UNIVERSITÉ DU QUÉBEC À MONTRÉAL

SYSTÈME DE QUESTION-RÉPONSE VISUELLE PAR CLASSIFICATION BASÉ SUR LES LLMS ET LES TRANSFORMATEURS D'IMAGES POUR LE HAOUSSA

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

ALI MIJIYAWA

MARS 2026

REMERCIEMENTS

Je souhaite tout d'abord remercier chaleureusement ma directrice de recherche, Professeure Fatiha Sadat, pour la confiance qu'elle m'a témoignée et pour m'avoir guidé dans le choix de mon sujet de recherche. Sa disponibilité et ses conseils avisés m'ont permis d'affiner mes idées et de structurer ma réflexion tout en développant mes compétences. Son regard critique et sa manière méthodique de résoudre les problèmes m'ont poussé à repousser mes limites. Grâce à son accompagnement bienveillant, j'ai pu explorer de nouvelles pistes et acquérir une compréhension solide des notions abordées dans ce travail. Je suis persuadé que cet enrichissement constituera un atout précieux dans la suite de mon parcours professionnel.

Je tiens également à adresser ma reconnaissance à :

- L'ensemble des membres du jury, pour le temps consacré à l'évaluation de ce mémoire, ainsi que pour la pertinence de leurs commentaires et recommandations ;
- L'ensemble du corps enseignant et administratif de l'UQAM, dont l'encadrement pédagogique de qualité et la disponibilité constante ont grandement facilité mon cheminement académique.

Enfin, je remercie sincèrement toutes les personnes qui, directement ou indirectement, ont apporté leur aide, leur soutien ou leurs encouragements tout au long de cette aventure.

DÉDICACE

Je dédie ce mémoire à mes parents, piliers de ma vie, dont l'amour, les sacrifices et les encouragements m'ont toujours poussé à me dépasser.

À ma tante Mintoumba Yendoumban, pour son aide et son soutien inconditionnel.

À l'ensemble de ma famille et à mes proches, pour leur présence bienveillante et leur appui indéfectible.

Que ce mémoire soit l'aboutissement de vos espoirs et le reflet de votre foi en moi.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

ACRONYMES

AA apprentissage automatique (machine learning).

AAM affinage adaptatif multilingue (multilingual adaptive fine-tuning).

ADS augmentation de données simplifiée (easy data augmentation).

ADSE augmentation de données en ligne supervisée (supervised data augmentation).

AI artificial intelligence (intelligence artificielle).

ARRH apprentissage par renforcement à partir de retours humains (reinforcement learning from human feedback).

ASR automatic speech recognition (reconnaissance automatique de la parole).

BERT bidirectional encoder representations from transformers (représentations encodeurs bidirectionnelles issues des transformateurs).

BUTD bottom-up and top-down (ascendant et descendant).

CLIP contrastive language-image pretraining (préentraînement contrastif langage-image).

CNN convolutional neural network (réseau de neurones convolutif).

CPU central processing unit (unité centrale de traitement).

CVQA culturally-diverse visual question answering (question-réponse visuelle culturellement diversifiée).

EDA easy data augmentation (augmentation de données simplifiée).

GML grand modèle de langue (large language model).

GMLM grand modèle de langue multimodal (multimodal large language model).

GPT generative pre-trained transformer (transformateur génératif préentraîné).

GPU graphics processing unit (unité de traitement graphique).

GQV génération de question visuelle (visual question generation).

GRU gated recurrent unit (unité récurrente à portes).

HaVG hausa visual genome.

HaVQA hausa visual question answering.

HaVQAaug hausa visual question answering augmenté.

HITL human-in-the-loop (humain dans la boucle).

I image caractérisant un exemple du jeu de données $HaVQA$.

IA intelligence artificielle (artificial intelligence).

IAE intelligence artificielle explicable (explainable artificial intelligence).

IHM interface homme-machine (human-machine interface).

LCS least common subsumer (plus petit subsumeur commun).

LLaMA large language model meta AI.

LLM large language model (grand modèle de langue).

LSTM long short-term memory (mémoire à court et long terme).

MAFT multilingual adaptive fine-tuning (affinage adaptatif multilingue).

mBERT modèle BERT multilingue.

MCLT mémoire à court et long terme (long short-term memory).

ML machine learning (apprentissage automatique).

MLLM multimodal large language model (grand modèle de langue multimodal).

NER named entity recognition (reconnaissance d'entités nommées).

NLP natural language processing (traitement automatique du langage naturel).

OCR optical character recognition (reconnaissance optique de caractères).

PPS plus petit subsumeur commun (least common subsumer).

Qen question en anglais caractérisant un exemple du jeu de données $HaVQA$.

Qha question en haoussa caractérisant un exemple du jeu de données $HaVQA$.

QA question answering (question-réponse).

QCM questionnaire à choix multiples (multiple choice questionnaire).

QR question-réponse (question answering).

QRV question-réponse visuelle (visual question answering).

Ren réponse en anglais caractérisant un exemple du jeu de données $HaVQA$.

Rha réponse en haoussa caractérisant un exemple du jeu de données $HaVQA$.

RAP reconnaissance automatique de la parole (automatic speech recognition).

REN reconnaissance d'entités nommées (named entity recognition).

RLHF reinforcement learning from human feedback (apprentissage par renforcement à partir de retours humains).

RNC réseau de neurones convolutif (convolutional neural network).

ROC reconnaissance optique de caractères (optical character recognition).

SDA supervised data augmentation (augmentation de données supervisée).

SV synthèse vocale (text-to-speech).

TAL Traitement automatique de langue.

TALN traitement automatique du langage naturel (natural language processing).

TAM traduction automatique multimodale (multimodal machine translation).

TAQRV traduction automatique et question-réponse visuelle (machine translation and visual question answering).

TAT traduction automatique textuelle (textual machine translation).

TGP transformateur génératif préentraîné (generative pre-trained transformer).

TI transformateur d'image (image transformer).

TTS text-to-speech (synthèse vocale).

TV transformateur de vision (vision transformer).

UCT unité centrale de traitement (central processing unit).

URP unité récurrente à portes (gated recurrent unit).

UTG unité de traitement graphique (graphics processing unit).

ViT vision transformer (transformateur de vision).

VQA visual question answering (question-réponse visuelle).

VQAv2 visual question answering version 2.

XAI explainable artificial intelligence (intelligence artificielle explicable).

NOTATION

Accuracy Proportion de prédictions correctes par rapport au nombre total d'exemples évalués.

concept$_1$: Premier concept comparé dans la taxonomie.

concept$_2$: Deuxième concept comparé dans la taxonomie.

depth$(\cdot)$: Profondeur d'un concept dans la hiérarchie de la taxonomie.

F1-score Mesure F1 : moyenne harmonique entre précision et rappel.

FN : Nombre de faux négatifs (instances positives incorrectement prédites comme négatives).

FP : Nombre de faux positifs (instances négatives incorrectement prédites comme positives).

LCS : Plus petit subsumeur commun (Least Common Subsumer) entre deux concepts dans la taxonomie.

$\mathcal{L}_{\text{cross-entropy}}$: Perte d'entropie croisée, utilisée pour la classification supervisée.

loss : Valeur moyenne de la fonction de perte sur l'ensemble de test, utilisée comme métrique d'évaluation de la qualité du modèle.

N : Nombre total d'exemples utilisés dans l'entraînement ou l'évaluation.

Precision : Proportion de prédictions positives correctes parmi toutes les prédictions positives du modèle.

Recall : Proportion de prédictions positives correctes parmi toutes les instances réellement positives.

TN : Nombre de vrais négatifs (instances négatives correctement prédites comme négatives).

TP : Nombre de vrais positifs (instances positives correctement prédites).

Wu-Palmer : Score de similarité sémantique entre concepts basé sur la profondeur et le plus petit subsumeur commun.

y_i : Vraie étiquette de la i-ème instance (valeur binaire ou catégorielle).

$\hat{y}_i$: Probabilité prédite par le modèle pour la classe correcte de la i-ème instance.

RÉSUMÉ

Ce mémoire propose un système de question-réponse visuelle basé sur la classification de données pour le haoussa, une langue africaine peu dotée en ressources. L'approche combine des grands modèles de langue et des transformateurs d'images, en affinant les modèles linguistiques sur des textes en haoussa et en les associant aux représentations visuelles pour prédire des réponses dans un vocabulaire prédéfini. Trente-six combinaisons de modèles ont été évaluées selon trois stratégies d'apprentissage sur le corpus $HaVQA$ de Parida et al. (2023), composé de 6 022 paires questions-réponses et 1 555 images : sans augmentation des données, avec augmentation en ligne des données, et avec augmentation hors ligne des données. La stratégie d'augmentation hors ligne a permis de créer un nouveau jeu de données, $HaVQAaug$, doublant la taille du corpus original. Les meilleurs résultats sont obtenus avec le modèle pré-entraîné en haoussa Gemini combiné au transformateur d'images ViT-base-patch16-224-in21k, atteignant 35,85 % de précision, 35,89 % de Wu-Palmer et 15,32 % de F1-score, soit un gain de plus de 5 % par rapport à l'état de l'art. Ces résultats démontrent l'importance d'un préentraînement linguistique spécifique et d'un enrichissement des données pour développer des systèmes performants dans des contextes multilingues à faible ressource, en particulier pour les langues africaines.

Mots-clés : question-réponse visuelle, haoussa, traitement automatique des langues naturelles, transformateurs de vision, classification multimodale, langues peu dotées en ressources

INTRODUCTION

Le traitement automatique des langues naturelles (TALN) est une pierre angulaire dans l'élaboration de technologies linguistiques adaptées à une variété de situations. Bien que les grands modèles de langue (GML) aient connu des progrès récents, notamment grâce à des modèles tels que BERT (Représentations d'encodeur bidirectionnelles issues des transformateurs) ou GPT (transformeur génératif préentraîné), un déséquilibre persiste entre les langues qui bénéficient de ressources abondantes (comme l'anglais) et les langues sous-représentées, comme le haoussa, parlé par des millions de locuteurs en Afrique de l'Ouest (Joshi et al., 2020a). La marginalisation numérique désigne l'exclusion de certaines communautés linguistiques de l'accès aux technologies numériques modernes. Cette marginalisation entraîne une fracture technologique qui limite l'accès à des outils sophistiqués, tels que la traduction automatique, les assistants vocaux et les systèmes de question-réponse visuelle (QRV) (Agrawal et al., 2016). Ces derniers combinent la vision par ordinateur et la compréhension du langage afin de répondre à des questions exprimées en langage naturel à partir d'images, mais leur utilisation reste rare dans les contextes multilingues africains. En ce sens, l'intégration de modèles de langue multilingues tels que mBERT (version multilingue du modèle BERT) ou AfriBERTa (modèle de type BERT qui couvre plusieurs langues africaines dont le haoussa) et les transformateurs d'image comme ViT (transformateur d'image) ou CLIP (Préentraînement contrastif multimodal texte–image) ouvre une perspective prometteuse pour développer des systèmes de question-réponse (QR) adaptés aux langues à faibles ressources (Conneau et al., 2020 ; Devlin et al., 2019a). Des études récentes ont souligné l'importance de telles démarches pour promouvoir l'équité linguistique et l'inclusion numérique à l'échelle mondiale (Nekoto, Marivate, Matsila, Fasubaa, Fagbohungbe, Akinola et al., 2020). Ce mémoire s'inscrit dans cette dynamique inclusive en proposant le développement et l'analyse d'un système QRV par classification pour le haoussa combinant les apports des GML et des transformateurs d'image (TI) à travers des stratégies de fusion multimodale et d'augmentation de données. La fusion multimodale consiste à combiner des informations provenant de différentes sources (texte et image) pour améliorer la compréhension et la performance du système. L'objectif est de mesurer l'impact de chaque composant (texte, image, réponse) sur la performance du système et d'examiner les défis spécifiques liés à la modélisation du haoussa.

