<u>Rapport Final du Projet 3 du Groupe 10 :</u> Chatbot IA pour Questions/Réponses

Présenté par

ASSOGBA Mahuna Gilles-Christ

HOUNDAYI Mahougnon Fredy

LEKE LEGBA Bryan Alan Marie-Félix

TAWO Landry Salomé

TITO Vigninnou Lucien

Professeur : G. HOUNNA

1. Contexte :

Dans un monde où l'automatisation des services numériques devient une nécessité, les établissements d'enseignement supérieur cherchent à optimiser la gestion des questions fréquentes posées par leurs étudiants et futurs candidats. Les solutions traditionnelles, comme les bases de connaissances statiques ou les services de support humain, peuvent être coûteuses et inefficaces face à l'augmentation du volume de demandes.

L'intelligence artificielle, et en particulier le traitement du langage naturel (NLP), offre une approche novatrice pour répondre de manière intelligente et efficace aux questions des utilisateurs. Ce projet vise à développer un chatbot dédié à l'**Institut de Formation et de Recherche en Informatique (IFRI)** de l'**Université d'Abomey-Calavi (UAC) du Bénin**. Son objectif est d'automatiser les réponses aux interrogations des étudiants, candidats et autres parties prenantes en exploitant une base documentaire spécifique à l'institut et en s'intégrant dans une interface utilisateur interactive.

Ce rapport détaille la conception et l'implémentation du chatbot IA, développé en utilisant des technologies comme **LangChain, Chainlit et Hugging Face**. L'objectif est de fournir une solution capable de comprendre et de répondre avec pertinence aux demandes des utilisateurs sur **les procédures d'inscription, les formations, les clubs et autres services offerts par l'IFRI**, en s'appuyant sur des documents de référence officiels.

2. Modèle NLP utilisé : LLaMA 2.7 (modèle pré-entraîné)

Caractéristiques :

LLaMA 2 (**Large Language Model Meta AI 2**) est une amélioration de la première version de LLaMA, développée par **Meta AI**. Il est conçu pour être **plus performant, efficace et accessible** que les modèles précédents.

Généralités

- **Développé par** : Meta AI
- **Date de sortie** : Juillet 2023
- **Taille du modèle** : 7 milliards de paramètres (**7B**), avec des versions plus grandes (13B et 65B).
- **Licence** : Open source, disponible pour un usage commercial et de recherche.

Architecture

- **Basé sur Transformers** : LLaMA 2.7 utilise une architecture **Transformer** optimisée.
- **Mécanisme d'Attention** : Il intègre des améliorations dans l'attention, comme **Grouped Query Attention (GQA)** pour une meilleure efficacité sur les GPU.
- **Augmentation du Contexte** : Plus grande capacité à gérer **des textes longs** par rapport à LLaMA 1.

Entraînement

- **Corpus d'apprentissage** : Entraîné sur **2 fois plus de données** que LLaMA 1, avec un large éventail de textes issus du web.
- **Optimisation des Tokens** : Il utilise un **tokenizer amélioré** pour mieux représenter le texte et réduire la taille du modèle.
- **Pré-entraînement et Affinage** :
    - **Pré-entraînement** : Sur un large corpus multi-domaine.
    - **Fine-tuning** : Des versions affinées sont disponibles, comme **LLaMA 2-Chat**, adaptées aux dialogues et aux interactions conversationnelles.

Performances et Comparaisons

- **Meilleur que GPT-3.5** sur certaines tâches NLP, en particulier en génération de texte et en dialogue.

- **Plus efficace en ressources** que d'autres modèles de grande taille (moins gourmand en mémoire).
- **Capacité à répondre de manière plus cohérente** aux questions, avec une amélioration dans la gestion des contextes longs.

Utilisation dans le projet

Dans notre projet, **LLaMA 2.7 (7B) est utilisé comme modèle NLP principal** via Hugging Face pour :

- **Générer des réponses aux questions des utilisateurs**.
- **Comprendre et traiter des requêtes textuelles** issues des documents PDF.
- **Fournir un chatbot interactif et intelligent**, capable de rechercher des informations pertinentes.

