Les Modèles de Langage Face aux Théories Cognitivistes du Langage

Mathis Reymond

mai 2025

1 Introduction

À travers l'essor des modèles de langage de grande taille, il devient pertinent d'examiner leur place au regard des théories cognitivistes du langage. Bien que le terme réseau de neurones soit employé pour désigner à la fois les architectures biologiques et artificielles, il convient de rappeler qu'il s'agit de deux objets fondamentalement distincts. A plus forte raison, nous montrerons en quoi les fondements techniques à la base des modèles artificiels ne satisfont pas pleinement aux nombreux impératifs biologiques nécessaires pour légitimer les analogies courantes entre l'humain et la machine. Au regard de ces différences fonctionnelles, nous évoquerons également les écarts comportementaux observés entre les capacités humaines et artificielles, notamment en ce qui concerne l'abstraction et la manipulation de symboles. C'est sur la base de ces divergences, tant fonctionnelles que comportementales, que nous structurerons le rapprochement entre les modèles de langage de grande taille – portés par la linguistique computationnelle – et les théories cognitivistes du langage. Nous verrons notamment que le succès de ces modèles conforte les conceptions connexionnistes, au détriment de théories telles que la linguistique générative et ses prolongements, qui promeuvent une vision modulaire de l'esprit. Enfin, nous nuancerons le succès des modèles de langage de grande taille en soulignant leur difficile interprétabilité en supportant le besoin d'un ancrage symbolique, qui fait encore défaut aux modèles actuels.

2 Analogies humain – machine

Les réseaux de neurones artificiels et biologiques sont deux systèmes inspirés l'un de l'autre. Que ce soit dans les médias ou la littérature scientifique et philosophique, ils sont fréquemment comparés, et la question de la transférabilité des propriétés de l'un à l'autre ou de l'autre à l'un est souvent évoquée (capacité à comprendre, créativité, intentionnalité etc.). Mais ce sont également des systèmes fondamentalement différents dans leur structure et leur fonctionnement. Tandis que les réseaux de neurones biologiques sont à la base du traitement de l'information – sous forme d'impulsions électrique et de réactions chimiques – dans le cerveau des êtres vivants, les réseaux de neurones artificiels sont des modèles mathématiques conçus pour capturer certains aspects des neurones biologiques. Comprendre leurs différences permet de mieux appréhender les limites des analogies humain-machine.

Le cerveau humain repose sur des neurones biologiques qui communiquent par impulsions électrochimiques, s'organisent en structures dynamiques en constante restructuration, et intègrent une variété de signaux hormonaux et corporels non modélisés dans les systèmes actuels. De plus, il a été montré que certains types de neurones biologiques présentent des dynamiques plus proche de celles d'un réseau de neurones artificiel profond plutôt que de celle d'un neurone isolé (Beniaguev, Segev, and London 2021). Les dendrites apportent des dynaniques fondamentalement non linéaires et donc des neurones artificiels isolés, tel qu'ils sont modélisés, ne peuvent pas capturer ces dynamiques. Par ailleurs, le mécanisme d'apprentissage clé des réseaux de neurones artificiels est la rétropropagation de l'erreur (ou backpropagation, en anglais). Malgré de nombreuses tentatives de rendre ce mécanisme biologiquement plausible (Shervani-Tabar and Rosenbaum 2023) (Lv et al. 2024), aucune analogie claire dans la neurobiologie humaine n'est pour l'instant reconnue. La backpropagation pose de nombreux problèmes. Premièremennt, elle suppose l'existence de deux modes pour le réseau : un mode d'apprentissage, pendant lequel le réseaux ne peut être utilisé pour faire des inférences, et un second mode d'inférence, dans le cadre duquel les neurones sont figés et ne changent plus leurs poids. Une telle sépararation de l'apprentissage et de l'inférence n'est pas pas observée dans les réseaux de neurones biologiques. Deuxièmement, la backpropagation exige également une orchestration globale et structurée de la propagation de l'information à travers le réseau, tandis que les réseaux biologiques semblent fonctionner de manière plus locale et distribuée. Troisièmement, la backpropagation repose sur l'idée que les gradients d'erreur peuvent être calculés et propagés à travers le réseau, ce qui n'est pas le cas dans les réseaux biologiques, où les signaux sont souvent non linéaires et non différentiables. Enfin, on compte trois principaux développements qui, au cours des dernières années, ont conduit aux progrès des modèles de langues. En premier lieu, un nouveau mécanisme dit d'attention a été introduit et permet de mieux contextualiser les mots (position absolue dans la phrase

