RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

anr agence nationale de la recherche
AU SERVICE DE LA SCIENCE

LLM4ALL

LLM4ALL
ANR-23-IAS1-0008
up-to-date LLM for all

Date de démarrage : 01/10/2023 **Date de fin : 31/03/2027**

Coordinateur : CERISARA Christophe, LORIA/CNRS, *cerisara@loria.fr*

Consortium

Partenaires

- Académiques:
  - LORIA: CNRS, INRIA
  - LIX - DASCIM: CNRS
  - AP-HP: INRIA
- Industriels:
  - LINAGORA
  - With support from Huggingface

With support from:

HUGGING FACE

Compétences de chacun dans le projet

- LORIA: LLM modeling, pruning; dialogue processing; speech processing
- LIX: LLM distillation, low-cost training
- AP-HP: applications to healthcare domain; data provider

- LINAGORA: LLM training, finetuning, speech processing; data provider
- Huggingface: LLM training, hosting

Positionnement

Contexte et Objectifs

Comment mettre à jour un LLM ? Le finetuning ajoute difficilement de nouvelles connaissances et mène au *catastrophic forgetting*.

Objectifs:

- mettre à jour les modèles existants
- réduire les coûts d'apprentissage
- faciliter l'intégration de nouvelles informations
- appliquer au dialogue en santé et en réunion

Originalité et Verrous

Verrous liés à la mise à jour des LLM: Coût de l'apprentissage, catastrophic forgetting
Verrous liés au résumé de réunions: langue française, modélisation des dialogues inter-humains (pas de chatbot !)
Verrous liés aux appels d'urgence: dialogue oral en situation de stress, données sensibles difficiles à acquérir, LLM voix+texte

Originalité:

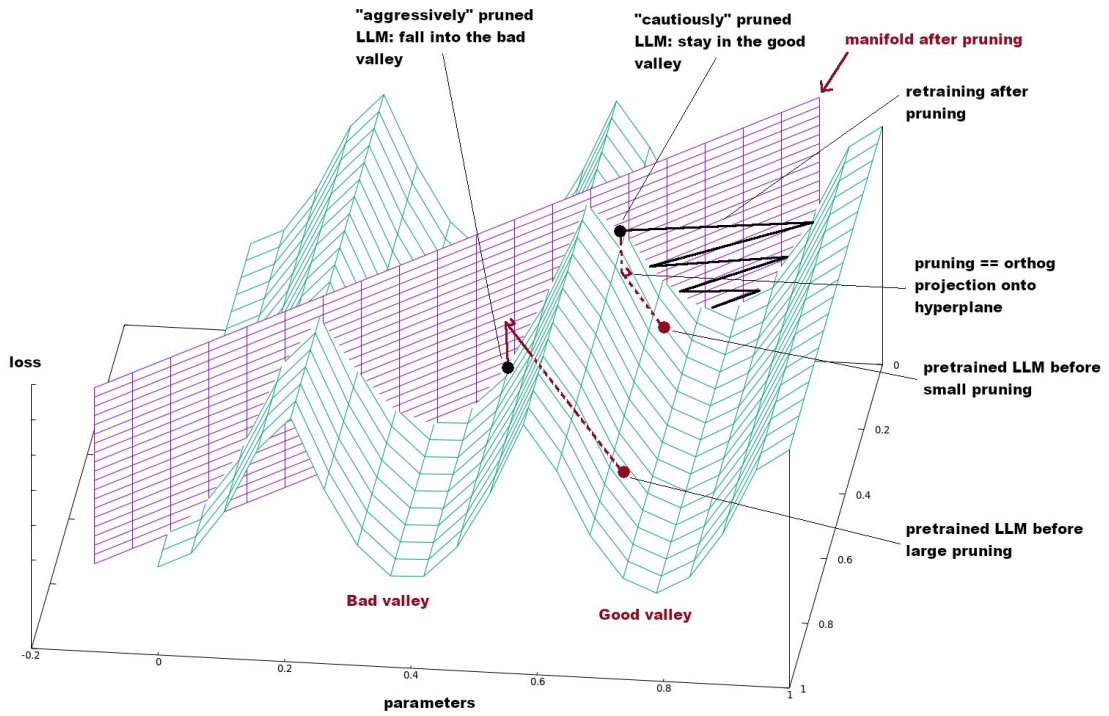
- Exploration et proposition de nouvelles approches combinant pruning / distillation / compression low-rank en vue de continuer le pré-training de modèles existants, et de réduire les coûts d'apprentissage en travaillant à (a) maintenir le niveau de performances tout en diminuant le nombre de paramètres et (b) cibler des données de meilleures qualité.
- Propositions de modèles permettant de traiter les données multimodales dans le domaine médical: voix + texte, données structurées et questionnaires semi-structurés

Réalisation

Méthodologie

Compression: poids = high-rank, activations = low-rank, mais compresser les activations demande plus de données; nous combinons pré-compression des poids avec SVD puis distillation

Simsamu dataset: simulation d'appels d'urgence en FR, 3h15 de données audio + texte, 61 sessions. Disponible sur huggingface hub: medkit/simsamu



Attendus et retombées

LLM4ALL produira en open-source:

- une version mise à jour d'un LLM fondation existant.
- un LLM adapté aux résumés de réunion
- un LLM adapté aux appels d'urgence du SAMU
- les études et scripts permettant de mieux comprendre les verrous et solutions pour construire ces LLM