers of Cosine Similarities of Pairs of Successive Core-Tokens (Interquartile range, GPT2-XL)

Layer 0			
Neuron	n	Activation of the First Core-Token	Pair of Tokens
16	10	2.75, 2.78, 2.87, 2.89, 2.99, 3, 3.23, 3.24, 3.38, 3.41	(' generates', ' Doesn'), (' delet', ' listens'), (' behaves', ' adop'), (' interven', ' opens'), (' Gets', ' overwhel'), (' overwhel', ' establishes'), (' becomes', ' convin'), (' convin', ' begins'), (' DOES', ' retrie'), (' retrie', ' Takes')
20	3	1.61, 1.63, 1.72	(' resides', ' fian'), (' amura', ' territ'), (' atana', ' worsh')
27	1	1.64	(' outper', ' Jazz')
33	2	1.99, 2.11	(' overc', ' ancient'), (' Late', ' nause')

Layer 1			
Neuron	n	Activation of the First Core-Token	Pairs of Tokens
3	1	2.33	(' unf', 'sensitive')
5	1	2.07	(' exposures', ' camoufl')
7	1	2.48	(' IRD', ' autobi')
57	2	1.53, 1.57	(' uding', ' quar'), (' quar', ' valuable')

A.5 Sample of Weak Cosine Similarities of Pairs of Successive Core-Tokens (GPT2-XL)

Layer 0			
Neuron	n	Activation of the First Core-Token	Pairs of Tokens
0	10	1.69, 1.73, 1.74, 1.76, 1.81, 1.82, 1.92, 1.94, 2.1, 2.15	(' Sessions', ' *.'), (' Memphis', 'thread'), (' Metall', ' url'), (' My', ' Nebraska'), (' WordPress', ' mah'), (' mah', '*/'), (' Moj', ' Shows'), (' Yiannopoulos', ' mc'), (' Sched', ' me'), (' Listener', ' Mog')
1	2	1.31, 1.54	(' videog', ' appellant'), (' parap', ' Crossref')
2	5	1.06, 1.08, 1.21, 1.45, 1.47	(' calm', 'ex'), (' guilty', 'let'), ('e', ' overdue'), (' fair', 'inn'), ('inn', ' low')
3	5	1.36, 1.4, 1.52, 1.55, 1.81	(' AGA', ' roadside'), (' furthe', '?",'), (' alike', ' arra'), (' arra', ' affiliated'), (' somew', ' related')

Layer 1			
Neuron	n	Activation of the First Core-Token	Pairs of Tokens
0	3	0.66, 0.76, 1.09	(' unrestricted', ' chees'), (' Voting', ' artificially'), (' fishing', ' handic')
1	2	0.73, 0.74	(' handling', ' Nicarag'), (' NYPD', ' les')
2	9	1.81, 1.84, 1.87, 1.99, 2.2, 2.24, 2.26, 2.69, 2.87	(' sched', ' mayors'), (' shouted', ' LOG'), (' Coun', ' planning'), (' Mourinho', ' log'), (' coun', ' Management'), (' logistics', ' padd'), (' padd', ' commander'), (' log', ' managing'), (' management', ' Coun')
3	4	2.07, 2.31, 2.32, 2.33	(' Resp', ' affordable'), (' soften', ' loudspe'), (' replied', ' voic'), (' unf', 'sensitive')

