Document de présentation du plugin “QuestForgeAI”

Sommaire

Présentation du plugin

L'objectif du plugin **QuestForgeAI** est de faciliter le travail des *game designers* en accélérant leur processus de création et en leur permettant de générer directement des quêtes au sein de l'environnement de développement Unreal Engine.

Intégré à UE, cet outil offre une interface permettant d'interagir avec un assistant virtuel spécialement optimisé pour la génération de quêtes destinées à des jeux vidéo de type RPG ou Action-Aventure.

Ce plugin répond également à un enjeu majeur de l'industrie du jeu vidéo : éviter l'utilisation de grands modèles d'IA publics susceptibles d'exposer des éléments narratifs confidentiels. En conservant le traitement en interne, il limite les risques de fuite d'informations ou de divulgation non contrôlée.

Pour cette raison, le modèle d'IA est conçu pour fonctionner **on-premise**, directement sur l'infrastructure locale du studio, permettant ainsi aux game designers d'accéder en toute sécurité à l'assistant intelligent via le réseau interne.

Installation du serveur

Veuillez installer Ollama : https://ollama.com/

Veuillez récupérer les fichiers sur le lien suivant : https://github.com/jpontoire/LLM_rag_serveur
Veuillez extraire le zip et le mettre dossier où vous le souhaitez.

Ensuite ouvrez l'invite de commande, rendez vous à l'endroit où vous avez mis le dossier "LLM_rag_serveur-main".

Faites (selon votre système d'exploitation) :
./pull-models_Windows.ps1 (Windows) ou ./pull-models_Mac.sh (Mac)

(NB : si vous ne souhaitez pas installer toutes les libraires python, qui seront installés en lançant ces scripts, directement sur votre serveur, il faudra modifier les scripts afin d'exécuter la commande "python -m venv [nom_de_votre_environnement_virtuel]")

puis faites : ./runRAG.bat (sur Windows) ou ./runRAG.sh (sur Mac)

Une fois que vous voyez s'afficher "uvicorn ready" votre serveur est prêt à recevoir des requêtes.

Installation du plugin

Veuillez récupérer les fichiers sur le lien suivant : .
Veuillez extraire le zip et le mettre dossier où vous le souhaitez.

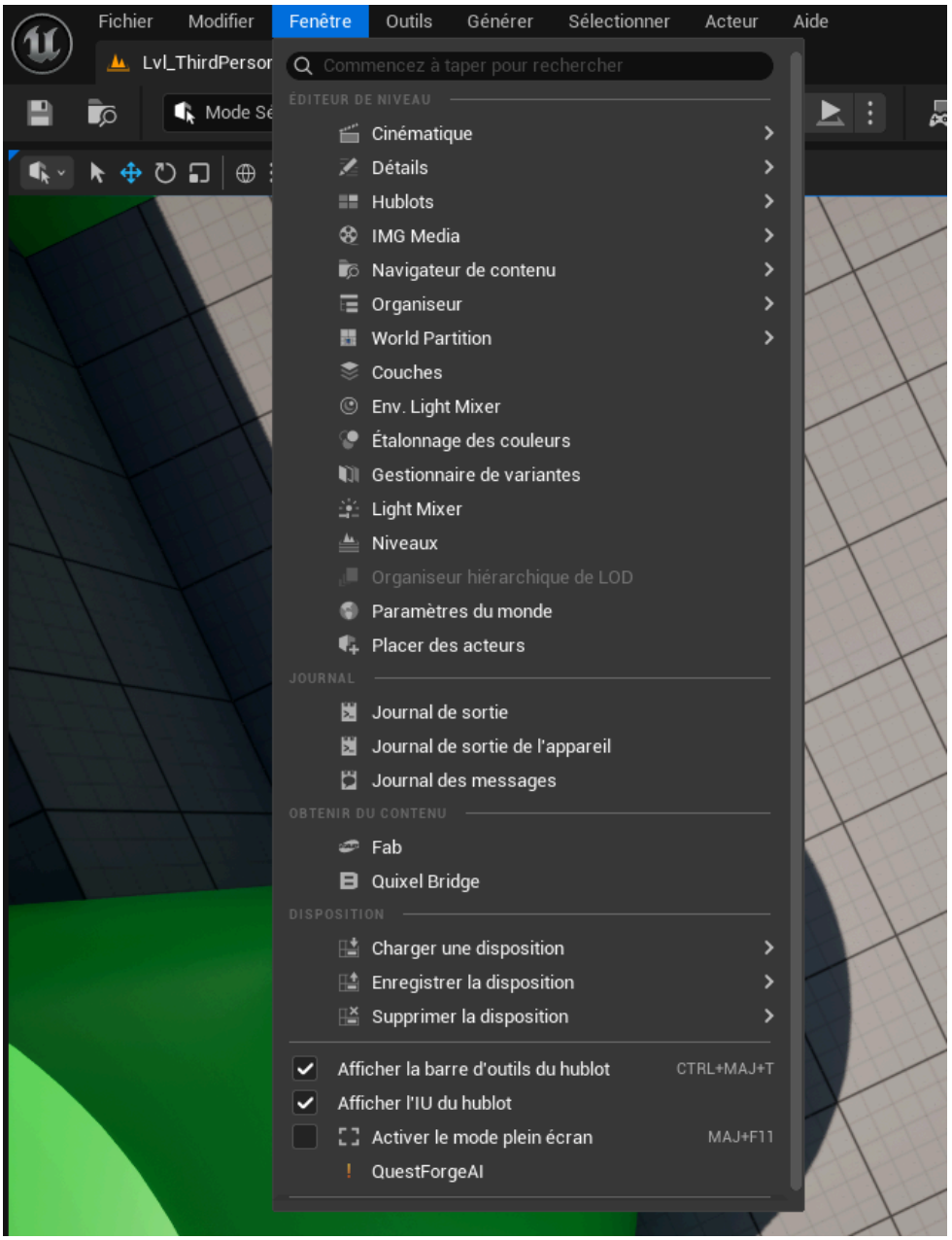
Ensuite ouvrez l'invite de commande, rendez vous à l'endroit où vous avez mis le dossier contenant le plugin.
Selon votre système d'exploitation, exécutez la commande : ./Install_Plugin_Windows.ps1 (Windows) ou ./Install_Plugin_Mac.sh (Mac)

