Par Antoine LORIN,
Product Owner, StarClay

Une IA Générative peut-elle dégénérer ?

Aujourd'hui, les Intelligences Artificielles (IA) génératives apprennent à créer des images ou du texte à partir d'autres textes ou à partir de textes décrivant des images disponibles sur internet.

Mais quel serait l'impact sur la qualité de générations, si la majorité des textes et images disponibles en ligne était elle-même générée par des intelligences artificielles ?

C'est ce qu'une équipe de chercheurs en informatique (computer science) d'Oxford a cherché à déterminer.

Une IA générative peut-elle dégénérer ?

Si les modèles LLMs (large language models) sont basés sur des modèles dont l'apprentissage est déjà fait, cela leur permet de faire moins d'erreurs. Mais si ChatGPT 2 entraine ChatGPT3.4 qui entraine ChatGPT4 qui lui-même entraine ChatGPTo1, c'est un peu comme si nous clouions le pied d'une IA, lui demandant d'avancer et d'apprendre à partir de sa position précédente.

Vulgarisons

- Un LLM, c'est une méthode qui consiste à demander à un enfant de reformuler la littérature du XIXème siècle, puis de demander à second enfant de reformuler la reformulation du premier enfant, et ce à l'infini.

- Un modèle GMM (Gaussian mixture models) ou VAE (variational autoencoders), c'est une méthode qui consiste à demander à un premier enfant de reformuler la littérature du XIXème siècle, puis à un deuxième enfant de reformuler cette même littérature, en y ajoutant les reformulations des enfants qui ont fait l'exercice précédemment.

L'équipe de chercheurs d'Oxford a donc poussé les deux méthodes dans ses retranchements, voyons le résultat. Nous conserverons ci-après la métaphore des méthodes d'apprentissage de la littérature de nos charmants chérubins :

- **Dans une première méthode**, ils ont entrainé leurs « enfants » sur 5 générations, chaque enfant reformulant à partir de la reformulation de l'enfant précédent.
- **Dans une deuxième méthode**, ils ont entrainé leurs enfants sur 10 générations, chaque enfant reformulant sur la base de 90% des reformulations des enfants précédents, et de 10% aléatoires de la base de connaissance originale.

Sans surprise, quel que soit le cas, au mieux, le dernier enfant n'est plus capable de réaliser une reformulation aussi correcte que celle qu'aurait faite le premier enfant, au pire celui-ci répète continuellement les mêmes phrases.

Pire encore, lorsque l'on sermonne les « enfants » en leur indiquant qu'ils ne doivent pas se répéter, ceux-ci commencent à dire n'importe quoi encore plus tôt.

Quittons notre exemple des enfants et passons à la réalité

Les premières IA génératives ont appris sur une base de connaissance saine : générée uniquement ou presque par des humains.
Mais de plus en plus de personnes aujourd'hui utilisent des modèles d'IA génératives pour les aider à rédiger du contenu, sans compter les médias d'information qui ne se cachent même plus d'utiliser à 100% des textes générés par IA pour remplir leurs sites.
La part « empoisonnée » (i.e. générée par une IA dont le modèle s'effondre) de la base de connaissances utilisée pour entrainer les IA ne va donc qu'aller croissante.

S'il est difficile – sinon impossible – de déterminer la part de l'internet générée par IA, ou simplement inspirée par de l'IA (seul Chuck Norris ayant lu l'intégralité d'internet deux fois), nous pouvons malgré tout prendre le risque de généraliser des exemples à moindres échelles.

Prenons le réseau social professionnel LinkedIn :
une étude de originality.ai estime que 54% des publications longues sont générées par une IA.IA pour remplir leurs sites.


Voici la version mise à jour pour 2025, toujours avec l'esprit *boomer cringe overload* :

🌟🚀 **2025 : Let's boost our impact and scale new heights!** 🚀🌟

Dear team,

👏✨ Alors que nous kickstartons cette nouvelle année, je tiens à saluer votre mindset disruptif 💡 et votre capacité à deliver des résultats exceptionnels en 2024. 🙌🎯 Cette année, nous allons **upgrader nos ambitions** 🆙 et **redéfinir notre value proposition** à tous les levels. 💪🌍