Le système QRV proposé pour le haoussa se base sur la classification, avec des GML préentraînés ou non préentraînés sur le haoussa et des TI de pointe. L'analyse porte sur le jeu de données $HaVQA$ constitué de 6 022 paires de questions-réponses en haoussa-anglais accompagnées de 1 555 images. Trois schémas d'entraînement sont examinés pour analyser l'impact de l'accroissement des données sur le texte et les images :

1. Sans augmentation des données : paradigme de base ;
2. Augmentation en ligne des données : par transformations dynamiques pendant l'entraînement ;
3. Augmentation hors ligne des données : grâce à la génération préalable de données (réécriture de textes, modification visuelle).

Les principales contributions de ce mémoire sont les suivantes :

— Évaluation à grande échelle de modèles multimodaux : Comparaison de neuf GML et de quatre TI, ce qui donne lieu à 36 variantes de modèles dans un cadre unifié pour le QRV en haoussa.

— Stratégies d'augmentation pour les ressources limitées : Deux méthodes (en ligne et hors ligne) sont proposées, adaptées aux spécificités linguistiques et culturelles du haoussa, en s'appuyant sur des techniques existantes. Ces méthodes ont enrichi le jeu de données $HaVQA$ et amélioré les performances dans un contexte de données limitées (Parida et al., 2023).

— Un nouveau jeu de données $HaVQAaug$: Extension en double du jeu de données $HaVQA$ obtenue par augmentation hors ligne.

— Développement et évaluation de pipelines multimodaux : Conception de pipelines d'amélioration pour le texte et l'image, évalués par des métriques telles que la précision, la similarité Wu-Palmer et le F1-score.

Ce mémoire est organisé de la manière suivante : le chapitre 1 présente le contexte général, les concepts clés (GML, TI, QRV) ainsi que les enjeux liés aux langues sous-représentées, en mettant l'accent sur le haoussa ; le chapitre 2 offre une revue de la littérature portant sur l'histoire de la langue haoussa, les avancées récentes dans l'étude des langues africaines grâce aux technologies du traitement automatique du langage naturel, les approches multimodales en question-réponse visuelle, ainsi que les jeux de données disponibles. Le chapitre 3 décrit la méthodologie

employée, incluant la préparation des données, les métriques d'évaluation, les modèles choisis et les stratégies d'apprentissage. Le chapitre 4 présente les résultats expérimentaux, leur analyse et une discussion critique sur les performances, les erreurs, ainsi que les avantages et limites du système ainsi que la synthèse des défis rencontrés. Finalement, le chapitre 5 résume les contributions principales, les limites constatées et les perspectives d'avenir, en particulier pour l'adaptation des systèmes de question-réponse visuelle aux contextes linguistiques variés en Afrique.

CHAPITRE 1
CONTEXTE ET PROBLÉMATIQUE

1.1 Introduction

Ce chapitre présente le contexte général de cette recherche et expose la problématique liée au développement de systèmes de question-réponse visuelle (QRV) pour les langues peu dotées en ressources. La première section introduit les grands modèles de langue (GML) et leur rôle dans le traitement automatique du langage naturel (TALN) moderne. La deuxième section décrit les transformateurs d'image (TI) et leur application à l'analyse visuelle. La troisième section définit les systèmes QRV et présente les différentes approches existantes. Les sections suivantes dressent un état des lieux du TALN et de l'intelligence artificielle (IA), analysent les inégalités linguistiques qui affectent les langues sous-représentées, et examinent le cas spécifique du haoussa. La dernière partie explore le potentiel des GML pour l'inclusion linguistique et présente les applications multimodales en QRV. Ce chapitre se termine par un historique de la langue haoussa, afin de mieux situer les enjeux culturels et technologiques de cette recherche.

1.2 Les grands modèles de langue (GML)

Un GML est un modèle d'apprentissage profond conçu pour traiter et générer du langage naturel de manière cohérente, fluide et contextuellement appropriée. Entraînés sur d'immenses volumes de données textuelles, souvent multilingues, ces modèles disposent d'une capacité de modélisation remarquable, soutenue par un nombre considérable de paramètres, généralement plusieurs centaines de millions, voire plusieurs milliards (Bommasani et al., 2021). Cette échelle leur permet de capturer des régularités linguistiques complexes ainsi que des dépendances à long terme au sein des séquences textuelles. Reposant principalement sur l'architecture du transformateur, les GML constituent aujourd'hui la base des performances les plus avancées en TALN (Vaswani et al., 2023). Leur efficacité repose non seulement sur un préentraînement massif sur des corpus diversifiés, mais aussi sur leur capacité d'adaptation à des tâches spécifiques via des techniques comme l'apprentissage par consigne ou l'apprentissage par renforcement avec retour humain (ARRH) (Ouyang et al., 2022). Toutefois, leur complexité soulève d'importants défis en matière

d'opacité, de biais algorithmiques et de consommation énergétique. Les GML sont capables de réaliser un large éventail de tâches en TALN, parmi lesquelles :

— La génération de texte (rédaction, complétion automatique),

— La traduction automatique,

— Le résumé de documents,

— La réponse à des questions,

— La classification de textes.

Parmi les modèles emblématiques, on retrouve le GML GPT (Transformateur générative préentraîné), reconnu pour ses performances en génération de texte , ainsi que BERT (Bidirectional Encoder Representations from Transformers), qui a révolutionné la compréhension contextuelle des séquences linguistiques (Radford et al., 2019), (Devlin et al., 2019a). Ces modèles, et leurs nombreux dérivés occupent désormais une place centrale dans les applications modernes d'intelligence artificielle, en améliorant significativement la qualité et la polyvalence des systèmes de TALN dans divers domaines.

1.3 Les transformateurs d'image

Un TI est un type de réseau de neurones conçu pour analyser des images en s'inspirant d'un modèle très performant utilisé dans le TALN : le transformateur. Plutôt que de traiter l'image en une seule fois, le transformateur d'image ViT la découpe en petites régions appelées *patchs*, qu'il considère ensuite comme une séquence, à la manière des mots dans une phrase. Cette approche permet au modèle d'identifier des relations et des motifs à travers toute l'image, ce qui le rend particulièrement efficace pour des tâches comme la classification ou la reconnaissance d'objets (Dosovitskiy et al., 2021a). Contrairement aux approches classiques de vision par ordinateur reposant sur les réseaux de neurones convolutifs (RNC), les TI traitent les images en les divisant en des *patchs*, qui sont ensuite linéarisées, encodées en vecteurs, et enrichies par des informations de position. Ces vecteurs sont ensuite transmis à un transformateur standard, similaire à celui utilisé pour le texte, permettant de modéliser les relations globales entre les différentes parties de l'image. L'un des avantages majeurs des transformateurs d'images (ViT) réside dans leur capacité à capturer des dépendances à longue distance entre les régions d'une image, ce qui peut s'avérer

plus difficile pour les RNC, souvent limités à des champs réceptifs locaux. Cependant, l'efficacité des transformateurs d'image dépend fortement de la quantité et de la diversité des données d'entraînement disponibles. Des variantes et améliorations telles que DeiT (*Transformateurs d'images à efficacité accrue en données*) ont été proposées pour réduire cette dépendance, en rendant l'entraînement plus accessible à moindre coût en données et en ressources (Touvron et al., 2021a). De nos jours, les transformateurs d'images occupent une place de plus en plus importante dans les tâches de classification d'images, de détection d'objets, ainsi que dans les systèmes multimodaux combinant vision et langage.

1.4 Système de question-réponse visuelle (QRV)

1.4.1 Définition

Le système de question-réponse visuelle (QRV) est un système d' intelligence artificielle (IA) qui combine la vision par ordinateur et le traitement automatique du langage naturel. L'objectif principal est de permettre à un modèle de fournir une réponse pertinente à une question formulée en langage naturel, en s'appuyant sur l'analyse d'une image comme support contextuel (Agrawal et al., 2016). Ce système requiert une compréhension conjointe du contenu visuel (objets, relations spatiales, scènes) et du contenu textuel (intention de la question, raisonnement sémantique), ce qui en fait un système multimodal. Les systèmes QRV sont généralement constitués de trois composantes principales : un extracteur de caractéristiques visuelles (souvent un TI ou un RNC), un encodeur de texte pour la question, et un module de fusion multimodale permettant d'intégrer les deux représentations avant la prédiction de la réponse (Anderson et al., 2018 ; Lu et al., 2019). En fonction des applications, la sortie du modèle peut être une réponse libre (génération) ou une classification parmi un ensemble de réponses prédéfinies. Les systèmes QRV sont utilisés dans des domaines variés, notamment l'accessibilité numérique, l'assistance robotique, l'analyse d'images médicales ou encore l'éducation assistée par l'IA.

1.4.2 Différentes types de systèmes question-réponse visuelle

Les architectures des systèmes QRV ont considérablement évolué au cours des dernières années, passant de modèles modulaires spécialisés à des architectures unifiées, préentraînées sur de grandes

quantités de données multimodales. Une architecture QRV typique repose sur trois composants fondamentaux : un encodeur visuel pour extraire les caractéristiques de l'image, un encodeur textuel pour modéliser la question, et un module de fusion multimodale qui combine les deux représentations pour produire une réponse (Agrawal et al., 2016 ; Anderson et al., 2018). Dans les premières approches, les encodeurs visuels reposaient sur des réseaux de neurones convolutifs (RNC) tels que ResNet, tandis que les questions étaient encodées par des modèles séquentiels comme les LSTM ou GRU. La fusion multimodale s'effectuait par concaténation, fusion bilinéaire ou attention croisée (Fukui et al., 2016). Avec l'essor des transformateurs, de nouvelles architectures ont émergé : BERT ou RoBERTa pour le texte, ViT pour les images, et des modèles à co-attention comme ViLBERT ou LXMERT , qui favorisent une interaction croisée entre les modalités dès les couches intermédiaires. Plus récemment, des modèles unifiés tels que Flamingo ou BLIP-2 intègrent vision et langage dans une même architecture et montrent des performances accrues sur des tâches complexes, notamment en contexte multilingue ou à faibles ressources. Au-delà des aspects architecturaux, les systèmes QRV se distinguent également par la nature de la sortie attendue (Lu et al., 2019),(Tan & Bansal, 2019) , (Alayrac et al., 2022), (J. Li et al., 2023). On peut regrouper les approches en trois grandes catégories :

— Système QRV par classification : la réponse est choisie parmi un ensemble fixe de réponses possibles, défini à l'avance (vocabulaire fermé). Ce type d'approche reformule la tâche comme un problème de classification multiclasse (Agrawal et al., 2016 ; Tan & Bansal, 2019). Elle est particulièrement adaptée aux contextes à faibles ressources, car elle permet une modélisation plus stable, une évaluation plus directe (précision, , etc.) et une réduction du risque de réponses incohérentes. En revanche, elle limite la capacité du système à générer des réponses ouvertes ou nuancées. Le présent travail s'inscrit dans ce paradigme, en développant un système QRV basé sur la classification pour la langue haoussa.

