Bonjour à tous,  
  
Je suis ravie de vous présenter ma soutenance de fin d’étude dont le thème est l’exploration de l’utilisation de l’IA générative dans le domaine bancaire.  
  
Now, I m going to switch in english to give a short introduction about this presentation.

Nowdays, the advancement of technology has significantly transformed the banking industry. Especially during the recent pandemic, there has been an increasing demand for online banking services due to their convenience, accessibility, and efficiency. To meet the needs of customers, banks are required to provide instant responses to queries and concerns, which has led to the adoption of new strategies for efficient customer service.

One of these innovative strategies is the use of AI-powered Chatbots and Assistants. By leveraging AI and chatbot technologies, banks can provide immediate responses to customer inquiries. These chatbots effectively handle common queries, offer personalized recommendations, and ensure 24/7 assistance, guaranteeing timely customer support.

Now, let's delve into the question: " how generative AI can be used to generate consistent and relevant conversations in banking?"

Qu’est ce que l’inteeligence artificielle et plus particulièrement l’IA générative ?  
  
On dit qu’un objet est artificiel lorsque ce dernier a été produit et crée par l’homme. En effet, il est le résultat des activités humaines et n’est donc pas naturel.

Ainsi, l’intelligence artificielle (IA), peut se définir comme une « copie » de l’intelligence humaine.

L’IA générative s’inspire du concept de la « générativité présent dans la nature ». Ce concept a pour but de donner la capacité à un système de créer un contenu nouveau, significatif, cohérent et pertinent, souvent en apprenant à partir d’un jeu de données existantes. L’intelligence artificielle générative vise à de la réplication et à générer de nouveaux résultats.

l’intelligence artificielle générative est utilisée dans de nombreux domaines tels que la création d’art, la production de musique, le design, la traduction automatique, la génération de textes, jeux vidéo, réalité virtuelle, modélisation 3D, analyse de données etc.

Le secteur bancaire est un secteur des affaires et de la finance qui existe depuis des milliers d’années. En effet, les prémices du bancaire existe depuis la Grèce antique et c’est notamment le Moyen Orient et l’Asie qui ont joué un rôle important dans son développement. Par exemple le mot « chèque », dérive d’un mot arabe signifiant « promesse de paiement à la livraison ».

Le but principal d’une banque est avant tout de fournir aux clients un lieu où ils peuvent accumuler leur argent via un compte courant auquel il est facile d’accéder pour les retraits et les dépôts.

Le premier canal de communication populaire dans le domaine bancaire est les interaction orales. Cela comprend divers moyens tels que les

L’un des principaux avantages de l’interaction visuelle avec les clients est qu’elle renforce la confiance et les relations que l’on peut avoir avec ces clients. Cela est notamment due aux indices corporels et aux expressions faciales.. Cet aspect va donc favoriser l’empathie et aider à mieux comprendre le souhait du client.

L’un des avantages du courrier électronique est qu’il est de nature asynchrone ce qui permet aux clients et aux conseillers de répondre quand ils le souhaitent.

Aujourd’hui avec l’arrivé des chatbots, nous avons une nouvelle façon de répondre aux besoins du client. En effet, les progrès de la technologie de l’intelligence artificielle en particulier dans la génération du langage naturel ont considérablement amélioré les capacités des chatbots.

L’un de ces principaux avantages du chatbot, c’est qu’il est disponible vingt-quatre heures sur vingt-quatre et sept jours sur sept et de plus, il fournit une réponse instantanée. Il a la capacité de gérer plusieurs conversations en même temps et garantissent ainsi une assistance aux clients. Cette disponibilité contribue à améliorer la satisfaction des clients et l’expérience globale des utilisateurs.

MOD7LES DE LANGAGES

L'apprentissage supervisé est une méthode d'intelligence artificielle où un modèle est entraîné à partir de données étiquetées, afin de prédire des résultats futurs en fonction de nouvelles données similaires

C’est ce type d’apprentissage qui est utilisé pour la classification ou de la régression.

L'apprentissage supervisé peut également être utilisé pour la génération de texte. Pour cela, on va entraîner un modèle sur un jeu de données labélisées, où chaque exemple est associé à une sortie attendue.

Cette approche peut produire des résultats précis, mais elle nécessite des données d'entraînement étiquetées, ce qui peut être coûteux et chronophage.