🗓️ Notre roadmap 2025 est ambitieuse : 💼 co-création 🤝, activation de nouvelles synergies 🔗, et maximisation de nos insights 🧠 pour générer un impact 360° sustainable 🌿. Chaque projet 🌈 sera une opportunité de penser out of the box 📦, de créer du flow 🌊 et d'aligner nos KPIs 📊 avec nos objectifs stratégiques. ⏳✨

🔥 Je vous invite à **embrasser le challenge** 💪, à être des game-changers 🎮 et à incarner notre DNA d'innovation 🚀. Ensemble, nous allons scaler nos initiatives 📈, transformer nos pain points 💔 en opportunités 🌟 et drive l'excellence à tous les touchpoints 📍. 🌟🌟

💖 Merci de rester engagés et purpose-driven. Let's co-build a legendary 2025! 🌈🌟🌍

💼 [Nom du dirigeant]
🚀 [Poste du dirigeant]

Si nous généralisons le cas LinkedIn à l'ensemble d'internet, le volume de textes générés par IA augmente très sensiblement, et donc la part de l'internet dans son ensemble sur laquelle les nouvelles générations d'IA vont s'appuyer pour apprendre.

Ce qui augmente le risque que les IA s'empoisonnent de plus en plus avec leurs propres textes.

L'IA n'est donc pas un problème en elle-même, le bât blesse plutôt sur la base d'apprentissage.
Il est très couteux d'entrainer une IA à chaque fois que l'on met à jour l'algorithme, qui plus est, il faut que la base d'apprentissage soit suffisamment fiable (et donc humaine) pour que le modèle ne s'effondre pas.
Afin de répondre à ce problème, Google – conscient des biais – a pondéré les sources réputées d'origine humaines plus fortement dans sa base d'apprentissage, ce qui constitue peut-être une première piste pour conserver une bonne qualité de génération par IA.

Toutefois, une telle pondération est-elle une solution viable ou n'est-ce qu'un frein à un inévitable effondrement des modèles ? Quelle serait la part maximale de données générées par IA pour entrainer une autre IA sans prendre le risque d'empoisonner les modèles ?

Autant de questions qu'il serait bon de se poser lorsque l'on conçoit des IA génératives. Sans oublier évidemment le principal sujet : l'humain.

Celui-ci saura-t-il identifier un début d'effondrement des modèles génératifs pour les corriger à temps, ou n'est-il – tel une IA – que la dernière génération des enfants de notre exemple : déjà intellectuellement effondré ?

Définitions, précisions méthodologiques et concepts utilisés dans cet article

« Généré par IA » originality.ai considère dans ce terme tout ce qui a été créé par IA et modifié ou non par un humain, ainsi que tout ce qui a été écrit par un humain et corrigé par l'IA.

« Généré humainement » originality.ai considère dans ce terme tout ce qui a été uniquement écrit et édité par un humain ou écrit par un humain basé sur un résultat de recherche généré par une IA.
Cette nuance se justifie par le fait que le moteur de recherche google par exemple est une forme d'IA.

Il est à noter ici qu'un angle mort existe quant aux textes écrits par des humains et légèrement corrigés par une IA, du fait qu'il est difficile de les détecter.

« Effondrement du modèle » : processus de dégénérescence des générations d'apprentissage menant à la perte d'informations.

Deux facteurs mènent à l'effondrement d'un modèle :

- L'oubli : L'IA oublie totalement des éléments fondamentaux de sa base d'apprentissage
- L'empoisonnement de la donnée : les données deviennent corrompues et mènent à des comportements inattendus dans les réponses générées

Littérature du XIXème siècle : ce siècle n'est pas choisi ici au hasard, il est celui des révolutions, politiques, sociales et intellectuelles qui cherchent une fracture totale avec le passé, tout en oubliant qu'une révolution complète est étymologiquement le retour d'un corps à son état initial.

Références

https://www.nature.com/articles/s41586-024-07566-y
https://originality.ai/blog/ai-content-published-linkedin
https://fr.wikipedia.org/wiki/Grand_mod%C3%A8le_de_langage
https://fr.wikipedia.org/wiki/Mod%C3%A8le_de_m%C3%A9lange_gaussien
https://fr.wikipedia.org/wiki/Auto-encodeur_variationnel
https://intelligence-artificielle.developpez.com/actu/347237/L-IA-serait-sur-le-point-detransformer-l-Internet-en-un-veritable-cauchemar-un-outil-a-double-tranchant-pour-le-webcreativite-ou-manipulation/

starclay

www.starclay.com
hello@starclay.com