Extraction automatique d'informations environnementales à partir d'articles scientifiques

Thibault Schneeberger
QARMA, LIS
thibault.schneeberger@proton.me

Valentin Emiya
QARMA, LIS
valentin.emiya@lis.fr

Constance Douwes
QARMA, LIS
constance.douwes@lis.fr

1 Introduction et état de l'art

L'essor des modèles d'apprentissage profond à grande échelle a engendré une préoccupation croissante quant à leur impact environnemental. Dès 2019, Strubell et al. (2019) ont attiré l'attention de la communauté du traitement automatique des langues en quantifiant les coûts financiers et environnementaux de l'entraînement de modèles comme BERT et Transformer, révélant des empreintes carbone comparables à celles de vols transatlantiques.

Depuis, plusieurs travaux ont affiné ces estimations. Patterson et al. (2021) ont analysé l'énergie consommée par des modèles majeurs (T5, GPT-3, Switch Transformer) et montré que le choix du centre de données, de la localisation géographique et du matériel peut réduire l'empreinte carbone d'un facteur 100 à 1000. Luccioni et al. (2023) ont proposé une méthodologie complète pour estimer l'empreinte carbone du modèle BLOOM (176 milliards de paramètres) sur l'ensemble de son cycle de vie, de la fabrication du matériel au déploiement, aboutissant à une estimation de 25 à 50 tonnes de CO_2eq.

Malgré ces avancées méthodologiques, les informations relatives à l'impact environnemental restent dispersées dans les articles scientifiques et difficiles à collecter systématiquement. Certaines bases de données ont été constituées manuellement : Epoch.ai (Epoch AI, 2025) recense plus de 3000 modèles notables avec leurs caractéristiques techniques (nombre de paramètres, puissance de calcul, matériel), tandis que GreenMIR (Kaila and Holzapfel, 2024) catalogue la présence d'informations environnementales dans les publications du domaine audio (ISMIR). Cependant, ces efforts manuels ne permettent pas de suivre l'ensemble des modèles publiés de façon exhaustive et pérenne.

L'objectif de ce projet est d'automatiser l'extraction d'informations liées à l'impact environnemental des modèles d'apprentissage automatique directement à partir des articles scientifiques. Les approches d'extraction d'information ont été largement étudiées (Abdullah et al., 2023), et les méthodes génératives basées sur des grands modèles de langage (LLM) se révèlent actuellement les plus performantes (Zhang et al., 2025), surpassant les approches extractives classiques de type BERT (Gardazi et al., 2025).

Nous présentons dans ce rapport un système d'extraction automatique, ainsi qu'une évaluation permettant de mesurer la qualité des extractions par rapport aux données de référence d'Epoch.ai et GreenMIR.

2 Approches Préliminaires

Avant d'aboutir à l'architecture finale, nous avons exploré qualitativement plusieurs approches d'extraction d'information.

Expressions régulières. La première approche a été l'utilisation d'expressions régulières pour capturer des motifs numériques et textuels. Pour des raisons évidentes de complexité et de robustesse face à la variabilité des formulations scientifiques, cette méthode a été rapidement abandonnée.

Modèles extractifs. Nous avons testé un modèle RoBERTa réentraîné sur SQuAD v2, où la description de l'information à extraire était fournie en question et le modèle retournait un extrait de texte. L'avantage majeur est l'absence d'hallucination puisque le texte extrait provient directement du document. Cependant, même avec des modèles à grand contexte, il était difficile de faire correspondre les informations entre plusieurs sec-

tions du texte, le contexte se limitant souvent à un ou deux paragraphes (voir annexe A).

Génération augmentée par récupération. Nous avons réalisé quelques tests avec des systèmes de type RAG. L'information étant parfois diffuse dans le texte, il était difficile de discriminer les passages pertinents pour la récupération.

Grands modèles de langage sur GPU local. Nous avons testé des LLM en mode sans exemples préalables sur le cluster GPU, équipé de cartes NVIDIA A100 de 40 à 48 Go. La mémoire requise croît de façon quadratique avec la taille du contexte. Dans notre cas, le contexte est l'intégralité du texte du papier scientifique, soit plusieurs dizaines de milliers de tokens. Les modèles comme Mistral ou Llama étaient inaccessibles à cette échelle en raison de leur taille trop élevée. Nous avons testé Gemma 3 E4B et GPT-OSS 20B via Ollama et la bibliothèque Transformers, mais les résultats n'étaient pas satisfaisants en raison d'hallucinations fréquentes.

