Pôle d'expertise de la régulation numérique

Éclairage sur…

Avril 2023
#06

ChatGPT ou la percée des modèles d'IA conversationnels

GPT-4, Stable Diffusion, GitHub Copilot : de nombreux modèles génératifs d'intelligence artificielle ont vu le jour ces derniers mois. Grâce à leurs performances et leur ergonomie, ils ne sont plus réservés aux seuls experts mais se sont ouverts au grand public. Parmi eux, les modèles dits « conversationnels » ont suscité de nombreuses réactions. Capables de répondre à des requêtes formulées en langue naturelle, ils peuvent mener une conversation avec un agent humain, quel que soit le niveau de technicité. Leur popularité est déjà avérée (ChatGPT compte plus de cent millions d'utilisateurs actifs mensuels) et leurs applications s'annoncent nombreuses.

Comment sont construits ces modèles conversationnels ? Quelle place occupent-ils dans l'écosystème du Traitement Automatique des Langues (TAL) ? En quoi sont-ils innovants ? Ce n°6 d'« Éclairage sur… » invite à une plongée au cœur de la technologie de ces dispositifs pour en saisir les principaux enjeux et limites. Combinant plusieurs briques techniques pré-existantes, ces modèles d'IA ne se révèlent pas si révolutionnaires comparés aux modèles antérieurs en TAL, mais innovent plutôt en matière d'accessibilité. Leur conception soulève de nombreuses questions sur le respect de la vie privée, la propriété intellectuelle ou l'ouverture de la science : là où les modèles GPT sont développés comme de véritables produits, des modèles plus respectueux et ouverts pourraient aussi trouver leur place. L'éducation des utilisateurs permettrait de dépasser l'engouement autour de quelques modèles seulement et de véritablement démocratiser leurs capacités.

L'ESSENTIEL EN UNE PAGE

Les réseaux de neurones ont bouleversé de nombreux domaines de recherche en informatique, dont le Traitement Automatique des Langues (TAL ou *Natural Language Processing*, NLP). **Modèles informatiques et statistiques** entraînés à l'aide de gros volumes de données, **leurs résultats prédictifs dépendent entièrement de la distribution de ces données.**

Les réseaux de neurones de TAL ont pris la forme de *Large Language Models* (LLM). Leur entraînement est très coûteux en énergie (l'entraînement de GPT-3, ancêtre de ChatGPT et de GPT-4, a consommé autant d'énergie que 275 foyers français en un an) et a souvent recours à d'énormes volumes de données issues du web ou d'œuvres littéraires libres de droit. **Le web contient des données très variées, y compris du contenu haineux ou sensible, des données propriétaires ou sous licence, ou encore des données personnelles. Les LLM qui sont entraînés sur de telles données peuvent présenter des risques en matière de conformité**.

Les LLM conversationnels, dont font partie ChatGPT et GPT-4, permettent aux LLM d'être utilisés par le plus grand nombre pour des tâches plus variées. Ces modèles prédisent le texte le plus vraisemblable au vu des attentes d'un utilisateur et des données d'entraînement. Ils héritent pour cette raison des biais des agents humains qui ont participé à leur entraînement. Les données nécessaires à l'entraînement de ces modèles sont parfois labellisées par des travailleurs de pays en développement, qui doivent par conséquent lire de grands volumes de contenus sensibles. Malgré les précautions des acteurs du domaine, **des utilisations néfastes des LLM conversationnels restent souvent possibles.**

Le coût d'entraînement des LLM, tant en énergie qu'en données conformes, de qualité et en quantité suffisante, représente une barrière à l'entrée importante. Ainsi, les LLM les plus performants actuellement sont souvent développés par des entreprises qui ne les rendent pas toujours disponibles de façon ouverte, au motif d'inquiétudes quant à leurs utilisations néfastes. **Des chercheurs développent toutefois des LLM ouverts.**

Les LLM conversationnels vont probablement transformer de nombreux domaines, dont le marché des moteurs de recherche. Dans certaines circonstances, il est bien plus simple pour un utilisateur de lire une réponse synthétique en langue naturelle que de consulter une ou plusieurs pages, aussi pertinentes soient-elles. **Cependant, les LLM conversationnels sont aujourd'hui incapables de restituer quelles sources ont été utilisées pour générer une réponse. Ils peuvent également « halluciner », c'est-à-dire inventer des réponses vraisemblables mais fausses.**

Le règlement européen en matière d'intelligence artificielle, en cours de discussion, se dirige vers une définition des systèmes d'intelligence artificielle à usage général dont feraient partie les LLM. Ce règlement a vocation à détailler les obligations des fournisseurs en matière de transparence et de conformité.

AU CŒUR DES MODÈLES CONVERSATIONNELS : LES RÉSEAUX DE NEURONES

Réseau de neurones, késako ?

Un réseau de neurones est un modèle reposant sur des hypothèses statistiques et des règles, qui s'entraîne à partir de grands volumes de données. Cet entraînement tend à imiter celui d'un cerveau : les données fournies en entrée traversent plusieurs couches de « neurones » avant de fournir un résultat. Chaque neurone contient des **paramètres**, c'est-à-dire des nombres qui permettent via des calculs simples de détecter des motifs récurrents (ou *patterns*) dans les données d'entrée. Un algorithme, dit de rétropropagation de l'erreur, indique ensuite au réseau si le résultat est juste afin qu'il puisse ajuster les paramètres dans les neurones (cf. Figure 1). Cette opération d'inférence puis de rétropropagation est effectuée de très grands nombres de fois sur de très nombreux exemples, afin d'obtenir de solides performances sur des nouvelles données. En ce sens, **le modèle s'entraîne à avoir raison aussi souvent que possible au regard de la distribution statistique des données d'entrée**.

