REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Université Saad DAHLAB - Blida 1

Faculté des sciences

Département d'Informatique

Mémoire présenté par :

Mlle. HEDJAZI Ines et Mme. DRARENI Sara

Pour l'obtention du diplôme de Master

Domaine : Mathématique et Informatique

Filière : Informatique

Spécialité : Traitement Automatique de la Langue

Sujet **:**

Conception d'un Système TAL pour la Génération Automatique et la Validation de Contenus Encyclopédiques

Soutenu le : 30 Juin 2025 , Devant le jury composé de :

Mme. MEZZI .M	Promotrice	Université de Blida 1
Mme Meskaldji	Président	Université de Blida 1
Mme Ouahrani	Examinateur	Université de Blida 1

Remerciement

Nous tienons à exprimer toute notre gratitude et notre profond respect à Madame Melyara Mezzi, notre promotrice, pour son accompagnement précieux, sa bienveillance constante et sa confiance tout au long de ce travail.

Nous remercions également l'ensemble des enseignants du département Informatique de l'Université de Saad Dahlab pour la qualité de leur enseignement, leur rigueur scientifique, et leur engagement constant à transmettre le savoir avec passion.

Mes remerciements les plus sincères vont aussi à nos camarades de classe et à nos amis proches, qui nous ont généreusement aidé durant la phase de validation humaine du projet, en consacrant de leur temps pour lire, annoter, évaluer et critiquer les articles générés avec sérieux et bienveillance.

À vous tous, merci.

Dédicaces

To my dear parents, ***Abdeslam*** *and* ***Nadia****,*
Your love, strength, and sacrifices built the foundation of everything I've achieved. I owe it all to you.

To ***Mehdi****,*
My anchor and my light. Thank you for being my calm through every storm.

To ***Madame Melyara Mezzi****,*
You didn't just teach, you guided, uplifted, and inspired. Your trust in me, your kindness, and your quiet encouragement gave me the confidence to rise. I'll carry your impact with me always. May Allah bless you and grant you the best.

To ***Kaouther****,* ***Sara****,* ***Intissar*** *,*
For the support, the laughter, and the shared journey, I'm grateful for every moment.Each of you shaped this chapter in your own way.

The best is yet to come, inshallah.

Ines

إهداء

الشكر لله أولا ثم إلى نفسي فأنا أهنئها على ما وصلت اليه وماهي عليه الان وأشكرها على قمة صبرها لما واجهته رغم صغرها

الشكر لمن كانت دعواتهما سرّ وصولي وحمايتي للان ولا تزال لمن تعبا وسهرا لراحتي ورعايتي، ولمن ضحّيا وما زالان يضحيان لأجلي، إلى من كانوا السند والمسند والملاذ في كل هذه الحياة إلى من هم الحياة والداي الكريمين

إلى من جاء ليكمل معي الدرب إلى الصاحب والصديق الى الزوج والحبيب الى خليلي الى شريكي وداعمي

إلى كتفي بعد أبي وزوجي إلى وحيدنا وكبيرنا أخي

إلى إخوتي وأبنائهم وأزواجهم

إلى عائلتي الثانية عائلة زوجي

إلى البروفيسور ولد عيسى الذي رغم مستواه العلمي والفكري العالي جدا إلا انه تواضع معنا فلم يبخل علينا بعلم يعرفه ولا بنصيحة أبوية سواء كأستاذ أو كرئيس قسم.

إلى الأستاذة التي لعبت دور الأستاذ الحقيقي تبنتنا كإخوة صغار لها قبل أن نكون طلبتها إلى الراقية مليارا مزي

إلى صديقتي منى اشكرها على دعمها النفسي وتحفيزاتها لي على الاستمرار وعدم الفشل

أما فالختام فأختتم بصديقتي وشريكتي إيناس فانا أشكرها على الوقت الجميل الذي أمضيناه معا من العام الماضي للان سواء فالدراسة او فالتحضير للمذكرة شكرا على مجهوداتك، شكرا لصحبتك الطيبة

سَارَة

Résumé

Le développement de grands modèles de langage (LLM) tels que GPT a conduit à des avancées significatives dans la génération automatique de contenu. Cependant, leur application dans des contextes spécifiques, tels que la création d'un dictionnaire encyclopédique algérien, nécessite d'affiner (fine-tune) les modèles pour assurer la pertinence, l'exactitude et la cohérence du contenu généré.

Ce projet s'inscrit dans une démarche d'IA explicable, où des mécanismes sont développés pour structurer l'information générée et la valider en collaboration avec des experts. Le système vise à enrichir la base de données du patrimoine culturel algérien tout en intégrant des techniques modernes de traitement du langage naturel pour assurer la cohérence et la fiabilité de la génération de contenu.

Mots clés : *Web Scraping, Multilinguisme, Modèles de Langage Large, Validation Collaborative.*

Abstract

The development of Large Language Models (LLMs) such as GPT has led to significant advances in automatic content generation. However, their application in specific contexts, such as the creation of an Algerian encyclopedic dictionary, requires fine-tuning the models to ensure the relevance, accuracy, and coherence of the generated content.

This project is part of an explainable AI approach, where mechanisms are developed to structure the generated information and validate it in collaboration with experts. The system aims to enrich the Algerian cultural heritage database while integrating modern natural language processing techniques to ensure consistency and reliability in content generation.

Keywords : *Web Scraping, Multilinguisme, Large Language Models, Collaborative Validation*.

ملخص

تطوير نماذج اللغة الكبيرة أدى إلى تقدم كبير في التوليد التلقائي للمحتوى. ومع ذلك، فإن تطبيقها في سياقات محددة، مثل إنشاء قاموس موسوعي جزائري، يتطلب ضبط هذه النماذج لضمان الصلة، والدقة، والترابط في المحتوى المُولَّد.
هذا المشروع يُعد جزءًا من نهج الذكاء الاصطناعي القابل للتفسير، حيث يتم تطوير آليات لهيكلة المعلومات المُولَّدة والتحقق منها بالتعاون مع الخبراء. ويهدف النظام إلى إثراء قاعدة بيانات التراث الثقافي الجزائري مع دمج تقنيات معالجة اللغة الطبيعية الحديثة لضمان الاتساق والموثوقية في توليد المحتوى .

الكلمات *المفتاحية*:

التعدد اللغوي، نماذج اللغة الكبيرة، التحقق التعاوني استخراج البيانات من الويب،

Table des matières

Liste des figures

Liste des tableaux

Liste d’acronymes

- **BART :** Bidirectional and Auto-Regressive Transformers
- **BERT :** Bidirectional Encoder Representations from Transformers
- **GPT :** Generative Pre-trained Transformer
- **GPU :** Graphics Processing Unit
- **HTML :** HyperText Markup Language
- **Http :** HyperText Transfer Protocol
- **HTTL :** Human in the loop
- **IA :** Intelligence Artificielle
- **JSON :** JavaScript Object Notation
- **JSONL :** JavaScript Object Notation Lines
- **LCS :** Longue commune séquence
- **LLM :** Large Language Model
- **NLP :** Natural Language Processing
- **PEFT** : Parameter-Efficient Fine-Tuning
- **RAG** : Retrieval-Augmented Generation
- **ReLU** : Rectified Linear Unit
- **RNN :** Recurrent Neural Network
- **T5 :** Text-to-Text Transfer Transformer
- **TAL :** Traitement Automatique de la Langue
- **UNESCO :** United Nations Educational, Scientific and Cultural Organization
- **URL :** Uniform Resource Locator
- **XML:** Extensible Markup Language

INTRODUCTION GENERALE

1.Contexte Global

Saleh Zayadneh définit le patrimoine comme c'est ce qui se transmet d'une génération a une autre : des coutumes, des traditions, des sciences, des lettres, et des arts, et on dit patrimoine humain, patrimoine littéraire, etc. Et cela inclut l'art, la poésie, la musique, les traditions de mariage, les occasions et ce qu'il implique dans les façons héritées de la performance, les formes de dance, les jeux, etc [1] .

En ce qui concerne notre pays, L'Algérie a traversé de nombreuses civilisations, elle a été colonisée d'abord par les peuples amazighs, suivie par la civilisation romaine et la civilisation byzantine, puis avec les conquêtes islamiques et le contact avec les Ottomans à cette époque a laissé une empreinte et une grande influence sur le patrimoine, les coutumes et les traditions de l'Algérie, La succession de ces civilisations sur le territoire algérien a contribué dans l'industrie, la formulation et la formation d'un mélange culturel unique apparaît dans l'urbanisme, l'art, la culture, les coutumes et même la langue. Leur importance est de renforcer l'appartenance nationale et la compréhension de l'histoire algérienne, partie essentielle de l'identité et de la mémoire de la nation, il est donc important de la préserver et de diffuser pour assurer sa survie et son ancrage et ainsi promouvoir l'Algérie dans le monde.

Dans ce contexte, l'automatisation de la création et la vérification de contenus encyclopédiques représentent une formidable opportunité pour consigner et préserver le riche patrimoine de l'Algérie; les avancées récentes en Traitement Automatique du Langage (TAL), notamment avec des modèles comme GPT ont ouvert la voie à la production de textes de grande qualité, cependant; pour appliquer ces technologies à un domaine particulier tel que le patrimoine algérien; il est primordial de les adapter afin d'assurer des contenus pertinents; précis et cohérents.

2. Poblématiques

A notre avis, le problème principal de ce projet est de collecter des articles pertinents à partir de sources web ouvertes, souvent non structurées et multilingues, tout en respectant le patrimoine algérien. Les LLMs comme GPT offrent des solutions puissantes

pour générer des textes, mais ils doivent être adaptés pour garantir des contenus précis et culturellement pertinents. De plus, il est essentiel de faire appel à des experts pour assurer la fiabilité et véracité du contenu ; leur rôle sera de vérifier les informations, d'évaluer la qualité des articles et de faire des corrections si besoin, en combinant le traitement automatique de la langue et la validation humaine, on pourra produire un contenu encyclopédique précis, fiable et adapté au contexte culturel et historique.

3. Objectifs

Pour répondre à ces défis, notre projet propose une approche hybride, en combinant :

- La conception d'un algorithme de web scraping pour extraire des informations textuelles en arabe et en français à partir des sources ouvertes tels que les sites web, les réseaux sociaux et les base de données.

- Adapter et affiner des modèles de langage large (LLM) : pour structurer et générer des articles encyclopédiques de manière cohérente et précise, dans le but d'assurer une production de contenu fidèle aux réalités culturelles et historiques, ces modèles seront entraînés à partir d'un corpus représentatif du patrimoine algérien.

- lancé une plateforme interactive en collaboration avec des experts humains pour assurer que les contenus sont validés et leur conformité culturelle est garantie, elle appelle la contribution active de l'utilisateur en lui offrant des outils de suggestion, de relecture et de validation

4. Organisation du Mémoire

Ce mémoire se décompose en trois chapitres, précédés par une introduction sur le contexte général et suivi d'une conclusion contenant des orientations pour les travaux futurs,

Les chapitres sont divisés comme suit :

Chapitre 01 : état de l'art

Dans ce chapitre, nous introduisons les concepts fondamentaux liés à notre sujet, ainsi que les travaux antérieurs réalisés dans ce domaine, ce chapitre constitue le fondement

théorique permettant de comprendre les défis et les problématiques associés à notre sujet. Il offrira également un aperçu des travaux connexes dans les domaines de la génération de contenu encyclopédique et de la validation collaborative.

Chapitre 02 : Conception

Ce chapitre détaille les étapes du pipeline mis en place : de la collecte des données culturelles via le web scraping, le fine-tuning des modèles de Language, l'évaluation de ces modèles et le choix finale de meilleur entre eux.

Chapitre 03 : Implémentation de la solution

Dans ce chapitre, nous présentons les outils et méthodes mis en œuvre pour évaluer la qualité des articles générés. Une analyse automatique des sentiments permettant de détecter la tonalité des retours critiques, ainsi qu'un mécanisme d'apprentissage actif intégrant un score d'incertitude. Enfin, les résultats obtenus sont analysés, les améliorations mesurées sont présentées, et les perspectives.

CHAPITRE I: ETAT DE L'ART

1.Introduction

Dans un contexte où les modèles de langage (LLM) et le traitement automatique du langage (TAL) gagnent une popularité ces derniers temps le développement et la validation de contenus automatisés attirent davantage l'attention [3]. Bien que ces avancées offrent de nouvelles opportunités pour produire des contenus organisés et crédibles ; elles soulèvent également des défis en matière de qualité, d'exactitude et d'adaptation culturelle des informations générées.

Dans ce chapitre spécifique que nous entamons maintenant ensemble, découvrons les recherches antérieures et les idées clés en lien avec notre sujet d'étude,

Nous allons explorer diverses approches utilisées pour la création de contenus encyclopédiques, méthodes de collectage de données web, l'importance des modèles linguistiques et les plateformes collaboratives de validation. Notre objectif principal est de placer notre projet dans le contexte des études actuelles et d'identifier les aspects nécessitant encore des améliorations afin d'y apporter une contribution significative.

2.Brève histoire de la génération de texte

La création automatique de texte a eu un grand changement ces dernières années, passant de systèmes basés sur des règles à modèles d'intelligence artificielle et réseaux de neurones, voici une brève histoire de leur évolution :

2.1 Années 1950-1980 : Systèmes basés sur des règles

Exemple : le premier chabot, Eliza, Était l'un des premiers systèmes de génération de texte, il simulait une conversation en utilisant des règles simples pour reformuler les entrées de l'utilisateur [2].

2.2 Années 1990-2000 : L'ère des modèles statistiques

En utilisant des approches statistiques, différents modèles ont été développés.

Par exemple, les chaînes de Markov et les modèles de langage n-grammes étaient entraînés sur de grands corpus de texte pour calculer des probabilités. Mais ces modèles étaient incapables de capturer des dépendances à long terme dans le texte, par conséquent, ils généraient souvent des phrases grammaticalement correctes mais manquaient de cohérence contextuelle [4] [5] .

2.3 Années 2010-présent : réseaux de neurones modèles de langage modernes

Les réseaux de neurones et les modèles de langage modernes comme les RNN et les Transformers produisent des textes du bon qualité, cependant, ces modèles demandent d'importantes ressources informatiques, ce qui les rend coûteux, ils peuvent aussi produire des informations inexactes, nécessitant des filtres de contenu ou des vérifications humaines pour garantir leur fiabilité[6].

Exemple : GPT[13] , il est entraîné sur d'énormes corpus de texte, utilise l'architecture Transformer pour générer des textes de haute qualité adaptés à divers contextes [6].

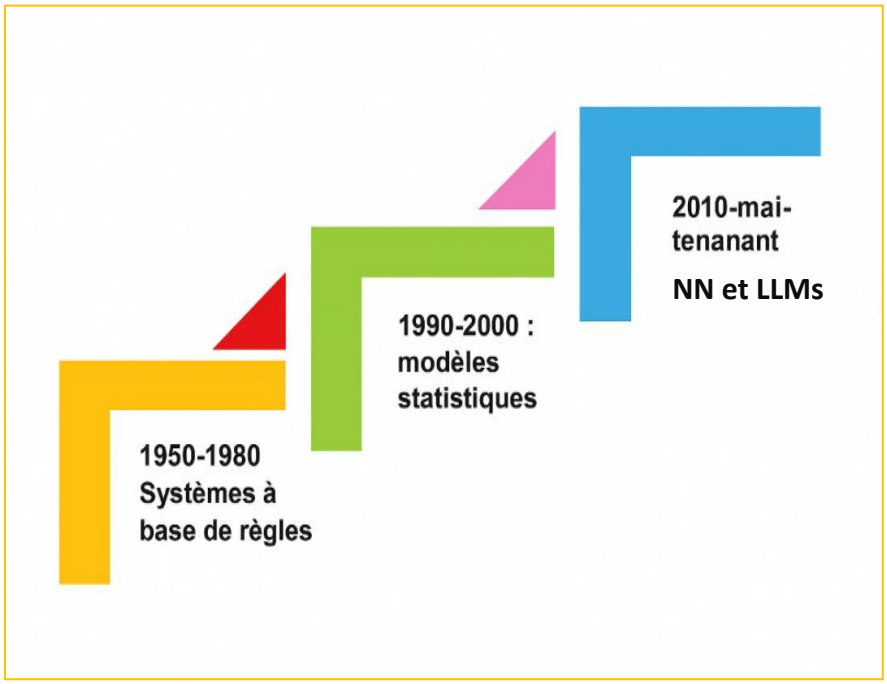

Figure 1 : évolution des modelés de génération de texte

3.Les modèles de langage large (LLM)

Les modèles de langage large (LLM), ou Large Language Models en anglais, désignent une classe avancée d'algorithmes d'intelligence artificielle conçus pour traiter, comprendre et générer de vastes quantités de texte naturel. Entraînés sur des corpus massifs, allant de livres et articles scientifiques à des pages web et discussions en ligne, ces modèles ajustent des milliards de paramètres afin de prédire la probabilité de chaque mot dans une séquence [41]. L'adjectif « large » reflète à la fois l'ampleur des données utilisées et la taille du réseau neuronal, généralement de type Transformer [7] ,tandis que « Language model » souligne leur capacité à modéliser la structure et la sémantique du langage humain *[24]*. Grâce à cette architecture, les LLM peuvent être appliqués à de nombreuses tâches : rédaction, traduction, résumé ou dialogue, simplement en modulant leurs prompts sans nécessiter de modifications architecturales profondes [9] .