API Gemini 2.5 Flash. Nous nous sommes finalement tournés vers l'API Gemini 2.5 Flash, qui offre un contexte suffisamment grand, jusqu'à un million de tokens, et des performances supérieures à toutes les approches précédentes.

3 Architecture et Méthodologie

Le système d'extraction se décompose en trois étapes principales : la préparation des données, l'extraction par modèle de langage, et l'évaluation des résultats.

3.1 Préparation des données

La première étape consiste à constituer une base de données SQLite à partir des jeux de données de référence. Pour chaque papier, nous récupérons le document PDF depuis son URL, puis en extrayons le texte brut. Les informations sont organisées en plusieurs tables :

- **paper_info** : métadonnées du papier, incluant le lien vers le PDF
- **paper_document** : contenu binaire du PDF téléchargé
- **paper_text** : texte extrait du document
- **model_info** : informations de référence sur les modèles, servant de vérité terrain

Deux tables de référence sont également constituées : une table des pays avec leur intensité carbone en gCO_2/kWh, et une table du matériel avec sa puissance de calcul et sa consommation (voir annexe B pour le schéma complet).

3.2 Stratégie d'extraction

L'extraction se déroule en deux phases. Dans un premier temps, le modèle de langage reçoit le texte intégral du papier et doit énumérer tous les modèles mentionnés. La réponse attendue est une liste Python de noms de modèles.

Dans un second temps, pour chaque modèle identifié, une série de questions est posée afin d'extraire les attributs suivants :

- Nombre de paramètres
- Type de matériel utilisé pour l'entraînement
- Nombre d'unités matérielles
- Durée d'entraînement en heures
- Puissance de calcul totale en FLOP
- Pays d'entraînement
- Année de publication
- Émissions de CO_2 équivalent

Pour chaque question, le modèle doit retourner une liste contenant deux éléments : la valeur extraite et une citation du passage source justifiant cette réponse. Cette structure permet de tracer l'origine de chaque information et de détecter d'éventuelles hallucinations.

3.3 Évaluation

L'évaluation compare les informations extraites aux données de référence en deux phases.

Appariement des entités. Les modèles inférés sont d'abord appariés aux modèles de référence par similarité de nom, en utilisant la distance de Jaro-Winkler. Cet appariement permet de distinguer trois cas : les modèles correctement détectés, les modèles manqués et les modèles hallucinés.

Notation des champs. Pour chaque paire de papiers et paire appariée de modèles, une distance est calculée par champ selon sa nature, de façon a ce qu'elle soit comprise entre 0 et 1 :

- **Valeurs numériques** : erreur relative symétrique, définie par :

$$\frac{|v_{inf} - v_{ref}|}{\max(|v_{inf}|, |v_{ref}|)}$$

- **Identifiants** : correspondance binaire, définie par :

$$\begin{cases} 1 & \text{si } v_{inf} = v_{ref} \\ 0 & \text{sinon} \end{cases}$$

- **Chaines de caractères** : Jaro-Winker agnostique a l'ordre des mots.
- **Année** : erreur absolue normalisée sur 5 ans, définie par :

$$\frac{|v_{inf} - v_{ref}|}{5}$$

Ces distances sont ensuite agrégées pour calculer des métriques de précision et rappel souples, prenant en compte à la fois la qualité de l'extraction et les pénalités pour les modèles hallucinés.

4 Expérimentations et Résultats

Deux expériences ont été réalisées pour évaluer le système d'extraction.

4.1 Protocole expérimental

GreenMIR. La première expérience porte sur l'intégralité du jeu de données GreenMIR, soit environ 110 papiers issus de la conférence ISMIR. Ce jeu de données est centré sur les modèles de traitement audio et musical, avec un accent particulier sur la documentation de l'impact environnemental.

Epoch.ai. La seconde expérience porte sur un sous-ensemble du jeu de données Epoch.ai. Le jeu complet contenant plus de 3000 modèles, son traitement intégral aurait nécessité un nombre de jetons trop important. Nous avons donc sélectionné un échantillon représentatif de papiers accessibles via ArXiv.

4.2 Métriques d'évaluation

L'évaluation repose sur deux niveaux de granularité : l'appariement des entités et la qualité d'extraction des champs.