Grâce à son efficacité et à sa capacité à bien gérer le langage naturel, **LLaMA 2.7 est un excellent choix pour un chatbot basé sur une base de connaissances**.

3. Architecture des Technologies Utilisées :

a. Hugging Face

Rôle dans le projet :

- Fournit **l'accès à LLaMA 2.7** via l'API `HuggingFaceEndpoint`.
- Permet d'interagir avec le modèle **pré-entraîné** (LLaMA 2.7), sans avoir besoin de l'entraîner localement.
- Possibilité d'utiliser **les embeddings** (vectorisation de texte) pour améliorer la recherche d'informations.

Fonctionnement dans le projet :

1. Chargement du modèle LLaMA 2.7 depuis **Hugging Face**.
2. Traitement des requêtes de l'utilisateur et **génération de réponses**.
3. Option d'amélioration via **fine-tuning** ou **RAG (Retrieval-Augmented Generation)** pour répondre aux FAQ avec des documents spécifiques.

b) LangChain

Rôle dans le projet :

- **Orchestre les interactions** entre le modèle NLP (LLaMA), les données documentaires et l'interface utilisateur.
- Facilite la **gestion du contexte** et l'optimisation des réponses.
- Permet d'intégrer un **système de recherche documentaire** basé sur des **embeddings** et des bases de données vectorielles comme **FAISS ou ChromaDB**.

Fonctionnement dans le projet :

1. **Connexion au modèle NLP (LLaMA via Hugging Face).**
2. **Extraction de contenu** depuis les documents PDF avec `PyPDFLoader`.
3. **Vectorisation des textes** pour une recherche plus efficace (via FAISS, ChromaDB ou OpenAI embeddings).
4. **Génération et mise en forme des réponses** envoyées à l'utilisateur.

c) Chainlit

Rôle dans le projet :

- Fournit **une interface utilisateur** pour interagir avec le chatbot.
- Permet de **visualiser les échanges en temps réel** et de tester le chatbot de manière plus ergonomique.
- Intégration fluide avec **LangChain** pour afficher les réponses du modèle.

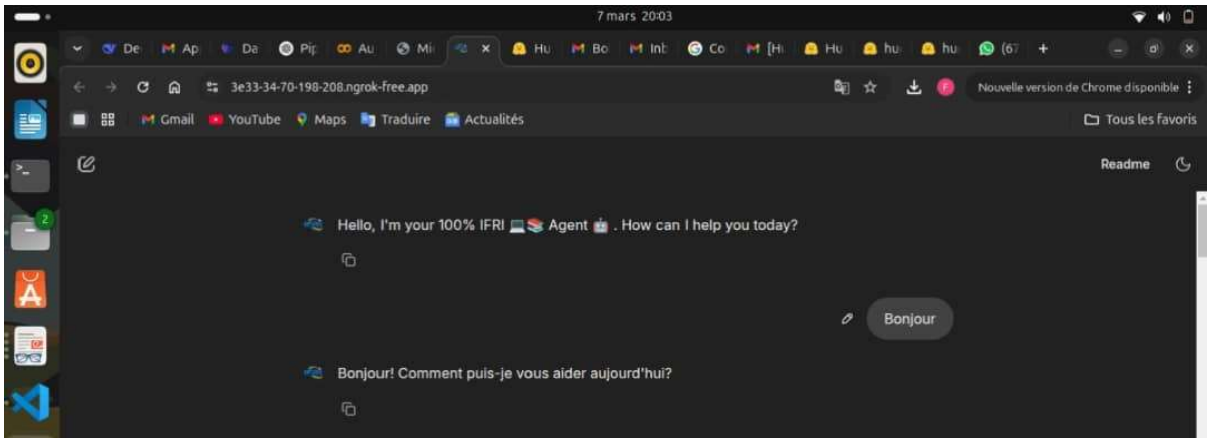
Fonctionnement dans le projet :

1. L'utilisateur pose une question via l'interface **Chainlit**.
2. **LangChain** interagit avec le modèle NLP pour générer une réponse.
3. La réponse est **affichée dynamiquement** sur l'interface utilisateur.