— Système QRV de type QCM (questions à choix multiples) : une variante spécialisée du système par classification, où les réponses candidates sont explicitement fournies avec la question. Le modèle doit sélectionner la réponse correcte parmi les options proposées (généralement 2 à 5 choix), en évaluant la cohérence entre l'image, la question et chaque option de réponse. Cette approche est particulièrement pertinente pour les évaluations standardi-

sées et les benchmarks de raisonnement visuel, car elle permet une évaluation objective et facilite la comparaison entre différents modèles.

— Système QRV par génération : la réponse est générée mot à mot, en langage naturel, grâce à un décodeur conditionné sur les représentations textuelles et visuelles. Cette approche, plus expressive, permet de traiter des questions ouvertes, mais nécessite davantage de données et des métriques d'évaluation adaptées (BLEU, CIDEr, etc.) (Alayrac et al., 2022 ; J. Li et al., 2023).

— Approche hybride : certaines architectures combinent les deux paradigmes, en générant des réponses au sein d'un espace restreint ou en adaptant le type de sortie à la nature de la question. Ces approches visent à tirer parti de la robustesse de la classification et de la flexibilité de la génération.

Ces évolutions témoignent d'une convergence vers des systèmes QRV plus généralistes, adaptables à une grande variété de tâches visio-linguistiques, tout en tenant compte des contraintes spécifiques aux langues sous-représentées et aux données limitées.

1.5 État des lieux du traitement automatique du langage naturel (TALN) et de l'intelligence artificielle (IA)

Le développement du TALN s'inscrit aujourd'hui comme un pilier fondamental de l' IA moderne, notamment à travers l'émergence des GML. En combinant des avancées majeures en apprentissage profond, en représentation du langage et en architectures de transformateurs, le TALN a permis de franchir des seuils de performance inédits dans de nombreuses tâches linguistiques : génération de texte, traduction automatique, résumé, classification et question-réponse (Otter et al., 2021; Young et al., 2018). L'intégration des GML dans les systèmes d'interfaces homme-machine (IHM) transforme profondément la manière dont les utilisateurs interagissent avec les technologies numériques. Ces modèles permettent une interaction plus naturelle, contextuelle et fluide, en réduisant la complexité des commandes formelles au profit d'un langage libre. Les agents conversationnels (chatbots), assistants vocaux intelligents, interfaces de recherche sémantique ou systèmes de recommandation illustrent cette révolution (Bommasani et al., 2021). Au-delà de l'amélioration de l'ergonomie et de l'accessibilité, les GML favorisent également l'adaptabilité des

systèmes aux intentions de l'utilisateur, y compris dans des contextes multilingues ou peu standardisés. Toutefois, ces avancées soulèvent aussi des défis majeurs en matière d'interprétabilité, de robustesse, de biais sociolinguistiques et de respect de la vie privée. L'état actuel du TALN et des GML appelle donc à une réflexion éthique et technique sur le développement de nouvelles interfaces centrées sur l'humain, fiable et inclusif.

1.6 Inégalités linguistiques et défis des langues sous-représentées

Malgré les avancées remarquables du TALN et des GML, une fracture persistante subsiste entre les langues dominantes, notamment l'anglais, le mandarin ou l'espagnol et les langues dites sous-représentées ou à faibles ressources. La majorité des progrès technologiques repose sur des données massives, standardisées et largement disponibles, ce qui crée une inégalité structurelle au détriment des langues locales, souvent peu documentées, faiblement numérisées, ou absentes des grands corpus d'entraînement (Joshi et al., 2020a ; Nekoto, Marivate, Matsila, Fasubaa, Fagbohungbe, Akinola et al., 2020). Cette situation entraîne des biais linguistiques majeurs dans les systèmes d'IA, qui se traduisent par une moindre qualité de traduction, de compréhension ou de génération pour les langues africaines, autochtones ou régionales. Elle limite également l'accès équitable à des outils technologiques performants, renforçant ainsi les inégalités numériques et linguistiques à l'échelle mondiale (Blasi et al., 2022 ; Joshi et al., 2020b). Les défis sont multiples : absence de ressources annotées, manque de standardisation orthographique, diversité dialectale et faible soutien institutionnel pour la documentation linguistique (D. Adelani et al., 2022 ; Nekoto, Marivate, Matsila, Fasubaa, Fagbohungbe, Kolawole et al., 2020). Des initiatives récentes visent toutefois à combler cette fracture, notamment par la création de corpus multilingues ouverts, l'implication des communautés locales dans la collecte de données, ou l'adaptation des modèles multilingues existants à des contextes spécifiques (D. I. Adelani et al., 2021 ; Jude Ogundepo et al., 2022). Ces efforts soulignent l'importance d'un TALN éthique, inclusif et durable, capable de valoriser la richesse linguistique mondiale tout en répondant aux besoins technologiques des communautés marginalisées (Jude Ogundepo et al., 2022).

1.7 Le cas du haoussa : ressources et enjeux

La langue haoussa, parlée par plus de 60 millions de locuteurs principalement en Afrique de l'Ouest (notamment au Nigeria, au Niger et dans certaines régions du Cameroun, du Ghana et du Tchad), occupe une place linguistique importante, mais reste largement sous-représentée dans les technologies du TALN. Ce paradoxe entre vitalité sociolinguistique et invisibilité numérique illustre de manière frappante les inégalités structurelles qui affectent de nombreuses langues africaines dans le domaine de l'IA (Jude Ogundepo et al., 2022 ; Nekoto, Marivate, Matsila, Fasubaa, Fagbohungbe, Akinola et al., 2020). Les ressources existantes pour le haoussa demeurent très limitées, tant en quantité qu'en diversité : absence de corpus annotés de grande échelle, pénurie de lexiques standardisés, rareté des jeux de données parallèles pour la traduction automatique, et faiblesse des outils de segmentation, d'étiquetage morphosyntaxique ou d'analyse syntaxique (Muhammad et al., 2025). En outre, la coexistence de plusieurs systèmes d'écriture (alphabet latin et ajami) ainsi que la forte variation dialectale posent des défis supplémentaires pour la standardisation linguistique et l'entraînement de modèles robustes. Quelques initiatives récentes cherchent à combler ce déficit. Le projet MasakhaNER a permis la constitution d'un jeu de données d'annotation d'entités nommées pour le haoussa, tandis que des efforts comme AfriTeVa visent à produire des corpus variés pour le préentraînement et l'adaptation de GML multilingues aux langues africaines (D. I. Adelani et al., 2021), (Jude Ogundepo et al., 2022). Néanmoins, ces efforts restent encore insuffisants au regard des besoins, ce qui justifie la poursuite de travaux ciblés pour renforcer l'inclusion du haoussa dans l'écosystème numérique global.

1.8 Potentiel des grands modèles de langues (GML) pour l'inclusion linguistique

Les GML, tels que BERT, GPT, mBERT ou XLM-R, représentent une opportunité majeure pour atténuer les déséquilibres linguistiques dans le domaine du TALN. Grâce à leur capacité à être préentraînés sur de vastes corpus multilingues, ces modèles peuvent apprendre des représentations sémantiques partagées entre les langues, y compris celles qui disposent de peu de ressources annotées (Conneau et al., 2020 ; Devlin et al., 2019a). Par transfert interlangue, ils permettent d'améliorer la performance sur des langues sous-représentées comme le haoussa, en exploitant des similarités structurelles ou typologiques avec d'autres langues mieux dotées. Des modèles

comme mBERT (Multilingual BERT) ou XLM-RoBERTa ont été spécifiquement conçus pour couvrir un grand nombre de langues simultanément, ce qui ouvre la voie à une meilleure généralisation dans des contextes multilingues ou à faibles ressources (Conneau et al., 2020). Par ailleurs, les modèles autorégressifs comme GPT, bien qu'originellement entraînés sur des données principalement anglophones, ont été adaptés à des contextes multilingues, ce qui permet leur réutilisation dans des tâches en haoussa, notamment lorsqu'ils sont combinés à des transformateurs d'image dans des architectures multimodales (Workshop et al., 2022). L'un des principaux avantages de ces GML est leur potentiel de réutilisation à faible coût : il est possible d'adapter un modèle généraliste à une langue spécifique sans nécessiter d'énormes volumes de données, en recourant à des méthodes comme l'apprentissage par transfert, l'adaptation continue ou les techniques de *prompting*. Néanmoins, pour que ce potentiel se traduise en inclusion réelle, il demeure essentiel de renforcer les ressources linguistiques disponibles et d'impliquer les communautés locales dans les processus de collecte et de validation des données (D. I. Adelani, Alabi et al., 2022).

1.9 Applications multimodales et système question-réponse visuelle (QRV)

Si les avancées récentes bénéficient largement aux langues dominantes, telles que l'anglais, le chinois ou l'espagnol, les langues locales et sous-représentées demeurent souvent en marge des innovations. Le haoussa, par exemple, l'une des langues les plus parlées en Afrique de l'Ouest, est confronté à un déficit notable de ressources linguistiques et d'outils d'analyse adaptés (Joshi et al., 2020a). L' émergence des GML tels que BERT, mBERT ou GPT offre de nouvelles perspectives pour l'inclusion de ces langues dans des applications d'IA avancées. Ces modèles, grâce à leur capacité à apprendre à partir de vastes corpus multilingues, ouvrent la voie à des systèmes plus inclusifs et performants (Brown et al., 2020 ; Conneau et al., 2020 ; Devlin et al., 2019a). Dans ce cadre, les systèmes QRV, qui associent l'analyse d'images à la compréhension du langage naturel pour répondre à des questions, représentent une avancée révolutionnaire (Agrawal et al., 2016). Cependant, la majorité de ces systèmes demeure optimisée pour des langues largement répandues, créant ainsi une barrière pour les locuteurs de langues moins représentées (Nekoto, Marivate, Matsila, Fasubaa, Fagbohungbe, Akinola et al., 2020). Face à ce déséquilibre technologique, il apparaît essentiel de développer des solutions adaptées aux spécificités linguistiques

et culturelles des langues locales, telles que le haoussa. L'intégration des GML dans ces systèmes pourrait contribuer de manière significative à réduire les inégalités d'accès à l'information et à promouvoir une utilisation plus équitable des technologies avancées.

