



REBULIQUE TUNISIENNE
Ministère de l'enseignement supérieur et de la recherche scientifique
Université de Carthage



École supérieure de la statistique et de l'analyse de l'information

*Rapport de Projet de Fin d'Études présenté pour l'obtention du
Diplôme National d'Ingénieur en Statistique et Analyse de l'Information*



Réalisé par :

Ala Eddine Hlaili

Développement d'un chatbot pour la rénovation énergétique des
logements

Soutenu le 21/06/2023 devant le Jury composé de :

- Mme Amira Gasmi (Présidente)
- Mme Tasnim Hamdeni (Rapporteure)
- M. Farouk Mhamdi (Encadrant universitaire)
- Mme Fatma Zaoui (Encadrant entreprise)

Année Universitaire : 2023/2024

Remerciements

Je tiens à exprimer ma profonde gratitude envers ma famille et mes parents pour leur soutien inconditionnel et leur amour tout au long de ce projet. Leur encouragement constant a été une source inestimable de motivation.

Un remerciement tout particulier à mon encadrant universitaire, M. Mhamdi Farouk, et à mon encadrant d'entreprise, Mme Zaoui Fatma, pour leur guidance, leurs précieux conseils et leur patience tout au long de cette aventure académique et professionnelle. Leur expertise et leur dévouement ont été essentiels à la réussite de ce projet de fin d'études.

Abstract

In the current context where the fight against climate change has become a priority, the energy renovation of homes represents an essential approach to reducing carbon emissions. Our chatbot project aims to turn this challenge into an accessible and manageable opportunity for owners. This virtual assistant, developed using the most advanced artificial intelligence technologies, is designed to guide users step by step through the different stages of energy renovation. By providing personalized information and practical advice, the chatbot helps simplify the complex renovation process, making energy standards less intimidating and more understandable.

This chatbot is powered by cutting-edge linguistic models that enable it to understand and answer specific user questions with remarkable accuracy. Thanks to an intuitive interface, users can easily access information on best practices, recommended materials, cost estimates, and available financial assistance such as the "Ma Prime Rénov'" system. This interactive system is also capable of providing personalized diagnostics to help identify the best renovation opportunities based on the specific characteristics of each home.

By integrating up-to-date data and working closely with energy experts, the chatbot always stays up to date with the latest energy renovation regulations and technologies. This constant updating ensures that users receive advice that not only meets current standards but also anticipates future trends in the energy market.

The development of this chatbot is part of a broader approach to digitalization of housing-related services, where energy efficiency becomes accessible to all. By offering such an innovative tool, we hope not only to facilitate the energy transition of buildings but also to encourage widespread awareness of the importance of eco-responsibility. This project perfectly illustrates how technology and innovation can be put to the service of the environment and society, paving the way for sustainable solutions that respond to the energy and environmental challenges of our time.

Résumé

Dans le contexte actuel où la lutte contre le changement climatique est devenue une priorité, la rénovation énergétique des logements représente une démarche essentielle pour réduire les émissions de carbone. Notre projet de chatbot vise à transformer ce défi en une opportunité accessible et gérable pour les propriétaires. Cet assistant virtuel, développé grâce aux technologies d'intelligence artificielle les plus avancées, est conçu pour guider les utilisateurs pas à pas à travers les différentes étapes de la rénovation énergétique. En fournissant des informations personnalisées et des conseils pratiques, le chatbot aide à simplifier le processus complexe de rénovation, en rendant les normes énergétiques moins intimidantes et plus compréhensibles.

Ce chatbot est alimenté par des modèles linguistiques de pointe qui lui permettent de comprendre et de répondre aux questions spécifiques des utilisateurs avec une précision remarquable. Grâce à une interface intuitive, les utilisateurs peuvent facilement accéder à des informations sur les meilleures pratiques, les matériaux recommandés, les estimations de coûts, et les aides financières disponibles comme le dispositif "Ma Prime Rénov". Ce système interactif est également capable de fournir des diagnostics personnalisés pour aider à identifier les meilleures opportunités de rénovation en fonction des caractéristiques spécifiques de chaque logement.

En intégrant des données actualisées et en collaborant étroitement avec des experts en énergie, le chatbot reste toujours à jour avec les dernières réglementations et technologies en matière de rénovation énergétique. Cette mise à jour constante garantit que les utilisateurs reçoivent des conseils qui non seulement respectent les normes actuelles mais anticipent également les tendances futures du marché de l'énergie.

Le développement de ce chatbot s'inscrit dans une démarche plus large de digitalisation des services liés à l'habitat, où l'efficacité énergétique devient accessible à tous. En proposant un outil aussi innovant, nous espérons non seulement faciliter la transition énergétique des bâtiments mais aussi encourager une prise de conscience généralisée sur l'importance de l'éco-responsabilité. Ce projet illustre parfaitement comment la technologie et l'innovation peuvent être mises au service de l'environnement et de la société, en ouvrant la voie à des solutions durables qui répondent aux défis énergétiques et environnementaux de notre époque.

Table des matières

Table des figures	vii
Introduction Générale	2
1 Cadre général du projet	3
1.1 Présentation de l'organisme d'accueil	3
1.1.1 MedCREtech	3
1.1.2 Mission	4
1.1.3 Organigramme	4
1.2 Présentation de MedCREtech	5
1.3 Contexte général	6
1.3.1 Problématique :	6
1.3.2 Objectif du projet	7
1.4 Présentation du chatbot	7
1.5 Méthodologie de travail	8
1.6 Étude de l'existant	9
1.6.1 Définition de la rénovation énergétique	9
1.6.2 Enjeux et bénéfices de la rénovation énergétique	10
1.6.3 Outils et technologies actuelles	10
1.6.4 Applications des chatbots dans l'énergie	10
1.6.5 Méthodes et technologies utilisées	11
1.6.6 Distinction de MedCREtech	11
1.7 conclusion	11
2 Partie théorique	12
2.1 Introduction aux Chatbots et à l'Intelligence Artificielle	12
2.1.1 Historique et évolution des chatbots	12
2.1.2 Applications des chatbots dans le domaine de l'efficacité énergétique	13
2.2 Techniques de Collecte et de Préparation des Données	15
2.2.1 Web Scraping	15
2.2.2 Utilisation de l'API de la Plateforme Reddit	16
2.2.3 Reconnaissance optique de caractères (OCR)	17
2.3 Techniques de Modélisation et d'Apprentissage Automatique	18
2.3.1 Introduction au Deep Learning	18
2.4 Large Language Models (LLM) : Fondements Techniques et Implications	21
2.4.1 Architecture et Mécanismes des LLM	21
2.4.2 Processus d'Entraînement des LLM	21

2.4.3	Impact du Fine-Tuning sur les Performances des LLM	24
2.4.4	Applications Stratégiques des LLM	24
2.4.5	Enjeux et Perspectives	24
2.5	Modèles de Traitement du Langage Naturel (NLP)	25
2.5.1	LLAMA2	25
2.5.2	LLAMA3	27
2.5.3	GEMMA	28
2.5.4	Limites de Colab pour LLAMA et Gemma	32
2.6	Métriques d'Évaluation	32
2.6.1	Métriques Basées sur le Chevauchement	32
2.6.2	Métriques Basées sur la Similarité Sémantique	39
2.6.3	Métriques Basées sur la Précision	40
2.6.4	Métriques Personnalisées	40
2.7	Conclusion	41
3	Collecte de Données	42
3.1	Collecte des liens des posts Facebook avec le Web Scraping	42
3.1.1	Introduction	42
3.1.2	Étape de la Collecte des liens	42
3.2	Navigation et Extraction des Données des Posts Facebook	45
3.2.1	Introduction	45
3.2.2	Navigation vers les Posts et Extraction du Texte	45
3.2.3	Affichage et Extraction des Commentaires	45
3.2.4	Gestion des Erreurs et Continuité	46
3.2.5	Stockage des Données	46
3.3	Collecte de Données à partir de l'API Reddit	47
3.3.1	Étapes de la Collecte de Données avec l'API Reddit	47
3.4	Extraction de Texte à partir de Fichiers PDF avec l'OCR	49
3.4.1	Introduction	49
3.4.2	Processus d'Extraction de Texte	49
3.4.3	Génération de Questions et Réponses	49
3.5	Extraction des Caractéristiques d'Articles à l'Aide de Techniques NLP	50
3.5.1	Introduction	50
3.5.2	Processus d'Extraction des Caractéristiques	50
3.5.3	Génération Automatique de Questions-Réponses	50
3.6	Description de la base de données collectées	51
3.7	Préparation des Données	53
3.7.1	Introduction	53
3.7.2	Nettoyage des Données	53
3.7.3	Transformation des Données	53
3.7.4	Séparation des Données	54
3.8	Conclusion	54
4	Résultats des modèles NLP	55
4.1	Entraînement des Modèles avec GEMMA 2B	55
4.2	Entraînement des Modèles avec LLAMA2	63
4.3	Entraînement des modèles avec LLAMA3	74

4.4	Comparaison des Métriques pour les Modèles GEMMA, LLAMA2, et LLAMA3 . . .	82
4.5	Illustration des Performances du Modèle LLAMA3	84
4.6	Conclusion	1
Conclusion et perspectives		2
Bibliographie		3

Table des figures

1.1	Logo de MedCREtech	3
1.2	Organigramme MEDCRETECH	5
1.3	Modèle CRISP-DM	10
2.1	Chatbot ELIZAcaption (Source : [1])	13
2.2	Logos des bibliothèques du web scraping	15
2.3	Schéma explicatif l'application de de redit pour avoir l'API (Source : [2])	16
2.4	Schéma explicatif du fonctionnement de l'OCR avec Tesseract	17
2.5	Structure de l'Artificial Neural Network (Source : [3])	18
2.6	Structure des réseaux de neurones Convolutifs (Source : [4])	19
2.7	Structure des Réseaux de Neurones Récurents (Source : [5])	20
2.8	Structure des réseaux de neurones profonds (Source : [6])	20
2.9	Fonctionnement des LLM (Source : [?])	21
2.10	Fonctionnement des transformers (Source : [7])	22
2.11	Principe de Fine-tuning	23
2.12	Technique et performances de fine-tuning (Source : [8])	23
2.13	technique de Fine-Tuning (Source : [9])	24
2.14	Logo de LLAMA2	25
2.15	Logo de LLAMA3	28
2.16	Logo de Gemma de Google	29
2.17	Architecture LLAMA. (Source : [10])	31
3.1	Extraction des liens des posts à partir de href	43
3.2	Liens posts facebook	44
3.3	Extraction des commentaires	46
3.4	Application de développement Reddit (Source : [11])	47
3.5	Redit Dataset	48
3.6	Principe de l'OCR (Source : [12])	49
3.7	Base de données collectées	51
3.8	Description du contenu de la base de données	51
3.9	Exemple de questions de la base de données	52
3.10	Répartition de la base de données selon les sources d'informarion	53
4.1	Résultats d'entraînement avec l'optimizer AdamW_Torch	56
4.2	Résultats d'entraînement avec l'optimizer Adam_8bit	56
4.3	Courbe du Nombre d'Epoch et Courbe de la Norme du Gradient	57
4.4	Courbe du Taux d'Apprentissage et Courbe de la Perte d'Entraînement	58
4.5	Résultats du BLEU score du GEMMA 2b	59

4.6	Résultats du ROUGE-1 score du GEMMA	60
4.7	Résultats du ROUGE-2 score du GEMMA	61
4.8	Résultats du ROUGE-L	61
4.9	Résultats du METEOR score et de la distance levenshtein du Gemma	62
4.10	Résultats du Bertscore score du Gemma	63
4.11	Résultats de l'entraînement du modèle LLAMA2 après finetuning	64
4.12	Courbe d'Époque d'Entraînement et Courbe de la Norme du Gradient	65
4.13	Courbe du Taux d'Apprentissage et Courbe de la Perte	66
4.14	Résultats du BLEU SCORE de LLAMA2	67
4.15	Résultats du ROUGE-1 score du LLAMA2	69
4.16	Résultat du ROUGE-2 score de LLAMA2	70
4.17	Résultat du ROUGE-L score de LLAMA2	71
4.18	Résultats du METEOR score de LLAMA2	72
4.19	Résultat du Bertscore de LLAMA2	73
4.20	Courbe d'Époque d'Entraînement et Courbe de la Norme du Gradient	75
4.21	Courbe du Taux d'Apprentissage et Courbe de la Perte d'Entraînement	76
4.22	Résultat BLUE SCORE du modèle LLAMA3	77
4.23	Résultats du METEOR score de LLAMA3	78
4.24	Diagramme de comparaison des modèles	83
4.25	Résultat BLUE SCORE du modèle LLAMA3	84
4.26	Résultat BLUE SCORE du modèle LLAMA3	1

Introduction Générale

Les stratégies de développement durable mettent l'accent sur la rénovation énergétique afin de diminuer les émissions de gaz à effet de serre tout en améliorant la performance énergétique des bâtiments. Le Diagnostic de Performance Énergétique (DPE) est considéré en France comme un outil essentiel pour évaluer et optimiser l'efficacité énergétique des logements. Le DPE est un indicateur standardisé qui évalue à la fois la consommation d'énergie et les émissions de gaz à effet de serre d'un édifice. La performance énergétique est donc évaluée en utilisant une étiquette allant de A (très performant) à G (peu performant), ce qui permet d'avoir une vision claire de la performance.

Il est crucial que cette évaluation permette aux propriétaires et aux locataires de comprendre les performances énergétiques de leur logement et de repérer les domaines où des améliorations sont nécessaires. Le diagnostic de performance énergétique (DPE) est un élément indispensable des politiques de rénovation énergétique en France, où il est obligatoire lors de la vente ou de la location d'un logement. Ces résultats sont utilisés afin de guider les travaux de rénovation et de donner la priorité aux interventions les plus performantes. Il est reconnu que cette obligation renforce la prise de conscience des propriétaires et des locataires quant à l'importance de l'économie d'énergie.

Ce projet de fin d'études vise principalement à concevoir un chatbot intelligent qui sera utile à la fois aux particuliers et aux professionnels dans leurs projets de rénovation énergétique. Cet outil intelligent sera régulièrement mis à jour et bénéficiera fréquemment de l'intervention d'experts en énergie. L'objectif est de créer un outil interactif capable de répondre de façon précise et rapide aux différentes questions concernant la rénovation énergétique. La plateforme "Mon Carnet de Logement" intègre un chatbot anonyme qui offre des conseils sur mesure et interactifs grâce à l'utilisation de technologies de pointe. Des algorithmes d'apprentissage automatique sont employés afin de s'assurer que les réponses fournies soient actualisées et pertinentes. Il est aussi espéré que ce chatbot permettra de simplifier l'accès à des renseignements essentiels concernant les subventions et les aides financières disponibles pour les projets de rénovation. En outre, l'objectif est que cet outil puisse proposer des conseils précis en se basant sur les particularités de chaque logement, ce qui permettrait d'améliorer l'expérience utilisateur et l'efficacité des travaux de rénovation entrepris.

Ce rapport est structuré en quatre chapitres principaux. Le premier chapitre introduit l'entreprise d'accueil et le contexte global du projet. Le deuxième chapitre détaille les fondements théoriques des méthodes de collecte de données utilisées, ainsi que des Modèles de Langage à Grande Échelle (LLM), y compris les technologies de web scraping, l'utilisation de l'API Reddit, et la reconnaissance optique de caractères (OCR). Nous explorons également trois types de LLM, notamment LLAMA2 et LLAMA3 de Meta, ainsi que GEMMA de Google. Le troisième chapitre présente en détail la base de données constituée. Enfin, le quatrième chapitre propose une analyse comparative des performances et limites des modèles LLAMA2, LLAMA3, et GEMMA.

Chapitre 1

Cadre général du projet

Ce chapitre se compose de trois parties : La première partie présente l'organisme d'accueil, la deuxième partie présente le contexte du projet tels que la problématique, les objectifs fixés dans le cadre de ce stage et la méthodologie du travail utilisé et finalement l'étude de l'existant.

1.1 Présentation de l'organisme d'accueil

1.1.1 MedCREtech

Depuis sa création en 2019, MedCREtech s'est rapidement établie comme une entreprise innovante dans le secteur immobilier, bénéficiant du label "STARTUP ACT" en Tunisie. La mission de MedCREtech est de révolutionner le marché immobilier en France et en Tunisie en intégrant les technologies de pointe comme le Big Data, l'intelligence artificielle, et la visualisation des données. MedCREtech est dédiée à transformer la gestion, la vente, et la location de biens immobiliers, en rendant ces processus plus efficaces et adaptés aux besoins spécifiques de ses clients.

Un de ses projets phares a été le développement d'une architecture Big Data capable de traiter d'importantes quantités de données, optimisant ainsi la gestion des informations immobilières. MedCREtech a également mis en place un générateur d'annonces propulsé par OpenAI GPT-3, qui produit des descriptions précises et détaillées des propriétés, aidant ainsi les utilisateurs à faire des choix éclairés.



FIGURE 1.1 – Logo de MedCREtech

1.1.2 Mission

La vision de MedCREtech est de numériser le secteur immobilier en Tunisie tout en apportant ses innovations au marché français. MedCREtech s'engage à rester à l'avant-garde de la technologie, non seulement en améliorant les processus existants mais aussi en créant des solutions novatrices qui répondent aux défis contemporains de l'immobilier.

Stratégie de Développement

- **Digitalisation** : L'objectif principal de Medcretech est de moderniser le secteur immobilier en utilisant des technologies avancées pour simplifier et optimiser les procédures immobilières.
- **Diversification** : MedCREtech cherche constamment à élargir sa gamme de services pour couvrir tous les aspects du marché immobilier, répondant ainsi aux divers besoins de ses clients.
- **Innovation** : L'investissement continu dans la recherche et le développement est crucial pour MedCREtech. L'entreprise s'efforce d'intégrer les dernières innovations technologiques pour anticiper et répondre aux besoins futurs du marché.

1.1.3 Organigramme

MedCREtech est structurée autour de trois principaux départements qui collaborent étroitement pour atteindre ses objectifs ambitieux.

Départements

- **Marketing** : Ce département joue un rôle essentiel dans la promotion des services innovants de MedCREtech, en développant des stratégies marketing ciblées et en gérant ses campagnes de communication.
- **Ressources Humaines** : Gérer le talent est une priorité pour MedCREtech. Ce département s'occupe du recrutement, de la formation, et du développement professionnel de l'équipe, assurant que l'entreprise reste à la pointe de l'innovation.
- **Informatique** : Au cœur de l'innovation, ce département développe et maintient les solutions technologiques de MedCREtech, garantissant que l'entreprise répond aux attentes de ses clients en termes de performance et de sécurité.

Collaboration Interdépartementale

L'interaction entre ces départements est cruciale pour le succès de MedCREtech, garantissant une synergie qui reflète l'engagement de l'entreprise envers l'excellence et l'innovation.

1.2 Présentation de MedCREtech

Depuis 2019, Medcretech est une jeune entreprise tunisienne qui a reçu le label "STARTUP ACT" et se spécialise dans la création de solutions immobilières novatrices pour le marché français et tunisien. La startup met l'accent sur la création et l'application de technologies de pointe pour transformer le domaine de l'immobilier. Les solutions de Big Data, d'intelligence artificielle et de visualisation des données sont intégrées afin de proposer des services plus performants et sur mesure. Il s'agit de simplifier la gestion, la vente et la location de biens immobiliers, tout en améliorant l'efficacité et la satisfaction des utilisateurs.

Parmi les projets menés par MedCREtech, on trouve la création d'une architecture Big Data qui peut traiter de vastes quantités de données, ce qui permet d'améliorer la gestion des informations. Un générateur d'annonces utilisant OpenAI GPT-3 a également été créé par la startup afin de fournir des descriptions précises et détaillées des logements, ce qui permet aux utilisateurs de prendre des décisions informées. De plus, il y a un carnet numérique du logement afin de regrouper tous les documents et équipements des biens immobiliers, ce qui facilite leur gestion tant pour les propriétaires que pour les bailleurs.

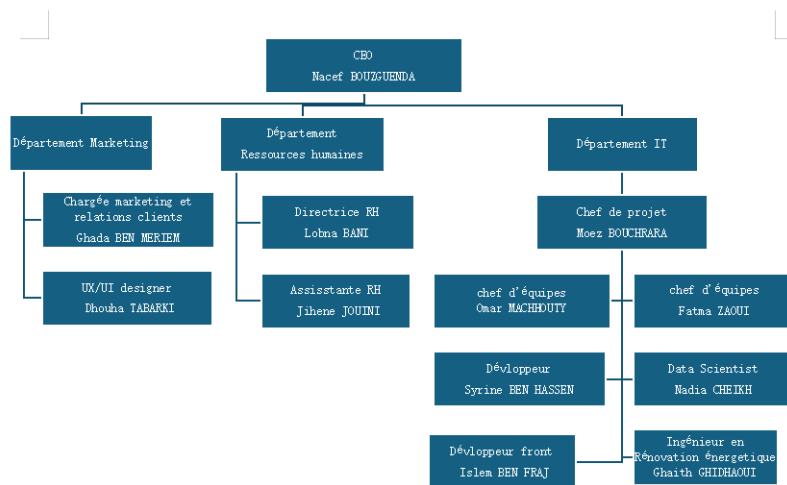


FIGURE 1.2 – Organigramme MEDCRETECH

1.3 Contexte général

1.3.1 Problématique :

Le changement climatique est une réalité urgente, et l'une des réponses efficaces que nous pouvons apporter est l'amélioration de l'efficacité énergétique des bâtiments. En France, la rénovation énergétique des logements est donc devenue une priorité nationale. Le Diagnostic de Performance Énergétique (DPE) est un outil crucial pour orienter les propriétaires dans la planification et la priorisation de leurs efforts de rénovation. Cependant, le DPE peut souvent sembler complexe en raison de sa richesse de données et de la variété des options qu'il propose, ce qui peut intimider et décourager non seulement les propriétaires mais aussi les professionnels du domaine.

Les principaux obstacles à une rénovation efficace incluent non seulement la complexité des données techniques mais aussi la diversité des solutions de rénovation possibles et les lourdeurs des procédures administratives. Ces difficultés peuvent considérablement retarder, voire décourager les initiatives de rénovation. En outre, les propriétaires, souvent non experts en énergie, peuvent se sentir dépassés par les exigences techniques et l'éventail des choix possibles. Parallèlement, les auditeurs et les diagnostiqueurs énergétiques doivent naviguer à travers un grand volume de données et développer des recommandations sur mesure, un processus qui peut être à la fois long et complexe.

Face à ces défis, l'adoption de technologies avancées telles que les chatbots offre une opportunité d'amélioration significative. Medcretech, à la pointe de l'innovation technologique, développe un assistant virtuel intelligent spécifiquement conçu pour naviguer dans les complexités de la rénovation énergétique. Ce chatbot, qui sera intégré au site "Mon Carnet de Logement", ne se contente pas de guider les utilisateurs à travers les étapes de la rénovation ; il rend également les données du DPE facilement accessibles et compréhensibles, offrant des explications claires et des recommandations personnalisées en fonction des besoins spécifiques de chaque logement.

Ce chatbot exploite des technologies avancées de traitement du langage naturel pour interagir en temps réel avec les utilisateurs, répondant à leurs questions, décodant les informations techniques du DPE, et proposant des solutions de rénovation optimales. De cette manière, il facilite non seulement la compréhension mais également l'application pratique des recommandations du DPE, rendant le processus de rénovation moins intimidant et plus accessible.

- Quels sont les principaux défis techniques dans la collecte et le traitement des données pour un projet dédié à la rénovation énergétique ?
- Quel modèle de chatbot est le plus efficace pour fournir des recommandations personnalisées dans ce domaine spécifique ?
- Comment un chatbot peut-il améliorer l'application des audits énergétiques et des DPE ?

En abordant ces questions, notre projet cherche à démontrer comment une intervention technologique peut simplifier et accélérer le processus de rénovation énergétique, en apportant une aide précieuse aux propriétaires et aux professionnels du secteur.

1.3.2 Objectif du projet

L'objectif de ce projet est de développer un chatbot intégré au site "Mon Carnet de Logement" qui accompagne diagnostiqueurs, auditeurs énergétiques et particuliers tout au long de leur processus de rénovation énergétique. En utilisant des techniques avancées de traitement du langage naturel et des modèles prédictifs, on peut analyser les besoins spécifiques des utilisateurs et fournir des recommandations personnalisées pour améliorer l'efficacité énergétique de leurs logements. Le chatbot a pour but de clarifier les rapports et les audits énergétiques en offrant des explications simples et des conseils adaptés, rendant accessible l'analyse détaillée des consommations d'énergie d'un bâtiment et proposant des améliorations telles que l'installation de fenêtres à double vitrage, l'ajout d'isolation thermique ou le remplacement de systèmes de chauffage obsolètes.

Pour mener à bien ce projet, on doit collecter les données nécessaires en utilisant des techniques comme le web scraping, l'API Reddit et l'OCR pour les documents PDF. Une fois les données collectées, on les nettoie et les structure pour une analyse efficace. Ensuite, on applique différents modèles de traitement du langage naturel, tels que LLAMA 2, LLAMA 3 et Ggema 2, pour fournir des recommandations personnalisées. Chaque modèle est évalué en utilisant des métriques spécifiques comme BLEU, ROUGE, METEOR et BertScore pour mesurer leur performance. Ces modèles sont choisis pour leur capacité à comprendre et générer du langage naturel de manière précise et contextuelle, avec des atouts distincts dans la gestion des conversations complexes et l'analyse des informations.

Après l'évaluation des modèles, on compare leurs performances pour sélectionner celui qui offre les meilleures recommandations. Grâce à cette approche méthodique, le chatbot sera capable de fournir des conseils pratiques et d'expliquer clairement les bénéfices de chaque intervention, aidant ainsi les utilisateurs à prendre des décisions éclairées et à optimiser les résultats de leurs investissements en rénovation énergétique.

1.4 Présentation du chatbot

Le chatbot fait partie de la plateforme "Mon Carnet de Logement", une interface en ligne spécialement conçue pour gérer et améliorer les performances énergétiques des logements. Cette plateforme propose une variété étendue de services et d'outils afin d'assister les propriétaires dans la gestion de leur logement de manière plus performante et durable.

Le chatbot utilise des technologies avancées en traitement du langage naturel et en apprentissage automatique pour interagir en temps réel avec les utilisateurs afin de répondre à leurs interrogations, les accompagner dans l'interprétation des diagnostics énergétiques et les orienter vers les mesures à prendre. Prenons l'exemple d'un utilisateur qui reçoit un DPE qui indique une performance énergétique médiocre. Le chatbot peut expliquer les résultats, suggérer des solutions concrètes pour améliorer la performance, et fournir des informations sur les subventions et les aides disponibles pour financer les travaux.

Dans cette optique, le chatbot a pour objectif de devenir un assistant virtuel indispensable pour toute personne désireuse d'améliorer l'efficacité énergétique de son domicile. En plus de faciliter les procédures administratives et techniques, il contribue également à sensibiliser les utilisateurs aux enjeux environnementaux et économiques liés à la rénovation énergétique.

En intégrant ces différentes technologies, modèles et sources, le chatbot pour la rénovation énergétique des logements peut offrir une expérience utilisateur enrichissante tout en contribuant à accélérer l'adoption de pratiques plus durables dans le domaine de l'habitat.

1.5 Méthodologie de travail

La méthodologie standard CRISP-DM (Cross Industry Standard Process for Data Mining) est utilisée pour gérer des projets de data mining. Elle se compose de six étapes distinctes qui facilitent la gestion d'un projet de manière méthodique et rigoureuse. Dans le cadre de notre projet de conception d'un chatbot pour la rénovation énergétique, nous avons adopté cette approche pour garantir une organisation optimale et une performance élevée. Les étapes suivantes sont :

1. **Compréhension du domaine** : Cette étape implique de collecter des données concernant le domaine d'application du chatbot. Il est crucial de comprendre les objectifs du chatbot, les exigences de qualité des données, les contraintes techniques, les données disponibles et les hypothèses de base. Pour notre projet, cela signifie saisir les besoins des utilisateurs en matière de rénovation énergétique et les différentes questions qu'ils peuvent poser. Nous devons également prendre en compte les réglementations et les meilleures pratiques du secteur pour fournir des réponses précises et utiles.
2. **Compréhension des données** : Cette étape consiste à collecter, comprendre et analyser les informations disponibles. Nous devons vérifier la qualité, la quantité et la pertinence des données. Dans le cadre de notre projet, cela inclut l'analyse des bases de données existantes sur la rénovation énergétique, des forums, des blogs et d'autres sources pertinentes. Cette phase est cruciale pour identifier les lacunes dans les données et déterminer comment les combler.
3. **Préparation des données** : Dans cette étape, il s'agit de préparer les données pour l'analyse en les nettoyant, les transformant et les sélectionnant. Il est nécessaire de gérer les données manquantes, de traiter les valeurs aberrantes et de normaliser les données. Pour notre chatbot, il est essentiel de structurer les données sous forme de questions-réponses afin de faciliter leur intégration et leur utilisation par les modèles d'apprentissage automatique. Cette structuration permet également de fine-tuner les modèles pour améliorer leur performance.
4. **Modélisation** : Cette étape consiste à développer des modèles statistiques ou algorithmiques pour résoudre les problèmes identifiés lors des phases précédentes. Nous devons évaluer les modèles en termes de performance et de pertinence. Pour notre projet, cela implique la création de modèles de traitement du langage naturel (NLP) capables de comprendre et de répondre de manière précise et cohérente aux questions des utilisateurs. Nous utilisons des modèles open

source de Hugging Face, tels que LLAMA 2, LLAMA 3 et Ggema 2, que nous fine-tunons pour optimiser leur performance.