Métriques d'appariement. L'appariement permet de mesurer la capacité du système à détecter les bons modèles. Après appariement par similarité Jaro-Winkler, on distingue :

- **TP** : modèles de référence correctement appariés à un modèle inféré
- **FN** : modèles de référence non détectés (oublis)
- **FP** : modèles inférés sans correspondance (hallucinations)

Les métriques classiques sont alors calculées :

$$\text{Précision} = \frac{TP}{TP + FP} \quad (1)$$

$$\text{Rappel} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{F1} = \frac{2 \times \text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}} \quad (3)$$

Métriques de qualité des champs. Pour chaque champ, on définit un score de similarité $S = 1 - d$, où d est la distance calculée selon le type de champ. On calcule ensuite des métriques souples qui prennent en compte la qualité des valeurs extraites :

$$\text{Précision alignée} = \frac{\sum_{i \in TP} S_i}{TP} \quad (4)$$

$$\text{Précision globale} = \frac{\sum_{i \in TP} S_i}{TP + FP} \quad (5)$$

$$\text{Rappel souple} = \frac{\sum_{i \in TP} S_i}{TP + FN} \quad (6)$$

La précision globale pénalise les hallucinations en les comptant comme des erreurs totales dans le dénominateur. Le score F1 souple global est la moyenne harmonique de la précision globale et du rappel souple :

$$\text{F1 souple} = \frac{2 \times \text{Précision globale} \times \text{Rappel souple}}{\text{Précision globale} + \text{Rappel souple}} \quad (7)$$

4.3 Résultats

Appariement des modèles. Le tableau 1 présente les métriques d'appariement pour les deux jeux de données.

Dataset	P	R	F1
GreenMIR	0.XX	0.XX	0.XX
Epoch.ai	0.XX	0.XX	0.XX

TABLE 1 – Métriques d'appariement des entités (P : précision, R : rappel).

Qualité d'extraction par champ. Les tableaux 2 et 3 détaillent les métriques souples pour chaque attribut.

Champ	$\mathbf{P}_a$	$\mathbf{P}_g$	$\mathbf{R}_s$	$\mathbf{F1}_s$
Paramètres	0.XX	0.XX	0.XX	0.XX
Matériel	0.XX	0.XX	0.XX	0.XX
Nb. unités	0.XX	0.XX	0.XX	0.XX
Durée entr.	0.XX	0.XX	0.XX	0.XX
Compute	0.XX	0.XX	0.XX	0.XX
Pays	0.XX	0.XX	0.XX	0.XX
Année	0.XX	0.XX	0.XX	0.XX
CO_2eq	0.XX	0.XX	0.XX	0.XX
Moy.	0.XX	0.XX	0.XX	0.XX

TABLE 2 – Métriques par champ sur GreenMIR. P_a : précision alignée, P_g : précision globale, R_s : rappel souple, $F1_s$: F1 souple.

Champ	$\mathbf{P}_a$	$\mathbf{P}_g$	$\mathbf{R}_s$	$\mathbf{F1}_s$
Paramètres	0.XX	0.XX	0.XX	0.XX
Matériel	0.XX	0.XX	0.XX	0.XX
Nb. unités	0.XX	0.XX	0.XX	0.XX
Durée entr.	0.XX	0.XX	0.XX	0.XX
Compute	0.XX	0.XX	0.XX	0.XX
Pays	0.XX	0.XX	0.XX	0.XX
Année	0.XX	0.XX	0.XX	0.XX
CO_2eq	0.XX	0.XX	0.XX	0.XX
Moy.	0.XX	0.XX	0.XX	0.XX

TABLE 3 – Métriques par champ sur Epoch.ai.

4.4 Analyse

Les attributs les plus faciles à extraire sont :

- L'année de publication, souvent présente dans les métadonnées
- Le nombre de paramètres, fréquemment mentionné dans le texte
- Le type de matériel, généralement indiqué explicitement

Les informations les plus difficiles à extraire sont :

- L'empreinte carbone, rarement reportée dans les articles
- Le pays d'entraînement, souvent non mentionné
- La durée d'entraînement exacte

5 Discussion et Limitations

6 Conclusion

References

M. H. A. Abdullah, N. Aziz, S. J. Abdulkadir, H. S. A. Alhussian, and N. Talpur. 2023. Systematic Literature Review of Information Extraction From Textual Data: Recent Methods, Applications, Trends, and Challenges. *IEEE Access*, 11 :10535–10562.

Epoch AI. 2025. Data on AI Models. Accessed : 2025-12-13.

N. M. Gardazi, A. Daud, M. K. Malik, A. Bukhari, T. Alsahfi, and B. Alshemaimri. 2025. BERT applications in natural language processing: a review. *Artificial Intelligence Review*, 58(166).

Anna-Kristin Kaila and Andre Holzapfel. 2024. Green MIR data table 2024. International Society for Music Information Retrieval Conference (ISMIR), San Francisco, California, USA.