A.6 Sample of Sucessive Core-Tokens with Similar Activations

LAYER 0					
Neuron	n	Activation value of the first token of identical activation pairs	Tokens of identical activation pairs of sucessive tokens	Sucessive tokens with Max(COS)	d mean / Range
0	45	1.63, 1.64, 1.64, 1.67, 1.67, 1.68, 1.69, 1.73, 1.73, 1.73, 1.74, 1.75, 1.75, 1.76, 1.76, 1.76, 1.78, 1.79, 1.79, 1.81, 1.82, 1.83, 1.84, 1.84, 1.84, 1.87, 1.89, 1.92, 1.92, 1.93, 1.94, 1.96, 1.96, 1.98, 1.98, 1.99, 1.99, 1.99, 2.04, 2.04, 2.06, 2.1, 2.11, 2.11, 2.45	(' Mast', '*.'), (' Mahar', ' Buch'), (' Buch', ' Mahm'), (' Prague', ' McCull'), (' McCull', ' Miz'), (' Springfield', ' Mania'), (' albums', ' Sessions'), ('rences', ' MG'), (' MG', ' Memphis'), (' Memphis', 'thread'), (' Amon', 'Metall'), (' url', ' Nak'), (' Nak', 'MW'), (' My', ' Nebraska'), (' Nebraska', ' MY'), (' MY', ' Mick'), (' Manz', ' McGill'), (' Oz', ' Main'), (' Main', 'Meng'), (' posts', ' WordPress'), (' mah', '*/'), (' req', ' Us'), ('rack', ' Events'), (' Events', ' Ame'), (' Ame', 'lb'), (' Monroe', ' Melbourne'), ('*.', ' MLG'), (' Moj', ' Shows'), (' Shows', 'settings'), ('archives', ' MSG'), (' Mack', ' Yiannopoulos'), (' *', ' Munich'), (' Munich', 'Us'), (' venues', ' *'), (' *', ' Hoff'), (' sessions', 'uploads'), ('uploads', 'mes'), ('mes', ' Meh'), (' Scenes', 'scene'), ('scene', 'RM'), (' Mek', ' mes'), ('meg', ' Sched'), (' me', ' threads'), (' threads', 'mA'), (' Meyer', ' Manson')	(' Me', 'Me') 0.923	0,67
1	54	1.3, 1.3, 1.3, 1.31, 1.31, 1.31, 1.31, 1.32, 1.32, 1.33, 1.33, 1.33, 1.33, 1.33, 1.33, 1.35, 1.37, 1.37, 1.38, 1.38, 1.39, 1.39, 1.4, 1.4, 1.4, 1.4, 1.42, 1.46, 1.5, 1.51, 1.52, 1.53, 1.53, 1.56, 1.56, 1.58, 1.58, 1.58, 1.59, 1.59, 1.6, 1.64, 1.64, 1.64, 1.65, 1.66, 1.66, 1.67, 1.69, 1.69, 1.7, 1.76, 1.93, 2	(' Sega', ' Bungie'), (' Bungie', ' dwellings'), (' dwellings', ' Marketplace'), (' recre', ' videog'), (' videog', ' appellant'), (' appellant', 'eanor'), ('eanor', ' Goods'), ('Bio', ' Cafe'), (' Cafe', ' Ga'), ('ancers', ' Cobb'), (' Cobb', 'aine'), ('aine', ' Dexter'), (' Dexter', ' para'), (' para', 'enge'), ('enge', ' EMBER'), (' parole', ' Gamer'), (' Dell', ' Alley'), (' Alley', ' Ball'), (' apparel', ' Gabe'), (' bathing', ' robe'), (' robe', ' Madden'), (' Madden', ' cane'), (' Gle', ' Device'), (' Device', ' Beach'), (' Beach', ' attractions'), (' attractions', ' Fant'), (' Aero', 'ameda'), (' Beard', ' Brigham'), ('Age', ' Recreation'), (' Mansion', 'antes'), (' Pavilion', ' Trop'), ('ognitive', 'endor'), ('endor', 'biomark'), ('antle', ' avatar'), (' avatar', 'rero'), (' Activities', 'adden'), ('adden', ' Doll'), (' Doll', 'agus'), (' restroom', ' dementia'), (' dementia', ' Devices'), (' Fan', ' Anaheim'), (' Paddock', ' Bio'), (' Bio', 'arine'), ('arine', ' LDS'), (' biography', ' disturbance'), ('abilia', ' wearable'), (' wearable', 'affle'), (' Aki', ' Badge'), (' recreation', 'raine'), ('raine', 'EGA'), (' Bing', ' beverage'), (' Blend', ' Disorder'), (' Barbie', ' Ala'), (' affidavit', ' Masquerade')	(' Mansion', 'antes') 0.724	0,444
