Travail pratique # 3
Étude comparative de petits LLMs

Automne 2024

Proposé par l'Autorité des marchés financiers (AMF)

1. Description du projet

Pour certains projets, nous devons utiliser des cartes graphiques de puissance faible ou modérée à cause de contraintes internes. Nous avons mené quelques expérimentations préliminaires sur des tâches d'analyse de données confidentielles. Nous avons obtenu de bons résultats avec certains petits LLMs, comme le modèle *Qwen 2.5-3B*.

Nous vous proposons de nous aider à identifier des petits LLMs de 4 milliards (ou moins) de paramètres qui fonctionnent bien pour des tâches d'analyse de documents. Les 4 tâches d'intérêt que nous proposons sont :

a) Le résumé d'un document en quelques lignes qui contiennent l'essentiel du contenu;
b) La description d'un document en 1 à 2 lignes de texte au maximum (idéalement un passage très court);
c) La classification de documents en fonction de classes prédéterminées (par ex. message légitime ou frauduleux);
d) L'extraction d'informations personnelles (des entités nommées) comme des adresses courriel, des numéros de téléphone ou des noms.

2. Jeu de données – *Enron Fraud Email*

Le jeu de données à utiliser pour le projet est le jeu de données publiques de Enron. Des versions de ce *dataset* sont disponibles sur différents sites, dont les suivants :

- https://www.justice.gov/archive/index-enron.html
- https://www.cs.cmu.edu/~./enron/
- https://www.kaggle.com/datasets/advaithsrao/enron-fraud-email-dataset

Vous pouvez utiliser la version (ou une autre version) de ce jeu de données qui vous convient le mieux. Vous pouvez nous contacter en cas de doute.

3. Liste de modèles proposés

Voici une liste de 20 petits LLM, avec 4 milliards de paramètres ou moins, pouvant être considérés pour cette étude :

1. Minitron : https://huggingface.co/nvidia/Minitron-4B-Base
2. Llama 3.2 : https://huggingface.co/meta-llama/Llama-3.2-1B
3. SmolLM : https://huggingface.co/HuggingFaceTB/SmolLM-1.7B
4. Phi-3 : https://huggingface.co/microsoft/Phi-3-small-128k-instruct
5. Qwen2.5 : https://huggingface.co/Qwen/Qwen2.5-3B

6. Gemma : https://huggingface.co/google/gemma-2b-it
7. T5 : https://huggingface.co/google-t5/t5-3b
8. MobileLLM : https://huggingface.co/facebook/MobileLLM-1B
9. Opt : https://huggingface.co/facebook/opt-2.7b
10. OpenELM : https://huggingface.co/apple/OpenELM-3B-Instruct
11. Stablelm-zephyr : https://huggingface.co/stabilityai/stablelm-zephyr-3b
12. Cerebras : https://huggingface.co/cerebras/Cerebras-GPT-2.7B
13. Gpt-neo : https://huggingface.co/EleutherAI/gpt-neo-2.7B
14. Dolly : https://huggingface.co/databricks/dolly-v2-3b
15. h2o-danube3-4b-chat : https://huggingface.co/h2oai/h2o-danube3-4b-chat
16. Pythia : https://huggingface.co/EleutherAI/pythia-2.8b
17. TinyLlama : https://huggingface.co/TinyLlama/TinyLlama-1.1B-Chat-v1.0
18. RedPajama : https://huggingface.co/togethercomputer/RedPajama-INCITE-Chat-3B-v1
19. Allenai OLMO : https://huggingface.co/allenai/OLMo-1B-hf
20. Opt-iml-max : https://huggingface.co/facebook/opt-iml-max-1.3b

Pour ces familles de modèles, différentes versions peuvent être testées tant que le modèle a 4 milliards de paramètres ou moins. Il pourrait également être intéressant de tester plusieurs modèles d'une même famille ayant des tailles différentes (p. ex. Qwen 0.5B, Qwen 1.5B et Qwen 3.0B).

Pour toute autre famille de modèle non mentionnée dans cette liste, merci de nous contacter pour approbation.

4. Consignes pour ce travail

Nombre de modèles et de tâches : L'envergure minimale exigée pour cette étude varie en fonction de la taille des équipes :

Taille de l'équipe	**Nombre de modèles (minimum)**	**Nombre de tâches (minimum)**
1 personne	2	2
2 personnes	3	2
3 personnes	4	3

Bonus : 5 points par modèle supplémentaire. Maximum de 20 points bonus au total. Les points seront ajoutés à la note finale du cours de chacun des membres de l'équipe qui aura obtenu un bonus.

Choix de modèles : Afin de couvrir un grand nombre de modèles dans cette étude, chaque équipe doit choisir au moins 1 modèle parmi les 10 premiers et 1 autre parmi les 10 derniers de la liste proposée. Par exemple, les modèles Gemma (6) et OLMO (19). Aucune contrainte pour les autres choix que vous ferez dans cette liste.

Formulation des tâches et des *prompts* :

- Vous pouvez vous limiter à une seule formulation de problème par tâche. Par exemple, faire l'extraction de 3 types d'informations que vous aurez choisis pour la tâche d) ou une classification binaire de votre choix pour la tâche c).
- Comme le *Enron email dataset* a souvent été utilisé dans la littérature en traitement automatique de la langue, vous trouverez sur le Web des exemples de tâches à accomplir avec ce jeu de données.
- Vous pouvez présenter, pour chaque tâche, les résultats obtenus avec un seul canevas (*template*) de *prompt*. Vous pouvez vous inspirer d'exemples dans la littérature ou sur le Web. On s'attend toutefois que vous faites l'effort de bien structurer ces *prompts*.

Évaluation des modèles :

- Résumé de texte : BLEU, ROUGE et BERTScore
- Classification de texte : *accuracy*, F1 (precision et rappel)
- Extraction d'informations : *exact match*, F1 (precision et rappel)
- Temps de réponse moyen par prédiction - indiquez également avec quelle carte graphique et/ou dans quel environnement les modèles sont exécutés.
- Appréciation de l'impact de la longueur des textes sur les performances des modèles.

Format de remise : Des *notebooks* Jupyter remis dans un fichier ZIP. Aucun fichier de départ n'est fourni, vous les construisez vous-même. Merci de nous contacter pour valider tout autre format de remise.

5. ÉVALUATION DU TRAVAIL

Respect des consignes sur le choix des modèles et des tâches	10%
Code, *prompts* et expérimentations (avec explications)	50%
Présentation des résultats. Évaluation comparative. Analyse d'erreurs.	30%
Qualité des *notebooks*	10%
Bonus pour l'étude de modèles supplémentaires	*20%*