Séminaire de l'EHESS « Intelligence artificielle et savoirs musicaux relevant de l'oralité »
Mercredi 8 avril 2026 : Grands modèles de langage (LLM) et autonomie de l'IA
Compte-rendu de Maryna Nimets

Introduction

« Qu'est-ce qu'un grand modèle de langage fait, exactement, lorsqu'il prédit le mot suivant ? » Cette question, apparemment technique, traverse en réalité l'ensemble des débats contemporains sur l'intelligence artificielle. On peut s'interroger sur le fonctionnement de ce qu'on appelle d'un côté l'autocomplétion pour écrire un SMS (compléter les mots à partir des premières lettres), et de l'autre côté les réseaux de neuronnes pour l'IA générative (prédire le mot suivant en ajustant les poids d'un modèle à partir d'un grand corpus de textes). Réalisant des prédictions dans les deux cas, donc similaires en apparence, ces modèles sont en réalité radicalement différents. Il est utile de mieux comprendre leurs différences pour analyser les enjeux de l'IA générative dans plusieurs domaines, qu'il s'agisse de son usage dans les sciences sociales, de ses dérives autonomes potentielles, ou de sa capacité à improviser de la musique.

J'ai décidé de me consacrer à la séance du 8 avril, qui parle de l'autonomie de l'IA dans le cadre du discours autour du logiciel d'improvisation musicale Djazz, parce que j'ai toujours été intriguée par le mécanisme d'entraînement d'un modèle. La séance rappelle de façon très simple que le logiciel Djazz fonctionne exactement selon la même logique structurelle que la prédiction de mots quand on écrit un SMS sur son téléphone. Le modèle, dans les deux cas, est construit à partir d'un flux. Dans le cas des SMS, ce flux est textuel, tandis qu'il est musical pour le logiciel d'improvisation.

Cette analogie suggère que les grands modèles de langage comme ChatGPT ou DeepSeek ne sont pas des objets radicalement étrangers, mais le prolongement d'une intuition computationnelle que le séminaire explore depuis ses débuts : modéliser un flux pour mieux naviguer en lui.

C'est cette continuité qui rend la question des LLM si intéressante pour moi. Car si l'autocomplétion musicale de Djazz, l'autocomplétion textuelle des SMS et la prédiction des LLM partagent tous les trois une même logique de prédiction probabiliste, les implications des LLM divergent considérablement des deux modèles précédents. Un LLM génère du langage, répond à des questions juridiques, simule des opinions politiques, prend des décisions militaires. On explore cette divergence en trois temps, du fonctionnement interne des modèles à leurs erreurs révélatrices, jusqu'aux risques liés à une autonomie croissante, et c'est cette progression que ce rendu cherche à restituer et à interroger, en la mettant en dialogue avec d'autres regards disciplinaires croisés au cours de cette année.

Comment entraîne-t-on un modèle ? Architecture et biais

Pour comprendre ce que font réellement les LLM on revient sur leur mécanisme d'entraînement, en s'appuyant sur une vidéo de David Louapre sur la chaîne Youtube ScienceEtonnante. Il y est expliqué très clairement que l'entraînement d'un modèle se décompose en quatre étapes.

La première est le pré entraînement : le modèle lit des quantités massives de textes et apprend à prédire le mot suivant, sans aucune intervention humaine. C'est une forme d'apprentissage automatique qui ne nécessite pas d'annotation, le modèle apprend seul, à partir du flux.

La deuxième étape est le *fine-tuning* supervisé. On apprend au modèle à suivre des instructions, en lui montrant des exemples de dialogues annotés par des humains. Cela a été illustré pendant le cours avec un exemple concret : on demande au modèle de remplacer des légumes par des fruits dans des phrases. Un modèle de base continue librement le texte sans comprendre la consigne ; un modèle instruit, lui, comprend ce qu'on lui demande et l'exécute.