et position relative des mots entre eux), ensuite, l'architecture hautement parallélisable des transformers (Vaswani et al. 2023) a rendu possible l'entrainement de modèle avec plusieurs milliards de neurones, et surtout, l'augmentation considérable à la fois de la puissance de calcul et de taille des données a permi un entrainement beaucoup plus intensif et général. Ces avancées capitales ont été opérées complètement indépendamment de tout rapprochement aux réseaux de neurones biologiques.

A la lumière de ces divergences structurelles et fonctionnelles il apparait plus clairement que les rapprochements entre intelligence artificielle et cognition humaine relève davantage d'analogies interprétatives du comportement que de réelles correspondances techniques rigoureusement établies.

Les modèles de langue artificiels possèdent la faculté de générer des textes qui imitent ceux rédigés par l'être humain. Cette faculté leur permet de générer des contenus que nous, en tant qu'êtres humains, sommes capables d'interpréter et d'analyser. Ce qui est intéressant, c'est que cette capacité à produire des textes ait suscité, chez les scientifiques et les philosophes, un intérêt pour l'étude d'une sorte de psychologie des modèles de langage. L'ambition de cette entreprise consiste à établir un parallèle entre le comportement linguistique de ces systèmes et certaines dispositions propres à l'esprit humain, telles que la créativité, l'intentionnalité, voire une certaine forme de conscience. Or, une telle démarche interprétative ne repose guère sur des fondements techniques robustes, comme nous l'avons établi précédemment. Elle ne s'impose ni par sa nécessité théorique ni par des évidences empiriques : elle s'inscrit plutôt dans une dynamique spéculative, portée par une forme d'enthousiasme intellectuel selon laquelle tout ce qui est susceptible d'être étudié mérite *ipso facto* de l'être. Il semble ainsi que cette orientation relève moins d'une exigence scientifique que d'une fascination pour les potentialités mimétiques des machines. Il convient de rappeler que la faculté qu'ont certains systèmes artificiels de duper l'humain en produisant des énoncés perçus comme authentiquement humains n'est en rien une découverte récente. Le célèbre test de Turing (Turing 1950), conçu pour évaluer la capacité d'un dispositif à simuler le comportement humain, a été passé avec succès dès les premières générations de programmes. L'exemple d'ELIZA (Weizenbaum 1966), développé en 1966, en témoigne de manière éloquente : ce modèle de langage rudimentaire, fondé sur de simples règles grammaticales, parvenait déjà à créer l'illusion d'un véritable échange humain en reformulant les affirmations sous forme de questions.

Dans Prodiges et vertiges de l'analogie (Bouveresse 1999), Jacques Bouveresse critique l'usage excessif et souvent abusif du raisonnement par analogie, notamment dans les sciences humaines et la philosophie contemporaine. Il développe deux arguments principaux

pour en souligner les limites. D'abord, il dénonce l'analogie comme un outil de légitimation fallacieuse, en montrant comment certains auteurs s'appuient sur des concepts scientifiques complexes — comme le théorème d'incomplétude de Gödel — pour donner une apparence de profondeur à des discours qui n'ont en réalité aucun lien avec les domaines concernés. Dans le cadre de la comparaison entre réseaux biologiques et artificiels, un récit à forte teneur anthropomorphique est souvent mis en avant, tant dans les médias que, parfois, par les scientifiques eux-mêmes (Metz and Grant 2022), lequel s'appuie sur la prétendue opacité des aspects mathématiques et algorithmiques à l'oeuvre. Ensuite, Bouveresse met en lumière une tendance à privilégier les analogies séduisantes au détriment de la rigueur conceptuelle. Selon lui, cette approche conduit à survoler les différences fondamentales entre les domaines comparés, au profit de ressemblances superficielles qui affaiblissent la clarté de la pensée. Il s'appuie notamment sur Wittgenstein pour rappeler que la philosophie doit clarifier et délimiter nettement les pensées qui autrement sont, pour ainsi dire, opaques et floues. En somme, Bouveresse appelle à un usage plus critique et discipliné de l'analogie, soulignant que celle-ci ne peut se substituer à une analyse rigoureuse ni compenser un déficit de conceptualisation.

