Publication Scientifique Mission R&D

ESII : Agent virtuel d'accueil

Élèves :
CRUVELLIER Chloé
LEGAREZ Lucie
TERRASSON Ludovic

Tuteurs :
LERGENMULLER Philippe (entreprise)
TCHECHMEDJIEV Andon (école)

Table des matières

I- Résumé (Abstract)

Dans un contexte de modernisation des services municipaux, l'intégration d'agents virtuels d'accueil représente une réponse innovante aux défis de gestion des flux de visiteurs dans les mairies. Cette étude se concentre sur l'évaluation de la technologie de chatbot Llama cpp pour cette application spécifique.

L'étude compare Llama cpp à d'autres technologies de chatbot émergentes telles que ChatGPT, Copilot, Rasa, IBM Watson Assistant et Microsoft Azure Bot Service. Chaque technologie est évaluée en termes de pertinence pour une utilisation sur une borne de mairie et d'interaction avec les visiteurs.

Après une analyse approfondie, Llama cpp est choisi pour le déploiement du chatbot en raison de sa capacité à fonctionner localement sans connexion Internet constante. Nous avons exploré différentes techniques d'ingénierie du prompt, de sélection de modèle et d'optimisation pour répondre aux attentes en termes d'efficacité et de personnalisation de l'accueil des visiteurs. Cependant, des défis potentiels liés à la compréhension des demandes complexes des utilisateurs sont identifiés.

Ainsi, l'intégration d'agents virtuels d'accueil dans les mairies présente un potentiel prometteur pour améliorer l'expérience des visiteurs. L'étude met en évidence l'importance de choisir une technologie de chatbot adaptée aux besoins spécifiques de l'application et des contraintes de déploiement

II- Introduction :

Les établissements accueillant du public sont souvent confrontés à des défis dans la gestion de l'affluence de leurs clients. Les files d'attente peuvent rapidement s'allonger, ce qui rend leur gestion difficile. Face à ce problème, la société ESII développe de nombreuses solutions afin de moderniser les systèmes d'accueil des établissements et d'améliorer la rapidité de traitement des demandes des clients. Ainsi est né le sujet de ce projet. Nous avons décidé de nous concentrer principalement sur le cas des mairies car ce sont des établissements publics fréquemment sujets à des problèmes liés à l'attente. Lancé en 2018, le projet a mobilisé plusieurs groupes de personnes et continue d'évoluer. Il repose sur l'utilisation des bornes de redirection automatiques existantes, auxquelles une fonctionnalité de reconnaissance vocale est ajoutée. Ainsi, les clients peuvent être dirigés efficacement sans perdre de temps à chercher le bon ticket ou la bonne option, ce qui améliore la fluidité de l'accueil.

Lors d'études précédentes, le sujet a pu être abordé par des modèles d'apprentissage automatique tels que les arbres de décisions, forêts aléatoires et modèle de régression logistique. Nous avons donc décider de nous tourner vers des modèles génératifs de type LLM (Large Language Model). Ces modèles utilisent ainsi des techniques de NLP (Natural Language Processing) afin de comprendre et générer du texte en langage naturel. Ainsi, puisque le but de ce projet est de créer un outil permettant de définir le motif de la venue d'un client à partir de sa demande orale, ce processus se décomposera en deux parties : une partie

reconnaissance vocale et transcription écrite, et une deuxième partie d'analyse de la demande et d'association à un motif. En ce qui concerne la première partie, l'entreprise E.S.I.I. possède déjà un outil de reconnaissance vocale. Ainsi, nous avons décidé de nous concentrer sur la deuxième partie pour notre projet, à savoir l'identification d'un motif. Afin de sélectionner la meilleure méthode, il était nécessaire de comparer différents modèles de chatbot afin d'en sélectionner le plus optimal pour son cas d'application. Nous privilégions ainsi un modèle rapide et qui puisse fonctionner en local.

L'organisation de la publication est la suivante : après cette introduction, nous examinerons dans la section suivante l'état de l'art des technologies de chatbot, afin de situer notre recherche par rapport aux travaux existants dans ce domaine. Ensuite, nous détaillerons les matériaux et méthodes utilisés pour mener notre analyse comparative. Les résultats obtenus seront discutés en profondeur, mettant en lumière les avantages et les limites de chaque technologie évaluée. Enfin, nous conclurons en proposant des perspectives et des recommandations pratiques pour l'intégration réussie d'un agent virtuel d'accueil dans les mairies.