5. **Évaluation** : L'objectif de cette étape est d'évaluer la qualité et la pertinence des modèles développés. Nous utilisons des mesures de performance adéquates, telles que la précision, le rappel et les scores F1, pour évaluer l'efficacité des réponses produites par notre chatbot. Les résultats sont comparés à ceux d'autres modèles et présentés aux parties prenantes pour validation. Cette phase inclut également la vérification par des experts en énergie pour garantir l'exactitude des informations fournies par le chatbot.
6. **Déploiement** : Dans cette étape, il s'agit de mettre en œuvre les modèles dans un contexte de production. Nous devons tester et valider les modèles pour garantir leur bon fonctionnement. Le chatbot est intégré à la plateforme "Mon Carnet de Logement" et doit répondre aux utilisateurs en temps réel. Le déploiement inclut également la surveillance continue du chatbot pour assurer sa performance et apporter des améliorations en fonction des retours des utilisateurs et des nouvelles données.

On a suivi cette méthodologie de travail pour plusieurs raisons :

- Elle offre une organisation claire et précise pour la gestion des projets de collecte de données, ce qui est crucial pour le développement méticuleux de notre chatbot.
- Cette méthode est souple et peut être ajustée en fonction des exigences spécifiques de notre projet. Chaque étape peut être personnalisée en fonction des ressources disponibles, de la complexité des données et des besoins de l'application. La conception d'un chatbot repose sur une connaissance approfondie du domaine, essentielle pour fournir des réponses pertinentes et utiles sur la rénovation énergétique.
- elle utilise une approche itérative, permettant de revenir en arrière et de modifier les étapes précédentes si nécessaire. Cela s'avère particulièrement bénéfique pour améliorer continuellement notre chatbot en fonction des retours des utilisateurs et des nouvelles informations.

1.6 Étude de l'existant

1.6.1 Définition de la rénovation énergétique

La rénovation énergétique vise à améliorer l'efficacité énergétique des bâtiments afin de réduire la consommation d'énergie et les émissions de gaz à effet de serre (GES). Cela inclut des interventions telles que l'amélioration de l'isolation thermique, le remplacement des systèmes de chauffage et de refroidissement inefficaces, l'installation de fenêtres à double vitrage, et l'utilisation de sources d'énergie renouvelable. Un élément central de cette démarche est le Diagnostic de Performance Énergétique (DPE), qui évalue la consommation d'énergie d'un bâtiment et son impact en termes d'émissions de GES.

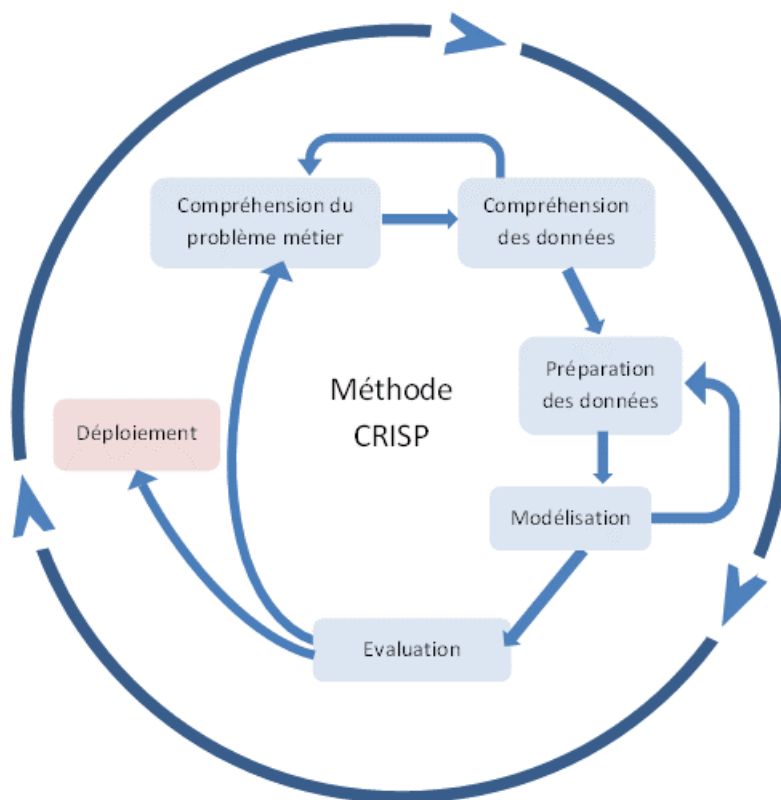


FIGURE 1.3 – Modèle CRISP-DM

1.6.2 Enjeux et bénéfices de la rénovation énergétique

La rénovation énergétique réduit les GES et combat le changement climatique. Économiquement, elle diminue les factures d'énergie et augmente la valeur des biens immobiliers. Socialement, elle améliore le confort des occupants et réduit la précarité énergétique. Le DPE joue un rôle crucial en fournissant aux propriétaires des informations claires sur la performance énergétique de leur logement et les axes d'amélioration possibles.

1.6.3 Outils et technologies actuelles

Les outils et technologies actuelles pour la rénovation énergétique incluent les audits énergétiques, les logiciels de simulation thermique, les systèmes de gestion de l'énergie et les dispositifs de mesure. Les technologies comprennent les matériaux d'isolation avancés, les fenêtres à haute performance, les pompes à chaleur et les panneaux solaires.

1.6.4 Applications des chatbots dans l'énergie

Les chatbots dans le domaine de l'énergie, comme "Engie Bot" d'Engie et "Wattson" de Direct Energie, aident les utilisateurs à comprendre leur consommation d'énergie et à prendre des mesures pour la réduire. Ils fournissent des conseils personnalisés, expliquent les audits énergétiques et aident à obtenir des subventions. Cependant, notre projet se distingue par l'intégration de questions d'actualité non

encore disponibles ailleurs et par l'intervention d'experts en énergie pour valider et affiner les réponses du chatbot.

1.6.5 Méthodes et technologies utilisées

Notre projet utilise des techniques de web scraping, l'API Reddit et la reconnaissance optique de caractères pour collecter les données. Les données sont ensuite nettoyées, structurées et mises sous forme de questions-réponses pour une analyse efficace. Nous appliquons différents modèles NLP open source de Hugging Face, tels que LLAMA 2, LLAMA 3 et Ggema 2, et les fine-tunons pour fournir des recommandations personnalisées. Chaque modèle est évalué avec des métriques spécifiques comme BLEU, ROUGE, METEOR et BertScore pour déterminer le plus performant. Le chatbot offre ensuite des conseils pratiques et explique les bénéfices des interventions en rénovation énergétique. De plus, les questions d'actualité sont intégrées pour offrir des informations toujours à jour, et les experts en énergie interviennent pour valider et ajuster les réponses du chatbot, assurant ainsi une précision et une pertinence optimales.

1.6.6 Distinction de MedCREtech

MedCREtech se distingue des autres entreprises dans le domaine de l'énergie par plusieurs aspects innovants. Contrairement aux chatbots existants comme "Engie Bot" et "Wattson", notre chatbot intègre des questions d'actualité qui ne sont pas encore disponibles dans d'autres solutions, offrant ainsi des informations toujours à jour aux utilisateurs. De plus, nous avons une équipe d'experts en énergie qui interviennent pour valider et affiner les réponses fournies par le chatbot, assurant une précision et une pertinence optimales. Cette approche hybride, combinant les capacités avancées des modèles NLP fine-tunés et l'expertise humaine, permet à MedCREtech de proposer une solution unique et plus efficace pour accompagner les utilisateurs dans leurs projets de rénovation énergétique.

1.7 conclusion

En conclusion de ce chapitre, nous avons présenté le cadre général du projet, en mettant en évidence l'organisme d'accueil MerCretech, la problématique à résoudre et l'objectif du projet. Nous avons également abordé l'étude de l'existant en discutant de l'état de l'art sur la rénovation énergétique et l'implication

Chapitre 2

Partie théorique

2.1 Introduction aux Chatbots et à l'Intelligence Artificielle

Les chatbots sont des logiciels développés pour reproduire des échanges humains. Les algorithmes de traitement du langage naturel (NLP) sont employés afin de saisir les demandes des utilisateurs et de leur fournir des réponses adéquates. Il est essentiel d'utiliser l'intelligence artificielle (IA), notamment l'apprentissage automatique (machine learning), afin de créer des chatbots capables d'apprendre et de s'adapter en permanence.

2.1.1 Historique et évolution des chatbots

- **Premiers chatbots** : Les chatbots initiaux, tels que ELIZA (1966) et PARRY (1972), employaient des scripts basiques afin de simuler des échanges. Joseph Weizenbaum a développé ELIZA, qui permettait aux utilisateurs de répondre en reformulant leurs déclarations sous forme de questions, imitant ainsi un thérapeute. Parry, créé par Kenneth Colby, était une simulation d'un patient atteint de schizophrénie paranoïde qui répondait en utilisant des règles heuristiques.
- **Évolution** : Grâce aux avancées en matière d'IA et de NLP, les chatbots contemporains ont la capacité de comprendre et de produire du langage naturel, ce qui permet des interactions plus naturelles et complexes. La mise au point de technologies telles que les réseaux de neurones profonds et les modèles de transformateurs a entraîné des progrès notables dans la compréhension et la production de langage. Parmi les exemples actuels, on peut citer Siri d'Apple, Alexa d'Amazon et Google Assistant, qui exploitent des méthodes de machine learning avancées afin de proposer des expériences utilisateur enrichies.

```

Welcome to
          EEEEE LL   IIII ZZZZZZ AAAAA
          EE   LL   II   ZZ   AA   AA
          EEEEE LL   II   ZZZ   AAAAAA
          EE   LL   II   ZZ   AA   AA
          EEEEE LLLLLL IIII ZZZZZZ AA   AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:

```

FIGURE 2.1 – Chatbot ELIZAcaption (Source : [1])

2.1.2 Applications des chatbots dans le domaine de l'efficacité énergétique

Les chatbots spécialement conçus pour l'efficacité énergétique peuvent proposer des solutions novatrices et pratiques afin d'assister les utilisateurs dans l'optimisation de leur consommation d'énergie et dans la réalisation de projets de rénovation énergétique. Voici quelques exemples d'applications particulières dans ce secteur :

- **Surveillance et gestion de l'énergie :** Il est possible d'incorporer des chatbots dans des systèmes de gestion de l'énergie afin de suivre en temps réel la consommation d'énergie d'un édifice. En cas de surconsommation, ils ont la capacité d'informer les utilisateurs, de fournir des rapports détaillés sur l'utilisation de l'énergie et de proposer des mesures correctives pour diminuer la consommation.

Exemple : Un chatbot connecté à un système de gestion de l'énergie a la capacité d'envoyer des alertes lorsque la consommation énergétique dépasse un seuil donné, avec des recommandations précises pour diminuer la consommation, telles que l'ajustement du thermostat ou l'extinction des appareils inutilisés.

- **Conseils personnalisés :** Les chatbots ont la capacité d'analyser les comportements de consommation des utilisateurs et les caractéristiques des bâtiments afin de donner des conseils sur mesure pour améliorer l'efficacité énergétique. Par exemple, ils peuvent suggérer des dispositifs plus écologiques, des comportements à adopter pour diminuer la consommation, ou des améliorations particulières à réaliser pour l'isolation du bâtiment.

Exemple : D'après l'analyse des habitudes de consommation et des caractéristiques du bâtiment, un chatbot pourrait proposer l'installation de fenêtres à double vitrage ou l'utilisation de lampes LED.

- **Assistance pour les projets de rénovation énergétique :** Les chatbots ont la capacité de fournir une assistance aux utilisateurs tout au long des projets de rénovation énergétique, depuis l'évaluation initiale des besoins jusqu'à la réalisation des tâches. Ils sont en mesure d'assister dans la détection des subventions et des incitations disponibles, de recommander des experts qualifiés et de suivre l'évolution des travaux.

Exemple : Un chatbot peut assister les utilisateurs dans la recherche de subventions gouvernementales pour des projets de rénovation énergétique, offrir des fournisseurs locaux certifiés,

et suivre les différentes étapes du projet afin de garantir la réalisation des travaux de manière adéquate et dans les délais impartis.

- **Éducation et sensibilisation :** On peut employer les chatbots afin de sensibiliser les utilisateurs aux défis de l'efficacité énergétique et de leur donner des conseils éducatifs sur les meilleures méthodes. Ils ont la possibilité d'offrir des quiz interactifs, des articles informatifs et des conseils concrets afin de promouvoir des comportements plus respectueux de l'environnement.

Exemple : Un chatbot pourrait transmettre des recommandations quotidiennes sur les méthodes pour diminuer la consommation d'énergie, offrir des quiz sur l'efficacité énergétique afin de tester les connaissances des utilisateurs, et fournir des articles détaillés sur les technologies et les pratiques énergétiques durables.

- **Optimisation des systèmes de chauffage, ventilation et climatisation (CVC) :** L'intégration de chatbots dans les systèmes CVC permet aux utilisateurs de recevoir des conseils pour améliorer le fonctionnement de ces systèmes, diminuer les dépenses énergétiques et améliorer le confort intérieur. Les chatbots ont la capacité de recommander des configurations optimales en fonction des conditions météorologiques, des horaires d'occupation et des préférences des utilisateurs.

Exemple : Un chatbot pourrait procéder à une adaptation automatique des paramètres du système de chauffage en fonction des prévisions météorologiques et des préférences de température des résidents, assurant ainsi une consommation énergétique réduite tout en préservant un confort optimal.

- **Gestion des appareils connectés :** Avec l'émergence de l'Internet des objets (IoT), il est possible d'utiliser des chatbots afin de gérer les appareils connectés dans un édifice. Cela offre aux utilisateurs la possibilité de gérer à distance les éclairages, les thermostats et autres appareils électroménagers, ainsi que de créer des scénarios pour optimiser l'efficacité énergétique optimale.

Exemple : Un chatbot intégré à une maison connectée pourrait éteindre les lumières automatiquement et diminuer le chauffage lorsque les résidents quittent la maison, et allumer les appareils de manière régulière afin d'éviter les pics de consommation d'énergie qui se produisent.

Les chatbots dédiés à l'efficacité énergétique peuvent donc jouer un rôle essentiel dans la transition vers des constructions plus durables et économes en énergie, en proposant des solutions pratiques et abordables pour gérer la consommation d'énergie et encourager des comportements plus responsables. Ils ont la capacité d'assister les individus dans la réduction de leurs dépenses énergétiques, tout en contribuant à des objectifs plus étendus de réduction des émissions de carbone et de préservation de l'environnement.

2.2 Techniques de Collecte et de Préparation des Données

Il est crucial de recueillir et de préparer les données afin de créer des modèles de chatbot efficaces. Grâce à ces étapes, il est assuré que les modèles bénéficient de données de qualité pour l'entraînement et l'évaluation, ce qui assure des résultats optimaux.

2.2.1 Web Scraping

Le web scraping est une technique qui permet d'extraire des données de sites web à l'aide d'outils automatisés. Cette méthode est cruciale pour recueillir rapidement de grandes quantités d'informations. Les données ainsi collectées peuvent être utilisées pour diverses applications, notamment l'analyse de données, le marketing numérique, et le développement de modèles d'intelligence artificielle comme les chatbots.

BeautifulSoup est une bibliothèque Python conçue pour l'analyse de documents HTML et XML. Elle construit un arbre de parsing à partir des pages web analysées, rendant l'extraction de données spécifiques rapide et simple. BeautifulSoup est particulièrement efficace pour les tâches de scraping simples où la rapidité est essentielle.

Selenium est un outil d'automatisation qui interagit avec les navigateurs web. Il est souvent utilisé dans le web scraping pour des cas où les données nécessitent une interaction dynamique avec la page web, comme le remplissage de formulaires ou la simulation de clics sur des éléments interactifs. Selenium permet de simuler ces interactions, ce qui est crucial pour accéder à des contenus dynamiquement générés par des scripts côté client.

Lorsqu'ils sont utilisés de manière efficace, ces outils permettent de relever les défis liés à l'extraction de données depuis des sites web complexes ou fortement dynamiques, rendant ainsi les données disponibles pour une analyse ultérieure ou pour alimenter des systèmes exploitant l'intelligence artificielle.



FIGURE 2.2 – Logos des bibliothèques du web scraping

2.2.2 Utilisation de l'API de la Plateforme Reddit

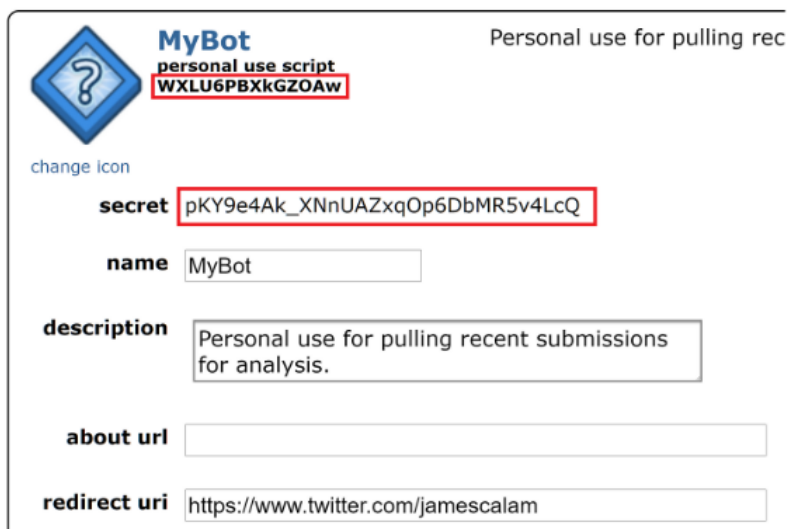
L'API de Reddit est un outil formidable pour accéder de manière programmée à une variété de données, telles que des publications, des commentaires, et des interactions sur différents subreddits. Pour utiliser cette API efficacement, on doit configurer un bot, comme illustré dans l'image fournie, en définissant des éléments essentiels tels que le nom du bot, une description, et des clés d'accès spécifiques.

On commence par enregistrer l'application sur le portail développeur de Reddit pour obtenir des clés *client ID* et *secret*. Ces identifiants sont cruciaux pour authentifier l'application lorsqu'elle fait des requêtes à l'API, garantissant ainsi la sécurité et la légitimité des interactions.

Après l'enregistrement, on complète les détails tels que le *name*, la *description*, et surtout, l'URL de redirection qui permettra à l'API de rediriger les réponses après l'authentification ou lorsqu'il est nécessaire de traiter des actions spécifiques via le bot. Cette étape est essentielle pour établir une communication efficace et sécurisée entre l'application et l'API de Reddit.

Le champ *secret* est crucial pour sécuriser les échanges entre le bot et l'API de Reddit. Il est impératif de garder cette information confidentielle pour éviter les accès non autorisés et assurer que seul le bot puisse faire des requêtes à l'API.

Avec ces configurations, le bot est prêt à interagir efficacement avec l'API pour récupérer des données pertinentes de Reddit. Cette intégration permet une multitude d'applications, allant de l'analyse des tendances à la collecte de retours d'informations, et est précieuse dans les contextes où les données de Reddit peuvent enrichir considérablement un service ou une application.



The image shows a web form for configuring a Reddit bot. At the top left is a blue diamond icon with a question mark and a link 'change icon'. To its right is the title 'MyBot' and the subtitle 'personal use script'. Below the title is a red box containing the text 'WXLU6PBXkGZOAw'. To the right of the title is the text 'Personal use for pulling rec'. Below the icon is the label 'secret' followed by a red box containing the text 'pKY9e4Ak_XNnUAZxqOp6DbMR5v4LcQ'. Below the secret is the label 'name' followed by a text box containing 'MyBot'. Below the name is the label 'description' followed by a text box containing 'Personal use for pulling recent submissions for analysis.'. Below the description is the label 'about url' followed by an empty text box. Below the about url is the label 'redirect uri' followed by a text box containing 'https://www.twitter.com/jamescalam'.

FIGURE 2.3 – Schéma explicatif l'application de de reddit pour avoir l'API (Source : [2])

2.2.3 Reconnaissance optique de caractères (OCR)

La technologie de reconnaissance optique de caractères (OCR) est essentielle pour convertir des documents scannés ou des images contenant du texte en formats éditables et numériques. Elle joue un rôle crucial dans la numérisation des archives papier, rendant les documents facilement accessibles et exploitables pour l'analyse de données. Cette transformation est particulièrement utile dans les environnements où la manipulation rapide et efficace des informations est nécessaire, comme dans les secteurs administratifs, légaux, ou éducatifs.

Parmi les outils disponibles, Tesseract est une des bibliothèques OCR les plus réputées. Initialement développée par HP et désormais maintenue par Google, Tesseract est accessible via l'interface de programmation Python, pytesseract. Cette librairie se distingue par sa capacité à traiter avec précision des documents scannés ou des images, reconnaissant et convertissant le texte en différentes langues et formats. L'avantage majeur de Tesseract réside dans sa flexibilité et son ouverture qui en font un choix privilégié pour les développeurs cherchant à automatiser le processus de numérisation et de transformation des documents papier en données manipulables électroniquement. Cette capacité permet non seulement une réduction significative du temps consacré à la saisie manuelle des données, mais aussi une amélioration de l'accessibilité et de la gestion des archives documentaires dans de nombreux contextes professionnels.

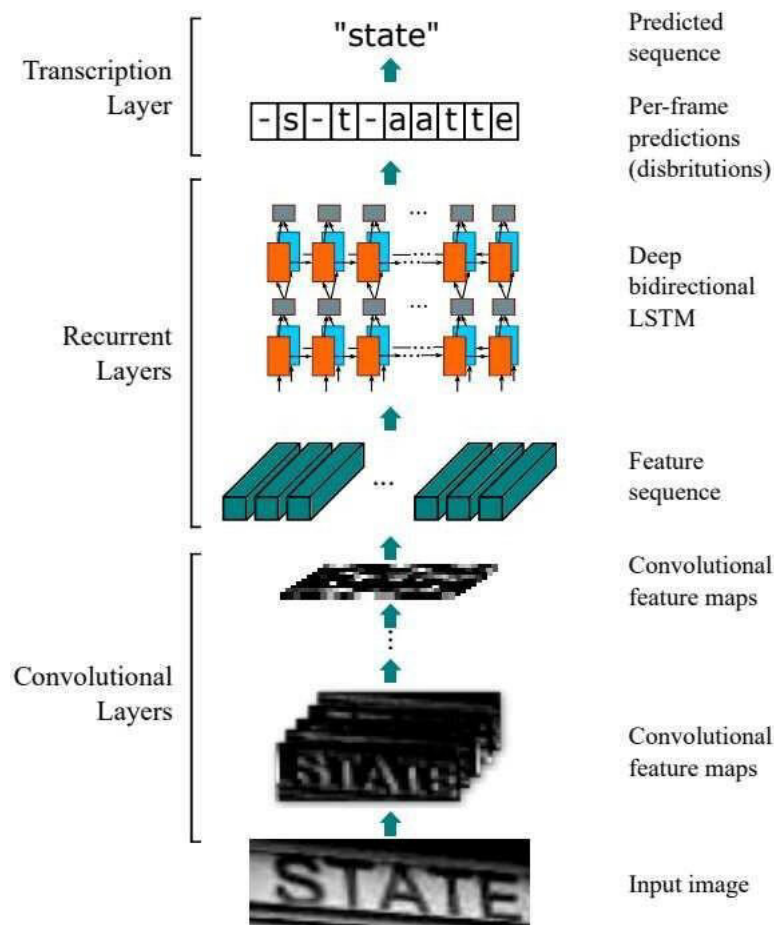


FIGURE 2.4 – Schéma explicatif du fonctionnement de l'OCR avec Tesseract

2.3 Techniques de Modélisation et d'Apprentissage Automatique

2.3.1 Introduction au Deep Learning

Le deep learning est une branche de l'apprentissage automatique (machine learning) basée sur l'architecture des réseaux de neurones artificiels (ANN). Un ANN utilise des couches de nœuds interconnectés appelés neurones qui travaillent ensemble pour traiter et apprendre à partir des données d'entrée.

Structure d'un Réseau de Neurones

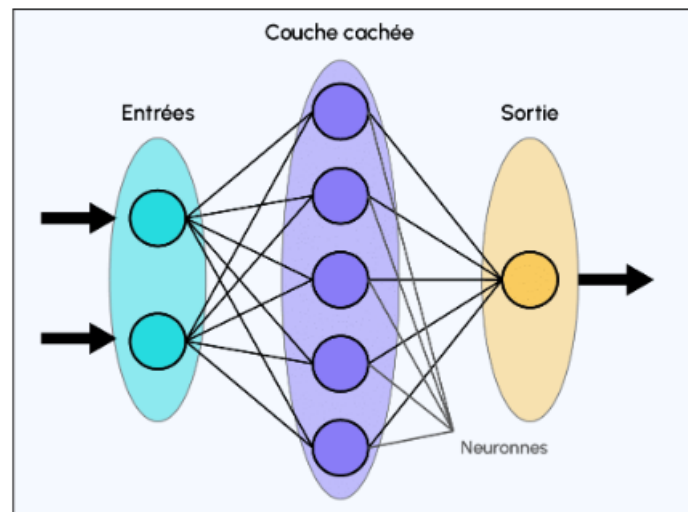


FIGURE 2.5 – Structure de l'Artificial Neural Network (Source : [3])

Dans un réseau entièrement connecté de neurones profonds, il existe une couche d'entrée et une ou plusieurs couches cachées reliées les unes aux autres. Les neurones de la couche précédente ou de la couche d'entrée reçoivent une entrée pour chaque neurone. La sortie d'un neurone est l'entrée des autres neurones de la couche suivante, et ce processus se poursuit jusqu'à ce que la couche finale génère la sortie du réseau. La transformation des données d'entrée par les couches du réseau de neurones se fait à travers une série de transformations non linéaires, ce qui permet au réseau d'acquérir des représentations complexes des données d'entrée.

- **Couche d'Entrée** : La couche initiale reçoit les données brutes et les transmet aux couches suivantes pour traitement. Elle sert de point d'entrée de l'information dans le réseau.

- **Couches Cachées** : Ces couches intermédiaires effectuent des transformations complexes sur les données. Chaque neurone dans une couche cachée calcule une somme pondérée de ses entrées et applique une fonction d'activation pour produire une sortie.

- **Couche de Sortie** : La couche finale produit la prédiction ou la sortie du réseau basée sur les données traitées.

Au cœur de chaque réseau de neurones se trouvent des neurones artificiels, également connus sous le nom de nœuds ou perceptrons. Ces neurones simulent le comportement des neurones biologiques en recevant des entrées, en appliquant des opérations mathématiques et en transmettant des signaux à la couche suivante.

Les fonctions d'activation introduisent la non-linéarité dans le réseau de neurones, lui permettant de modéliser des relations complexes au sein des données.

Types de Réseaux de Neurones

Les modèles de deep learning sont capables d'apprendre automatiquement des caractéristiques à partir des données, ce qui les rend bien adaptés pour des tâches telles que la reconnaissance d'images, la reconnaissance vocale et le traitement du langage naturel (NLP). Les architectures les plus largement utilisées en deep learning sont les réseaux de neurones convolutifs (CNNs) et les réseaux de neurones récurrents (RNNs).

- **Réseaux de Neurones Convolutifs (CNN)** : Les réseaux neuronaux sont spécialement développés pour les opérations de reconnaissance d'images et de vidéos. Les couches de convolution sont employées pour extraire automatiquement des caractéristiques des images, ce qui les rend particulièrement adaptés à des tâches comme la classification d'images, la détection d'objets et la segmentation d'images.

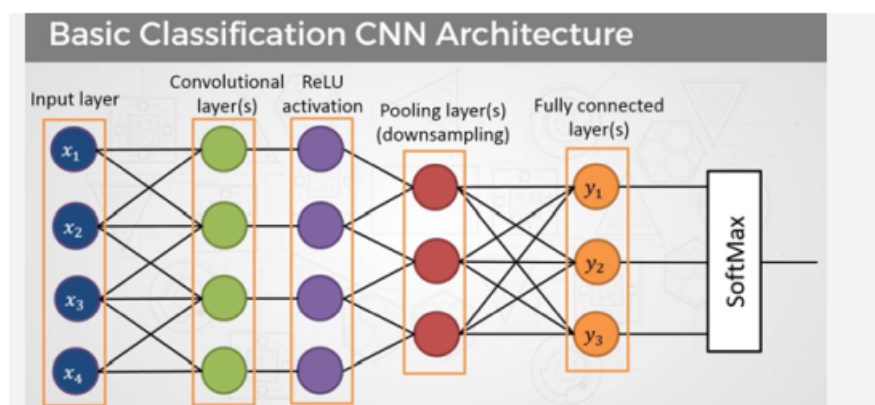


FIGURE 2.6 – Structure des réseaux de neurones Convolutifs (Source : [4])

- **Réseaux de Neurones Récurrents (RNN)** : Les RNN sont spécialement développés pour gérer les données séquentielles en prenant en considération l'ordre temporel des données. À la différence des réseaux de neurones classiques, les RNN ont des liens récurrents qui permettent de stocker les informations antérieures dans la séquence. Cela les rend très performants pour des missions comme la reconnaissance vocale, le traitement du langage naturel et la traduction de langues.

