

Génération de données synthétiques pour des domaines spécialisés – H/F

Stage de 6 mois @ EDF R&D Saclay (91120)

Début du stage : mars 2026

Contexte

La R&D d'EDF (2000 chercheurs) a pour missions principales de contribuer à l'amélioration de la performance des unités opérationnelles du groupe EDF ainsi que d'identifier et de préparer les relais de croissance à moyen et long terme. Dans ce cadre, le département Services, Economie, Outils Innovants et IA (SEQUOIA) est un département pluridisciplinaire (sciences de l'ingénieur, sciences humaines et sociales) qui fournit un appui à l'élaboration et au portage des offres, des services et des outils de relation client aux directions opérationnelles du groupe EDF.

Au sein de ce département, ce stage sera rattaché au groupe « Statistiques et Outils d'Aide à la Décision » (SOAD) : cette équipe compte une vingtaine d'ingénieurs chercheurs spécialisés en IA et data science avec des compétences fortes autour du machine learning et du deep learning, du web sémantique, de l'IA symbolique et de l'IA générative (texte, voix, image, multimodalité...), en particulier du NLP (LLM, RAG, data mining...).

Objectifs

Les modèles d'intelligence artificielle pour le traitement des données textuelles ont connu des avancées techniques importantes depuis la création de l'architecture *Transformer*. D'une part, les grands modèles de langue (LLM) génératifs démontrent des capacités raisonnement avancé pour des tâches complexes. D'autre part, les modèles encodeurs (type BERT) permettent de calculer des représentations vectorielles capables de représenter finement la sémantique des textes.

Toutefois, l'entraînement et l'adaptation (fine-tuning) de ces modèles à des tâches spécifiques nécessitent une quantité importante de données. Dans les domaines industriels hautement spécialisés, l'acquisition de ces données annotées est souvent complexe et coûteuse. Pour pallier ce défi, la génération de données synthétiques s'impose comme une piste de recherche prometteuse, notamment pour l'entraînement de modèles de similarité sémantique dédiés à la recherche d'information (Information Retrieval) [1,2,3,4].

L'accès efficace à l'information est un enjeu stratégique pour EDF. Notre système d'informations héberge plusieurs millions de documents techniques. Optimiser la recherche d'information au sein de ce vaste corpus est crucial pour le travail quotidien de nos ingénieurs. C'est dans ce contexte que s'inscrit ce stage, dont l'objectif est d'explorer et d'implémenter des méthodes de génération de données synthétiques pour entraîner des modèles de similarité performants, spécifiquement adaptés à notre corpus technique dans le domaine du nucléaire.

Encadré(e) par nos ingénieurs-rechercheurs, vous aurez pour missions principales de :

- **Réaliser un état de l'art** approfondi sur les méthodes de génération de données synthétiques (ex : usage de LLM génératifs, back-translation, etc.) pour l'adaptation de modèles de langue à des tâches de similarité.
- **Sélectionner et implémenter** les approches les plus prometteuses pour notre cas d'usage industriel.
- **Développer le pipeline d'entraînement** (fine-tuning) des modèles encodeurs sur les données (réelles et synthétiques) générées.
- **Concevoir et mettre en œuvre** un protocole d'évaluation rigoureux pour mesurer la performance des modèles entraînés sur des tâches de recherche d'information réelles.

Ce stage est une immersion complète en R&D, couvrant l'ensemble du cycle de vie d'un projet de Deep Learning, de la recherche bibliographique à l'évaluation en contexte industriel.

Le/la stagiaire développera une triple compétence pointue :

- **Expertise Technique en NLP (Deep Learning)** : Maîtrise des architectures SOTA (Transformers), des techniques de génération de données synthétiques, et de l'implémentation de pipelines de fine-tuning et d'évaluation de modèles de similarité.
- **Compétence MLOps (Calcul Intensif)** : Capacité à opérer dans un environnement de calcul haute performance (HPC), incluant la gestion de jobs d'entraînement distribué sur plusieurs dizaines de GPU.
- **Compétence Méthodologique** : Capacité à mener un état de l'art sur un sujet de recherche avancé, à proposer et tester des solutions innovantes, et à appliquer une méthodologie scientifique rigoureuse pour répondre à un enjeu industriel critique (la recherche d'information en domaine spécialisé).

Références

- [1] Wu, Ian, Sravan Jayanthi, Vijay Viswanathan, et al. « Synthetic Multimodal Question Generation ». In Findings of the Association for Computational Linguistics: EMNLP 2024. Association for Computational Linguistics, 2024. <https://doi.org/10.18653/v1/2024.findings-emnlp.759>.
- [2] Kim, Seunghee, Changhyeon Kim, et Taeuk Kim. « FCMR: Robust Evaluation of Financial Cross-Modal Multi-Hop Reasoning ». In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2025. <https://doi.org/10.18653/v1/2025.acl-long.1138>.
- [3] Yang, Yuming, et al. "Benchmarking Multimodal RAG through a Chart-based Document Question-Answering Generation Framework." *arXiv preprint arXiv:2502.14864* (2025).
- [4] Shen, Zhiyu, et al. "HopWeaver: Synthesizing Authentic Multi-Hop Questions Across Text Corpora." *arXiv preprint arXiv:2505.15087* (2025).

Profil recherché

- Étudiant(e) en dernière année d'École d'Ingénieur ou en Master 2 Recherche, avec une spécialisation en Data Science, Machine Learning ou Traitement Automatique du Langage (NLP).
- Solide maîtrise de Python et des librairies de Deep Learning (PyTorch).
- Bonnes connaissances théoriques en Machine Learning et Deep Learning (notamment les Transformers).

- Une première expérience (projets académiques ou stages) en NLP sera appréciée.
- Bon niveau de rédaction en français et en anglais
- Curiosité scientifique, intérêt pour la recherche

Informations pratiques

Début du stage : mars/avril 2026

Durée du stage : 6 mois

Unité d'accueil : Groupe Statistique et Outils d'Aide à la Décision (SOAD), département Services, Economie, Outils Innovants et IA (SEQUOIA) – EDF Lab Paris-Saclay, 7 boulevard Gaspard Monge, 91120 Palaiseau.

Télétravail 2j/semaine

Transmettre par mail un CV et une lettre de motivation à :

leila.hassani@edf.fr, julien.tourille@edf.fr