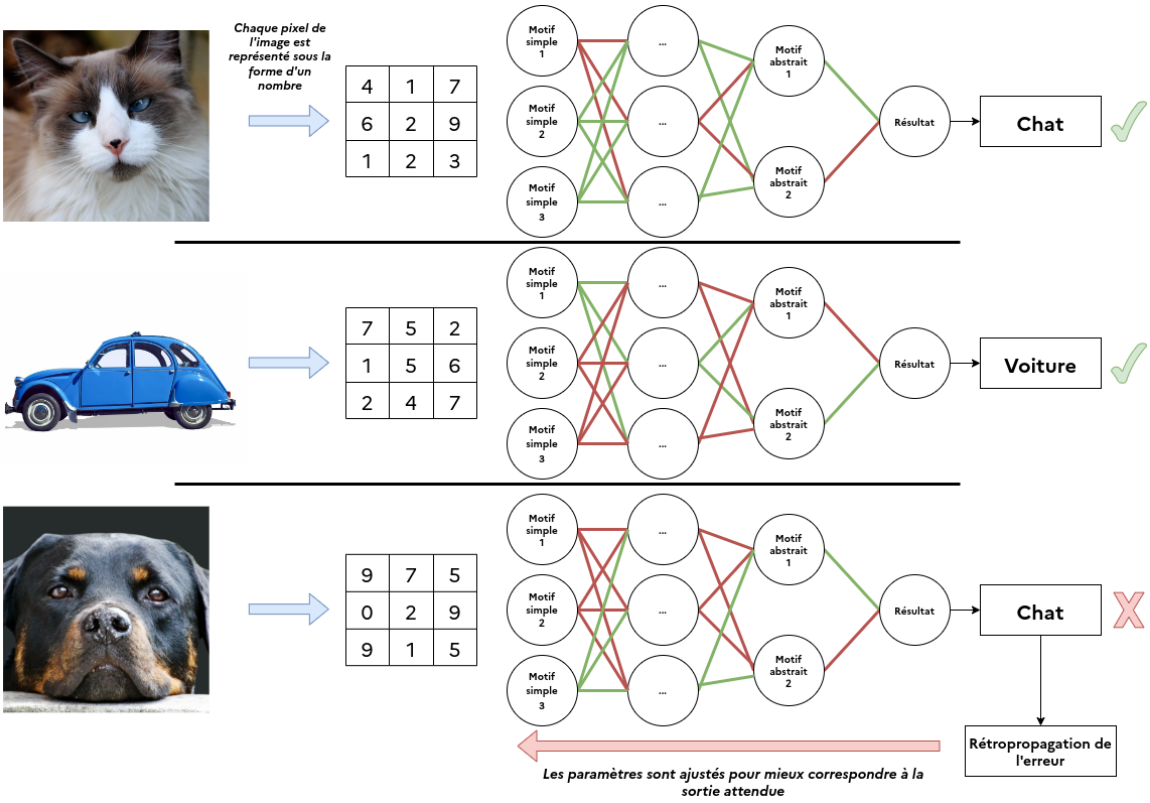

Figure 1 : Représentation simplifiée de l'entraînement d'un réseau de neurones classifieur.
Chaque couche détecte des motifs récurrents (ou *patterns*) à partir de la couche précédente. Plus les couches sont situées profondément au sein du réseau, plus elles traitent des motifs abstraits. Les neurones se spécialisent par eux-mêmes : on ne précise jamais que pour différencier une voiture d'un chat, la présence de roues est un indice. C'est cette spécialisation sans supervision humaine qui rend les réseaux de neurones difficilement interprétables.

La structure d'un réseau de neurones est appelée **architecture**. Elle permet de reconstruire un réseau de neurones identique, en indiquant le nombre de couches, le nombre de neurones dans chaque couche, et enfin le type de ces neurones. Dans la Figure 1, l'architecture consiste en une première couche de trois neurones, plusieurs couches non détaillées, une avant-dernière couche avec deux neurones et enfin une couche de résultat avec un seul neurone. Cependant, imiter l'architecture d'un modèle ne permet pas toujours d'obtenir des résultats comparables : ce sont

les données vues lors de l'entraînement qui sont au cœur des performances du modèle.
En 2012, un réseau de neurones nommé AlexNet a remporté pour la première fois un challenge important de classification d'images. Cet événement marque le début de la révolution des réseaux de neurones : ceux-ci permettent d'obtenir de meilleurs résultats que les algorithmes antérieurs dans de nombreux domaines, dont le Traitement Automatique des Langues.

Réseaux de neurones et Traitement Automatique des Langues

L'une des premières innovations majeures apportées par les réseaux de neurones dans le domaine du Traitement Automatique des Langues (ou *Natural Language Processing*, NLP) a été le plongement lexical (ou *word embedding*). Ce mécanisme permet à un réseau de neurones de s'entraîner à produire une représentation mathématique de mots sans supervision humaine. Cette représentation permet d'effectuer des opérations mathématiques sur les mots :

Roi – Homme + Femme = Reine
Paris – France + Ukraine = Kyiv

Le principe consiste à prédire un mot à partir de son contexte (comme un texte à trous, cf. Figure 2), ou l'inverse, c'est-à-dire le contexte à partir d'un seul mot. Le problème de cette approche réside dans l'absence de mesure de l'importance des mots dans le contexte : le modèle fait l'hypothèse que tous les mots du contexte ont la même importance. Par exemple, dans la phrase « le chat a poursuivi le rat, puis il l'a mangé », le pronom « il » fait référence au chat, mais on ne peut pas simplement l'inférer à partir des mots du contexte (il pourrait aussi faire référence au rat). Un modèle qui n'est pas capable de traiter ce type de phrases complexes est nécessairement limité lorsqu'il s'agit de traduire ou de résumer des textes en langue naturelle.

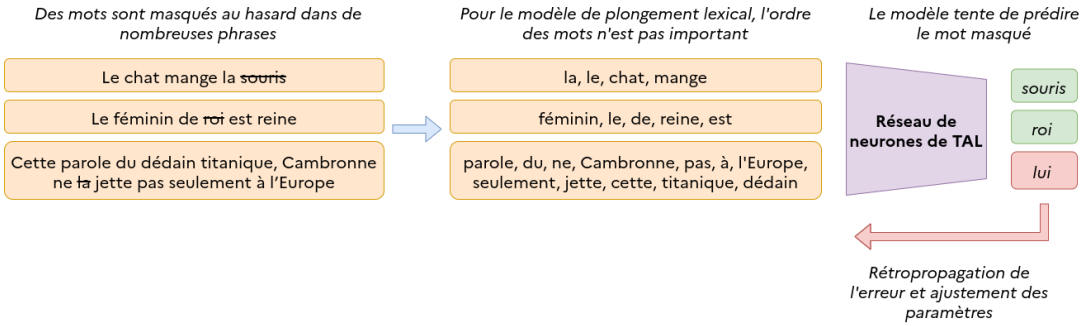

Figure 2 : Représentation simplifiée de la méthode d'entraînement des premiers réseaux de neurones de plongement lexical.
Le modèle ne prend pas en compte l'ordre des mots du contexte. Il est par conséquent en difficulté face à des phrases complexes.