2	44	1.01, 1.01, 1.01, 1.02, 1.04, 1.04, 1.04, 1.04, 1.05, 1.06, 1.06, 1.06, 1.06, 1.07, 1.07, 1.08, 1.08, 1.08, 1.1, 1.1, 1.14, 1.15, 1.15, 1.15, 1.19, 1.19, 1.19, 1.2, 1.2, 1.23, 1.23, 1.24, 1.25, 1.27, 1.27, 1.28, 1.33, 1.34, 1.34, 1.35, 1.37, 1.37, 1.37, 1.41	('n', 'ny'), ('ny', 'day'), ('day', ' on'), (' closed', 'id'), ('nut', ' Wait'), (' Wait', ' farewell'), (' farewell', 'ay'), ('ay', ' cold'), (' not', 'er'), ('re', 'more'), ('more', 'allow'), ('allow', ' calm'), (' calm', 'ex'), ('out', 'ick'), ('ick', 'atts'), ('tight', 'illy'), ('illy', 'an'), ('an', ' guilty'), ('man', 'un'), ('un', 'att'), ('gross', 'yes'), ('very', 'een'), ('een', 'ood'), ('ood', ' well'), ('ent', ' sure'), (' sure', 'good'), ('good', 'false'), (' okay', ' true'), (' true', ' ok'), (' clean', 'onest'), ('onest', 't'), ('ow', ' open'), ('oops', 'Safe'), (' happy', 'safe'), ('safe', 'a'), ('oon', 'night'), ('un', 'not'), ('ine', 'free'), ('free', 'ty'), ('less', ' quiet'), ('wait', ' good'), (' good', ' sorry'), (' sorry', 'all'), ('true', ' Safe')	('ick', 'atts') 0.712	0,499
3	53	1.36, 1.37, 1.37, 1.37, 1.38, 1.38, 1.38, 1.38, 1.38, 1.39, 1.4, 1.4, 1.41, 1.41, 1.42, 1.42, 1.42, 1.43, 1.44, 1.44, 1.44, 1.44, 1.45, 1.45, 1.47, 1.49, 1.51, 1.52, 1.55, 1.55, 1.55, 1.55, 1.59, 1.59, 1.59, 1.59, 1.6, 1.6, 1.6, 1.61, 1.64, 1.7, 1.71, 1.75, 1.75, 1.76, 1.78, 1.81, 1.84, 1.85, 1.87, 1.89, 1.89	(' AGA', ' roadside'), ('?!"', 'chn'), ('chn', ' Styles'), (' Styles', 'ROR'), (' kinds', ' favorites'), (' favorites', 'gomery'), ('gomery', ' alongside'), (' alongside', ' along'), (' along', ' On'), ('outheastern', ' attach'), (' CTRL', 'rosso'), ('rosso', ' furthe'), ('?",', ' underside'), (' underside', ' Sly'), (' side', 'tein'), ('tein', 'opez'), ('opez', ' ways'), ('aldi', ' Along'), ('????????', 'ype'), ('ype', '?"'), ('?"', ' Af'), (' Af', ' hither'), ('versely', 'arest'), ('arest', ' affinity'), ('associated', '????'), (' Want', 'ites'), ('...?', ' Everywhere'), ('opp', ' alike'), (' arra', ' affiliated'), (' affiliated', 'ipher'), ('ipher', 'rou'), ('rou', 'esides'), (' closely', ' favorite'), (' favorite', ' accuse'), (' accuse', ' sidebar'), (' sidebar', 'ateg'), ('?', 'want'), ('want', 'ategories'), ('ategories', ' Favorite'), ('?!', 'Brow'), (' Modes', 'ivan'), ('afi', ' prefer'), (')?', ' symp'), ('types', ' vying'), (' vying', ' Cous'), (' hates', ' harder'), ('Side', 'aple'), (' somew', ' related'), (' prefers', 'ategor'), ('Favorite', '"?'), (' Moh', ' like'), (' Ways', 'awi'), ('awi', ' Types')	('Like', ' Like') 0.963	0,677

