BioLLM

LLM au service de la biodiversité

RAPPORT DE PROJET EXPLORATOIRE

A-IGÉco

ITTECOP

Décembre 2025

Contact : Sylvain Moulherat
Mail : sylvain.moulherat@a-igeco.fr

1. DESCRIPTION DU PROJET

Le secteur de l'ingénierie et du génie écologique est un secteur jugé stratégique par l'Etat, en forte croissance mais souffrant de difficultés de recrutement majeures. Les derniers travaux conduits par la filière en 2022, ont ainsi montré qu'avec un volume d'affaire national doublant chaque année depuis 10 ans et une tendance à l'accélération, la filière devra faire face à un déficit d'environ 50 000 postes d'ici 5 ans. Face à ce constat, le secteur fait régulièrement remonter des difficultés majeures à faire certaines tâches particulièrement chronophages telles que l'extraction d'information issues de documents réglementaires comme les études d'impacts.

Avec le développement très rapide ces dernières années de l'intelligence artificielle générative et plus particulièrement des LLM (*Large Langage Model*), il est désormais possible de développer des applications susceptibles de faciliter l'extraction de d'informations précises à partir de corpus documentaires spécifiques. Par exemple, le modèle Climate Q&R est une intelligence artificielle (IA) conversationnelle dérivée de ChatGPT spécialisée dans le domaine de l'environnement. Elle est entrainée à ne répondre qu'au sujets environnementaux en mobilisant les corpus documentaires produits le GIEC et l'IPBES.

En Occitanie, la Communauté Régionale ERC d'Occitanie (CRERCO) a travaillé ces deux dernières années à constituer une première base de corpus documentaire de plus de 800 études d'impacts qui pourrait être mobilisée dans BioLLM pour développer une IA capable d'extraire et constituer une base de données visant à suivre l'application de la séquence ERC en Occitanie à travers ces études d'impacts.

BioLLM s'appuiera donc sur ces travaux pour :

- Elargir le corpus documentaire constitué par la CRERCO en se rapprochant d'autres territoires ayant adoptés des approches comparables et de l'ITTECOP qui a déjà abordé cette problématique.
- Développer un prototype de LLM spécialisée permettant d'extraire les informations d'intérêt dans les dossiers d'évaluations environnementales
- Constituer une base de données contribuant au suivi de la mise en œuvre de la séquence ERC.

Pour atteindre cette objectif, BioLLM mobilisera les membres des réseaux membres de l'A-IGEco et s'est appuyé sur l'expertise technique de TerrOïko dont l'équipe de chercheurs en intelligence artificielle est membre de l'A-IGEco et impliqués sur ce projet.

2. CALENDRIER DE REALISATION DU PROJET

Démarrage du projet et lancement du recrutement de stagiaire : Janvier 2025

Stage dédié : Avril-Août 2025

Production du livrable : Automne 2025

3. RESULTATS PRINCIPAUX DU PROJET

RESUME

L'annexe I détail le développement d'un prototype d'extraction et de recherche d'informations appliqué à un corpus de documents réglementaires en écologie. L'objectif était de concevoir un pipeline complet, allant de l'acquisition des données à leur restitution, en intégrant des briques de traitement avancées telles que l'OCR, le nettoyage et la segmentation des textes, la vectorisation par *embeddings*, la recherche sémantique et l'enrichissement par reconnaissance d'entités nommées (NER). L'approche s'appuie sur les modèles de langage (LLM) et le *framework Retrieval-Augmented Generation* (RAG). Après définition d'une référence, plusieurs itérations ont permis d'améliorer les performances à l'aide de métriques de recherche d'information (HitRate, MRR, DCG). Les résultats démontrent la faisabilité technique du concept et ouvrent la voie à des extensions dans d'autres domaines tels que la santé publique ou le droit du travail.