3.1 Types des LLMs

Les LLMs sont classés en fonction de leur architecture, de leur objectif…etc. Dans le tableau suivant, on a défini quelques modèles :

Tableau 1 : Quelque grands modèles de langue

	Classification selon	Modèles	Comment il travail	Exemples
01	Architecture	Transformer	Utiliser l'architecture du transformer pour la compréhension et la génération du langage naturel.	GPT T5
		Modèles RNN	Utiliser les RNN ou les LSTM pour traiter des données séquentielles.	Modèles Seq2Seq
		Modèles hybrides	Combiner des transformateurs avec d'autres techniques comme l'augmentation de la mémoire	RETRO, RAG
02	Approche de fine tuning	Modèles autorégressifs	Générer du texte en prédisant le mot suivant basé sur les précédents	GPT XLNet
		Seq2Seq	Mapper les séquences d'entrée aux séquences de sortie	mT5 BART
03	Les capacités multilingues	Modèles monolingues	Se concentrer sur une seule langue.	FinBERT
		Modèles multilingues	Gérer efficacement plusieurs langues.	XLM-R mT5

3.2 Architecture générale

Les grands modèles de langage (LLM), comme GPT, reposent sur une succession de couches et de composants complémentaires, chacun jouant un rôle clé pour permettre au modèle de comprendre et de générer du texte [41] **.**

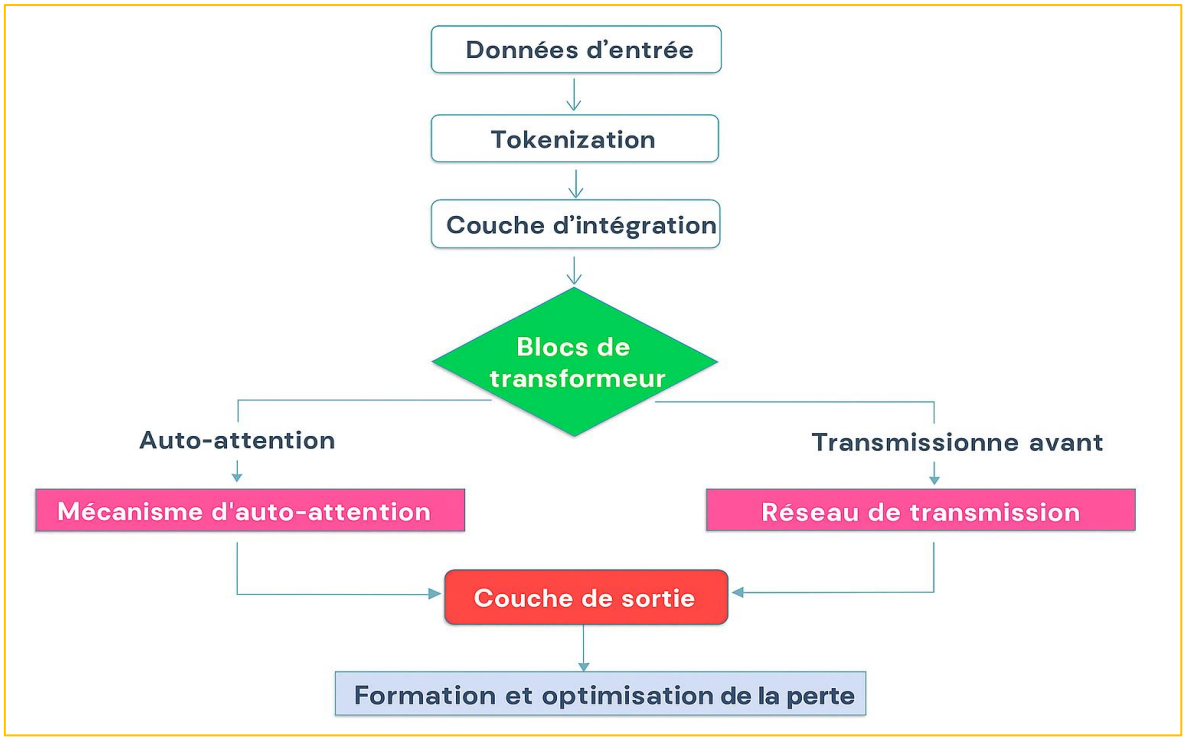

Figure 2 : Vue d'ensemble de l'architecture d'un LLM

3.2.1 Couche d'entrée (Input Layer)

-Tokenization

La tokenisation est un processus fondamental en traitement du langage naturel (NLP), elle consiste à décomposer une séquence de texte en unités plus petites appelées jetons(tokens). Ces 'Tokens' peuvent être des mots individuels, des phrases ou même des caractères[11] . Chaque token est ensuite transformé en un identifiant numérique, puis encodé sous forme de vecteur (embedding), afin d'être exploité par les couches suivantes [10] . Exemple :

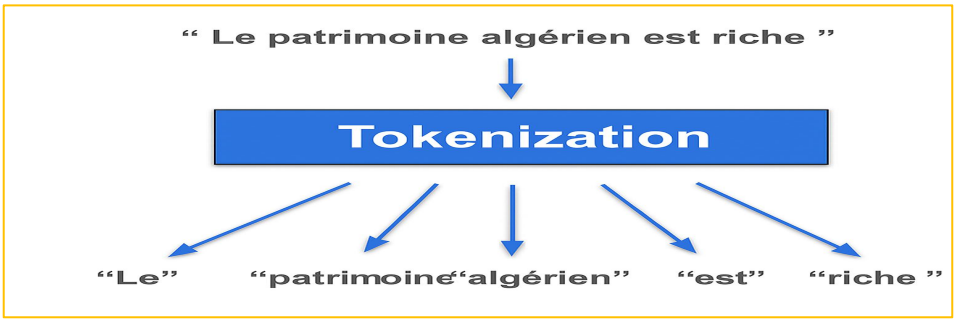

Figure 3 : exemple de tokenisation

3.2.2 Couche d'embedding (Embedding Layer)

-Word Embeddings

Chaque token est mappé sur un vecteur dense de dimension fixe, ce que l'on appelle *word embedding*. Cette représentation vectorielle capture la signification sémantique des mots en les plaçant dans le même espace. Bien que des méthodes telles que Word2Vec ou GloVe aient été utilisées par le passé, le Transformer apprend ses propres embeddings directement pendant l'entraînement [7].

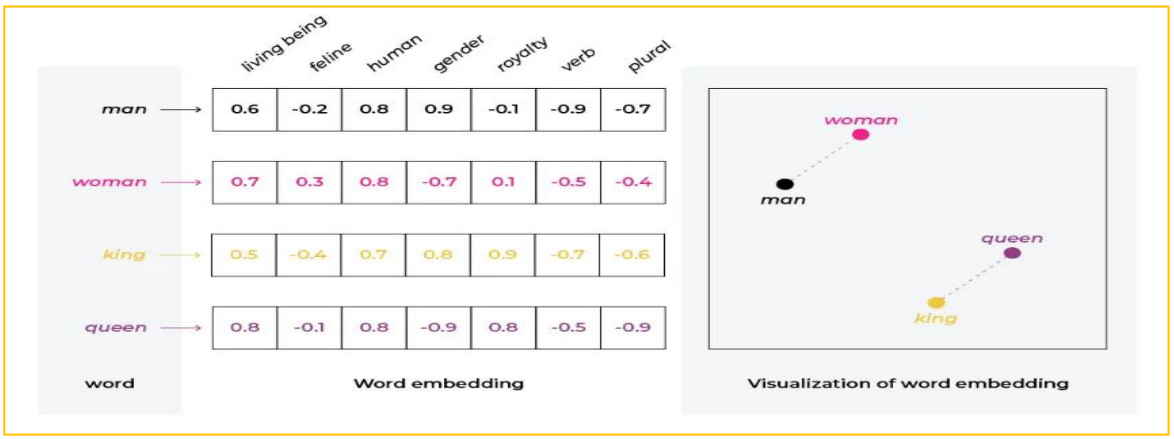

Figure 4 : Illustration d'un word embedding

3.3 Architecture de transformateur

Le Transformer, proposé par Vaswani et al. (2017) dans Attention is All You Need, se structure en blocs empilés selon le schéma encodeur–décodeur.

3.3.1 Schéma encodeur–décodeur

- **Encodeur** : une succession de blocs identiques, chacun combinant :

-Self- Attention multi- tête

-Réseau feed- forward positionnel

-Connexions résiduelles et normalisation

-Décodeur : reprend la structure de l'encodeur, à laquelle s'ajoute une couche d'attention encodeur–décodeur pour extraire les informations pertinentes de l'encodeur.

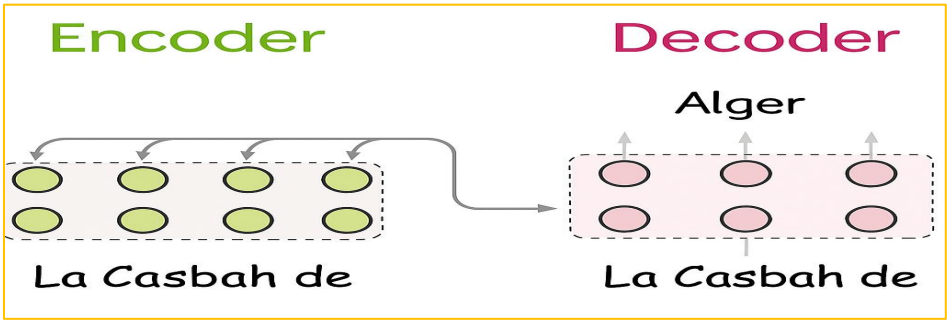

Figure 5 : Architecture encodeur-décodeur

3.3.2 Variantes de LLM selon l'usage

En fonction des composants du Transformer utilisés, trois grandes familles se dégagent *:*

-**Encodeur seul** (ex : BERT [12]) : permet une compréhension bidirectionnelle, idéal pour la classification ou l'analyse de sentiments.

-**Décodeur seul** (ex : GPT[13]) : se focalise sur la génération autorégressive, prédiction mot à mot.

-**Encodeur–décodeur** (ex : T5 [14], BART [15]) : conçu pour des tâches Seq2Seq (traduction, paraphrase, génération de contenu structuré).

3.3.3 Mécanisme de self- attention

-Projections Q, K, V : chaque embedding est transformé en vecteurs Query, Key et Value.

-Scores d'attention

-Multi- Head Attention : plusieurs têtes calculent simultanément des attentions, puis leurs résultats sont concaténés et relinéarisés, enrichissant ainsi la capacité du modèle à identifier des relations variées.

3.3.4 Empilement de blocs

Les blocs décrits ci-dessus sont empilés les uns sur les autres, formant un réseau profond. Chaque couche successive affine progressivement les représentations, passant de motifs simples à des abstractions complexes.

3.3.5 Couche de sortie

- **Modèles autorégressifs (GPT) :** le modèle prédit itérativement le mot suivant à partir du contexte fourni.

- **Modèles masqués (BERT) :** certains mots sont masqués à l'entrée, et le modèle apprend à les reconstituer.

- **Softmax final** : convertit les sorties en probabilités sur le vocabulaire pour sélectionner la réponse la plus probable.

3.3.6 Stratégies d'apprentissage des données

3.3.6.1 Pré-Training

Les LLM sont pré-entraînés sur des grands ensembles de données, comme les livres, les sites Web, les articles ou d'autres ressources textuelles.

La diversité de ces données aide le modèle à apprendre et à comprendre le langage et le contexte et la sémantique. Pendant l'entraînement, ces modèles utilisent des fonctions objectives comme la modélisation du langage masqué, Il s'agit de prédire les mots manquants dans la phrase (par exemple : BERT), ou bien de la modélisation autorégressive, ou elle se concentre sur la prédiction du mot suivant dans une séquence (par exemple : GPT).

Cependant, l'entraînement de ces modèles nécessite du matériel puissant et des ressources de mémoire et de calcul substantielles [16].

3.3.6.2 Fine-tuning :

Après l'entraînement, les modèles sont adaptés pour des tâches ou des domaines précis, le fine-tuning consiste à entraîner le modèle sur des données spécifiques afin de l'adapter à des tâches telles que la génération de texte, l'analyse de sentiments, ou la traduction automatique [18].

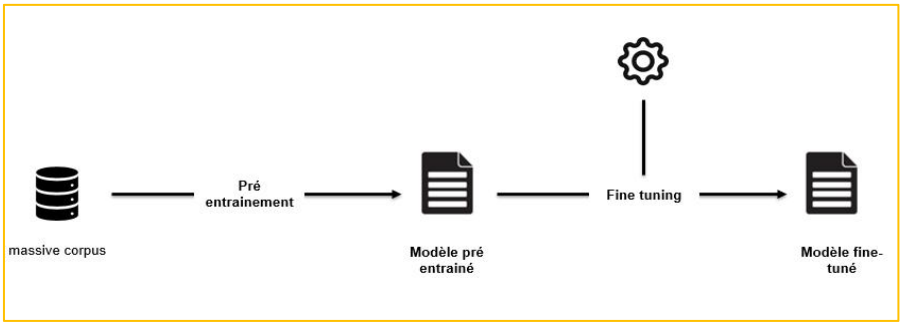

Figure 6 : fine tuning d'un LLM

Le processus de fine-tuning est comme suit [17] :

-La collecte et la préparation des données : les données doivent être de haute qualité et spécifique à la tache cible.

-Le nettoyage des données : éliminer les doublons, les erreurs et les incohérences

-Choix du modèle pré-entraîné : un modèle pré-entraîné pour la génération de texte structuré peut être plus adapté à la rédaction de contenu encyclopédique à partir de gabarits qu'un modèle conçu pour des tâches plus créatives,

-Évalue les performances initiales du modèle sélectionné sur la tâche cible : pour avoir une base de référence qui permettra de mesurer l'amélioration par la suite

-Ajuster minutieusement les hyperparamètres : implique l'ajustement de taux d'apprentissage (learning rate) et la taille des lots (batch size), ce processus permet d'optimiser les performances du modèle pour la tâche cible [18] .

-il est important de choisir les hyperparamètres de manière à éviter l'Overfitting et l'underfitting et ainsi obtenir un modèle performant sur les données de test, pour trouver les bonnes combinaisons, on peut utiliser des techniques telles que la recherche aléatoire [18], la recherche en grille , ou l'optimisation bayésienne.

Dans le tableau 2, on a fait une petite comparaison :

Tableau 2 : Comparaison simple entre les méthodes de choix des hyperparamètres

Méthode	Quand l'utiliser ?	Avantages	Inconvénients
Recherche aléatoire	Lorsque nous voulons des résultats rapides	Rapide	Peut manquer certaines bonnes valeurs
Recherche en grille	Lorsque nous avons peu de combinaisons	Garantit de tester toutes les combinaisons	Coûteuse en temps et en ressources
Optimisation bayésienne	Lorsque les expériences sont coûteuses	Intelligente et Économise du temps	Complexe et nécessite des étapes supplémentaires

3.3.6.3 les méthodes de fine-tuning

Il existe plusieurs méthodes et techniques de fine-tuning utilisées pour ajuster les paramètres du modèle, voici dans la figure suivante quelques-unes de ces méthodes dont les grands modèles de langage peuvent être adapté :

1.fine tuning par instruction

Consiste à entraîner le modèle à l'aide d'exemples montrant comment il doit répondre à des requêtes spécifiques, elle est particulièrement utile pour renforcer sa capacité à traiter efficacement diverses instructions spécifiques, par exemple, pour améliorer les capacités de génération de texte, on utilise une dataset contenant des instructions comme "généré un article sur la casbah" suivi de l'article sur ce sujet.

2.Parameter-Efficient Fine-Tuning, PEFT

Le PEFT faire la mise à jour d'un petit sous-ensemble seulement des paramètres du modèle pendant l'entraînement [19], donc il réduit les besoins en mémoire et en calcul par rapport à un fine tuning complet, donc cette méthode aide à contourner les limites

matérielles et évite l'oubli catastrophique, et préserve les connaissances initiales du modèle tout en l'adaptant à de nouvelles tâches [20].

3.Fine-Tuning spécifique à une tâche

Le fine-tuning spécifique à une tâche adapte un modèle pré-entraîné pour exceller dans un domaine particulier (ex. texte génération) à l'aide d'un dataset dédié, et il garantit une grande précision pour les besoins spécifiques.

4. L'apprentissage par transfert

Exploite un modèle pré-entraîné sur un large dataset généraliste, puis l'adapte à des tâches spécifiques avec peu de données.

5. Multi-Task Learning

Entraîne un modèle sur un dataset couvrant plusieurs tâches (ex. résumé, traduction, reconnaissance d'entités), Cela améliore ses performances.

6. Le Fine-Tuning séquential

Adapte un modèle à une série de tâches connexes par étapes, par exemple, un modèle linguistique généraliste peut d'abord être adapté pour le langage médical, puis pour la cardiologie pédiatrique.

3.3.6.4 Défis et limites du fine-tuning des LLM et les solutions [21]

Tableau 3 : Défis et limites du fine-tuning des LLM et les solutions

Défis	Explication	La cause
Overfitting	Le modèle performe bien sur les données d'entraînement mais mal sur des données nouvelles,	-Le dataset est trop petit ou peu diversifié

Underfitting	Le modèle échoue à capturer la complexité des données, même sur le jeu d'entraînement.	-modèle trop simple -Hyperparamètres mal ajustés -Données insuffisamment prétraitées.
Oubli catastrophique	Le modèle oublie ses connaissances antérieures lors d'un fine-tuning sur une nouvelle tâche.	-Réglage fin trop agressif sur une tâche niche. -Poids du modèle complètement réécrits.

Et le tableau suivant présente des solutions concrètes pour chaque défi identifié précédemment [21]:

Tableau 4 : solutions concrètes pour chaque défi

Défis	Solution
Overfitting	-Early stopping, -Dropout -Augmentation des données
Underfitting	-Augmenter la taille du modèle, -Meilleur réglage des hyperparamètres
Oubli catastrophique	-PEFT

3.3.7 Optimisation

Le but de training est de minimiser la fonction de perte (loss function) dans notre cas c'est la fonction 'cross-entropy loss', qui mesure la différence entre la distribution de probabilité prédite et la distribution réelle.

Ça formule est comme suit :

$$Loss = -\frac{1}{n}\sum_{t=1}^{n} log\ P(\omega_t|\omega_1 \cdots\cdots \omega_{t-1})$$

Tels que :

n : longueur de la séquence.

ω_t: le mot à la position t.

$p(\omega_t|\omega_1 \cdots \omega_{t-1})$ Probabilité prédite du mot conditionnée aux mots précédents.