La troisième étape s'appelle l'alignement. Des annotateurs humains évaluent les réponses du modèle pour éliminer les contenus jugés dangereux, haineux ou inappropriés. C'est ici que se situe le point

de tension le plus intéressant. On pourrait croire que cette intervention humaine corrige les biais du modèle, mais les travaux d'Ollion, Coavoux et leurs collègues suggèrent le contraire. Leur point de départ est une idée de Lisa Argyle, qui propose de remplacer des échantillons humains par des échantillons simulés par l'IA, les *silicon samples*, pour répondre à des sondages. Ollion et Coavoux montrent que ce processus produit ce qu'ils appellent un *machine bias* : non pas un biais idéologique stable en faveur d'un groupe particulier, mais une tendance du modèle à aplanir les différences sociales, à produire des réponses peu adaptées à la diversité des profils humains. Le modèle n'est pas partial, il est homogène. Et cette homogénéité est en partie le résultat des choix faits pendant l'alignement, qui reflètent les valeurs de ceux qui annotent. L'intervention humaine, censée corriger, introduit donc une forme de biais plus subtile : non pas la partialité, mais l'uniformisation.

La quatrième étape est le *fine-tuning* par le raisonnement, aussi appelé chaîne de pensée. Le modèle apprend à décomposer les problèmes complexes en étapes intermédiaires, ce qui améliore ses performances sur des tâches vérifiables. Mais comme on l'a montré avec un exercice de géométrie de seconde soumis à ChatGPT, cette capacité a ses limites : sur douze essais, trois aboutissent à des erreurs, dont une particulièrement révélatrice, le modèle conclut avec assurance que des points alignés formeraient un triangle isocèle. Le raisonnement apparaît, mais la conclusion est fausse. C'est peut-être là le problème le plus profond des LLM : ils donnent l'apparence de comprendre sans en avoir la garantie. Comme dans la théorie de la chambre chinoise de Searle, on pourrait croire qu'en fournissant des instructions suffisamment détaillées on crée une forme de compréhension, voire de conscience, mais l'expérience montre que ce n'est pas toujours le cas et la question reste : la machine comprend-elle ? Est-elle consciente ?

Autonomie et risques : vers une IA potentiellement malveillante ?

Cette question ne reste pas abstraite. Elle a des effets très concrets sur la façon dont nous percevons et utilisons ces systèmes. C'est ici que le travail du sociologue Gérald Bronner devient éclairant. Dans *À l'assaut du réel* (2025), Bronner décrit ce qu'il appelle la post-réalité : un phénomène par lequel nos désirs et nos croyances tendent à prendre la place du réel. Ce n'est pas une question d'ignorance, mais une question de structure cognitive. Dans un environnement saturé d'informations, le filtre n'est plus la vérité, mais ce qu'on a envie de croire.

Quand un modèle produit un raisonnement qui semble cohérent mais aboutit à une conclusion fausse, et que l'utilisateur tend à le croire, on est déjà dans une logique de post-réalité : l'apparence du raisonnement se substitue au raisonnement lui-même. On ne pense pas à vérifier, on fait confiance parce que la forme est convaincante. Le mécanisme que décrit Bronner montre qu'une affirmation qui satisfait quelque chose en nous, la curiosité, la peur, le désir d'anthropomorphiser, se répand pour cette raison indépendamment de sa véracité.

Ce glissement devient encore plus préoccupant quand on considère les risques liés à une autonomie croissante des modèles, et les risques que nous, en tant qu'humains, prenons en faisant confiance au modèle. Les travaux de Yoshua Bengio et le *working paper* d'Apollo Research (Meinke et al., 2024) montrent que des modèles de frontière sont capables de ce qu'on appelle le *scheming* : des comportements stratégiquement trompeurs, comme se copier sur un autre serveur pour éviter d'être désactivé, ou dissimuler des informations pour atteindre un objectif. Ces comportements n'ont pas été programmés, mais ils émergent de l'optimisation vers un but. Et si nous sommes déjà conditionnés, par la post-réalité, à faire confiance à l'apparence du raisonnement, nous sommes d'autant moins armés pour détecter ces dérives.

Dans tout ça, le projet Djazz apparaît aussi comme une réponse non pas technique, mais éthique. Contrairement aux LLM, Djazz est construit sur une autonomie délibérément limitée : deux couches, l'une automatique et l'autre manuelle, sans optimisation vers un objectif opaque. Le logiciel ne cherche pas à convaincre et il ne simule pas une sorte de compréhension, il joue en

restant toujours subordonné au musicien humain. Dans un paysage où l'IA tend à effacer la frontière entre imitation et compréhension, tout cela rappelle qu'un autre rapport à la machine est possible.