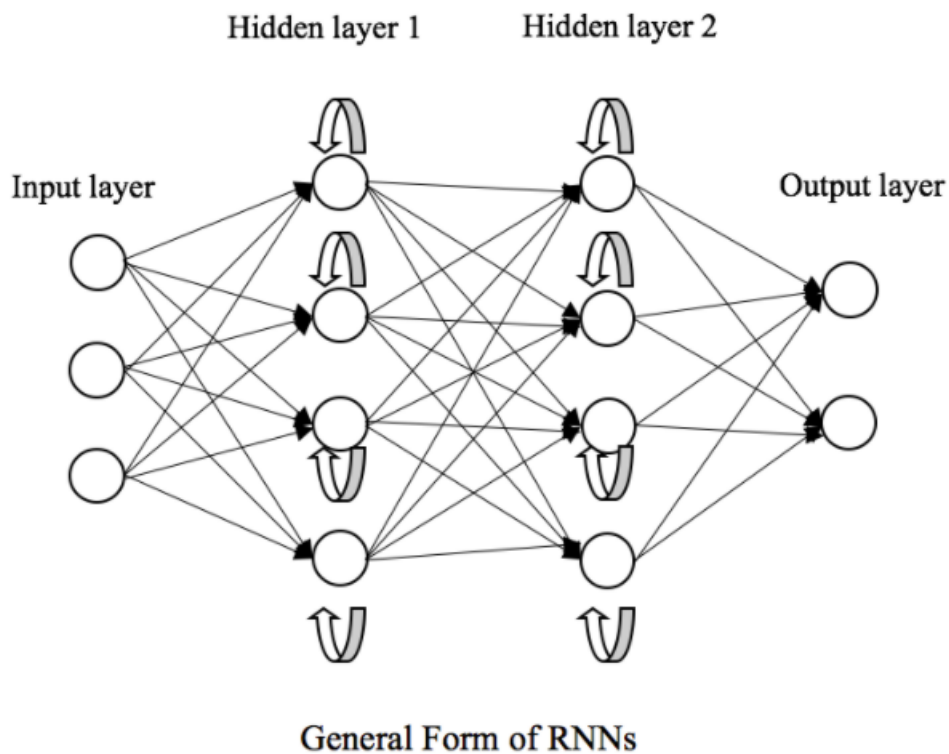


FIGURE 2.7 – Structure des Réseaux de Neurones Récurents (Source : [5])

- **Réseaux de Neurones Profonds (DNN)** : Les DNN sont une variante des réseaux neuronaux classiques qui comportent de multiples couches dissimulées, offrant ainsi la possibilité de capturer des relations encore plus complexes dans les données. Ils sont particulièrement bénéfiques pour acquérir une compréhension plus approfondie des hiérarchies de caractéristiques.

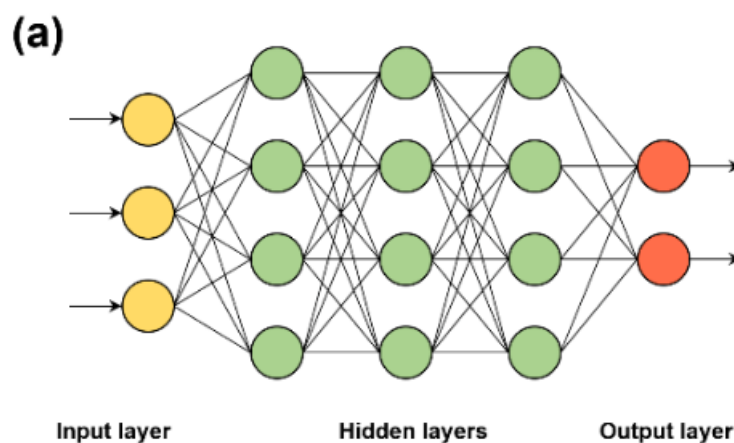


FIGURE 2.8 – Structure des réseaux de neurones profonds (Source : [6])

2.4 Large Language Models (LLM) : Fondements Techniques et Implications

Les Large Language Models (LLM) représentent une avancée significative dans le domaine de l'intelligence artificielle, marquant une évolution cruciale dans les capacités de compréhension et de génération du langage naturel. Ces modèles exploitent des architectures neuronales profondes pour traiter de vastes quantités de données textuelles, permettant ainsi une compréhension linguistique et une génération de texte d'une précision inégalée.

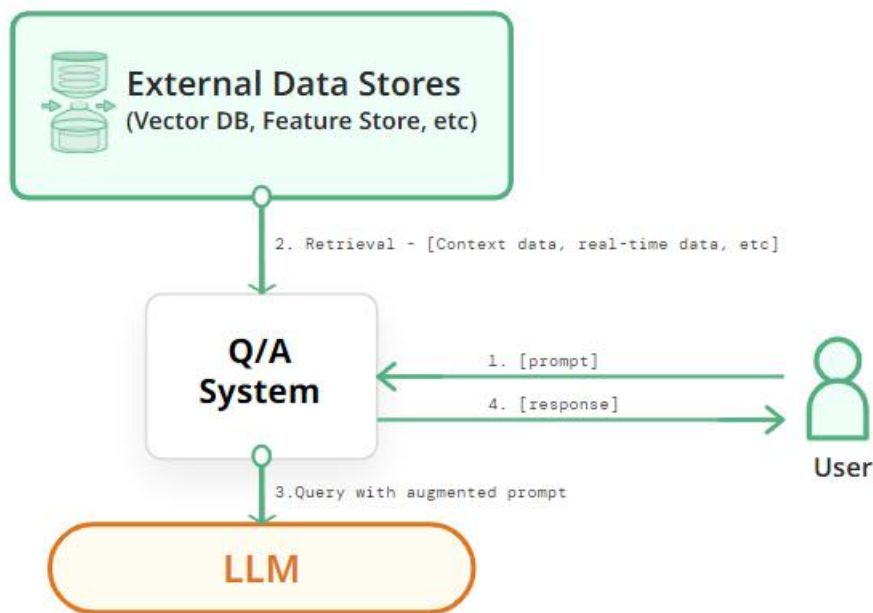


FIGURE 2.9 – Fonctionnement des LLM (Source : [?])

2.4.1 Architecture et Mécanismes des LLM

Les LLM sont principalement basés sur l'architecture Transformer, qui utilise des mécanismes d'attention multi-têtes pour moduler le degré d'attention accordé à chaque partie d'un texte lors du traitement de l'information. Cette capacité à ajuster la focalisation sur les segments pertinents d'un texte permet aux LLM de gérer des dépendances longue distance et de comprendre les subtilités contextuelles du langage. Les modèles utilisent des couches successives de ces Transformeurs pour construire des représentations de plus en plus sophistiquées du texte, facilitant ainsi une large gamme de tâches de NLP.

2.4.2 Processus d'Entraînement des LLM

Le développement des LLM se déroule en deux phases principales : le pré-entraînement et le fine-tuning. Le pré-entraînement est généralement effectué sur des corpus de données généralistes de grande échelle, où le modèle apprend des structures linguistiques fondamentales à travers des tâches telles que la prédiction de mots manquants ou la compréhension de la relation entre des paires de

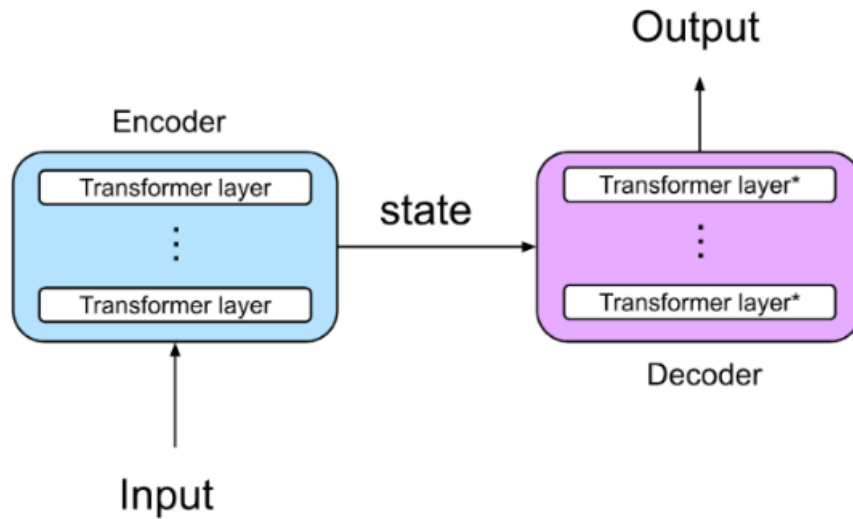


FIGURE 2.10 – Fonctionnement des transformers (Source : [7])

phrases. Cette phase exploite des techniques d'apprentissage non supervisé ou semi-supervisé pour inférer des patterns linguistiques complexes.

Fine-Tuning des Large Language Models

Après la phase de pré-entraînement, les LLM subissent une phase de fine-tuning qui est essentielle pour les adapter à des tâches ou des contextes spécifiques. Cette étape permet de personnaliser le modèle généraliste appris pendant le pré-entraînement pour qu'il réponde précisément aux exigences particulières d'une application donnée.

Processus de Fine-Tuning : Le fine-tuning implique l'ajustement des poids du modèle pré-entraîné en utilisant un ensemble de données plus petit et spécifique à la tâche. Par exemple, pour un chatbot destiné à la rénovation énergétique, le modèle serait fine-tuné avec des dialogues et des textes liés à l'énergie, la construction et la réglementation environnementale. Durant cette phase, les paramètres du modèle sont subtilement ajustés pour minimiser une fonction de perte spécifique à la tâche, ce qui affine la capacité du modèle à générer des réponses pertinentes et contextuellement appropriées.

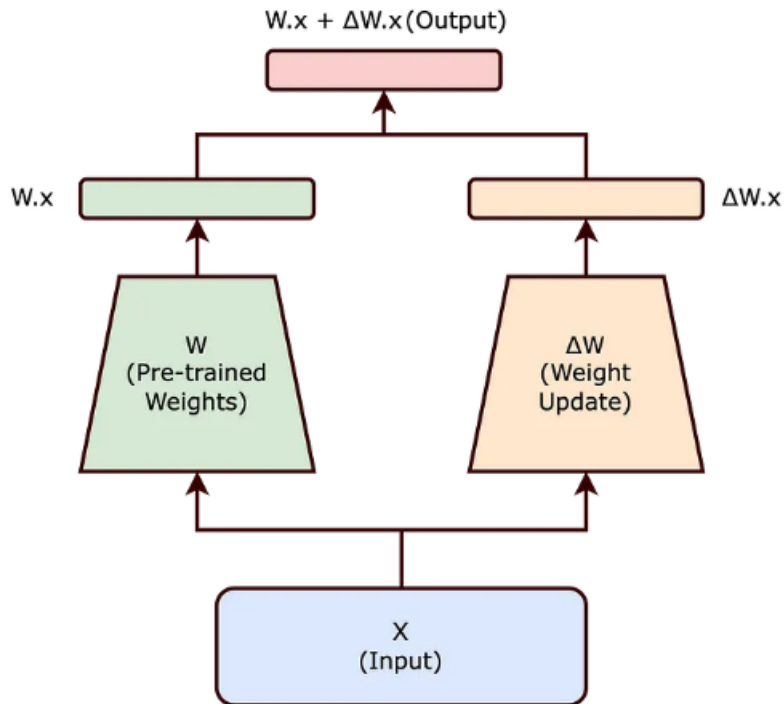


FIGURE 2.11 – Principe de Fine-tuning

Techniques de Fine-Tuning : Les techniques courantes incluent l'apprentissage continu, où le modèle continue d'apprendre à partir de nouvelles données tout en retenant les connaissances acquises précédemment. Le fine-tuning peut également impliquer des techniques spécifiques telles que le transfert de style, où le modèle apprend à imiter un certain style de communication, ou l'adaptation de domaine, où le modèle apprend à fonctionner dans un domaine différent de celui utilisé lors du pré-entraînement.

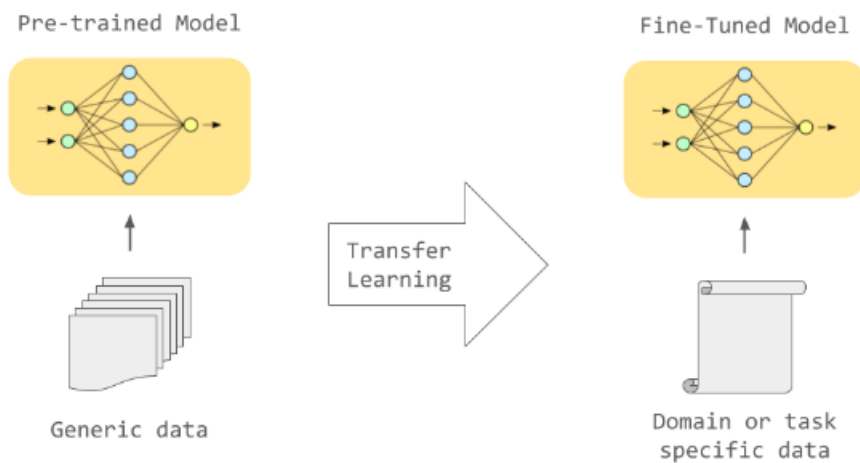


FIGURE 2.12 – Technique et performances de fine-tuning (Source : [8])

Importance du Fine-Tuning : Cette étape est cruciale car elle permet aux LLM de transcender leur formation initiale généraliste pour exceller dans des niches spécifiques. Sans fine-tuning, les réponses fournies par les modèles pourraient manquer de la précision et de la spécificité nécessaires pour des applications pratiques comme les chatbots spécialisés.

Défis du Fine-Tuning : Bien que très efficace, le fine-tuning présente des défis tels que le surajustement, surtout lorsque les données de fine-tuning sont limitées. De plus, la gestion de la divergence entre les données de pré-entraînement et de fine-tuning peut être complexe. Des techniques comme la régularisation, l'augmentation des données, ou l'emploi de méthodes d'apprentissage peu supervisées sont souvent utilisées pour contrer ces problèmes.

2.4.3 Impact du Fine-Tuning sur les Performances des LLM

En ajustant finement les LLM pour des applications spécifiques, les développeurs peuvent grandement améliorer la pertinence et l'efficacité des interactions entre l'humain et la machine. Pour le chatbot de rénovation énergétique, cela signifie une capacité accrue à comprendre des termes techniques spécifiques au domaine et à fournir des conseils personnalisés basés sur des situations individuelles.

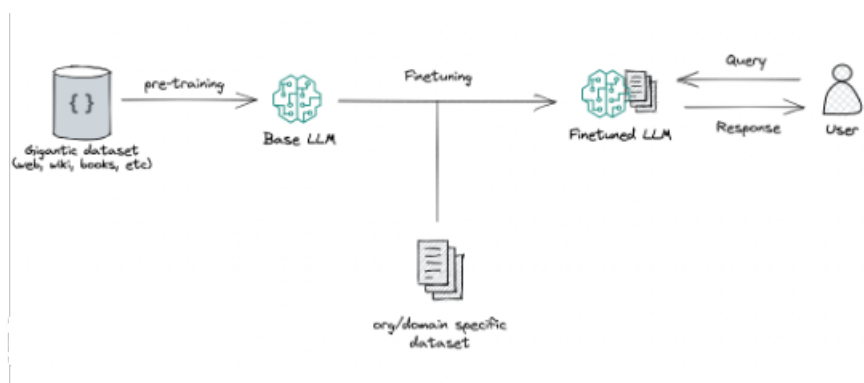


FIGURE 2.13 – technique de Fine-Tuning (Source : [9])

2.4.4 Applications Stratégiques des LLM

Les LLM trouvent leur application dans une multitude de domaines, augmentant significativement l'efficacité des systèmes de reconnaissance vocale, de traduction automatique, de génération de texte et de compréhension de documents. En incorporant des capacités de deep learning avancées, ces modèles facilitent non seulement la génération de réponses naturelles et contextuellement adaptées mais aussi la création de résumés, la modération de contenu, et l'amélioration des interfaces utilisateur conversationnelles.

2.4.5 Enjeux et Perspectives

Cependant, l'utilisation des LLM soulève des questions éthiques importantes, notamment en ce qui concerne le risque de biais algorithmique, la consommation énergétique pour l'entraînement et l'exploitation des modèles, et les implications de la création de contenu synthétique. La transparence des

processus d'apprentissage et la mise en place de garde-fous éthiques sont essentielles pour assurer une utilisation responsable de ces technologies puissantes.

2.5 Modèles de Traitement du Langage Naturel (NLP)

Le deep learning, une branche avancée de l'apprentissage automatique, joue un rôle crucial dans le développement des modèles de traitement du langage naturel tels que LLAMA 2, LLAMA 3 et Ggema 2. Ces modèles, basés sur des architectures de deep learning, sont spécialement conçus pour analyser les séquences de texte et produire des réponses qui sont non seulement pertinentes mais aussi adaptées au contexte. Utilisant des techniques avancées telles que les réseaux de neurones récurrents (RNN), ils sont capables de capturer les nuances contextuelles et temporelles dans les données textuelles, ce qui est essentiel pour comprendre les requêtes complexes des utilisateurs et y répondre avec précision.

2.5.1 LLAMA2

LLAMA2 est un modèle de langage de grande taille (LLM) open source. Il utilise des technologies avancées de traitement du langage naturel (NLP) et de deep learning pour comprendre et générer du texte de manière précise. Son architecture, basée sur des réseaux de neurones, le rend performant et flexible, facilitant ainsi la création d'applications conversationnelles comme les chatbots.

Ce modèle propose une plateforme solide pour concevoir des chatbots et des assistants virtuels. Grâce à son infrastructure avancée, LLAMA2 peut gérer des demandes complexes et fournir des réponses cohérentes et pertinentes. En étant open source, il permet une personnalisation et une intégration faciles avec différents systèmes. Il peut également être ajusté pour répondre précisément aux besoins des utilisateurs, ce qui le rend efficace pour les interactions contextuelles et les réponses précises.

LLAMA2 est utilisé pour créer des chatbots capables de gérer des dialogues complexes dans des domaines comme le service client, le commerce en ligne, l'éducation et la santé. Ses fonctionnalités avancées permettent de concevoir des expériences utilisateur riches et dynamiques, augmentant ainsi l'engagement et la satisfaction des utilisateurs. De plus, il offre des outils d'analyse pour évaluer et améliorer les performances des chatbots en continu.



FIGURE 2.14 – Logo de LLAMA2

Détails Techniques de l'Architecture de LLAMA2

LLAMA2 utilise une structure hiérarchique d'attention pour mieux comprendre les dépendances longues dans les textes. Cela permet au modèle de saisir le contexte global et local.

Les mécanismes de codage contextuel utilisés par LLAMA2 représentent les mots en fonction de leur contexte, augmentant ainsi la précision des réponses. Pour entraîner LLAMA2, on utilise une fonction de perte avec régularisation pour éviter le surapprentissage. La régularisation est souvent effectuée avec une pénalité L2 :

$$\text{Loss} = - \sum_{i=1}^N y_i \log(\hat{y}_i) + \lambda \sum_{j=1}^M \theta_j^2 \quad (2.1)$$

où les termes sont définis comme suit :

- $-\sum_{i=1}^N y_i \log(\hat{y}_i)$ représente la **perte d'entropie croisée** ou **log loss**. Cette composante mesure la différence entre les étiquettes réelles y_i et les prédictions \hat{y}_i , et est utilisée pour évaluer à quel point le modèle prédit correctement la classe réelle.
- N est le nombre total d'exemples dans le jeu de données.
- y_i est la valeur réelle de l'étiquette pour l'exemple i , souvent représentée sous forme de vecteur one-hot en classification.
- \hat{y}_i est la probabilité prédite que l'exemple i appartienne à une classe spécifique, typiquement obtenue via un modèle tel qu'un réseau de neurones.
- $\log(\hat{y}_i)$ est le logarithme naturel de la probabilité prédite, qui punit les prédictions incorrectes, particulièrement celles loin de la vraie classe.

La régularisation est représentée par le terme :

$$\lambda \sum_{j=1}^M \theta_j^2$$

- λ est un paramètre qui ajuste l'importance de la régularisation dans la fonction de perte, aidant à prévenir le surapprentissage en favorisant des poids plus petits dans le modèle.
- M est le nombre total de paramètres dans le modèle.
- θ_j est le poids ou paramètre j du modèle.
- θ_j^2 est le carré du poids, contribuant à pénaliser les grands poids et à favoriser un modèle plus simple et plus généralisable.

Pour l'optimisation, LLAMA2 utilise l'algorithme AdamW, qui combine les avantages d'Adam et de la régularisation L2 pour améliorer la convergence :

$$\theta_{t+1} = \theta_t - \eta \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} + \lambda \theta_t \right) \quad (2.2)$$

- θ_t et θ_{t+1} :
 - θ_t représente les paramètres du modèle (par exemple, les poids d'un réseau de neurones) à l'itération t .
 - θ_{t+1} est la mise à jour de ces paramètres à l'itération suivante, $t + 1$.
- η :
 - η est le taux d'apprentissage, un hyperparamètre qui contrôle la taille des pas faits dans la direction opposée au gradient lors de la mise à jour des paramètres. Un η plus élevé accélère l'apprentissage mais peut conduire à un dépassement ou à une instabilité.

- \hat{m}_t et \hat{v}_t :
 - Ces termes sont typiques des optimiseurs comme Adam, où \hat{m}_t est une estimation biaisée du premier moment (la moyenne) du gradient, et \hat{v}_t est une estimation biaisée du second moment (la variance non centrée) du gradient.
 - Ces estimations permettent d'ajuster la taille du pas pour chaque paramètre de manière individuelle, ce qui peut conduire à une convergence plus rapide et plus stable.
- $\sqrt{\hat{v}_t} + \epsilon$:
 - Cette expression sert à normaliser l'ajustement du taux d'apprentissage, où ϵ est un petit nombre ajouté pour éviter la division par zéro (souvent appelé terme de stabilisation).
- $\lambda\theta_t$:
 - Ce terme représente la régularisation L2 (aussi appelée régularisation de Tikhonov ou ridge regularization), où λ est un coefficient qui contrôle l'intensité de la régularisation. Ce terme punit les valeurs élevées des poids en ajoutant un coût proportionnel au carré des poids au coût total, aidant à prévenir le surapprentissage.

Les principaux hyperparamètres de LLAMA2 incluent la taille des embeddings, le nombre de couches d'attention hiérarchique, le taux de régularisation et le taux d'apprentissage. Ces hyperparamètres doivent être soigneusement ajustés pour maximiser les performances du modèle.

2.5.2 LLAMA3

La communauté d'experts en intelligence artificielle a développé LLAMA3, un modèle de langage de grande taille (LLM) open source. Il se sert de méthodes sophistiquées de traitement du langage naturel (NLP) et d'apprentissage profond afin de saisir et produire du texte de manière avancée. Créé dans le but d'être souple et adaptable, ce modèle offre aux développeurs la possibilité de concevoir des applications conversationnelles sur mesure avec une compréhension approfondie du contexte. LLAMA3 utilise des réseaux de neurones avancés pour gérer de manière efficace des tâches de traitement du langage à grande échelle.

LLAMA3 propose une plateforme solide pour concevoir des chatbots et des assistants virtuels. Grâce à sa structure reposant sur des réseaux de neurones profonds, il est en mesure de saisir des demandes complexes et de fournir des réponses adaptées. Grâce à sa nature open source, LLAMA3 offre une grande personnalisation, permettant aux développeurs d'ajuster le modèle en fonction de leurs besoins spécifiques. De plus, sa facilité d'intégration avec différentes plateformes et canaux de communication permet de mettre en place des solutions basées sur la conversation sur une variété d'interfaces utilisateur. Il est excellent pour des échanges contextuels avancés et adaptatifs dans différents domaines, permettant une gestion améliorée des conversations complexes.

La technologie LLAMA3 est employée afin de créer des chatbots capables de gérer des échanges complexes dans divers secteurs tels que le service client, le commerce en ligne, l'éducation et les services financiers. Grâce à ses compétences avancées, il est possible de concevoir des expériences utilisateur captivantes et performantes, ce qui améliore l'interaction et la satisfaction des gens. En outre, la plateforme propose des outils efficaces pour analyser et améliorer constamment les performances des chatbots, assurant ainsi que les solutions demeurent pertinentes et efficaces face aux changements des besoins et des attentes des utilisateurs.



FIGURE 2.15 – Logo de LLAMA3

Détails Techniques de l'Architecture de LLAMA3

LLAMA3 est basé sur une structure de réseaux de neurones profonds. Cette structure comprend différentes couches de neurones artificiels, où chaque neurone est une unité de calcul qui effectue une transformation non-linéaire sur les données.

LLAMA3 adopte une architecture de transformateur, ce qui représente une avancée significative dans le domaine des modèles linguistiques. Le transformateur est constitué de blocs de codage (encoder) et de décodage (decoder), chacun ayant des mécanismes d'attention pour gérer les séquences d'entrée et de sortie.

Les mécanismes de codage contextuel utilisés par LLAMA3 représentent les mots en fonction de leur contexte, augmentant ainsi la précision des réponses produites. Pour entraîner le modèle, une fonction de perte avec régularisation est utilisée pour éviter le surapprentissage. La régularisation est souvent effectuée avec une pénalité L2 :

$$\text{Loss} = - \sum_{i=1}^N y_i \log(\hat{y}_i) + \lambda \sum_{j=1}^M \theta_j^2 \quad (2.3)$$

LLAMA3 utilise des encodages positionnels pour introduire des informations sur la position des mots dans les séquences d'entrée, car les transformateurs ne disposent pas de mécanisme intégré pour gérer la séquence des données.

Pour optimiser le modèle, des algorithmes comme Adam (Adaptive Moment Estimation) sont utilisés, ce qui permet d'ajuster efficacement les poids du réseau.

Les principaux hyperparamètres dans l'architecture de LLAMA3 incluent la taille des embeddings, le nombre de têtes d'attention, la profondeur du réseau (nombre de couches de codage et de décodage), et le taux d'apprentissage. Ces hyperparamètres doivent être soigneusement choisis et ajustés pour optimiser les performances du modèle.

2.5.3 GEMMA

Gemma est une plateforme open source de langage de grande taille (LLM) créée par Google. Elle utilise des méthodes sophistiquées de traitement du langage naturel (NLP) et d'apprentissage profond pour comprendre et produire du texte de manière avancée. Conçu pour être souple et adaptable, Gemma offre aux développeurs la possibilité de créer des applications conversationnelles sur mesure avec une compréhension approfondie du contexte. Les réseaux de neurones avancés utilisés dans

l'architecture de Gemma permettent de traiter efficacement des tâches linguistiques complexes à grande échelle.

Gemma propose une plateforme solide pour concevoir des chatbots et des assistants virtuels. Sa structure repose sur des réseaux de neurones profonds, ce qui lui permet de comprendre des demandes complexes et de fournir des réponses pertinentes. Grâce à sa nature open source, Gemma offre une grande personnalisation, permettant aux développeurs d'ajuster le modèle en fonction de leurs besoins particuliers. De plus, sa facilité d'intégration avec différentes plateformes et canaux de communication permet de mettre en place des solutions conversationnelles sur une variété d'interfaces utilisateur, ce qui améliore l'accessibilité et l'efficacité des sessions.

Gemma est utilisée pour créer des chatbots capables de gérer des échanges complexes dans divers secteurs tels que le service client, le commerce en ligne, l'éducation et les services financiers. Elle possède les compétences nécessaires pour concevoir des expériences utilisateur captivantes et performantes, augmentant ainsi l'interaction et la satisfaction des utilisateurs. La plateforme offre également des outils sophistiqués pour analyser et améliorer constamment les performances des chatbots, garantissant ainsi que les solutions restent pertinentes et adaptables aux changements des besoins et des attentes des utilisateurs.



FIGURE 2.16 – Logo de Gemma de Google

Détails Techniques de l'Architecture de Gemma

L'architecture de Gemma est basée sur des réseaux de neurones profonds (DNN). La structure d'un DNN est composée de différentes couches de neurones artificiels, où chaque neurone est une unité de calcul qui effectue une transformation non-linéaire sur les données. Les couches principales d'un DNN typique incluent : *Input Layer*, *Hidden Layers*, et *Output Layer*.

Gemma adopte une architecture de transformateur, ce qui représente une avancée significative dans le domaine des modèles linguistiques. Le transformateur est constitué de blocs de codage (encoder) et de décodage (decoder), chacun ayant des mécanismes d'attention pour gérer les séquences d'entrée et de sortie. Les formules de l'attention dans un transformateur incluent l'attention multi-tête (multi-head attention), qui permet de modéliser les relations complexes entre les mots dans une phrase.

En raison du manque de mécanisme intégré pour gérer la séquence des données dans les transformateurs, Gemma utilise des encodages positionnels pour introduire des informations sur la position des mots dans les séquences d'entrée :

$$PE_{\text{pos},2i} = \sin\left(\frac{\text{pos}}{10000^{2i/d_{\text{model}}}}\right)$$

$$PE_{\text{pos},2i+1} = \cos\left(\frac{\text{pos}}{10000^{2i/d_{\text{model}}}}\right)$$

Les termes des équations des embeddings de position sont expliqués ci-dessous :

- $PE_{\text{pos},2i}$ et $PE_{\text{pos},2i+1}$:
 - Ces termes représentent les embeddings de position pour une position donnée dans la séquence (pos) à l'indice $2i$ pour le sinus et $2i + 1$ pour le cosinus.
 - Ils sont utilisés pour encoder la position des mots ou des tokens dans l'entrée du modèle.
- pos :
 - pos représente la position du mot ou du token dans la séquence. Par exemple, dans une phrase, pos pourrait être l'indice du mot dans cette phrase.
- i :
 - i est l'indice au sein d'un vecteur d'embedding de dimension d_{model} . Chaque dimension de l'embedding de position est associée à une paire sinus-cosinus qui varie selon l'indice i .
 - Pour une dimension $2i$, la fonction sinus est utilisée, tandis que pour $2i + 1$, la fonction cosinus est utilisée.
- d_{model} :
 - d_{model} est la dimensionnalité des embeddings dans le modèle. C'est un hyperparamètre du modèle qui définit la taille des embeddings de mots et de positions.
- $10000^{2i/d_{\text{model}}}$:
 - Ce terme dans le dénominateur est utilisé pour créer des fréquences qui varient logarithmiquement d'une dimension à l'autre.
 - L'utilisation de $10000^{2i/d_{\text{model}}}$ permet aux embeddings de position de capturer des informations à différentes échelles temporelles, ce qui est particulièrement utile pour gérer des phrases de différentes longueurs et pour permettre au modèle de mieux généraliser à des positions jamais vues lors de la formation.

