

Proposition de stage de fin d'études Bac+5 en informatique

Simuler le comportement humain par une approche RAG LLM

Contexte et enjeux

La transition énergétique est un enjeu majeur pour répondre aux défis du changement climatique. Pour étudier des scénarios de décarbonation de l'énergie, d'intégration des énergies renouvelables et de flexibilité énergétique, le département SEQUOIA de la R&D d'EDF développe des approches basées sur la simulation et l'IA, et notamment la plateforme SMACH (Simulation de l'activité humaine et des consommations d'énergie dans l'habitat) [1], développée en collaboration avec le Laboratoire Interdisciplinaire des Sciences du Numérique.

SMACH s'appuie sur une enquête statistique de l'INSEE appelée « Emploi du Temps » (*Time Use Survey* - TUS) décrivant les activités de vie quotidienne d'une ou deux journées de chaque répondant. La limite des TUS est qu'ils décrivent des journées éparpillées des français, sans continuité de jour en jour. Des travaux récents menés à EDF R&D ont montré que l'utilisation de grands modèles de langage (LLM) permet de générer des emplois du temps à l'échelle de la semaine, avec davantage de diversité que dans l'approche SMACH actuelle. On souhaite toutefois mieux cadrer la cohérence et la plausibilité des emplois du temps générés par LLM.

Objectif du stage

L'objectif de ce stage est d'**explorer comment l'injection de connaissances qualitatives sur le comportement humain, issues des sciences cognitives et des sciences humaines et sociales, peut améliorer la génération d'emplois du temps par LLM. L'approche envisagée est la Retrieval-Augmented Generation (RAG)**. Le but est de dépasser la simple imitation statistique des données TUS données en prompt au LLM, pour produire des narratifs de vie plus cohérents et crédibles. Le LLM ne se basera plus uniquement sur une ou deux journées d'exemple, mais aussi sur des connaissances expertes décrivant des logiques comportementales identifiées lors d'études terrain (ex : coordination au sein du ménage, comportements si présence d'enfants, gestion des imprévus, articulation vie professionnelle/vie personnelle...).

Déroulé

Le stage allie recherche bibliographique, ingénierie des données et mise en œuvre pratique. Il vise à créer une preuve de concept et d'en faire une analyse critique :

- **Analyse bibliographique et constitution d'un corpus de connaissances** : Dans un premier temps, vous vous familiariserez avec les travaux précédents sur la génération d'activité par LLM. Vous réaliserez une synthèse sur les techniques de RAG existant dans le contexte du stage [2],[3]. Une tâche clé sera ensuite d'identifier, collecter et structurer un corpus de documents (rapports d'études, articles scientifiques en sciences humaines et sociales, publications internes EDF) décrivant les routines et les logiques d'organisation de la vie quotidienne des ménages. Ce travail pourra se faire avec l'appui des chercheurs en ergonomie et sociologie du Département SEQUOIA.
- **Conception et implémentation d'une chaîne RAG** : Vous devrez ensuite concevoir et implémenter une chaîne RAG complète. Cela impliquera de mettre en place une base

de données vectorielle pour indexer le corpus de connaissances, de choisir une stratégie pour extraire les informations les plus pertinentes en fonction du profil de l'individu simulé, et d'intégrer ces informations dans le prompt du LLM, avec une approche de prompt engineering permettant de gérer les hallucinations potentielles.

- **Évaluation** : L'enjeu final sera d'évaluer l'apport de cette approche. En vous appuyant sur le cadre d'évaluation existant et les échanges avec les chercheurs en sciences humaines et sociales, vous proposerez des approches pour analyser l'apport de l'approche RAG sur la cohérence des comportements simulés (ex : analyse quantitative des habitudes de vie, analyse qualitative manuelle sur des exemples emblématiques...).

Les résultats du stage pourront faire l'objet d'une publication scientifique.

Profil recherché

Bac+5 en informatique : étudiant en dernière année de Master 2 ou école d'ingénieur orienté intelligence artificielle, traitement du langage naturel (NLP) ou sciences des données. Le candidat doit avoir une bonne compréhension des modèles de langage et des techniques d'apprentissage. Une connaissance des techniques de RAG (Retrieval-Augmented Generation) est un atout. Une appétence ou expérience sur le sujet du comportement humain et les sciences humaines et sociales est un plus.

Encadrement, informations pratiques

Le stage se déroulera principalement à EDF Lab Paris Saclay (Palaiseau) mais vous pourrez aussi bénéficier des infrastructures du LISN, situé à 10 minutes en bus sur le campus de l'université voisine.

Encadrement : Mathieu Schumann (EDF R&D), Julien Tourille (EDF R&D) et Nicolas Sabouret (Université Paris-Saclay)

Lieu du stage : EDF R&D, 7 boulevard Gaspard Monge, 91120 Palaiseau

Durée : 6 mois

Rémunération indicative : 80-85% du SMIC brut par mois (donnée au 1^{er} janvier 2026)

Contact

Mathieu SCHUMANN

Tél : 06 98 25 85 83

E-mail : mathieu.schumann@edf.fr

Références

[1] *A multi-sourced data and agent-based approach for complementing Time Use Surveys in the context of residential human activity and load curve simulation*. Mathieu Schumann, Quentin Reynaud et al. 2023, In Proc. 18th International IPBSA Conference and Exhibition, Building Simulation (BS). <https://perso.limsi.fr/sabouret/ps/BS2023.pdf>

[2] *Human Mobility Modeling with Household Coordination Activities Under Limited Information via Retrieval-Augmented LLMs*. Yifan Liu, Xishun Liao et al. 2025, In 2025 IEEE 28th International Conference on Intelligent Transportation Systems (ITSC). <https://dx.doi.org/10.1109/itsc60802.2025.11423539>

[3] *Large language models empowered agent-based modeling and simulation: A survey and perspectives*. Chen Gao et al., 2024, In Humanities and Social Sciences Communications (11:1). <https://github.com/tsinghua-fib-lab/LLM-Agent-Based-Modeling-and-Simulation>