Une fois cela réalisé vous aurez le plugin de disponible sur Unreal Engine.

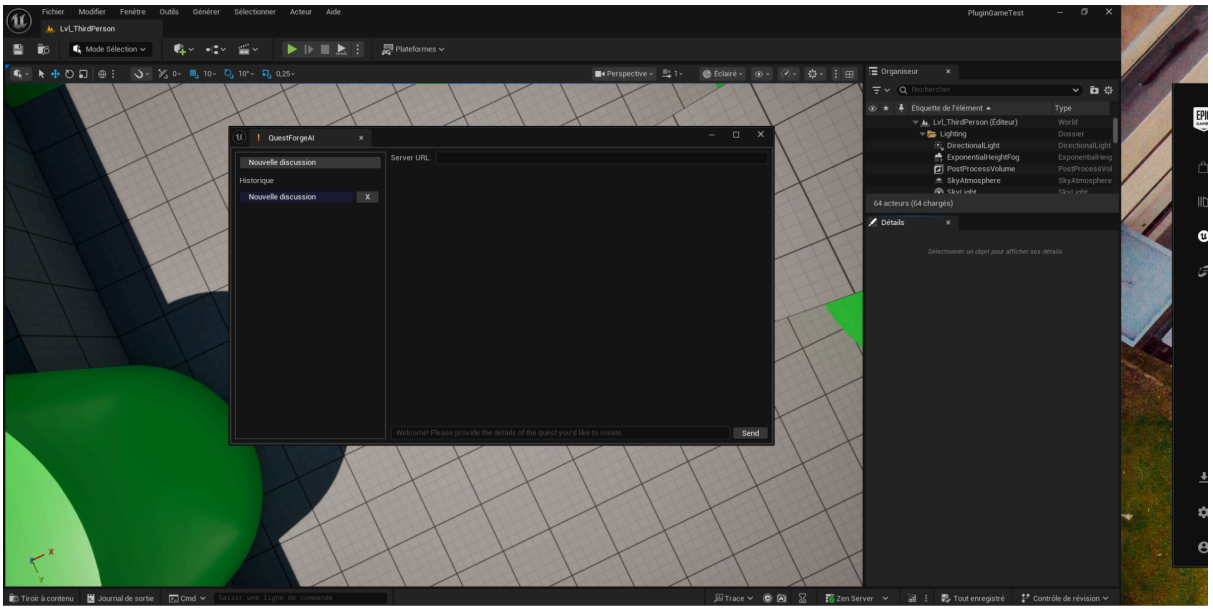
Guide d’utilisation du plugin

Afin que les étapes suivantes fonctionnent bien, il est nécessaire que le serveur avec l’IA soit démarré et contactable sur le réseau, dans le cas contraire veuillez le démarrer ou l’installer puis le démarrer en vous référant à Installation du serveur.

