

OFFRE DE STAGE EN NLP

Mission : Développer un système d'IA générative pour interroger une base de données SQL

Durée : 6 mois - Début du stage souhaité : mars 2026 - Lieu : EDF R&D Lab Saclay (91120)

Contexte et objectifs

La R&D d'EDF a pour missions principales de contribuer à l'amélioration de la performance des unités opérationnelles du Groupe EDF, d'identifier et de préparer les relais de croissance à moyen et long terme. Dans ce cadre, le département SEQUOIA (Services, Economie, Questions humaines, Outils innovants et IA) d'EDF R&D est un département pluridisciplinaire (sciences des données et IA, offres et services innovants aux clients, sciences humaines et sociales, économie et stratégies) qui fournit un appui à l'élaboration et au portage des offres, des services et des outils aux Directions opérationnelles du Groupe EDF. Au sein de ce département, le stage sera rattaché au groupe SOAD (Statistiques et Outils d'Aide à la Décision) qui compte une vingtaine d'ingénieurs chercheurs spécialisés en Data Science, Data Engineering, Informatique Décisionnelle et NLP. Cette équipe a pour mission de construire et mettre en œuvre les méthodes d'analyse, de fouille et d'enrichissement de données volumineuses d'origines multiples, structurées ou complexes. Le(a) stagiaire sera amené(e) à interagir et évoluer dans un cadre collaboratif avec d'autres chercheurs travaillant sur des problématiques communes au Groupe EDF.

L'ambition de ce stage serait de concevoir un système qui permettrait à un utilisateur d'interagir en langage naturel avec une base de données SQL. Les données manipulées seront spécifiques au réseau de distribution d'électricité. Les LLMs sont utilisés pour générer les requêtes SQL à partir des questions métiers et pour synthétiser une réponse à partir des résultats d'exécution des requêtes générées.

Ce sujet présente plusieurs défis :

- Spécificité des données qui s'inscrivent dans un contexte métier (connaissances, norme CIM, vocabulaire).
- Mise à l'échelle du système pour interaction avec une grande base de données.
- Prise en charge éventuelle de données au format graphe.
- Evaluation du système.

Objectifs

L'objectif général du stage sera donc d'explorer et de tester des méthodes qui permettraient d'une part d'arriver à un système performant de génération de requêtes SQL et d'autre part de synthétiser une réponse pour l'utilisateur. Des travaux sont en cours sur le sujet, l'idée est de pousser les limites identifiées en collaborant avec les data scientists d'EDF R&D.

Pour cela, le stagiaire devra :

- Réaliser un état de l'art sur les méthodes de génération de requêtes SQL et identifier les pistes pertinentes.
- Prendre en main des données spécifiques au réseau de distribution d'électricité.
- Proposer et tester des méthodes pour améliorer les résultats existants ainsi que des méthodes d'évaluation du système.

- Participer aux réunions avec le métier pour comprendre leurs besoins spécifiques au cas d'usage, et aux points d'avancement et démonstrations.
- Enrichir le « gold standard » d'évaluation et éventuellement tester le système sur des benchmarks existants de génération de SQL.

Profil recherché :

- Formation : M2, Ecole d'Ingénieur ou équivalent dans les domaines de compétences : IA générative, NLP, Data Science.
- Bonne connaissance en Python, SQL, frameworks LLM / agents comme Langchain.
- Bonne connaissance des LLMs, prompt engineering, fine-tuning.
- Confortable avec la prise en main de nouvelles bibliothèques et outils.
- Esprit scientifique.
- Motivation pour la prise en main du cas d'usage et des données métiers.
- Autonome, force de proposition et orienté résolution de problème.

Informations pratiques

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

Le stage sera encadré par des Data Scientists du département SEQUOIA.

Transmettre par mail un CV et une lettre de motivation à: eric.lavergne@edf.fr et kovila.pauvaday@edf.fr (Département SEQUOIA).