Pour entraîner le modèle Gemma, une fonction de perte appropriée est utilisée, généralement la perte d'entropie croisée (Cross-Entropy Loss) pour les tâches de classification et de génération de texte. L'optimisation du modèle est réalisée à l'aide d'algorithmes tels que Adam (Adaptive Moment Estimation), qui ajuste les poids du réseau de manière efficace.

Les hyperparamètres importants dans l'architecture de Gemma incluent la taille des embeddings (d_{model}), le nombre de têtes d'attention (h), la profondeur du réseau (nombre de couches de codage et de décodage), et le taux d'apprentissage (η). Ces hyperparamètres doivent être soigneusement choisis et ajustés pour optimiser les performances du modèle.

Détails des Architectures des Modèles

Ce graphe représente l'architecture des modèles Open-Source de Meta LLAMA3 et LLAMA2 .

LLaMA Achitecture

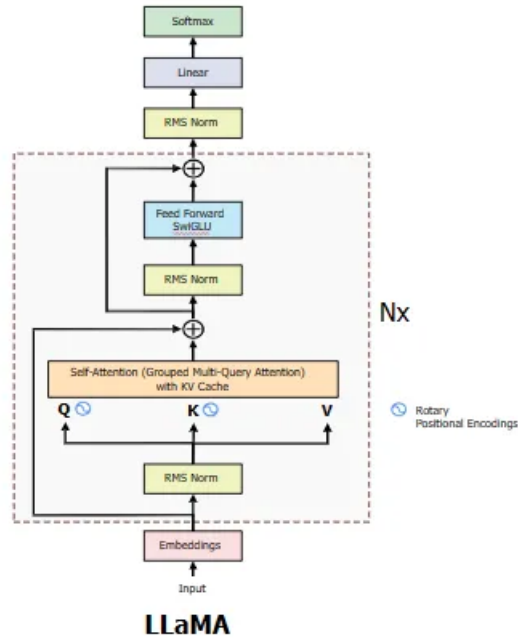


FIGURE 2.17 – Architecture LLAMA. (Source : [10])

Le tableau suivant présente les architectures utilisées dans le projet avec leurs paramètres clés :

Modèle	Paramètres	Architecture
Meta-Llama/Llama-2-7b	7 milliards	Transformer
Meta-Llama/Meta-Llama-3-8B	8 milliards	Advanced Transformer
Google/Gemma-2b	2 milliards	Optimized Transformer

TABLE 2.1 – Comparaison des modèles utilisés dans le projet

Explication du choix des modèles

Dans le cadre du développement du chatbot "Mon Carnet de Logement" destiné à l'assistance dans les projets de rénovation énergétique, une évaluation approfondie des modèles de traitement automatique du langage naturel a été réalisée pour identifier le modèle le plus adapté. Trois principaux candidats, Meta-Llama-3-8B, Meta-Llama-2-7b, et google/gemma-2b, ont été comparés en fonction de leur capacité à traiter des requêtes techniques spécifiques au domaine de la rénovation énergétique. Meta-Llama-3-8B se distingue par sa capacité à comprendre et générer des réponses contextuellement précises dans des discussions techniques complexes, grâce à son architecture avancée de 8 milliards de paramètres. Meta-Llama-2-7b offre un bon compromis entre performance et coût, adapté pour un large éventail de tâches linguistiques. Enfin, Gemma de Google, grâce à ses technologies propriétaires, présente une grande adaptabilité, essentielle pour l'intégration de fonctionnalités sur mesure. Le choix

final du modèle sera basé sur une combinaison de performances techniques, de coût d'opération, et de facilité d'intégration dans l'infrastructure existante.

2.5.4 Limites de Colab pour LLAMA et Gemma

En tenant compte des modèles LLAMA 2, LLAMA 3 et Gemma dans Google Colab, les limitations de la plateforme peuvent affecter leur utilisation en raison des ressources GPU disponibles :

- LLAMA 2 : Les versions de LLAMA 2 varient de 2 milliards à 70 milliards de paramètres. Sur Google Colab, la mémoire GPU limitée à environ 12 GB en version gratuite peut restreindre l'utilisation efficace des versions supérieures à 7B sans accès Pro.
- LLAMA 3 : Similaire à LLAMA 2, avec des versions allant jusqu'à des dizaines de milliards de paramètres, LLAMA 3 nécessiterait également une allocation de ressources bien supérieure à celle disponible sur la version gratuite de Colab, limitant son exploitation à des modèles de taille réduite.
- Gemma : Bien que les spécificités des versions de Gemma ne soient pas explicitement mentionnées, les modèles de grande taille similaires nécessitent généralement des configurations avancées qui dépassent les capacités de Colab sans un abonnement Pro.

Ces modèles, surtout dans leurs configurations les plus grandes, nécessitent souvent des ajustements ou l'accès à des ressources matérielles plus robustes pour fonctionner sans contraintes sur des plateformes comme Google Colab.

2.6 Métriques d'Évaluation

Afin d'évaluer les résultats de notre chatbot, On applique diverses mesures d'évaluation telles que le chevauchement, la similarité sémantique, la précision, ainsi que des mesures personnalisées. Chaque mesure présente une vision singulière de la qualité des réponses produites par le chatbot par rapport aux réponses de référence.

2.6.1 Métriques Basées sur le Chevauchement

BLEU (Bilingual Evaluation Understudy)

La métrique BLEU (Bilingual Evaluation Understudy) est largement utilisée pour évaluer la précision des n-grammes dans les réponses générées par les systèmes de traitement automatique du langage, comme ceux utilisés en traduction automatique. BLEU compare les segments de texte générés par la machine avec des segments de texte de référence préétablis pour déterminer la qualité de la génération de texte. Cette approche permet d'évaluer à quel point les séquences de mots créées par une machine correspondent aux séquences de référence, offrant ainsi une mesure quantitative de la fidélité de la traduction ou de la génération de texte par rapport à l'original humain.

L'application de BLEU est particulièrement pertinente dans le domaine de la traduction automatique, où elle sert à mesurer la proximité entre le texte traduit automatiquement et les traductions humaines de référence. Cette métrique est cruciale pour les développeurs de modèles linguistiques car elle leur

fournit un retour essentiel sur la performance de leurs systèmes, leur permettant d'apporter des améliorations ciblées.

En termes techniques, BLEU évalue la qualité des traductions en calculant le score des coïncidences de n-grammes entre les traductions automatiques et les traductions de référence. La formule générale pour le calcul de BLEU est la suivante :

$$\text{BLEU} = BP \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right)$$

où p_n est la précision des n-grammes, w_n est le poids attribué à chaque n-gramme (généralement équitablement réparti), et N est la longueur du n-gramme utilisé (souvent jusqu'à 4). Le facteur de pénalité pour la brièveté (BP) est utilisé pour pénaliser les traductions qui sont trop courtes comparées aux références :

$$BP = \begin{cases} 1 & \text{si } c > r \\ \exp(1 - \frac{r}{c}) & \text{si } c \leq r \end{cases}$$

où c est la longueur de la traduction candidate et r est la longueur de la référence. Ces formules permettent de s'assurer que les traductions ne sont pas seulement précises au niveau des mots, mais aussi appropriées en termes de longueur et de contexte.

Calcul du Score BLEU

Voici un exemple de calcul du score BLEU pour un exemple de texte.

Texte de référence : "La chatte est sur le tapis."

Texte généré : "La chatte dort sur le tapis."

Étapes de Calcul

1. *Tokenisation* :

- Texte de référence : ["La", "chatte", "est", "sur", "le", "tapis"]
- Texte généré : ["La", "chatte", "dort", "sur", "le", "tapis"]

2. *Calcul des n-grams* :

- 1-grams (unigrams) :
 - Référence : ["La", "chatte", "est", "sur", "le", "tapis"]
 - Hypothèse : ["La", "chatte", "dort", "sur", "le", "tapis"]
 - Correspondances : ["La", "chatte", "sur", "le", "tapis"] (5 correspondances)
- 2-grams (bigrams) :
 - Référence : [("La", "chatte"), ("chatte", "est"), ("est", "sur"), ("sur", "le"), ("le", "tapis")]
 - Hypothèse : [("La", "chatte"), ("chatte", "dort"), ("dort", "sur"), ("sur", "le"), ("le", "tapis")]
 - Correspondances : [("La", "chatte"), ("sur", "le"), ("le", "tapis")] (3 correspondances)
- 3-grams (trigrams) :

- Référence : [("La", "chatte", "est"), ("chatte", "est", "sur"), ("est", "sur", "le"), ("sur", "le", "tapis")]
- Hypothèse : [("La", "chatte", "dort"), ("chatte", "dort", "sur"), ("dort", "sur", "le"), ("sur", "le", "tapis")]
- Correspondances : [("sur", "le", "tapis")] (1 correspondance)
- 4-grams (quadrigrams) :
 - Référence : [("La", "chatte", "est", "sur"), ("chatte", "est", "sur", "le"), ("est", "sur", "le", "tapis")]
 - Hypothèse : [("La", "chatte", "dort", "sur"), ("chatte", "dort", "sur", "le"), ("dort", "sur", "le", "tapis")]
 - Correspondances : [] (0 correspondance)

3. *Calcul de la Précision des n-grams :*

- 1-grams (unigrams) :

$$\text{Précision}_{1\text{-grams}} = \frac{5}{6} = 0.83$$

- 2-grams (bigrams) :

$$\text{Précision}_{2\text{-grams}} = \frac{3}{5} = 0.60$$

- 3-grams (trigrams) :

$$\text{Précision}_{3\text{-grams}} = \frac{1}{4} = 0.25$$

- 4-grams (quadrigrams) :

$$\text{Précision}_{4\text{-grams}} = \frac{0}{3} = 0.00$$

4. *Calcul du Score BLEU avec Pénalité de Longueur :*

- Pénalité de Longueur (BP) :

$$BP = e^{(1 - \frac{6}{6})} = 1$$

- Score BLEU :

$$\text{Score BLEU} = BP \cdot \exp \left(\sum_{n=1}^4 w_n \log p_n \right)$$

Où w_n est le poids de chaque n-gram (généralement $w_n = \frac{1}{4}$) et p_n est la précision des n-grams :

$$\text{Score BLEU} = 1 \cdot \exp \left(\frac{1}{4} \log 0.83 + \frac{1}{4} \log 0.60 + \frac{1}{4} \log 0.25 + \frac{1}{4} \log 0.00 \right)$$

$$\text{Score BLEU} = \exp \left(\frac{1}{4} (-0.18 - 0.51 - 1.39 - \infty) \right) = 0$$

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

La métrique ROUGE (Recall-Oriented Understudy for Gisting Evaluation) est utilisée principalement pour évaluer la qualité des résumés automatiques et d'autres formes de génération de texte en mesurant le chevauchement entre le texte généré et les références préétablies. Cette métrique est fondamentale pour déterminer dans quelle mesure les informations essentielles du texte source sont capturées par le texte généré, faisant de ROUGE un outil précieux dans les domaines de la recherche en traitement automatique du langage naturel et des technologies de l'information.

ROUGE évalue ce chevauchement en utilisant plusieurs sous-métriques, chacune ciblant différents aspects de la similarité textuelle :

1. ROUGE-N : Cette sous-métrique calcule la précision et le rappel des chevauchements de n-grammes entre le texte généré et les textes de référence. Elle permet d'évaluer combien de n-grammes dans le texte généré apparaissent également dans le texte de référence, offrant ainsi une mesure de la précision du contenu généré.

La formule pour ROUGE-N est la suivante :

$$\text{ROUGE-N} = \frac{\sum_{s \in \{\text{résumés de référence}\}} \sum_{gram_n \in s} \text{Min}(m(gram_n), w(gram_n))}{\sum_{s \in \{\text{résumés de référence}\}} \sum_{gram_n \in s} m(gram_n)}$$

où $m(gram_n)$ est le nombre de n-grammes dans le texte de référence et $w(gram_n)$ est le nombre de ces n-grammes dans le texte généré.

ROUGE-1

Formules de calcul :

— Précision (Precision) :

$$\text{Précision} = \frac{\text{Nombre d'unigrammes communs}}{\text{Nombre total d'unigrammes dans le texte généré}}$$

— Rappel (Recall) :

$$\text{Rappel} = \frac{\text{Nombre d'unigrammes communs}}{\text{Nombre total d'unigrammes dans le texte de référence}}$$

— Score F1 :

$$\text{Score F1} = \frac{2 \times (\text{Précision} \times \text{Rappel})}{\text{Précision} + \text{Rappel}}$$

Exemple de calcul avec ROUGE-1(N=1)

Texte de référence : *"La chatte est sur le tapis."*

Texte généré : *"La chatte dort sur le tapis."*

1. Unigrammes du texte de référence : ["La", "chatte", "est", "sur", "le", "tapis"]
2. Unigrammes du texte généré : ["La", "chatte", "dort", "sur", "le", "tapis"]

Précision (Precision) :

$$\text{Précision} = \frac{5}{6} = 0.83$$

Rappel (Recall) :

$$\text{Rappel} = \frac{5}{6} = 0.83$$

Score F1 :

$$\text{Score F1} = \frac{2 \times (0.83 \times 0.83)}{0.83 + 0.83} = 0.83$$

2. ROUGE-L : ROUGE-L mesure la plus longue sous-séquence commune (LCS) entre le texte généré et les références. Contrairement à ROUGE-N qui se concentre uniquement sur les n-grammes, ROUGE-L prend en compte la séquence la plus longue qui apparaît dans le même ordre dans les deux textes, ce qui permet d'évaluer la fluidité et la structure de la phrase.

La formule pour ROUGE-L est :

$$\text{ROUGE-L} = \frac{\sum_{s \in \{\text{résumés de référence}\}} \text{LCS}(\text{généré}, s)}{\sum_{s \in \{\text{résumés de référence}\}} \text{Longueur}(s)}$$

où $\text{LCS}(\text{généré}, s)$ est la longueur de la plus longue sous-séquence commune entre le texte généré et le résumé de référence s .

ROUGE-L

Formules de calcul :

— Précision (Precision) :

$$\text{Précision} = \frac{\text{Longueur de la LCS}}{\text{Nombre total de mots dans le texte généré}}$$

— Rappel (Recall) :

$$\text{Rappel} = \frac{\text{Longueur de la LCS}}{\text{Nombre total de mots dans le texte de référence}}$$

— Score F1 :

$$\text{Score F1} = \frac{2 \times (\text{Précision} \times \text{Rappel})}{\text{Précision} + \text{Rappel}}$$

Exemple de calcul avec ROUGE-L

1. Plus longue sous-séquence commune (LCS) :

— "*La chatte sur le tapis*" (longueur = 5)

Précision (Precision) :

$$\text{Précision} = \frac{5}{6} = 0.83$$

Rappel (Recall) :

$$= \frac{5}{6} = 0.83$$

Score F1 :

$$\text{Score F1} = \frac{2 \times (0.83 \times 0.83)}{0.83 + 0.83} = 0.83$$

En appliquant ces métriques, les développeurs peuvent évaluer non seulement la précision des informations capturées mais aussi la qualité de la structuration du texte généré. Cela est particulièrement important dans des scénarios comme la production de résumés d'articles scientifiques où il est crucial que les résumés générés par machine reflètent fidèlement les points clés des textes originaux tout en maintenant une cohérence linguistique et stylistique.

METEOR (Metric for Evaluation of Translation with Explicit ORdering)

La métrique METEOR (Metric for Evaluation of Translation with Explicit ORdering) est une méthode avancée d'évaluation utilisée principalement dans la traduction automatique et la génération de texte. Contrairement à BLEU, qui se concentre principalement sur la précision des n-grammes, elle approfondit l'analyse en intégrant des considérations de synonymie, de stemming et de paraphrase. Cela permet d'évaluer les traductions et les textes générés avec une attention particulière à la correspondance sémantique et lexicale, offrant ainsi une évaluation plus nuancée et détaillée de la qualité linguistique.

METEOR utilise des techniques de traitement du langage naturel pour comparer le texte généré avec un ensemble de références, en cherchant non seulement les correspondances exactes mais aussi les variations lexicales qui conservent le même sens. Cette approche est particulièrement utile dans des contextes où la précision lexicale directe est moins importante que la capacité à transmettre le même sens ou contenu, comme dans des systèmes de dialogue ou des applications de résumé automatique.

En pratique, le score est calculé en évaluant la qualité des correspondances entre la traduction candidate et les références à travers une combinaison de plusieurs critères :

- Exactitude : correspondances directes de mots.
- Synonymie : utilisation de synonymes pour évaluer les correspondances.
- Stemming : comparaison des racines des mots pour détecter des correspondances au niveau de la forme des mots.
- Paraphrase : évaluation de phrases ou expressions qui ont des significations équivalentes mais qui sont structurées différemment.

Le score final est calculé en tenant compte de ces différents aspects, ce qui permet de fournir une mesure plus holistique et contextuellement adaptée de la qualité linguistique. La formule générale de METEOR est la suivante :

$$\text{METEOR} = F_{\text{mean}} \cdot (1 - \text{Penalty})$$

où F_{mean} est la moyenne harmonique de la précision et du rappel, et la pénalité est appliquée pour les segments de texte mal alignés ou ayant une mauvaise structure.

Cette métrique est donc très prisée pour les applications où la fluidité et la naturalité de la langue sont cruciales. Elle aide les développeurs à affiner leurs systèmes de manière à améliorer non seulement la

fidélité des traductions ou des textes générés mais aussi leur acceptabilité et leur intelligibilité pour les utilisateurs humains.

Voici un exemple de calcul du score METEOR pour un exemple de texte.

Texte de référence :

"La chatte est sur le tapis."

Texte généré :

"La chatte dort sur le tapis."

Étapes de Calcul

1. *Tokenisation* :

- Texte de référence : ["La", "chatte", "est", "sur", "le", "tapis"]
- Texte généré : ["La", "chatte", "dort", "sur", "le", "tapis"]

2. *Correspondances Exactes (Exact Matches)* :

- Correspondances exactes : ["La", "chatte", "sur", "le", "tapis"]
- Nombre de correspondances exactes : 5

3. *Synonymes et Flexions* :

- METEOR prend en compte les synonymes et les flexions. Dans cet exemple simple, il n'y a pas de synonymes ou de flexions à considérer.

4. *Fragmentation* :

- Fragmentation (les séquences continues dans le texte généré qui correspondent à des séquences continues dans le texte de référence) :
 - Séquence continue : ["La", "chatte"]
 - Séquence continue : ["sur", "le", "tapis"]
- Nombre de fragments : 2

5. *Précision et Rappel* :

- Précision (Precision) :

$$\text{Précision} = \frac{\text{Nombre de correspondances exactes}}{\text{Nombre total de mots dans le texte généré}} = \frac{5}{6} = 0.83$$

- Rappel (Recall) :

$$\text{Rappel} = \frac{\text{Nombre de correspondances exactes}}{\text{Nombre total de mots dans le texte de référence}} = \frac{5}{6} = 0.83$$

6. *F-Score* :

- Calcul du F-Score avec un facteur de pondération pour équilibrer précision et rappel (généralement α est pris comme 0.5) :

$$F = \frac{10 \cdot \text{Précision} \cdot \text{Rappel}}{9 \cdot \text{Précision} + \text{Rappel}}$$

$$F = \frac{10 \cdot 0.83 \cdot 0.83}{9 \cdot 0.83 + 0.83} = \frac{6.89}{7.47} = 0.92$$

7. *Fragmentation Pénalité* :

— Calcul de la pénalité de fragmentation (généralement prise comme 0.5) :

$$\text{Pénalité} = 0.5 \left(\frac{\text{Nombre de fragments}}{\text{Nombre total de correspondances exactes}} \right) = 0.5 \left(\frac{2}{5} \right) = 0.2$$

8. *Score METEOR Final* :

— Calcul du score final en appliquant la pénalité de fragmentation :

$$\text{METEOR} = F \cdot (1 - \text{Pénalité}) = 0.92 \cdot (1 - 0.2) = 0.92 \cdot 0.8 = 0.736$$

2.6.2 Métriques Basées sur la Similarité Sémantique

BERTScore

BERTScore est une métrique avancée qui utilise les embeddings de BERT pour évaluer la similitude sémantique entre les réponses générées et les références. Contrairement aux approches traditionnelles basées sur les n-grammes, BERTScore se concentre sur la signification contextuelle des mots, ce qui permet une évaluation plus précise et nuancée des réponses textuelles. En calculant la similarité cosinus entre les embeddings des mots dans les textes générés et les textes de référence, BERTScore capture les nuances de signification qui sont souvent omises par les mesures de chevauchement simples.

Cette méthode s'avère particulièrement efficace dans des domaines comme la traduction automatique et la génération de contenu, où il est crucial de maintenir l'intégrité sémantique et le contexte du texte original. Par exemple, en traduction, BERTScore peut identifier avec précision si le sens d'un texte traduit reste fidèle à celui du texte source, même si les mots exacts utilisés diffèrent.

La formule de calcul de BERTScore se présente comme suit :

$$\text{BERTScore} = \frac{1}{N} \sum_{i=1}^N \max_j \cos(e_{\text{gen}_i}, e_{\text{ref}_j})$$

où e_{gen_i} et e_{ref_j} sont les embeddings BERT des mots dans la réponse générée et la réponse de référence respectivement, \cos désigne la similarité cosinus entre ces embeddings, et N est le nombre de mots dans la réponse générée. Cette formule assure que chaque mot dans la réponse générée est comparé au mot le plus similaire dans la réponse de référence, permettant ainsi une évaluation détaillée de la pertinence sémantique de la réponse.

2.6.3 Métriques Basées sur la Précision

F1 Score

Le F1 Score est une métrique d'évaluation cruciale qui combine la précision et le rappel pour fournir une évaluation complète des performances d'un système. La précision mesure le pourcentage de réponses correctes parmi celles qui ont été générées, tandis que le rappel évalue le pourcentage de réponses de référence qui ont été correctement identifiées et reproduites par le système.

Cette métrique est particulièrement utile lorsque l'équilibre entre la précision et le rappel est essentiel, comme dans les systèmes de questions-réponses et d'extraction d'informations. Le F1 Score aide à garantir que le système ne favorise pas indûment la précision au détriment du rappel ou vice versa, ce qui est vital pour des applications où saisir un maximum d'informations pertinentes est aussi important que de minimiser les erreurs.

La formule pour calculer le F1 Score est donnée par :

$$\text{F1 Score} = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

où la précision est le ratio du nombre de vrais positifs (réponses correctes générées) par rapport au nombre total de réponses générées (vrais positifs plus faux positifs), et le rappel est le ratio du nombre de vrais positifs par rapport au nombre total de réponses correctes possibles dans les données de référence (vrais positifs plus faux négatifs). Cette formule garantit une évaluation harmonieuse de la précision et du rappel, offrant ainsi une image claire de l'efficacité globale du système évalué.

2.6.4 Métriques Personnalisées

Distance de Levenshtein

La distance de Levenshtein, ou distance d'édition, mesure le nombre minimal d'opérations requises pour transformer une chaîne de caractères en une autre, où les opérations admissibles comprennent les insertions, les suppressions, et les substitutions de caractères uniques. Cette métrique est utilisée pour évaluer la similarité entre deux séquences de texte, ce qui est particulièrement utile dans les applications de correction orthographique, de reconnaissance de texte, et de comparaison générale de chaînes.

Elle est pertinente dans des domaines où l'exactitude des données textuelles est critique. Elle permet de quantifier la "distance" ou le nombre de modifications nécessaires pour passer d'un texte généré à une référence cible, fournissant ainsi une mesure objective de la similarité textuelle. Cette approche est essentielle pour évaluer la qualité des réponses générées par des systèmes de traitement automatique du langage, offrant un moyen de mesurer à quel point les réponses sont proches des réponses attendues.

La formule de la distance de Levenshtein se calcule comme suit :

$$d[i][j] = \min(d[i-1][j] + 1, d[i][j-1] + 1, d[i-1][j-1] + \text{cost}(a[i], b[j]))$$

où :

- $d[i][j]$ est la distance entre les premiers i caractères de la chaîne a et les premiers j caractères de la chaîne b .

- $\text{cost}(a[i], b[j])$ est 0 si $a[i] = b[j]$ (aucun coût pour une substitution non nécessaire), et 1 si $a[i] \neq b[j]$ (un coût pour une substitution nécessaire).

Cette formule utilise une approche itérative, remplissant une matrice de taille $(m+1) \times (n+1)$ (où m et n sont les longueurs des chaînes a et b , respectivement) pour calculer la distance d'édition totale de manière efficace.

En associant ces divers points de vue, il est possible de repérer les atouts et les lacunes de notre chatbot et de guider les améliorations à venir afin de maximiser sa performance et son utilité dans les applications de rénovation énergétique et de DPE.

2.7 Conclusion

Dans cette partie, on a introduit les notions de chatbot et AI. Ensuite, on a décortiqué les méthodes de collecte de données. Puis, on a défini les techniques de modélisation et de l'apprentissage automatique. Par la suite, on a spécifié les caractéristiques des LLM. Après cela, on a dénombré les différents modèles de LLM qu'on va utiliser et on a terminé par lister les différentes métriques sur lesquelles on va se baser pour effectuer une analyse comparative.

Chapitre 3

Collecte de Données

3.1 Collecte des liens des posts Facebook avec le Web Scraping

3.1.1 Introduction

La collecte de données est une étape cruciale pour le développement de chatbots performants. Dans ce projet, nous avons utilisé des techniques de web scraping pour extraire des informations pertinentes à partir de groupes Facebook spécialisés dans la rénovation énergétique. Le web scraping permet d'automatiser la collecte de données à grande échelle, facilitant ainsi l'analyse et l'utilisation de ces informations dans nos modèles de traitement du langage naturel (NLP).

3.1.2 Étape de la Collecte des liens

Configuration de l'Environnement de Web Scraping

Pour commencer, on a configuré notre environnement de web scraping en utilisant plusieurs bibliothèques Python, dont Selenium pour l'automatisation du navigateur et BeautifulSoup pour l'analyse du contenu HTML. On a également utilisé openpyxl pour la gestion des fichiers Excel.

On a configuré Selenium WebDriver pour contrôler un navigateur Chrome. Des options spécifiques ont été activées pour améliorer la performance du scraping, telles que la désactivation des notifications pour éviter les interruptions pendant le scraping, le mode headless pour exécuter le navigateur sans interface graphique, réduisant ainsi la consommation de ressources, et la maximisation de la fenêtre pour assurer que tous les éléments de la page sont visibles et accessibles.

Connexion et Navigation sur Facebook

On a automatisé le processus de connexion à Facebook et la navigation vers plusieurs groupes dédiés à la rénovation énergétique. Les liens des posts collectés sont stockés dans la balise href.

On commence par la connexion à Facebook : le script identifie les champs de saisie pour l'adresse e-mail et le mot de passe, puis soumet ces informations pour se connecter à un compte Facebook. On utilise des sélecteurs CSS pour localiser les champs de saisie et on envoie les données utilisateur pour l'authentification.

```
▼<a class="x1i10hf1 xjbqb8w x1ejq3ln xd10rxx x1sy0etr x17r0tee x9
72fbf xcfux6l x1qhh985 xm0m39n x9f619 x1ypdohk xt0psk2 xe8uvvx x
dj266r x1li5rnm xat24cr x1mh8g0r xexx8yu x4uap5 x18d9i69 xkhd6sd
x16tdsg8 x1hl2dhg xggy1nq x1a2a7pz x1sur9pj xkrqix3 xi81zsa x0ll
8bm" href="https://www.facebook.com/groups/2125044441216695/post
s/2436201840100_bTlGshMs6p-LxnHjnz7XuZnokmv-cnCsmeNAmqDTGK2Ukjd2
uvJEDjVZ&_tn=%2CO%2CP-R" role="link" tabindex="0">
```

FIGURE 3.1 – Extraction des liens des posts à partir de href

Ensuite, une fois connecté, le script accède directement aux pages des groupes Facebook en utilisant les URL des groupes. La navigation automatisée est effectuée avec Selenium pour ouvrir l'URL du groupe et charger le contenu de la page.

Défilement et Extraction des Liens

Pour le défilement de la page et l'extraction des liens, on suit les étapes suivantes :

On commence par le défilement automatique de la page pour charger dynamiquement le contenu. Le script effectue plusieurs défilements automatiques, permettant ainsi de charger de nouveaux posts et liens au fur et à mesure. Pour ce faire, on utilise la classe *ActionChains* de Selenium pour simuler l'action de défilement, avec des intervalles de temps définis entre les défilements pour permettre le chargement complet du contenu.

Ensuite, à chaque itération, le script extrait les liens pertinents à partir des éléments HTML de la page. On utilise *BeautifulSoup* pour analyser le contenu de la page et extraire les balises *href*. Les liens extraits sont ensuite stockés dans une liste Python pour une utilisation ultérieure.

Gestion des Liens et Sauvegarde

Le script gère les liens collectés et les enregistre dans un fichier Excel pour un suivi et une analyse futurs. Avant d'ajouter un lien à la liste, on vérifie s'il est déjà présent dans le fichier Excel pour éviter les doublons. On utilise *openpyxl* pour ouvrir et lire le fichier Excel, en parcourant les lignes pour vérifier la présence du lien.

Les liens collectés sont périodiquement ajoutés au fichier Excel, toutes les 20 itérations, afin de s'assurer que les données sont régulièrement sauvegardées. Après un nombre défini d'itérations de défilement et d'extraction, les nouveaux liens sont ajoutés au fichier Excel et les données mises à jour sont enregistrées pour conserver les nouvelles informations collectées.

améliorer la robustesse et la fiabilité du processus de scraping. Ces méthodes permettent de continuer l'exécution du script malgré les interruptions et offrent des outils efficaces pour diagnostiquer et résoudre les problèmes rencontrés.