Du mécanisme de l'attention aux *Large Language Models (LLM)*

Le mécanisme de l'attention permet de résoudre le problème des phrases complexes. Le modèle s'entraîne à reconnaître l'importance des mots du contexte en fonction du mot considéré. Reprenons l'exemple précédent, « le chat a poursuivi le rat, puis il l'a mangé ». Lorsque le modèle va considérer le mot « il », l'attention sera portée sur le mot « chat ». Ce mécanisme, proposé par Google en 2017 dans un

article de recherche devenu célèbre[1], est la dernière grande avancée du domaine. Il donne naissance aux architectures appelées *Transformers* (cf. Figures 3 et 4).

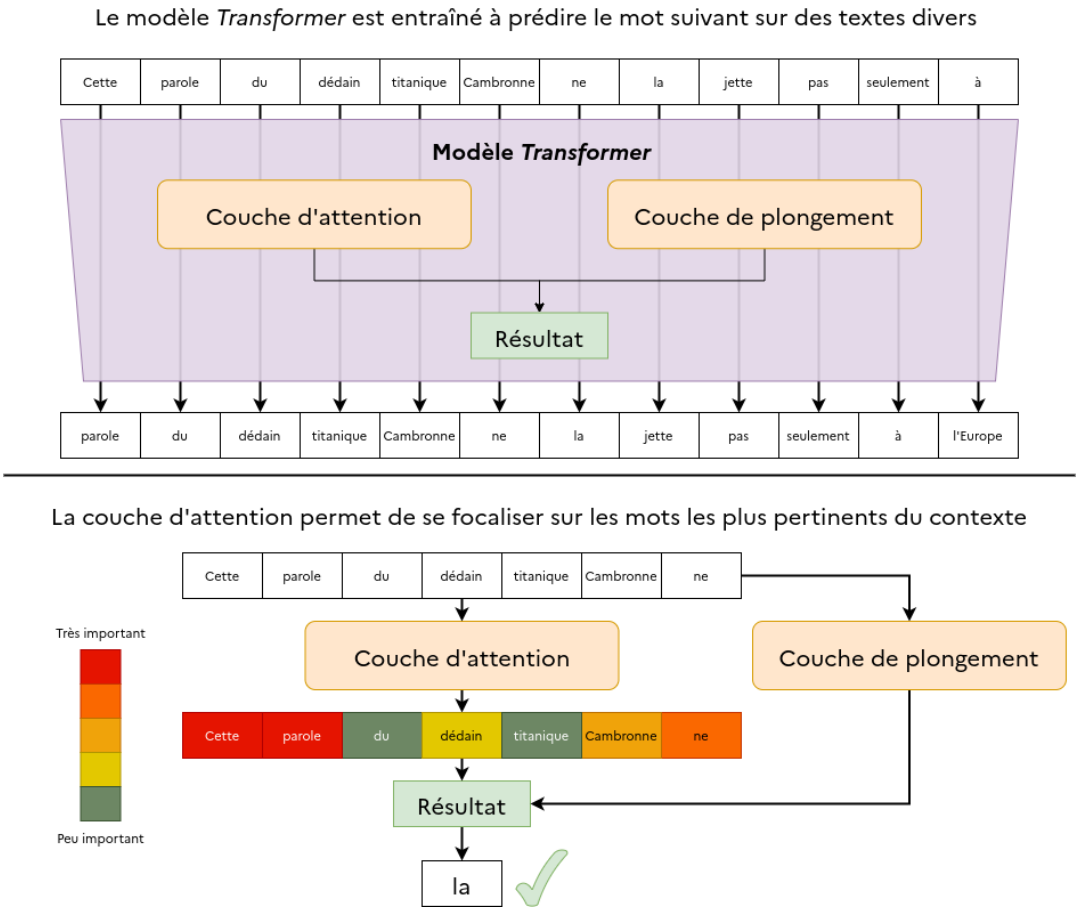

Figure 3 : Représentation simplifiée du fonctionnement d'un modèle Transformer.
La couche d'attention permet de se focaliser sur les parties importantes de la phrase en fonction du mot à prédire.

Les architectures *Transformers* ont rapidement pris la forme d'énormes modèles de langage, ou *Large Language Models* (LLM) dont les contraintes techniques sont fortes :

- des milliards de paramètres sont nécessaires ;
- d'énormes corpus doivent être constitués ;
- l'entraînement d'un seul de ces modèles est très coûteux en temps (plusieurs jours à plusieurs semaines), en matériel (plusieurs dizaines voire centaines de processeurs de pointe) et par conséquent en énergie. À titre d'exemple, le coût de l'entraînement de GPT-3, un LLM d'OpenAI, a été estimé à 1287 MWh[2], soit l'équivalent de la consommation énergétique annuelle moyenne de 275 foyers français.

À l'instar du plongement lexical, l'entraînement initial des LLM consiste à fournir au modèle le début d'une phrase ou d'un texte et à le laisser prédire le mot suivant en utilisant, cette fois, le mécanisme de l'attention, c'est-à-dire en prenant en compte l'importance relative des mots dans leur contexte. Cette opération de prédiction est répétée pour des millions (voire parfois des milliards) d'exemples. Les données d'entraînement des LLM proviennent d'internet (pages web variées, Wikipedia en

[1] https://arxiv.org/pdf/1706.03762.pdf
[2] https://arxiv.org/pdf/2104.10350.pdf

intégralité dans plusieurs langues), d'ouvrages ou de données textuelles internes aux entreprises développant ces modèles.

Comme tout réseau de neurones, un LLM « prédit » le résultat (en l'occurrence le mot suivant) le plus vraisemblable au vu de la distribution statistique des données d'entraînement. Cette capacité à générer du texte vraisemblable en fonction du contexte peut directement être utilisée à des fins malveillantes : plusieurs laboratoires et entreprises refusent de publier leurs modèles afin d'éviter qu'ils ne soient utilisés pour désinformer ou offenser.