Finalement, à ce jour, nous ne disposons d'aucun algorithme d'apprentissage connu qui reproduise fidèlement le fonctionnement du cerveau humain ; tout au plus, certaines approches comme le predictive coding — qui, notons-le, ne sont pas mises en œuvre dans les modèles de langage actuels — s'en rapprochent partiellement. Dès lors, il paraît hasardeux de supposer l'existence de propriétés émergentes communes entre les réseaux biologiques et les modèles artificiels, alors même que ces derniers diffèrent radicalement des premiers à plusieurs égards essentiels : ils ne reposent ni sur le même substrat matériel, ni sur les mêmes modalités de transmission de l'information, ni sur des mécanismes d'apprentissage équivalents, ni sur des données similaires. Une telle spéculation, qui ambitionne d'établir des parallèles sur la seule base d'analogies comportementales, semble d'autant plus fragile que les notions psychologiques invoquées demeurent elles-mêmes mal définies du point de vue théorique et difficilement étudiées du du point de vue expérimental.

3 Raisonnement

Des études récentes ont démontré que les modèles de langage de grande taille obtiennent de bonnes performances dans des tâches impliquant du raisonnement. Cela inclut des tâches liées à l'exploration spatiale, à l'ordre temporel, à l'inférence causale et aux mathématiques en général (Chen et al. 2024) (Xiong et al. 2024). Ces tâches sont particulièrement bien réalisées par les modèles dits « de raisonnement » – des modèles de langage spécifiquement conçus pour accomplir ce type de tâches. La sortie de Claude 4 Opus et Sonnet 4 par An-

thropic a introduit des modèles hybrides de raisonnement capables de résoudre des problèmes complexes, de coder et d'exécuter des tâches autonomes prolongées, avec des fonctionnalités telles que la « pensée étendue » et les « résumés de pensée » améliorant leurs performances. Le modèle de raisonnement o3 d'OpenAI, successeur du o1, met l'accent sur un raisonnement interne délibéré, atteignant de meilleurs résultats dans les benchmarks scientifiques et mathématiques grâce à l'apprentissage par renforcement et à des stratégies de décodage en plusieurs étapes. AlphaGeometry de DeepMind a combiné la logique symbolique avec les LLMs pour résoudre des problèmes de géométrie complexes de niveau olympiade. Grok-3 de xAI, entraîné avec d'importantes ressources de calcul, a démontré un raisonnement avancé en surpassant ses pairs lors d'évaluations mathématiques et scientifiques. En outre, des projets de recherche comme LLaVA-CoT et LongRePS ont amélioré le raisonnement multimodal et sur de longs contextes, respectivement, en mettant en œuvre des techniques de traitement structurées, étape par étape. Ces modèles de raisonnement reposent sur un mécanisme intégré de « chaîne de pensée » (Chain-of-Thought, ou CoT). Le mécanisme CoT consiste à faire générer par les LLMs une analyse pas à pas d'un problème avant de fournir une réponse. Les modèles de raisonnement sont désormais spécifiquement conçus pour « penser » en générant une chaîne de pensée avant de produire la réponse finale. Il a été démontré que, de manière générale, le CoT améliore significativement les performances des LLMs dans les tâches complexes (Wei et al. 2022). Cependant, certaines découvertes récentes suggèrent que le CoT pourrait parfois être contre-productif, car il s'agirait d'un type de raisonnement bruité influencé, lui aussi, par une mémorisation probabiliste (Cuadron et al. 2025)(McCoy et al. 2024).