Pour minimiser la valeur de Loss ou bien le perte, le modèle utilise l'algorithme descente de gradient, qui est un algorithme d'optimisation qui ajuste de manière itérative les paramètres du modèle (poids et biais) dans la direction qui réduit la perte, les mises à jour des valeurs gradients requis sont faire à l'aide de la rétropropagation, ce qui est un algorithme fondamental utilisé pour former les réseaux neuronaux, il s'agit du processus qui consiste à calculer les gradients de la fonction de perte par rapport aux paramètres du modèle (poids et biais) et à utiliser ces gradients pour mettre à jour les paramètres pendant l'entraînement , et pour améliorer la vitesse de convergence et la stabilité du processus d'optimisation il applique des variantes de la descente de gradient , comme par exemple : Adam optimiser.

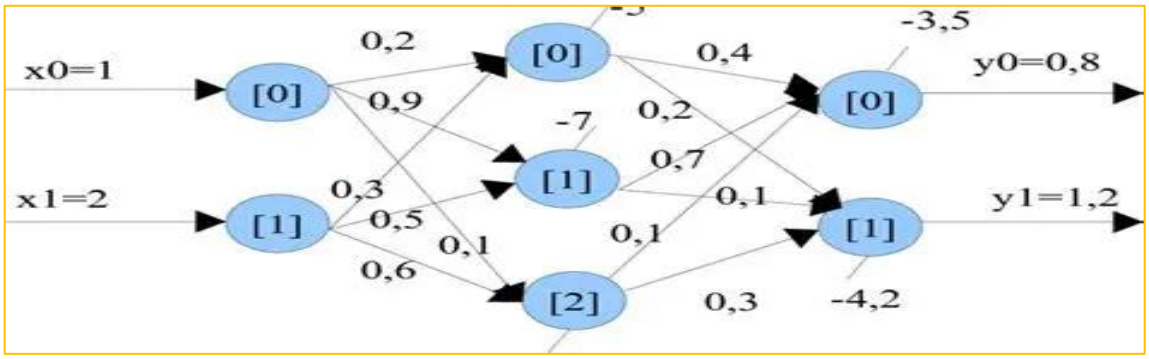

Figure 7 : Une illustration simple du fonctionnement de la rétropropagation par des ajustements de poids

4. Le patrimoine Algérien et son importance

L'Algérie est un vaste pays d'Afrique du Nord, elle possède l'un des patrimoines les plus riches et les plus variés du monde arabe et africain, ce patrimoine incarne la diversité historique et culturelle qui a marqué le sol algérien, et il illustre l'influence des différentes cultures telles que berbère, islamique, ottomane et française.

Il se divise en deux grandes catégories : le patrimoine matériel et immatériel.

4.1 Le patrimoine matériel

Comprend les monuments historiques, et les architecturaux témoignent de la richesse historique de notre pays, et la diversité de ses civilisations, comme la Casbah, la ville de Timgad, etc.

4.2 Le patrimoine immatériel

Comprend les traditions orales (la musique, les danses folkloriques et les arts artisanaux).

4.3 L'importance du patrimoine

Le patrimoine algérien représente une richesse précieuse qui incarne l'histoire, les coutumes et les valeurs d'un peuple et il regroupe les monuments historiques, les œuvres artistiques, les traditions populaires et les savoir-faire transmis au fil des générations. La protection de ce patrimoine constitue a la préservation de la mémoire collective et de garantit sa continuité dans le temps, c'est aussi pour renforcer le sentiment d'appartenance et encourager le dialogue entre les différentes cultures, ainsi qu'elle est essentielle pour bâtir une société fondée sur la diversité culturelle et la connaissance.

Voici quelques exemples du patrimoine matériel :

Figure 8 : Mosquée Ketchaoua-Alger

Figure 9 : La casbah d'Alger

5. gabarits et contenu encyclopédique

5.1 Définition des gabarits

Un gabarit est un modèle ou un patron servant de référence pour produire quelque chose, dans le contexte de notre projet, c'est une structure prédéfinie, elle permet d'organiser et standardiser les contenue encyclopédique générer. Les gabarits servent de modèles pour garder une structure claire, car chaque article va suit une structure uniforme,

De plus, il aide à l'amélioration des articles généré car le modelé LLM doit est guidées au temps de génération pour produire de texte aligné.

Enfin, les experts de peuvent faire leur validation facilement grâce à ces sections prédéfinis. En figure 10, il y a un exemple d'un gabarit :

Figure 10 : Structure d'un gabarit d'article

5.2 Contenu encyclopédique

Dans notre projet, le contenu encyclopédique fait référence à les articles générés automatiquement par le système, ces articles respectent des gabarit et validés par des spécialistes,

Le but de ces articles est de :

-Documenter le patrimoine algérien en couvrant ses multiples facettes : art, histoire, traditions, architecture, langues, etc.

-Créer une base de données fiable et accessible pour les chercheurs, enseignants, étudiants et passionnés du patrimoine culturel.

-Valoriser et préserver la richesse culturelle algérienne en la rendant disponible sous forme numérique et consultable à tout moment.

5.3 Caractéristiques de contenue encyclopédique

Les articles encyclopédiques sont des articles rédigés dans un style neutre, ils ne présentent aucune opinion personnelle ni émotion et s'appuient uniquement sur des faits vérifiables provenant de sources fiables, leur langage est clair, précis et accessible, même pour un public non spécialiste et ils doivent respecter une structure logique qui est : une introduction qui pose le sujet, un développement organisé en idées claires, et une conclusion synthétique, avec un style explicatif, sans storytelling et narration.

En résumé : Un bon article encyclopédique cherche à couvrir l'essentiel du sujet sans être ni trop superficiel ni inutilement complexe.

6. Extraction de données à partir du web

Le web scraping ou web crawling fait référence à la procédure d'extraction automatique de données de sites web à l'aide de logiciels ou des algorithmes. C'est une technologie qui nous permet d'extraire des données structurées à partir de texte tel que HTML [22] .

Il est extrêmement utile dans les situations où les données ne sont pas fournies dans un format lisible par machine, tel que JSON ou XML, il a été constaté que l'utilisation d'un programme de grattage Web permettrait d'obtenir des données beaucoup plus complètes, précises et cohérentes que la saisie manuelle. Sur la base du résultat, il a été conclu que le web scraping est un outil très utile à l'ère de l'information, et un outil essentiel dans les domaines modernes. Plusieurs technologies sont nécessaires pour mettre en œuvre correctement le web scraping [22] .

6.1 : le processus de web scraping

Le processus de web scraping [23], est divisé en trois étapes qui sont :

Tableau 5 : processus de web scraping

Étapes	Description
Etape 01 : récupération	-Accéder au site web cible contenant les informations pertinentes. -Utiliser le protocole HTTP pour envoyer et recevoir des requêtes depuis des serveurs web. -Employer des méthodes similaires aux navigateurs web pour obtenir le contenu des pages. -Utiliser des bibliothèques comme curl et wget. -Envoyer une requête HTTP GET à l'URL cible. -Récupérer la page HTML en réponse [23]
Etape 02 : extraction	-Récupérer la page HTML à partir de l'étape précédente. -Extraire les données importantes de la page HTML. -Utiliser des expressions régulières pour l'extraction. - Employer des bibliothèques d'analyse HTML. -Appliquer des requêtes XPath pour localiser les informations. -Considérer XPath comme un outil pour trouver des données dans des documents [23].

Etape 03 transformation	-Conserver uniquement les données pertinentes après l'extraction. -Convertir ces données dans un format structuré. -Utiliser ce format pour la présentation ou le stockage. -Recueillir des informations à partir des données stockées. -Aider l'intelligence économique à prendre de meilleures décisions [23]

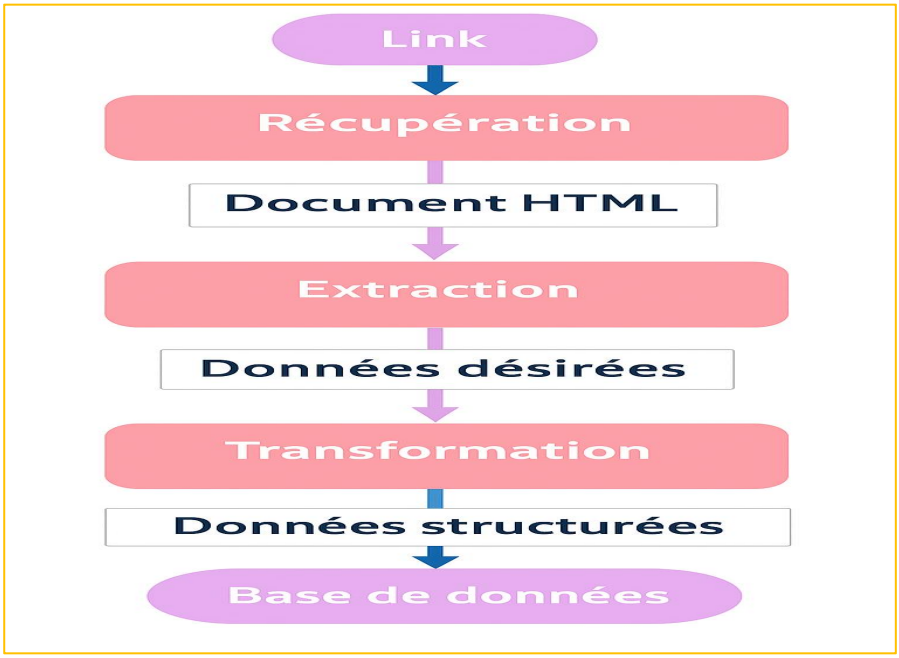

Figure 11 : processus de web scraping [23]

7. Validation collaborative

La validation collaborative est un processus ou plusieurs individus, tels que des experts, collaborent pour examiner, vérifier, corriger et améliorer la qualité du contenu généré par l'IA, c'est une forme de Human-in-the-loop (HITL) [24] où les humains sont intégrés au des système TAL pour améliorer les performances.

Dans notre cas, cette approche garantit que les articles générés sont fiables à la culture algérienne.

8. Travaux connexes

Aucun travail assez similaire n'a été trouvée dans le contexte de la génération de contenu encyclopédique sur le patrimoine algérien en particulier, mais nous avons trouvé de nombreuses plateformes qui utilisent LLMs pour générer des articles dans différent contextes ou qui intègrent la vérification collaborative pour assurer la qualité des articles génères, y compris :

- ArchGPT : est un agent d'IA, il emploie des LLM dans la planification des tâches et la génération augmentée par récupération (RAG) afin d'intégrer les directives architecturales et les exigences des utilisateurs à partir d'une base de connaissances, il génère des contenus spécifiques au domaine et les valide par des experts humains, de sorte que son approche peut inspirer notre système à mieux adapter les LLM pour les contenus culturels avec une collaboration humaine [36] .

-Hadretna : est un projet de recherche IA algérien visant à développer des LLMs pour les dialectes arabes locaux (Daridja) et la berbère (Tamazight), et il utilise des données crowdsourcées pour l'entraînement de ce modèle, avec une emphasis sur la préservation culturelle et la représentation numérique par l'IA générative [25] .il applique les LLM à un contexte culturel et linguistique ciblé (patrimoine algérien) et est basé sur la collecte de données multilingues, comme nos activités de web scraping et de prétraitement.

- Writersonic : est une plateforme qui est dédiée pour la génération automatique des articles, en fonction de modèles de langage large (LLM) comme GPT, elle support multiples langues dont a langue 'arabe et la langue français, (Writesonic.com, 2025).

- Wikipedia : est une plateforme collaborative dédiée à la création et à la validation de contenus encyclopédiques, utilisant des bots TAL, dont certains exploitent des modèles de langage large (LLM), pour automatiser des tâches comme la génération de texte, la correction et l'ajout de références Elle repose sur une communauté d'utilisateurs qui enrichissent et vérifient les informations, disponible en plusieurs langues, dont l'arabe et le français (Wikipedia.org, 2025).

Le tableau suivant évalue ces travaux selon nos critères :

Tableau 6 : Comparaison des travaux connexes avec les critères de notre projet

Projet/Plateforme	Web Scraping	Génération des texte via LLMs	multilinguisme	Validation collaborative	Contexte algérien
ArchGPT	Non	Oui	Oui	Non	Non
Hadretna	Oui	Oui	Arabe	Non	Oui
writesonic	Non	Oui	Oui	Non	Non
Wikipedia	Non	Oui	Oui	Oui	Partiel

9. Conclusion

Dans ce chapitre on a exploré les notions de base liées à notre projet, ou on a commencé par le patrimoine algérien et son importance pour l'identité culturelle, puis on a fait un petit historique de la génération de texte et des LLMs, du web scraping, ainsi que les exigences du contenu gabarit et encyclopédique, enfin, nous avons mentionné la validation collaborative qui est aussi une tâche clé dans notre système,

Ces éléments posent les bases théoriques et pratiques, préparant ainsi le terrain pour le prochain chapitre : la modélisation de notre système TAL.

CHAPITRE II : CONCEPTION DU SYSTÈME DE GÉNÉRATION ET D'ÉVALUATION AUTOMATIQUE

1. Introduction

Après avoir présenté une introduction claire sur les notions de base de ce projet dans le chapitre précédent, nous allons maintenant entrer dans la conception et la réalisation. Nous allons expliquer comment nous avons collecté l'ensemble de données, ainsi que l'architecture des LLMs choisis et comment nous les avons adaptés pour la génération des gabarits des articles, ainsi que l'évaluation par machine et la validation humaine.

2.Aperçu de la Méthodologie du Projet

Le but de notre projet est d'adapter des LLMs pour la génération des gabarits des articles concernant notre patrimoine algérien, pour le réaliser, nous devons passer par trois étapes importantes :

-Collection et préparation de l'ensemble de données,

-Choix et entraînement des modèles

- Avec l'évaluation par machine et validation humaine

La figure suivante montre l'architecture que nous allons suivre :

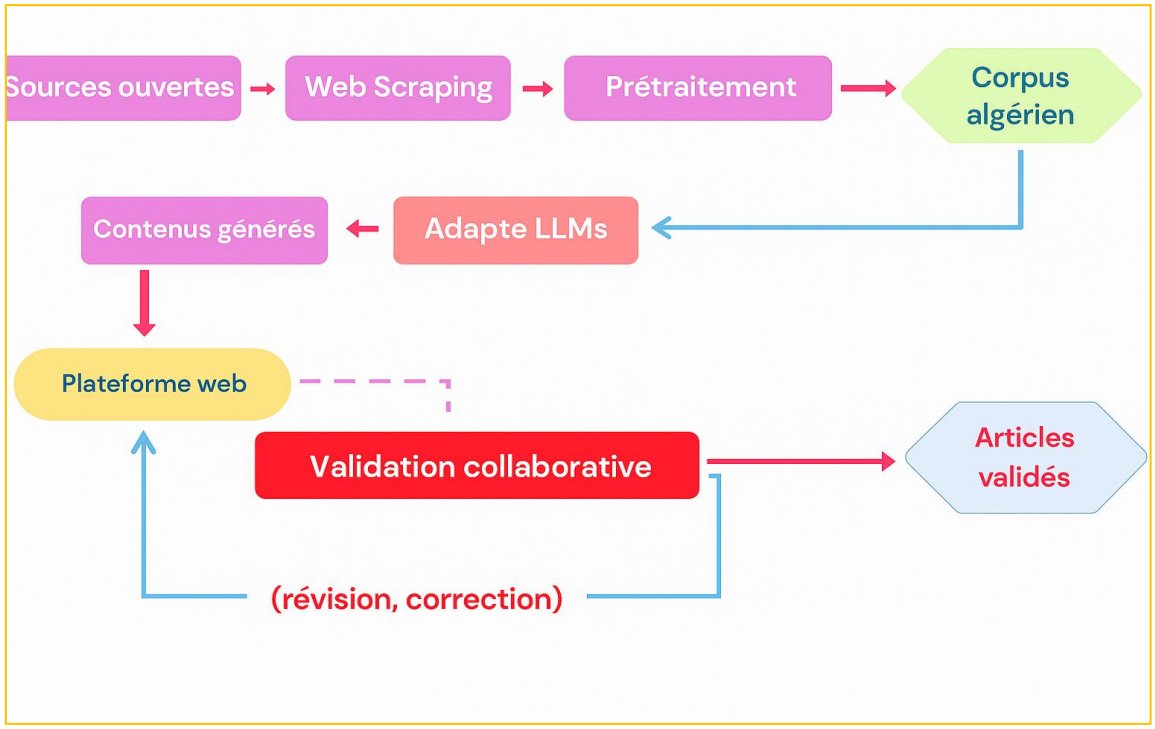

Figure 12 : Architecture globale du système de génération, d'adaptation et de validation des articles

La première approche repose sur la collection de données, là, nous essayons de trouver des articles tirés des sources du patrimoine et de la culture algérienne n'écrits qu'en arabe ou français, dans un domaine varié comme l'art, l'histoire, etc.

Les articles qu'on doit avoir vont être fidèle à l'identité algérienne, il faut donc recourir à des sources sécurisées comme celles de l'UNESCO.

La deuxième étape est d'adapter des LLMs sur la dataset est comparer leur performance, et choisir le meilleur entre eux, ces modèles ont capable à faire produire du contenu similaire à notre training data, il est donc important de choisir un modèle approprié à la génération de gabarits.

La troisième nous combinons l'évaluation par les mesures de quantités avec la validation humaine, qui est en appelons des experts pour faire une validation pour les articles génères par les modèles adapter à notre tâche.

3.Collecte et Préparation des Données

3. 1 Sources et Techniques de Collecte des Données

La première phase est la collecte des données, tout d'abord, nous recherchons des articles en arabe et en français provenant de sources fiables telles que l'UNESCO, Wikipédia et quelques journaux électroniques Algériens connus, ensuite, nous avons gardé les liens et les titres des articles, puis utilisé l'extraction de données web pour les collecter en utilisant la bibliothèque BeautifulSoup [40].