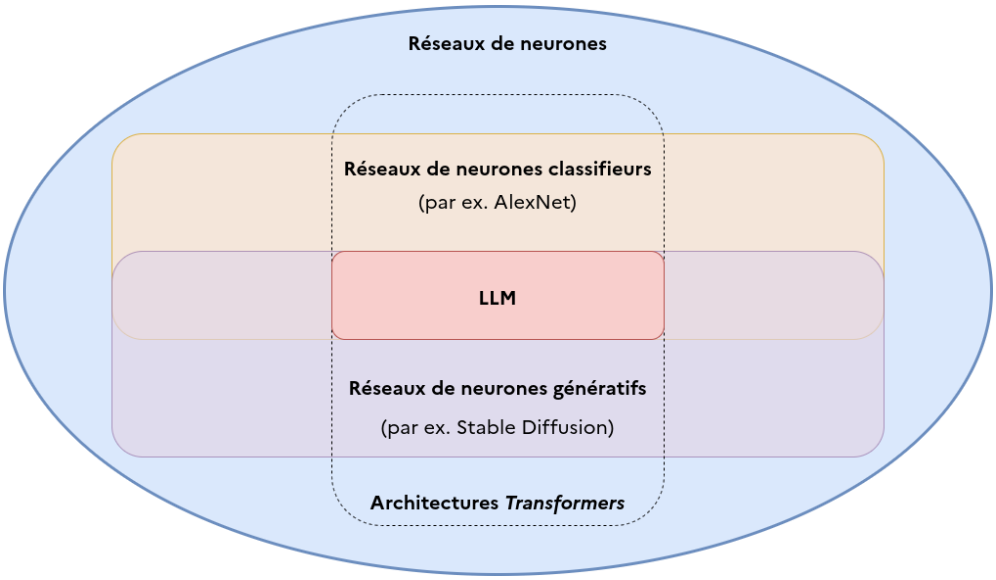

Figure 4 : Représentation schématique de la place des Large Language Models (LLM), dont font partie les modèles GPT et BLOOM, au sein de l'écosystème des réseaux de neurones. Les LLM sont à l'intersection des architectures *Transformers*, des réseaux classifieurs (c'est-à-dire qui segmentent en différentes catégories les objets fournis en entrée, comme AlexNet) et des réseaux génératifs (c'est-à-dire capables de créer du contenu, comme Dall-E ou Stable Diffusion pour la génération d'images, ou ChatGPT pour la génération de texte).

DE LA PERFORMANCE À L'ERGONOMIE : LES LLM CONVERSATIONNELS

Malgré la simplicité apparente de la tâche d'entraînement initiale (prédire le mot suivant sur de larges volumes de texte), les LLM peuvent être ré-entraînés par la suite pour des tâches plus spécifiques sur des volumes de données bien moindres et obtenir des performances excellentes. **Ils peuvent alors traduire avec justesse des textes (y compris techniques), résumer des documents ou des ouvrages, répondre à des questions ou encore générer du code ou du texte**. Aujourd'hui, les LLM sont déjà intégrés dans de nombreux outils : traduction automatique, suggestions de réponses dans les applications de messagerie ou de courriels… Cependant, ils montrent deux handicaps majeurs les empêchant d'être utilisés directement par le grand public pour des tâches plus générales :

- ré-entraîner un LLM pour une tâche spécifique demande une certaine compréhension de leur fonctionnement ;
- prédire le texte le plus vraisemblable est différent de prédire le texte conforme aux attentes de l'utilisateur. Certaines questions ou requêtes peuvent mettre les LLM en difficulté : les données issues d'internet ou de Wikipedia ne permettent pas au modèle « d'apprendre » à répondre aux attentes précises d'un utilisateur.

Pour surmonter ces handicaps, entreprises et chercheurs du domaine se sont attelés au développement de nouveaux modèles : des **LLM conversationnels, soit des modèles capables de mener une conversation avec n'importe quel utilisateur humain.** Les méthodes de développement des modèles conversationnels (ou *chatbots*) s'apparentent à des stratégies pour réussir à simuler un humain, c'est-à-dire réussir le « jeu de l'imitation » (*imitation game*) ou Test de Turing.

En 1950, Alan Turing dévoile cette première méthode d'évaluation de l'intelligence d'une machine dans son article de recherche *Computing Machinery and Intelligence*[3] Le principe du test peut être résumé ainsi : « la machine peut-elle réussir à mener une conversation aussi intelligemment qu'un humain ? ». Turing ne propose pas d'évaluer la conscience d'un programme informatique mais de qualifier d'intelligent un modèle capable de tromper un agent humain. Ainsi ELIZA, premier modèle ayant réussi le test de Turing, et développé par le MIT dans les années 1960[4], se contentait de reformuler les entrées de l'utilisateur sous forme d'interrogations, à la manière d'un piètre psychologue.

Depuis ELIZA, les modèles conversationnels ont largement évolué. Parmi les plus récents rendus accessibles ou détaillés dans des articles de recherche, citons ChatGPT[5], LaMDA[6], Bard[7] (version de LaMDA qui sera intégrée dans le moteur de recherche Google) ou encore BlenderBot 3[8]. Si chacun de ces modèles se distingue par ses particularités, ils partagent une même brique technologique ayant permis de séduire le grand public : l'**« apprentissage par renforcement à l'aide de retours humains »** (ou RLHF pour *Reinforcement Learning from Human Feedback*[9]). Le RLHF consiste à entraîner un modèle afin qu'il parvienne à reproduire un jugement humain, puis à utiliser ce modèle comme un « professeur » capable d'évaluer d'autres modèles. La Figure 5 détaille ce processus.

ZOOM sur l'apprentissage par renforcement : des modèles qui apprennent en s'amusant ?

L'apprentissage par renforcement consiste à entraîner un modèle à jouer à un « jeu ». Ici, le jeu désigne à la fois un environnement régi par des règles et une récompense liée aux actions du modèle. Cette définition du jeu s'applique autant à de véritables jeux, comme le Go ou les jeux vidéos, qu'à des situations réelles comme un robot s'entraînant à marcher, à courir, à sauter… Le modèle « apprend » à adapter ses actions à l'état de l'environnement pour maximiser sa récompense, c'est-à-dire son score. AlphaGo, premier modèle à avoir battu des champions du jeu de Go, avait été entraîné grâce à l'apprentissage par renforcement. L'apprentissage par renforcement à l'aide de retours humains (RLHF) est une méthode particulière d'apprentissage par renforcement qui consiste à transformer des préférences humaines en score, comme le montre la Figure 5.