En dépit du vocabulaire propre à décrire des capacités humaines dont nous avons fait l'emploi dans le paragraphe précédent, nous voulons exprimer de fortes réserves quant à la capacité des modèles de raisonnement à accomplir des tâches de raisonnement au sens communément entendu. Les modèles de langage, échouent très fréquemment dès qu'il s'agit de manipuler des symboles en dehors de la distribution de données sur laquelle ils ont été entraînés. Ils sont notamment en incapacité de manipuler avec pertinence de nouveaux symboles dont on leur donne la définition — comme des chaînes inventées du type "ghjk" McCoy et al. 2024. De plus, les difficultés d'abstraction et de raisonnement sont de plus en plus mise en lumière par des tests comme le ARC Challenge (Chollet et al. 2024), dans le cadre desquels les modèles de langage échouent à résoudre des puzzles pourtant triviaux pour un humain. Encore un exemple de limitation est l'incapacité des modèles à lire l'heure sur une montre analogique ou, inversément, à générer une image de montre indiquant une certaine heure spécifique. Ces échecs reflètent la nature prodondément statistique et non symbolique de l'entrainement de ces modèles de langage. Ils ne sont pas capables de manipuler des symboles abstraits de manière cohérente et fiable, ce qui est pourtant une caractéristique essentielle du raisonnement humain.

Il est certes impossible de démontrer de manière absolue que les modèles de langage ne tentent pas de raisonner, mais il est en revanche possible de constater qu'ils échouent fréquemment à le faire. Ce type d'échec observable sur des tâches bien définies permet d'évaluer objectivement les limites des modèles en matière de raisonnement et d'être prudent quant à l'attribution de concepts psychologiques humain à ces modèles. En revanche, si le raisonnement est par essence une notion précisément formalisable, il n'en n'est rien des propriétés telles que la créativité, l'intentionnalité ou la conscience ; la situation est donc bien différente. En l'absence de définition claire — et a fortiori de définition formelle — de ces notions, il est impossible de prouver que les modèles ne les possèdent pas. On ne peut pas non plus montrer qu'un modèle "échoue à être conscient" car cela n'a simplement pas de sens. Tout au plus peut-on soutenir que les comportements observés chez ces systèmes n'exigent pas, en eux-mêmes, la présence de telles facultés.

Les modèles de langage génératifs peuvent produire des textes indicernables de ceux écrits par un humain. Cependant, cela ne signifie pas qu'ils accomplissent réellement la tâche dans toute sa complexité. Comme nous l'avons établit plus haut, nous n'avons pas accès au fonctionnement causal interne des modèles de langue. Donc, contrairement à un algorithme classique, écrit de manière séquentiel, avec des étapes explicites dans la réalisation d'un objectif non ambigu, on ne peut apprécier les tâches réalisées par les modèles de langages qu'à travers les résultats de ces tâches. Pourtant il y a une différence entre faire une tâche et imiter le résultat de cette tâche. On peut apprécier cela en considérant le paradoxe du singe savant. Si un singe tape au hasard sur un clavier pendant une durée infinie, il finira par reproduire l'intégralité de tous les textes jamais écrits par l'homme. En particulier, en lui accordant un temps arbitrairement long, mais fini, il écrira l'intégralité de l'œuvre de Fodor. Cependant, cela ne signifie pas qu'il a compris ni même qu'il a accompli la tâche d'écrire Fodor. Il n'a pas réfléchi, ni planifié, ni même eu l'intention de produire une œuvre philosophique. Il a simplement tapé au hasard, et le résultat est une coïncidence statistique. Cette réflexion résulte d'un résultat mathématiques. On peut toutefois la prolonger par une expérience de pensée en substituant le singe savant par les modèles de langue actuels. Ces modèles ne produisent pas des mots ou des phrases au hasard, mais génèrent des séquences de mots en se basant sur leur probabilité statistique d'apparition, telle qu'elle est dérivée d'un immense corpus de textes existants. Ainsi, on pourrait les comparer à un singe savant qui ne taperait pas complètement au hasard, mais qui aurait réduit son champ d'action en éliminant les combinaisons de lettres ou de mots qui n'ont quasiment jamais été observées dans les textes qu'il a analysé. Par exemple, ce singe ne produirait presque jamais dix consonnes consécutives, non pas parce que cela viole une règle sématnique ou de syntaxique explicite, mais simplement parce qu'une telle séquence est extrêmement rare, voire absente, dans les textes qu'il connaît. De cette

manière, le modèle de langue se base sur des patterns statistiques pour générer un texte qui semble cohérent, fluide et conforme aux structures observées, mais cela ne signifie pas qu'il comprend réellement le contenu ou qu'il possède une intention. Il suit des probabilités, non des significations.