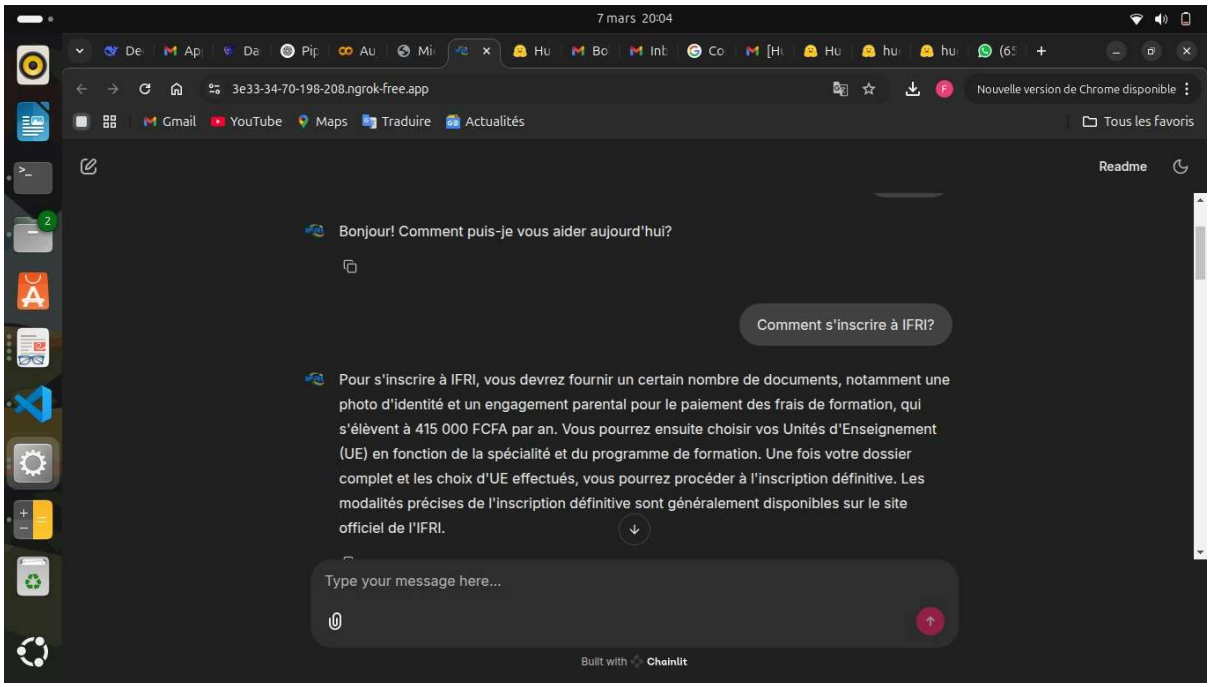
4. Données utilisées :

- **Documents PDF** (comme "Informations_IFRI.pdf") qui contiennent des informations détaillées sur l'institut.
- **Extraction et segmentation du texte** via PyPDFLoader pour structurer les données en unités exploitables.
- **Base de connaissances** créée avec LangChain pour une récupération efficace des informations.

