

Sujets de Stages/Alternances : Consolidation et Développement d'une Plateforme d'IA Générative

Contexte

Depuis l'arrivée fracassante de ChatGPT dans le grand public en novembre 2022, l'intelligence artificielle générative a rapidement suscité une prise de conscience quant à son potentiel de bouleversement dans les métiers et l'innovation. Au CEA, cette révolution technologique a été perçue comme une opportunité majeure pour transformer nos processus et accélérer notre capacité d'innovation. Cependant, l'utilisation de ces outils dans un cadre sécurisé et souverain s'est avérée être un défi de taille. Face à cette situation, nous avons initié un recueil des besoins au sein de l'entreprise, en organisant des sessions de brainstorming pour explorer les usages possibles de l'IA générative. Ces discussions ont permis de définir des cas d'usage concrets et de poser les bases d'une expérimentation au niveau corporate.

Avec le soutien fort de la Direction Innovation et de la Direction des Systèmes d'Information, nous avons lancé une expérimentation visant à étudier l'intégration de l'IA générative dans nos processus internes. Cette initiative vise non seulement à exploiter pleinement le potentiel de cette technologie, mais aussi à garantir que son adoption se fasse de manière sécurisée et conforme aux exigences de souveraineté numérique. Cette phase a permis de créer une première version d'une plateforme d'IA générative mettant différents applicatifs de type IA générative (texte, image, code).

Objectifs des Stages/Alternances

Nous proposons une série de trois stages/alternances visant à consolider et développer notre plateforme d'IA générative. Chaque stage/alternance se concentrera sur un aspect spécifique de la plateforme :

1. **Robustification de la Plateforme d'IA Générative**
2. **Développement de Workflows Métiers Basés sur LangFlow/LangChain**
3. **Consolidation d'une Approche RAG/GraphRAG avec Connecteurs à Nos Bases de Données**

1. Robustification de la Plateforme d'IA Générative

Objectif

L'objectif de ce stage/alternance est de consolider et de robustifier notre plateforme d'IA générative basée sur l'utilisation de modèles de langage de grande taille (LLM) comme MistralLarge, de moteurs d'inférence comme Ollama ou Vllm, de proxy LiteLLM et d'une interface de type ChatUI Openwebui.

Missions

- **Consolidation de l'architecture** : Analyser et optimiser l'architecture actuelle de la plateforme pour garantir sa scalabilité, sa performance et sa sécurité.

- **Robustification des conteneurs Docker** : Améliorer la robustesse des conteneurs Docker en optimisant les configurations, en mettant en place des mécanismes de surveillance et de redémarrage automatique, et en assurant la sécurité des conteneurs.
- **Intégration des LLM et moteurs d'inférence** : Assurer l'intégration fluide des modèles de langage de grande taille et des moteurs d'inférence dans la plateforme, en veillant à ce que les interactions entre les différents composants soient efficaces et sécurisées.
- **Documentation et formation** : Rédiger des documents techniques et des guides d'utilisation pour les utilisateurs et les développeurs, et organiser des sessions de formation pour assurer une adoption en douceur de la plateforme.

Profil Recherché

- **Formation** : Étudiant en dernière année d'école d'ingénieur ou en master spécialisé en informatique, avec une spécialisation en DevOps.
- **Compétences techniques** :
 - Maîtrise des conteneurs Docker et des orchestrateurs comme Kubernetes.
 - Connaissances en développement logiciel (Python, JavaScript, etc.).
 - Expérience avec les modèles de langage de grande taille (LLM) et les moteurs d'inférence.
 - Connaissance des outils de CI/CD et des pratiques DevOps.
 - Bonnes compétences en communication et en rédaction technique.
- **Qualités personnelles** :
 - Rigueur et sens de l'organisation.
 - Capacité à travailler en équipe et à communiquer efficacement.
 - Curiosité intellectuelle et goût pour l'innovation.

2. Développement de Workflows Métiers Basés sur LangFlow/LangChain

Objectif

L'objectif de ce stage/alternance est de développer des workflows métiers basés sur les frameworks LangFlow et LangChain, afin d'automatiser et d'optimiser nos processus internes.