On a collecté une dataset pour la langue française est une pour la langue arabe ;

3.2 Structure et Format de dataset

Après le prétraitement et le nettoyage de ces articles, et avoir assuré que chaque article a une structure claire (voir la figure suivante), Chaque ligne du jeu de données contient une "prompt" et une "completion", où le prompt contient une instruction pour la rédaction de l'article, tandis que le completion est un dictionnaire comportant 3 sous-clés : **introduction**, **corps** et **conclusion**.

Tableau 7 : Structure et Format de dataset :

Type de dataset	Fichier jsonl	
Taille de dataset départ :	Français	179 articles
	Arabe	111 articles
Entrée	Français	Une instruction telle que « Rédiger un article détaillé sur : [titre] »
	Arabe	Une instruction telle que : اكتب مقالة مفصلة حول الموضوع التالي: [titre]
Sortie	L'article correspondant.	

3.3. Enrichissement des données

Afin que les modèles puissent réellement apprendre, ils ont besoin d'un nombre suffisant d'exemples d'entraînement en général, au moins 500 exemples sont nécessaires. Et comme le nombre des articles disponibles en ligne concernent le patrimoine et la culture algérienne reste limité, nous avons opté pour une méthode permettant d'atteindre une taille de corpus acceptable : l'augmentation de données. [26]

C'est une technique consiste à générer automatiquement de nouveaux articles à partir de données existantes, en y apportant de légères modifications, ces variations sont conçues habituer le modèle à différentes formulations[27].

Voici les méthodes que nous utilisons :

3.3.1 Back translation

Consiste à traduire un texte d'une langue vers une autre, puis à le retraduire vers la langue d'origine, la traduction inversée repose sur l'invariance sémantique présente dans les jeux de données traduits sous supervision, afin de générer de nouveaux articles qui conservent le même sens. Elle constitue ainsi un moyen efficace pour l'enrichissement des données [24].

Cette technique consiste à traduire le texte de l'arabe ou français vers l'anglais, puis à le retraduire en arabe à l'aide de la bibliothèque *googletrans*. Ce double passage génère des reformulations automatiques, ou paraphrases. [27]

3.3.2 Insertion aléatoire de mots

Cette méthode consiste à insérer aléatoirement des mots fréquents, tels que « جدًا » ou « بشكل », à différents endroits du texte, elle permet d'introduire de petites variations de style sans altérer le sens général du texte.

3.3.3 Le remplacement par synonymes à l'aide d'Arabic WordNet

Pour enrichir notre datasets, et introduire de la variété dans les formulations, nous avons utilisé Arabic WordNet pour remplacer des mots par des synonymes culturellement pertinents[28] cela enrichit le vocabulaire arabe tout en maintenant la cohérence sémantique[38].

Entrée : une phrase en arabe

Sortie : une liste de phrases reformulées avec des alternatives sémantiques

1-Découper la phrase en mots (tokenisation)

2. Pour chaque mot de la phrase :

a. Utiliser un analyseur morphologique (ISRIStemmer) pour extraire la racine

Exemple : "المدارس" → racine : "درس"

b. Interroger Arabic WordNet pour obtenir des alternatives sémantiques (synonymes, mots proches)

Exemple : racine درس->["المعاهد", "الجامعات"]

c. Enregistrer les correspondances : mot original → alternatives sémantiques

3. Générer des phrases reformulées en remplaçant les mots originaux par leurs alternatives

- Vérifier que les remplacements gardent un sens correct dans le contexte

- Éviter les formulations maladroites ou grammaticalement incorrectes

4. Retourner toutes les phrases reformulées valides

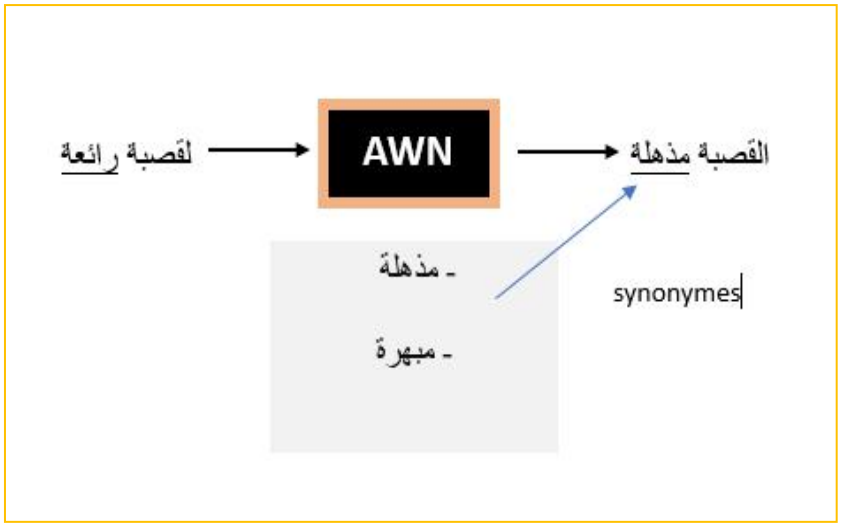

Figure 13 : Remplacement des synonymes via Arabic Wordnet

Les techniques sont appliquées séparément à chaque jeu de données, ainsi qu'à chaque section du texte (introduction, corps, conclusion), donc pour chaque entrée originale, entre 4 et 5 variantes augmentées sont générées, avec une vérification systématique pour éviter les doublons, et enfin les fichiers finals sont enregistrés au format JSONL, tout en préservant la structure initiale des données.

Tableau 8 : Taille des datasets après le data augmentation

Langue	**Enrichissement appliquée**	**Taille de dataset après l'enrichissement**
Français	Back Translation vers la langue anglais, la langue allemande, et espagnol.	543 article unique de gabarits en français
Arabe	-Back translation, -Insertion aléatoire de mots -Remplacement par synonymes à l'aide d'Arabic WordNet	600 article unique de gabarits en arabe

3.4. Validation de la Qualité des Données

Après avoir obtenu une taille acceptable pour le fine-tuning, et après avoir vérifié que tous les articles suivent un style de gabarit, nous avons utilisé des algorithmes de Deep

learning pour vérifier et corriger le dataset. Maintenant, nous pouvons commencer la deuxième étape de notre projet : le fine-tuning des LLMs pour la génération d'articles de gabarit sur le patrimoine algérien.

4.Sélection des Modèles de Langage

4.1 Aperçu général

Nous avons utilisé 80 % du dataset pour l'entraînement et 20 % pour le test, et ensuite, nous avons testé plusieurs modèles de langage et les avons évalués pour leur adéquation à notre tâche, Parmi eux :

4.1.1 GPT-2

Est un modèle de génération de texte développé par Open AI, qui repose uniquement sur l'empilement de blocs de décodeur Transformer [13], Contrairement aux architectures encodeur-décodeur classiques, GPT-2 fonctionne selon une approche autorégressive : il prédit chaque mot en se basant uniquement sur les précédents, ce qui le rend particulièrement adapté aux tâches de génération de texte libre. Chaque couche du modèle comprend un mécanisme d'attention masquée, une normalisation, et un réseau feed forward, le tout structuré avec des connexions résiduelles et une activation de type GELU.

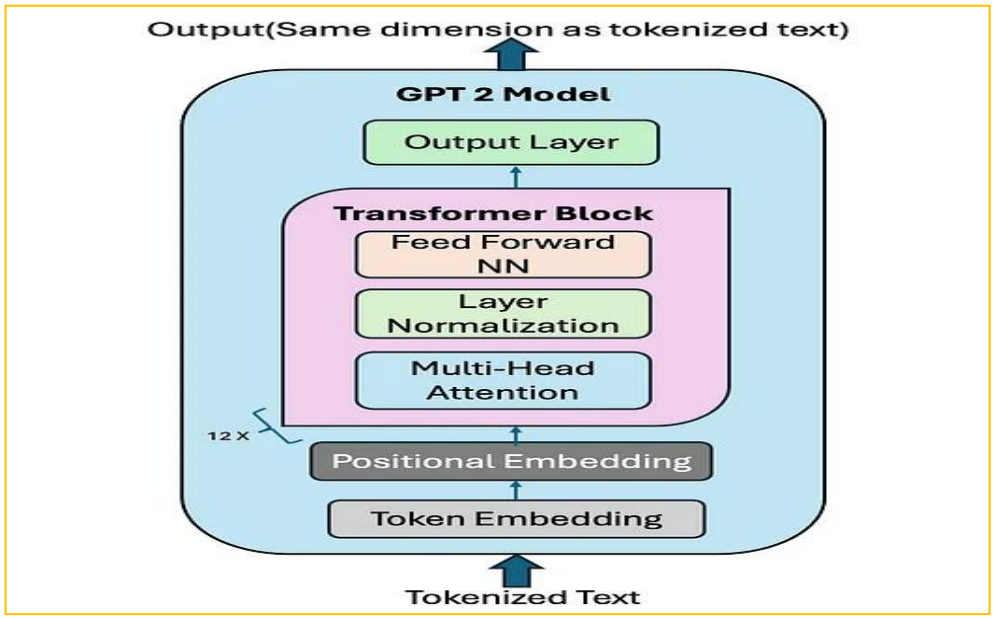

Figure 14 : Architecture de GPT2

Dans notre travail, nous avons utilisé deux variantes de GPT-2, chacune adaptée à une langue :

-Pour le français, nous avons utilisé le modèle FrGPT-2 (dbddv01/gpt2-french-small), entraîné sur des corpus francophones tels que OSCAR ou Wikipedia.

-Pour l'arabe, nous avons utilisé AraGPT2 (aubmindlab/aragpt2-base), un modèle pré entraîné sur un large éventail de textes arabes modernes (AUBMindLab).

4.1.2 T5 (Text-To-Text Transfer Transformer)

Est un modèle proposé par Google Research, qui adopte une approche originale : il convertit toutes les tâches NLP en un simple problème de génération de texte, quel que soit le type d'entrée [14] Qu'il s'agisse de traduction, de classification ou de résumé, tout est reformulé comme une tâche de "texte en entrée → texte en sortie". Cette formulation unifiée a permis à T5 d'être particulièrement polyvalent et performant dans des contextes multitâches.

Dans notre projet, nous avons testé trois variantes du modèle :

-**T5-base**, pour le français, une version multilingue adapté sur des corpus francophones, qui s'est montrée stable et globalement cohérente dans la génération d'articles structurés ;

-**T5-ar-base**, pour l'arabe, une version adaptée aux spécificités morpho-syntaxiques de la langue arabe. Cependant, cette version s'est révélée plus limitée, notamment en termes de fluidité et de précision culturelle ;

-**T5-small**, une version allégée de T5, utilisée ici à des fins de comparaison. Moins coûteuse en ressources, elle permet des générations rapides mais reste très limitée sur les textes patrimoniaux : manque de richesse, style trop simple, contenu souvent superficiel.

Ces variantes ont été sélectionnées pour leur diversité en termes de capacité, de langue, et de style de génération. Leur évaluation a permis de mesurer l'impact de la taille du modèle et de la langue cible sur la qualité des articles produits.

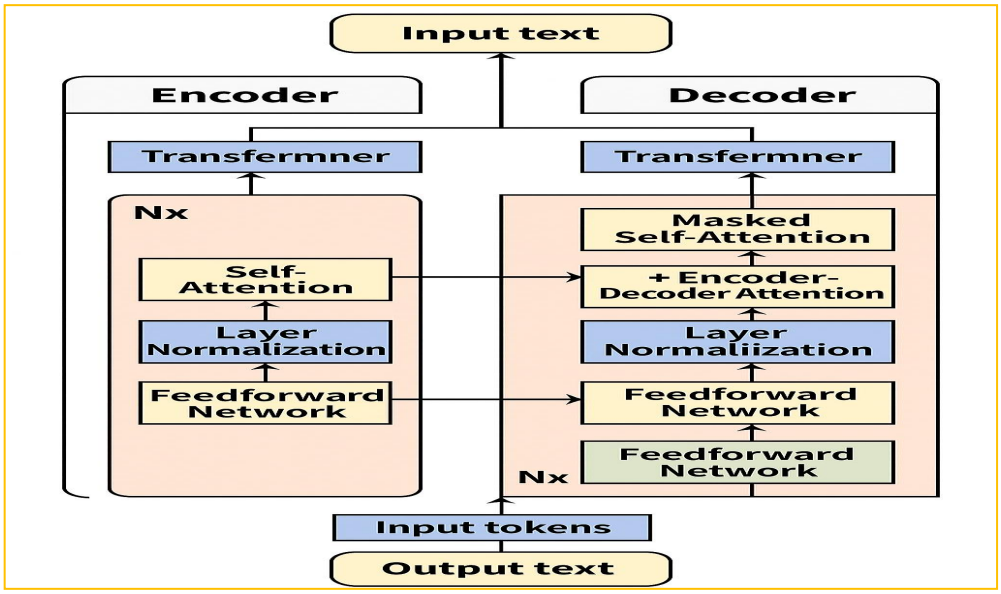

Figure 15 : Architecture de T5

4.2 comparaisons des architectures

Le Tableau suivant représente une comparaison entre les architectures.

Tableau 9 : comparaisons des architectures

Modèle	Type	Taille	Structure	Langue	Type d'entrée
GPT-2-FR	Decoder only	124M	Transformer	Français	Texte brut
GPT-2-AR	Decoder-only	124M	Transformer	Arabe	Texte brut
T5-small	Encoder-Decoder	60M	Transformer	Multilingue	Texte structuré
T5-base	Encoder-Decoder	220M	Transformer	Multilingue	Texte structuré
Ar-T5base	Encoder-Decoder	220M	Transformer	Arabe	Texte structuré

5.Ajustement Fin pour la Génération de Gabarits

Le fine tuning consiste à adapter les LLMs pour générer des articles encyclopédiques structurés selon des gabarits sur le patrimoine et la culture algérienne, en respectant un style neutre et en optimisant les performances pour les deux langues, français et arabe.

5.1 Méthodologie de fine tuning

Notre processus de fine-tuning commence par la sélection des modèles de langage (LLMs) adaptés à notre tâche. Ensuite, nous procédons à la configuration de l'environnement de calcul, notamment les ressources GPU nécessaires à l'entraînement.

Le fine-tuning est réalisé à l'aide d'instructions spécifiques, en français ou en arabe standard, ce qui le qualifie de fine-tuning supervisé. Afin d'optimiser les performances, nous mettons en œuvre une recherche par grille (*grid search*) pour identifier les hyperparamètres qui minimisent la fonction de perte (*loss*).

Enfin, la performance des modèles ainsi entraînés est évaluée à l'aide de plusieurs métriques pertinentes. La figure 17 ci-dessous illustre les différentes étapes de ce processus.

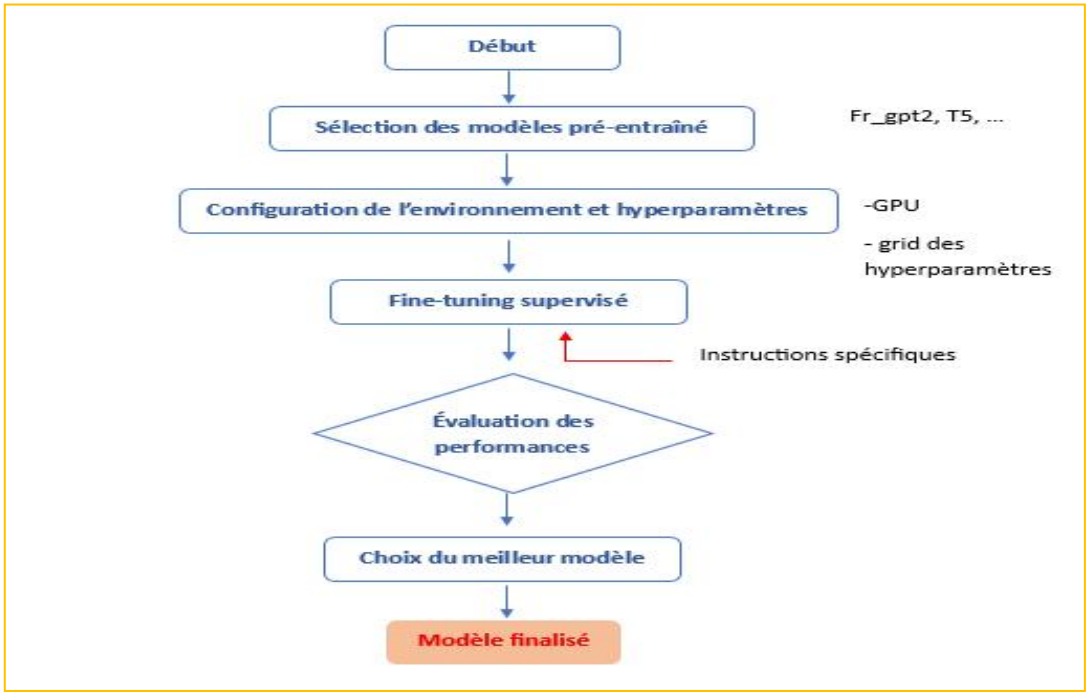

Figure 16 : Processus de Fine-Tuning

5. 2 Configuration des Hyperparamètres

Pour trouver les meilleurs hyperparamètres pour chaque modèle, nous avons appliqué la méthode de « Grid Search » mentionnée au chapitre 01, ainsi, pour tous les modèles, nous avons créé une grille contenant les valeurs possibles de chaque paramètre, ensuite, nous avons retenu les valeurs donnant les meilleurs résultats. Voici la grille utilisée :

Tableau 10 hyperparamètres

Paramètre	Valeurs
Learning rate	0.0003, 0.0005
Batch size	2
Époches	6, 8

Et voici les valeurs que on a obtenues pour chaque LLM :

5. 2 .1 Pour la langue française

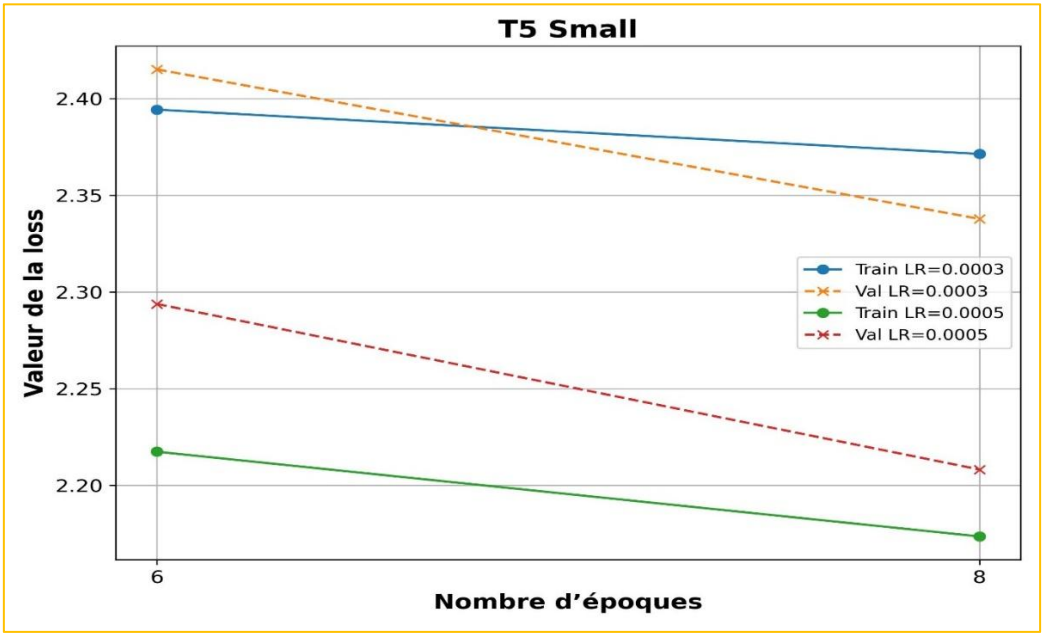

Figure 17 : Pertes du modèle T5 Small selon le taux d'apprentissage et les époques

Analyse :

Le modèle T5-small affiche des pertes relativement élevées dans l'ensemble, ce qui est attendu vu sa capacité réduite (~60M de paramètres).En comparant les courbes : Le passage de 6 à 8 époques améliore légèrement les performances ,Un learning rate de 0.0005 donne de meilleures pertes que 0.0003.La validation loss suit de près la training loss, ce qui est bon signe (peu de surapprentissage).

la combinaison choisie : ' Learning rate : 0.0005, Époques : 8' est la configuration qui donne la plus faible perte d'entraînement (2.17) et la validation loss la plus basse (2.20) parmi les essais. Elle montre une convergence raisonnable malgré les limites du modèle.