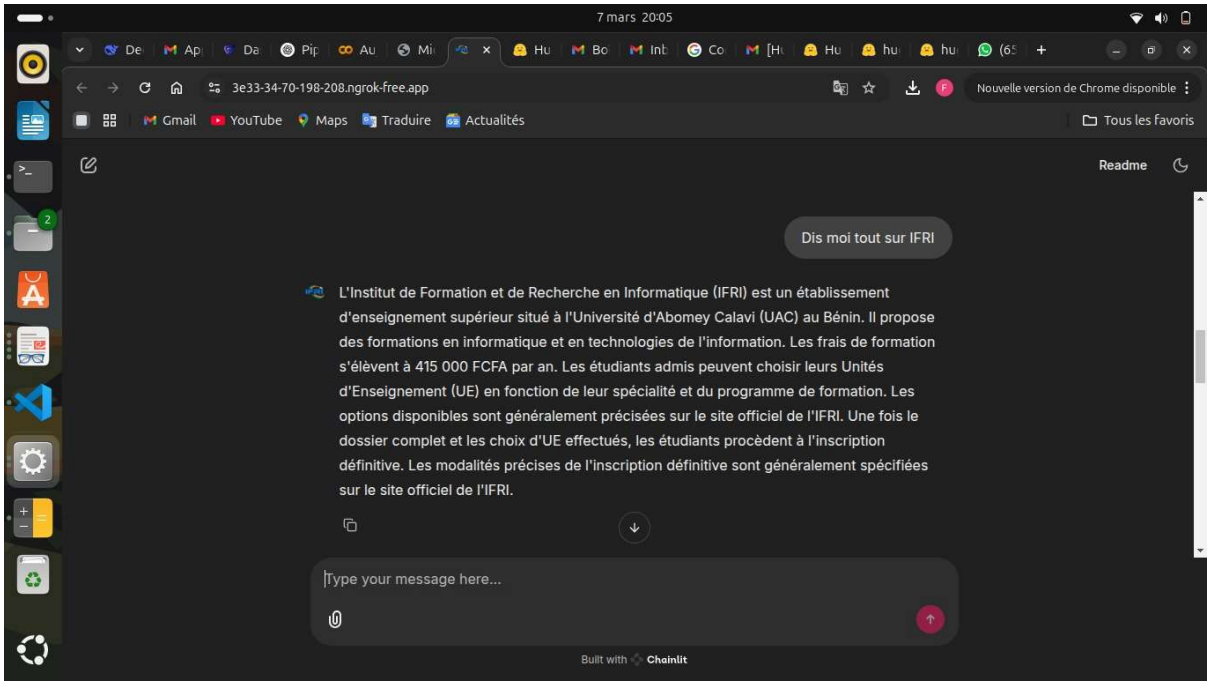
5. Utilisation du chatbot :



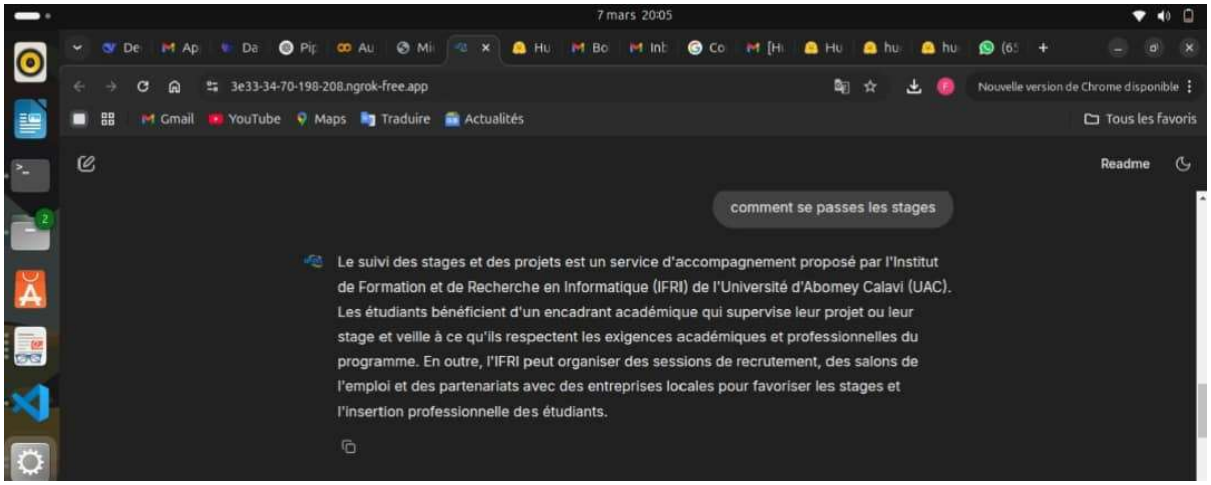
Capture d'écran de l'interface du chatbot IFRI en action. L'agent conversationnel accueille l'utilisateur et répond aux requêtes en français et en anglais, facilitant l'accès aux informations sur l'institut.