Missions

- **Analyse des besoins métiers** : Collaborer avec les différents départements pour identifier les besoins spécifiques et définir les workflows à automatiser.
- **Développement des workflows** : Utiliser LangFlow (no-code) et LangChain pour concevoir et implémenter des workflows métiers automatisés.

- **Intégration avec la plateforme d'IA générative** : Assurer l'intégration fluide des workflows développés avec la plateforme d'IA générative existante.
- **Tests et validation** : Effectuer des tests rigoureux pour valider les workflows et garantir leur fiabilité et leur efficacité.
- **Documentation et formation** : Rédiger des documents techniques et des guides d'utilisation pour les utilisateurs et les développeurs, et organiser des sessions de formation pour assurer une adoption en douceur des workflows.

Profil Recherché

- **Formation** : Étudiant en dernière année d'école d'ingénieur ou en master spécialisé en informatique, avec une spécialisation en développement logiciel.
- **Compétences techniques** :
 - Maîtrise des frameworks LangFlow et LangChain.
 - Connaissances en développement logiciel (Python, JavaScript, etc.).
 - Expérience avec les modèles de langage de grande taille (LLM) et les moteurs d'inférence.
 - Connaissance des outils de CI/CD et des pratiques DevOps.
 - Bonnes compétences en communication et en rédaction technique.
- **Qualités personnelles** :
 - Rigueur et sens de l'organisation.
 - Capacité à travailler en équipe et à communiquer efficacement.
 - Curiosité intellectuelle et goût pour l'innovation.

3. Consolidation d'une Approche RAG/GraphRAG avec Connecteurs à Nos Bases de Données

Objectif

L'objectif de ce stage/alternance est de consolider une approche RAG (Retrieval-Augmented Generation) et GraphRAG avec des connecteurs à nos bases de données, afin d'enrichir les réponses générées par notre plateforme d'IA générative.

Missions

- **Analyse des bases de données** : Étudier les différentes bases de données disponibles et identifier les sources d'information pertinentes pour enrichir les réponses générées.
- **Développement des connecteurs** : Concevoir et implémenter des connecteurs pour interfacer notre plateforme d'IA générative avec nos bases de données.

- **Intégration RAG/GraphRAG** : Mettre en place une approche RAG/GraphRAG pour enrichir les réponses générées par les modèles de langage de grande taille.
- **Tests et validation** : Effectuer des tests rigoureux pour valider l'intégration et garantir la qualité des réponses générées.
- **Documentation et formation** : Rédiger des documents techniques et des guides d'utilisation pour les utilisateurs et les développeurs, et organiser des sessions de formation pour assurer une adoption en douceur de l'approche RAG/GraphRAG.

Profil Recherché

- **Formation** : Étudiant en dernière année d'école d'ingénieur ou en master spécialisé en informatique, avec une spécialisation en bases de données et en intelligence artificielle.
- **Compétences techniques** :
 - Maîtrise des bases de données et des systèmes de gestion de bases de données (SQL, NoSQL, etc.).
 - Connaissances en développement logiciel (Python, JavaScript, etc.).
 - Expérience avec les modèles de langage de grande taille (LLM) et les moteurs d'inférence.
 - Connaissance des approches RAG et GraphRAG.
 - Bonnes compétences en communication et en rédaction technique.
- **Qualités personnelles** :
 - Rigueur et sens de l'organisation.
 - Capacité à travailler en équipe et à communiquer efficacement.
 - Curiosité intellectuelle et goût pour l'innovation.

Modalités

- **Durée** : 6 à 12 mois.
- **Lieu** : [Lieu du stage/alternance].
- **Rémunération** : Selon les barèmes en vigueur.

Pointeurs

- **Openwebui** : <https://github.com/open-webui/open-webui>
- **GraphRag** : <https://github.com/microsoft/graphrag>
- **LightRag** : <https://github.com/HKUDS/LightRAG>
- **Langflow** : <https://github.com/langflow-ai/langflow>
- **Gradio** : <https://www.gradio.app/>

Si vous êtes passionné par l'IA générative et souhaitez contribuer à un projet innovant et stratégique au CEA, rejoignez-nous pour cette aventure technologique.

Pour postuler, merci d'envoyer votre CV et une lettre de motivation à francois.le-fevre@cea.fr en précisant la référence du stage/alternance.