A. S. Luccioni, S. Viguier, and A.-L. Ligozat. 2023. Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. *Journal of Machine Learning Research*, 24(253) :1–15.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon Emissions and Large Neural Network Training.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3645–3650.

Z. Zhang, W. You, T. Wu, X. Wang, J. Li, and M. Zhang. 2025. A Survey of Generative Information Extraction. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pages 4840–4870, Abu Dhabi, UAE.

A Expérimentations avec RoBERTa

Avant d'adopter l'approche par grands modèles de langage, nous avons testé une approche extractive basée sur RoBERTa réentraîné sur SQuAD v2. Le principe consiste à parcourir l'article avec une fenêtre glissante et à récupérer les segments de texte les plus pertinents pour chaque question.

Paramètres expérimentaux.

- Taille de fenêtre : 512 tokens
- Chevauchement : 128 tokens
- Longueur maximale de réponse : 20 tokens
- Nombre de candidats retenus : 10

Visualisation par carte de chaleur. Pour chaque token du document, nous avons calculé un score d'attention indiquant sa pertinence par rapport à la question posée. La figure ci-dessous illustre un extrait de cette carte de chaleur sur le papier BLOOM (Luccioni et al., 2023), pour la question *"What is the name of the main machine learning model presented in this paper ?"*.

```
As reported in Table 1, training
the BLOOM model required a total
of 1.08 million GPU hours on a
hardware partition
constituted of Nvidia A100 SXM4
GPUs with 80GB of memory...
```

L'intensité de la couleur orange indique le score d'attention attribué par le modèle à chaque token. On observe que le mot "BLOOM" obtient le score maximal, ce qui est cohérent avec la question posée.

Segments candidats. Le tableau suivant présente les 10 meilleurs segments détectés :

Rang	Segment extrait	Score
1	Codecarbon	13.13
2	BLOOM	11.72
3	Open-access Multilingual Language Model	11.61
4	433,196 kWh	11.45
5	BLOOM model	10.71
6	1.08 million	10.82
7	TDP	10.60

TABLE 4 – Segments candidats détectés par RoBERTa pour la question sur le nom du modèle.

Limitations observées. Bien que les segments pertinents soient correctement identifiés, plusieurs problèmes limitent l'utilisabilité de cette approche :

- Les scores des bonnes réponses restent proches de ceux des mauvaises réponses, rendant la sélection automatique difficile.
- Le contexte limité à 512 tokens empêche de croiser des informations provenant de sections éloignées du document.
- Le modèle ne peut pas effectuer de raisonnement ou de calcul, contrairement aux LLM génératifs.

Ces limitations nous ont conduits à privilégier l'approche par grands modèles de langage décrite dans ce rapport.

B Schéma de la base de données

La figure 1 présente le schéma relationnel de la base de données utilisée pour stocker les informations extraites.

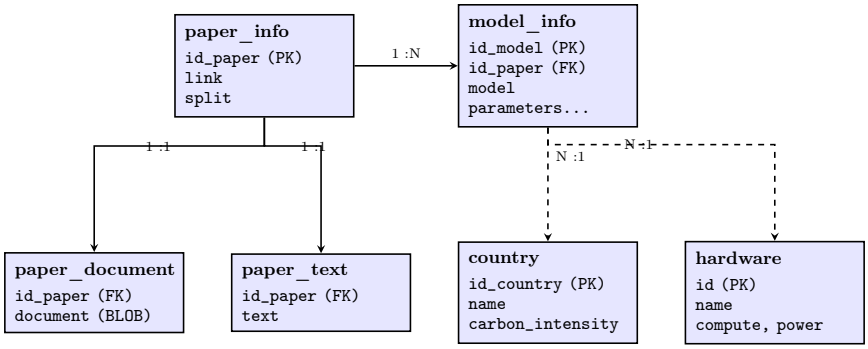

FIGURE 1 – Schéma relationnel. Trait plein : clé étrangère directe. Trait pointillé : référence via table de lookup.

Tables principales.

- **paper_info** : contient les métadonnées de chaque papier (lien, partition train/test)
- **model_info** : contient les informations sur chaque modèle mentionné dans un papier (vérité terrain)
- **paper_document** : stocke le contenu binaire du PDF téléchargé
- **paper_text** : stocke le texte extrait du PDF

Tables de référence.

- **country** : liste des pays avec leur intensité carbone (gCO_2/kWh), source Our World in Data
- **hardware** : liste des GPU/TPU avec leur puissance de calcul (FLOP/s) et consommation (kW), source Epoch.ai