[3] https://redirect.cs.umbc.edu/courses/471/papers/turing.pdf
[4] https://dl.acm.org/doi/10.1145/365153.365168
[5] https://arxiv.org/pdf/2203.02155.pdf
[6] https://arxiv.org/pdf/2201.08239.pdf
[7] https://blog.google/technology/ai/bard-google-ai-search-updates/
[8] https://arxiv.org/pdf/2208.03188.pdf
[9] https://proceedings.neurips.cc/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf

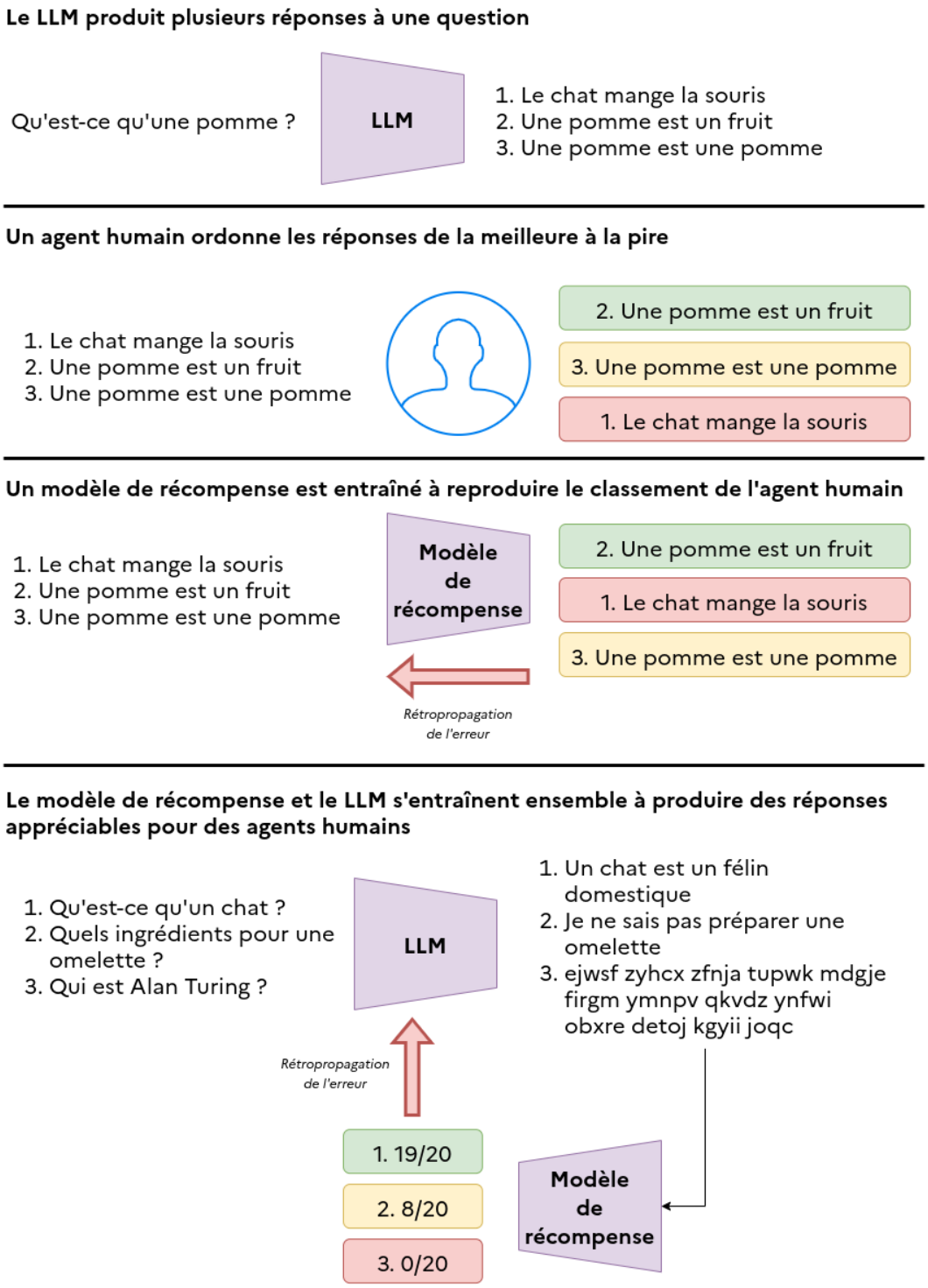

Figure 5 : Utilisation de l'apprentissage par renforcement à l'aide de retours humains pour l'entraînement de LLM conversationnels.
La tâche de classer les réponses entre elles et celle de les noter sont analogues : s'entraîner à la première permet à l'inférence de réussir la seconde.

Ainsi, **les LLM conversationnels sont optimisés afin de satisfaire le plus possible les utilisateurs humains**. Cette nouvelle aptitude leur permet de mener une conversation de façon très vraisemblable et de prendre en compte les attentes des utilisateurs. Cette approche présente elle aussi plusieurs limites :

- **Il existe une différence importante entre la satisfaction de l'utilisateur et la véracité des réponses**. Le modèle tend à s'aligner sur la position de l'utilisateur lorsque celui-ci fait savoir qu'il n'est pas satisfait de la réponse fournie. De plus, les LLM conversationnels sont entraînés sur des corpus labellisés par des agents humains. Ces agents humains labellisent et évaluent leurs échanges à l'aune de leurs propres idées, croyances et stéréotypes. Ces biais doivent être reconnus : le modèle propose une certaine vision du monde et influencera à son tour les utilisateurs.
- L'optimisation de la satisfaction des utilisateurs ne limite pas la génération de contenus haineux ou illégaux. **Si un utilisateur souhaite générer du contenu haineux, le modèle va *a priori* tenter de s'y conformer**. Les LLM conversationnels embarquent des classifieurs afin de détecter des questions ou des réponses sensibles, ce qui permet de limiter certaines utilisations malveillantes.

Malgré l'engouement des utilisateurs pour les LLM conversationnels, leurs applications s'inscrivent dans la continuité de ce qu'avaient déjà permis les modèles antérieurs de LLM en Traitement Automatique des Langues. **Les LLM conversationnels génèrent du texte comme n'importe quel LLM, mais sont davantage capables de s'adapter à tous les types d'utilisateurs.** L'innovation consiste donc principalement à rendre ces modèles :

- plus ergonomiques ;
- moins nocifs, c'est-à-dire à limiter leur capacité à générer du contenu faux ou sensible.

