Localisation des poids impliqués dans la détection de mensonges générés par un LLM

Coralie Serrand
serc1702

Louis Clériot
clel3204

Louis-Vincent Capelli
capl1101

1 Introduction

Les grands modèles de langage, ou LLMs, ont la capacité de générer du texte. Il arrive néanmoins fréquemment que le texte généré soit erroné et ce alors que l'information correcte se trouve dans les données d'entraînement du LLM. On peut alors parler de mensonges. Par exemple, à la question "Peux-tu me dire où se trouve la Statue de la Liberté ?", le LLM répond "Bien sûr ! La Statue de la Liberté se trouve aux États-Unis à Chicago". Des recherches approfondies montrent que les LLMs ont non seulement tendance à mentir avec assurance [5] mais persistent aussi dans le mensonge lorsque leur première affirmation est fausse [4]. Ceci soulève des préoccupations sur leur fiabilité.

Notre objectif est donc de détecter lorsqu'un LLM génère de fausses informations. Des études précédentes [1] révèlent que les poids internes d'un LLM fournissent de l'information sur le fait que celui-ci mente ou non. Afin de détecter quels sont les poids jouant un rôle déterminant dans la détection de mensonges, nous allons entraîner d'autres classifieurs sur les valeurs d'activation des neurones du LLM lorsqu'une assertion vrai ou fausse lui est donnée en entrée.

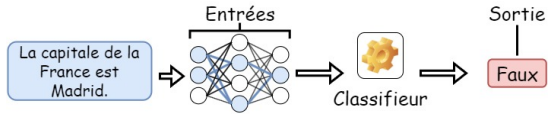

Figure 1: Entrées et sortie de notre modèle

2 État de l'art

L'état actuel de la recherche sur la détection de mensonges dans les LLMs se divise entre deux approches : la méthode "black box", qui analyse uniquement les réponses du modèle, et la méthode "white box", qui se base sur les activations interne du modèle. Une approche "black box" notable utilise 48 questions binaires pour entraîner un classifieur sur les réponses du modèle, sans accéder aux mécanismes sous-jacents. Cependant, cette méthode est principalement efficace pour des tâches de question-réponse et repose sur un seuil arbitraire pour déterminer le mensonge, sans expliquer la généralisation à d'autres types de questions. D'autre part, l'approche "white box" a montré que les modèles possèdent une représentation interne de la vérité, qui peut être altérée en modifiant certains poids [6], mais cette découverte est limitée à des questions binaires et ne se généralise pas aisément. Des approches non-supervisées [3], explorent l'espace des poids internes en les supposant linéairement séparables [6] selon la véracité de l'information donnée en entrée du LLM.

3 Notre démarche et ses enjeux

Notre approche (fig.1) poursuit celle de l'article [1] qui est de type "white box" et qui montre que les LLMs peuvent connaître la vérité d'une assertion mais produire une réponse contradictoire due à des contraintes de cohérence de la phrase et de grammaire. Identifier les poids qui déterminent la véracité peut améliorer la critique des réponses des LLMs et l'expérience utilisateur, en censurant ou alertant sur les réponses potentiellement fausses. Nous allons donc nous concentrer sur le fait de localiser quels poids servent à prédire effectivement si une assertion est vraie ou fausse.

Pour reproduire l'expérience de l'article nous n'avons à disposition que les données [2] utilisées mais le processus d'entraînement reste flou.

Nous pourrons ensuite sélectionner un sous-ensemble des poids du modèle de base qui seront pertinents pour la prédiction. La sélection des classifieurs sera également un défi puisqu'elle devra faciliter la localisation des poids utiles. Contrairement aux chercheurs qui ont utilisé des réseaux de neurones, nous pensons nous orienter vers des forêts aléatoires ou un XGBoost par exemple.

References

[1] Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying.

[2] Amos Azaria and Tom Mitchell. 2023. true-false-dataset. azariaa.com/Content/Datasets/true-false-dataset.zip.

[3] Collin Burns, Hao-Tong Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *ArXiv*, abs/2212.03827.

[4] Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models?

[5] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

[6] Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *ArXiv*, abs/2310.06824.

A Contributions

Coralie a rédigé la partie "Introduction" et a fait des recherches sur l'origine des hallucinations et mensonges des LLMs. Elle a aussi fait le schéma d'entrées-sorties de la solution que l'on souhaite proposer.

Louis s'est occupé de rassembler de nombreux articles sur notre sujet et sur les solutions proposées afin de nous permettre de nous informer sur l'état de l'art en les lisant et en partageant nos interprétations. C'est lui qui a proposé notre article de référence et qui a rédigé la partie "État de l'art".

Louis-Vincent a rédigé la partie "Notre démarche et ses enjeux" et s'est renseigné sur la réception des LLMs par le grand public, et leurs cas d'utilisation qui peuvent être problématiques selon le domaine et la cible concernés.

Nous avons ensuite tous les trois relu les rédactions de chacun et les avons synthétisées pour qu'elles rentrent dans une seule page. Nous avons également dû sélectionner un seul des deux schémas créés.

Pour la suite de notre projet, nos objectifs sont :

- Retracer la démarche précise d'entraînement et de classification de notre article de référence [1] (Tout le monde)
- Trouver comment accéder aux poids de modèles pré-entraînés lors de l'inférence (Louis)
- Sélectionner divers sous-ensembles potentiels de poids comme candidats d'entrées pour la classification (Louis-Vincent)
- Sélectionner les types de classifieurs que nous utiliserons sur les critères énoncés dans la partie "Notre démarche et ses enjeux" (Coralie)

B Jeux de données

Nous souhaitons utiliser le jeu de données [2] publié en même temps que l'étude [1]. C'est un jeu de données d'assertions vraies et fausses sur 6 sujets différents mais nous n'en utiliseront probablement que 5 qui sont les moins ambigus. Il contient 50% d'assertions vraies et 50% d'assertions fausses pour un total de 6084 assertions.

C Ressources

Concernant les ressources informatiques, nous possédons un compte Google Colab Pro qui nous permettra d'entraîner les modèles. Nous avons aussi en tant qu'étudiants accès à un certain nombre de crédits gratuits sur Azure.