Pertinence et stabilité de l'Auto-Explicabitilité des LLMs comparée aux méthodes Lime et SHAP

Alexandre Pascault*[1]

[1]Astek - Direction de la Recherche et de l'Innovation (Astek - DRI) – entreprise privé – 81ter Rue Marcel Dassault 77, 92100 Boulogne-Billancourt, France

Résumé

Les avancées récentes des grands modèles de langue (LLM) ainsi que leur très forte démocratisation,
rend ces technologies accessibles à des developpeurs non expert ; notament dans les produits de
type agents conversationnels. Maitriser ces systèmes est un enjeu important vis à vis par exemple
des hallucinations (Ji et al., 2023), et leur complexité croissante fait de la transparence une qualité
cruciale pour leur intégration éthique (Barman et al., 2024) et leur acceptabilité sociale.
Ces travaux de recherche évaluent l'efficacité de l'auto-explicabilité (Huang et al., 2023) intégrée
aux LLM par rapport aux techniques d'explication externe, telles que LIME (Local Interpretable
Model-agnostic Explanations) (Salih et al., 2023) et SHAP (SHapley Additive exPlanations) (Mosca
et al., 2022).
Les analyses ont été effectuées sur les modèles GPT-2 (Wang et al., 2022) et LLaMA2 (Touvron
et al., 2023), utilisés pour des tâches de classification et question/réponses. La stabilité et la qualité de
1000 justifications fournies conjointement aux réponses choisies par les modèles ont été comparées
avec des explications générées par les méthodes externes. La stabilité a été évaluée en étudiant la
différence induite dans les explications à partir d'un petit changement du texte initial, la qualité a été
évaluée subjectivement.
Les résultats montrent que l'auto-explicabilité est bien intrinsèquement alignée avec le processus
décisionnel des modèles. Elle présente une bonne stabilité pour Llama2 (similarité cosinus allant
jusqu'à 75%) mais peut présenter des limitations en termes de précision et de pertinence, particulièrement faibles pour GPT2. En revanche, LIME et SHAP offrent une granularité et une transparence

*Intervenant

supérieure (évaluées à 98% d'explications pertinentes pour Llama2), mais leur stabilité peut être
affectée par de petites variations dans le vocabulaire utilisé.
L'étude conclut que l'auto-explicabilité des modèles testés est relativement fiable pour les tâches de
classification et question réponse confiées à Llama2, et a l'avantage d'être intrinséquement présente
donc accessibles aux non-experts, un atout indéniable pour assurer une adoption responsable et
éthique des technologies d'IA. Les méthodes externes LIME ou SHAP pourraient fournir un cadre
explicatif plus précis et complet, notament au niveau de l'influence de mots individuels, malgrès leur
mise en oeuvre plus compliquée pour un public non expert.

Mots-Clés: Grands Modèles de Langue, Explicabilité de l'Intelligence Artificielle, LIME, SHAP