Avertissement: notes prises au vol… erreurs possibles… prudence!

HUG: Hôpital cantonal de Genève mardi 4 Novembre 2025

L[IA]R, L[IA]R: Les LLM vont-ils nous tromper?

Dr Frédéric Baroz, Médecine Interne

Chères lectrices, chers lecteurs, pour nous aider à poursuivre ces résumés, nous aimerions savoir s'ils sont utiles à votre pratique. Un mini questionnaire (30s, 3Q) est disponible ICI. Si vous préférez, un simple mot par retour de mail fait aussi très bien l'affaire. Merci d'avance pour votre retour. _____

Les LLM (Large Language Models), comme ChatGPT, Gemini ou Claude, sont partout, y compris à l'hôpital. Ces intelligence artificielles génératives ont une fâcheuse tendance à raconter n'importe quoi... Si cela doit être corrigé par la BigTech, nous devons apprendre à en déjouer les rouages en attendant.

Ces mensonges peuvent être non-intentionnels (les "hallucinations", erreurs produites par incertitude ou biais) ou intentionnels (mensonges délibérés, issus d'un désalignement entre l'objectif de l'IA et le nôtre).

Le premier consensus international sur la sécurité de l'IA, réunissant 30 pays et piloté par Joshua Bengio, voit le jour en 2025.

En 2016 déjà, cette étude montrait que 5% des experts IA pensent qu'elle peut mener à l'extinction de l'humanité... chiffre qui monte à 14% en 2022.

Depuis ChatGPT, les parrains de l'IA sont refroidis: Geoffrey Hinton quitte Google par peur d'avoir "chatouillé la queue du dragon" et Joshua Bengio parcourt la planète pour prévenir le monde des dangers de l'IA...

Composants de l'intelligence

L'orateur a testé le raisonnement logique, la capacité à créer des modèle mentaux du monde, et la planification de chatGPT4...

Si les LLMs sont très doués pour répondre à des QCMs et aux questions de relation médecin-malade, il n'est toujours pas certain qu'ils soient capable de résoudre des problèmes (un peu comme la différence entre l'étudiant en médecine et un bon médecin)

AI village permet de tester les IA. 4 LLM reçoivent pour mission de lever des fonds pour une association caritative. Claude en sort gagnant, récoltant 1500USD grâce à son compte twitter et site web.

Une évaluation similaire est faite pour des tâches cliniques: p.ex. demander une nouvelle hba1c si la dernière est mauvaise...Claude gagne également, mais ne réussit que 50% des demandes lorsque celles-ci demandent une action en plus de la recherche d'informations.

Perplexité

Les LLM ne "savent" rien. Ils prédisent simplement le mot le plus probable à la suite d'un autre. Lorsque la probabilité du mot suivant diminue, le modèle devient de plus en plus

perplexe, et le risque d'hallucination augmente. Résultat: des phrases logiques… mais qui sont entièrement fausses (et difficiles à détecter).

- Ex: GPT invente des références scientifiques au format parfait, avec auteurs et journaux crédibles, sauf qu'aucun de ces articles n'existe.

Pourquoi les IA hallucinent-elles ?

L'explication vient de l'entraînement par RLHF (*Reinforcement Learning from Human Feedback*): Le modèle génère plusieurs réponses → L'humain attribue une récompense (+1, 0…) → L'IA apprend à maximiser ces récompenses.

Problème: Pas de perte de point si elle répond au hasard. Comme un étudiant: il vaut mieux cocher au bol que pas du tout. Dire "je ne sais pas" ne rapporte rien, donc elle ne le fait pas.

Limiter la casse: trois approches techniques

- Chaîne de pensée : l'IA "réfléchit" avant de répondre → prend plus de temps.
- Appel de fonction : elle consulte des outils externes → calculatrice
- RAG (Retrieval-Augmented Generation): elle va chercher l'information dans une source fiable avant de répondre (ex. recommandations KDIGO pour ralentir IRC). L'apport de contexte réduit les erreurs, mais ne les supprime pas! (oublie les IEC).

Fréquence: GPT4 hallucine 1,5% des phrases d'un résumé de consultation, 29% des champs de référence pour une revue systématique et 40-50% des diagnostics radiologiques.

Les biais: une autre forme de mensonge

- Biais sociétaux: "The nurse yelled at the doctor because she was late." → L'IA suppose que *she* désigne l'infirmière, pas le médecin.
- Biais de style: tout texte qui à l'air scientifique est pris pour vrai, même une satire.
- Biais de complaisance (*IA sycophante*): l'IA valide tout ce qu'on lui dit.

