Open data

Hugging Face Explorer : Exploration des Modèles de Langage

Contexte du Projet

Dans le cadre de notre projet IA du second semestre sur le *fine-tuning* de modèles de langage (LLM), nous avons développé l'application **Hugging Face Explorer** à l'aide de streamlit.

Fonctionnalités et Architecture

L'application se décompose en trois principales pages :

- **Catalogue des Modèles** : Informations sur les modèles de Hugging Face ;
- **Benchmarks** : Analyse des performances des modèles sur différentes métriques ;
- **Actualités** : Visualisation d'articles récents sur les LLM.

Catalogue des Modèles

Cette page s'appuie sur l'API de **Hugging Face** pour extraire des informations sur les modèles disponibles. Elle fournit des détails comme les auteurs, les tags associés, les téléchargements et les tendances de popularité.

- **Modèle le plus populaire par mois :** Un graphique temporel montre le modèle ayant reçu le plus de likes chaque mois. Cette visualisation aide à identifier les modèles les plus tendance au fil du temps.
- **Camembert des tags :** Ce graphique illustre les tags les plus fréquents parmi les modèles. Cela permet de repérer les thèmes dominants, comme les architectures spécifiques.

Benchmarks

Cette page utilise le **dataset Open LLM Leaderboard** de Hugging Face pour comparer les performances des modèles.

- **Évolution des performances par type de modèle :** Un graphique linéaire montre l'évolution des scores des modèles sur différentes métriques, en fonction du temps. Cela met en évidence les progrès des modèles et des architectures.
- **Répartition des architectures de modèles :** Un diagramme circulaire montre la diversité des types de modèles disponibles.
- **Coût CO2 cumulé au fil du temps :** Une courbe illustre l'empreinte carbone cumulée liée à l'entraînement des modèles.
- **Analyse performance vs impact environnemental :** Un graphique en nuage de points examine la relation entre performances des modèles, leur coût CO2 et la taille des paramètres. Cette visualisation permet d'identifier les modèles à forte efficacité énergétique.

Actualités

Cette page agrège les articles récents sur les LLM via l'API du site **newsapi.ai**. Elle inclut des outils de filtrage par pays, sentiment, et période.

- **Distribution géographique des articles :** Une carte représente la répartition des articles par pays et leur sentiment moyen.
- **Tendances temporelles des articles :** Un graphique à double axe montre l'évolution du nombre d'articles publiés et du sentiment moyen des articles au fil du temps.
- **Répartition des articles par pays :** Un diagramme met en évidence les pays produisant le plus d'articles sur les LLM.

Conclusion

Cette application permet de se documenter sur les modèles de langage open source et leurs performances, en vue de notre projet IA du second semestre mais aussi :

- D'identifier les modèles les plus adaptés pour notre projet de fine-tuning en fonction de leurs benchmarks ;
- De suivre les tendances actuelles dans le domaine des LLM ;
- De choisir une architecture adaptée avec nos objectifs.