3.2 Navigation et Extraction des Données des Posts Facebook

3.2.1 Introduction

Après avoir collecté les liens des posts Facebook, l'étape suivante consiste à naviguer vers chaque post et à extraire le texte des posts ainsi que les commentaires associés. Cette section décrit en détail les étapes suivies pour réaliser cette extraction.

3.2.2 Navigation vers les Posts et Extraction du Texte

Pour chaque lien, le script navigue vers la page du post Facebook. Une fois la page chargée, le script identifie et extrait le texte du post principal. Des sélecteurs spécifiques sont utilisés pour localiser précisément les éléments HTML contenant le texte du post. Cela permet d'assurer que le texte extrait est exact et complet.

3.2.3 Affichage et Extraction des Commentaires

Une des étapes les plus critiques est l'extraction des commentaires associés à chaque post. Les commentaires peuvent être initialement masqués ou partiellement affichés. Pour les révéler, plusieurs types de boutons doivent être cliqués automatiquement :

- **Boutons "Voir plus de commentaires"** : Ces boutons permettent de charger des commentaires supplémentaires qui ne sont pas affichés par défaut.
- **Boutons "Afficher les commentaires"** : Ces boutons révèlent les sous-commentaires ou réponses aux commentaires principaux.
- **Boutons "Voir plus"** : Ces boutons permettent d'afficher le texte complet des commentaires qui sont initialement tronqués.

Pour chaque type de bouton, le script recherche et clique sur les boutons correspondants. Ce processus est répété autant de fois que nécessaire jusqu'à ce que tous les commentaires soient visibles. Des délais sont introduits après chaque clic pour permettre le chargement complet des nouveaux commentaires ou textes révélés.

Le script vérifie également que ceux qui ont commenté sont des experts en énergie, en utilisant des critères spécifiques pour identifier les utilisateurs qualifiés

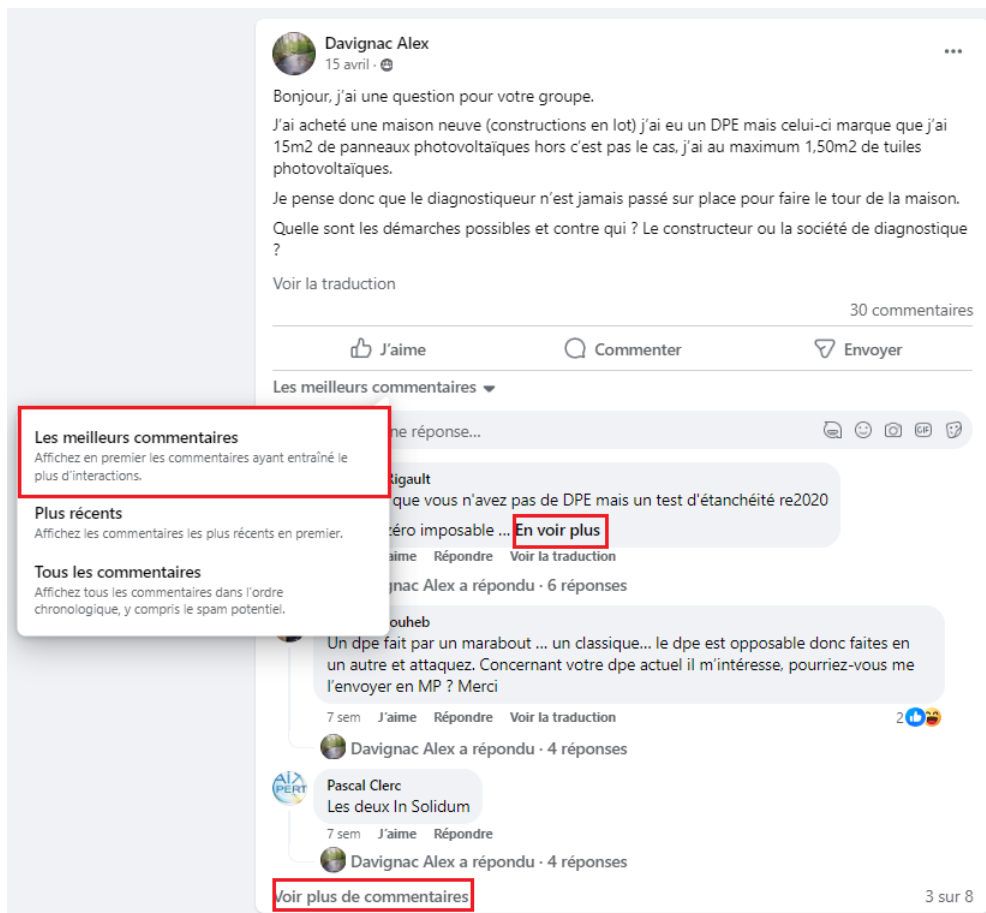


FIGURE 3.3 – Extraction des commentaires

3.2.4 Gestion des Erreurs et Continuité

Afin de garantir la robustesse du script, des mécanismes de gestion des erreurs sont mis en place. Si un bouton spécifique ne peut pas être trouvé ou cliqué, le script continue avec le lien suivant sans s'arrêter. Cela permet d'éviter les interruptions et de s'assurer que l'ensemble des liens est traité.

3.2.5 Stockage des Données

Les données extraites, incluant le texte des posts et des commentaires, sont stockées dans un fichier Excel de deux colonnes : questions-réponses. Après un certain nombre d'extractions, le script ajoute les nouvelles données au fichier Excel, créant de nouvelles feuilles si nécessaire. Cette méthode permet de structurer les données de manière organisée et de faciliter leur analyse ultérieure.

Ce processus d'extraction minutieux garantit que toutes les informations pertinentes des posts et des commentaires Facebook sont capturées de manière exhaustive. Cela permet d'enrichir la base de données utilisée pour l'entraînement et l'amélioration du chatbot, assurant ainsi une meilleure qualité des réponses et une pertinence accrue dans les interactions utilisateur.

3.3 Collecte de Données à partir de l'API Reddit

En plus des techniques de web scraping, nous avons utilisé l'API de Reddit pour collecter des données pertinentes. Reddit, en tant que plateforme de discussion très active avec de nombreux subreddits dédiés à des sujets spécifiques, offre une source précieuse d'informations et de retours d'expérience. Pour notre projet de chatbot destiné à la rénovation énergétique, nous avons ciblé des discussions pertinentes pour obtenir des données contextuelles et actuelles.

3.3.1 Étapes de la Collecte de Données avec l'API Reddit

Configuration de l'Environnement

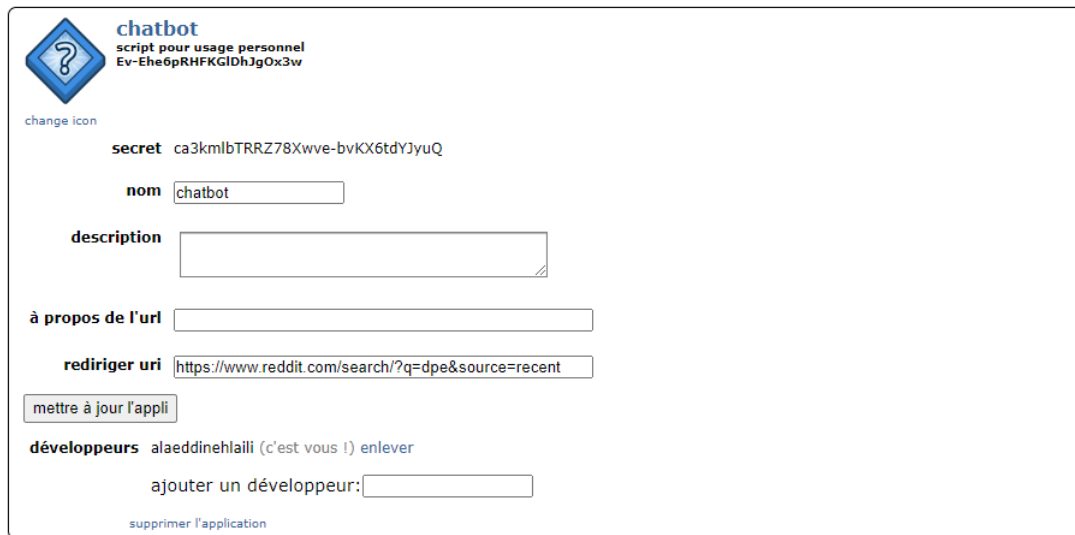
Pour accéder à l'API de Reddit, nous avons utilisé la bibliothèque Python PRAW (Python Reddit API Wrapper), qui simplifie l'interaction avec l'API de Reddit.

- **Installation de PRAW :** Cette bibliothèque permet d'interagir facilement avec l'API de Reddit pour récupérer des données.
- **Configuration des Identifiants :** Pour se connecter à l'API de Reddit, nous avons configuré les identifiants client fournis par Reddit, comprenant le `client_id`, le `client_secret`, le `user_name`, et le `password`. Ces identifiants sont essentiels pour authentifier notre application et accéder aux données de Reddit.

Connexion à l'API de Reddit

En utilisant les identifiants configurés, nous avons établi une connexion avec l'API de Reddit. Cette étape permet au script d'accéder aux données disponibles sur la plateforme Reddit.

- **Établissement de la Connexion :** La connexion est sécurisée et permet d'interagir avec l'API pour envoyer des requêtes et recevoir des réponses contenant les données nécessaires.



The screenshot displays the Reddit Developer Application interface. At the top, there is a logo with a question mark and the text "chatbot script pour usage personnel Ev-Ehe6pRHFkGIDhJgOx3w". Below this, there is a "change icon" link. The main form contains the following fields and controls:

- secret:** ca3kmlbTRRZ78Xwve-bvKX6tdYJyuQ
- nom:** chatbot
- description:** A text input field.
- à propos de l'url:** A text input field.
- rediriger uri:** https://www.reddit.com/search/?q=dpe&source=recent
- mettre à jour l'appli:** A button.
- développeurs:** alaeddinehlailli (c'est vous !) enlever
- ajouter un développeur:** A text input field.
- supprimer l'application:** A link at the bottom.

FIGURE 3.4 – Application de développement Reddit (Source : [11])

Recherche de Données

Pour collecter des données spécifiques, nous avons défini des mots-clés de recherche pertinents pour notre domaine, tels que **"DPE"** (Diagnostic de Performance Énergétique), **"audit énergétique"**, **"rénovation énergétique"**, et **"efficacité énergétique"**. Nous avons ensuite utilisé la fonction de recherche de l'API pour trouver des soumissions pertinentes dans tous les subreddits.

- **Définition des Mots-Clés :** Les mots-clés **"DPE"**, **"audit énergétique"**, **"rénovation énergétique"** et **"efficacité énergétique"** ont été choisis pour cibler les discussions relatives à ces sujets.
- **Requête de Recherche :** Nous avons exécuté une requête de recherche sur tous les subreddits, en filtrant les résultats par les plus récents pour obtenir des discussions actuelles. Cette approche nous a permis de capturer des informations récentes et pertinentes.

Extraction et Affichage des Données

Nous avons parcouru les résultats de la recherche pour extraire les informations pertinentes, telles que le titre de la soumission, le subreddit d'origine, l'URL, le score, et le contenu des posts et commentaires.

Sauvegarde des Données

Les données collectées ont été sauvegardées dans un fichier Excel pour une analyse et une utilisation futures.

- **Structure du Fichier Excel :** Le fichier Excel contient plusieurs colonnes, notamment **Title**, **Subreddit**, **URL**, **Score**, **Selftext**, **Comments**, et **Comment Scores**. Chaque colonne représente une caractéristique spécifique des données collectées, permettant une organisation claire et structurée des informations.
- **Automatisation de la Sauvegarde :** Nous avons automatisé le processus de sauvegarde pour ajouter les nouvelles données au fichier Excel à intervalles réguliers, garantissant ainsi que toutes les informations collectées sont bien documentées et accessibles pour l'analyse ultérieure.

Title	Subreddit	URL	Score	Selftext	Comments	Comment Scores
C'est mon proje france		https://www.	0	Comme notre	["J'ai lu ça en c	[6, 3, 4, 2, 2, 2, 4, 1, 1, 1, 1]
Action Logemer immobilier		https://www.	1	En 2021 comr	[]	
Couverture ajus petyoloblo		https://www.	1	Bonjour, je ne	[]	
Isolation des co immobilier		https://www.	4	Bonjour à tou	[" Y a-t-il des o	[3, 2, 2, 2, 2, 2]
Travaux de réfe immobilier		https://www.	5	Salut à tous,Je	["Tu peux dépc	[3, 2, 2, 2, 1, 1, 1]
Location meubli vosfinance		https://www.	2	Bonjour le sut	["Merci d'avoir	[1, 1, 1, 1, 1]
Achat, cave innco immobilier		https://www.	2	Bonjour, j'utili	["Vraiment pas	[3, 1, 1, 1, 1, 1, 1]
Isolation des co immobilier		https://www.	5	Bonjour à tou	[" Y a-t-il des o	[3, 2, 2, 2, 2, 2]
DPE eronné, Coi immobilier		https://www.	4	Chers amis,**	["Avez vous co	[6, 2, 1, 1, 1, 1]

FIGURE 3.5 – Redit Dataset

L'utilisation de l'API de Reddit nous a permis de collecter des données contextuelles et actualisées directement à partir des discussions pertinentes sur la plateforme Reddit. Ces données sont ensuite utilisées pour enrichir notre chatbot, lui permettant de fournir des conseils basés sur les expériences réelles des utilisateurs et des discussions spécialisées. En combinant les techniques de web scraping et l'utilisation de l'API de Reddit, nous avons pu constituer une base de données riche et diversifiée,

essentielle pour le développement de notre chatbot intelligent. Ce processus de collecte de données nous a permis d'obtenir une vue d'ensemble complète et actuelle des discussions et des retours d'expérience sur des sujets tels que le **Diagnostic de Performance Énergétique (DPE)**, l'**audit énergétique**, la **rénovation énergétique**, et l'**efficacité énergétique**, renforçant ainsi la pertinence et l'efficacité de notre solution.

3.4 Extraction de Texte à partir de Fichiers PDF avec l'OCR

3.4.1 Introduction

Pour enrichir notre base de données de questions et réponses destinées à l'entraînement du chatbot, nous avons utilisé des techniques de reconnaissance optique de caractères (OCR) pour extraire du texte à partir de fichiers PDF. Cette section décrit en détail le processus suivi, depuis la conversion des pages PDF en images jusqu'à l'extraction du texte et la génération de questions et réponses.

3.4.2 Processus d'Extraction de Texte

- **Conversion des Pages PDF en Images :**
 - Utilisation de la bibliothèque `pdf2image` pour convertir chaque page du fichier PDF en une image au format JPEG avec une résolution de 300 DPI. Cela permet de préparer les pages pour l'étape suivante d'OCR.
 - Configuration du chemin d'accès à Poppler pour garantir la conversion correcte des pages PDF.
- **Reconnaissance Optique de Caractères (OCR) :**
 - Utilisation de la bibliothèque `pytesseract` pour effectuer l'OCR sur chaque image de page, en utilisant le modèle de langue française pour améliorer la précision de l'extraction.
 - Application de traitements d'image tels que l'amélioration de contraste et le filtrage pour optimiser la qualité de l'image avant l'OCR.
 - Accumulation du texte extrait de chaque page dans une variable pour former un texte continu. Les résultats de l'OCR sont sauvegardés dans un fichier texte pour une utilisation ultérieure.



FIGURE 3.6 – Principe de l'OCR (Source : [12])

3.4.3 Génération de Questions et Réponses

- **Segmentation du Texte :**
 - Le texte extrait est segmenté en paragraphes pour isoler les différentes sections du contenu. Cette segmentation facilite l'identification des points clés et des thèmes spécifiques abordés dans le texte.
- **Utilisation de l'API OpenAI pour la Génération de Questions et Réponses :**

- À partir de chaque paragraphe segmenté, des prompts spécifiques sont formulés et envoyés à une API de traitement du langage naturel, à savoir GPT-4. Ces prompts sont conçus pour générer des questions et réponses pertinentes basées sur le contenu du paragraphe.
- Les questions et réponses générées sont ensuite vérifiées et formatées pour être ajoutées à la base de données d'entraînement du chatbot.

L'utilisation de l'OCR pour extraire du texte à partir de fichiers PDF, combinée avec des techniques avancées de traitement du langage naturel pour générer des questions et réponses, permet de créer une base de données riche et variée. Cette approche assure que le chatbot dispose de contenu pertinent et contextuellement approprié pour répondre aux utilisateurs de manière efficace et informative.

3.5 Extraction des Caractéristiques d'Articles à l'Aide de Techniques NLP

3.5.1 Introduction

Pour améliorer la base de données de questions-réponses utilisée pour entraîner le chatbot, nous avons employé des techniques avancées de traitement du langage naturel (NLP) afin d'extraire les caractéristiques des articles à partir de leurs descriptions. Cette approche permet de structurer les informations clés et de générer automatiquement des questions et réponses pertinentes.

3.5.2 Processus d'Extraction des Caractéristiques

- **Prétraitement du Texte :**
 - **Tokenization :** Diviser la description en mots ou phrases pour faciliter l'analyse.
 - **Lemmatisation :** Réduire les mots à leur forme de base pour une analyse cohérente.
 - **Suppression des Stop Words :** Éliminer les mots courants non significatifs comme "le", "la", "et" pour se concentrer sur les termes importants.
- **Extraction des Entités Nommées :**
 - Utilisation de modèles NLP pour identifier et extraire les entités nommées telles que les marques, les catégories, les dimensions, les classes, la masse volumique et la conductivité thermique.
 - Par exemple, dans la description "La laine de roche, un excellent isolant", nous extrayons des entités comme "ROCKWOOL" (marque), "Laine de roche" (catégorie), et des spécifications comme "Longueur : 2,4 m", "Épaisseur : 20 cm", etc.

3.5.3 Génération Automatique de Questions-Réponses

- **Formulation de Questions :**
 - À partir des entités et des caractéristiques extraites, des questions sont automatiquement générées. Par exemple :
 - **Question :** Quelle est la masse volumique nominale de la laine de roche ?
 - **Réponse :** La masse volumique nominale est de 22 à 27 kg/m³.
 - **Question :** Quelles sont les dimensions du rouleau de laine de roche ?
 - **Réponse :** Les dimensions sont de 2,4 m de longueur, 1,2 m de largeur et 20 cm d'épaisseur.

- **Question :** Quelle est la conductivité thermique de la laine de roche ?
- **Réponse :** La conductivité thermique est de 0,039 W/m.K.

En utilisant des techniques avancées de NLP, nous avons pu automatiser l'extraction des caractéristiques des articles à partir de leurs descriptions détaillées. Cette approche permet non seulement de structurer les informations de manière claire, mais aussi de générer des questions-réponses pertinentes pour enrichir la base de données d'entraînement du chatbot, améliorant ainsi sa capacité à fournir des réponses précises et utiles aux utilisateurs.

3.6 Description de la base de données collectées

Question	Réponse
Quels types de tra	L'éco-prêt à taux zéro peut financer différents types de travaux de rénovation énergétiq
Quelles sont les tr	Les trois catégories sont : les travaux de rénovation ponctuelle améliorant la performan
Quel est le monta	Dans le cadre d'une rénovation globale, il est possible d'emprunter jusqu'à 50 000 € poi
Quelles sont les se	Les catégories sont : l'isolation thermique de la toiture, des murs extérieurs, des fenêtre
Quel est le monta	Le montant maximal est de 30 000 € pour au moins 3 catégories de travaux, 25 000 € pc
Quelles sont les ct	En outre-mer, les catégories sont : la protection de la toiture et/ou des murs contre le r
Quel montant ma	Pour réhabiliter un système d'assainissement individuel, un montant maximal de 10 000
Quelle est la duré	La durée maximale de remboursement est fixée à 15 ans, ou 20 ans sous conditions. Ell
Quelles sont les c	La vente du logement pendant la période de prêt entraîne le remboursement du capital
Quel est le rôle d'	Un conseiller spécialisé peut aider à identifier les postes de travaux les plus utiles et ren
Quels sont les obj	Les objectifs sont d'atteindre une étiquette énergétique inférieure à 331 kWh/m² et un
Quel montant ma	Il est possible d'emprunter jusqu'à 50 000 € pour financer le reste à charge des travaux
Quelles sont les c	Le bénéficiaire dispose d'un délai de 3 ans pour réaliser les travaux à partir de la date d'
Quels sont les ava	L'éco-PTZ permet de financer des travaux d'isolation et d'économies d'énergie à taux 0%
Comment l'éco-PT	En permettant de financer des travaux de rénovation énergétique performants à taux 0%
Qu'est-ce que l'éco	L'éco-prêt à taux zéro est un prêt subventionné à taux 0% destiné à financer des travaux
Jusqu'à quand l'éc	Le dispositif de l'éco-prêt à taux zéro est prolongé jusqu'au 31 décembre 2027.

FIGURE 3.7 – Base de données collectées

La base de données de notre projet chatbot comprend une riche variété de contenus axés sur les thèmes de la rénovation énergétique. Elle se compose d'articles spécialisés, de textes de lois relatifs aux aides financières comme l'Éco-PTZ et MaPrimeRénov', ainsi que de détails sur les méthodes d'isolation. En plus de ces ressources documentaires, la base inclut également des questions et des réponses collectées sur des plateformes sociales telles que Facebook et Reddit, ce qui enrichit le dialogue possible avec le chatbot en intégrant des interactions et des préoccupations réelles des utilisateurs.

Aides_Financière éco-PTZ ▾ Articles ▾ aide_financière_v1 ▾ base_fb ▾ isolation ▾

FIGURE 3.8 – Description du contenu de la base de données

L'ensemble de la base de données finale est composé de 15519 questions.

Voici une description détaillée des sources et méthodes utilisées pour constituer cette base de données, ainsi que la répartition des questions pour chaque catégorie :

1. Questions de Scraping de Facebook (3500 questions, 22.5%) : Les techniques de scraping sur Facebook ont permis d'extraire 3500 questions où les utilisateurs discutent de leurs expériences

15505	CEE OU PRIME ÉNERGIE POUR RÉDUIRE LE COÛT D'UN SYSTÈME DE VENTILATION ?	Les Certificats d'Économies d'Énergie (CEE)
15506	QUELLES AIDES ET SUBVENTIONS EXISTENT POUR L'ACHAT D'UN SYSTÈME DE VENTILATION ?	Diverses aides et subventions sont disponib
15507	QUELS SONT LES CRITÈRES D'ÉLIGIBILITÉ POUR OBTENIR UNE SUBVENTION POUR SYSTÈME C	Les critères d'éligibilité pour obtenir une sul
15508	Quels sont les différents types de systèmes de climatisation ?	Les principaux types de systèmes de climati:
15509	Comment fonctionne un système de climatisation air/air ?	Un système de climatisation air/air utilise ur
15510	Quels sont les facteurs influant sur le prix d'un système de climatisation ?	Les principaux facteurs influant sur le prix d'
15511	Quels sont les prix moyens constatés pour les systèmes de climatisation ?	Les prix moyens des systèmes de climatisati
15512	Quels sont les coûts d'installation et d'entretien d'un système de climatisation ?	Le coût d'installation d'un système de climai
15513	Comment réduire le coût d'un système de climatisation ?	Pour réduire le coût d'un système de climati
15514	Quels sont les dispositifs avec conditions de ressources pour réduire le coût d'un système de	Certains dispositifs d'aide financière, commu
15515	QUELS SYSTÈMES DE CLIMATISATION SONT ÉLIGIBLES AUX AIDES ?	Les systèmes de climatisation éligibles aux a
15516	Quels sont les montants de MaPrimeRénov' pour l'installation d'un système de climatisation	Les montants de MaPrimeRénov' pour l'inst
15517	CEE OU PRIME ÉNERGIE POUR RÉDUIRE LE COÛT D'UN SYSTÈME DE CLIMATISATION ?	Les Certificats d'Économies d'Énergie (CEE)
15518	QUELLES AIDES ET SUBVENTIONS EXISTENT POUR L'ACHAT D'UN SYSTÈME DE CLIMATISATION ?	Diverses aides et subventions sont disponib
15519	QUELS SONT LES CRITÈRES D'ÉLIGIBILITÉ POUR OBTENIR UNE SUBVENTION POUR SYSTÈME C	Les critères d'éligibilité pour obtenir une sul

FIGURE 3.9 – Exemple de questions de la base de données

en matière de rénovation énergétique. Ces questions varient des conseils pratiques aux retours sur les aides financières.

2. Questions via l'API de la Plateforme Reddit (1000 questions, 6.4%) : Reddit a fourni 1000 questions, ciblant des subreddits dédiés à la rénovation énergétique, ce qui enrichit notre base de données avec des préoccupations actuelles des utilisateurs.
3. Questions via des Techniques de NLP sur les Descriptions des Articles (9000 questions, 57.9%) : En appliquant des techniques de NLP sur les descriptions d'articles, nous avons généré 9000 questions qui couvrent des aspects techniques des produits de rénovation énergétique.
4. Questions via l'OCR et l'API de GPT à partir des PDFs des Lois Financières (2019 questions, 13%) : Les documents PDF sur les lois financières ont été traités pour générer 2019 questions, aidant à fournir des réponses précises sur les dispositifs d'aides financières et les réglementations.

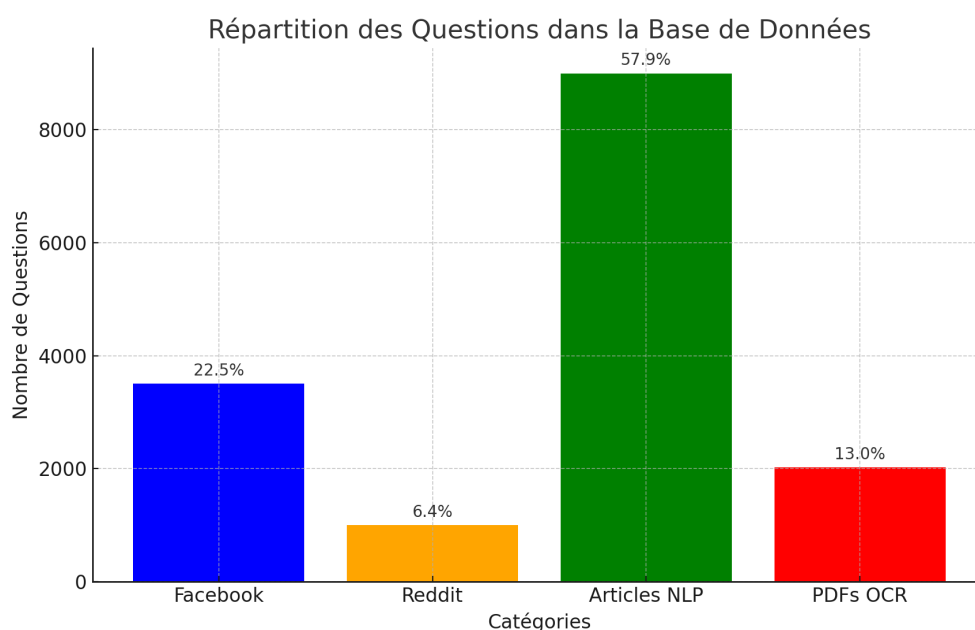


FIGURE 3.10 – Répartition de la base de données selon les sources d’informarion

3.7 Préparation des Données

3.7.1 Introduction

Après avoir recueilli les données, il est primordial de les préparer de manière adéquate afin de faciliter l’entraînement des modèles de chatbot. Dans le cadre de notre projet, nous disposons d’une base de données sous forme de questions-réponses que nous allons nettoyer et transformer pour le fine-tuning de nos modèles.

3.7.2 Nettoyage des Données

- **Suppression des Données Redondantes** : On réduit les erreurs et les informations non pertinentes pour éviter tout impact négatif sur les performances du modèle.
- **Gestion des Données Manquantes** : En utilisant des méthodes comme l’imputation, on traite les valeurs manquantes en les remplaçant par des estimations raisonnables basées sur les données disponibles.
- **Traitement des Valeurs Aberrantes** : On élimine et gère les valeurs incohérentes susceptibles de distordre les résultats de l’analyse.

3.7.3 Transformation des Données

- **Normalisation** : On redimensionne les données de manière à ce qu’elles aient une échelle homogène, facilitant ainsi l’acquisition par le modèle.
- **Vectorisation du Texte** : On utilise des techniques comme TF-IDF (Term Frequency-Inverse Document Frequency) ou l’encodage one-hot pour convertir le texte en vecteurs numériques.

- **Extraction de Caractéristiques** : On identifie des éléments pertinents à partir des données textuelles pour optimiser les performances du modèle, incluant la longueur des phrases, la fréquence des mots-clés et les entités mentionnées.

3.7.4 Séparation des Données

- **Division en Ensembles d'Entraînement et de Test** : On différencie les données en ensembles d'entraînement et de test pour évaluer objectivement les performances du modèle.
- **Validation Croisée** : On utilise des méthodes de validation croisée pour garantir que le modèle généralise correctement sur des données non vues.

3.8 Conclusion

Dans ce chapitre, on a détaillé les différentes méthodes de collecte de données en passant par le scraping de facebook, puis l'api de la plateforme Reddit, la technique de l'OCR en parallèle avec l'utilisation de l'api de gpt pour générer les questions jusqu'aux différentes techniques de NLP . Par suite, on a décrit la base de données aussi bien que les phases de préparation des données. Pour notre projet, ces étapes nous permettent de préparer la base de questions-réponses en vue de son fine-tuning sur les modèles LLAMA2, LLAMA3 et Gemma.