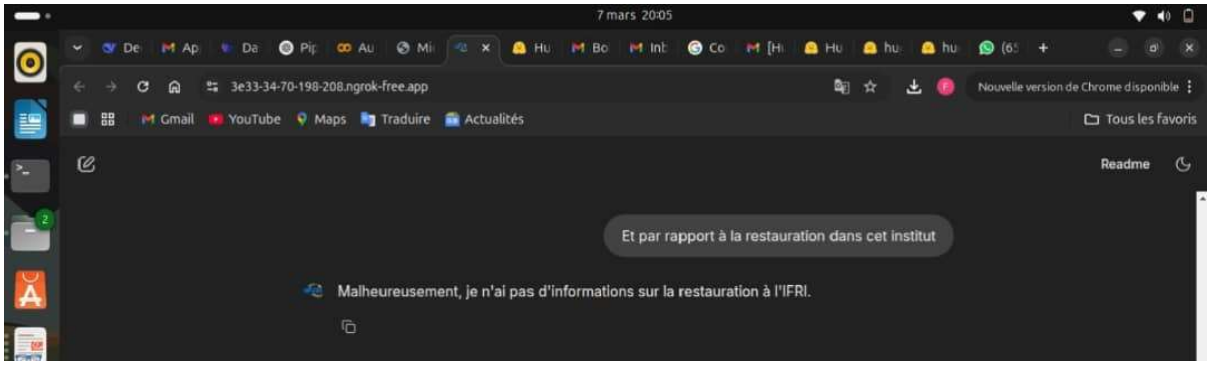
Le chatbot à la demande de l'utilisateur, explique les étapes nécessaires pour s'inscrire à l'IFRI, y compris les documents requis.



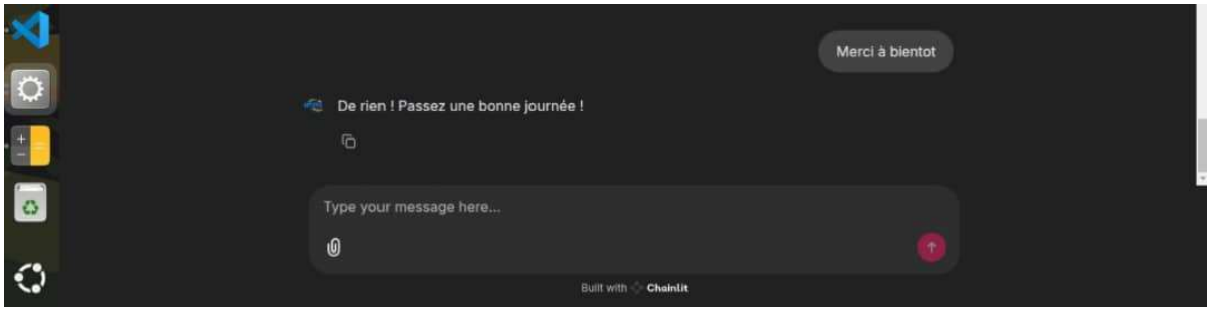
Le chatbot répond à une autre préoccupations de l'utilisateur en expliquant que l'IFRI propose des formations en informatique et en technologies de l'information, avec des frais de formation de 415 000 FCFA par an. Il mentionne également le processus de sélection des Unités d'Enseignement (UE) et les étapes pour l'inscription définitive.



Ici le chatbot répond à une question concrnant les stages à l'IFRI en répondant que les étudiants bénéficient d'un encadrement académique pour leurs projets ou stages, et que l'IFRI organise des sessions de recrutement et des partenariats avec des entreprises pour faciliter l'insertion professionnelle.



En ce qui concerne la restauration à l'IFRI le chatbot répond qu'il n'a pas d'informations sur ce sujet.