1.10 Historique de la langue haoussa

Le haoussa est généralement considéré comme l'une des langues les plus parlées en Afrique, se classant souvent comme la troisième langue du continent en termes de nombre de locuteurs natifs, après l'arabe et le swahili (Campbell, 2008). C'est l'une des principales langues d'Afrique de l'Ouest, parlée majoritairement au Nigeria, au Niger et dans plusieurs autres pays de la région. Appartenant au groupe tchadique, le haoussa possède une longue histoire marquée par une tradition orale riche, progressivement transcrite à l'écrit. À l'origine, la langue était notée à l'aide de l'alphabet arabe sous la forme de l'ajami, avant d'être adaptée, sous l'influence coloniale, à l'alphabet latin (Kaye, 2002). Historiquement, le haoussa a joué un rôle crucial comme *lingua franca* dans le commerce transsaharien et dans la diffusion de l'Islam, contribuant ainsi à son expansion et à sa pérennité (Furniss, 1996). Sa tradition orale, riche en poésie, contes et musique, constitue un patrimoine culturel d'une grande valeur, toujours présent dans la vie quotidienne et l'identité des locuteurs (Furniss, 1995). Dans le contexte du TALN et de l'inclusion numérique, la valorisation du haoussa dépasse la seule préservation culturelle. Elle répond à un besoin urgent de développer des systèmes adaptés aux langues sous-représentées, tels que les systèmes de QRV. En effet, l'implémentation de solutions en TALN pour le haoussa permet non seulement de rendre le contenu numérique accessible à un public diversifié, mais aussi de renforcer l'impact des technologies de l'information dans la région (Jude Ogundepo et al., 2022).

1.11 Conclusion

Ce premier chapitre a permis d'établir les fondations théoriques et contextuelles nécessaires à la compréhension du présent travail. Nous avons introduit les concepts clés que sont les GML, les TI ainsi que les systèmes QRV, en soulignant leur rôle central dans l'évolution des technologies de l'IA. Nous avons ensuite mis en lumière les inégalités linguistiques qui persistent dans le domaine du TALN, notamment au détriment des langues africaines, telles que le haoussa. Malgré

sa large diffusion en Afrique de l'Ouest, cette langue souffre d'un manque criant de ressources numériques, freinant son intégration dans les systèmes modernes de TALN. Dans ce contexte, les GML apparaissent comme une opportunité stratégique pour favoriser l'inclusion linguistique, grâce à leurs capacités de transfert interlangue et leur adaptabilité à des contextes multilingues à faibles ressources. Toutefois, cette promesse ne peut se concrétiser qu'en renforçant les efforts de collecte de données, d'adaptation linguistique et d'implication communautaire. Ces constats justifient pleinement le développement d'un système QRV dédié à la langue haoussa, tel que proposé dans ce mémoire. Le chapitre suivant propose une revue de la littérature scientifique sur les approches existantes en QRV et les ressources linguistiques disponibles pour le haoussa.

CHAPITRE 2
REVUE DE LITTÉRATURE

2.1 Introduction

Le présent chapitre propose une revue critique de la littérature pertinente à la problématique du système question-réponse visuelle (QRV) en haoussa, une langue africaine à faibles ressources. Il vise à situer la recherche dans le contexte plus large des avancées récentes en traitement automatique du langage naturel (TALN), en intelligence artificielle (IA) multimodale et en modélisation des langues sous-représentées. Trois axes principaux structurent cette synthèse à savoir :
- les défis et progrès du TALN pour les langues africaines, en mettant en évidence les initiatives qui ont permis de développer des corpus, des modèles et des benchmarks adaptés à ces contextes linguistiques ;
- l'évolution des systèmes QRV, depuis les approches pionnières en langues à ressources élevées jusqu'aux premiers efforts d'extension vers les langues à faibles ressources, avec un focus particulier sur le jeu de données $HaVQA$ et ses implications pour le haoussa ;
- le rôle croissant des grands modèles de langue (GML) multilingues et africains, ainsi que les stratégies d'augmentation de données (textuelles, visuelles et multimodales) qui visent à compenser la rareté des ressources annotées. Une section dédiée à la langue haoussa précise son statut sociolinguistique, sa vitalité et sa position actuelle dans les travaux en TALN et en multimodalité. Enfin, la tâche de QRV en haoussa est explicitement définie, accompagnée d'une présentation des baselines établies sur $HaVQA$.

Cette revue permet d'identifier les lacunes persistantes, en particulier le manque de données multimodales annotées en haoussa et la sous-représentation des langues tchadiques dans les GML et constitue le fondement théorique et empirique sur lequel s'appuient les expérimentations décrites dans les chapitres suivants.

2.2 Traitement automatique du langage naturel (TALN) pour les langues sous-représentées

Le traitement automatique des langues sous-représentées, comme le haoussa, présente des défis uniques par rapport aux langues dominantes. La littérature sur le traitement automatique du

langage naturel (TALN) pour les langues africaines a mis en évidence deux points. D'abord, elle a souligné le manque de ressources linguistiques (données annotées, dictionnaires, corpus). Ensuite, elle a mis en lumière les défis liés à la diversité linguistique des langues africaines (Joshi et al., 2020a). L'un des principaux obstacles consiste en la rareté de corpus annotés pour des tâches telles que la traduction automatique, l'annotation syntaxique et la reconnaissance d'entités nommées (D. Adelani et al., 2022). Cependant, des travaux récents ont cherché à combler cette lacune en développant des approches pour la création de corpus de langues sous-représentées et en adaptant des modèles préexistants à ces contextes (Conneau et al., 2020 ; Nekoto, Marivate, Matsila, Fasubaa, Fagbohungbe, Akinola et al., 2020). Des cadres d'évaluation équitables pour les modèles linguistiques appliqués aux langues africaines ont également été proposés (Adebara et al., 2024a), ainsi que des modèles préentraînés multilingues couvrant jusqu'à 23 langues africaines (Reid et al., 2024). Une analyse récente de l'état actuel du TALN pour le haoussa révèle les défis persistants et les directions futures (Muhammad et al., 2025). La recherche sur le TALN pour les langues africaines a connu un essor significatif ces dernières années, malgré les défis liés aux ressources limitées. Voici un aperçu de l'état de l'art :

2.2.1 Traduction automatique de langue (TAL)

Des initiatives collaboratives comme le projet Masakhane ont permis de développer des systèmes de traduction adaptés à de nombreuses langues africaines, contribuant ainsi à réduire la fracture numérique linguistique (Nekoto, Marivate, Matsila, Fasubaa, Fagbohungbe, Akinola et al., 2020). Plusieurs jeux de données ont vu le jour, notamment MAFAND-MT et AFRIMTE, tandis que des modèles comme Masakhane NMT illustrent les progrès réalisés (D. I. Adelani, Alabi et al., 2022 ; J. Wang et al., 2024). Des benchmarks reproductibles pour la traduction de 8 langues africaines (Reid et al., 2021) et des systèmes de traduction multimodale massivement multilingues (Babu et al., 2024) ont également été développés.

2.2.2 Reconnaissance automatique de la parole (RAP)

Des chercheurs ont développé des corpus pour des langues comme le maninka, le susu et le pular, ainsi que pour l'amharique, le swahili et le wolof, facilitant le développement de modèles RAP

(ASR) adaptés (Doumbouya et al., 2021; Gauthier et al., 2016). Des corpus vocaux massivement multilingues incluant plusieurs langues africaines ont également été créés (Ardila et al., 2023).

2.2.3 Analyse morpho-syntaxique et reconnaissance d'entités nommées (REN)

Des outils spécifiques ont émergé pour traiter la richesse morphologique des langues africaines et pour extraire des entités pertinentes (D. I. Adelani et al., 2021). Des approches centrées sur l'Afrique pour l'apprentissage par transfert et des ressources pour l'étiquetage morpho-syntaxique de langues africaines typologiquement diverses ont également été proposées (D. I. Adelani, Neubig et al., 2022), (Dione et al., 2023).

2.2.4 Analyse de sentiments et classification de textes

Des travaux, par exemple pour le nigerian Pidgin, ont permis d'analyser les opinions exprimées sur les réseaux sociaux (Muhammad et al., 2022). Des benchmarks couvrant 14 langues africaines avec des données Twitter annotées et des approches exploitant le contenu cinématographique nigérian ont élargi la portée de l'analyse de sentiments (Muhammad et al., 2023), (Winata et al., 2023). Des ensembles de données à grande échelle pour la classification thématique couvrant plus de 200 langues et dialectes ont également été développés (D. I. Adelani, Liu et al., 2024a).

2.2.5 Identification de langue et code-switching

La détection automatique de la langue et du changement de code dans des contextes multilingues s'avère cruciale (D. I. Adelani, Alabi et al., 2022).

2.2.6 Synthèse vocale (TTS)

Des systèmes de synthèse vocale multilingues comme *AfroTTS* ont été proposés afin d'améliorer l'accessibilité numérique (Mellouk et al., 2021). Des approches pour la synthèse vocale dans les langues à faibles ressources ont également été développées (Gupta et al., 2024).

2.2.7 Systèmes multimodaux pour les langues africaines

Des travaux récents sur l'apprentissage multimodal pour les langues bantoues (Siaminwe et al., 2024), des benchmarks de réponse à des questions visuelles culturellement diversifiés (Romero et al., 2024) et des approches de légende d'images tenant compte de la diversité des données (Hacheme & Sayouti, 2021) ouvrent de nouvelles perspectives pour l'inclusivité linguistique dans les technologies vision-langage. L'émergence de modèles multimodaux de nouvelle génération comme Gemini (Team et al., 2025), Llama 3 (AI, 2024) et XGen-MM (Singh et al., 2024) offre de nouvelles opportunités pour les langues africaines. Des approches de génération de données synthétiques facilitent également le développement de systèmes multimodaux pour les langues à faibles ressources (Y. Li et al., 2024).