LAYER 1					
Neuron	n	Activation value of the first token of identical activation pairs	Tokens of identical activation pairs of sucessive tokens	Sucessive tokens with Max(COS)	d mean / Range
0	40	0.51, 0.51, 0.52, 0.52, 0.52, 0.53, 0.53, 0.53, 0.54, 0.54, 0.54, 0.55, 0.55, 0.55, 0.56, 0.57, 0.58, 0.6, 0.6, 0.61, 0.61, 0.61, 0.62, 0.62, 0.63, 0.63, 0.63, 0.63, 0.66, 0.68, 0.69, 0.72, 0.72, 0.77, 0.83, 0.87, 0.94, 0.97, 1, 1.2	(' dividends', ' disguised'), (' disguised', ' glued'), (' booze', ' deduction'), (' deduction', ' socialist'), (' socialist', ' bob'), (' Jazz', ' retarded'), (' retarded', ' Rubio'), (' Rubio', ' Hispanic'), (' cuts', ' copied'), (' copied', ' Fruit'), (' Fruit', ' pumped'), (' Bernie', ' pinned'), (' pinned', ' zombies'), (' zombies', ' donors'), (' Controlled', ' Cookies'), (' Cookies', 'Flo'), (' fisher', ' jammed'), (' manipulated', ' tricked'), (' tricked', ' Zombies'), (' bottled', ' messing'), (' messing', ' vetoed'), (' vetoed', ' pumping'), (' scripting', ' Disabled'), (' Disabled', ' nailed'), (' dunk', ' charitable'), (' charitable', ' Powers'), (' Powers', ' modifications'), (' modifications', ' sway'), (' jumper', ' unrestricted'), (' chees', ' tweaking'), (' Medicare', ' divorced'), (' Bowling', ' boxing'), (' boxing', ' adjusted'), (' artificially', ' Hispanics'), (' dyed', 'mining'), (' Cuban', ' Negro'), (' tweaked', ' delegates'), (' boosted', ' restricted'), (' screwed', ' modify'), (' colored', 'modified')	(' Casino', ' casino') 0.864	0,551
1	41	0.73, 0.74, 0.74, 0.74, 0.74, 0.75, 0.78, 0.79, 0.8, 0.82, 0.82, 0.82, 0.83, 0.83, 0.84, 0.84, 0.85, 0.85, 0.85, 0.86, 0.9, 0.92, 0.92, 0.92, 0.93, 0.93, 0.93, 0.93, 0.95, 0.95, 1.02, 1.09, 1.1, 1.13, 1.14, 1.15, 1.2, 1.26, 1.28, 1.32, 2.2	(' redevelopment', ' handling'), (' Nicarag', ' Ramirez'), (' Ramirez', ' NYPD'), (' NYPD', ' les'), (' les', ' incumb'), (' Rica', ' Blanc'), (' Buenos', ' separ'), (' Tuc', ' Suffolk'), (' Mexicans', ' Dallas'), (' departing', ' Salvador'), (' Salvador', ' Villa'), (' Villa', ' patrol'), (' Cardiff', ' Honduras'), (' Honduras', ' departed'), ('Mexico', ' stint'), (' stint', ' stops'), (' Sacramento', ' crossings'), (' crossings', ' Roberto'), (' Roberto', ' Juan'), (' Islanders', ' squad'), (' LAPD', ' resident'), (' suffix', ' goalkeeper'), (' goalkeeper', ' Handling'), (' Handling', ' relegation'), (' arra', ' delinqu'), (' delinqu', 'resident'), ('resident', ' relocation'), (' relocation', ' Luis'), (' Carib', ' Staten'), (' Staten', ' Sic'), (' stranded', ' Confeder'), (' encamp', ' transfers'), (' Southampton', ' itiner'), (' Haitian', ' sheltered'), (' Hung', ' departures'), (' apprehended', ' jurisdiction'), (' Roma', 'encia'), (' Hait', ' composure'), (' entrusted', ' los'), (' subdivision', ' Sergio'), (' Guatem', ' Marcos')	(' Roberto', ' Juan') 0.754	0,485
2	24	1.81, 1.81, 1.81, 1.81, 1.84, 1.86, 1.86, 2.18, 2.2, 2.22, 2.34, 2.44, 2.54, 2.61, 2.64, 2.66, 2.67, 2.69, 2.69, 2.87, 2.91, 2.92, 2.94, 2.94	(' Stamford', ' Pep'), (' Pep', ' paramed'), (' paramed', ' sched'), (' sched', ' mayors'), (' shouts', ' shouted'), (' LOG', ' midfield'), (' midfield', 'Plan'), (' chefs', ' Managing'), (' Managing', 'ateg', ' coun'), (' planners', 'Coach'), (' systemd', ' Strategy'), (' yell', 'fc'), (' Guardiola', ' councillor'), (' coaching', ' captain'), (' cmd', 'Msg'), (' Wenger', ' Chef'), (' Log', ' chef'), (' yells', ' dispatcher'), (' dispatcher', 'log'), (' management', ' Coun'), (' Strateg', ' strategy'), (' Sergeant', ' captains'), ('Management', 'management'), ('management', ' coach')	('Manager', 'manager') 0.937	0,64
3	25	1.97, 2.08, 2.11, 2.12, 2.12, 2.19, 2.24, 2.24, 2.31, 2.32, 2.32, 2.32, 2.33, 2.33, 2.41, 2.5, 2.51, 2.61, 2.64, 2.69, 2.75, 2.82, 2.93, 3.14, 3.3	(' balanced', 'unung'), (' affordable', 'swers'), ('adjusted', 'utral'), (' softened', ' loud'), (' loud', ' amplification'), (' answers', ' responses'), (' speakers', ' voice'), (' voice', ' dial'), (' softer', ' loudspe'), (' loudspe', ' Yo'), (' Yo', 'Adjust'), ('Adjust', ' replied'), (' voic', ' unf'), (' unf', 'sensitive'), ('sounding', ' sounding'), (' responding', 'adjust'), ('adjust', 'audio'), (' amplify', ' headphones'), (' headphone', ' listeners'), (' responded', ' audio'), ('umi', ' respond'), (' ear', ' responds'), (' responsiveness', ' muted'), (' attentive', 'Audio'), ('ono', ' impedance')	(' listened', ' listen') 0.947	0,601