Ainsi, on ne prouve pas que les modèles de langues ne comprennent pas ou n'ont pas d'intention, mais on prouve qu'il est possible de produire les même résultats que les humains sans ces qualités. En d'autres termes, on peut produire des textes qui semblent indiscernables de ceux écrits par un humain sans pour autant accomplir la tâche de manière consciente ou intentionnelle. Et plus généralement, l'approximation des résultats ne garantit pas le déploiement fidèle des mécanismes cognitifs ou créatifs à l'origine de ces résultats.

4 Language of Thought Hypothesis

L'hypothèse du langage de la pensée (abréviée LOTH, en anglais) (Rescorla 2024b), popularisée par Jerry Fodor dans les années 1970, soutient que penser implique l'utilisation d'un langage mental interne appelé "mentalais". Ce langage inné permettrait de manipuler des représentations mentales structurées selon des règles syntaxiques et sémantiques. La théorie voit la pensée comme un processus computationnel. Elle vise à expliquer comment des opérations mentales complexes sont possibles sans recours à une langue naturelle. Les défenseurs actuels de cette théorie défendent une certaine compositionalité des représentations mentales ; c'est à dire qu'elles possèdent une sémantique compositionnelle. Les représentations complexes sont formées à partir de représentations plus élémentaires et leur sens dépend à la fois du sens de ces représentations de base et de la manière dont elles sont structurées entre elles. Les versions compositionnelles de l'hypothèse du langage de la pensée s'inscrivent dans la théorie computationnelle de l'esprit classique (Rescorla 2024a). Cette théorie compare l'esprit à une machine de Turing, manipulant des symboles mentaux selon des règles syntaxiques et se veut expliquer la cohérence sémantique du raisonnement humain. LOTH prévoit une distinction entre représentations perceptuelles – traduction des stimuli provenant de l'environnement à travers les sens – et représentation mentales. Toutefois, la ressemblance entre ces deux types de représentation est dure à estimer. Et bien que l'intelligence artifielle se soit développée en puisant son inspiration dans les processus cognitifs humains, les tentatives actuelles d'extraction de concept ou d'une organisation sémantique logiques de ces concepts à partir des représentations internes des réseaux de neurones ne sont pas fructueuses et trouvent deux issues. Certaines tentatives ne présentent simplement pas de résultats positifs tandis que d'autres ont montré que si les modèles artificiels possèdent une capacité à manipuler des concepts de manière structurée et cohérente comme les humains en sont capa-

bles selon LOTH, alors la structure de ces concepts est instable et lacunaire (Ameisen et al. 2025) et aucune méthode générale ne permet de les extraire. Cette absence de preuve d'une sructure logique des concepts n'est pas surprenante dans la mesure où elle n'est pas contrôlée lors de l'entrainement des modèles, lequel se base sur l'apprentissage des correlations de l'apparition de mots de textes variés, sans distinction explicite pour la nature et le contexte de ces textes.