ELEMENTS TECHNIQUES CLEFS DU PROJET

Bibliographie et choix méthodologique :

- Méthode "pure-LLM" d'un intérêt limité :
  - il faudrait soit traiter tous les docs par un LLM pour chaque requête (capacité de traitement massif limité) ou fine-tuner un LLM sur les documents existants annotés (construction des données d'entraînement longue à mettre en œuvre)
  - toute l'information ne serait pas mémorisée,
  - pas de traçabilité des sources d'informations,
  - fort risque d'hallucinations.
- Utilisation de la méthode RAG (*Retrieval-Augmented Generation*):
  - Bonne capacité de traitement de données massives enrichies fréquemment,
  - Toutes les informations sont mémorisées,
  - Traçabilité des données permettant de générer les réponses,
  - Réduction du risque d'hallucinations en basant les réponses sur les documents du corpus documentaire.

Collecte des données :

Les données ont été collectées via des services web gouvernementaux et comprennent des études d'impact, des textes de lois, des demandes de dérogations espèces protégées.... Ce sont ainsi 5234 documents mis à disposition via les portails gouvernementaux qui ont été utilisés (Annexe II).

Métriques de performances des pipelines d'analyse

Mise en place un protocole d'évaluation des performances des pipelines de traitements IA lors des requêtages formulés, basé sur des métriques, permettant de mesurer la performance du système et l'impact des expérimentations effectuées. La mesure des métriques est effectuée sur les sources de données et permettent d'estimer la capacité du pipeline à trouver les informations pertinentes au sein du corpus. Elles n'évaluent pas la qualité de la réponse générée par le modèle de language.

- HitRate@k : moyenne de la présence d'au moins une source pertinente dans la liste de sources retournées.
- Mean Reciprocal Rank (MRR) - rang moyen de la première source pertinente dans la liste de résultat retourné.
- Discounted Cumulative Gain (DCG) – mesure globale de la qualité d'une liste ordonnée de résultats.

Conception des pipelines et leur amélioration

Méthodologie générale.

1. Préparation des données (traitement des documents pdf pour mise en base).
2. Traitement des requêtes (table des requêtes évaluatives en annexe III).
3. Evaluation des métriques de performances

- Pipeline de référence - Implémentation d'un RAG basique, servant de référence en matière de performance.
- Amélioration itération 1 : Utilisation du cosinus au lieu d'une distance euclidienne pour la mesure de similarité entre vecteurs. Le cosinus est plus approprié pour une distance "sémantique" et est souvent utilisé dans le cadre de la récupération d'information textuelle.

- Amélioration itération 2 : Ajout de métadonnées. L'ajout de métadonnées ("réglementation", "avis et expertise", "études et données", "référentiels", "code l'environnement"...) permet de filtrer les segments retrouvés et d'éviter les segments "hors-sujet".
- Amélioration itération 3 : Utilisation de Weaviate au lieu de FAISS comme base de données vectorielles. Itération purement technique : Weaviate est plus complexe à mettre en œuvre mais est plus flexible que FAISS et plus appropriée à notre contexte opérationnel.
- Amélioration itération 4 : Extraction d'entités nommées (Named Entity Recognition = NER) Utilisation de modèles NER dédiés à la détection d'entités clés: localisations géographiques, espèces animales et dates. Une fois détectées, ces entités sont ajoutées aux métadonnées et peuvent être utilisées lors de la phase de filtrage.

Synthèse des résultats

	Référence	Evaluation prototype final (itération 4)
HitRate@5	0.38	0.5
MRR	2.85	2.13
DCG@5	0.71	0.78

POINTS FORTS DE L'APPROCHE UTILISEE DANS BIOLLM

- Le RAG permet de construire la réponse à une requête à partir d'éléments, sourcés, des documents de la base de données (au contraire d'un LLM qui génère entièrement toute la réponse).
- L'utilisation d'une distance « sémantique » pour mesurer la similarité entre vecteurs.
- Approche appropriée pour le traitement de données massives : à chaque ajout de documents il faut juste les ajouter à la base de données documentaires (au contraire du besoin de réentraînement avec l'approche « pure LLM »).
- L'utilisation de métadonnées et de NER permet d'indexer les éléments critiques et d'améliorer la recherche de segments pertinents.