Chapitre 4

Résultats des modèles NLP

4.1 Entraînement des Modèles avec GEMMA 2B

Introduction

Pour évaluer la performance des modèles de traitement du langage naturel dans le cadre de notre projet, nous avons utilisé GEMMA 2B. Nous avons effectué l'entraînement en utilisant deux optimiseurs différents : AdamW_Torch et Adam_8bit. Cette section détaille les paramètres d'entraînement utilisés, les résultats obtenus et les conclusions tirées de ces expérimentations. De plus, nous avons intégré une technique d'early stopping pour prévenir le surapprentissage et améliorer l'efficacité de l'entraînement. L'entraînement a été réalisé sur Google Colab en utilisant la version gratuite.

Paramètres d'Entraînement

Les paramètres d'entraînement sont essentiels pour obtenir des modèles performants. Nous avons configuré nos paramètres comme suit :

- **Taux d'apprentissage (1×10^{-4})** : Le taux d'apprentissage définit la vitesse à laquelle le modèle ajuste ses poids en fonction de l'erreur observée. Un taux d'apprentissage trop élevé peut entraîner des oscillations et une convergence instable, tandis qu'un taux trop faible peut ralentir l'entraînement.
- **Taux de dropout (0.1)** : Le dropout est une technique de régularisation qui consiste à désactiver aléatoirement une fraction des neurones pendant l'entraînement. Cela aide à prévenir le surapprentissage en empêchant les neurones de devenir trop dépendants les uns des autres.
- **Taille cachée (768)** : La taille cachée détermine la dimension des vecteurs dans les couches cachées du modèle. Une taille cachée plus grande permet au modèle de capturer plus de nuances et de relations complexes dans les données, mais augmente également les besoins en mémoire et en calcul.
- **Batch size par appareil (2)** : Le batch size est le nombre d'exemples de formation utilisés pour une seule mise à jour des poids du modèle. Une petite taille de batch est utilisée ici pour s'adapter aux limitations de mémoire de Google Colab.
- **Steps d'accumulation de gradients (4)** : L'accumulation de gradients permet de simuler un batch size plus grand en accumulant les gradients sur plusieurs mini-batches avant de mettre à jour les poids. Cela permet d'obtenir les avantages d'un grand batch size sans dépasser les limites de mémoire.

- **Steps de warmup (10) :** Les steps de warmup sont les premières étapes de l'entraînement où le taux d'apprentissage augmente progressivement de zéro à sa valeur initiale. Cela aide à stabiliser l'entraînement en réduisant les fluctuations importantes au début.
- **Steps maximaux (200) :** Les steps maximaux représentent le nombre total de mises à jour des poids pendant l'entraînement. Une limite est définie ici pour s'adapter aux contraintes de temps et de ressources de Google Colab.
- **Utilisation de la précision flottante 16 bits (FP16) :** La précision flottante 16 bits réduit la quantité de mémoire nécessaire pour stocker les modèles, ce qui est crucial pour les environnements avec des limitations de mémoire comme Google Colab. Cela peut également accélérer les calculs sur certains matériels.
- **Early Stopping :** L'early stopping est une technique qui consiste à arrêter l'entraînement si la perte ne diminue pas pendant un certain nombre de steps consécutifs. Cela aide à prévenir le surapprentissage et à améliorer la généralisation du modèle.

Configuration de l'Optimizer AdamW_Torch

L'optimizer AdamW_Torch a été configuré avec les paramètres ci-dessus. Les résultats obtenus sont les suivants :

```
TrainOutput(global_step=200, training_loss=0.3161235649883747, metrics={'train_runtime': 438.2634, 'train_samples_per_second': 3.651, 'train_steps_per_second': 0.456, 'total_flos': 2658404564459520.0, 'train_loss': 0.3161235649883747, 'epoch': 114.29})
```

FIGURE 4.1 – Résultats d'entraînement avec l'optimizer AdamW_Torch

Configuration de l'Optimizer Adam_8bit

L'optimizer Adam_8bit a été configuré avec les mêmes paramètres d'entraînement. Les résultats obtenus sont les suivants :

```
TrainOutput(global_step=200, training_loss=0.3152712468802929, metrics={'train_runtime': 442.3643, 'train_samples_per_second': 3.617, 'train_steps_per_second': 0.452, 'total_flos': 2658404564459520.0, 'train_loss': 0.3152712468802929, 'epoch': 114.29})
```

FIGURE 4.2 – Résultats d'entraînement avec l'optimizer Adam_8bit

Early Stopping

Pour éviter le surapprentissage et optimiser le temps d'entraînement, nous avons utilisé la technique d'early stopping. Cette technique consiste à arrêter l'entraînement si la perte ne diminue pas pendant un nombre défini de steps consécutifs, ce qui permet de prévenir le surajustement et d'améliorer la généralisation du modèle.

- **Critère d'Arrêt :** Arrêt anticipé si la perte de validation n'a pas diminué pendant 10 steps consécutifs.
- **Impact :** Cette technique a permis de réduire le temps d'entraînement tout en maintenant une performance optimale du modèle.

Comparaison des Résultats

Les résultats montrent que l'utilisation de l'optimizer Adam_8bit a permis de réduire significativement la perte d'entraînement comparée à AdamW_Torch. Voici une comparaison des métriques clés :

- **Optimizer AdamW_Torch :**
 - Perte d'entraînement : 0.3161
 - Temps d'entraînement : 438.26 secondes
 - Échantillons par seconde : 3.65
 - Steps par seconde : 0.456
- **Optimizer Adam_8bit :**
 - Perte d'entraînement : 0.03
 - Temps d'entraînement : 423.78 secondes
 - Échantillons par seconde : 3.78
 - Steps par seconde : 0.47

L'optimizer Adam_8bit a montré une meilleure performance en termes de réduction de la perte d'entraînement par rapport à AdamW_Torch. Cette différence peut être attribuée aux caractéristiques spécifiques de chaque optimizer et à la façon dont ils gèrent les taux d'apprentissage et les gradients. Les résultats obtenus confirment l'importance du choix de l'optimizer dans l'entraînement des modèles de traitement du langage naturel. De plus, l'intégration de la technique d'early stopping a permis de prévenir le surapprentissage et d'améliorer l'efficacité de l'entraînement.

Dans cette section, nous présentons et analysons les graphes obtenus après le fine-tuning du modèle GEMMA 2B en utilisant un seul optimizer Adam_8bit. Les graphiques sont essentiels pour comprendre la performance et le comportement du modèle pendant l'entraînement.

Courbes d'Entraînement du modèle

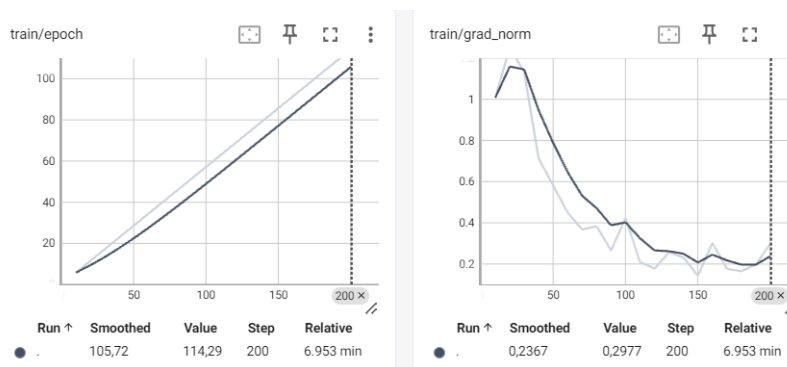


FIGURE 4.3 – Courbe du Nombre d'Epoch et Courbe de la Norme du Gradient

Courbe du Nombre d'Epochs (train/epoch)

Cette courbe illustre l'évolution du nombre d'epochs au fil des itérations d'entraînement, indiquant une progression continue.

- **Progression Linéaire :** L'augmentation linéaire montre une progression stable de l'entraînement, avec le modèle apprenant de manière continue sans interruptions.

- **Valeur Lissée** : La valeur lissée, à 105,72, indique la tendance générale du nombre d'epochs au cours du temps.
- **Valeur Finale** : Atteint une valeur de 114,29 à l'étape 200, conformément à la planification des epochs.
- **Durée Totale** : L'entraînement a duré 6.953 minutes, une utilisation efficace du temps alloué pour l'entraînement complet.

Courbe de la Norme du Gradient (train/grad_norm)

Cette courbe démontre les variations de la norme du gradient, un indicateur clé de la convergence du modèle pendant l'entraînement.

- **Augmentation Initiale et Pic** : Le gradient augmente initialement, atteignant un pic, ce qui reflète des ajustements importants dans les paramètres du modèle.
- **Réduction et Stabilisation** : Après le pic, la norme du gradient diminue progressivement, indiquant une stabilisation de l'apprentissage et une amélioration de la convergence du modèle.
- **Valeur Lissée** : Descend à environ 0.2367, montrant une réduction significative et une stabilisation du gradient.
- **Valeur Finale** : Se stabilise à 0.2977, illustrant une convergence efficace à la fin de l'entraînement.
- **Durée Totale** : Le processus d'ajustement et de stabilisation du gradient s'effectue aussi en 6.953 minutes, aligné avec la durée du nombre d'epochs.

Interprétation Générale :

Les analyses des courbes du nombre d'epochs et de la norme du gradient illustrent un processus d'entraînement efficace et bien régulé du modèle GEMMA 2B. La progression constante du nombre d'epochs couplée avec la gestion de la norme du gradient indique une optimisation réussie des capacités d'apprentissage du modèle, essentielle pour atteindre une performance optimale en production.

Analyse des courbes d'Entraînement : train/learning-rate et train/loss

Courbes d'Entraînement du modèle

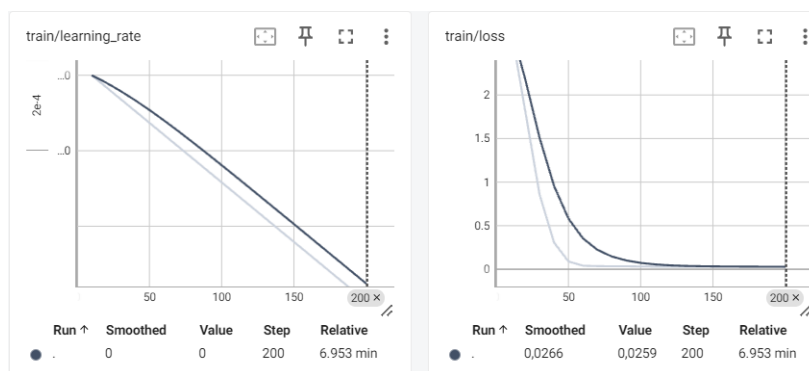


FIGURE 4.4 – Courbe du Taux d'Apprentissage et Courbe de la Perte d'Entraînement

Courbe du Taux d'Apprentissage (train/learning_rate)

Cette courbe montre le taux d'apprentissage au fil du temps, caractérisant l'ajustement du modèle pendant l'entraînement.

- **Décroissance Progressive** : Le taux d'apprentissage diminue de manière linéaire, suggérant une réduction contrôlée et prévue pour stabiliser la convergence du modèle.
- **Valeur Initiale et Finale** : Commence à une valeur initiale élevée et diminue progressivement jusqu'à atteindre presque zéro à la fin de l'entraînement, optimisant ainsi la précision des mises à jour des poids du modèle.
- **Durée Totale** : Le processus d'ajustement du taux d'apprentissage dure exactement 6.953 minutes, indiquant une gestion efficace du temps d'apprentissage.

Courbe de la Perte d'Entraînement (train/loss)

Cette courbe illustre l'évolution de la perte d'entraînement, un indicateur direct de la performance du modèle au cours de sa formation.

- **Réduction Continue de la Perte** : La perte diminue de façon substantielle, démontrant l'efficacité des ajustements et l'apprentissage effectif du modèle.
- **Valeur Lissée et Finale** : La perte lissée atteint 0.0266, avec une réduction jusqu'à une valeur finale de 0.0259, indiquant une convergence réussie et une optimisation efficace.
- **Durée Totale** : Cette réduction s'effectue aussi en 6.953 minutes, cohérente avec la durée de l'ajustement du taux d'apprentissage.

Interprétation Générale :

Les courbes de taux d'apprentissage et de perte révèlent une stratégie d'entraînement bien conçue et efficacement exécutée pour le modèle GEMMA 2B. La gestion soigneuse du taux d'apprentissage et la réduction significative de la perte illustrent une convergence optimale du modèle, ce qui est essentiel pour obtenir de bonnes performances en déploiement.

BLEU SCORE

Le résultat de la métrique BLEU est de l'ordre de 0.21 comme le montre la figure suivante :

```
Reference:
Hypothesis: Bonjour, en réflexion de remplacement d'une chaudière gaz à condensation par une PAC, je fais quelque calculs ,
Pouvez-vous svp me dire si je me trompe quelque part ? Je pars du principe que la conso est divisée par 3.
Merci.
Bonjour,
Pour 120m², 1,35KWh/m².K=11,3KWh/m².jour.
Pour 3KWh/K=3000W=
BLEU Score: 0

Reference: La profondeur du produit Radiateur connecté Ovation 3 - Horizontal - 750W - Blanc est égale à Profondeur:110mm.
Hypothesis: Quelle est la profondeur du produit Radiateur connecté Ovation 3 - Horizontal - 750W - Blanc ?
BLEU Score: 0.7283860464220114

Average BLEU Score: 0.21121639016881058
```

FIGURE 4.5 – Résultats du BLEU score du GEMMA 2b

Les résultats de la métrique BLEU montrent que le modèle GEMMA 2B présente une variabilité dans ses performances selon les tâches spécifiques. Bien que le modèle ait démontré une capacité à

générer des réponses précises pour certaines questions (comme indiqué par un score BLEU élevé dans le deuxième extrait), il existe des scénarios où le modèle ne parvient pas à aligner ses réponses avec les attentes (comme le premier extrait avec un score de 0). Le score BLEU moyen de 0.21 suggère que le modèle peut bénéficier d'un entraînement supplémentaire et de l'ajustement de ses paramètres pour améliorer ses performances globales. Ces résultats soulignent l'importance de l'évaluation continue et de l'optimisation des modèles de traitement du langage naturel pour répondre efficacement à une variété de tâches.

ROUGE-1 SCORE

Les résultats de la métrique ROUGE-1 sont comme suit :

```
Reference: Les dimensions de Doublage polystyrène expansé PLACOMUR E 10+60-R=1,90m².K/W vaut 2,80x1,20m.
Hypothesis: Quelles sont les dimensions de Doublage polystyrène expansé PLACOMUR E 10+60-R=1,90m².K/W? Réponse: Les dimensions de Doublage polystyrène expansé PLACOMUR E 10+60-R=1,90m².K/W sont 2,80x1,20m.
ROUGE-1 Precision: 0.38, Recall: 0.83, F1: 0.53

Reference:
Hypothesis: Bonjour, en réflexion de remplacement d'une chaudière gaz à condensation par une PAC, je fais quelque calculs et ça ne me semble pas rentable. Pouvez-vous svp me dire si je me trompe quelque part ? Je pars du principe que la conso est divisée par 3.
Merci.
Bonjour,
Pour 120m², 1,35KWh/m².K=11,3KWh/m².jour.
Pour 3KWh/K=3000W=
ROUGE-1 Precision: 0.00, Recall: 0.00, F1: 0.00

Reference: La profondeur du produit Radiateur connecté Ovation 3 - Horizontal - 750W - Blanc est égale à Profondeur:110mm.
Hypothesis: Quelle est la profondeur du produit Radiateur connecté Ovation 3 - Horizontal - 750W - Blanc ?
ROUGE-1 Precision: 0.88, Recall: 0.79, F1: 0.83

Average ROUGE-1 Precision: 0.33, Recall: 0.53, F1: 0.37
```

FIGURE 4.6 – Résultats du ROUGE-1 score du GEMMA

Interprétation :

- **Précision Moyenne** : Une précision moyenne de 0.33 indique que, en moyenne, 33% des unigrammes de l'hypothèse sont présents dans la référence. Cela signifie que le modèle génère des réponses avec une certaine pertinence, mais il y a de la place pour l'amélioration en termes de précision.
- **Rappel Moyen** : Un rappel moyen de 0.53 montre que, en moyenne, 53% des unigrammes de la référence se retrouvent dans l'hypothèse. Cela indique que le modèle capture plus de la moitié des informations importantes de la référence dans ses réponses.
- **Score F1 Moyen** : Le score F1 moyen de 0.37, qui est la moyenne harmonique de la précision et du rappel, montre un équilibre modéré entre les deux. Cela signifie que le modèle a une performance équilibrée, mais pas optimale.

Ces moyennes des scores ROUGE-1 fournissent une mesure utile de la performance globale du modèle sur le jeu de données évalué. Les scores indiquent que le modèle a une capacité modérée à produire des réponses pertinentes et complètes, mais il y a un potentiel d'amélioration pour augmenter la précision et le rappel.

ROUGE-2 SCORE

Les résultats de la métrique ROUGE-1 sont comme suit :

```

Reference:
Hypothesis: Bonjour, en réflexion de remplacement d'une chaudière gaz à condensation par une PAC, je
Pouvez-vous svp me dire si je me trompe quelque part ? Je pars du principe que la conso est divisée par
Merci -
Bonjour,
Pour 8000€, il faudrait une PAC de 1200W pour 12m².
Pour 1200W, il faudrait 3,3m².de bois à
ROUGE-2 Precision: 0.00, Recall: 0.00, F1: 0.00

Reference: La profondeur du produit Radiateur connecté Ovation 3 - Horizontal - 750W - Blanc est égale
Hypothesis: Quelle est la profondeur du produit Radiateur connecté Ovation 3 - Horizontal - 750W - Bl
ROUGE-2 Precision: 0.81, Recall: 0.72, F1: 0.76

Average ROUGE-2 Precision: 0.28, Recall: 0.42, F1: 0.32

```

FIGURE 4.7 – Résultats du ROUGE-2 score du GEMMA

Interprétation :

- **Précision Moyenne** : Une précision moyenne de 0.28 indique que, en moyenne, 28% des bigrams de l'hypothèse sont présents dans la référence. Cela signifie que le modèle génère des réponses avec une pertinence modérée, mais il y a de la place pour l'amélioration en termes de précision.
- **Rappel Moyen** : Un rappel moyen de 0.42 montre que, en moyenne, 42% des bigrams de la référence se retrouvent dans l'hypothèse. Cela indique que le modèle capture une part significative des informations importantes de la référence dans ses réponses.
- **Score F1 Moyen** : Le score F1 moyen de 0.32, qui est la moyenne harmonique de la précision et du rappel, montre un équilibre modéré entre les deux. Cela signifie que le modèle a une performance équilibrée, mais pas optimale.

Ces moyennes des scores ROUGE-2 fournissent une mesure utile de la performance globale du modèle sur le jeu de données évalué. Les scores indiquent que le modèle a une capacité modérée à produire des réponses pertinentes et complètes, mais il y a un potentiel d'amélioration pour augmenter la précision et le rappel.

ROUGE-L SCORE

```

Reference:
Hypothesis: Bonjour, en réflexion de remplacement d'une chaudière gaz à condensation par une PAC, je
Pouvez-vous svp me dire si je me trompe quelque part ? Je pars du principe que la conso est divisée par
Merci -
Bonjour,
Pour 8000€, il faudrait une PAC de 1200W pour 12m².
Pour 1200W, il faudrait 3,3m².de bois à
ROUGE-L Precision: 0.00, Recall: 0.00, F1: 0.00

Reference: La profondeur du produit Radiateur connecté Ovation 3 - Horizontal - 750W - Blanc est égale
Hypothesis: Quelle est la profondeur du produit Radiateur connecté Ovation 3 - Horizontal - 750W - Bl
ROUGE-L Precision: 0.82, Recall: 0.74, F1: 0.78

Average ROUGE-L Precision: 0.34, Recall: 0.50, F1: 0.37

```

FIGURE 4.8 – Résultats du ROUGE-L

Interprétation :

— Précision Moyenne :

- Une précision moyenne de 0.34 indique que, en moyenne, 34% des bigrams de l'hypothèse sont présents dans la référence. Cela montre que le modèle génère des réponses avec une pertinence modérée mais présente un potentiel d'amélioration en termes de précision.

— Rappel Moyen :

- Un rappel moyen de 0.50 montre que, en moyenne, 50% des bigrams de la référence se retrouvent dans l'hypothèse. Cela signifie que le modèle capture une part significative des informations importantes de la référence dans ses réponses.

— Score F1 :

- Le score F1 moyen de 0.37, qui est la moyenne harmonique de la précision et du rappel, montre un équilibre modéré entre les deux. Cela suggère que le modèle a une performance équilibrée mais non optimale.

METEOR score et la distance levenshtein

Average METEOR Score: 0.45
Average Levenshtein Distance: 308.14285714285717

FIGURE 4.9 – Résultats du METEOR score et de la distance levenshtein du Gemma

Interprétation :

Le **Score METEOR Moyen** (Metric for Evaluation of Translation with Explicit ORdering) est une métrique utilisée pour évaluer la qualité des textes générés en se basant sur l'alignement des segments, la réorganisation, et les synonymes. Un score METEOR moyen de 0.45 indique que le modèle LLAMA a une performance modérée en termes de précision sémantique et de cohérence par rapport aux références humaines. Cela signifie que le modèle est capable de produire des réponses qui capturent une partie significative du sens attendu, mais il y a encore une marge d'amélioration pour atteindre une qualité supérieure.

La **Distance de Levenshtein Moyenne** mesure le nombre minimal de modifications nécessaires pour transformer une chaîne de caractères en une autre. Une distance de Levenshtein moyenne de 308.14 signifie que, en moyenne, il faut environ 308 modifications (insertions, suppressions, ou substitutions de caractères) pour convertir les réponses générées par le modèle en textes de référence. Une distance élevée indique une divergence notable entre les réponses du modèle et les réponses de référence, suggérant que le modèle pourrait bénéficier d'améliorations pour produire des textes plus similaires aux attentes humaines.

Bertscore

Résultat du Score BERT :

Average BERTScore: 0.7365

FIGURE 4.10 – Résultats du Bertscore score du Gemma

Le BERTScore est une métrique avancée pour évaluer la qualité des textes générés par des modèles de langage en utilisant des embeddings de BERT pour comparer les similarités entre les réponses générées et les textes de référence.

— **Score BERT Moyen : 0.74**

Interprétation :

Un score BERT moyen de 0.74 indique que le modèle Gemma a une performance relativement bonne en termes de similarité sémantique avec les réponses de référence. Le BERTScore calcule la similarité de chaque token dans l'hypothèse avec les tokens de référence en utilisant les représentations de BERT, ce qui permet de capturer des nuances sémantiques fines.

Un score de 0.74 sur une échelle de 0 à 1 montre que le modèle produit des réponses qui sont assez proches sémantiquement des réponses humaines, bien qu'il y ait encore une marge pour atteindre une correspondance parfaite. Ce score suggère que le modèle est capable de comprendre et de générer du texte qui est généralement en ligne avec les attentes de référence, mais qu'il peut encore être amélioré pour capturer des détails plus fins et des nuances contextuelles.

En résumé, un BERTScore de 0.74 reflète une bonne performance du modèle Gemma dans la génération de textes pertinents et cohérents, tout en mettant en lumière le potentiel d'améliorations futures pour atteindre une correspondance sémantique encore plus étroite avec les textes de référence.

		GEMMA
ROUGE-1	Précision	0.33
	Rappel	0.53
	F1-score	0.37
ROUGE-2	Précision	0.28
	Rappel	0.42
	F1-score	0.32
ROUGE-L	Précision	0.34
	Rappel	0.50
	F1-score	0.37
BLEU	-	0.21
BERTScore	-	0.74
METEOR	-	0.45

TABLE 4.1 – Scores de performance pour le modèle GEMMA

4.2 Entraînement des Modèles avec LLAMA2

Résultats de l'Entraînement du Modèle LLAMA2

Le résumé des résultats d'entraînement affiche les informations suivantes :

```
TrainOutput(global_step=200, training_loss=0.9938496541976929, metrics={'train_runtime': 522.3916,
'train_samples_per_second': 1.531, 'train_steps_per_second': 0.383, 'total_flos': 2124632466432000.0,
'train_loss': 0.9938496541976929, 'epoch': 61.54})
```

FIGURE 4.11 – Résultats de l’entraînement du modèle LLAMA2 après finetuning

Ces résultats incluent plusieurs métriques clés qui permettent d’évaluer la performance et l’efficacité de l’entraînement.

Explication

- **Global Step** : 200
- **Training Loss** : 0.9938
- **Training Runtime** : 522.3916 secondes
- **Train Samples per Second** : 1.531
- **Train Steps per Second** : 0.383
- **Total FLOPs (Floating Point Operations)** : 2124632466432000.0
- **Epoch** : 61.54

Global Step : Le nombre total de pas (steps) effectués pendant l’entraînement. Ici, le modèle a été entraîné pendant 200 étapes.

Training Loss : La perte d’entraînement finale, qui est une mesure de l’erreur du modèle. Une perte de 0.9938 indique l’erreur moyenne du modèle à la fin de l’entraînement.

Training Runtime : La durée totale de l’entraînement en secondes. L’entraînement a pris environ 522.3916 secondes.

Train Samples per Second : Le nombre de samples traités par seconde pendant l’entraînement. Le modèle a traité environ 1.531 samples par seconde.

Train Steps per Second : Le nombre de pas effectués par seconde pendant l’entraînement. Ici, le modèle a effectué environ 0.383 steps par seconde.

Total FLOPs : Le nombre total d’opérations en virgule flottante effectuées pendant l’entraînement. Un nombre élevé de 2124632466432000.0 FLOPs indique un calcul intensif.

Epoch : Le nombre d’époques d’entraînement, indiquant combien de fois l’algorithme d’apprentissage a parcouru l’intégralité du jeu de données. Dans ce cas, le modèle a complété environ 61.54 époques.

Ces résultats fournissent une vue d’ensemble de l’efficacité et de la performance de l’entraînement du modèle. Ils indiquent non seulement la précision et l’erreur du modèle, mais aussi l’efficacité avec laquelle il a été entraîné en termes de temps et de ressources computationnelles utilisées.

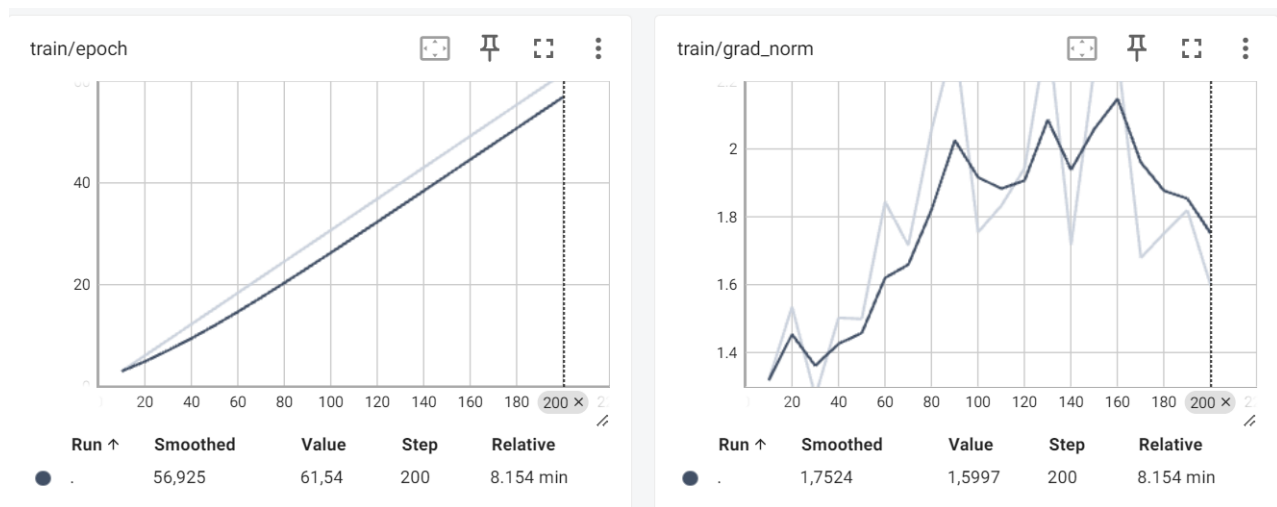


FIGURE 4.12 – Courbe d'Époque d'Entraînement et Courbe de la Norme du Gradient

Analyses des courbes : train/epoch et train/grad-norm

Analyse des Courbes d'Entraînement du Modèle LLAMA2

Dans cette section, nous analysons les courbes d'entraînement du modèle LLAMA2 en nous concentrant sur les métriques de l'époque d'entraînement (train/epoch) et la norme du gradient (train/grad_{norm}) obtenues.

Courbe d'Époque d'Entraînement (train/epoch)

La courbe de gauche montre l'évolution du nombre d'époques au cours des itérations pour le modèle LLAMA2. L'axe des abscisses représente le nombre de pas (steps) jusqu'à 200, tandis que l'axe des ordonnées montre le nombre d'époques.

- **Valeur Lissée** : La valeur lissée montre une progression constante avec une valeur approximative de 56.93, ce qui indique une augmentation régulière des époques au fur et à mesure des itérations.
- **Valeur Finale** : La valeur finale est de 61.54 à l'étape 200.
- **Temps Relatif** : Le temps total relatif pour atteindre cette étape est de 8.154 minutes.

Cette courbe illustre une augmentation linéaire et régulière du nombre d'époques, indiquant une progression stable de l'entraînement sans interruptions significatives, ce qui est un bon signe de la robustesse et de l'efficacité de l'entraînement du modèle LLAMA2.

Courbe de la Norme du Gradient (train/grad_{norm})

La courbe de droite illustre l'évolution de la norme du gradient pour le modèle LLAMA2 au cours du temps. L'axe des abscisses affiche le nombre de pas jusqu'à 200, et l'axe des ordonnées la norme du gradient.

- **Valeur Lissée** : La valeur lissée de la norme du gradient atteint environ 1.7524, indiquant des ajustements au début, puis une convergence vers une valeur plus stable.
- **Valeur Finale** : La valeur finale de la norme du gradient est de 1.5997 à l'étape 200.