2.3 Évolution des systèmes de question-réponse visuelles (QRV)

2.3.1 Des approches de fusion classiques aux architectures transformateurs

La recherche en QRV a connu une évolution rapide grâce aux avancées conjointes en vision par ordinateur et en traitement du langage naturel. Les premières approches se concentraient sur la fusion des caractéristiques visuelles et textuelles pour répondre à des questions visuelles, atteignant des performances élevées sur des ensembles de données annotées en langues à ressources élevées (Agrawal et al., 2016). L'avènement des modèles de langue fondés sur les transformateurs, tels que BERT, finement ajustés pour des tâches spécifiques, a amélioré la précision des systèmes QRV (Devlin et al., 2019b). De son côté, l'utilisation des transformateurs d'images a permis d'obtenir des représentations visuelles plus riches (Dosovitskiy et al., 2021a). Ensemble, ces innovations ont donné lieu à des architectures multimodales performantes capables de relever des défis QRV de plus en plus complexes. L'émergence d'architectures vision-langage préentraînées a marqué un tournant majeur. Des modèles comme LXMERT, VisualBERT et ViLBERT ont démontré l'efficacité de l'apprentissage conjoint de représentations texte-image (L. H. Li et al., 2019 ; Lu et al., 2019 ; Tan & Bansal, 2019). Plus récemment, les modèles BLIP et BLIP-2 ont introduit des approches de bootstrapping pour l'apprentissage vision-langage, permettant une meilleure intégration des modalités visuelles et textuelles (J. Li et al., 2022, 2023). Par ailleurs, LLaVA et InstructBLIP ont démontré l'efficacité de l'ajustement par instructions pour améliorer la compréhension multimodale (Dai et al.,

2023 ; H. Liu et al., 2023). L'avènement de modèles multimodaux à grande échelle comme GPT-4V, Gemini et XGen-MM ouvre de nouvelles perspectives pour la QRV, notamment par leur capacité de raisonnement visuel avancé et leur apprentissage few-shot (OpenAI, 2023 ; Singh et al., 2024 ; Team et al., 2025).

2.3.2 Les systèmes de question-réponse visuelles (QRV) pour les langues à faibles ressources et les langues africaines

Malgré ces avancées, la recherche en QRV se concentre encore largement sur des langues à fortes ressources comme l'anglais, laissant de côté les langues à faibles ressources (G. K. Kumar et al., 2022). Ce déséquilibre est particulièrement marqué pour les langues africaines comme le haoussa, qui manquent de corpus annotés et de ressources linguistiques. Pour y remédier, le corpus multimodal $HaVQA$ a été introduit comme première base essentielle pour les systèmes QRV en contexte de faible ressource (Parida et al., 2023). Plus récemment, le benchmark CVQA, un corpus QRV multilingue culturellement diversifié couvrant 10 000 paires question–réponse (QR) en 31 langues, a été proposé en expérimentant sur des grands modèles multimodaux de langage (MLLM) pour la tâche QRV par génération (Romero et al., 2024). Cependant, le haoussa n'y figure pas, soulignant encore le besoin de jeux de données spécifiquement dédiés aux langues africaines peu dotées. Des benchmarks multilingues ont également été développés pour évaluer le transfert d'apprentissage entre modalités, tâches et langues, bien que la couverture des langues africaines reste limitée (Bugliarello et al., 2022). L'émergence de modèles multimodaux multilingues à grande échelle offre néanmoins de nouvelles opportunités pour réduire cette fracture numérique (Singh et al., 2024).

2.3.3 Stratégies d'augmentation de données pour le système de question-réponse visuelle (QRV)

En parallèle du développement de corpus, l'augmentation de données est devenue une stratégie clé pour pallier le manque de ressources. Des techniques comme la rétrotraduction, la substitution de synonymes ou les perturbations visuelles enrichissent les corpus d'entraînement, tout en renforçant la robustesse et la généralisation des modèles (Z. Wang et al., 2021). Des approches récentes exploitent également l'apprentissage curriculaire pour améliorer l'augmentation de don-

nées en QRV (Zheng et al., 2024). La génération de données synthétiques assistée par des modèles de langage constitue une avenue prometteuse pour les tâches multimodales à faibles ressources (Y. Li et al., 2024).

2.4 Les grands modèles de langue (GML) en général et les grands modèles de langue (GML) pour les langues africaines

2.4.1 Évolution des grands modèles de langue multilingues

Les grands modèles de langue (GML) comme BERT, GPT-3 ou T5 ont révolutionné la compréhension et la génération du langage naturel (Brown et al., 2020 ; Devlin et al., 2019b ; Raffel et al., 2023). Leurs variantes multilingues, comme mBERT et XLM-R, prennent en charge, respectivement, plus de 100 et 50 langues. Néanmoins, malgré cette couverture étendue, ces modèles affichent souvent des performances faibles pour les langues à faibles ressources, notamment celles parlées en Afrique, en raison de leur faible représentation dans les corpus d'entraînement (Conneau et al., 2020 ; Devlin et al., 2019b). Des analyses récentes ont quantifié ce déséquilibre, montrant que toutes les langues n'ont pas le même coût de traitement dans les modèles commerciaux, ce qui pénalise davantage les langues africaines (Ahia et al., 2023).

2.4.2 Modèles de langue spécialisés pour les langues africaines

Pour combler cette lacune, des méthodes comme le *affinage adaptatif multilingue* (AAM) permettent de raffiner des modèles préentraînés sur du texte monolingue africain, obtenant ainsi des gains notables sans entraîner un modèle distinct pour chaque langue (Alabi et al., 2022a). En parallèle, des modèles entraînés depuis zéro sur des données africaines montrent des performances remarquables. Par exemple, le GML AfriBERTa, préentraîné sur moins de un gigaoctet de texte en onze langues africaines, rivalise avec des modèles multilingues plus grands sur des tâches comme la reconnaissance d'entités nommées et la classification de texte (Ogueji et al., 2021b). Le modèle de langue AfroLM poursuit dans cette voie en utilisant un cadre d'apprentissage actif pour préentraîner sur 23 langues africaines, avec des gains en performance malgré des données limitées (Dossou et al., 2022). Plus récemment, le modèle `UBC-NLP/cheetah-base` a été préentraîné sur des textes couvrant plus de 500 langues africaines, démontrant d'excellents résultats

dans des contextes à faibles ressources (Adebara et al., 2024b). Le modèle Aya, finement ajusté sur des instructions multilingues, a également élargi la couverture à 101 langues, incluant plusieurs langues africaines (Elmadany et al., 2024).

2.4.3 Benchmarks et évaluation des GML pour les langues africaines

Au-delà des architectures, des initiatives communautaires ont permis la création de jeux de données et de benchmarks essentiels. MasakhaNER propose un corpus de qualité pour la reconnaissance d'entités nommées dans dix langues africaines (D. I. Adelani, Neubig et al., 2022 ; D. I. Adelani et al., 2021). Le projet Masakhane favorise quant à lui la recherche ouverte et collaborative en TALN sur tout le continent (Nekoto, Marivate, Matsila, Fasubaa, Fagbohungbe, Akinola et al., 2020). Le benchmark AfroBench (ensemble de tests standardisés conçu pour évaluer la performance des GML sur les langues africaines) a évalué systématiquement mT0, Aya, LLaMA 2 et GPT-4 sur six tâches (classification, traduction, résumé, questions-réponses, NER, etc.) couvrant 60 langues africaines (Ojo et al., 2025). Cela a révélé des écarts persistants avec les langues à fortes ressources. Enfin, l'outil IrokoBench fournit une évaluation complète de 16 GML (10 open source et 6 propriétaires) sur 17 langues africaines (haoussa, yoruba, somali, zoulou, swahili, etc.) à travers trois tâches principales (D. I. Adelani, Ojo et al., 2024). Ces dernières incluent AfriXNLI (inférence), AfriMGSM (raisonnement mathématique) et AfriMMLU (QCM de connaissances). Il met en évidence des écarts de performance importants, où le meilleur modèle open source (Gemma 2 27B) atteint seulement 63 % de la performance de GPT_4O. Des benchmarks plus récents ont élargi cette analyse à des GML tels que mT0-base/large, AfriBERTa-large, Afro-XLMR-large-76L, Gemini, BloomZ-560m/1b7, Llama-3.2-1B ou DeepSeek-R1-1.5B (D. I. Adelani, Liu et al., 2024a ; AI, 2024 ; DeepSeek-AI et al., 2025 ; Muennighoff, Wang, Sutawika, Roberts, Biderman, Le Scao et al., 2023 ; Ogueji et al., 2021b ; Team et al., 2025). Le benchmark AfroLM a également évalué les capacités de ces modèles sur 23 langues africaines, révélant des performances prometteuses pour les modèles spécialisés (Reid et al., 2024). Plusieurs de ces modèles sont utilisés au cours des expérimentations pour analyser et réduire ces écarts de performance.

2.4.4 Augmentation de données

2.4.4.1 Augmentation de données textuelles

Des chercheurs ont proposé diverses techniques d'augmentation de données textuelles pour améliorer la généralisation et la robustesse des modèles. Les premiers travaux ont introduit la substitution de synonymes à l'aide d'un thésaurus (Zhang et al., 2016). La technique d'augmentation de données simplifiée (ADS), qui inclut des insertions, suppressions, permutations aléatoires et remplacements par synonymes, a ensuite été présentée (Wei & Zou, 2019). La rétrotraduction s'est également imposée comme une méthode populaire, de même que la génération automatique de mots (Dong et al., 2017 ; Mallinson et al., 2017 ; Sennrich et al., 2016b). Pour les besoins de l'inférence causale, l'augmentation contrefactuelle modifie un attribut spécifique d'une phrase tout en préservant le contexte global (Kaushik et al., 2020). Plus récemment, des méthodes exploitant les capacités des GMLs utilisent le masquage et la reconstruction (V. Kumar et al., 2020 ; X. Wu et al., 2018). Ces méthodes génèrent ainsi des transformations riches et cohérentes qui surpassent les simples modifications de jetons. D'autres approches émergentes tirent parti des capacités génératives des GML pour élargir et diversifier davantage les jeux de données (Ding et al., 2024). Ainsi, nous avons utilisé des techniques comme la rétrotraduction et l'ADS pour relever les défis liés aux données faibles (Nekoto, Marivate, Matsila, Fasubaa, Fagbohungbe, Akinola et al., 2020 ; Wei & Zou, 2019).

2.4.4.2 Augmentation de données visuelles

L'augmentation de données visuelles rassemble des méthodes visant à accroître artificiellement la taille et la diversité d'un jeu de données d'images. Elles cherchent ainsi à améliorer la robustesse et la capacité de généralisation des modèles de vision par ordinateur (Yang et al., 2023). Popularisées par des travaux démontrant leur efficacité dans l'entraînement de réseaux neuronaux profonds, ces techniques incluent des transformations géométriques (telles que la rotation, les renversements horizontal et vertical, le zoom), des ajustements de luminosité, des déformations élastiques, ainsi que des approches de masquage partiel (*mixup* (mélange d'images), *cutout* (découpage), *random erasing* (effacement aléatoire)) (Krizhevsky et al., 2012 ; Simard et al., 2003). Elles permettent d'augmenter la diversité des exemples tout en réduisant le surapprentissage et

en évitant des coûts d'annotation élevés (K. He et al., 2016 ; Shorten & Khoshgoftaar, 2019b). Ces stratégies sont devenues des pratiques standard dans la préparation des données pour des tâches de reconnaissance d'images, des petits corpus aux collections à grande échelle (Simonyan & Zisserman, 2015).

