ChatGPT, modèles de langage et données personnelles : quels risques pour nos vies privées ?

Gaspard Berthelier, Antoine Boutet

ChatGPT, modèles de langage et données personnelles : quels risques pour nos vies privées ?

Gaspard Berthelier, Antoine Boutet
Univ. Lyon, INSA-Lyon, Inria, CITI, Lyon, France
gaspard.berthelier@student-cs.fr, antoine.boutet@insa-lyon.fr

Les grands modèles de langage [1] ont récemment attiré beaucoup d'attention, notamment grâce à l'agent conversationnel ChatGPT [2]. Cette plate-forme est devenue virale en seulement quelques mois et a déclenché une course effrénée pour développer de nouveaux modèles de langage toujours plus efficaces et puissants, rivalisant avec l'humain pour certaines tâches.

Cette croissance phénoménale est d'ailleurs jugée dangereuse par de nombreux acteurs du domaine [3], qui plaident pour une pause afin d'avoir le temps de débattre sur l'éthique en IA et de mettre à jour les réglementations.

Une des grandes questions qui se pose est l'articulation entre intelligence artificielle et vie privée des utilisateurs. En particulier, les prouesses des grands modèles de langage sont dues à un entraînement intensif sur d'énormes ensembles de données, qui contiennent potentiellement des informations à caractère personnel, car il n'y a pas d'obligation d'anonymiser les données d'entraînement.

Il est alors difficile de garantir en pratique que le modèle ne compromet pas la confidentialité des données lors de son utilisation. Par exemple, un modèle pourrait générer des phrases contenant des informations personnelles qu'il a vues pendant sa phase d'entraînement.

Apprendre à imiter le langage humain

Les modèles de traitement du langage sont une famille de modèles basés sur l'apprentissage automatique (machine learning en anglais), entraînés pour des tâches telles que la classification de texte, le résumé de texte et même des chatbots [4].

Ces modèles apprennent d'une part à encoder les mots d'une phrase sous la forme de vecteurs, en tenant compte de l'ensemble du contexte. Dans les phrases "J'ai mangé une orange" et "Son manteau orange est beau", le mot

[1] https://theconversation.com/fr/topics/modeles-de-langage-133746
[2] https://theconversation.com/chatgpt-pourquoi-tout-le-monde-en-parle-197544
[3] https://futureoflife.org/open-letter/pause-giant-ai-experiments/
[4] https://theconversation.com/fr/topics/chatbots-67347

"orange" se verra attribuer deux encodages différents, puisque la position et le sens ne sont pas les mêmes.

Ces modèles apprennent également à décoder ces ensembles de vecteurs contextualisés et leurs relations, pour générer de nouveaux mots. Une phrase est générée séquentiellement, en prédisant le prochain mot en fonction de la phrase d'entrée et des mots prédits précédemment.

L'architecture de ces modèles peut être spécialisée pour certaines tâches. Par exemple, les modèles de type BERT [1] sont souvent affinés en apprenant sur des données spécialisées, par exemple sur des dossiers de patients pour développer un outil de diagnostic médical, et sont plus performants sur des tâches de classification de texte tandis que les modèles GPT sont utilisés pour générer de nouvelles phrases. Avec l'essor des applications exploitant les modèles de langage de langage, les architectures et les algorithmes d'entraînement évoluent rapidement. Par exemple, ChatGPT est un descendant du modèle GPT-3, son processus d'apprentissage ayant été étendu pour se spécialiser dans la réponse aux questions.

Confidentialité des informations utilisées pendant la phase d'entraînement du modèle

Les modèles de traitement du langage naturel ont besoin d'une quantité énorme de données pour leur entraînement. Pour ChatGPT par exemple, les données textuelles du web tout entier ont été récoltées pendant plusieurs années [2].

Dans ce contexte, la principale préoccupation en matière de confidentialité est de savoir si l'exploitation de ces modèles ou les informations qu'ils produisent peuvent dévoiler des données personnelles ou sensibles utilisées pendant la phase d'apprentissage et "recrachées" ou inférées pendant la phase d'utilisation.

Considérons d'abord les chatbots (exploitant les modèles de type GPT) qui ont appris à générer des phrases à partir d'un texte d'entrée. D'un point de vue mathématique, chaque mot est prédit séquentiellement, sur la base de probabilités qui auront été apprises durant la phase d'entraînement.