Ainsi, dans l’apprentissage non supervisé, l’algorithme acquiert des connaissances uniquement à partir des données d’entrée sans avoir connaissance de leur signification. La phase de prédiction consiste donc à fournir une nouvelle entrée et à la classer dans un des groupes (regroupement et association).

Les architectures de réseaux de neurones sont largement utilisées pour la génération de texte en raison de leur capacité à apprendre des modèles complexes à partir de données textuelles.

Un réseau de neurones est un dispositif informatique qui s’inspire du fonctionnement du cerveau humain pour apprendre. Le but de cette technologie est « de simuler l’activité du cerveau humain c’est-à-dire de reconnaître des images et de transmettre des informations entre les différentes couches de connexions » ([6], s.d.).

Le réseau de neurones informatique est lui composé de plusieurs couches : une couche d’entrée, une ou plusieurs couches cachées et une couche de sortie. Les différentes données sont traitées de façon progressive pour réaliser la prédiction dans les couches cachées.

Le résultat prédit est transmis à la couche suivante du neurone grâce à une connexion synaptique. Cette connexion (synapse) est associée à un poids qui définit « l’influence des données sur le résultat ».

TRANSFORMERS  
  
Les Transformers ont été introduits en 2017 par Vaswani et al

L'élément central d'un modèle Transformer est le mécanisme d'attention comme nous pouvons sur la figure ci-dessous, qui permet au modèle de se concentrer sur différentes parties de la séquence d'entrée lorsqu'il fait des prédictions. Ce qui signifie que le poids de chaque mot ici dépend des autres mots de la même phrase.

Voici un exemple qui illustre ce concept d’auto-attention :

Dans la phrase, *« L’athlète n’a pas fini son marathon car il était fatigué. »,* le *« il »* fait référence à l’athlète, le processus d’auto-attention aura donc pour but de détecter le lien entre *« l’athlète »* et *« il ».* Ainsi, son rôle est de faire garder l’interdépendance de chaque mot dans la représentation des séquences.

En plus de l'auto-attention, les transformers comprennent des réseaux neuronaux de type "feed-forward" comme deuxième composant majeur et des couches entièrement connectées (fully connected neural network) pour faire la prédiction.

BERT emploie une architecture de réseau de neurones basée sur des transformers

La principale innovation de BERT réside dans sa capacité à apprendre des représentations bidirectionnelles des mots, ce qui signifie qu’il peut prendre en compte à la fois les mots précédents et les mots suivants dans une phrases pendant l’apprentissage

Le modèle BERT est entraîné à l’aide de deux concepts principaux de pré-traitement :

* La modélisation du langage masqué (MLM)
* La prédiction de la phrase suivante (NSP)

La modélisation du langage masqué (MLM) consiste à masquer aléatoirement certains mots d’une phrase pour apprendre à retrouver ces mots masqués en se basant sur le contexte fourni par les autres mots.

En effet, un mot peut avoir plusieurs sens dans toutes les langues. Dans la langue française, si on prend le mot « orange », celui-ci peut avoir trois sens : représentation du fruit, représentation de la couleur et représentation de l’entreprise.

Concernant maintenant la prédiction de la phrase suivante (NSP), celle-ci a pour but de détecter si deux phrases peuvent être placés l’une après l’autre

Vicuna atteint plus de 90% de la qualité d'OpenAI ChatGPT

**Première approche** : utilisation des modèles de langage pré-entraîné et fourniture d’un jeu de données labélisées à la main pour adapter ce modèle au domaine bancaire. Pour pouvoir évaluer la performance du modèle, on pourra effectuer une comparaison avec le Chabot de la Société Générale.

LangChain est un outil de développement d'applications alimentées par des modèles de langage. Ce Framework permet aux applications de :

* Connecter un modèle de langage à d'autres sources de données
* Utiliser un modèle de langage pour interagir avec son environnement

Le framework LangChain fournit deux principaux outils de valeur :

Dans notre cas d’étude, nous allons utiliser les chaînes pour interagir avec Vicuna car c’est un modèle open source .

Dans la section chaîne, nous allons effectuer du Question Answering sur la base de connaissances du Chabot de la Société Générale en liant ce document avec le modèle Vicuna.

En conclusion, on peut noter certaines différences dans leurs réponses et fonctionnalités. LangChain Vicuna fournit des réponses plus directes et spécifiques, donnant aux utilisateurs les informations recherchées sans nécessiter des recherches supplémentaires. Par exemple, LangChain Vicuna fournit une liste complète des pays de la zone SEPA, ce qui est plus précis et pratique pour les utilisateurs qui ont besoin de connaître ces informations pour leurs transactions internationales.