2.4.4.3 Augmentation de données pour les systèmes de question-réponse visuelle

Dans le contexte de la QRV, plusieurs travaux ont proposé des techniques spécifiques d'augmentation. Une combinaison de transformations d'images et de reformulations de questions pour enrichir les jeux d'entraînement a été proposée (Kafle et al., 2017). Cette méthode applique des modifications comme des miroirs horizontaux ou des occultations partielles sur les images, tout en reformulant les questions de manière paraphrasée. Cette stratégie met l'accent sur l'importance du maintien de la cohérence entre le contenu visuel modifié et la description textuelle. Les résultats montrent que des augmentations soigneusement conçues peuvent améliorer significativement les performances des modèles. Des méthodes revenant sur les approches conventionnelles d'augmentation soulignent l'importance de préserver la cohérence multimodale (D. Chen et al., 2022 ; L. Chen et al., 2022). Des transformations d'images ciblées et des ajustements textuels sur mesure, plutôt que des manipulations génériques, ont été proposés. Les résultats expérimentaux montrent que ces augmentations cohérentes conduisent à de meilleures performances et à une robustesse accrue. Des approches basées sur l'apprentissage multimodal en espace de caractéristiques ont également démontré leur efficacité pour générer des augmentations cohérentes (Z. Liu et al., 2023). Dans le cadre de la QRV en contexte de données faiblement annotées, une méthode d'augmentation générant de nouvelles paires question-réponse à partir d'annotations existantes a été introduite (Askarian et al., 2022). Elle exploite des métadonnées (attributs d'objets) et des modèles de questions pour injecter des biais inductifs ciblés dans l'entraînement. Cela permet d'améliorer significativement les performances jusqu'à +34 % par rapport aux modèles entraînés uniquement sur les données initiales. Des techniques de mixup conditionnel ont également été appliquées avec succès dans le domaine médical, démontrant l'efficacité de la combinaison d'échantillons pour la QRV spécialisée (Gong et al., 2022). La méthode KDDAug, une approche d'augmentation des données pour la question-réponse visuelle qui repose sur la distillation des

connaissances, a été introduite (L. Chen et al., 2022). Contrairement aux approches synthétiques traditionnelles, qui éditent des régions visuelles ou des mots au risque de produire des échantillons artificiels et bruités, KDDAug génère de nouvelles paires image-question et leur assigne des pseudo-réponses de manière robuste. Cette approche évite la dépendance à des règles heuristiques limitées et améliore la capacité de généralisation des modèles. Des approches récentes exploitent également l'apprentissage curriculaire pour optimiser l'augmentation de données en QRV (Zheng et al., 2024). Une méthode d'augmentation évitant les transformations classiques (rotations, miroirs) susceptibles d'altérer l'intégrité sémantique des triplets image-question-réponse a été proposée (Tang et al., 2020). À la place, des exemples adversariaux soigneusement conçus pour préserver les caractéristiques visuelles et le sens de la question sont générés. Ces données sont ensuite intégrées dans un cadre d'apprentissage adversarial pour entraîner un modèle classique (BUTD), conduisant à des gains significatifs sur le benchmark VQAv2 et à une meilleure résistance aux attaques adversariales. L'utilisation de petits modèles multimodaux augmentés pour guider des modèles de langage plus grands a également montré des résultats prometteurs (W. He et al., 2023).

2.5 À propos de la langue haoussa

2.5.1 Démographie et rayonnement géographique

Plus de 100 millions de personnes parlent le haoussa, ce qui classe cette langue parmi celles que les africains utilisent le plus (Inuwa-Dutse, 2021). Cette estimation s'harmonise bien avec d'autres évaluations qui précisent que la population de locuteurs natifs se compte par plus de 30 millions, dont une grande partie vit au Nigeria et au Niger (Jaggar, 2006a). Le haoussa figure également parmi les langues les plus parlées du continent africain, avec une présence significative dans plusieurs pays d'Afrique de l'Ouest et du Centre (Campbell, 2008).

> Le haoussa constitue la langue la plus influente et la plus répandue d'Afrique de l'Ouest, et il continue de s'affirmer comme une lingua franca transnationale. Son usage s'étend aux domaines commercial, administratif, éducatif et médiatique. On compte de nombreux journaux en haoussa, et le secteur de l'édition, de la télévision et de la production audiovisuelle se montre particulièrement dynamique. La langue est également

largement diffusée à la radio, tant en Afrique que sur la scène internationale, notamment par la BBC World Service, la Voice of America, la Deutsche Welle et la China Radio International, qui utilisent principalement le dialecte standard de Kano. Sur le plan académique, plusieurs universités au Nigeria et au Niger offrent des programmes de premier et de deuxième cycles en haoussa, tandis que des chercheurs spécialisés en langue et littérature haoussa participent à des programmes comparables dans des établissements en Europe, aux États-Unis, au Japon, en Chine et en Corée du Sud (Jaggar, 2006b).

2.5.2 Caractéristiques linguistiques et patrimoine culturel

Membre de la branche tchadique de la famille afroasiatique, le haoussa joue un rôle de langue maternelle et de lingua franca régionale en Afrique de l'Ouest et du Centre (Kaye, 2002). Il dispose d'une riche tradition littéraire, orale et écrite, comprenant une poésie et une prose anciennes et contemporaines (Furniss, 1996, 2019). La langue s'écrit à la fois en alphabet latin (boko) et en alphabet dérivé de l'arabe (ajami) (Bondarev, 2021). Cette tradition d'écriture ajami témoigne de l'influence historique et culturelle de l'islam dans la région, et constitue un patrimoine manuscrit important pour la préservation de la littérature et des savoirs locaux (Furniss, 1995).

2.5.3 Le haoussa dans le traitement automatique du langage naturel (TALN)

Le haoussa importe sur le plan linguistique et socioculturel, mais il reste largement sous-représenté dans les domaines du traitement automatique du langage naturel (TALN) et de l'intelligence artificielle multilingue. On le considère alors comme une langue à faible ressource d'un point de vue computationnel (Hedderich et al., 2021; Nekoto, Marivate, Matsila, Fasubaa, Fagbohungbe, Akinola et al., 2020). Une analyse récente et complète de l'état actuel du TALN pour le haoussa a identifié les défis persistants et les directions futures prometteuses pour cette langue (Muhammad et al., 2025). Ce manque de ressources a suscité des efforts récents pour développer des corpus dédiés et des systèmes d'IA multimodaux inclusifs favorisant la diversité linguistique en recherche. Ces dernières années, l'intérêt pour le haoussa s'est accru dans le cadre de plusieurs tâches de TALN. Parmi celles-ci, on compte la classification de textes, l'analyse de sentiments via

AfriSenti, la traduction automatique avec JW300, la reconnaissance d'entités nommées, l'étiquetage morphosyntaxique ainsi que les technologies vocales (Common Voice, CMU Wilderness et MLS) (D. I. Adelani et al., 2021; Agić & Vulić, 2019; Ardila et al., 2020; Black, 2019; Dione et al., 2023; Muhammad et al., 2023; Pratap et al., 2020). Le haoussa est également intégré dans des corpus multimodaux comme HaVQA et HaVG, ainsi que dans plusieurs modèles préentraînés tels qu'AfriBERTa, AfroXLM-R ou mT0 (Abdulmumin et al., 2022; Alabi et al., 2022a; Muennighoff, Wang, Sutawika, Roberts, Biderman, Le Scao et al., 2023; Ogueji et al., 2021b; Parida et al., 2023). Des modèles de langue plus récents ont élargi la couverture du haoussa, notamment AfroLM avec 23 langues africaines, Cheetah avec plus de 500 langues africaines, et Aya avec 101 langues (Adebara et al., 2024b; Elmadany et al., 2024; Reid et al., 2024). Des benchmarks systématiques comme IrokoBench et AfroBench ont également évalué les performances des grands modèles de langue sur le haoussa, révélant des écarts persistants mais des progrès encourageants (D. I. Adelani, Ojo et al., 2024; Ojo et al., 2025). L'avènement de modèles multimodaux à grande échelle comme Gemini et Llama 3 offre de nouvelles opportunités pour améliorer les systèmes multilingues incluant le haoussa (AI, 2024; Team et al., 2025). Ces développements contribuent ainsi à une meilleure inclusion multilingue dans les systèmes de TALN.

2.6 Système de question-réponse visuelle pour le haoussa

2.6.1 Problématique du système question-réponse (QRV) pour le haoussa

Le développement de systèmes de question-réponse visuelle pour le haoussa constitue une avancée importante pour l'intelligence artificielle inclusive, permettant d'étendre les bénéfices des technologies multimodales aux locuteurs de langues sous-représentées. Un système QRV est un système permettant de répondre à des questions en s'appuyant à la fois sur l'analyse d'une image et sur la compréhension du texte de la question (Agrawal et al., 2016). L'application de cette technologie au haoussa implique le développement de modèles capables de comprendre et de répondre à des questions posées en haoussa à propos de contenus visuels. Cette approche multimodale nécessite à la fois une compréhension du langage naturel en haoussa, une langue à faibles ressources, et un raisonnement visuel, afin de générer des réponses précises et contextuellement appropriées en intégrant l'information linguistique et visuelle.

Le défi principal réside dans le manque de ressources annotées pour le haoussa, une problématique commune aux langues africaines dans le domaine multimodal (G. K. Kumar et al., 2022). Une analyse récente de l'état du TALN pour le haoussa a souligné les défis spécifiques liés au développement de systèmes multimodaux dans ce contexte linguistique (Muhammad et al., 2025). Des travaux sur d'autres langues à faibles ressources ont démontré l'efficacité de stratégies d'apprentissage par transfert et d'augmentation de données pour pallier ces limitations (Bugliarello et al., 2022).

2.6.2 Le jeu de données HaVQA : première ressource de référence pour le QRV-haoussa

La publication du jeu de données $HaVQA$ constitue une avancée majeure pour la recherche dans ce domaine (Parida et al., 2023). Ce jeu de données contient des milliers de triplets image-question-réponse en haoussa, et il a permis d'établir des bases de référence pour l'évaluation des systèmes QRV en haoussa. Plusieurs approches ont été expérimentées sur ce jeu de données, notamment un système QRV par classification combinant quatre transformateurs d'image avec le modèle de langue BERT-base-Hausa (Parida et al., 2023). Ce dernier a été obtenu par raffinement du modèle bert-base-multilingual-cased sur la langue haoussa et compte 110 millions de paramètres (D. I. Adelani, 2023 ; Devlin et al., 2019b). Cette configuration de référence a permis d'établir des performances initiales sur le jeu de données $HaVQA$, ouvrant la voie à des améliorations ultérieures et à l'exploration de stratégies d'augmentation de données pour compenser la rareté des ressources annotées en haoussa. Le tableau 2.2 résume les résultats d'expérimentations du système QRV obtenus sur le jeu de données $HaVQA$ (Parida et al., 2023).