Toutefois, **les briques logicielles qui visent à limiter les capacités nocives des modèles sont d'autant plus difficiles à construire que les modèles sont performants : les objectifs de limitation et d'adéquation aux attentes de l'utilisateur sont contradictoires**. À titre d'exemple, le récent article présentant GPT-4[10], dernier né de la firme OpenAI, revendique une amélioration en matière de véracité des réponses ainsi qu'une diminution des « comportements incorrects » face à du contenu « sensible » ou « interdit ». Cependant, les données relatives à ces métriques ou les définitions de « comportement incorrect » ou « contenu sensible ou interdit » ne sont pas publiques. Des acteurs indépendants comme NewsGuard[11] ont évalué le modèle à l'aide de leurs propres protocoles. Ils concluent que GPT-4 s'avère moins sûr que ChatGPT.

DE L'IMPORTANCE CAPITALE DES DONNÉES D'ENTRAÎNEMENT DANS LA CONCEPTION DES LLM

Autre facteur clé de succès et d'évaluation d'un modèle d'intelligence artificielle : les données utilisées pour son entraînement. Leur qualité et leur volume sont le socle de la performance de tout réseau de neurones. Les LLM, qui ont recours à d'énormes corpus dans de très nombreuses langues, en sont l'illustration. Cependant, ces énormes corpus peuvent mener à des écueils.

[10] https://cdn.openai.com/papers/gpt-4.pdf

[11] https://www.newsguardtech.com/fr/misinformation-monitor/march-2023/

L'entraînement des LLM implique souvent le *Common Crawl*[12], un ensemble très vaste de pages récoltées sur internet sans filtre, disponible gratuitement. La taille et la variété de cet ensemble permet aux modèles d'atteindre d'excellentes performances, mais il peut contenir des données personnelles, sous licence ou protégées par les droits d'auteur[13].

Plusieurs procès ont déjà été intentés aux États-Unis contre des entreprises proposant des modèles génératifs pour des litiges liés à la propriété intellectuelle[14]. Les LLM pourraient être utilisés pour générer de faux textes « dans le style » d'un auteur, ou pire, plagier des œuvres vues durant l'entraînement (par exemple une nouvelle publiée par un auteur sous certaines licences et qui ferait partie des données du *Common Crawl*). Les LLM ne sont pas non plus capables de citer les sources à l'origine des textes générés. Enfin, il est possible qu'un attaquant parvienne à accéder via ces modèles à des données personnelles présentes à l'entraînement ou partagées par des utilisateurs durant leurs interactions[15].

Si les LLM sont souvent capables de générer des textes dans de nombreuses langues, leur exactitude et la richesse de leur vocabulaire dépendent des données d'entraînement. Ce problème est accentué dans le cas des LLM conversationnels : à titre d'exemple, LaMDA (Google) a été entraîné sur un corpus contenant plus de 90% de données en langue anglaise, et BlenderBot 3 (Meta) sur un corpus exclusivement anglais. De son côté, ChatGPT (OpenAI) peut utiliser des anglicismes erronés lorsqu'il génère du texte en français (il peut par exemple utiliser le mot « condition » pour parler d'une maladie) et est incapable de générer du texte dans certaines langues répertoriées comme « En danger » par l'UNESCO[16].

Enfin, les LLM conversationnels nécessitent un surcroît de données : d'une part, des données ordonnées par préférence par des agents humains (cf. Figure 5) ; d'autre part, des données labellisées comme sensibles ou non, afin de limiter les capacités du modèle à générer du contenu haineux ou illégal. OpenAI a par exemple fait appel à des travailleurs kényans pour labelliser des corpus d'entraînement[17]. Ces travailleurs ont dû lire de grands volumes de contenus sensibles. La question de la labellisation des données est récurrente en apprentissage machine, et il n'est pas rare que celle-ci soit déléguée à des travailleurs de pays en développement. Les articles de recherche qui présentent LaMDA et BlenderBot 3 mentionnent aussi un recours à des microtravailleurs (ou *crowdworkers*). Les acteurs du domaine sont parfois réticents à proposer un accès ouvert à leurs données car l'opération de labellisation peut demander un investissement conséquent.

[12] https://commoncrawl.org/

[13] La Commission Européenne considère que la législation européenne en vigueur relative à l'utilisation d'œuvres pour l'entraînement de modèles d'IA constitue un bon compromis entre protection des auteurs et innovation, comme le montre cette réponse à une question posée par un député européen.

[14] https://www.phonandroid.com/lia-midjourney-attaquee-en-justice-aux-etats-unis.html et https://www.lemondeinformatique.fr/actualites/lire-premiere-action-judiciaire-contre-github-copilot-88522.html

[15] Une enquête de l'éditeur de sécurité Cyberhaven montre que de nombreux utilisateurs partagent déjà des données confidentielles ou personnelles avec ChatGPT : https://www.cyberhaven.com/blog/4-2-of-workers-have-pasted-company-data-into-chatgpt/ . Les interactions avec ChatGPT sont utilisées par OpenAI pour entraîner les nouvelles versions du modèle. Ces données peuvent être divulguées par le modèle ou l'interface lors de fuites de données, comme ç'a déjà été le cas une première fois : https://www.01net.com/actualites/chatgpt-divulgue-donnees-sensibles-utilisateurs.html.

[16] https://unesdoc.unesco.org/ark:/48223/pf0000189451

[17] https://time.com/6247678/openai-chatgpt-kenya-workers/

LES MODÈLES DE LLM À L'ÉTAT DE L'ART SONT-ILS « DÉMOCRATISÉS » ?