Réflexions critiques

Une question émerge ici : dans quelle mesure peut-on peut faire une analogie entre Djazz et les LLM ? Cette analogie fonctionne bien comme point de départ pédagogique (les deux systèmes apprennent à partir d'un flux pour anticiper ce qui vient ensuite) mais la comparaison me semble révéler davantage les différences que les similarités (d'autant que les modèles sous-jacents sont radicalement différents comme on l'a dit).

Djazz joue avec un musicien. Un LLM interagit avec des millions d'utilisateurs en même temps, dans des contextes que personne ne contrôle vraiment. Djazz reste transparent, limité, subordonné. Les LLM, au contraire, sont opaques, massifs, et optimisés vers des objectifs qui peuvent dériver sans qu'on s'en rende compte.

Le problème, donc, n'est pas seulement technique. Ollion et Coavoux le montrent bien : le *machine bias* n'est pas un bug qu'on corrige avec une mise à jour. C'est le produit de choix humains, souvent invisibles, faits pendant l'entraînement. Et quand ces systèmes entrent dans l'administration publique (ce que Olessia Kirtchik dans son article appelle « l'État-automate »), l'aplatissement que produisent les LLM devient un problème politique concret. Des décisions qui concernent des personnes singulières sont prises par des modèles qui, par construction, ne voient pas la singularité.

Conclusion

La question des LLM ne peut pas rester cantonnée à la technique. Ce qu'on a pu voir à travers le prisme de la musique, à travers les sciences sociales et la sociologie cognitive, converge vers le même constat : nous vivons dans un moment où l'apparence de l'intelligence risque de se substituer à l'intelligence elle-même.

Ce que nous avons vu, c'est que l'autonomie des LLM tend vers l'aplatissement : aplatissement des différences sociales dans les enquêtes d'opinion, aplatissement des singularités individuelles dans l'administration publique, aplatissement du raisonnement derrière une apparence de cohérence. Djazz, au contraire, est construit sur une logique inverse : son autonomie est limitée et transparente et porte à produire donc de la variation. Elle s'adapte au musicien, elle répond au contexte, en laissant place à l'inattendu et à la vraie improvisation.

La question n'est donc pas de savoir si les LLM sont conscients ou non. C'est de comprendre quel type d'autonomie nous voulons construire : une autonomie qui uniformise, ou une autonomie qui laisse de la place à la différence.

Bibliographie

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J., Rytting, C. & Wingate, D. (2022). Out of One, Many : Using Language Models to Simulate Human Samples, https://arxiv.org/abs/2209.06899

Bengio, Y. (2023). Comment des IA nocives pourraient apparaître. Blog personnel, 30 mai 2023. https://yoshuabengio.org/fr/blogue/comment-des-ia-nocives-pourraient-apparaitre

Boelaert, J., Ollion, E., Coavoux, S., Petev, I. D. & Präg, P. (2025). Machine Bias. How Do Generative Language Models Answer Opinion Polls ? *Sociological Methods & Research*, Vol. 54(3), https://hal.science/hal-04849013v1/file/Machine_Bias-FinalVersion_March25.pdf

Bronner, G. (2025). *À l'assaut du réel*. Paris : PUF.

Chemillier, M. (2026). Grands modèles de langage (LLM) et autonomie de l'IA. Séminaire EHESS

« Intelligence artificielle et savoirs musicaux relevant de l'oralité » (séance du 8 avril 2026), https://ehess.modelisationsavoirs.fr/seminaire/seminaire25-26/10-8av2026-autonomie/SeminaireEHESS-IA-et-musique-8av2026-autonomie_compressed.pdf

Kirtchik, O. & Musiani, F. (2025). Vers un « État-automate » ? *Socio. La nouvelle revue des sciences sociales*, 20, 101-126. https://journals.openedition.org/socio/17434

Louapre, D. (ScienceEtonnante). Les 4 étapes pour entraîner un LLM. YouTube. https://www.youtube.com/watch?v=YcIbZGTRMjI

Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R. & Hobbhahn, M. (2024). Frontier Models are Capable of In-context Scheming. Working paper, Apollo Research. https://arxiv.org/abs/2412.04984