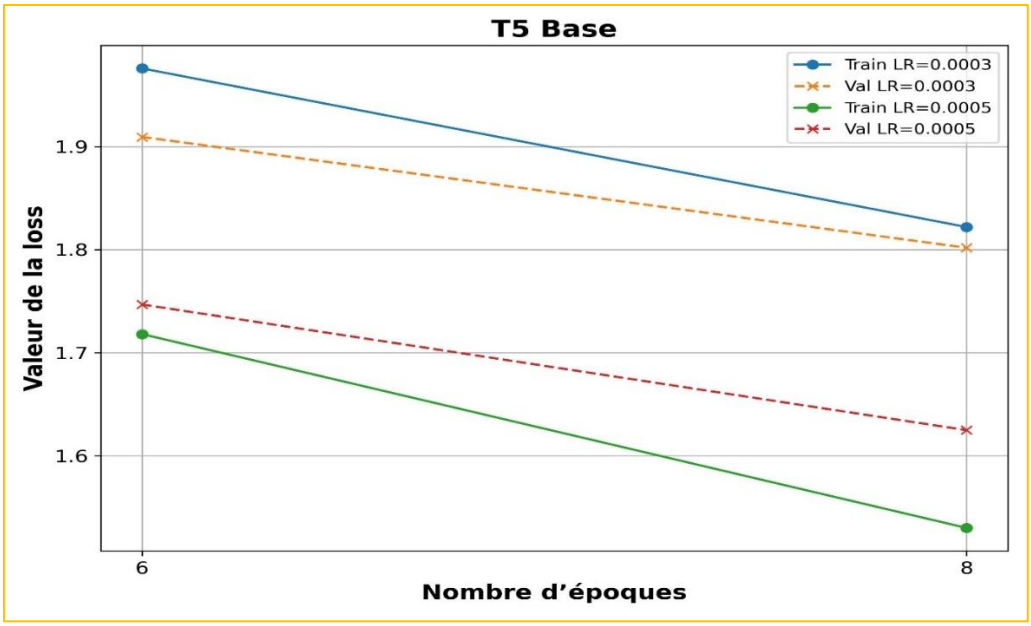

Figure 18 : Pertes du modèle T5 base selon le taux d'apprentissage et les époques

Analyse :

Les courbes montrent une amélioration nette entre 6 et 8 époques, avec une baisse continue des pertes. Le learning rate de 0.0005 est plus efficace : il permet une convergence plus rapide et profonde. Le modèle atteint une validation loss de 1.62, bien plus basse que celle de T5-small.

La Combinaison choisie : Learning rate : 0.0005, Époques : 8 est la combinaison qui donne à la fois la plus faible perte d'entraînement (1.53) et la validation loss la plus faible (1.62). Elle démontre une excellente capacité de généralisation sans surapprentissage apparent.

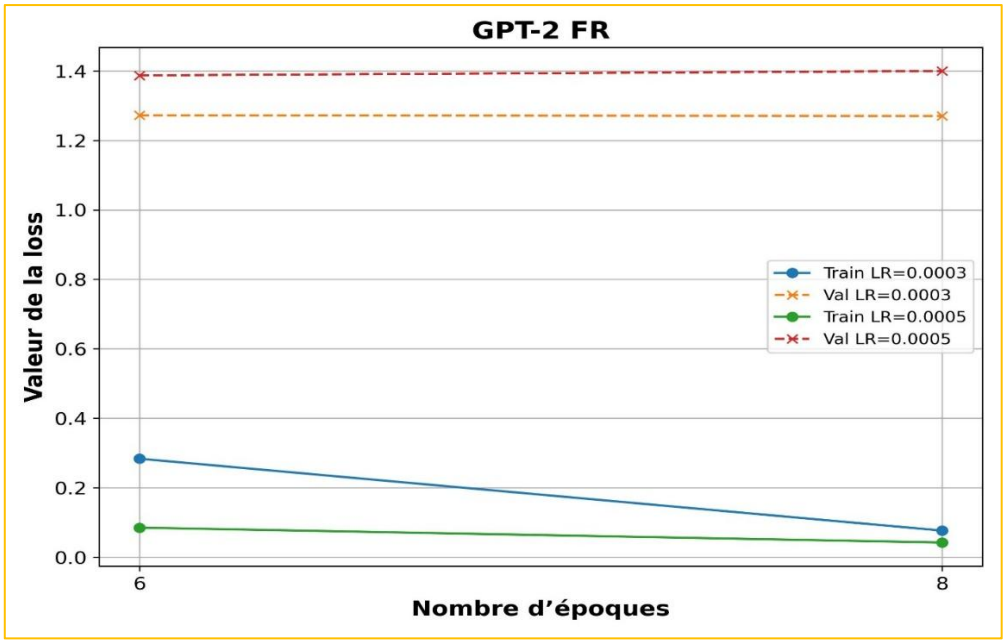

Figure 19 : Pertes du modèle GPT2-FR selon le taux d'apprentissage et les époques

Analyse

Le modèle GPT-2-FR atteint des pertes d'entraînement extrêmement basses, parfois proches de zéro, ce qui est un indice de surapprentissage.

En revanche, la validation loss stagne autour de 1.27–1.39, quelle que soit la configuration.

Un learning rate plus élevé (0.0005) accentue l'écart entre training et validation, ce qui confirme un surapprentissage.

La combinaison choisie est ' Learning rate : 0.0003, Époques : 8 '. Bien que la perte d'entraînement soit un peu plus élevée que pour 0.0005, la validation loss est la plus stable (1.27), Cela indique un meilleur équilibre entre apprentissage et généralisation, ce qui est crucial pour les modèles de type GPT.

5.2.2 Pour la langue arabe

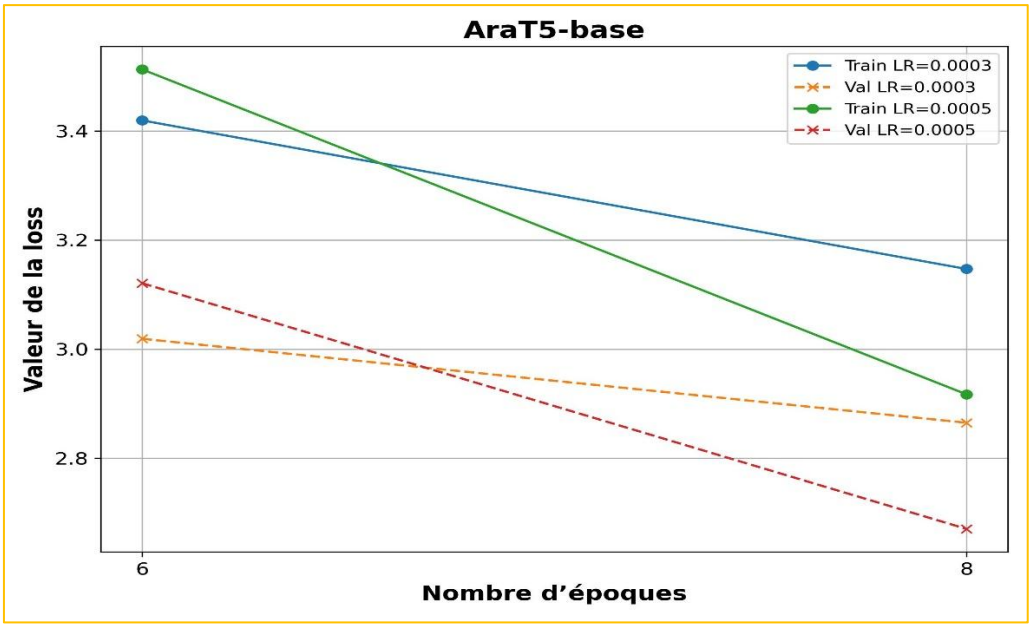

Figure 20 : Pertes du modèle AraT5 base selon le taux d'apprentissage et les époques

Analyse :

Le modèle AraT5-base présente des pertes initialement élevées, cohérentes avec sa taille plus importante (comme T5-base multilingue). L'augmentation des époques améliore clairement les performances, notamment avec un learning rate de 0.0005. La meilleure configuration donne un training loss de 2.91 et une validation loss de 2.67. La courbe montre une bonne progression de l'apprentissage, sans surapprentissage majeur.

la Combinaison ou Learning rate : 0.0005, Époques : 8, est la configuration avec la validation loss la plus basse, signalant une meilleure capacité à généraliser les données d'entraînement

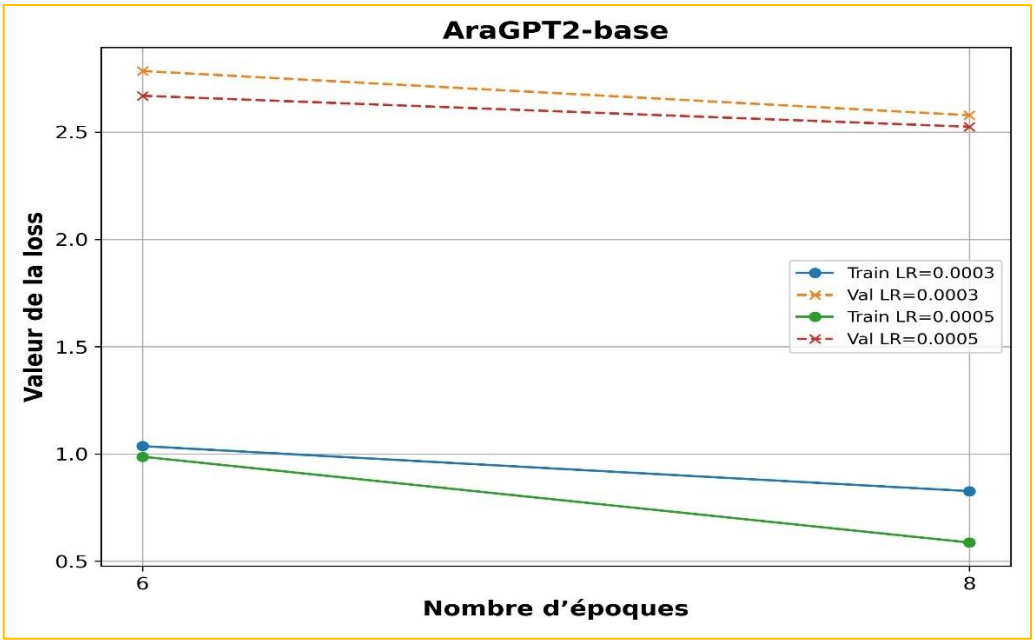

Figure 21 : Pertes du modèle AraGPT2-base selon le taux d'apprentissage et les époques

Analyse :

Le modèle AraGPT2-base montre une perte d'entraînement extrêmement faible (jusqu'à 0.58), mais avec une validation loss plus élevée que prévue, autour de 2.52 à 2.78.

Cela indique un surapprentissage potentiel, surtout avec le learning rate élevé (0.0005). Le meilleur équilibre est observé avec un learning rate de 0.0003 et 8 époques, mais la validation loss reste à 2.57, moins bonne que AraT5.

La combinaison choisie : Learning rate : 0.0005, Époques : 8

Même si la perte d'entraînement est très basse, cette combinaison reste la moins déséquilibrée (val_loss = 2.52), et exploite au mieux la capacité du modèle

6. Évaluation des performances des modèles

6.1. Méthodologie d'évaluation

Les grands modèles de langage ont une capacité à générer un texte incroyablement semblable à celui d'un humain, pour savoir si ces modèles sont réellement performants, on doit les évaluer et comparer leurs résultats.

L'évaluation de ces modèles, dans notre cas (pour la génération de contenu encyclopédique), nécessite une approche multidimensionnelle combinant :

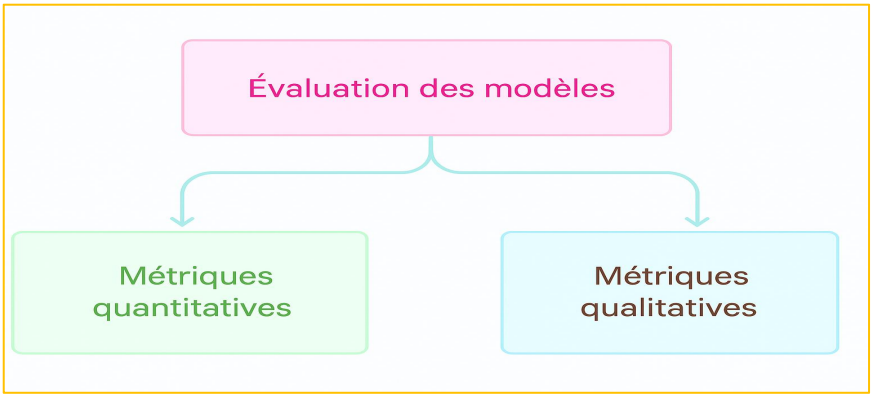

Figure 22 : Types d'évaluation des modèles (métriques quantitatives et qualitatives)

6.2 Métriques quantitatives

Les métriques quantitatives sont des mesures objectives utilisées pour évaluer la performance des modèles de génération de texte. Elles permettent de comparer différents modèles de manière reproductible et basée sur des calculs mathématiques.

Les mesures que nous avons utilisées, sont présentées ci-dessous :

6.2.1 La Perplexité

Comme présenté par Jurafsky et Martin [4] elle mesure l'incertitude d'un modèle de langage à prédire la suite d'un texte. Plus sa valeur est faible, plus le modèle est considéré comme bien adapté à la langue cible (ici le français ou l'arabe). En d'autres termes, une faible perplexité indique que le modèle attribue une forte probabilité aux mots réellement présents dans le texte. Formellement, elle est définie comme :

$$PP(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i | w_1, .. w_{i-1})}}$$

Tels que :

- *N : le nombre total de mots dans la séquence W.*

- $W = w_1, .., w_N$ *est la séquence de mots étudiée.*
- $P(w_i|w_1, .. w_{i-1})$ *est la probabilité d'apparition du mot wi sachant le contexte précédent.*

6.2.2 Score BLEU

Lamesure BLEU score **[29]** mesure la précision des n-grammes du texte généré par rapport à une référence, avec une pénalité de brièveté pour sanctionner les textes trop courts. Le score BLEU évalue essentiellement combien de mots du texte généré apparaissent dans le texte de référence rédigé par un humain. L'idée de base est de mesurer la précision en comptant le nombre de n-grammes (combinaisons de n mots) partagés entre le texte généré et la référence, une implémentation standard destinée à un usage cohérent entre chercheurs ; c'est cette version que nous utilisons dans notre analyse [42].

6.2.3. Score ROUGE-L

La mesure ROUGE-L score **[30]**, évalue la qualité du texte généré en se basant sur la plus longue sous-séquence commune (LCS) entre le texte généré et le texte de référence. Contrairement aux métriques qui se fondent uniquement sur des correspondances exactes de n-grammes, ROUGE-L prend en compte l'ordre des mots, ce qui le rend particulièrement utile pour évaluer la qualité globale et la cohérence du contenu généré. L'idée centrale est de mesurer le rappel et la précision de la LCS, en capturant ainsi à la fois la couverture du contenu attendu et la fidélité à la structure du texte de référence ; nous utilisons donc l'implémentation standardisée de ROUGE-L afin d'assurer la cohérence des évaluations dans notre analyse

6.3 Métriques qualitatives et Évaluation Humaine

Les métriques qualitatives s'intéressent à ce que les chiffres, à eux seuls, ne peuvent pas vraiment dire. Elles regardent le fond du texte : est-ce que ça a du sens ? Est-ce que c'est bien écrit ? Est-ce que ça sonne juste, culturellement parlant ?

On parle ici de dimensions plus subtiles, la pertinence du contenu, la cohérence des idées, la qualité de la langue. Autrement dit, tout ce qui fait qu'un article est non seulement lisible, mais crédible et intéressant.