III- État de l'art :

L'intégration d'un agent virtuel d'accueil dans le contexte des mairies pour la gestion des flux de visiteurs représente un domaine de recherche émergent, bénéficiant des avancées rapides dans le domaine des technologies de chatbot. Dans cette section, nous positionnerons notre étude par rapport à la revue de littérature pertinente, en examinant en détail les principales technologies de chatbot telles que Llama cpp, ChatGPT, Copilot, ainsi que cinq autres technologies émergentes, et leurs avantages, inconvénients et cas d'utilisation, en tenant compte de leur déploiement sur une borne de mairie et de leur interaction avec les visiteurs. [1]

1) Llama cpp :

Llama cpp offre une intégration optimale pour une borne de mairie, en raison de sa capacité à fonctionner localement sans nécessiter une connexion Internet constante. Cela garantit une réactivité accrue et une disponibilité sans faille, même dans les zones où la connectivité en ligne peut être limitée. De plus, Llama cpp est capable d'engager des conversations avec les visiteurs de manière fluide, en leur offrant une assistance personnalisée dès leur arrivée. Cependant, sa capacité à comprendre les demandes complexes des utilisateurs peut être limitée, ce qui peut entraîner des interactions moins satisfaisantes dans certains cas.

2) ChatGPT :

ChatGPT excelle dans la compréhension du langage naturel et peut fournir des réponses pertinentes et contextuellement appropriées aux visiteurs d'une mairie. Son déploiement sur une borne de mairie peut toutefois poser des défis, en raison de sa dépendance à une connexion Internet constante. Cela peut entraîner des temps de latence plus longs et une disponibilité réduite dans les zones où la connectivité en ligne est limitée. Cependant, une fois déployé, ChatGPT peut offrir des conversations engageantes et utiles aux visiteurs.

3) *Copilot :*

Copilot offre des fonctionnalités avancées de compréhension du langage naturel et de génération de réponses, ce qui en fait une option attrayante pour l'intégration dans une mairie. Son déploiement sur une borne de mairie peut toutefois poser des défis similaires à ceux de ChatGPT, en raison de sa dépendance à une connexion Internet constante. De plus, Copilot peut nécessiter des ressources informatiques plus importantes pour son fonctionnement, ce qui peut être un facteur limitant dans certains environnements.

4) *Rasa :*

Rasa offre une flexibilité et une personnalisation accrues dans le développement d'agents conversationnels, ce qui en fait une option attrayante pour les mairies cherchant à adapter leur agent d'accueil à leurs besoins spécifiques. Son déploiement sur une borne de mairie peut être réalisé avec succès, bien qu'il puisse nécessiter des compétences techniques avancées pour sa mise en œuvre et sa maintenance. Une fois déployé, Rasa peut offrir des conversations personnalisées et utiles aux visiteurs, en répondant à leurs besoins de manière efficace. [2]

5) *IBM Watson Assistant :*

IBM Watson Assistant offre des fonctionnalités avancées de compréhension du langage naturel et de gestion de conversation, ce qui en fait une option solide pour l'intégration dans une mairie. Son déploiement sur une borne de mairie peut être réalisé avec succès, bien que son coût élevé puisse être un facteur limitant pour certaines organisations. Une fois déployé, IBM Watson Assistant peut offrir des conversations engageantes et utiles aux visiteurs, en répondant à leurs demandes avec précision.

6) *Microsoft Azure Bot Service :*

Microsoft Azure Bot Service offre des outils puissants pour la création, le déploiement et la gestion des agents conversationnels, ce qui en fait une option attrayante pour les mairies cherchant à intégrer un agent d'accueil sur une borne. Son déploiement peut être réalisé avec succès, bien que sa dépendance à une connexion Internet constante puisse poser des défis dans certains environnements. Une fois déployé, Microsoft Azure Bot Service peut offrir des conversations interactives et utiles aux visiteurs, en répondant à leurs besoins de manière efficace.

7) *Comparaison :*

En comparant ces différentes technologies de chatbot, nous pouvons voir que chaque approche présente des avantages et des inconvénients distincts en termes de déploiement sur une borne de mairie et d'interaction avec les visiteurs. Llama cpp se distingue par sa capacité à fonctionner localement sans nécessiter une connexion Internet constante, offrant ainsi une réactivité accrue et une disponibilité sans faille. ChatGPT, Copilot, Rasa, IBM Watson Assistant et Microsoft Azure Bot Service offrent des fonctionnalités avancées de compréhension du langage naturel et de gestion de conversation, mais leur déploiement peut poser des défis en

termes de connectivité en ligne et de ressources informatiques requises. Le choix de la technologie appropriée dépendra donc des besoins spécifiques de l'application et des contraintes de déploiement de chaque mairie.