LIMITES DE L'APPROCHE UTILISEE DANS BIOLLM

- Si la réponse à une requête est distribuée entre plusieurs segments d'un document, les segments peuvent ne pas être retrouvés (similarité plus faible entre la requête et chaque segment que si toute la réponse était dans un unique segment).
- La performance du NER « espèces » (en français) est faible. Le développement d'un modèle dédié améliorerait la performance globale.
- L'intégration d'un arbre taxonomique serait utile. Par exemple une requête mentionnant « amphibiens » ne cherche que le taxon correspondant. Avec un arbre taxonomique tous les taxons d'un rang inférieur pourraientt être pris en compte.
- Pour les mêmes raisons, l'intégration d'un arbre représentant la hiérarchie des divisions administratives seraient utiles. Par exemple une requête mentionnant une région prendrait en compte tous les départements et communes de la région.

4. DEROULE DU PROJET ET DEVISATIONS PAR RAPPORT AUX ATTENDUS

Bien que le projet ait été conçu et réalisé avec la CRERCO, BioLLM n'a pas eu accès aux documents publiques détenus par la DREAL (co-pilote de la CRERCO). Malgré les annonces faites en 2024-2025, et le soutien fourni par le programme ITTECOP via Mme Millard (annexe IV), la DREAL Occitanie refuse toujours à ce jour de nous transmettre les documents publics relatifs aux dossiers de dérogations espèces protégées. Outre les contraintes sur l'accès et la qualité des données utilisées dans le projet, cette absence de transfert ne permet pas au projet de produire le *data paper* tel qu'espéré dans la proposition initiale.

Malgré les difficultés rencontrées dans la collaboration avec la DREAL, les travaux se poursuivent en collaboration avec les services de la région Occitanie (co-pilote de la CRERCO) et 2026 devrait voir les travaux initiés dans BioLLM se poursuivre et se concrétiser à travers l'analyse des documents relatifs aux SCoT de la région.

5. VALORISATION DES TRAVAUX

LIVRABLE

- Ce document, mis à disposition sur demande auprès de l'A-IGEco.
- Le rapport du stagiaire du projet (Annexe I), mis à disposition sur demande auprès de l'A-IGEco.

CODE SOURCE

- Le code source des travaux disponible sur https://oikolab.terroiko.fr/

COMMUNICATION

- Présentation des travaux lors du séminaire A-IGEco le 01/07/2025 à St-Féréol en marge du lancement de la feuille de route nationale sur l'ingénierie et le génie écologique.

6. CONCLUSION

Bien que le projet BioLLM n'ait put être réalisé dans des conditions optimales (non transfert des données par la DREAL Occitanie), le projet a permis de montrer l'intérêt à moyen terme des modèles de langage pour traiter massivement les documents réglementaires relatifs à la biodiversité. A court terme ces usages sont toutefois à considérer avec beaucoup de précautions car les IA utilisées ont encore des difficultés à trouver les éléments pertinents dans les documents lorsqu'elles parviennent à les trouver.

Des développements spécifiques doivent donc encore être envisagés pour améliorer les performances des LLM à exploiter les documents relatifs à la biodiversité. Deux pistes principales sont envisagées dans ces travaux :

- Des développements techniques des NER pour qu'ils puissent fonctionner efficacement avec les arbres taxinomiques. De tels travaux, sont déjà en cours au LECA.
- Des apprentissages spécialisés (fine-tuning) au contexte des documents réglementaires relatifs à la biodiversité.

Des acteurs académiques et économiques existent et opèrent déjà en France pour massifier ce type d'approches dans d'autres secteurs (santé, assurance, transport, etc.) qui peuvent être mobilisé pour développer des solutions rapidement opérationnelles sur le terrain et capable de traiter les grand volumes de données documentaires générées par les documents réglementaires relatifs à la biodiversité, mais une attention particulière doit être dédié à l'association avec des professionnels de la biodiversité pour en assurer le développement efficace pour une éventuelle utilisation en support de rédaction ou révision de documents réglementaires.