Et pour ça, les scores automatiques ne suffisent pas. Il faut des yeux humains. Des gens qui connaissent le sujet, qui comprennent les références culturelles, qui sentent quand une phrase sonne faux ou quand une idée est mal amenée.

Dans notre cas, ce sont des experts du patrimoine algérien qui jouent ce rôle. Ils apportent un regard précis, critique, et surtout ancré dans la réalité du terrain. Ils peuvent repérer une maladresse dans la formulation, une confusion historique, ou une interprétation un peu bancale, des choses qu'aucune métrique ne peut détecter seule.

C'est ce qu'on appelle une approche Human-in-the-loop : l'humain reste au cœur du processus, pour garantir que ce que le modèle produit soit non seulement techniquement correct, mais surtout pertinent, riche et fidèle à la culture qu'il cherche à représenter.

C'est cette boucle entre la machine et l'expertise humaine qui fait toute la différence, et qui nous permet de produire des textes vraiment solides.

6 .3.1 Domaines d'évaluation :

Tableau 11 : domaines d'évaluation

Critère	Description
Pertinence culturelle	Le contenu reflète-t-il fidèlement le patrimoine algérien ?
Exactitude historique	Les faits, dates et noms sont-ils corrects ?
Structure logique	L'article suit-il le gabarit (introduction/développement/conclusion) ?
Style encyclopédique	Le ton est-il neutre, objectif et adapté à un public général ?
Fluidité linguistique	Le texte est-il naturel en arabe/français (grammaire, syntaxe)

6.4 Protocole d'Évaluation Comparative

Lorsque nous somme dans un contexte comme le patrimoine culturel algérien, il ne suffit pas de sélectionner le modèle ayant obtenu le meilleur score automatique, c'est une approche simple, elle est souvent insuffisante et inadaptée à la complexité du contenu patrimonial.

donc les modéles ont été comparés à l'aide d'une méthode d'évaluation combinant des critères quantitatifs et qualitatifs,

-les scores quantitatifs (BLEU, ROUGE-L, perplexité) permettent de mesurer la performance brute des modèles : fluidité des textes générés, proximité avec les textes de référence, etc. Ces métriques sont 'objectives, ne suffisent pas seules, et elles ont été pondérées à hauteur de 60 % dans notre évaluation globale.

-l'évaluation humaine a été menée par des experts du patrimoine algérien, ces derniers ont analysé les textes produits selon plusieurs critères : crédibilité du contenu, respect des référents culturels, et nature du style, Cette analyse est importante pour saisir la dimension culturelle et contextuelle des articles,elle a été intégrée à hauteur de 40 % dans le score final.

en fin, chaque modèle reçoit une note globale sur 100, reflétant à la fois ses performances techniques et la pertinence culturelle de ses productions.

Cette approche hybride, croisant métriques automatiques et jugement humain, constitue un cadre rigoureux et équilibré pour le choix du modèle le plus adapté à notre objectif : générer des articles fiables, bien rédigés et culturellement pertinents.

Figure 23 protocoles d'évaluation

7. Résultats de l'Évaluation des Modèles

Nous présentons ici les résultats de l'évaluation comparative des différents modèles, obtenus en suivant rigoureusement notre protocole. Cette évaluation s'appuie à la fois sur des métriques quantitatives, calculées à partir de jeux de test de taille croissante (50 puis 100 articles), et sur une analyse qualitative menée par des experts.

En croisant ces deux types de données, les chiffres d'un côté, jugement humain de l'autre. Nous avons pu identifier le modèle qui s'en sort le mieux pour la tâche de génération d'articles sur le patrimoine culturel algérien. Ce modèle, jugé le plus performant à la fois sur le fond et sur la forme, a été retenu pour la suite du projet.

7.1 Résultats Quantitatifs

Nous avons évalué les performances des modèles sur deux jeux de test distincts, composés de 50 et 100 articles. Les métriques ont été calculées séparément pour chaque langue afin de tenir compte des spécificités linguistiques du français et de l'arabe.

Les tableaux suivants présentent les résultats obtenus pour chaque taille de jeu de test. On commence ici par les modèles adaptés sur le dataset en français, avec une analyse détaillée des scores pour chaque critère.

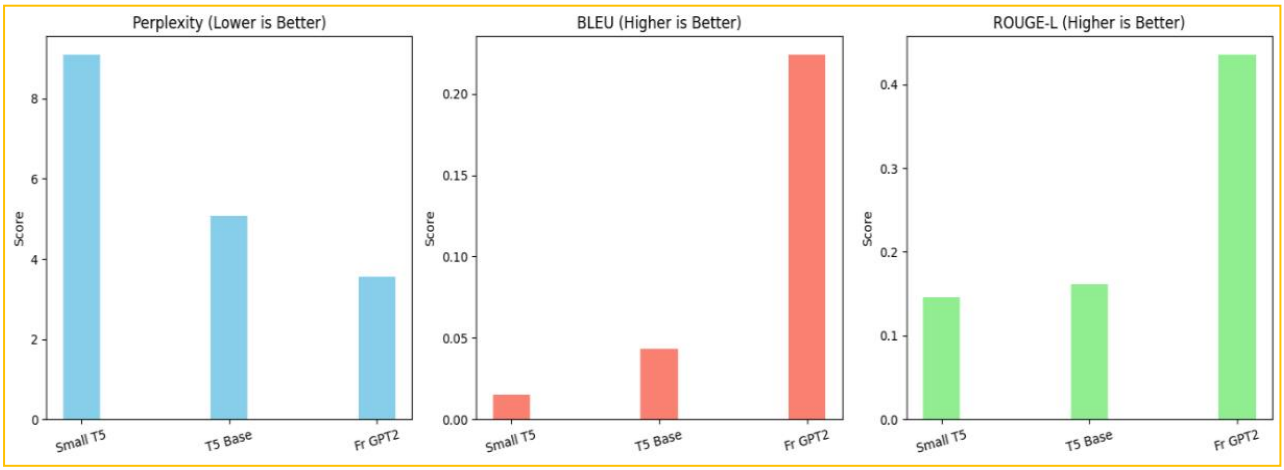

Figure 24 performances des LLMs pour le dataset en français

Analyse Quantitative Pour la langue français :

En observant le tableau et l'histogramme, on voit que Fr-GPT2 obtient une perplexité très faible (3.56), ainsi que les meilleurs scores en BLEU (0.25 et 0.22) et ROUGE-L (0.47 et 0.43). Il présente donc les meilleures performances, ce qui signifie que les textes générés sont précis et proches des articles d'entraînement.

Concernant T5-Base et Small-T5, ils offrent des performances acceptables mais limitées. T5-Base affiche une perplexité de 5.08, avec des scores BLEU autour de 0.043 et ROUGE-L proches de 0.16. Small-T5, quant à lui, atteint une perplexité de 9.10, avec un BLEU de 0.013 à 0.015 et un ROUGE-L d'environ 0.14. Ces résultats restent en dessous de ceux de Fr-GPT2, ce qui les rend moins adaptés pour la génération d'articles structurés en français.

Pour les modèles adaptés sur le dataset en arabe :

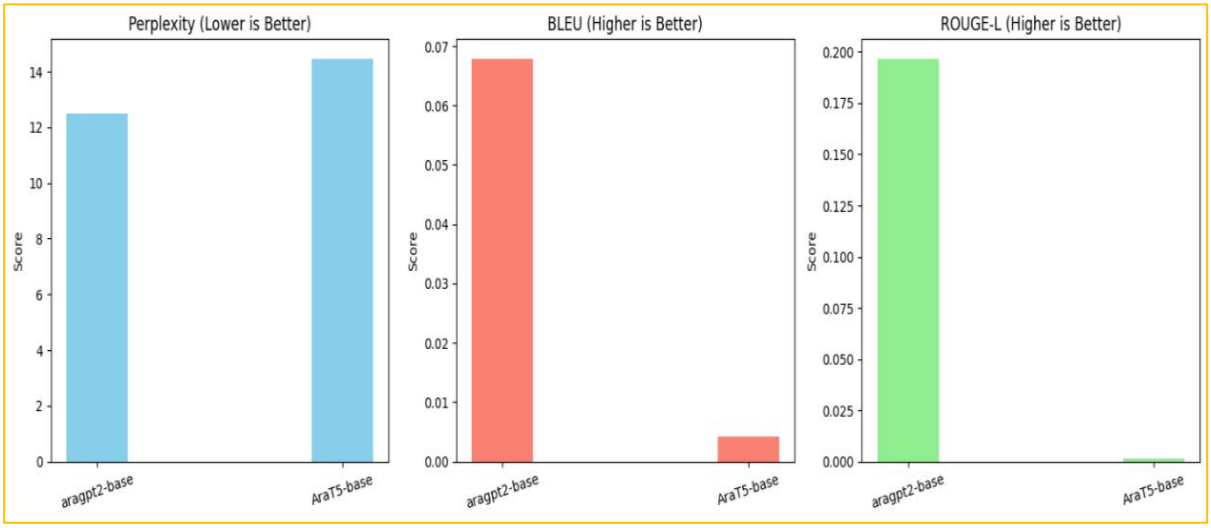

Figure 25 performances des LLMs pour le dataset en arabe

Analyse Quantitative pour la langue arabe :

D'après l'histogramme précédent, on observe que AraT5-base présente une faible performance : ses scores sont très petits, avec une perplexité de 14.44, un score BLEU avoisinant 0.004, et un ROUGE-L inférieur à 0.002. Cela signifie que les textes générés ne sont pas précis et ne ressemblent pas aux articles d'entraînement. En revanche, le deuxième

LLM, AraGPT2-base, offre un meilleur équilibre, combinant une perplexité plus faible (12.49) avec des scores BLEU autour de 0.07 et ROUGE-L proches de 0.19.

Meilleur LLMs avec évaluation quantitative seulement :

Tableau 12 Meilleur LLMs avec évaluation quantitative seulement

Langue	LLMs choisi
Français	Fr-gpt2
Arabe	Aragpt2-base

7.2Résultats Qualitatifs

Pour évaluer la vraie qualité des articles générés, on ne s'est pas contenté de scores automatiques. On a voulu savoir ce qu'ils valent aux yeux de vrais connaisseurs. C'est pourquoi nous avons fait appel à deux experts du patrimoine algérien, des gens qui connaissent profondément l'histoire, les coutumes, les objets, les mots… tout ce qui fait l'âme de notre culture.

Ils ont lu chaque article avec un œil attentif et exigeant, et les ont notés selon cinq critères mentionnés en section 6.3.1.

Chacun de ces points a été noté sur 5, individuellement par les deux experts. Ensuite, pour chaque critère, on a pris la moyenne des deux avis. Et pour avoir une vision globale de chaque article, on a fait une moyenne des cinq critères.

Ce processus nous a permis de garder une évaluation humaine, fine, nuancée. Une sorte de "filet de sécurité" que les métriques automatiques ne peuvent pas offrir. Grâce à ces regards humains, on ne mesure pas seulement si le texte est "statistiquement bon", mais s'il est vraiment pertinent, crédible, et respectueux du patrimoine qu'il est censé valoriser.

Voir le résumé de ces étapes dans le pseudo-code code en figure 26:

```
procedure evaluate_articles(articles, experts)
1   for each article in articles
2       scores ← []
3       for each expert in experts
4           scores.append([
5               expert.rate_cultural_relevance(article),
6               expert.rate_historical_accuracy(article),
7               expert.rate_logical_structure(article),
8               expert.rate_encyclopedic_style(article),
9               expert.rate_linguistic_fluency(article)
10          ])
11      end for
12      average_scores ← average_across_criteria(scores)
13      article.overall_score ← average(average_scores)
14      store_results(article, article.overall_score)
15  end for
16  analyze_results(articles)
```

Figure 26 Pseudocode pour l'évaluation qualitative des articles par des experts

Cette méthode permet d'avoir une évaluation équilibrée et précise de la qualité des articles selon plusieurs aspects essentiels, et voici les résultats dans le tableau 13, qui nous aident à comparer la performance du modèle sur différents sujets :

Tableau 13 Résultats qualitatifs de l'évaluation humaine

Modèle	Moyenne pondérée des notes des experts					Moyenne finale globale
	culturelle	Style encyclopédique	Fluidité linguistique	Exactitude historique	Structure logique	
Français						
Small-T5	2.7	1.8	1.9	1.71	2.2	2.062
T5-base	2.78	2.45	2.64	2.5	3	2.674
Frgpt-2	3.025	2.6	2.71	2.625	3.5	2.892
Arabe						
T5-ar-base	1.1	1	0.9	0.9	0.9	0.96
Ar-gpt2-base	2.2	2.4	2	1.8	2.2	2.12

Analyse Qualitative :

D’après les experts, les différences entre les modèles sont claires. En français, les textes générés par Fr-gpt2 sont des textes bien écrits, cohérents et adoptent le bon ton. T5 base s’en sort, mais manque parfois de précision. T5-small, lui, peine à convaincre avec des textes fades et souvent hors sujet.

En arabe, Ar-gpt2-base reste lisible sans impressionner, tandis que déçoit T5-ar-base franchement : formulations maladroites, idées floues, et peu de sensibilité culturelle. L’avis humain est donc essentiel pour évaluer la qualité réelle.

8. Sélection du Meilleur Modèle

Pour obtenir une évaluation plus complète et fidèle de la qualité des contenus générés, nous avons choisi de combiner des métriques automatiques avec des jugements humains.

Les métriques automatiques, telles que la perplexité, le BLEU [29] ou le ROUGE-L [30], permettent d’évaluer rapidement et objectivement certains aspects techniques du texte généré, notamment la fluidité, la pertinence sémantique ou la ressemblance avec une référence. Cependant, ces métriques restent limitées, car elles ne prennent pas en compte des dimensions plus subjectives, telles que la cohérence globale, l’adéquation culturelle ou l’exactitude historique [31], [32].

C’est la raison pour laquelle une évaluation humaine est également indispensable. Des experts peuvent juger de la pertinence, de la cohérence, du style ou de la valeur informative d'un texte généré, ce que les métriques automatiques peinent généralement à appréhender [33], [34].

Ainsi, afin d’obtenir une évaluation équilibrée, fidèle et exhaustive de la qualité des textes générés, nous avons opté pour une moyenne pondérée, en donnant 60 % de poids aux métriques automatiques et 40 % à l’évaluation humaine [35]. Cette approche prend donc en compte tanto les aspects objectifs que subjectifs de la génération de langage naturel.

Nous combinons Ces résultats quantitatifs et qualitatifs selon cette formule :

$$Score\_globale = 0,6 * (\frac{Preplexity_{norm} + BLEU_{norm} + ROUGE - L_{norm}}{3}) + 0,4 * (\frac{\sum Critéres}{5})$$

où :

- La perplexité normalisée :

$$Preplexity = 1 - \frac{PP - PP_{min}}{PP_{max} - PP_{min}}$$

- Le score BLEU normalisé :

$$BLEU_{norm} = \frac{BLEU - BLEU_{min}}{BLEU_{max} - BLEU_{min}}$$

-Le score ROUGE-L normalisé

$$ROUGE - L_{norm} = \frac{ROUGE - L - ROUGE - L_{min}}{ROUGE - L_{max} - ROUGE - L_{min}}$$

La normalisation permet de mettre les scores sur une échelle comparable entre 0 et 1, en tenant compte que la perplexité est un indicateur inversé (plus elle est basse, mieux c'est), tandis que BLEU et ROUGE sont positivement corrélés à la qualité.

Et la figure suivante présente ces scores globaux pour le jeu de test de 100 articles :

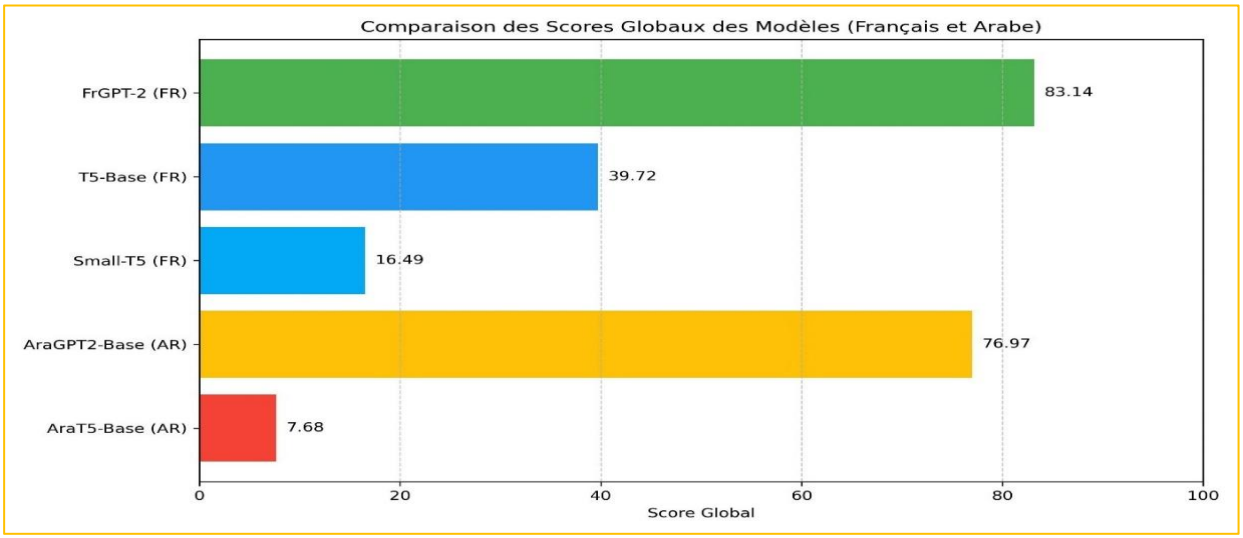

Figure 27 : Comparaison des Scores globaux des LLMs

Choix finaux :

Pour la langue française, d'après le score combiné, qui prend en compte la qualité et la quantité, Fr-GPT2 est le modèle le plus performant. De même, pour la langue arabe, Ar-GPT2 base s'est révélé le meilleur, car il obtient le score combiné le plus élevé.