5 General semantics

David Lewis, dans son article de 1970 General Semantics (Lewis 1970), propose une œuvre fondatrice en sémantique formelle. Il distingue deux types de considérations : d'une part, les systèmes sémantiques abstraits qui associent des symboles à des aspects du monde ; d'autre part, les facteurs psychologiques et sociologiques qui expliquent pourquoi des individus ou des populations adoptent un système sémantique donné. Il insiste sur l'importance de séparer ces deux dimensions afin d'éviter la confusion dans la théorie sémantique. Au cœur de l'approche de Lewis se trouve la conception du sens comme conditions de vérité. Il soutient que le sens d'une phrase est déterminé par les conditions dans lesquelles elle serait vraie ou fausse, en prenant en compte des facteurs comme le temps, le lieu ou le locuteur. Cela l'amène à introduire la notion d'intensions : des fonctions qui associent à chaque indice – c'est à dire à chaque ensemble d'informations contextuelles – une valeur de vérité. Ce cadre permet une compréhension plus fine du fonctionnement du sens en contexte. A ce compte, les modèles de langage actuels, qui se basent sur des statistiques de cooccurrence de mots, ne peuvent pas être considérés comme des systèmes sémantiques au sens de Lewis, car ils ne traitent pas le langage en terme de catégorie d'information (temps, lieux ou locuteur) et n'ont aucun égard pour les conditions de vérité. Ils ne font que capturer des régularités dans des données textuelles sans tenter de traiter la portée de leur contenu sémantique. Lewis propose également une théorie compositionnelle du sens, dans le prolongemnet du modèle syntaxique génératif de Chomsky (Chomsky 1965). Il adopte une approche de correspondance dans le cadre de laquelle chaque règle syntaxique correspond à une règle sémantique, ce qui permet une correspondance systématique entre structure syntaxique et signification. Une nouvelle fois, les modèles de langage sont en rupture avec cette approche, déjà parce qu'ils ne traitent pas le langage en terme de structures syntaxiques. Comme nous l'avons noté plus haut, leur méthode de génération de texte est dépourvue de règles formelles et ils cherchent encore moins à établir une correspondance systématique entre syntaxe et sémantique. En somme, aujourd'hui, Lewis serait en opposition frontale avec les affirmations selon lesquelles les modèles de langage font preuve d'une forme de compréhension des textes qu'ils manipulent. En effet, selon lui une théorie du sens ne peut faire l'économie des conditions de vérité des

énoncés et il critique les approches qui se contentent de traduire d'une langue à une autre sans aborder le fond du sens.

6 Conclusion

En somme, si les modèles de langage contemporains ont pu séduire par leurs performances spectaculaires, il reste fondamental de rappeler que leur architecture repose sur des principes très éloignés de ceux qui sous-tendent les réseaux neuronaux biologiques. La tentation de l'analogie entre machine et humain, fréquente dans le discours public et académique, mérite d'être accueillie avec prudence : la similarité des résultats comportementaux ne saurait valoir similarité des mécanismes sous-jacents et il convient de ne pas confondre coïncidence comportementale et équivalence fonctionnelle. C'est précisément ce qui justifie une certaine méfiance face aux prétentions cognitives qu'on attribue parfois aux modèles actuels. Leur incapacité à manipuler des symboles, à raisonner de manière structurée ou à former des représentations stables du monde souligne leur éloignement profond d'une cognition humaine véritable. Le passage du connexionnisme à une manipulation symbolique — c'est-à-dire l'émergence de systèmes capables non seulement de produire des structures langagières, mais de raisonner avec elles — reste un cap non franchi. À ce jour, ces modèles apprennent à générer du langage plausible en maximisant des probabilités conditionnelles, sans développement d'une économie conceptuelle interne maîtrisable ou interprétable.

Dans ce contexte, les théories cognitivistes du langage, comme l'hypothèse du language de la pensée ou la General Semantics de David Lewis, offrent un contraste éclairant. Là où l'hypothèse du language de la pensée postule une grammaire mentale manipulant des symboles clairement définis, les modèles de langage actuels n'ont ni syntaxe interne explicite, ni système symbolique accessible. De même, la General Semantics repose sur une correpsondance entre syntaxe et sémantique et défend la nécessité d'intégrer une notion de vérité pour construire une théorie sémantique sérieuse. Ces conditions sont absentes dans les systèmes artificiels qui ne sont pas ancré dans le monde et n'ont pas de notion de vérité, de référent ou d'engagement sémantique. Ce décalage découle directement de leur mode d'entraînement : les modèles se bornent à apprennent des régularités statistiques dans des corpus de textes.

Des pistes sont néanmoins explorées pour dépasser cette limite. Les approches dites world models, ou encore le projet JEPA (Joint Embedding Predictive Architecture) de Yann LeCun, cherchent à faire émerger une forme de modélisation interne du monde à partir de signaux sensoriels multimodaux (Garrido et al. 2025). Ces approches visent à concilier apprentissage auto-supervisé et formation de représentations utiles à la planification, à l'action, voire au

raisonnement. Si ces tentatives restent encore limitées en pratique, elles signalent une prise de conscience des failles actuelles et un retour possible vers une forme de symbolisme ancré.

