Injection de connaissance terminologique et de littérature grise en santé dans un LLM Génératif sur infrastructure Multi-GPUs

Badisse Dahamna[1,2], Romain Lelong[1,2], and Benoist Gaston*[3]

[1]DeSaN – CHU Rouen, Rouen – France
[2]Analyse intégrée multimodale en santé (AIMS) – Université de Rouen Normandie – France
[3]Centre Régional Informatique et d'Applications Numériques de Normandie – Région Normandie – France

Résumé

Motivations : L'intégration de connaissances spécialisées dans les modèles de langage de grande taille (LLM) est un enjeu majeur pour améliorer leur performance dans des domaines spécifiques comme la santé. Notre projet vise à enrichir un LLM génératif avec des données dites de connaissances en santé, afin de répondre aux objectifs suivants :
- améliorer la précision et la pertinence des réponses générées

- étudier l'impact de l'agrégation de connaissances complémentaires dans un même modèle

- envisager une alternative aux systèmes RAG et leur limitations en termes de tailles de contexte.

Travail Scientifique et Technique : Nos expérimentations consistent à injecter de la connaissance terminologique issue de notre portail termino-ontologique HeTOP (www.hetop.eu) et de la littérature grise en santé issue de nos portails LiSSa (www.lissa.fr et Doc'CISMeF (http://doccismef.chu-rouen.fr) dans un LLM. Nous utilisons en première intention un modèle Gemma 9B et effectuons un pré-entraînement continu de type MLM.

Résultats: Il s'agit de travaux en cours. L'expérimentation initiale avec Gemma 9B montre des besoins importants en VRAM qui ont nécessité la distribution du modèle sur plusieurs GPUs. Des premiers résultats qualitatifs sont attendus dans les prochaines semaines.

Originalité : Notre approche combine l'injection de connaissances terminologiques et de littérature grise dans un LLM génératif, ce qui n'a pas été largement exploré dans le domaine de la santé. De plus, la parallélisation sur plusieurs GPU pour gérer les contraintes de VRAM constituent un aspect technique avancé nécessitant une infrastructure matérielle conséquente et généralement peu accessible pour un établissement de santé.

Impact : Cette recherche a un impact potentiel majeur sur le domaine de la santé, en permettant aux professionnels de bénéficier de réponses plus précises et pertinentes générées par un unique LLM. Elle pourrait également ouvrir la voie à de nouvelles applications des

*Intervenant

LLM dans d'autres domaines spécialisés.
Choix des Ressources de Calcul et de Stockage : Pour mener à bien nos expérimentations, nous utilisons des ressources de calcul et de stockage du CRIANN, notamment ses serveurs à GPU de type A100 ou H200 (connectés en Nvlink) pour la parallélisation du modèle.