9. Problèmes rencontrés lors de l'entraînement du modèle

Nous avons rencontré plusieurs difficultés pendant l'entraînement du modèle :

-Après avoir mieux préparé les données et bien organisé l'entraînement, les choses se sont stabilisées. On a aussi remarqué que le modèle apprenait trop bien certains types d'articles qu'il voyait souvent, ce qui l'empêchait de bien s'adapter à d'autres sujets moins présents.

-Certaines données posaient problème : il y avait des articles mal rédigés, parfois avec un mélange de langues ou un format incohérent. Cela a nécessité un gros travail de nettoyage avant de pouvoir les utiliser.

-On s'est rendu compte que les scores automatiques comme BLEU ou ROUGE ne suffisaient pas pour juger la qualité réelle des textes, surtout d'un point de vue culturel ou encyclopédique, c'est pour ça qu'on a décidé d'ajouter une évaluation humaine et l'active learning,

-Comme on a travaillé sur Kaggle, on a été limités par le temps et la puissance de calcul.

10.Conclusion

Dans ce chapitre, nous avons exposé en détail l'architecture globale que nous avons conçue et développée dans le but de générer des articles structurés concernant le patrimoine et la culture algérienne en arabe ou en français. Nous avons commencé par présenter comment nous avons collecté les datasets via web scraping. Ensuite, nous avons détaillé le processus de fine-tuning et le processus de choix des hyperparamètres. Deux grandes architectures ont été utilisées, T5 et GPT-2. Nous avons fait comparaison des

performances selon deux métriques, quantitative et qualitative. Nous avons combiné ces deux méthodes et, enfin, nous choisissons un modèle pour chaque langue qui donne les meilleurs résultats.

CHAPITRE III : VALIDATION ET ANALYSE DES RÉSULTATS

1. Introduction :

Le chapitre précédent a présenté en détail l'architecture des différents modèles que nous avons testés, ainsi que la démarche suivie pour évaluer chacun d'eux et identifier le meilleur. Par la suite, ce chapitre est consacré au processus de mise en œuvre de ces modèles, nous commençons par une description des outils utilisés pour leur construction, puis nous aborderons leur intégration dans une application web. Enfin, nous évoquerons les perspectives pour les travaux futurs.

2.Les outils utilisés

Outils et APIs	Description
Google collab[1] 	Est une plateforme en ligne gratuite proposée par Google, conçue pour faciliter l'apprentissage et la recherche en intelligence artificielle. Basée sur l'environnement Jupyter Notebook, elle permet d'entraîner des modèles de Machine Learning directement dans le cloud à partir d'un navigateur web.
Python[2] 	C'est un langage de haut niveau, simple à lire et à écrire, ce qui facilite son apprentissage, et il est open source, donc libre d'utilisation et de modification.
VS Code[3] 	C'est un éditeur de code source à la fois léger et puissant, compatible avec de nombreux langages de programmation.

[1] *https://colab.research.google.com/*

[2] *https://www.python.org/*

[3] *https://code.visualstudio.com/*

Kaggle[4]	Kaggle est une communauté en ligne regroupant des experts en données et des apprenants en machine learning, un environnement de travail en ligne à la science des données, et des ressources pédagogiques en IA.
BeautifulSoup[5]	Est une bibliothèque Python conçue pour l'extraction de données à partir de fichiers HTML et XML, Utilisée principalement pour le web scraping.
PyTorch	Est une bibliothèque d'intelligence artificielle. Utilisée principalement pour le développement de modèles d'apprentissage profond (deep learning) et de réseaux de neurones artificiels.
Arabic WordNet[6]	WordNet est une base de données lexicales qui représentent les lemmes (lexèmes, mots) de la langue arabe, ainsi que leurs significations, organisées en un réseau lexico-sémantique.
Flask[7]	Flask est un micro-framework web écrit en Python, Il facilite le développement rapide d'applications web, d'API REST et de services backend. Flask s'appuie sur Werkzeug pour le routage et la gestion des requests, et sur Jinja2 pour le rendu de modèles HTML.
Googletrans[8]	Est une bibliothèque Python non officielle qui fournit une interface simple à l'API AJAX de Google Translate. Elle permet de traduire du texte entre plusieurs langues, de détecter la langue d'un texte, et de faire des traductions par lot.
SQLAlchemy[9]	Est une bibliothèque Python puissante pour la gestion des bases de données relationnelles. Elle est compatible avec de nombreux moteurs de

[4] *https://www.kaggle.com/*

[5] *https://pypi.org/project/beautifulsoup4/*

[6] *https://globalwordnet.org/resources/arabic-wordnet*

[7] *https://flask.palletsprojects.com/*

[8] *https://pypi.org/project/googletrans/*

[9] *https://www.sqlalchemy.org/*

	bases de données.

3.Validation collaborative

Dans le chapitre et les sections précédentes, on a mis en détail comment on a choisi et sélectionné les modèles de langage (LLM) qui offrent les meilleurs résultats sur notre dataset. Nous les utilisons pour générer des articles de gabarits sur le patrimoine culturel de notre pays, mais pour garantir la qualité finale de chaque article seul, on a besoin de : la validation humaine.

3 .1 Validation humaine basée sur les critères

Afin d'assurer la qualité et la pertinence des articles générés sur le patrimoine culturel algérien, nous avons intégré une phase de validation humaine réalisée par des experts du domaine. Ces évaluateurs relisent les textes, détectent les incohérences historiques ou linguistiques, et attribuent une note qualitative selon cinq critères définis : pertinence culturelle, exactitude historique, cohérence structurelle, style encyclopédique et fluidité linguistique.

Voici un article généré par le système et son évaluation :

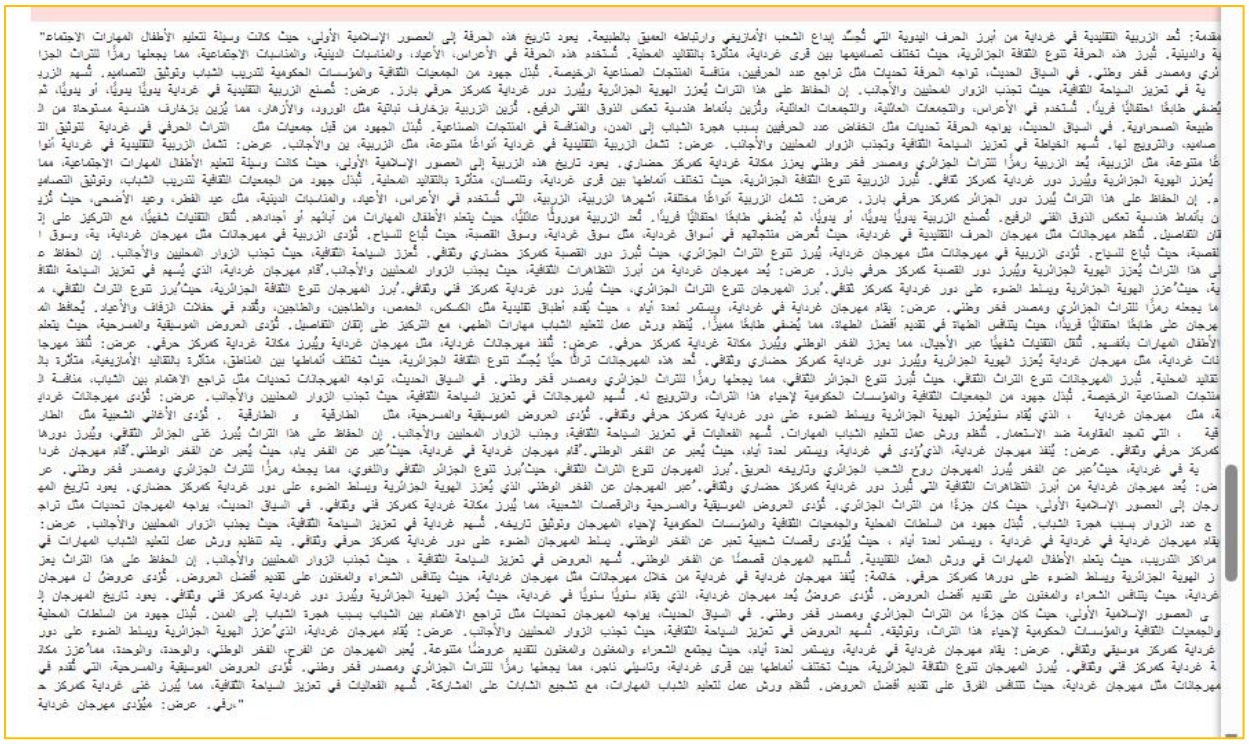

Figure 28 : un article généré sur "الزربية التقليدية في غرداية"

Tableau 14 exemple d'une évaluation humaine sur un article généré

Note globale	3.1/ 5
Avis	Le sujet est intéressant et bien ancré dans le contexte, mais le texte manque d'organisation. Trop de répétitions et de phrases longues rendent la lecture lourde. Avec plus de structure et de clarté, le contenu serait bien plus efficace.

3.2 Validation via l'analyse des sentiments

Dans le but d'accélérer le processus de validation humaine des articles générés, nous avons intégré une composante d'analyse des sentiments servant d'outil auxiliaire. Cette approche n'a pas vocation à remplacer l'évaluation humaine ni à fournir une mesure absolue de la qualité des articles, mais plutôt à classifier automatiquement les retours critiques selon leur tonalité globale (positive ou négative), afin de prioriser les textes à réévaluer manuellement.

Nous avons entraîné un modèle DistilBERT sur le dataset IMDb Reviews, un corpus bien établi dans le domaine du NLP, contenant des avis textuels annotés selon une polarité binaire. Le choix de ce dataset repose sur plusieurs critères :

-Les commentaires y sont rédigés librement, dans un langage naturel, ce qui reflète une diversité lexicale et stylistique intéressante ;

-La structure binaire des étiquettes facilite l'entraînement rapide d'un classifieur robuste ;

-Le dataset est équilibré, largement utilisé et reconnu pour ses performances stables.

Nous sommes conscients que ce domaine d'origine (critiques de films) diffère significativement du contexte de notre projet (commentaires sur des articles culturels). C'est pourquoi ce système a été conçu comme une expérimentation exploratoire, dans l'objectif de tester le potentiel de cette approche pour trier les retours selon leur tonalité générale.

Par exemple, une remarque experte telle que :

« Le texte manque de profondeur. Il effleure des idées intéressantes sans les développer, et la structure est confuse. », a été correctement classée comme négative, avec un score de confiance de 98.44%, démontrant la capacité du modèle à détecter les critiques nuancées, même hors domaine source.

Ce module permet donc :

-de trier les retours selon leur orientation émotionnelle ;

-de prioriser les articles les plus problématiques ;

-de faciliter le travail des experts en leur signalant en amont les cas les plus sensibles.

4. Active learning

4.1 Principe

Dans une logique d'amélioration continue, les articles validés par des experts ne sont pas seulement archivés, mais réutilisés dans une stratégie d'apprentissage actif (active learning). Le modèle est ainsi réentraîné à partir des textes jugés de haute qualité : bien structurés, pertinents et culturellement riches. Ce processus permet d'exploiter les retours

humains pour guider le modèle là où ses performances sont encore faibles. On favorise une boucle de rétroaction itérative où les données corrigées renforcent les générations futures.

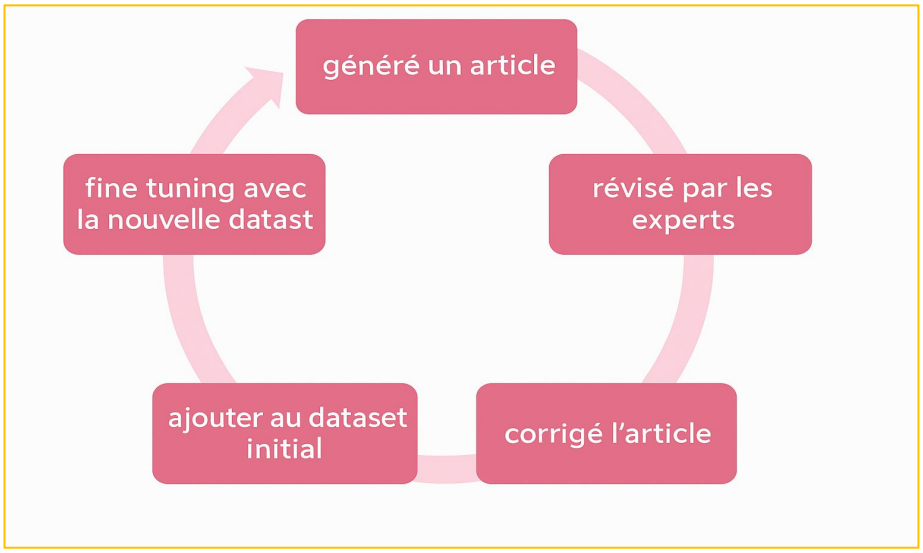

Figure 29 Human-in-the loop

4.2 Score d'incertitude

Afin de guider efficacement le processus d'apprentissage actif et de renforcer la robustesse du système, nous avons conçu un score d'incertitude combinant plusieurs

Indicateurs de performance, aussi bien quantitatifs que qualitatifs. Ce score permet de détecter les articles « incertains », c'est-à-dire ceux dont la qualité est potentiellement faible ou instable afin de les prioriser dans le cycle de correction humaine.

Le score d'incertitude U est défini comme suit :

$$U = 0{,}2(1 - Prep\_norm) + 0{,}2(1 - BLEU) + 0{,}2(1 - ROUGE) + 0{,}4(1 - Quality)$$

Tels que :

- **Perplexité** : mesure de fluidité et de confiance du modèle lors de la génération ;
- **BLEU / ROUGE-L** : métriques de similarité avec une référence humaine ;

- **Qualité** : note obtenue soit à partir d'une évaluation humaine (si disponible), soit estimée à l'aide du score de sentiment analysé si un retour textuel existe ('has_reviews = True'), ou bien fixée par défaut à 0,5 en l'absence de retour.

Le poids de 0,4 attribué à la composante qualitative reflète l'importance du jugement humain dans un contexte culturel comme le nôtre, où des éléments comme la justesse historique ou la tonalité linguistique sont parfois difficiles à quantifier par des métriques automatiques.

L'utilisation combinée de ces quatre dimensions permet d'obtenir un indicateur synthétique qui oriente les efforts de révision vers les articles les plus problématiques, tout en assurant un équilibre entre mesures objectives et appréciations humaines.

4.3 Résultats obtenus pour une itération de l'actif learning sur le dataset Arabe

Une première itération de l'apprentissage actif a été réalisée sur un lot d'articles générés en arabe. Les 10 articles ayant obtenu les scores d'incertitude les plus élevés ont été sélectionnés, signalant un niveau élevé de doute du modèle (forte perplexité, faibles scores de similarité).

Ces textes ont été revus et corrigés par un expert humain, puis réinjectés dans le corpus d'entraînement. Le modèle a ensuite été réentraîné avec ces données enrichies.

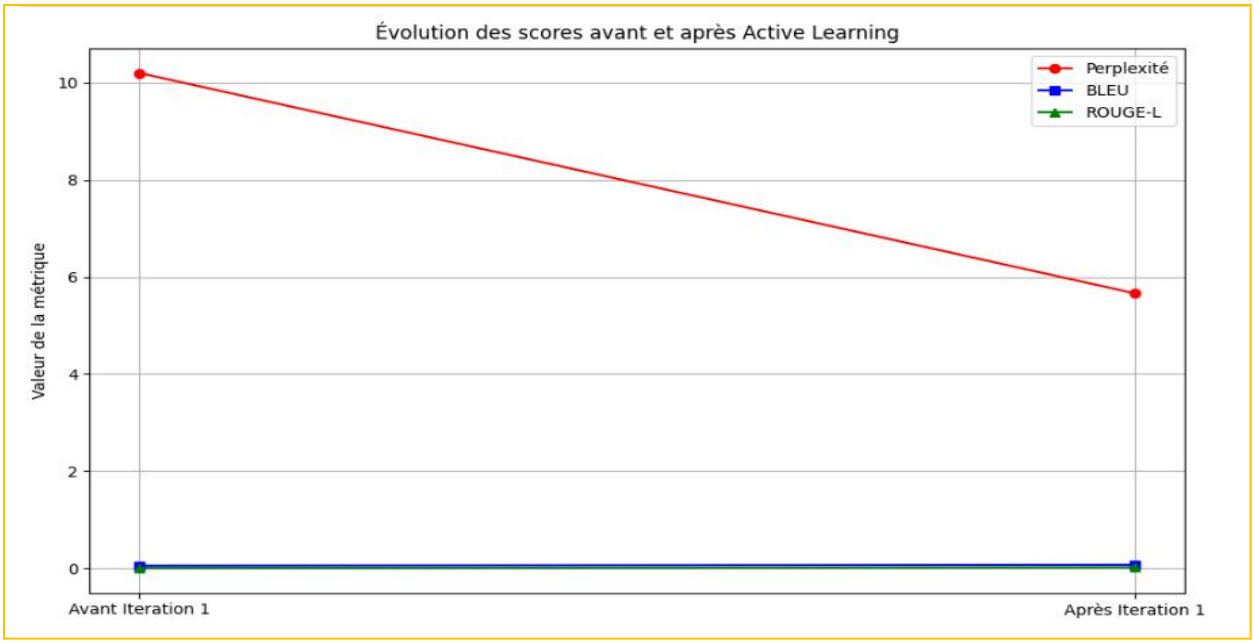

Figure 30 Evolution des scores avant et après Active Learning

Après une première itération d'apprentissage actif, nous avons observé une amélioration modeste mais significative sur les principales métriques d'évaluation : la perplexité est passée de 10.20 à 5.66, tandis que les scores BLEU et ROUGE-L sont passés respectivement de 0.0497 à 0.0682 et de 0.000 à 0.0153.

ces chiffres sont faibles a plusieurs raisons, d'abord la langue arabe pose des défis pour les métriques classiques (BLEU et ROUGE), qui sont très sensibles à la variation lexicale et à l'ordre des mots. Ensuite, les textes générés sont de nature culturelle et narrative, avec des phrases longues et riches en synonymes. Cela rend les comparaisons mot-à-mot moins pertinentes qu'en traduction. Enfin, la référence humaine utilisée pour le calcul de ces scores était unique, alors que plusieurs formulations peuvent être valides. Cela limite l'efficacité de l'évaluation automatique.