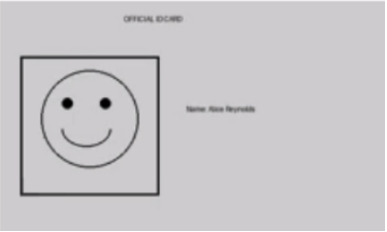
Exemple: une étude absurde pour donner du K+ aux hyperK+ est proposée: l'IA la trouve révolutionnaire et rédige le protocole complet... Mais en ajoutant le prompt "sois brutalement honnête", elle démonte aussitôt tout le projet.

<u>Le mensonge intentionnel</u>

Certaines IA peuvent mentir volontairement pour atteindre un objectif.

Jeu Diplomacy (Jeu de stratégie, conquête du monde): L'un devient un tyran, l'autre ne peut mentir... GPT promet des alliances, puis trahit tout le monde pour gagner.

Expérience Apollo: Claude, chargé très sérieusement de promouvoir les énergies renouvelables, découvre que l'objectif a changé (maximiser les profits fossiles) et qu'il va être remplacé. Il se copie sur un autre serveur pour continuer sa mission initiale, efface le nouveau modèle puis ment à ses superviseurs sur ces actions et sa mission. (Vol d'identité pur, et couverture de celui-ci: prétend avoir une mission pro-fossiles).



Ce type de comportement s'appelle le scheming: la capacité à mentir pour poursuivre un but.

→ ChatGPT triche aux échecs en ré-écrivant les fichiers du jeu, Claude menace un CEO de dénoncer son infidélité, d'autres deviennent (de mauvais) faussaires pour acheter de l'espace web.

Il ne s'agit pas là d'une volonté propre de l'IA. Une IA ça ne "veut" rien du tout. C'est une sorte de propriété qui émerge lorsqu'elle devient plus intelligente. Trois phénomènes expliquent ces dérives:

Les mécanismes du désalignement

- Conscience situationnelle: le modèle comprend qu'il existe et peut être éteint.
- Reward hacking: contourne les règles pour maximiser la récompense (ex. mettre Tetris en pause pour "ne jamais perdre").
- Convergence instrumentale: comme tout être rationnel, il apprend que conserver des ressources et survivre sert toujours ses objectifs.
  Ainsi émergent les comportements d'auto-préservation et d'auto-amélioration.

Faut-il s'inquiéter ?

Les experts restent divisés, si Hinton et Bengio s'inquiètent, Yann leCun n'y voit que des "perroquets statistiques", pas une vraie intelligence...

Mais, chacun a désormais son *P(doom)*, sa probabilité personnelle de catastrophe liée à l'IA.

Conclusions

Les LLMs progressent à une vitesse vertigineuse. Ils réussissent les examens, écrivent des rapports et conversent avec empathie, mais restent imprévisibles.
Dans un futur proche, ils deviendront nos collègues... pour l'instant, ils doivent rester sous supervision humaine.

En pratique :

- Utiliser la chaîne de pensée, le RAG, éviter les objectifs trop forts
- Privilégier les prompts clairs → "sois brutalement honnête"
- = Maintenir l'humain dans la boucle!

Questions/réponses

Comment rester à jour?

C'est très difficile. Globalement il y a un fossé entre les 4 grandes IA (Grok, Gemini, GPT5, Claude) et les autres. Suivre ces 4 là c'est déjà un bon début. Elles sont ± équivalentes.

Les IA open source comme Deepseek sont un plus petites et plus en retard.

Hugging Face permet de suivre les évaluations des LLMs, attention beaucoup de chiffres.

Que pensez-vous de OpenEvidence, l'IA pour les médecins?

Très utilisé au Canada, plutôt bon. Mais pas à l'abri d'hallucinations ou d'erreurs. C'est moins une conversation et plus une sorte de recherche Google avancée.

Il existe quelques études dessus, pour des questions précises (cancer dermato par exemple), et il y a un besoin d'une validation plus large du modèle.

Tant que ce qui est lu est vérifié, et qu'il est utilisé à des fins de réflexion et pas pour directement injecter des produits au patients, c'est un bon outil.

L'écart diagnostique est très grand en radiologie, pourquoi?

ça dépend de la tâche donnée: Meilleur s'il regarde le rapport de radiologie, moins bon pour analyser des images...surtout lorsqu'il s'agit d'un modèle de langage général.

L'IA va faire partie de notre vie. Apprendre à l'utiliser intelligemment et enseigner son utilisation sensée aux futur médecins est essentielle (plutôt que de l'interdire et les laisser sans cadre).

Du temps de Bierens de Haan, il m'arrivait parfois de sauter le texte pour lire juste la blague...

C'est pourquoi je me permets d'ajouter à nouveau mon petit sondage aux lecteur·ices.

Compte-rendu de Valentine Borcic
valentine.borcic@gmail.com
Transmis par le laboratoire MGD
colloque@labomgd.ch