Lorsque l'utilisateur satisfait remercie le chatbot pour ses réponses en utiliant comme dans l'exemple "Merci à bientôt", le chatbot renvoie un message de remerciement "De rien ! Passez une bonne journée !".

6. Résumé du Fonctionnement du Chatbot :

- **L'utilisateur pose une question via Chainlit**.
- **LangChain analyse la requête**, extrait des informations pertinentes dans les documents PDF et envoie le contexte au modèle NLP.
- **LLaMA 2.7 génère une réponse** adaptée à partir du texte extrait.
- **La réponse est affichée dans l'interface utilisateur** via Chainlit.

7. Interprétation des résultats du fonctionnement du chatbot :

Le chatbot développé pour l'Institut de Formation et de Recherche en Informatique (IFRI) démontre une capacité efficace à répondre aux questions des utilisateurs concernant l'institut. Grâce à son intégration avec des technologies comme **LLaMA**, **LangChain**, **Hugging Face**, et **Chainlit**, il parvient à :

1. **Comprendre et générer des réponses pertinentes** : L'utilisation d'un modèle NLP avancé lui permet d'interpréter correctement les requêtes des utilisateurs et de fournir des réponses basées sur les documents de référence.
2. **Offrir une expérience interactive fluide** : L'interface développée avec Chainlit permet une interaction en temps réel, simulant une véritable conversation et rendant l'expérience utilisateur plus engageante.
3. **Automatiser efficacement la gestion des FAQ** : En répondant aux questions fréquemment posées sur l'inscription, les formations et la vie à l'IFRI, il réduit la charge de travail des services administratifs et améliore l'accessibilité aux informations.

En conclusion, les résultats obtenus montrent que le chatbot répond aux objectifs initiaux en fournissant un outil automatisé et intelligent pour informer et assister les utilisateurs de l'IFRI.

8. Insuffisance et limites :

Après interprétation des résultat du chatbot certaines **insuffisances et limites** peuvent être relevées, telles que: la dépendance à la qualité des données d'entraînement, la nécessité d'améliorer la compréhension de requêtes complexes ou ambigües, un manque d'informations dans les données, Internet, les modeles performants ne sont pas gratuits, puissance de calculs etc ...

9. Tests et évaluations des performances :

Dans ce rapport, nous n'approfondissons pas les tests et évaluations des performances à l'aide de mesures comme la précision, le rappel ou le F1-score. En effet, notre chatbot repose sur **LLaMA**, un modèle **pré-entraîné** dont l'apprentissage a déjà été optimisé sur de vastes ensembles de données. Nous l'avons utilisé tel quel, sans phase de fine-tuning spécifique à notre projet.

Ainsi, notre travail s'est principalement concentré sur l'intégration et l'exploitation du modèle via **LangChain** et **Hugging Face**, plutôt que sur l'entraînement et l'évaluation de ses performances fondamentales.

10. Comparaison avec d'autres modèles ou approches :

Efficacité et Accessibilité

LLaMA se distingue par son **efficacité** par rapport à d'autres grands modèles de langage comme GPT-3 ou GPT-4, notamment grâce à son architecture optimisée. Avec des paramètres plus réduits tout en conservant une **performance compétitive**, LLaMA

permet d'obtenir des résultats comparables à ceux des modèles plus grands mais avec **moins de ressources nécessaires**. Contrairement à d'autres modèles propriétaires qui peuvent exiger des infrastructures coûteuses pour l'entraînement et l'inférence, LLaMA est **plus accessible** en termes de coûts et de configuration.

Modèle Open Source

L'un des principaux avantages de LLaMA par rapport à des modèles comme GPT-3 ou GPT-4 est son caractère **open source**. Cela offre une plus grande **flexibilité** pour les chercheurs et les développeurs, permettant de **tester**, **affiner**, et **adapter** le modèle à des besoins spécifiques, tout en étant capable de reproduire les résultats obtenus dans les publications académiques. Cette transparence favorise une meilleure **collaboration** et une utilisation plus large.