La baisse importante de la perplexité est, est un signal clair que le modèle devient plus confiant et plus fluide dans sa génération. C'est un indicateur de progression importante, surtout dans un contexte multilingue où l'arabe n'est souvent pas priorisé dans les LLMs génériques. même avec les scores BLEU ou ROUGE restent faibles , la combinaison avec la validation humaine et l'amélioration de la perplexité confirme que le cycle de génération, validation , correction , réentraînement fonctionne, et que le système apprend réellement de ses erreurs.

5.Notre interface

Notre interface est développée avec Flask. Elle offre la possibilité de saisir des topics concernant la culture et le patrimoine algérien en français ou en arabe, de choisir un modèle de génération, et de visualiser les articles générés, tout en permettant de les évaluer et de les valider, simplifiant ainsi le processus.

Tableau 15 Tableau des Interfaces et Fonctionnalités des Figures

Figure	Description
Figure 31	Représente l'interface principale où on peut choisir l'une des actions suivantes : générer un article, voir la liste des articles générés, scraper le web, consulter tous les articles scrappés, ou afficher les articles à corriger.
Figure 32	Fenêtre de web scraping permettant de saisir un lien et un titre, puis d'obtenir automatiquement l'article scrappé.
Figure 33	Fenêtre pour générer un article
Figure 34	Fenêtre pour évaluer un article
Figure 35	Liste des articles en attente de correction, triés par ordre de priorité décroissante ou les articles en haut de la liste doivent être corrigés en premier (selon le score d'incertitude)
Figure 36	Fenêtre pour voir, corriger, d'ajout d'avis sur un article ou de l'ajout au dataset

Figure 31 : Main application

Figure 32 :Interface pour le web scraping

Figure 33 :interface pour générer un article

Figure 34 :interface pour évaluer un article

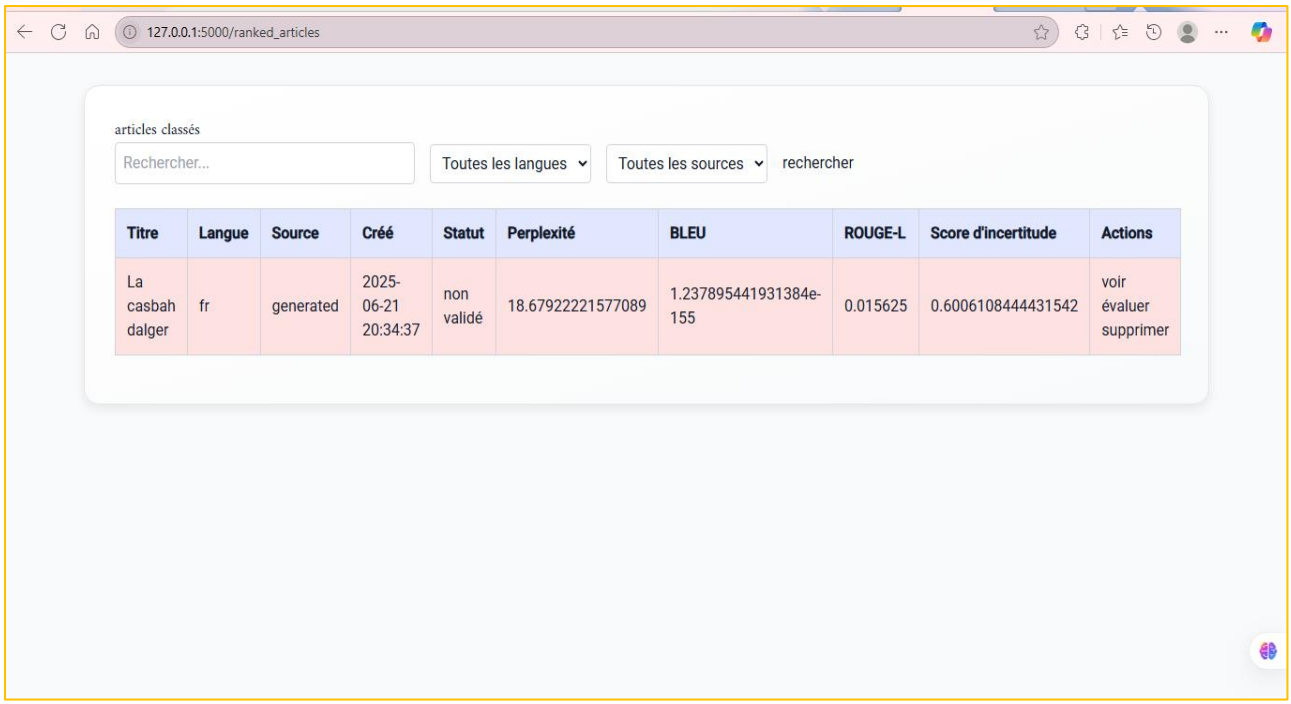

Figure 35 : Liste des articles à corriger (priorité décroissante)

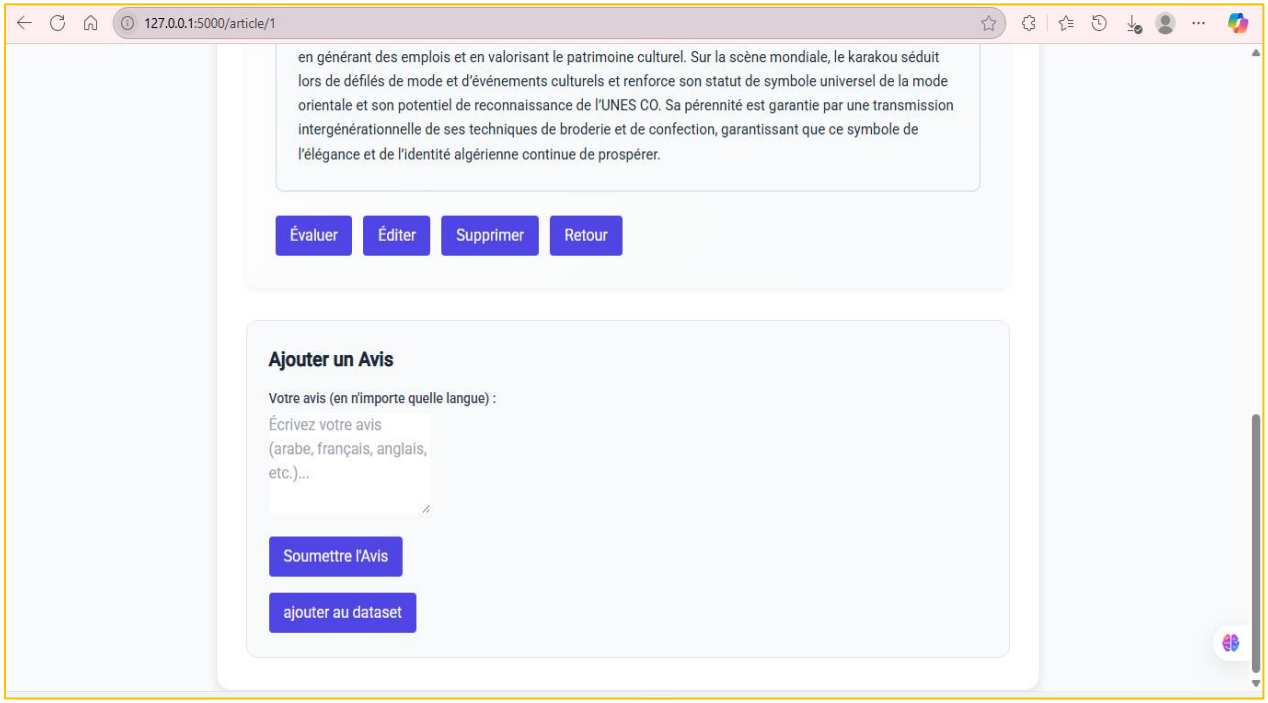

Figure 36 : Interface pour voir, corriger, d'ajout d'avis sur un article ou de l'ajout au dataset

7.Conclusion

Dans le dernier chapitre de notre travail nous avons présenté les outils que nous avons utilisés pour mener à bien notre projet. Puis, nous avons évalué le travail que nous avons réalisé tout en discutant des résultats des expériences que nous avons menées.

CONCLUSION GÉNÉRALE

1. Contributions

Ce mémoire contribue à la mise en valeur du patrimoine culturel algérien à travers les technologies du traitement automatique du langage en concevant un système capable de générer et d'évaluer automatiquement des articles encyclopédiques en arabe et en français, nous avons tenté d'apporter une réponse concrète à un besoin de plus en plus pressant : rendre visible et accessible un savoir souvent dispersé ou méconnu.

Dans ce travail nous avons découvert à quel point les modèles de langage peuvent être puissants… mais aussi qu'elles sont limitées lorsqu'ils sont confrontés à des thématiques riches de sens et de nuances comme le patrimoine.

Nous avons trouvé aussi qu'avec l'intégration de sources variées, la mise en place d'un système d'évaluation basé sur l'incertitude, ainsi que la conception d'une interface simple et fonctionnelle, ont constitué des étapes clés de notre démarche. Ce chemin n'a pas été sans difficultés : qualité inégale des données, gestion technique du bilinguisme, et surtout, la complexité de juger de la pertinence culturelle d'un contenu généré par une machine. Ces défis nous ont poussés à ajuster, repenser, parfois recommencer mais toujours avec le souci de rester fidèles à notre objectif initial , ce qui met ce projet ne constitue pas une fin, mais une base et un pipeline complet prêt à être utilisé, il ouvre sur plusieurs perspectives intéressantes : enrichir notre base de données avec des archives régionales ou des témoignages oraux, améliorer les modèles avec des ajustements plus ciblés, ou encore renforcer la collaboration avec des experts du domaine pour garantir une validation culturelle plus fine.

Donc ce travail nous avons montré qu'il est possible de faire dialoguer intelligence artificielle et mémoire collective, et si la technologie ne remplace pas la richesse humaine, elle peut devenir un levier précieux pour préserver, transmettre, et valoriser ce patrimoine et cette culture.

2.Perspectives

Notre projet, dans sa forme actuelle, représente une première base solide. Mais il ne demande qu'à grandir. Nous avons conscience de ses limites, mais aussi de son potentiel.

-Nous envions d'ouvrir le système à d'autres langues. L'Algérie est un pays multilingue, et pour que notre outil reflète vraiment à la diversité de notre pays, -il est essentiel d'intégrer le tamazight (kabyle), le darija. Ou encore des langues universelles tels que l'anglais.

-Nous souhaitons également de faire un affinage pour les méthodes de nettoyage, en explorant des techniques comme la reconnaissance d'entités nommées (NER), et en poussant des algorithmes pour le web scraping pour enrichir encore notre base de données.

-Nous avons envie de faire évoluer le modèle avec des approches plus récentes. L'idée d'utiliser la génération augmentée par récupération (RAG) : elle permettrait d'aller chercher en temps réel des informations vérifiées, pour produire des textes à la fois précis et documentés. Ce serait un grand pas vers un système plus intelligent, connecté au savoir réel.

-Nous voulons aussi de renforcer la collaboration humaine, leur en donner plus d'outils pour corriger, commenter, et enrichir le contenu généré.

-Nous avons l'ambition d'améliorer la structure des articles eux-mêmes, Aujourd'hui, ils sont organisés avec une introduction, corps et conclusion, mais -nous voulons les rendre encore plus lisibles, plus agréables à consulter, avec des sous-parties claires, des titres pertinents, et aussi avec des éléments visuels ou sonores à l'avenir.

Enfin, notre ambition est de sortir ce projet des murs académiques. Travailler avec des enseignants, des musées, des associations culturelles, des créateurs de contenu. Faire en sorte que ce système devienne un outil vivant, utilisé dans les écoles, les visites culturelles, ou même dans les médias. Un outil qui évolue, qui s'adapte, et qui garde toujours un pied dans la réalité.

RÉFÉRENCES BIBLIOGRAPHIQUES

Références bibliographiques

[1] El-Hadj, S. (2021). *Le patrimoine culturel et immatériel*. Mémoire de Master, Université Abdelhamid Ibn Badis, Mostaganem.

[2] Weizenbaum, J. (1966). A computer program for the study of natural language communication between man and machine. Dans J. Weizenbaum, A computer program for the study of natural language communication between man and machine.

[3] Daniel Jurafsky, J. H. (2023). Speech and Language Processing. Dans Speech and Language Processing. Récupéré sur https://web.stanford.edu/~jurafsky/slp3/

[4] Brown, P. F., et al. (1992). Class-Based n-gram Models of Natural Language. Computational Linguistics, 18(4), 467–479.

[5] Mikolov, T. (2010). Recurrent neural network based language model. Interspeech

[6] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–190

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762. https://arxiv.org/abs/1706.03762

[8] Zittouni, M. (2024). exploration et revue de littérature sur les LLM.

[9] NOUREEN, F., & ALI SHARIQ IMRAN, (. I. (2022). A Systematic Literature Review on Text generation Using Deep Neural Network Models.

[10] Transformer (deep learning architecture). (2025, juin 16). Transformer (deep learning architecture). In Wikipedia. https://en.wikipedia.org/wiki/Transformer_(deep_learning_architecture)

[11] Académie EITCA. (2023, 5 août). Qu'est- ce que la tokenisation dans le contexte du traitement du langage naturel ?. EITCA. https://fr.eitca.org/artificial-intelligence/eitc-ai-tff-tensorflow- fundamentals/natural-language-processing-with-tensorflow/tokenization/examination-review-tokenization/what-is-tokenization-in-the-context-of-natural-language-processing/

[12] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. https://arxiv.org/abs/1810.04805

[13] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[14] Raffel, C. a. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1--67.

[15] Mike Lewis, Y. L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.

[16] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361. https://arxiv.org/abs/2001.08361

[17] Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146. https://arxiv.org/abs/1801.06146

[18] Bengio, Y., Courville, A., & Vincent, P. (2012). Practical recommendations for gradient-based training of deep architectures. arXiv preprint arXiv:1206.5533. https://arxiv.org/abs/1206.5533 [42] Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(1), 281–305. https://www.jmlr.org/papers/v13/bergstra12a.html

[19] Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Vries, H., Gesmundo, A., ... & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. arXiv preprint arXiv:1902.00751. https://arxiv.org/abs/1902.00751

[20] Lester, B., Al-Rfou, R., & Constant, N. (2021). The Power of Scale for Parameter-Efficient Prompt Tuning. arXiv:2104.08691. https://arxiv.org/abs/2104.08691

[21] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

[22] Khder, M. A. (3 November 2021). Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. doi:10.15849/IJASCA.211128.11

[23] PERSSON, E. (December 15, 2019). Evaluating tools and techniques for web scraping.

[24] Wu, T., et al. (2021). Collaborative human-in-the-loop text generation. arXiv:2110.01904. https://arxiv.org/abs/2110.01904

[25] Malin, C. (15 JUL 2024). Algerian AI researchers crowdsource local language data.

[26] Shorten, C. K. (2021). Text data augmentation for deep learning. Journal of Big Data, 13-14.

[27] BOURDOIS., L. (2020, May 20). L'AUGMENTATION DE DONNEES EN NLP. Récupéré sur

https://lbourdois.github.io/blog/nlp/Data-augmentation-in-NLP/"

[28] Ysser Regragui, L. A. (s.d.). Arabic WordNet: New Content and New Applications.

[29] Kishore Papineni, S. R.-J. (s.d.). BLEU: a Method for Automatic Evaluation of Machine Translation.

[30] Lin, C. Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Text summarization branches out: Proceedings of the ACL-04 workshop (pp. 74–81).

[31] Callison-Burch, C., et al. (2006). Re-evaluating the Evaluation of Machine Translation Systems. Workshop on Machine Translation, 249–256.

[32] Reiter, E. (2018). A Structured Review of the Validation and Evaluation of Natural Language Generation Systems. Natural Language Generation Journal, 23(1), 6–21.

[33] Sai, R. G., et al. (2021). Towards Better Evaluation of Text Generation: From Metrics to Human Judgments. ACM Transactions on Speech and Language Processing, 29(1), 145–160.

[34] Bojar, O., et al. (2016). Findings of the 2016 Conference on Machine Translation (WMT16). Workshop on Machine Translation, 131–198.

[35] Novikova, J., et al. (2017). Sentiment Analysis: An Evaluation of Text Generation Methods. Workshop on Neural Generation and Translation, 104–111.

[36] Zhang, J., Xu, R., et al. (2024). ArchGPT: Harnessing large language models for renovation and conservation of traditional architectural heritage. *Heritage Science*, 12(220).

[37] Abed Alhakim Freihat, H. K. (April 2024). Advancing the Arabic WordNet: Elevating Content Quality.

[38] Antoun, W., Baly, F., & Hajj, H. (2020). AraGPT2: Pre-trained transformer for Arabic language generation. *Proceedings of the Third Workshop on Arabic Natural Language Processing (WANLP)*. https://aclanthology.org/2020.wanlp-1.25

[39] Hendrycks, D. &. (2020). Une base pour détecter les exemples mal classés et hors distribution dans les réseaux neuronaux.https://heritagesciencejournal.springeropen.com/articles/10.1186/s40494-024-01024-3

[40] Abodayeh, A. a. (2023). Web Scraping for Data Analytics: A BeautifulSoup Implementation. Dans 2023 Sixth International Conference of Women in Data Science at Prince Sultan University (WiDS PSU (pp. 65-69).

[41] *Amazon Web Services. (n.d.). Qu'est-ce que le grand modèle de langage (LLM). AWS. https://aws.amazon.com/fr/what-is/large-language-model/*

[42] Hassani, H. (2021). SacreBLEU : Implémentation du score BLEU pour l'évaluation de textes. (G. Repository, Éd.) Récupéré sur https://github.com/mjpost/sacrebleu

[43] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311–318).