IV- Matériels et Méthodes :

Pour mener à bien cette étude sur l'intégration d'un agent virtuel d'accueil dans une mairie pour la gestion des flux de visiteurs et de leur accueil, nous avons suivi une méthodologie rigoureuse comprenant plusieurs étapes clés. L'objectif principal était d'associer un motif, choisi parmi une liste prédéfinie, à chaque demande client. L'utilisateur exprime sa demande à la borne et celle-ci lui renvoie le motif correspondant sous la forme d'un ticket avec le nom du motif imprimé.

1) *Design de l'étude :*

Dans la conception de notre étude, nous avons opté pour l'utilisation de l'intelligence artificielle générative à travers la technologie de llama.cpp. Ce système open-source, écrit en C/C++, nous permet de mettre en œuvre des modèles pour déployer un chatbot prêt à l'emploi. [3]

2) *Population ou échantillons étudiés :*

Concernant la population étudiée, notre chatbot sera déployé dans plusieurs administrations et entreprises. Pour notre mission, nous nous consacrons à une mairie en particulier, la mairie de Roanne. Une mission R&D précédente a permis de recueillir les demandes les plus fréquentes rencontrées dans une mairie. Ces demandes ont été associées à des niveaux de difficulté (0, 1, 1.5, 2, 2.5) en fonction de la complexité de la formulation de la demande par l'utilisateur. De plus, chaque demande a été liée à un motif parmi une liste de 25 motifs différents. Au total, nous disposons d'environ 600 demandes d'utilisateurs pour notre étude.
Nous devrons par la suite tester les performances de notre chatbot sur cet échantillon.

3) *Procédures expérimentales :*

Pour déployer efficacement notre chatbot, nous avons effectué une sélection méticuleuse d'un modèle adapté et élaboré un prompt spécifique. Le modèle joue un rôle crucial car il est chargé de générer le texte du chatbot, tandis que le prompt fournit les instructions nécessaires à son fonctionnement optimal.

À cet effet, nous avons téléchargé plusieurs modèles depuis la plateforme Hugging Face, en nous concentrant sur les modèles français au format gguf. Nous avons pris connaissance des différentes variantes de modèles disponibles, notamment les modèles "instruct" et "chat" [4]. Bien que la création d'un modèle personnalisé ait été une option envisageable, nous avons décidé de privilégier l'utilisation de modèles préconçus, vu la complexité et la durée de ce processus.

Parallèlement, nous avons exploré différents prompts pour guider le chatbot dans ses réponses. Le prompt a été élaboré et enregistré dans un fichier texte distinct, comprenant les 25 motifs spécifiques ainsi que les directives nécessaires au bon fonctionnement du chatbot. Notre objectif principal était de configurer le chatbot pour qu'il associe exclusivement les motifs aux demandes des utilisateurs, afin d'optimiser son efficacité et sa pertinence dans le contexte d'accueil virtuel au sein d'une mairie.

4) *Méthodes d'analyse utilisées :*

Pour évaluer les performances de notre chatbot, nous avons développé un script Python dédié. Ce script ne se contente pas seulement de mesurer la précision globale du chatbot, mais il analyse également la précision selon différents critères, tels que le niveau de difficulté des demandes (évalué sur une échelle de 0 à 2.5) et la précision par motif spécifique.

En analysant la précision du chatbot par niveau de difficulté, nous pouvons obtenir des informations précieuses sur sa capacité à traiter efficacement les demandes de différentes complexités. De plus, en évaluant la précision par motif, nous pouvons identifier les domaines où le chatbot excelle et ceux où il pourrait nécessiter des améliorations.

En combinant ces différentes analyses, notre objectif est d'obtenir un aperçu complet des performances du chatbot et d'identifier les domaines où des ajustements sont nécessaires pour garantir une expérience utilisateur optimale.

V- Évaluation et analyse

Pour le déploiement de notre chatbot, nous avons opté pour l'utilisation d'un modèle de type "instruct". Ce choix s'est imposé car ce type de modèle offrait un contrôle total à l'utilisateur sur la réponse générée par le chatbot, contrairement aux modèles de type "chat" qui limitaient la conversation sans possibilité de diriger les réponses. En outre, nous avons constaté que les modèles de taille plus réduite, bien que présentant des performances moins élevées, offraient des temps de déploiement et de réponse plus courts. Étant donné que notre borne est assimilable à un ordinateur, le déploiement de modèles très volumineux pourrait entraîner des retards considérables. Il est donc impératif de trouver un équilibre entre le temps de réponse et la précision du modèle, en tenant compte des contraintes de performance spécifiques à notre environnement de déploiement.