Cependant, le chatbot SG offre des instructions détaillées pour certains processus, ce qui peut être bénéfique pour les utilisateurs qui ont besoin d'un accompagnement plus complet

Effectuer de l’analyse conversationnelle pour pouvoir construire un jeu de données. Dans cette approche, nous allons récupérer des conversations et effectuer de l’apprentissage non supervisé pour connaître les principaux topics des conversations. Avec les résultats obtenus de l’apprentissage non supervisé, on pourra construire un jeu de données pour effectuer de l’apprentissage supervisé et par la suite entraîner un réseau de neurones ou le fournir à un modèle de langage pré-entraîné (LLM).

Elle consiste à tracer le graphique de la variance en fonction du nombre de cluster et d’observer le point où la courbe consiste à fléchir en forme de coude.

BERTopic est une technique puissante de modélisation thématique non supervisée qui exploite les performances du modèle de langage BERT. Elle capture les relations sémantiques entre les mots et les documents, fournissant une représentation plus précise des sujets sous-jacents

Pour conclure, les cinq topics de conversations sont : Afficher, télécharger RIB ; Code sécurité, téléphone pass ; Carte bancaire client ; Code secret espace client et Virement.

Nous allons maintenant comparer ces résultats avec ce que nous obtenons avec BERTopic. Dans notre cas d’étude, nous avons obtenu 42 topics différents. Pour des raisons de visualisation, voici le résultat des 8 premiers clusters :

BERTopic est capable de capturer les relations sémantiques entre les mots dans une phrase, ce qui est un véritable avantage comparé à K-Means qui lui se base seulement sur une approche géométrique.

Un autre avantage de BERTopic c’est qu’il est capable de déterminer le nombre de cluster par lui-même et de capturer les seuls mots importants.

Construire par la suite un jeu de données pour l’entraînement d’un réseau de neurone ou d’un modèle de langage pré-entraînés (fine-tuning). L’entraînement d’un tel modèle nécessite des grandes infrastructures très coûteuses. C’est pour cela, qu’il est plus avantageux de partir sur l’approche avec LangChain et Vicuna.

La figure 1 montre un même ordre de grandeur de l’impact carbone des requêtes, cette valeur est autour de KgCO2eq. La réponse à une requête prend en moyenne une minute. Nous remarquons une proportionnalité entre le temps de calcul et les émissions carbones.

La figure 2 démontre le cumul des émissions carbones des requêtes. Nous remarquons que l’impact carbone est en augmentation exponentielle au fur et à mesure que le nombre de requêtes augmente. Les émissions passent de KgCO2eq à KgCO2eq en 10 requêtes.

il faut prendre nos précautions lors de l’utilisation de ces modèles de LLM. Il faut penser à réduire la complexité des requêtes voire éviter celles qui ne sont pas utiles.

Les engagements RSE sont à prendre en considération au sein de tous les projets IA et surtout ceux qui traitent les LLM.

La problématique de la génération de réponses cohérentes et instantanées aux clients dans le domaine bancaire à l’aide de l’intelligence artificielle générative est cruciale pour offrir une expérience utilisateur fluide et satisfaisante. En examinant les deux approches : approche 1 LangChain et Vicuna et approche 2 analyse conversationnelle, nous pouvons tirer quelques conclusions.

L’approche avec LangChain et le modèle de langage pré-entraînés ont montré des résultats cohérents aux questions posées

L’approche 2 nécessite une infrastrucre plus coutuese pour effectuer l’entrâinement

Cependant, on ne peut pas exclure définitivement l’approche 2 qui se repose sur l’analyse conversationnelle car on pourrait se servir de cette méthode pour construire une base de connaissances plus robuste que celle utilisée pour l’approche 1 LangChain et Vicuna. Cette nouvelle base de connaissances pourra donc être utilisée avec LangChain et Vicuna ou bien même un autre Larges Languages Model plus performant comme GPT-4, pour fournir probablement de meilleurs résultats. En effet, un modèle comme GPT-4 pourra offrir aux clients des réponses plus précises et cohérentes avec leur besoin.

Générer des réponses cohérentes et instantanées aux clients dans le domaine bancaire à l’aide de l’IA générative demande de se poser beaucoup de questions, notamment sur l’impact environnemental, économique et sur l’infrastructure