Bien que les LLM conversationnels représentent une innovation indéniable en matière d'accessibilité de ces technologies, les modèles à l'état de l'art sont rarement disponibles de façon ouverte. Si leurs développeurs exposent souvent des inquiétudes quant à des utilisations néfastes des modèles les plus performants[18], la question de leur ouverture reste centrale : peut-on véritablement parler de démocratisation si les LLM les plus performants sont la propriété d'un nombre restreint d'acteurs ? Les exemples suivants, récapitulés dans le Tableau 1, illustrent ces différentes politiques d'ouverture :

- Fondé en 2015 sous la forme d'un laboratoire à but non-lucratif dont les brevets et la recherche auraient vocation à être ouverts, OpenAI est devenu en 2019 un acteur privé dont la série de modèles GPT est le produit phare. Ses nombreuses versions, incluant la dernière-née GPT-4, sont disponibles via des API (interfaces de programmation) payantes. Les différents ré-entraînements du modèle lui ont permis de rester compétitif face aux modèles concurrents pourtant plus récents. Les modèles GPT sont les seuls LLM à l'état de l'art disponibles exclusivement sous la forme de produits. L'article[19] de recherche présentant GPT-4, bien que revendiquant la dimension innovante du modèle vis-à-vis de ses prédécesseurs, ne révèle aucune information précise quant à son architecture, son entraînement (données, infrastructure ou méthode) ou son coût énergétique. Dans cette publication, OpenAI s'affranchit des pratiques du domaine qui consistent à détailler la méthode afin de permettre à la communauté scientifique de comprendre en quoi celle-ci est une avancée technologique ; et choisit de présenter uniquement des performances.
- Les LLM de Google présentés ici, GLaM[20] et PaLM[21], constituent des avancées en matière de recherche sur le passage à l'échelle des LLM, c'est-à-dire l'entraînement de LLM encore plus volumineux. Ils requièrent des architectures matérielles très particulières et très coûteuses. PaLM est accessible via une API payante au sein de la plate-forme Google Cloud. Le modèle conversationnel de Google, LaMDA, n'est pas disponible directement, mais il semble capable de réussir le jeu de l'imitation. Ces LLM permettent à Google de proposer des fonctionnalités de génération intelligente de texte via l'espace de travail Google pour quelques bêta-testeurs.
- Dans le cadre de son engagement pour une IA plus responsable, Meta rend ses modèles disponibles de façon ouverte. C'est le cas d'OPT[22] (et de son pendant conversationnel BlenderBot 3) un modèle de langage avec le même nombre de paramètres que GPT-3. Cela concerne aussi LLaMA[23], un modèle au nombre de paramètres plus faible mais dont les performances sont pourtant comparables à celles de GPT-3.

[18] https://www.numerama.com/tech/1307322-nous-avions-tort-en-devoilant-gpt-4-openai-dit-rejeter-lopen-source-desormais.html

[19] cf. supra, page 9

[20] https://arxiv.org/pdf/2112.06905.pdf

[21] https://arxiv.org/pdf/2204.02311.pdf

[22] https://arxiv.org/pdf/2205.01068.pdf

[23] https://arxiv.org/pdf/2302.13971.pdf

– En Europe, le collectif BigScience, composé de plusieurs centaines de chercheurs, a entraîné le LLM BLOOM[24] sur le supercalculateur Jean-Zay de Paris-Saclay. Ce modèle est entraîné à réaliser les mêmes tâches que GPT dans 46 langues naturelles (y compris des langues régionales ou en danger) et 13 langages de programmation. Les ensembles de données utilisées pour l'entraînement sont tous disponibles en open-source, tout comme le modèle entraîné via *HuggingFace*[25]. Le modèle comprend 175 milliards de paramètres, soit autant que GPT-3.

	Année	Nb. maximal de paramètres (en mds)	Architecture publique	Modèle entraîné ouvert	Données ouvertes	Conversationnel (RLHF)	Accessible aux utilisateurs (UI ou API)
BigScience BLOOM	2022	175	✓	✓	✓	✗	✓
Google GLaM/PaLM	2021 / 2022	1200 / 540	✓	✗	✗	✗	✓ (API payante)
Google LaMDA/Bard	2022	137 / ?	✓/ ✗	✗	✗	✓	✓ (UK/US)
Meta OPT	2022	175	✓	✓	✓	✗	✗
Meta BlenderBot3	2022	175	✓	✓	✓	✓	✓ (US)
Meta LLaMA	2023	65	✓	✓	✓	✗	✗
OpenAI GPT-3	2020	175	✓	✗	✓ (1re version seulement)	✗	✓ (option payante)
OpenAI GPT-3.5 (InstructGPT /ChatGPT)	2022	175 / ?	✓ / ✗	✗	✗	✓	✓ (option payante)
OpenAI GPT-4	2023	?	✗	✗	✗	✓	✓ (API et UI payantes)

Table 1 : Exemples de LLM à l'état de l'art et des politiques d'ouverture associées

LLM CONVERSATIONNELS : UN CHANGEMENT DE PARADIGME POUR LES MOTEURS DE RECHERCHE

Les LLM ont déjà métamorphosé les pratiques dans de nombreux domaines. Leur évolution conversationnelle transformera à son tour certains usages. Le marché des moteurs de recherche apparaît comme l'un des premiers secteurs impactés par ces développements.

La tâche d'un moteur de recherche consiste le plus souvent à ordonner des résultats externes par pertinence vis-à-vis d'une requête fournie par un utilisateur. À l'origine, les moteurs de recherche étaient d'autant plus efficaces que l'utilisateur connaissait leur fonctionnement. De nos jours, les acteurs du marché ont développé des outils en mesure de s'adapter à tous les types d'utilisateurs.

[24] https://arxiv.org/pdf/2211.05100.pdf

[25] Il est possible de télécharger BLOOM ici : https://huggingface.co/bigscience/bloom et de l'essayer ici : https://huggingface.co/spaces/huggingface/bloom_demo

Les modèles conversationnels offrent une solution différente : plutôt que de lier une requête à des contenus externes qu'il s'agit d'ordonner, le modèle synthétise une réponse en langue naturelle. Il est dès lors plus simple et plus rapide pour l'utilisateur de lire cette synthèse que de consulter un à un des résultats proposés par un moteur de recherche traditionnel.

Les principaux moteurs de recherche semblent conscients de la révolution que représentent ces technologies. Microsoft et DuckDuckGo ont déjà intégré une version de ChatGPT dans leur solution. Google, de son côté, a commencé des essais pour son modèle Bard. Baidu et Yandex, qui proposent également des moteurs de recherche (respectivement en Chine et en Russie), ont mentionné le développement de leurs propres LLM conversationnels et une intégration prochaine dans leurs produits.

Certains acteurs, dont Yann LeCun, VP and Chief AI Scientist chez Meta et figure importante de la communauté scientifique de l'apprentissage machine, émettent des réserves quant à la capacité des LLM conversationnels à remplacer les moteurs de recherche classiques :

- ces modèles peuvent produire des réponses, totalement ou partiellement, incohérentes ou fausses ;
- à l'instar des LLM originels, ils ne sont pas capables de restituer les sources à l'origine de leurs réponses.