Après une analyse approfondie, nous avons choisi un modèle de taille modérée de 4 Go. Bien que ce modèle puisse être déployé en 2 minutes et 20 secondes sur un ordinateur, et qu'il offrait un temps de réponse raisonnable d'environ 5 secondes pour chaque demande, nous avons constaté que sa précision était seulement de 73%. Ce résultat, bien que respectable, ne répondait pas pleinement à nos attentes en termes de qualité de service. Il ne serait pas judicieux de déployer un chatbot sur une borne qui possède des performances aussi médiocres.

En examinant les performances du chatbot, nous avons remarqué que sa capacité à distinguer les phrases les plus simples des plus complexes n'était pas le principal problème. En revanche, nous avons identifié des difficultés dans la reconnaissance de certains motifs spécifiques. Par exemple, la distinction entre "Enregistrement de PACS" et "PACS (dépôt de dossier, modification ou dissolution)" était souvent floue. De plus, certains motifs, tels que "demande d'attestations diverses", présentaient des limites mal définies pour le chatbot, ce qui affectait sa capacité à fournir des réponses précises.

Pour améliorer la performance du chatbot, il faudrait peaufiner le prompt en fournissant des explications détaillées sur les motifs les moins bien distingués et en clarifiant leurs limites. Cela permettra de mieux guider le chatbot dans la sélection et l'attribution des motifs, améliorant ainsi la qualité globale des réponses fournies. Il est aussi nécessaire de trouver un compromis entre taille du modèle et performances.

VI- Conclusion et perspectives

Comme précisé dans l'introduction, notre objectif était de développé un modèle en locale et suffisamment rapide et précis pour détecter le motif de venue d'un client à partie de sa demande écrite. En effet, on considère la partie transcription de la demande orale comme un sujet à part. Pour ce faire nous avons donc étudié la solution de llama.cpp, qui est un outil local.
Cette étude a visé le développement d'un modèle local, rapide et précis pour déterminer le motif de visite des clients dans les mairies à partir de leurs demandes écrites, en utilisant l'outil llama.cpp pour sa capacité à fonctionner sans connexion internet constante. Bien que notre modèle ait montré une efficacité opérationnelle prometteuse, des défis demeurent, notamment dans la reconnaissance précise de motifs spécifiques et la gestion de demandes complexes. À l'avenir, il serait envisageable d'améliorer le prompt et le modèle pour une distinction plus claire des motifs, d'étendre les tests à d'autres mairies pour évaluer la performance dans divers contextes. Il serait également possible d'associer ces modèles aux fonctionnalités de reconnaissance vocale, déjà développées par l'entreprise, afin de simplifier les interactions. De plus, pour aller plus loin et avec les ressources suffisantes, il serait envisageable de créer son propre modèle afin d'en augmenter l'efficacité et la personnalisation à l'accueil.

VII- Remerciements

Nous exprimons notre sincère gratitude envers nos tuteurs, Philippe LERGENMULLER et TCHECHMEDJIEV Andon, pour leur accompagnement continu tout au long de cette mission. Leur expertise, leur soutien et leurs conseils ont été d'une importance capitale pour notre compréhension du sujet, nos recherches documentaires et le suivi de nos résultats. Leur engagement a grandement contribué à la réussite de ce projet.

Nous exprimons notre reconnaissance envers les documentalistes pour leur soutien et leur disponibilité tout au long de notre projet. Leur implication s'est manifestée par leurs conseils éclairés sur le fonctionnement des sources et leur aide dans notre recherche d'informations

Enfin, nous tenons à exprimer notre reconnaissance envers la société ESII pour la confiance qu'elle nous a accordée et pour avoir proposé un sujet de mission aussi captivant. Cette opportunité nous a permis d'approfondir nos compétences et de découvrir de nouveaux aspects passionnants dans le domaine de l'intégration d'un agent virtuel d'accueil.

VIII- Références bibliographiques

[1] Aixploria. (2024, 18 février). Liste des Meilleures IA Gratuites & Top 10 par Catégorie | Aixploria. https://www.aixploria.com/
[2] Conversational AI Platform | Superior customer experiences start here. (2024, 18 avril). Rasa. https://rasa.com/
[3] Admin. (2023, 30 décembre). What is llama cpp ? - AI Verse Info. AI Verse Info. https://aiverseinfo.com/what-is-llama-cpp/
[4] Computing For All. (2024, 8 mars). Chat LLM vs Instruct LLM — Differences and Similarities [Vidéo]. YouTube. https://www.youtube.com/watch?v=0V9OKfpNq64