Chapitre 2

Outils de recherche

Cette section de mon tutoriel porte sur des outils que j'ai trouvé utiles durant mes années d'études et de recherche.

2.1 Outils pour la recherche d'article

Google Scholar
ArXiv
Zotero
ResearchRabbit

2.2 Outils pour la rédaction

LaTeX/Typst
Overleaf

2.3 Outils pour le développement

Deep learning frameworks : TensorFlow, PyTorch, Jax, Keras (Tuto 2 : PyTorch)
Machine learning important tools : Numpy, Pandas, Matplotlib/Seaborn/Plotly, Scikit-Learn (Tuto 3 : All things machine learning)
Experiment tracking : Wandb, MLflow
Google Colab/Jupyter Notebook, Kaggle
Hyperparameter tuning : Optuna, Ray Tune, Hyperopt

2.4 LLMs

Je recommande d'utiliser un LLM UNIQUEMENT SI vous pouvez vérifier l'output.

Aussi, je recommande de ne jamais uploader du contenu sous copyright dans un LLM. J'ai demandé à Claude pour avoir une réponse dans la figure 2.1.

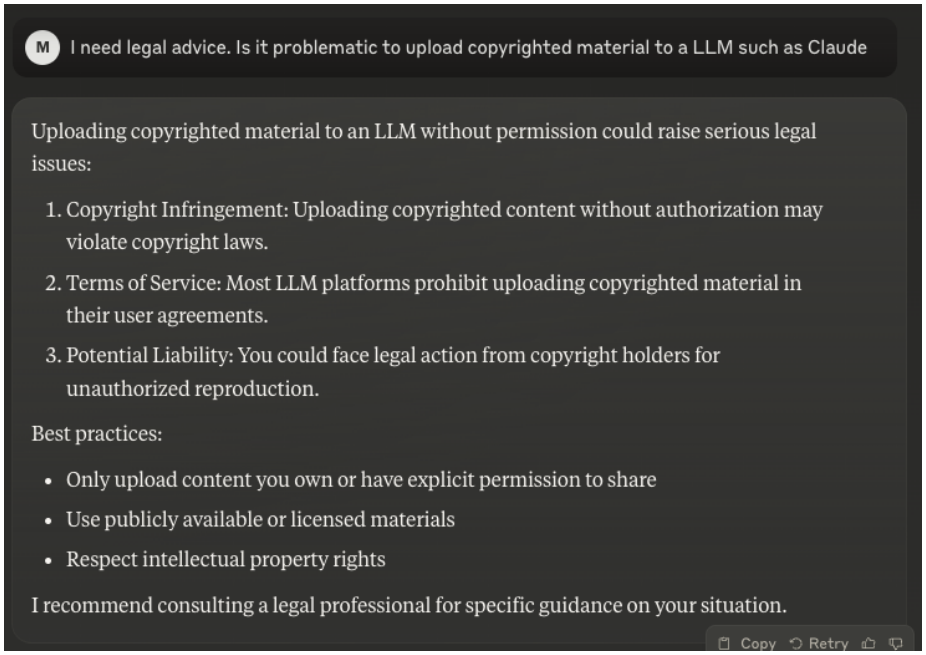

Figure 2.1: Réponse de Claude sur la question

J'ai aussi trouvé très intéressant la phrase suivante, du *Terms of Service* de ChatGPT. En effet, "[You may not] represent that Output was human-generated when it was not."

Finalement, même si les *Terms of service* de ChatGPT indique que l'output est la propriété de l'utilisateur (voir figure 2.2), généralement le contenu généré par une IA n'est possédée par personne. Quoique ce n'est pas du tout une source scientifique, j'avais trouvé ce vidéo par un streamer assez intéressant : `https://youtu.be/pt7GtDMTd3k?si=MQNskBNIoZRPN7OZ`.

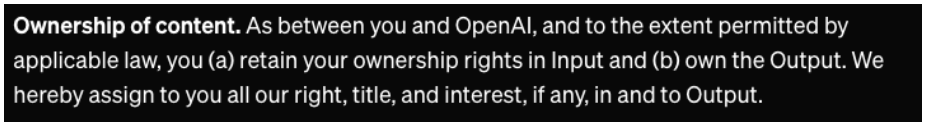

Figure 2.2: Propriété intellectuelle de la sortie d'une IA

LLMs disponibles :
ChatGPT (toujours désactivé l'entraînement sur les données, paramètres section Gestion des données, désactivez "Améliorer le modèle pour tous")

Claude (selon leurs conditions d'utilisation, ils n'utilisent pas les données par défaut. Toutefois, ils ont quand même utilisés les donnés dans le papier de Clio (https://www.anthropic.com/research/clio))

Le chat (Mistral) (ils disent que les données ne sont pas utilisées : https://help.mistral.ai/en/articles/156194-does-mistral-ai-exploit-users-data-to-train-its-models et que whatever données qui doivent être supprimées peuvent être fait en leur écrivant : https://help.mistral.ai/en/articles/154193-how-can-i-exercise-my-gdpr-rights)

Meta.ai (il ne semble pas y avoir d'options pour la vie privée)

DeepSeek (selon ce post, c'est vraiment pas une bonne idée : https://medium.com/data-science-in-your-pocket/dont-use-deepseek-v3-895be7b853b0.) You cannot use, copy, or even display any content or software from DeepSeek without permission. Any misuse, even unintentional, may lead to legal action from DeepSeek.

Gemini (might not be able to opt out of data being used to train the model : https://www.googlecloudcommunity.com/gc/AI-ML/Use-of-your-data-for-product-improvement-purposes-in-the-Google/m-p/723832)
Copilot (payé par l'université pour garantir la non-utilisation des données)

Avec un peu de volonté, il existe des modèles OpenSource qui peuvent être utilisés directement sur votre ordinateur. Par exemple, on a :
Lucie : `https://huggingface.co/OpenLLM-France/Lucie-7B-Instruct`
LLama : `https://www.llama.com/docs/llama-everywhere/running-meta-llama-on-mac/`
Ça m'a pris environ 30 secondes à installer Ollama et je suis capable d'utiliser Lucie et Llama sans problème. L'environnement dans un terminal est moins bien, mais j'ai téléchargé oterm et ça semble fonctionner sans problème.

2.5 Outils divers

`https://github.com/yuchenlin/rebiber`
`https://github.com/google-research/arxiv-latex-cleaner`
`https://capitalizemytitle.com/`
`https://flamingtempura.github.io/bibtex-tidy/index.html`