Le problème principal est que des données potentiellement personnelles peuvent parfois constituer la réponse la plus probable. Par exemple, si le modèle a vu la phrase "Monsieur Dupont habite 10 rue de la République" et qu'on lui demande "Où habite Monsieur Dupont ?", le modèle sera naturellement enclin à répondre l'adresse de celui-ci. Dans la pratique, le modèle aura aussi vu de nombreuses phrases de la forme "X habite à Y" et on s'attend plutôt à ce qu'il réponde des connaissances générales plutôt que des adresses spécifiques. Néanmoins, le risque existe et il est nécessaire de pouvoir le quantifier.

Évaluer les probabilités de fuites de données

Il existe tout d'abord des techniques pour évaluer en amont de l'entraînement final si des phrases rares ont le potentiel d'être anormalement mémorisées par le

modèle. On réalise pour cela des micro-entraînements, avec et sans ces phrases, et l'on se débarrasse de celles qui auraient une influence trop grande.

Mais les gros modèles de traitement du langage naturel sont non déterministes et très complexes de nature. Ils sont composés de milliards de paramètres et l'ensemble des résultats possibles étant infini, il est en pratique impossible de vérifier manuellement le caractère privé de toutes les réponses. Néanmoins, il existe des métriques qui permettent d'approximer ou de donner une borne maximale sur les fuites de données potentielles.

Une première métrique est *l'extractibilité*. Nous disons qu'un texte est *k-extractible* s'il est possible de le générer à partir d'une entrée de longueur k (en nombre de mots). Par exemple, si le modèle renvoie "10 rue république" lorsqu'on lui demande "Monsieur Dupont habite à", cette adresse est 3-extractible.

Pour les données personnelles ou sensibles, l'objectif est d'avoir un k le plus élevé possible, car un k faible implique une extraction facile. Une étude de ce type a été réalisée sur GPT-2 [3] : elle a permis d'extraire facilement des informations personnelles sur des individus.

Un autre risque qu'on peut évaluer est *l'inférence d'appartenance*. L'objectif ici est d'identifier si une donnée a été utilisée lors de l'apprentissage du modèle. Supposons par exemple qu'un hôpital entraîne un modèle pour détecter la présence de cancer à partir d'extraits médicaux de patients. Si vous parvenez à découvrir que le modèle a été entraîné sur les données de Monsieur Dupont, vous apprendrez indirectement qu'il est probablement atteint de cancer.

Pour éviter cela, nous devons nous assurer que le modèle ne donne aucun indice quant aux données sur lesquelles il a été entraîné, ce qu'il fait par exemple lorsqu'il se montre trop confiant vis-à-vis de certaines réponses (le modèle va mieux se comporter sur des données qu'il a déjà vu pendant la phase d'entraînement).

Trouver le bon compromis

Faire comprendre au modèle quelles données sont à caractère personnel n'est pas évident, puisque la frontière entre ces deux types de données dépend bien souvent du contexte (l'adresse d'Harry Potter est connue de tous, contrairement à celle de Monsieur Dupont).

L'entraînement d'un modèle qui respecte la confidentialité passe alors souvent par l'ajout de bruit à un moment ou un autre. L'ajout de bruit consiste à altérer l'information apprise ou bien les réponses du modèle, ce qui permet de réduire les risques d'extraction ou d'inférence. Mais cela implique aussi une légère baisse d'utilité. Il faut donc faire un compromis entre performance et respect des données personnelles.

Les applications potentielles des modèles de langage sont incroyablement vastes, mais il est nécessaire d'encadrer leur pratique en prenant compte les risques de fuites avant leur déploiement. De nouvelles méthodes d'entraînement, ainsi que l'anonymisation des données, voire l'utilisation de données synthétiques,

sont toutes des solutions prometteuses et en cours d'étude [5], mais il faudra de toute manière les accompagner de métriques et de méthodologies pour valider non seulement les performances mais aussi la confidentialité des informations personnelles utilisées lors de l'entraînement des modèles.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[3] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, A. Oprea, and C. Raffel, "Extracting training data from large language models," 2021.

[5] https://files.inria.fr/ipop/