2.6.3 Perspectives et approches émergentes

Depuis la publication de $HaVQA$, plusieurs approches prometteuses ont émergé dans le domaine de la QRV multilingue et pour les langues à faibles ressources. Des architectures vision-langage préentraînées comme BLIP-2 ont démontré leur capacité à faciliter l'apprentissage par transfert entre langues et modalités (J. Li et al., 2023). L'ajustement par instructions, illustré par des modèles comme LLaVA et InstructBLIP, offre également des perspectives intéressantes pour améliorer la compréhension multimodale dans des contextes à faibles ressources (Dai et al., 2023 ; H. Liu et al.,

2023).

Des benchmarks multilingues culturellement diversifiés comme CVQA ont élargi la portée de l'évaluation des systèmes QRV à 31 langues, bien que le haoussa n'y soit pas encore inclus (Romero et al., 2024). Cette absence souligne le besoin continu de développer des ressources spécifiques pour le haoussa et d'autres langues africaines. L'émergence de grands modèles multimodaux comme Gemini, Llama 3 et XGen-MM, capables de raisonnement visuel avancé et d'apprentissage few-shot, ouvre de nouvelles opportunités pour améliorer les performances des systèmes QRV en haoussa, notamment par des approches d'adaptation avec peu de données (AI, 2024 ; Singh et al., 2024 ; Team et al., 2025). Des techniques récentes de génération de données synthétiques pour les tâches multimodales à faibles ressources constituent également une avenue prometteuse pour enrichir les corpus d'entraînement en haoussa (Y. Li et al., 2024). De plus, l'apprentissage multimodal en contexte africain, tel que démontré pour les langues bantoues, suggère des stratégies adaptables au haoussa (Siaminwe et al., 2024).

Table 2.1 – Nombre de paramètres du modèle BERT-base affiné pour le haoussa

Modèle	Nombre de paramètres
BERT-base-hausa	110 millions

Source : Davlan/bert-base-Hausa

Table 2.2 – Résultats du système de base de question-réponse visuelle (QRV) sur le jeu de données $HaVQA$.

Encodeur d'images	Encodeur de texte	Score de Wu-Palmer
BEiT-large-P-224	BERT-base-Hausa	27,76
ViT-base-P-224	BERT-base-Hausa	28,91
ViT-large-P-224	BERT-base-Hausa	29,67
DeiT-base-P-224	**BERT-base-Hausa**	**30,86**

Source : (Parida et al., 2023)

2.7 Conclusion

Cette revue de littérature a permis de situer la problématique de la question-réponse visuelle en haoussa dans le contexte plus large des avancées récentes en traitement automatique du langage

naturel et en intelligence artificielle multimodale. Trois constats majeurs se dégagent de cette synthèse. Premièrement, bien que le TALN pour les langues africaines ait connu un essor significatif ces dernières années, avec le développement de corpus, de modèles et de benchmarks adaptés, les langues à faibles ressources comme le haoussa demeurent largement sous-représentées dans les systèmes multimodaux. Les initiatives telles que Masakhane, AfroLM, ou encore AfriBERTa témoignent d'un engagement croissant pour réduire cette fracture numérique linguistique, mais les efforts restent concentrés principalement sur des tâches unimodales (traduction, reconnaissance d'entités nommées, analyse de sentiments). Deuxièmement, l'évolution des systèmes QRV a été marquée par des avancées technologiques remarquables, notamment grâce aux architectures fondées sur les transformateurs et à la fusion de représentations visuelles et textuelles. Toutefois, ces progrès bénéficient essentiellement aux langues à ressources élevées comme l'anglais. Les rares initiatives visant les langues africaines, telles que le jeu de données $HaVQA$ pour le haoussa ou CVQA pour d'autres langues, constituent des premiers pas essentiels mais insuffisants pour combler l'écart persistant. Troisièmement, les techniques d'augmentation de données, qu'elles soient textuelles, visuelles ou multimodales, apparaissent comme des stratégies prometteuses pour pallier la rareté des ressources annotées. La rétrotraduction, l'augmentation de données simplifiée (ADS), les transformations géométriques d'images, ainsi que les méthodes spécifiques à la QRV préservant la cohérence multimodale, offrent des pistes concrètes pour enrichir les corpus d'entraînement et améliorer la robustesse des modèles dans les contextes à faibles ressources. Le haoussa, langue tchadique parlée par plus de 100 millions de locuteurs et jouant un rôle de lingua franca en Afrique de l'Ouest, mérite une attention particulière dans le développement de technologies inclusives. Malgré sa vitalité sociolinguistique et son importance culturelle, cette langue reste marginalisée dans les systèmes d'IA multimodaux. Le jeu de données HaVQA et les résultats de base établis par Parida et al. (2023), avec un score Wu-Palmer de 30,86 obtenu par la combinaison DeiT-base-P-224 et BERT-base-Hausa, constituent un point de départ pour les travaux présentés dans ce mémoire. Les lacunes identifiées dans cette revue justifient pleinement la démarche expérimentale des chapitres suivants. En combinant augmentation de données, GML africains et transformateurs d'images performants, ce travail vise à améliorer le système QRV-haoussa et à promouvoir une intelligence artificielle plus inclusive.

CHAPITRE 3
MÉTHODOLOGIE ET EXPÉRIENCES

3.1 Introduction

Ce chapitre présente la méthodologie complète mise en œuvre pour développer et évaluer un système de question-réponse visuelle (QRV) pour le haoussa. La première section décrit la préparation du jeu de données $HaVQA$, incluant son analyse détaillée et les étapes de nettoyage et de prétraitement. La deuxième section expose les métriques d'évaluation retenues pour mesurer les performances du système : Wu-Palmer, précision, F1-score et fonction de perte. La troisième section présente les modèles utilisés, comprenant neuf grands modèles de langue et quatre transformateurs d'image, ainsi que les stratégies de fusion multimodale adoptées. La quatrième section détaille les paramètres d'entraînement et le processus de raffinement des modèles. La cinquième section décrit les techniques d'augmentation de données, distinguant les approches en ligne et hors ligne. Finalement, la dernière section présente les trois architectures expérimentales correspondant aux trois stratégies d'entraînement : sans augmentation, avec augmentation en ligne et avec augmentation hors ligne. L'ensemble de ces éléments méthodologiques constitue le cadre expérimental permettant d'évaluer l'efficacité des différentes approches pour la question-réponse visuelle en haoussa.

3.2 Préparation du jeu de données

3.2.1 Jeu de données utilisé

La première étape consiste à créer et préparer le jeu de données multimodal nécessaire à l'entraînement et à l'évaluation des modèles. Afin de réaliser les expériences de ce travail de recherche sur la langue haoussa, le jeu de données $HaVQA$ a été utilisé (Parida et al., 2023). Ce jeu de données est disponible publiquement sur la plateforme Hugging Face (HausaNLP, 2023). Ce jeu de données $HaVQA$ constitue une innovation en offrant un ensemble de données multimodales dédié aux échanges question-réponse sous forme visuelle dans cette langue à faibles ressources. Il est élaboré à partir de 6 022 paires de questions-réponses en anglais, chacune étant reliée à

l'une des 1 555 images uniques provenant de *Visual Genome*. Chaque paire de question-réponse fait l'objet d'une traduction manuelle en haoussa, ce qui permet de garantir un lien sémantique entre la réponse et les données visuelles associées. Ainsi, ce corpus de référence de 12 044 phrases parallèles anglais-haoussa est créé, aligné sur les images. La figure 3.1 présente un exemple du jeu de données $HaVQA$ caractérisé par cinq éléments :

- L'image I
- Une question Qen en anglais
- Une réponse Ren en anglais correspondant à la question Qen
- Une question Qha en haoussa qui n'est autre que la traduction de la question Qen en haoussa
- Une réponse Rha en haoussa correspondant à la question Qha et qui n'est autre que la traduction de la réponse Ren en haoussa.

Dans ce jeu de données, il faut noter que plusieurs questions et réponses en anglais et en haoussa peuvent être associées à une même image. Le tableau 3.1 montre un aperçu des 5 premières lignes du jeu de données $HaVQA$: chaque image a un identifiant unique $image\_id$, et chaque quintuplet ($image\_id$, Qen, Ren, Qha, Rha) correspond à un identifiant unique qa_id.
Des expériences de base sont proposées par les chercheurs Parida et al. (2023) afin d'orienter l'usage du jeu de données $HaVQA$: Il s'agit des expériences sur le système QRV, le système de génération de question à partir du visuelle (GQV), ainsi que le système de traduction automatique textuelle (TAT) et le système de traduction automatique multimodale (TAM). Cet ensemble marque une étape importante dans la recherche sur les langues peu dotées, en particulier le haoussa, dans un contexte multimodal (Parida et al., 2023).

Question en anglais : What is up above the bank?
Réponse en anglais : Houses. (Label : 1168)
Question en hausa : Menene a sama a kan bakin ruwan?
Réponse en hausa : Gidaje. (Label : 1168)

Figure 3.1 – Exemple d'une donnée issue du jeu de données $HaVQA$

Table 3.1 – Extrait des 5 premières lignes du jeu de données *HaVQA*

#	image_id	image	qa_id	ques_en	ans_en	ques_ha	ans_ha
0	2335177	2335177.jpg	460732	What color is the man's shirt ?	White.	Menene launin rigar mutumin ?	Fari.
1	2392303	2392303.jpg	1494576	What is attached to the Westminster palace ?	A tall tower.	Menene yake makale a jikin fadar Westminster ?	Dogon hasumiya.
2	2321185	2321185.jpg	886868	What are the horses walking on ?	Sand.	Akan me dawakan suke tafiya ?	Kasa.
3	2392923	2392923.jpg	1436743	What are the ties on top of ?	Blanket.	A kan me ɗamarorin wuyan suke ?	Bargo.
4	2365369	2365369.jpg	631160	Why is the ground white ?	It's snowing.	Meyasa ƙasan yayi fari ?	Dusar ƙanƙara ke zuba.