Bibliographie

[1] 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

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

[3] Barsalou, L. W. (1995). *Storage side effects: studying processing to understand learning*. In: Ram, A. and Leake, D. (eds.) Goal-driven learning. MIT Press: Cambridge, MA, pp. 407-419.

[4] Beaufils, B. (1996). *Statistiques appliquées à la psychologie*. Editions Bréal.

[5] 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

[6] 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

[7] Bransford, J. D., & Franks, J. J. (1971). The abstraction of linguistic ideas. *Cognitive Psychology*, 2(4), 331–350.https://doi.org/10.1016/0010-0285(71)90019-3

[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] Brooks, N. (1987). No Silver Bullet Essence and Accidents of Software Engineering.*Computer*, *20*(4), 10-19. https://doi.org/10.1109/mc.1987.1663532

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

[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] Dalvi, F., Sajjad, H., Durrani, N., & Belinkov, Y. (2020, November). Analyzing redundancy in pretrained transformer models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP-2020)*(pp. 4908–4918). Online.

[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] 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

[16] 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

[17] 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

[18] 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

[19] 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

[20] Howell, D. C. (2008).*Méthodes statistiques en sciences humaines*. De Boeck Supérieur.

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

[22] 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

[23] 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

[24] 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

[25] Katz, M. B. (1972). Occupational Classification in History. The Journal Of Interdisciplinary History, 3(1), 63. https://doi.org/10.2307/202462

[26] Keil, F. C. (1989).*Concepts, kinds, and cognitive development*. The MIT Press.

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

[28] 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

[29] Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, 100(2), 254–278.https://doi.org/10.1037/0033-295X.100.2.254

[30] 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

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

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

[33] 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

[34] 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

[35] 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

[36] 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

[37] 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

[38] Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract idea. *Journal of Experimental Psychology*, 77(3), 353-363.https://doi.org/10.1037/h0025953

[39] 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

[40] Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, 3(3), 382-407.https://doi.org/10.1016/0010-0285(72)90014-X

[41] 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

[42] Roads, B. D., & Mozer, M. C. (2021). Predicting the Ease of Human Category Learning Using Radial Basis Function Networks. Neural Computation, 33(2), 376-397. https://doi.org/10.1162/neco_a_01349

[43] 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

[44] 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

[45] 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

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

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

[48] 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

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

[50] 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

[51] Thibaut, J. (1997). Similarité et catégorisation.*L'Année Psychologique,97*(4), 701-736.https://doi.org/10.3406/psy.1997.28989

[52] Tijus, C. (2004). Introduction à la Psychologie cognitive. Librairie Eyrolles. https://www.eyrolles.com/Loisirs/Livre/introduction-a-la-psychologie-cognitive-9782200340964/

[53] 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

[54] Whorf, B. L. (1941). Languages and logic. In J. B. Carroll (Ed.), *Language, thought, and reality: Selected papers of Benjamin Lee Whorf* (pp. 233-245). MIT Press.

[55] Wisniewski, E. J., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science*, 18(2), 221–281. https://doi.org/10.1207/s15516709cog1802_2

[56] Wittgenstein L. (1961 [1953]), Les Investigations philosophiques, Paris, Gallimard.

[57] 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

[58] 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