References

Ameisen, Emmanuel et al. (2025). "Circuit Tracing: Revealing Computational Graphs in Language Models". In: *Transformer Circuits Thread*. URL: https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

Beniaguev, David, Idan Segev, and Michael London (2021). "Single cortical neurons as deep artificial neural networks". In: *Neuron* 109.17, 2727–2739.e3. ISSN: 0896-6273. DOI: https://doi.org/10.1016/j.neuron.2021.07.002. URL: https://www.sciencedirect.com/science/article/pii/S0896627321005018.

Bouveresse, Jacques (1999). *Prodiges et vertiges de l'analogie: De l'abus des belles-lettres dans la pensée*. Paris: Éditions Liber-Raisons d'agir. ISBN: 2-912107-08-3.

Chen, Boyuan et al. (2024). "Spatialvlm: Endowing vision-language models with spatial reasoning capabilities". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465.

Chollet, Francois et al. (2024). *ARC Prize 2024*. https://kaggle.com/competitions/arc-prize-2024. Kaggle.

Chomsky, Noam (1965). *Aspects of the Theory of Syntax*. English. Special Technical Report. Cambridge, MA: MIT Press, p. 261. ISBN: 0262530074.

Cuadron, Alejandro et al. (2025). "The Danger of Overthinking: Examining the Reasoning-Action Dilemma in Agentic Tasks". In: *arXiv preprint arXiv:2502.08235*.

Garrido, Quentin et al. (2025). *Intuitive physics understanding emerges from self-supervised pretraining on natural videos*. arXiv: 2502.11831 [cs.CV]. URL: https://arxiv.org/abs/2502.11831.

Lewis, David K. (1970). "General Semantics". In: *Synthese* 22.1-2, pp. 18–67. DOI: 10.1007/bf00413598.

Lv, Changze et al. (2024). *Towards Biologically Plausible Computing: A Comprehensive Comparison*. arXiv: 2406.16062 [cs.NE]. URL: https://arxiv.org/abs/2406.16062.

McCoy, R. Thomas et al. (2024). "Embers of autoregression show how large language models are shaped by the problem they are trained to solve". In: *Proceedings of the National Academy of Sciences* 121.41, e2322420121. DOI: 10.1073/pnas.2322420121. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2322420121. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2322420121.

Metz, Cade and Nico Grant (July 2022). “Google Engineer Who Claimed AI Was Sentient Has Been Fired”. In: *The New York Times*. Accessed: 2025-05-28. URL: https://www.nytimes.com/2022/07/23/technology/google-engineer-artificial-intelligence.html.

Rescorla, Michael (2024a). “The Computational Theory of Mind”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Winter 2024. Metaphysics Research Lab, Stanford University.

— (2024b). “The Language of Thought Hypothesis”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Summer 2024. Metaphysics Research Lab, Stanford University.

Shervani-Tabar, Navid and Robert Rosenbaum (2023). “Meta-learning biologically plausible plasticity rules with random feedback pathways”. In: *Nature Communications* 14.1, p. 1805. ISSN: 2041-1723. DOI: 10.1038/s41467-023-37562-1. URL: https://doi.org/10.1038/s41467-023-37562-1.

Turing, Alan (Oct. 1950). “Computing Machinery and Intelligence”. In: *Mind* LIX.236. https://www.jstor.org/stable/2251299, pp. 433–460.

Vaswani, Ashish et al. (2023). *Attention Is All You Need*. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

Wei, Jason et al. (2022). *Emergent Abilities of Large Language Models*. arXiv: 2206.07682 [cs.CL].

Weizenbaum, Joseph (Jan. 1966). “ELIZA—A Computer Program for the Study of Natural Language Communication between Man and Machine”. In: *Communications of the ACM* 9.1. Open access, pp. 36–45. DOI: 10.1145/365153.365168.

Xiong, Siheng et al. (2024). “Large language models can learn temporal reasoning”. In: *arXiv preprint arXiv:2401.06853*.