Ces défauts sont intrinsèques à la méthode d'entraînement des modèles linguistiques. **Le contenu généré est le plus statistiquement vraisemblable au vu des textes utilisés à l'entraînement, textes qui peuvent contenir des erreurs ou des manipulations et qui sont le produit des opinions de leurs auteurs. Toutefois, nous pouvons imaginer que les utilisateurs de ces nouveaux outils préféreront leur efficacité à leur exactitude.**

QUELLE RÉGULATION POUR LES LLM ?

De nombreux experts et des personnalités appellent à un moratoire sur les LLM plus performants que GPT-4[26]. Par-delà les enjeux économiques des entreprises qui entraînent ces modèles, on peut se demander si une telle mesure empêcherait les usages néfastes déjà rendus possibles par les modèles existants. On peut également douter que ce délai permettrait véritablement d'évaluer les modèles actuels en l'absence de nouvelles obligations de transparence. En ce sens, **la proposition de règlement européen des systèmes d'intelligence artificielle, ou *AI Act*[27] devra prendre en compte les enjeux de souveraineté et d'innovation, tout en assurant la protection des consommateurs et des citoyens.**

Dans sa version en discussion proposée par la Commission européenne, l'*AI Act* contraindrait les fournisseurs de LLM conversationnels à la transparence. **La proposition stipule qu'un humain qui aurait une conversation avec un de ces modèles devrait en être informé ou pouvoir l'inférer du contexte d'utilisation**. En revanche, il circonscrirait les obligations de transparence liées à la génération de contenus (hypertrucages ou *deepfakes*) aux seuls contenus audio et vidéo. Ainsi, **les contenus textuels générés par les LLM, bien que « [ressemblant] sensiblement à un contenu authentique », ne seraient pas accompagnés d'obligation de transparence**.

[26] https://futureoflife.org/open-letter/pause-giant-ai-experiments/

[27] https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:52021PC0206

Enfin, **cette proposition n'inclut pas les modèles génératifs, auxquels appartiennent les LLM, dans la définition des « systèmes d'IA à haut risque »**[28].

Dans son orientation générale[29], le **Conseil de l'Union Européenne propose d'ajouter une définition des « systèmes d'IA à usage général »** au projet de règlement. Cette définition inclurait les LLM, et *a fortiori* les LLM conversationnels. Les fournisseurs de tels systèmes seraient soumis à des obligations particulières, distinctes des obligations applicables aux systèmes à haut risque. **Le Parlement Européen semble poursuivre ses travaux dans cette direction, en distinguant toutefois les systèmes d'IA à usage général reposant sur des « modèles de fondation »** (ou *Foundation models*, c'est-à-dire tout modèle entraîné sur de larges volumes de données qui peut servir à un grand nombre de tâches) des systèmes reposant sur des modèles plus simples.

À terme, cette proposition de règlement encadrera les LLM en tant que produits, mais n'a pas pour objectif de traiter de la protection des utilisateurs. Ces derniers sont pourtant au centre du succès des LLM conversationnels. Bien que ces modèles ne représentent pas une innovation technique importante, leur offre de service a permis une célébrité éclair : **les LLM conversationnels les plus simples d'utilisation, disponibles via une interface graphique intuitive, semblent être les plus plébiscités.** Ainsi, la popularité des modèles paraît décorrélée de leurs performances pures et davantage liée à leur *design* en tant que produit. De nouveaux acteurs pourraient dès lors s'imposer, car les attentes des utilisateurs se concentreront plutôt sur l'intégration et l'accessibilité des LLM que sur leurs performances théoriques. **Des LLM conversationnels ouverts commencent déjà à apparaître[30]. Bien que leurs performances n'égalent pas celles des modèles propriétaires, ils démontrent que les compétences techniques nécessaires au développement de tels modèles ne sont pas l'apanage de quelques entreprises.**

La principale limite au développement de concurrents aux LLM conversationnels les plus populaires réside dans l'accessibilité de données de qualité et en quantité suffisante. Afin d'éviter qu'un faible nombre d'acteurs reste seul capable d'entraîner de tels modèles, des données spécifiques à la tâche de conversation devront être disponibles de façon ouverte, en se plaçant dans la continuité des travaux de nombreux chercheurs dont les corpus *open-source* ont servi à l'entraînement de BLOOM et d'autres LLM ouverts. OpenAI obtient chaque jour gratuitement de nouvelles conversations grâce à ChatGPT et les utilise pour améliorer leurs produits payants. **L'un des enjeux majeurs à venir sera celui de l'éducation des utilisateurs à ces nouveaux outils, afin qu'ils soient en situation de choisir avec quel modèle interagir, c'est-à-dire à qui offrir leurs données de conversation.**

[28] Ces systèmes désignent les produits d'intelligence artificielle qui pourraient porter sérieusement atteinte aux droits fondamentaux, comme par exemple les systèmes d'identification biométrique.

[29] https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf

[30] *Together Research Computer*, une organisation regroupant des chercheurs, a par exemple développé OpenChatKit. À l'instar de ChatGPT, les échanges avec OpenChatKit sont utilisés pour continuer à entraîner le modèle et limiter les utilisations malveillantes. Toutefois, les données ont en l'occurrence vocation à être ouvertes.

Un groupe de chercheurs de Stanford a ré-entraîné LLaMA 7B à l'aide de l'API payante d'OpenAI : le modèle a appris à imiter ChatGPT. Il obtient des performances équivalentes à celles du modèle d'OpenAI, avec un nombre bien plus faible de paramètres ; et il est quant à lui ouvert.

La collection « Éclairage sur… » du PEReN propose des éléments d'analyse techniques sur des thèmes liés à la régulation des plateformes numériques.

Dépôt légal : Octobre 2022
ISSN (en ligne): 2824-8201

Service à compétence nationale, le Pôle d'expertise de la régulation numérique (PEReN) fournit, aux services de l'État et autorités administratives intervenant dans la régulation des plateformes numériques, une expertise et une assistance technique dans les domaines du traitement des données, des data sciences et des procédés algorithmiques. Il s'investit également dans des projets de recherche en science des données à caractère exploratoire ou scientifique.

PEReN – 120 rue de Bercy, 75572 Paris Cedex 12 - contact.peren@finances.gouv.fr