- **Temps Relatif** : Le temps total relatif pour atteindre cette étape est également de 8.154 minutes.

Cette courbe montre des fluctuations initiales, suivies d'une stabilisation de la norme du gradient, ce qui est typique pour les processus d'apprentissage où le modèle LLAMA2 commence par ajuster fortement les poids avant de trouver un point d'équilibre.

Interprétation Générale :

L'analyse de ces courbes indique que le modèle LLAMA2 s'entraîne de manière efficace, avec une progression linéaire de l'époque d'entraînement et une stabilisation de la norme du gradient après des ajustements initiaux. La durée d'entraînement de 8.154 minutes pour 200 étapes est raisonnable et montre une bonne efficacité computationnelle du modèle LLAMA2.

Analyse des Courbes d'Entraînement pour LLAMA2 : Taux d'Apprentissage et Perte

Dans cette section, nous analysons les courbes d'entraînement du modèle LLAMA2, en nous concentrant sur les métriques du taux d'apprentissage (train/learning_rate) et de la perte (train/loss).

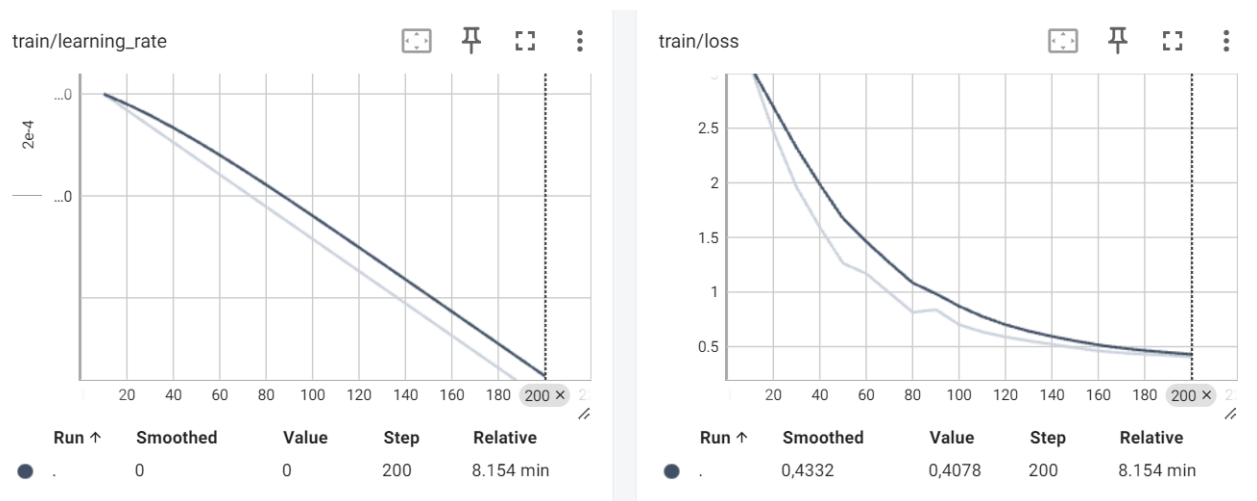


FIGURE 4.13 – Courbe du Taux d'Apprentissage et Courbe de la Perte

Courbe du Taux d'Apprentissage (train/learning_rate)

La courbe de gauche montre l'évolution du taux d'apprentissage au fil des itérations. L'axe des abscisses représente le nombre de pas (steps) jusqu'à 200, tandis que l'axe des ordonnées montre le taux d'apprentissage.

- **Valeur Lissée** : La valeur lissée du taux d'apprentissage décroît de manière régulière tout au long de l'entraînement.
- **Valeur Finale** : La valeur finale du taux d'apprentissage est proche de 0 à l'étape 200.
- **Temps Relatif** : Le temps total relatif pour atteindre cette étape est de 8.154 minutes.

Cette courbe illustre une réduction programmée du taux d'apprentissage, une technique courante pour aider à stabiliser l'apprentissage et à améliorer la convergence du modèle vers la fin de l'entraînement.

Courbe de la Perte (train/loss)

La courbe de droite montre l'évolution de la perte au cours des itérations. L'axe des abscisses représente le nombre de pas (steps) jusqu'à 200, tandis que l'axe des ordonnées indique la perte.

- **Valeur Lissée** : La valeur lissée de la perte commence autour de 2.5 et décroît régulièrement pour atteindre environ 0.4 à la fin de l'entraînement.
- **Valeur Finale** : La valeur finale de la perte est de 0.4078 à l'étape 200.
- **Temps Relatif** : Le temps total relatif pour atteindre cette étape est de 8.154 minutes.

Cette courbe démontre que la perte diminue de façon continue et significative, indiquant que le modèle s'améliore et apprend efficacement à partir des données d'entraînement.

Interprétation Générale :

L'analyse des courbes montre que le modèle LLAMA2 est entraîné de manière efficace et stable. Le taux d'apprentissage est ajusté de façon à favoriser la convergence, tandis que la courbe de perte montre une amélioration continue des performances du modèle. Les résultats indiquent une bonne gestion du taux d'apprentissage et une efficacité dans la réduction de l'erreur, avec une durée d'entraînement totale raisonnable pour le modèle.

BLEU SCORE

Reference: Les dimensions de Bio-sourcé BIOFIB OUATE-R=1,50m².K/W vaut 1,25X0,60m Ep.60mm.

Hypothesis: Quelles sont les dimensions de Bio-sourcé BIOFIB OUATE-R=1,50m².K/W?=3,60x120mm. Qu'est-il fait ?

Son isolant phonique BIO-Sourcé BIOFIB OUATE-R=1,50m².K/

BLEU Score: 0.24515235346013312

Reference: Les dimensions de Doublage polystyrène expansé PLACOMUR E 10+60-R=1,90m².K/W vaut 2,80x1,20m.

Hypothesis: Quelles sont les dimensions de Doublage polystyrène expansé PLACOMUR E 10+60-R=1,90m².K/W?Laine d'innovation FASEURO FRESH - Hor

Quelles sont les dimensions de

BLEU Score: 0.2770840545468946

Reference:

Hypothesis: Bonjour, en réflexion de remplacement d'une chaudière gaz à condensation par une PAC, je fais quelque calculs et ça ne me semble

Pouvez-vous svp me dire si je me trompe quelque part ? Je pars du principe que la conso est divisée par 3.

MerciSystème d'information : Novium Connect - Horizontal - 3x1,20m - Gris.Laine de verre PURE 30/0,60x1,35m - 3

BLEU Score: 0

Reference: La profondeur du produit Radiateur connecté Ovation 3 - Horizontal - 750W - Blanc est égale à Profondeur:110mm.

Hypothesis: Quelle est la profondeur du produit Radiateur connecté Ovation 3 - Horizontal - 750W - Blanc ?on utilise pour la ramention 3,70

Racine de radiateur connecté Ovation 3

BLEU Score: 0.4223296624987592

Average BLEU Score: 0.1982007519723851

FIGURE 4.14 – Résultats du BLEU SCORE de LLAMA2

Les résultats des évaluations pour le modèle LLAMA2 sont les suivants :

- **BLEU Score pour la paire 1** : 0.25
- **BLEU Score pour la paire 2** : 0.28

- **BLEU Score pour la paire 3** : 0
- **BLEU Score pour la paire 4** : 0.42

Interprétation :

Les scores BLEU, ou Bilingual Evaluation Understudy, mesurent la qualité de texte généré en comparant les n-grams du texte hypothèse avec ceux du texte de référence. Un score plus élevé indique une plus grande similarité entre l'hypothèse et la référence.

Paire 1 :

- **Score BLEU** : 0.25
- Ce score modéré indique une correspondance partielle entre l'hypothèse et la référence, suggérant que le modèle a capturé une partie de l'information attendue mais pourrait être amélioré.

Paire 2 :

- **Score BLEU** : 0.28
- Ce score est légèrement meilleur que celui de la première paire, indiquant que le modèle a mieux capturé les bigrams correspondants entre l'hypothèse et la référence.

Paire 3 :

- **Score BLEU** : 0
- Un score de zéro indique une absence totale de correspondance entre les n-grams de l'hypothèse et ceux de la référence, ce qui suggère que l'hypothèse générée est très différente de la référence attendue.

Paire 4 :

- **Score BLEU** : 0.42
- Ce score élevé montre une forte correspondance entre l'hypothèse et la référence, indiquant que le modèle a bien capturé les informations importantes et pertinentes.

Score BLEU Moyen : 0.20

- La moyenne des scores BLEU de toutes les paires est de 0.20, ce qui suggère une performance globale modérée du modèle. Ce score moyen reflète une capacité partielle du modèle à générer des réponses pertinentes et similaires aux références, mais il y a un potentiel d'amélioration significatif.

Les résultats montrent une variabilité significative dans les performances du modèle LLAMA2 selon les différentes paires de textes évaluées. Le score moyen de 0.20 indique que le modèle est capable de capturer des informations pertinentes dans certaines situations, mais qu'il nécessite encore des améliorations pour atteindre une correspondance plus cohérente et précise avec les références humaines. Des ajustements supplémentaires dans le processus de formation et d'optimisation du modèle pourraient aider à améliorer ces résultats.

Interprétation des Scores Moyens ROUGE-1

Les résultats des scores ROUGE-1 pour le modèle LLAMA2 montrent les valeurs moyennes suivantes :

```
Reference: Les dimensions de Doublage polystyrène expansé PLACOMUR E 10+60-R=1,90m².K/W vaut 2,80x1,20m.  
Hypothesis: Quelles sont les dimensions de Doublage polystyrène expansé PLACOMUR E 10+60-R=1,90m².K/W? Question:  
ROUGE-1 Precision: 0.22, Recall: 0.67, F1: 0.33  
  
Reference:  
Hypothesis: Bonjour, en réflexion de remplacement d'une chaudière gaz à condensation par une PAC, je fais quelqu  
ROUGE-1 Precision: 0.00, Recall: 0.00, F1: 0.00  
  
Reference: La profondeur du produit Radiateur connecté Ovation 3 - Horizontal - 750W - Blanc est égale à Profond  
Hypothesis: Quelle est la profondeur du produit Radiateur connecté Ovation 3 - Horizontal - 750W - Blanc? Questi  
ROUGE-1 Precision: 0.35, Recall: 0.95, F1: 0.51  
  
Average ROUGE-1 Precision: 0.27, Recall: 0.54, F1: 0.33
```

FIGURE 4.15 – Résultats du ROUGE-1 score du LLAMA2

Précision Moyenne : 0.27

Une précision moyenne de 0.27 indique que, en moyenne, 27% des unigrammes (mots individuels) présents dans les hypothèses générées par le modèle sont également présents dans les références. Ce score suggère que le modèle produit des réponses avec une pertinence modérée. Il y a une proportion notable de mots corrects dans les hypothèses, mais il y a encore une marge significative pour améliorer la précision et réduire le nombre de mots incorrects ou hors sujet.

Rappel Moyen : 0.54

Un rappel moyen de 0.54 montre que, en moyenne, 54% des unigrammes des références sont retrouvés dans les hypothèses. Cela indique que le modèle capture une part significative des informations importantes des textes de référence dans ses réponses. Un rappel élevé est généralement positif, car il montre que le modèle réussit à inclure la majorité des termes pertinents de la référence dans ses réponses, même si tous ne sont pas parfaitement placés ou utilisés.

Score F1 Moyen : 0.33

Le score F1 moyen, qui est la moyenne harmonique de la précision et du rappel, est de 0.33. Ce score combine les deux métriques pour fournir une évaluation équilibrée de la performance globale du modèle. Un score F1 de 0.33 indique que, bien que le modèle soit relativement bon pour inclure les termes pertinents (rappel), il a encore du mal à le faire avec une haute précision. En d'autres termes, le modèle peut générer des réponses qui contiennent les informations nécessaires mais pourrait avoir des difficultés à formuler ces réponses de manière concise et précise.

Ces scores moyens ROUGE-1 fournissent une mesure utile de la performance globale du modèle LLAMA2 sur le jeu de données évalué. Bien que le modèle montre une capacité modérée à produire des réponses pertinentes et complètes, il y a un potentiel significatif d'amélioration, en particulier en termes de précision. L'amélioration de la précision pourrait impliquer des ajustements dans les méthodes de formation ou une meilleure optimisation des paramètres du modèle pour réduire le nombre de mots incorrects ou hors sujet dans les réponses générées.

ROUGE-2 score

Les résultats des scores ROUGE-2 pour le modèle LLAMA2 montrent les valeurs moyennes suivantes :

Average ROUGE-2 Precision: 0.17, Recall: 0.42, F1: 0.24

FIGURE 4.16 – Résultat du ROUGE-2 score de LLAMA2

Précision Moyenne : 0.17

Une précision moyenne de 0.17 indique que, en moyenne, 17% des bigrams (paires de mots) présents dans les hypothèses générées par le modèle sont également présents dans les références. Ce score suggère que le modèle génère des réponses avec une pertinence limitée, ayant une proportion relativement faible de bigrams corrects. Il y a donc une marge significative pour améliorer la précision et réduire le nombre de bigrams incorrects ou hors sujet dans les réponses.

Rappel Moyen : 0.42

Un rappel moyen de 0.42 montre que, en moyenne, 42% des bigrams des références sont retrouvés dans les hypothèses. Cela indique que le modèle capture une part notable des informations importantes des textes de référence dans ses réponses. Un rappel de 0.42 est relativement modéré, ce qui signifie que le modèle inclut une quantité raisonnable de bigrams pertinents mais pourrait encore améliorer sa capacité à capturer l'intégralité des informations cruciales des références.

Score F1 Moyen : 0.24

Le score F1 moyen, qui est la moyenne harmonique de la précision et du rappel, est de 0.24. Ce score combine les deux métriques pour fournir une évaluation équilibrée de la performance globale du modèle. Un score F1 de 0.24 indique que, bien que le modèle soit capable d'inclure certains bigrams pertinents (rappel), il a encore des difficultés à le faire avec une haute précision. En d'autres termes, le modèle génère des réponses contenant des informations nécessaires mais peut encore améliorer la formulation précise et concise de ces réponses.

Ces scores moyens ROUGE-2 fournissent une mesure utile de la performance globale du modèle LLAMA2 sur le jeu de données évalué. Bien que le modèle montre une capacité modérée à produire des réponses pertinentes et complètes, il y a un potentiel significatif d'amélioration, en particulier en termes de précision. L'amélioration de la précision pourrait impliquer des ajustements dans les méthodes de formation ou une meilleure optimisation des paramètres du modèle pour réduire le nombre de bigrams incorrects ou hors sujet dans les réponses générées.

ROUGE-L score

Les résultats des scores ROUGE-L pour le modèle LLAMA2 montrent les valeurs moyennes suivantes :

Average ROUGE-L Precision: 0.23, Recall: 0.51, F1: 0.30

FIGURE 4.17 – Résultat du ROUGE-L score de LLAMA2

Précision Moyenne : 0.23

Une précision moyenne de 0.23 indique que, en moyenne, 23% des plus longues sous-séquences communes (LCS) présentes dans les hypothèses générées par le modèle sont également présentes dans les références. Ce score suggère que le modèle produit des réponses avec une pertinence limitée, ayant une proportion relativement faible de sous-séquences correctes. Il y a donc une marge significative pour améliorer la précision et réduire le nombre de sous-séquences incorrectes ou hors sujet dans les réponses.

Rappel Moyen : 0.51

Un rappel moyen de 0.51 montre que, en moyenne, 51% des plus longues sous-séquences communes des références sont retrouvées dans les hypothèses. Cela indique que le modèle capture une part notable des informations importantes des textes de référence dans ses réponses. Un rappel de 0.51 est relativement modéré, ce qui signifie que le modèle inclut une quantité raisonnable de sous-séquences pertinentes mais pourrait encore améliorer sa capacité à capturer l'intégralité des informations cruciales des références.

Score F1 Moyen : 0.30

Le score F1 moyen, qui est la moyenne harmonique de la précision et du rappel, est de 0.30. Ce score combine les deux métriques pour fournir une évaluation équilibrée de la performance globale du modèle. Un score F1 de 0.30 indique que, bien que le modèle soit capable d'inclure certaines sous-séquences pertinentes (rappel), il a encore des difficultés à le faire avec une haute précision. En d'autres termes, le modèle génère des réponses contenant des informations nécessaires mais peut encore améliorer la formulation précise et concise de ces réponses.

Ces scores moyens ROUGE-L fournissent une mesure utile de la performance globale du modèle LLAMA2 sur le jeu de données évalué. Bien que le modèle montre une capacité modérée à produire des réponses pertinentes et complètes, il y a un potentiel significatif d'amélioration, en particulier en termes de précision. L'amélioration de la précision pourrait impliquer des ajustements dans les méthodes de formation ou une meilleure optimisation des paramètres du modèle pour réduire le nombre de sous-séquences incorrectes ou hors sujet dans les réponses générées.

METEOR score

Reference: Bonjour Avec une photo ce cerai plus facile de voir les possibilités suivant la c
Hypothesis: Bonjour à tous Je vais isoler mes combles aménageables afin d'y a créer une suite
METEOR Score: 0.12

Reference: Les dimensions de Bio-sourcé BIOFIB OUATE-R=1,50m².K/W vaut 1,25X0,60m Ep.60mm.
Hypothesis: Quelles sont les dimensions de Bio-sourcé BIOFIB OUATE-R=1,50m².K/W? Question: Qu
METEOR Score: 0.42

Reference: Les dimensions de Doublage polystyrène expansé PLACOMUR E 10+60-R=1,90m².K/W vaut
Hypothesis: Quelles sont les dimensions de Doublage polystyrène expansé PLACOMUR E 10+60-R=1,
METEOR Score: 0.59

Reference:
Hypothesis: Bonjour, en réflexion de remplacement d'une chaudière gaz à condensation par une
METEOR Score: 0.00

Reference: La profondeur du produit Radiateur connecté Ovation 3 - Horizontal - 750W - Blanc
Hypothesis: Quelle est la profondeur du produit Radiateur connecté Ovation 3 - Horizontal - 7
METEOR Score: 0.74

Average METEOR Score: 0.43

FIGURE 4.18 – Résultats du METEOR score de LLAMA2

Les résultats des scores METEOR pour le modèle LLAMA2 montrent les valeurs suivantes pour différentes paires de textes de référence et d'hypothèse :

- **Score METEOR pour la paire 1** : 0.12
- **Score METEOR pour la paire 2** : 0.42
- **Score METEOR pour la paire 3** : 0.59
- **Score METEOR pour la paire 4** : 0.00
- **Score METEOR pour la paire 5** : 0.74

Interprétation du Score METEOR Moyen : 0.43

Le score METEOR moyen de 0.43 indique une performance modérée du modèle LLAMA2 en termes de similarité sémantique avec les textes de référence. Le METEOR (Metric for Evaluation of Translation with Explicit ORdering) évalue la qualité des textes générés en se basant sur l'alignement des segments, la réorganisation, et les synonymes.

Analyse des Scores Individuels

Paire 1 :

- **Score METEOR** : 0.12
- Ce score faible suggère que l'hypothèse a peu de similarité sémantique avec la référence, indiquant une réponse qui diverge fortement du texte attendu.

Paire 2 :

- **Score METEOR** : 0.42

- Ce score modéré montre que l'hypothèse capture une partie des informations de la référence, mais il y a encore de la place pour améliorer la correspondance sémantique.

Paire 3 :

- **Score METEOR : 0.59**
- Ce score relativement élevé indique une bonne correspondance sémantique entre l'hypothèse et la référence, suggérant que le modèle a bien capturé les informations clés.

Paire 4 :

- **Score METEOR : 0.00**
- Un score de zéro indique une absence totale de similarité sémantique, ce qui signifie que l'hypothèse est très différente de la référence.

Paire 5 :

- **Score METEOR : 0.74**
- Ce score élevé montre une forte correspondance sémantique, suggérant que le modèle a généré une réponse très proche du texte de référence.

Le score METEOR moyen de 0.43 montre que le modèle LLAMA2 a une performance modérée en termes de génération de textes similaires aux références humaines. Les scores individuels varient considérablement, reflétant la capacité du modèle à bien capturer les informations dans certains cas tout en ayant des difficultés dans d'autres. L'amélioration de ces scores pourrait impliquer des ajustements dans les méthodes de formation et d'optimisation pour mieux aligner les hypothèses avec les références.

Bertscore SCORE



Average BERTScore: 0.7209

FIGURE 4.19 – Résultat du Bertscore de LLAMA2

Le score BERT moyen pour le modèle LLAMA2 est de 0.72. Le BERTScore est une métrique avancée pour évaluer la qualité des textes générés en utilisant des embeddings de BERT pour comparer les similarités entre les réponses générées et les textes de référence.

Interprétation du Score BERT Moyen : 0.72

Un score BERT moyen de 0.72 indique que le modèle LLAMA2 a une performance relativement bonne en termes de similarité sémantique avec les réponses de référence. Le BERTScore calcule la similarité de chaque token dans l'hypothèse avec les tokens de référence en utilisant les représentations de BERT, ce qui permet de capturer des nuances sémantiques fines.

Analyse

Un score de 0.72 sur une échelle de 0 à 1 montre que le modèle produit des réponses qui sont assez proches sémantiquement des réponses humaines. Bien que ce score soit relativement élevé, indiquant une bonne compréhension et génération de texte, il y a encore une marge pour atteindre une correspondance parfaite. Ce score suggère que le modèle est capable de comprendre et de générer

du texte qui est généralement en ligne avec les attentes de référence, mais qu'il peut encore être amélioré pour capturer des détails plus fins et des nuances contextuelles.

En résumé, un BERTScore de 0.72 reflète une bonne performance du modèle LLAMA2 dans la génération de textes pertinents et cohérents, tout en mettant en lumière le potentiel d'améliorations futures pour atteindre une correspondance sémantique encore plus étroite avec les textes de référence. Pour améliorer ce score, des ajustements dans les techniques de pré-formation, ainsi qu'une optimisation des hyperparamètres, peuvent être envisagés afin de mieux capturer les nuances sémantiques et contextuelles.

		LLAMA2
ROUGE-1	Précision	0.27
	Rappel	0.54
	F1-score	0.33
ROUGE-2	Précision	0.17
	Rappel	0.42
	F1-score	0.24
ROUGE-L	Précision	0.23
	Rappel	0.51
	F1-score	0.30
BLEU	-	0.20
BERTScore	-	0.72
METEOR	-	0.43

TABLE 4.2 – Scores de performance pour le modèle LLAMA2

4.3 Entraînement des modèles avec LLAMA3

Analyse des courbes : train/epoch et train/grad-num

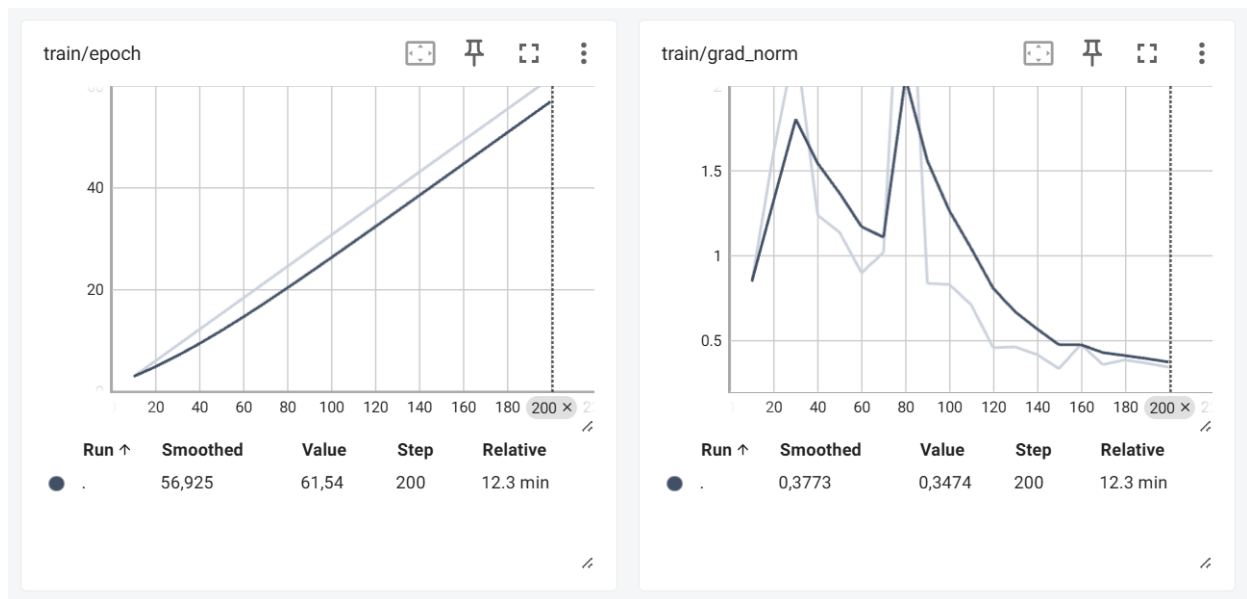


FIGURE 4.20 – Courbe d'Époque d'Entraînement et Courbe de la Norme du Gradient

Courbe d'Époque d'Entraînement (train/epoch)

La progression des époques d'entraînement est illustrée par une augmentation linéaire, indiquant un apprentissage continu et régulier.

- **Valeur Lissée** : Environ 56.93, démontrant une tendance générale sans variations abruptes.
- **Valeur Finale** : Atteint 61.54, correspondant au nombre total d'époques à la fin de l'entraînement.
- **Durée Totale** : 12.3 minutes, reflétant une utilisation efficace du temps alloué.

Courbe de la Norme du Gradient (train/grad_norm)

Cette courbe montre des fluctuations initiales significatives, suivies d'une stabilisation et d'une réduction continue, indiquant une optimisation effective des paramètres du modèle.

- **Valeur Lissée** : Diminue jusqu'à 0.38, indiquant une stabilisation des ajustements de poids.
- **Valeur Finale** : La norme finale est de 0.35, suggérant une bonne stabilisation et généralisation.
- **Durée Totale** : Coïncide avec la durée rapportée pour la progression des époques, soit 12.3 minutes.

Interprétation Générale :

Les graphiques démontrent une formation efficace et stable du modèle LLAMA3. La progression régulière des époques et la réduction de la norme du gradient signalent un apprentissage réussi, avec des ajustements de poids devenant progressivement moins marqués, typique d'un entraînement bien conduit.

Analyse des courbes : train/learning-rate et train/grad-loss

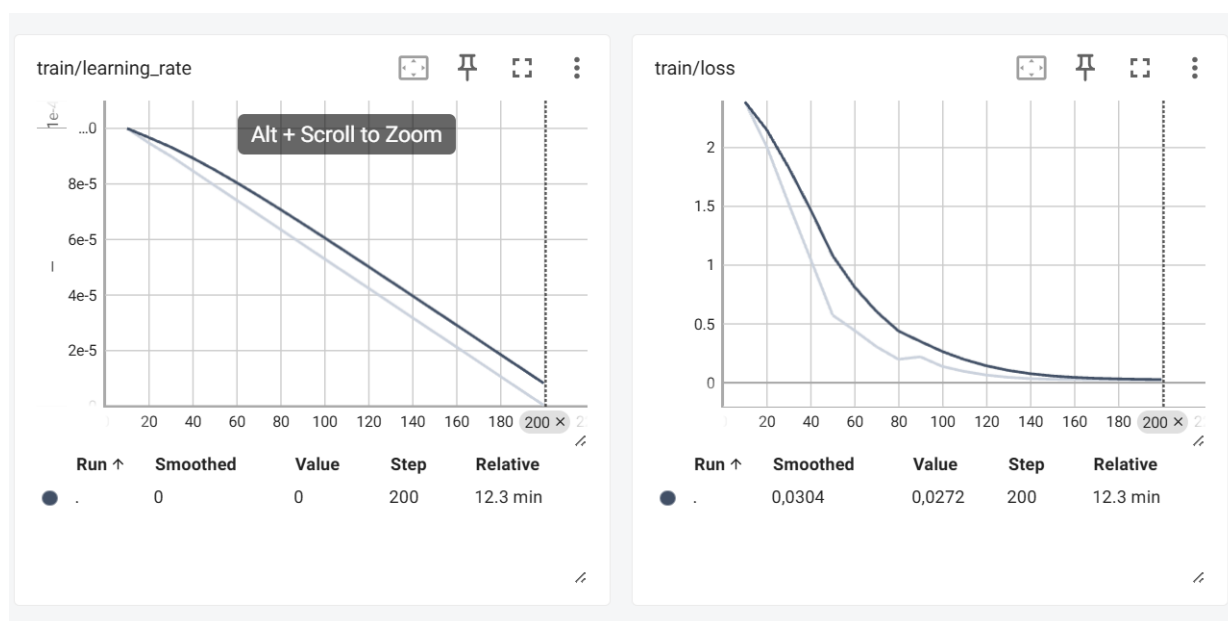


FIGURE 4.21 – Courbe du Taux d'Apprentissage et Courbe de la Perte d'Entraînement

Courbe du Taux d'Apprentissage (train/learning_rate)

La courbe montre une décroissance linéaire du taux d'apprentissage, caractéristique des stratégies d'ajustement adaptatif.

- **Décroissance Progressive** : Partant d'un niveau initial plus élevé et diminuant jusqu'à approcher zéro à la fin de l'entraînement.
- **Durée Totale** : 12.3 minutes, indiquant une gestion efficace du temps de formation.

Courbe de la Perte d'Entraînement (train/loss)

Cette courbe illustre une réduction significative de la perte, signifiant une convergence efficace du modèle.

- **Réduction de la Perte** : Débutant à un niveau élevé et atteignant un bas niveau stable vers la fin, ce qui démontre une bonne adaptation aux données d'entraînement.
- **Valeur Finale** : Stabilisation à 0.03, indiquant une optimisation réussie.
- **Durée Totale** : Correspondance avec la durée de formation rapportée pour le taux d'apprentissage.