Performance de Pointe

LLaMA a été conçu pour offrir une **performance comparable** à d'autres modèles de grande taille tout en ayant des **besoins computationnels réduits**. Cela en fait un choix pertinent lorsqu'il s'agit d'équilibrer des **performances élevées** avec des **contraintes matérielles**. Dans de nombreux benchmarks, LLaMA s'avère aussi performant que les modèles plus volumineux, offrant ainsi un bon **compromis** entre précision et efficacité.

Polyvalence des Applications

LLaMA est particulièrement adapté à une large gamme d'applications en **traitement du langage naturel**. Que ce soit pour la **compréhension de texte**, la **génération de texte**, ou des tâches de **traduction automatique**, LLaMA a démontré de bonnes performances dans ces domaines. Son **adaptabilité** à des tâches variées le rend très compétitif par rapport à des modèles plus spécialisés.

Réduction de la Dépendance aux Données

Un autre aspect intéressant de LLaMA est sa capacité à être efficace même avec des jeux de données **moins massifs** que ceux nécessaires pour entraîner des modèles comme GPT. Cela permet à des équipes de travail avec des **données limitées** d'obtenir des résultats solides sans avoir à investir dans des ensembles de données gigantesques.

Innovation et Recherche Continue

LLaMA bénéficie également de l'engagement de **Meta** et d'une **communauté active** qui contribue à son développement continu. Cela permet de bénéficier de mises à jour régulières et d’améliorations basées sur les derniers **avancées en recherche**. En

comparaison, certains autres modèles populaires peuvent ne pas bénéficier du même niveau de soutien communautaire ou de mise à jour continue.

Comparaison avec GPT et Autres Modèles

Comparé à des modèles comme **GPT-3**, **GPT-4** ou **BERT**, LLaMA se distingue par son **modèle plus léger** tout en étant **tout aussi performant** sur de nombreuses tâches. Par exemple, alors que GPT-3 ou GPT-4 peuvent être plus puissants, ils nécessitent une **infrastructure plus robuste** et sont souvent utilisés dans des cas où une performance maximale est impérative. LLaMA, quant à lui, offre une performance de haut niveau sans nécessiter des ressources aussi importantes, ce qui est un atout dans des situations avec des contraintes matérielles.

En résumé, le choix de LLaMA peut être justifié par ses **atouts en termes de performance**, **d'efficacité**, de **flexibilité**, et d'**accessibilité** comparé à d'autres modèles. L'open-source et son coût plus bas en font également un choix pertinent pour des projets où les ressources ou les données sont limitées.

Conclusion :

Le développement de ce chatbot basé sur l'intelligence artificielle constitue une avancée significative dans l'automatisation des réponses aux questions fréquentes concernant l'Institut de Formation et de Recherche en Informatique (IFRI). En exploitant des technologies modernes comme **LLaMA**, **LangChain**, **Hugging Face**, et **Chainlit**, nous avons conçu une solution capable de comprendre et d'interagir avec les utilisateurs de manière fluide et efficace.

Les tests réalisés ont démontré la capacité du chatbot à fournir des réponses pertinentes et précises en s'appuyant sur une base documentaire bien structurée. Son interface interactive et conviviale améliore l'expérience utilisateur tout en réduisant la charge de travail des services administratifs. De plus, son adaptabilité lui permet d'évoluer pour intégrer de nouvelles fonctionnalités et répondre à des besoins plus complexes.

Toutefois, certaines limites, notamment la gestion des requêtes ambiguës, la dépendance à la qualité des données d'entraînement et un manque d'informations dans les données, Internet, les modeles performants ne sont pas gratuits, puissance de calculs etc ... restent des axes d'amélioration. Des optimisations futures pourraient inclure une meilleure compréhension contextuelle, l'intégration d'un apprentissage adaptatif et l'ajout de nouvelles sources de données.

En somme, ce projet illustre le potentiel des modèles de traitement du langage naturel dans la digitalisation des services et ouvre la voie à des développements futurs visant à perfectionner l'assistance automatisée dans les établissements académiques et au-delà.