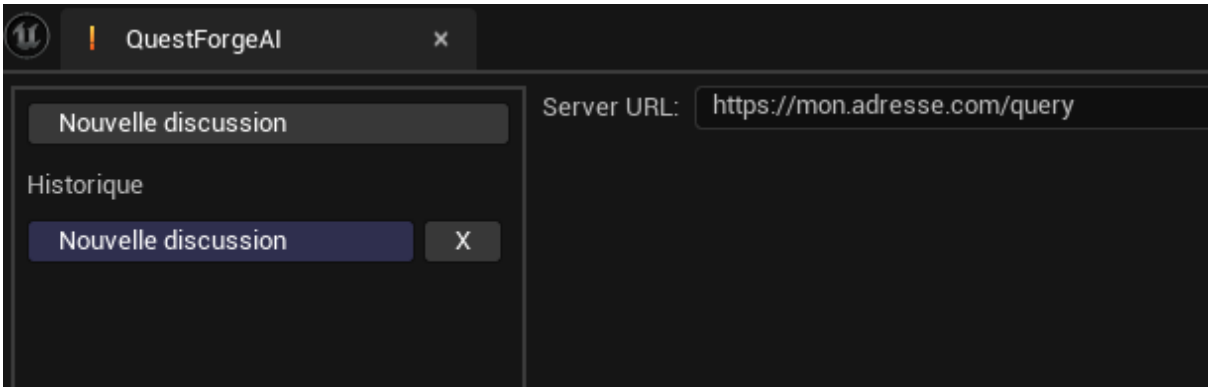
Dans Unreal Engine, veuillez cliquer sur “Fenêtre” ou “Window” puis sur *QuestForgeAI*, afin d’ouvrir la fenêtre du plugin.



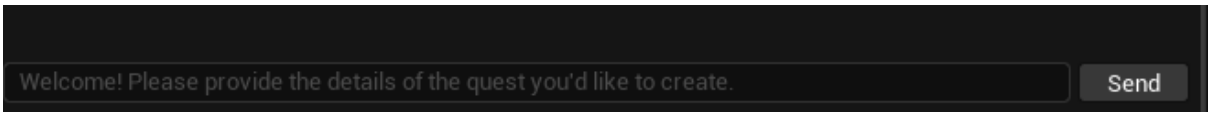
Vous aurez ensuite une fenêtre qui s’ouvre comme suit que vous pourrez docker où vous le souhaitez dans Unreal.



Veuillez entrer l’adresse du serveur à contacter (celui sur lequel fonctionne l’IA) suivi de “/query”.
Par exemple, si mon adresse est la suivante : “https://mon.adresse.con”,
il faut alors que je rentre : “https://mon.adresse.com/query”.



Vous pouvez écrire la quête que vous souhaitez générer dans l’encart “Welcome! Please provide the details of the quest you’D like to create.” et communiquer avec l’IA en écrivant vos messages dans ce même encart.
Appuyez sur “Entrée” pour confirmer l'envoi du message.



Une fois la réponse reçue de la part de l’IA vous pourrez de nouveau écrire un message.

Lorsqu’un premier message a été envoyé, il est possible de créer une nouvelle conversation en cliquant sur le bouton “Nouvelle discussion”, les anciennes discussions peuvent être supprimées en client sur le bouton “X” à côté de celle ci.

Fonctionnement

Lors du démarrage du serveur, les documents (ensemble de quêtes provenant de jeux RPG et action/aventure variés) du dossier DATA vont être chargés.

Une fois le contenu extrait, il est découpé en blocs de K “chunks” de tailles fixées (par défaut 20 chunks d’une taille de 5000). Ce découpage permet d’améliorer la recherche et d’optimiser la pertinence lors des requêtes.

Chaque chunk est ensuite converti en vecteur numérique grâce à un modèle d’embedding (de base bge-m3).

Tous les vecteurs générés sont ensuite insérés dans un index FAISS (FAISS est une bibliothèque optimisée, conçue par Meta).

Scénario d’envoi d’une requête

L’utilisateur appuie sur “Entrée” dans Unreal Engine Unreal pour envoyer un message au serveur :

1. Engine va envoyer une requête HTTP au serveur de calcul que vous avez entré.
2. Le texte de la question est converti en vecteur via le même modèle d’embedding.
3. Ce vecteur est comparé à tous les vecteurs de l’index et le système récupère les chunks les plus pertinents.
4. Les chunks récupérés sont ensuite envoyés au modèle LLM (par défaut llama3.1:8b)
5. Le LLM utilise la question de l’utilisateur et les informations retrouvées dans les documents afin de générer la réponse