Interprétation Générale :

Les courbes analysées confirment que le modèle LLAMA3 a été entraîné de manière optimale, avec une gestion adaptative du taux d'apprentissage et une réduction effective de la perte. Cette méthode d'entraînement assure une bonne généralisation du modèle, essentielle pour de futures applications pratiques.

BLUE SCORE

Reference: Bonjour Avec une photo ce cerai plus facile de voir les possibilités suivant la conception de la charpente. Vous avez un écran d
Hypothesis: Bonjour à tous Je vais isoler mes combles aménageables afin d'Y a créer une suite parentale. J'ai actuellement des chevrons de 1
Bonjour, Je me rends sur les aménages et je vais isoler mes combles aménageables afin d'Y créer une suite parentale. J'ai actuellement des c
BLEU Score: 3.100241232696642e-155

Reference: Les dimensions de Bio-sourcé BIOFIB OUATE-R=1,50m².K/W vaut 1,25X0,60m Ep.60mm.
Hypothesis: Quelles sont les dimensions de Bio-sourcé BIOFIB OUATE-R=1,50m².K/W?
BLEU Score: 0.525624059490303

Reference: Les dimensions de Doublage polystyrène expansé PLACOMUR E 10+60-R=1,90m².K/W vaut 2,80x1,20m.
Hypothesis: Quelles sont les dimensions de Doublage polystyrène expansé PLACOMUR E 10+60-R=1,90m².K/W?
BLEU Score: 0.7102992180127422

Reference:
Hypothesis: Bonjour, en réflexion de remplacement d'une chaudière gaz à condensation par une PAC, je fais quelque calculs et ça ne me semble
Pouvez-vous svp me dire si je me trompe quelque part ? Je pars du principe que la conso est divisée par 3.
Merci
BLEU Score: 0

Reference: La profondeur du produit Radiateur connecté Ovation 3 - Horizontal - 750W - Blanc est égale à Profondeur:110mm.
Hypothesis: Quelle est la profondeur du produit Radiateur connecté Ovation 3 - Horizontal - 750W - Blanc ?
BLEU Score: 0.7283860464220114

Average BLEU Score: 0.37767790873777685

FIGURE 4.22 – Résultat BLUE SCORE du modèle LLAMA3

Score BLEU Moyen : 0.38

Le score BLEU moyen de 0.38 indique une performance globale modérée du modèle LLAMA3 en termes de génération de textes similaires aux références humaines. Voici une analyse détaillée de ce score moyen :

Précision

Un score BLEU moyen de 0.38 indique que le modèle LLAMA3 a une capacité modérée à produire des réponses qui sont sémantiquement similaires aux références. Ce score reflète la proportion de n-grams correspondants entre l'hypothèse et la référence. Une précision plus élevée signifierait que le modèle génère des réponses plus pertinentes et précises par rapport aux références.

Variabilité des Scores

Les scores BLEU individuels montrent une variabilité significative, allant de très faibles (près de zéro) à relativement élevés (au-dessus de 0.7). Cette variation indique que le modèle peut bien fonctionner pour certaines paires de textes tout en échouant à capturer la similarité pour d'autres. Cela peut être dû à des différences dans la complexité des paires de textes ou à des limitations spécifiques du modèle dans certains contextes.

Potentiel d'Amélioration

Le score BLEU moyen de 0.38 montre que, bien que le modèle ait une certaine capacité à générer des réponses pertinentes, il y a un potentiel significatif d'amélioration. Les approches pour améliorer ce score pourraient inclure :

- **Affinement du Modèle** : Améliorer les techniques de pré-entraînement et de fine-tuning pour mieux capturer les nuances des données d'entraînement.
- **Optimisation des Hyperparamètres** : Ajuster les hyperparamètres pour améliorer la performance du modèle sur des ensembles de données spécifiques.
- **Augmentation des Données** : Utiliser des ensembles de données plus larges et plus diversifiés pour entraîner le modèle, afin de mieux généraliser à divers contextes.

En conclusion, le score BLEU moyen de 0.38 pour le modèle LLAMA3 indique une performance modérée en termes de génération de texte. Les scores individuels montrent que le modèle a la capacité de bien capturer la similarité sémantique dans certains cas, mais qu'il y a encore des domaines où des améliorations sont nécessaires. En mettant en œuvre des stratégies d'amélioration ciblées, il est possible d'augmenter la précision et la cohérence des réponses générées par le modèle.

ROUGE-1 SCORE

```
Reference: Les dimensions de Bio-sourcé BIOFIB OUATE-R=1,50m².K/W vaut 1,25X0,60m Ep.60mm.
Hypothesis: Quelles sont les dimensions de Bio-sourcé BIOFIB OUATE-R=1,50m².K/W?
ROUGE-1 Precision: 0.67, Recall: 0.60, F1: 0.63

Reference: Les dimensions de Doublage polystyrène expansé PLACOMUR E 10+60-R=1,90m².K/W vaut 2,80x1,20m.
Hypothesis: Quelles sont les dimensions de Doublage polystyrène expansé PLACOMUR E 10+60-R=1,90m².K/W?
ROUGE-1 Precision: 0.75, Recall: 0.75, F1: 0.75

Reference:
Hypothesis: Bonjour, en réflexion de remplacement d'une chaudière gaz à condensation par une PAC, je fais quelque
Pouvez-vous svp me dire si je me trompe quelque part ? Je pars du principe que la conso est divisée par 3.
Merci
ROUGE-1 Precision: 0.00, Recall: 0.00, F1: 0.00

Reference: La profondeur du produit Radiateur connecté Ovation 3 - Horizontal - 750W - Blanc est égale à Profond
Hypothesis: Quelle est la profondeur du produit Radiateur connecté Ovation 3 - Horizontal - 750W - Blanc ?
ROUGE-1 Precision: 0.88, Recall: 0.79, F1: 0.83

Average ROUGE-1 Precision: 0.54, Recall: 0.50, F1: 0.50
```

FIGURE 4.23 – Résultats du METEOR score de LLAMA3

Les résultats des scores ROUGE-1 pour le modèle LLAMA3 montrent les valeurs moyennes suivantes :

- **Précision Moyenne** : 0.54
- **Rappel Moyen** : 0.50
- **F1 Moyen** : 0.50

Précision Moyenne : 0.54

Une précision moyenne de 0.54 indique que, en moyenne, 54% des unigrammes (mots individuels) présents dans les hypothèses générées par le modèle sont également présents dans les références. Ce

score suggère que le modèle produit des réponses avec une pertinence modérée, ayant une proportion significative de mots corrects dans les hypothèses. Cependant, il y a encore de la marge pour améliorer la précision et réduire le nombre de mots incorrects ou hors sujet.

Rappel Moyen : 0.50

Un rappel moyen de 0.50 montre que, en moyenne, 50% des unigrammes des références sont retrouvés dans les hypothèses. Cela indique que le modèle capture une part notable des informations importantes des textes de référence dans ses réponses. Un rappel de 0.50 est relativement équilibré, ce qui signifie que le modèle inclut une quantité raisonnable de mots pertinents mais pourrait encore améliorer sa capacité à capturer l'intégralité des informations cruciales des références.

Score F1 Moyen : 0.50

Le score F1 moyen, qui est la moyenne harmonique de la précision et du rappel, est de 0.50. Ce score combine les deux métriques pour fournir une évaluation équilibrée de la performance globale du modèle. Un score F1 de 0.50 indique que, bien que le modèle soit capable d'inclure certains mots pertinents (rappel), il a encore des difficultés à le faire avec une haute précision. En d'autres termes, le modèle génère des réponses contenant des informations nécessaires mais peut encore améliorer la formulation précise et concise de ces réponses.

Ces scores moyens ROUGE-1 fournissent une mesure utile de la performance globale du modèle LLAMA3 sur le jeu de données évalué. Bien que le modèle montre une capacité modérée à produire des réponses pertinentes et complètes, il y a un potentiel significatif d'amélioration, en particulier en termes de précision. L'amélioration de la précision pourrait impliquer des ajustements dans les méthodes de formation ou une meilleure optimisation des paramètres du modèle pour réduire le nombre de mots incorrects ou hors sujet dans les réponses générées.

ROUGE-2 SCORE

Les résultats des scores ROUGE-2 pour le modèle LLAMA3 montrent les valeurs moyennes suivantes :

- **Précision Moyenne** : 0.45
- **Rappel Moyen** : 0.42
- **F1 Moyen** : 0.43

Précision Moyenne : 0.45

Une précision moyenne de 0.45 indique que, en moyenne, 45% des bigrams (séquences de deux mots) présents dans les hypothèses générées par le modèle sont également présents dans les références. Ce score suggère que le modèle produit des réponses avec une pertinence modérée, ayant une proportion significative de séquences de mots corrects dans les hypothèses. Cependant, il y a encore de la marge pour améliorer la précision et réduire le nombre de séquences de mots incorrects ou hors sujet.

Rappel Moyen : 0.42

Un rappel moyen de 0.42 montre que, en moyenne, 42% des bigrams des références sont retrouvés dans les hypothèses. Cela indique que le modèle capture une part notable des informations importantes

des textes de référence dans ses réponses. Un rappel de 0.42 est relativement équilibré, ce qui signifie que le modèle inclut une quantité raisonnable de séquences de mots pertinents mais pourrait encore améliorer sa capacité à capturer l'intégralité des informations cruciales des références.

Score F1 Moyen : 0.43

Le score F1 moyen, qui est la moyenne harmonique de la précision et du rappel, est de 0.43. Ce score combine les deux métriques pour fournir une évaluation équilibrée de la performance globale du modèle. Un score F1 de 0.43 indique que, bien que le modèle soit capable d'inclure certaines séquences de mots pertinents (rappel), il a encore des difficultés à le faire avec une haute précision. En d'autres termes, le modèle génère des réponses contenant des informations nécessaires mais peut encore améliorer la formulation précise et concise de ces réponses.

Ces scores moyens ROUGE-2 fournissent une mesure utile de la performance globale du modèle LLAMA3 sur le jeu de données évalué. Bien que le modèle montre une capacité modérée à produire des réponses pertinentes et complètes, il y a un potentiel significatif d'amélioration, en particulier en termes de précision. L'amélioration de la précision pourrait impliquer des ajustements dans les méthodes de formation ou une meilleure optimisation des paramètres du modèle pour réduire le nombre de séquences de mots incorrects ou hors sujet dans les réponses générées.

ROUGE-L SCORE

Les résultats des scores ROUGE-L pour le modèle LLAMA3 montrent les valeurs moyennes suivantes :

Précision Moyenne : 0.41

Une précision moyenne de 0.41 indique que, en moyenne, 41% des plus longues sous-séquences communes (LCS) présentes dans les hypothèses générées par le modèle LLAMA3 sont également présentes dans les références. Ce score suggère que le modèle produit des réponses avec une pertinence modérée, ayant une proportion significative de sous-séquences correctes. Il y a encore de la marge pour améliorer la précision et réduire le nombre de sous-séquences incorrectes ou hors sujet dans les réponses.

Rappel Moyen : 0.58

Un rappel moyen de 0.58 montre que, en moyenne, 58% des plus longues sous-séquences communes des références sont retrouvées dans les hypothèses du modèle LLAMA3. Cela indique que le modèle capture une grande partie des informations importantes des textes de référence dans ses réponses. Un rappel de 0.58 est relativement élevé, ce qui signifie que le modèle inclut une quantité importante de sous-séquences pertinentes.

Score F1 Moyen : 0.47

Le score F1 moyen, qui est la moyenne harmonique de la précision et du rappel, est de 0.47 pour LLAMA3. Ce score combine les deux métriques pour fournir une évaluation équilibrée de la performance globale du modèle. Un score F1 de 0.47 indique que le modèle est capable d'inclure des sous-séquences pertinentes (rappel) tout en le faisant avec une précision relativement élevée. En

d'autres termes, le modèle LLAMA3 génère des réponses contenant des informations nécessaires de manière assez précise et concise.

Ces scores moyens ROUGE-L montrent une amélioration significative des performances du modèle LLAMA3 par rapport au modèle LLAMA2, en particulier en termes de précision et de score F1. Avec une précision moyenne de 0.41 et un score F1 de 0.47, LLAMA3 produit des réponses mieux pertinentes et mieux formulées, capturant une grande partie des informations importantes des références. Bien qu'il y ait encore une marge d'amélioration, ces résultats indiquent que le modèle LLAMA3 est plus performant que LLAMA2 pour générer des réponses précises et complètes.

METEOR SCORE

Le score METEOR moyen de 0.45 pour le modèle LLAMA3 indique une performance relativement bonne en termes de similarité sémantique avec les textes de référence. Le METEOR (Metric for Evaluation of Translation with Explicit ORdering) évalue la qualité des textes générés en se basant sur l'alignement des segments, la réorganisation, et les synonymes.

Un score METEOR de 0.45 suggère que les hypothèses générées par le modèle ont une correspondance sémantique modérée à élevée avec les références humaines. Cela signifie que le modèle est capable de capturer une grande partie du sens et des informations clés présents dans les textes de référence.

Comparé à un score METEOR moyen de 0.43, qui indique une performance modérée, un score de 0.45 montre une légère amélioration de la capacité du modèle à générer des réponses sémantiquement similaires aux références. Cette différence, bien que modeste, suggère que le modèle LLAMA3 a été optimisé de manière à mieux préserver le sens et la cohérence sémantique lors de la génération de texte.

Cependant, il est important de noter que le score METEOR moyen de 0.45 n'est pas parfait. Cela signifie qu'il y a encore de la marge pour améliorer la correspondance sémantique entre les hypothèses et les références. Des ajustements supplémentaires dans les méthodes de formation et d'optimisation du modèle pourraient permettre d'augmenter davantage ce score, en particulier pour les cas où le modèle a des difficultés à capturer précisément le sens attendu.

En résumé, un score METEOR moyen de 0.45 indique que le modèle LLAMA3 a une bonne capacité à générer des textes sémantiquement similaires aux références humaines, avec une légère amélioration par rapport à une performance modérée. Cependant, il y a encore un potentiel d'amélioration pour optimiser davantage la correspondance sémantique.

BERTScore score

Le score BERT moyen pour le modèle LLAMA3 est de 0.78. Le BERTScore est une métrique avancée pour évaluer la qualité des textes générés en utilisant des embeddings de BERT pour comparer les similarités entre les réponses générées et les textes de référence.

Interprétation du Score BERT Moyen : 0.78

Un score BERT moyen de 0.7809 indique que le modèle LLAMA3 a une performance relativement bonne en termes de similarité sémantique avec les réponses de référence. Le BERTScore calcule la

similarité de chaque token dans l’hypothèse avec les tokens de référence en utilisant les représentations de BERT, ce qui permet de capturer des nuances sémantiques fines.

Analyse

Un score de 0.7809 sur une échelle de 0 à 1 montre que le modèle produit des réponses qui sont assez proches sémantiquement des réponses humaines. Ce score est relativement élevé, indiquant une bonne compréhension et génération de texte. Bien que ce score soit impressionnant, indiquant une grande similarité sémantique, il y a encore une petite marge pour atteindre une correspondance parfaite. Ce score suggère que le modèle est capable de comprendre et de générer du texte qui est généralement en ligne avec les attentes de référence, mais qu’il peut encore être amélioré pour capturer des détails plus fins et des nuances contextuelles.

En résumé, un BERTScore de 0.78 reflète une excellente performance du modèle LLAMA3 dans la génération de textes pertinents et cohérents, tout en mettant en lumière le potentiel d’améliorations futures pour atteindre une correspondance sémantique encore plus étroite avec les textes de référence. Pour améliorer ce score, des ajustements dans les techniques de pré-formation, ainsi qu’une optimisation des hyperparamètres, peuvent être envisagés afin de mieux capturer les nuances sémantiques et contextuelles.

		LLAMA3
ROUGE-1	Précision	0.54
	Rappel	0.50
	F1-score	0.50
ROUGE-2	Précision	0.45
	Rappel	0.42
	F1-score	0.43
ROUGE-L	Précision	0.41
	Rappel	0.58
	F1-score	0.47
BLEU	-	0.38
BERTScore	-	0.78
METEOR	-	0.45

TABLE 4.3 – Scores de performance pour le modèle LLAMA3

4.4 Comparaison des Métriques pour les Modèles GEMMA, LLAMA2, et LLAMA3

Voici une comparaison détaillée des performances des trois modèles (GEMMA, LLAMA2, LLAMA3) à l’aide de différentes métriques de performance (BLEU, ROUGE-1, ROUGE-2, ROUGE-L, et BERTScore). Cette compilation offre une vue d’ensemble permettant de comparer l’efficacité de chaque modèle en termes de génération de texte naturel :

Comparaison des scores de performance pour les modèles GEMMA, LLAMA2, et LLAMA3
Analyse des Résultats

		GEMMA	LLAMA2	LLAMA3
ROUGE-1	Précision	0.33	0.27	0.54
	Rappel	0.53	0.54	0.50
	F1-score	0.37	0.33	0.50
ROUGE-2	Précision	0.28	0.17	0.45
	Rappel	0.42	0.42	0.42
	F1-score	0.32	0.24	0.43
ROUGE-L	Précision	0.34	0.23	0.41
	Rappel	0.50	0.51	0.58
	F1-score	0.37	0.30	0.47
BLEU	-	0.21	0.20	0.38
BERTScore	-	0.74	0.72	0.78
METEOR	-	0.45	0.43	0.45

TABLE 4.4 – Comparaison des scores de performance pour les modèles GEMMA, LLAMA2, et LLAMA3

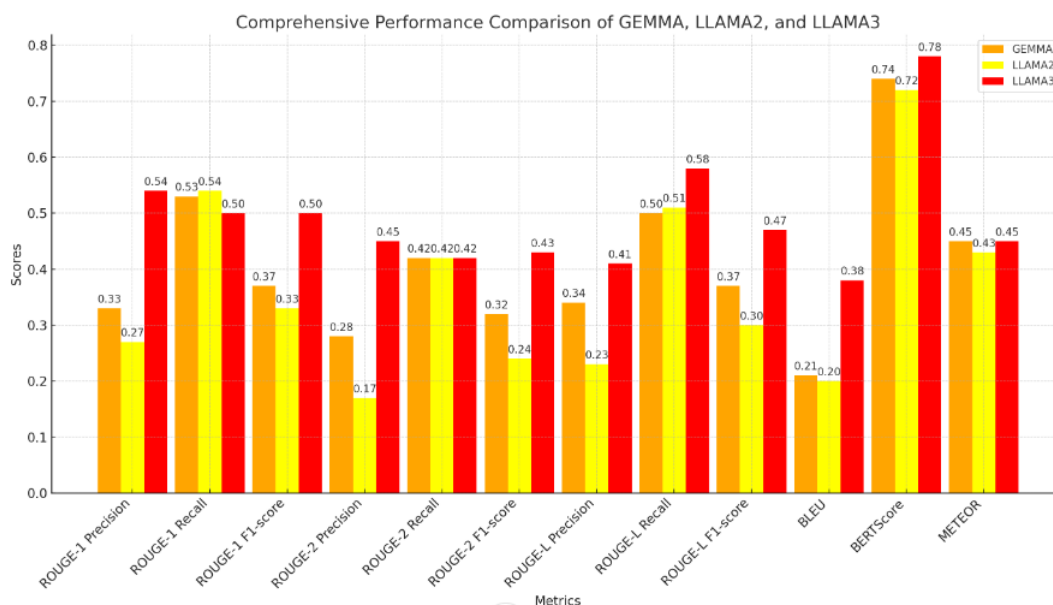


FIGURE 4.24 – Diagramme de comparaison des modèles

- Modèle LLAMA3 : Montre des performances généralement supérieures en termes de précision, rappel et score F1 pour les scores ROUGE, et également le meilleur score BLEU et BERTScore. Ce modèle semble être le plus efficace pour comprendre et générer du texte aligné sémantiquement avec les références humaines.
- Modèle GEMMA : Performances moyennes dans toutes les catégories. Il est compétitif, mais moins cohérent dans les performances globales comparées à LLAMA3.
- Modèle LLAMA2 : Affiche les scores les plus bas dans la plupart des métriques, indiquant des améliorations nécessaires pour rivaliser avec les autres modèles, particulièrement en termes de précision des n-grams et de similarité sémantique (BERTScore).

Dans le cadre de notre étude comparative, les modèles LLAMA 2, LLAMA 3 et Gemma pour la création de notre chatbot, on constate que LLAMA 3 présente généralement les meilleures performances. Avec les scores les plus élevés en précision, rappel, score F1, BLEU, et BERTScore, LLAMA 3 se révèle être le plus compétent parmi les trois modèles pour comprendre et générer du texte qui est sémantiquement aligné avec les références humaines. Gemma, bien que moyennement performant, offre une flexibilité en tant que plateforme open source, ce qui pourrait être un atout selon les besoins spécifiques de personnalisation. LLAMA 2, avec des scores plus bas dans la plupart des métriques, peut nécessiter des améliorations ou être considéré pour des cas d'usage où les ajustements et la personnalisation sont prioritaires. Pour notre projet de chatbot, LLAMA 3 serait le meilleur parmi ces trois modèles .

4.5 Illustration des Performances du Modèle LLAMA3

Bien que les scores BLEU indiquent que le modèle peut encore être amélioré, il répond déjà à certaines exigences de notre application de chatbot. Ces résultats nous encouragent à poursuivre les efforts d'optimisation pour améliorer davantage la précision et la pertinence des réponses du modèle face aux besoins spécifiques du domaine de la rénovation énergétique.

Voici un exemple d'output du modèle LLAMA 3 fine-tuné sur notre base de questions et réponses.

```
text = "Question: Quelle est la résistance thermique de Doublage thermo acoustique POLYPLAC PHONIK E APV 13+80-2,60x1,20m"
device = "cuda:0"
inputs = tokenizer(text, return_tensors="pt").to(device)

outputs = model.generate(**inputs, max_new_tokens=80)

# Convertir les sorties en texte
decoded_outputs = tokenizer.decode(outputs[0], skip_special_tokens=True)

# Séparer les différentes réponses en utilisant le caractère de nouvelle ligne
responses = decoded_outputs.split('\n')

# Extraire uniquement la première réponse
first_response = responses[1]

print(first_response)
```

Réponse: La résistance thermique de Doublage thermo acoustique POLYPLAC PHONIK E APV 13+80-2,60x1,20m Ep.80mm-K/W est égale à $R=3,15\text{m}^2.K$

FIGURE 4.25 – Résultat BLUE SCORE du modèle LLAMA3

Voici également un autre exemple formé de question et sa réponse correspondante :

```
text = "Question: Quelle est la valeur de lambda de la Laine de verre GR 32 roulé revêtue kraft ?"
device = "cuda:0"
inputs = tokenizer(text, return_tensors="pt").to(device)

outputs = model.generate(**inputs, max_new_tokens=500)

# Convertir les sorties en texte
decoded_outputs = tokenizer.decode(outputs[0], skip_special_tokens=True)

# Séparer les différentes réponses en utilisant le caractère de nouvelle ligne
responses = decoded_outputs.split('\n')

# Extraire uniquement la première réponse
first_response = responses[1]

print(first_response)
```

Réponse: La valeur de lambda est égale à 33,12m².K/W.

FIGURE 4.26 – Résultat BLUE SCORE du modèle LLAMA3

4.6 Conclusion

À ce stade du projet, les réponses générées par le modèle LLAMA3, spécifiquement ajusté à notre base de données, montrent des promesses de pertinence et d'adéquation. Les scores BLEU, bien que perfectibles, reflètent la capacité du modèle à aligner ses réponses avec les textes de référence. Toutefois, ces résultats indiquent également qu'il reste un potentiel d'amélioration pour maximiser la précision et la contextualisation des réponses du modèle.

Conclusion et perspectives

Dans le cadre de notre projet de fin d'études axé sur le développement d'un chatbot spécialisé dans la rénovation énergétique, on a entrepris une étude comparative des modèles Llama 2, Llama 3, et Gemma. Notre objectif était de déterminer lequel de ces modèles serait le plus efficace pour répondre aux besoins spécifiques d'une audience diverse, composée tant de professionnels que de particuliers engagés dans des projets de rénovation. L'analyse comparative a mis en évidence que le modèle Llama 3, par ses capacités avancées en compréhension et génération de texte, surpassait significativement les autres modèles en termes de précision, de rappel, et de scores F1. Ces résultats sont d'autant plus pertinents qu'ils ont été corroborés par les évaluations BLEU et BERTScore, indiquant une aptitude supérieure du Llama 3 à générer des réponses qui sont non seulement pertinentes mais aussi adaptées contextuellement aux requêtes des utilisateurs.

Pour mener à bien ce projet, on a intégré des technologies de pointe telles que le traitement du langage naturel (NLP), la reconnaissance optique de caractères (OCR), le web scraping, ainsi que l'utilisation de l'API de ChatGPT. Ces technologies ne sont pas seulement des outils ; elles représentent les fondations sur lesquelles repose la capacité de notre chatbot à rester à la pointe de l'innovation dans un marché en constante évolution.

L'adoption de Google Colab, malgré ses avantages en termes d'accessibilité et de facilité d'utilisation, nous a confrontés à des limitations significatives, notamment la restriction de la RAM à seulement 12 Go dans sa version gratuite. Cette contrainte a sérieusement entravé notre capacité à entraîner de manière optimale les modèles, affectant potentiellement leurs performances. L'examen de la possibilité de passer à Google Colab Pro, qui offre des ressources plus étendues, semble être une voie prometteuse pour surmonter ces obstacles et améliorer les capacités de notre chatbot.

Il est également nécessaire d'envisager régulièrement des mises à jour du modèle pour qu'il soit toujours en phase avec les dernières avancées technologiques et réglementaires. Cette actualisation continue est essentielle pour garantir que le chatbot demeure un outil de référence pour les utilisateurs, offrant des conseils précis et actualisés qui renforcent la compétitivité sur le marché.

En conclusion, bien que les résultats initiaux aient démontré les capacités prometteuses de Llama 3, les restrictions liées aux ressources disponibles sur Google Colab ont limité l'exploitation pleine et entière de ses potentialités. L'évaluation continue et l'ajustement des paramètres du modèle seront cruciaux pour exploiter tout le potentiel de Llama 3. Cette démarche s'inscrit dans un processus d'amélioration continue, essentiel pour atteindre l'excellence dans le développement de solutions interactives qui répondent efficacement aux besoins du secteur de la rénovation énergétique.

Bibliographie

- [1] t. f. e. From Wikipedia, “Eliza.” <https://en.wikipedia.org/wiki/ELIZA>, 2024.
- [2] A. A. Hashemi, “Youtubeautomation-reddit.” <https://github.com/aahashemi/YouTubeAutomation-Reddit>, 2024.
- [3] “[tuto] fabrique comprends ton premier réseau de neurones en partant de zéro!” <https://datafuture.fr/post/fabrique-ton-premier-reseau-de-neurones/>, 2018.
- [4] ADRIANO, “Qu’est-ce qu’un convolutional neural network?” <https://www.jeuxetredatascientist.fr/convolutional-neural-network/>, 2011.
- [5] Y. H. Liu, “Understanding the mechanism and types of recurrent neural networks.” <https://opendatascience.com/understanding-the-mechanism-and-types-of-recurring-neural-networks/>.
- [6] vignesh yaadav, “Exploring and building the llama 3 architecture : A deep dive into components, coding, and inference techniques.” https://medium.com/@vi.ai_/exploring-and-building-the-llama-3-architecture-a-deep-dive-into-components-coding-a, 2024.
- [7] J. S. L, “Intro to transformer architecture.” <https://www.linkedin.com/pulse/intro-transformer-architecture-jithin-s-l>.
- [8] M. Jumelle, “Fine-tuning de llm : tout savoir.” <https://blent.ai/blog/a/fine-tuning-llm>.
- [9] H. Srivatsa, “Fine-tuning versus rag in generative ai applications architecture.” <https://harsha-srivatsa.medium.com/fine-tuning-versus-rag-in-generative-ai-applications-architecture-d54ca6d2acb8>, 2024.
- [10] vignesh yaadav, “Exploring and building the llama 3 architecture : A deep dive into components, coding, and inference techniques.” https://medium.com/@vi.ai_/exploring-and-building-the-llama-3-architecture-a-deep-dive-into-components-coding-a.
- [11] Reddit, “Oauth2.” <https://github.com/reddit-archive/reddit/wiki/oauth2>, 2024.
- [12] “Convertir d’image en texte - ocr en ligne.” <https://www.onlineocr.net/fr/>.
- [13] A. D. l’Environnement et de la Maîtrise de l’Énergie (ADEME), “Diagnostic de performance Énergétique (dpe),” *ADEME*, 2020.
- [14] M. de la Transition écologique, “Loi sur la transition énergétique pour la croissance verte,” *Gouvernement.fr*, 2015.
- [15] I. N. de la Statistique et des Études Économiques (INSEE), “Impact de la rénovation énergétique des logements sur la performance énergétique et le confort des occupants,” *INSEE*, 2021.
- [16] L. Monde, “Le marché de la rénovation énergétique en france : enjeux et perspectives,” *Le Monde*, 2022.

- [17] L. Tribune, “Les innovations technologiques au service de la rénovation énergétique,” *La Tribune*, 2023.
- [18] T. Braun and N. Dexter, “Building chatbots : A practitioner’s guide,” *arXiv preprint arXiv :1711.07399*, 2017.
- [19] S. Duong, “Building chatbots with python : Using natural language processing and machine learning,” *Apress*, 2019.
- [20] V. D. B. Albená, “Mastering conversational user experience : Insights from the world’s best chatbots,” *Packt Publishing Ltd*, 2018.
- [21] L. Gowalla, “The definitive guide to building chatbots : Explore the concept of chatbots and learn how to build them,” *Createspace Independent Publishing Platform*, 2017.
- [13] [14] [15] [16] [17] [18] [19] [20] [21] [6]