3.2.2 Bref analyse du jeu de données $HaVQA$

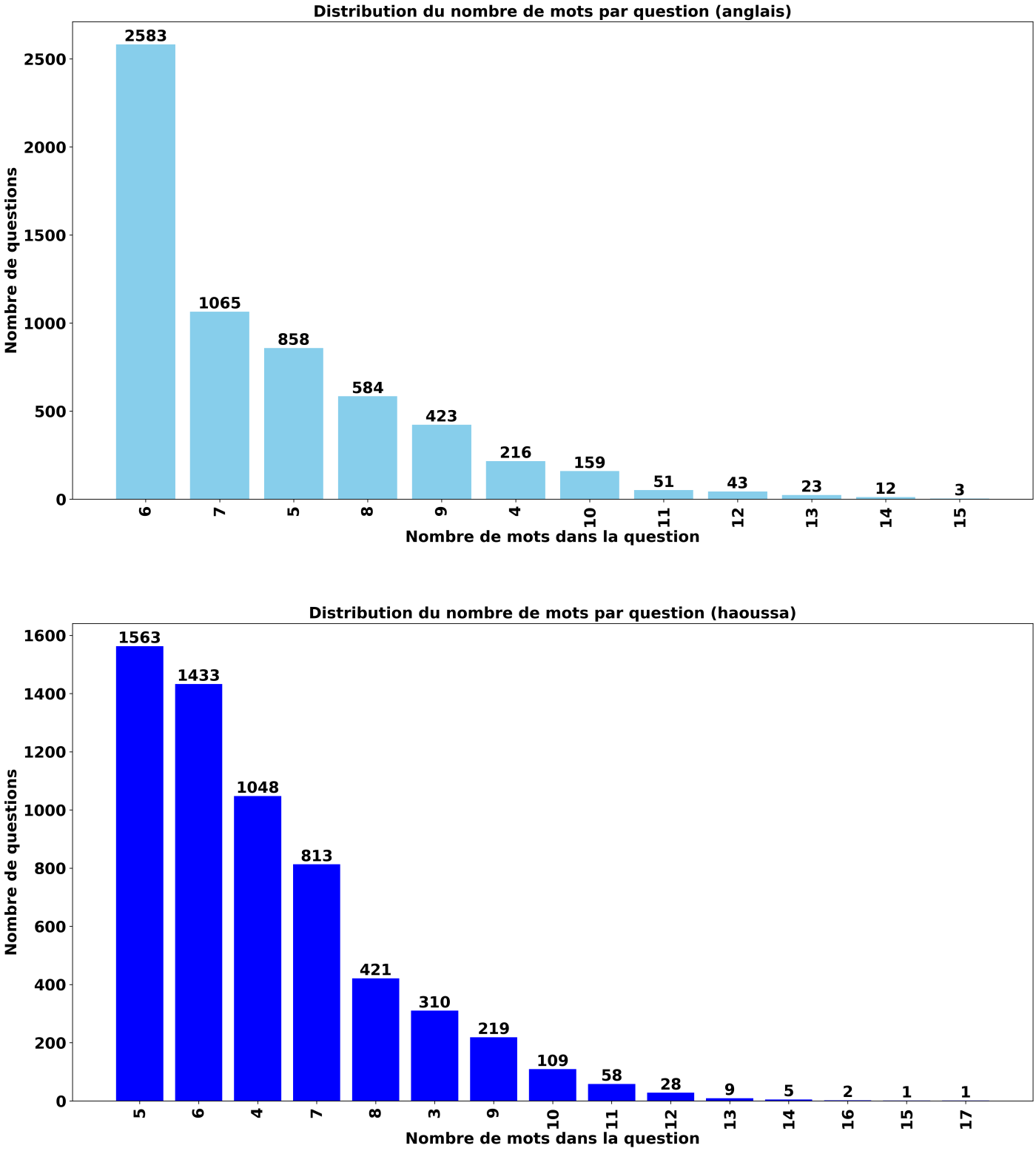

Figure 3.2 – Distribution de la longueur des questions dans $HaVQA$ (EN vs. HA).

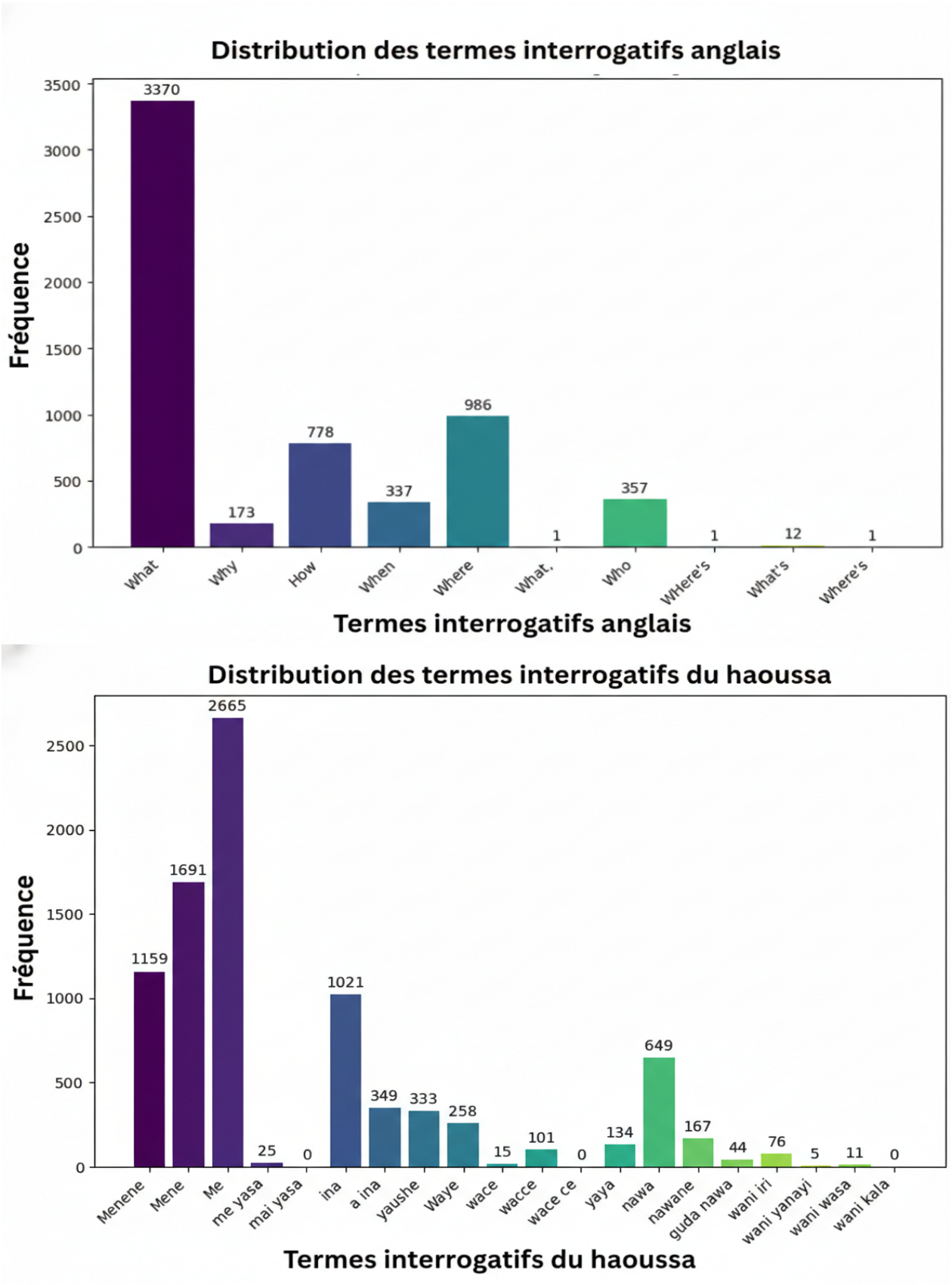

Figure 3.3 – Distribution des termes interrogatifs dans $HaVQA$ (EN vs HA).

3.2.2.1 Analyse de la longueur des questions.

L' analyse de la répartition de la longueur des questions en anglais et en haoussa, comme le montrent la figure 3.2, révèle des disparités marquées entre ces deux langues. En effet, on constate que la plupart des interrogations en anglais contiennent cinq termes, tandis qu'on observe une proportion significative de questions composées de six ou quatre mots. Les phrases complexes, c'est-à-dire celles qui comptent dix mots ou plus, sont quant à elles assez peu nombreuses, ne représentant pas plus de 2% du total. En haoussa, les questions sont légèrement plus courtes, avec une proportion plus élevée de questions de quatre ou cinq mots. Cette caractéristique linguistique s'explique par le fait que le haoussa est une langue agglutinante, qui a la capacité de transmettre des idées en utilisant moins de mots que le serait le cas en anglais. Cette caractéristique peut directement influencer les modèles multimodaux : des entrées plus courtes entraînent une simplification de la complexité syntaxique, mais peuvent aussi restreindre le contexte clair disponible pour la clarification.

3.2.2.2 Analyse des termes interrogateurs.

L'utilisation des termes interrogatifs (Figures 3.3) met en évidence un déséquilibre frappant entre les différents types de questions. En anglais, *What* domine largement avec plus de 3300 occurrences, suivi par *Where*, *How*, et dans une moindre mesure *Who* et *When*. Cette prépondérance indique que la majorité des questions du corpus visent à identifier ou décrire des entités visuelles. En haoussa, la répartition est plus diversifiée mais demeure concentrée autour de quelques formes dominantes, notamment *me*, *mene* et *ina*, qui couvrent respectivement les questions de type "quoi", "quel(le)" et "où". Des mots plus précis, comme yaya («comment»), wane («qui ?») et yaushe («quand ?»), se rencontrent nettement moins souvent. Cette disparité crée une distorsion dans la langue utilisée par les modèles formés avec le jeu de données $HaVQA$. En effet, ils seront davantage habitués à certaines classes de questions, telles que l'identification et la description, plutôt qu'à d'autres, comme le raisonnement temporel et la causalité.

3.2.2.3 Conséquences sur l'apprentissage.

Ces observations mettent en évidence deux défis. D'une part, la longueur moyenne des questions peut varier selon la langue. Cette variation peut affecter la robustesse du modèle, notamment si le prétraitement (tokenisation, padding) n'est pas harmonisé. De plus, une distribution inégale des termes interrogatifs peut fausser les résultats : un modèle peut obtenir de bons résultats globaux en s'appuyant sur des questions fréquentes (par exemple, la détection d'objets en utilisant*What/me*), tout en échouant sur des questions plus rares qui nécessitent une analyse plus approfondie (*Why*/Pourquoi cela s'est-il produit ? *When/yaushe*, Pourquoi cela s'est-il produit plus tard ?). Par conséquent, il est crucial, lors de l'évaluation, d'ajouter une analyse détaillée par type de question aux mesures globales (Accuracy, F1-score). Cela permettra de mettre en évidence les limites du système.

3.2.3 Préparation du jeu de données $HaVQA$

Pour la préparation des données, le nettoyage des textes (questions et réponses) et des images a d'abord été effectué, ainsi qu'un prétraitement partiel du jeu de données. Le reste du prétraitement est ensuite réalisé dans le pipeline d'entraînement en fonction de l'approche utilisée.

— Pour le nettoyage et le prétraitement partiel des textes, les lignes vides ou celles présentant au moins une valeur manquante ont d'abord été supprimées. Ensuite, une nouvelle colonne nommée *label* a été ajoutée qui n'est autre que l'encodage (labélisation) de la colonne de réponse en haoussa avec la classe *LabelEncoder()* de la bibliothèque scikit-learn (scikit-learn developers, 2024). Cette colonne *label* comportant les différentes classes, où chaque classe correspond à une question-reponse, constitue la variable de sortie (variable cible) pour l'entraînement des modèles QRV haoussa dans le cadre de ce travail de recherche. Ainsi, chaque paire (I, Qen) ou (I, Qha) correspond à une étiquette unique. Pour l'entraînement des modèles, la paire (I, Qha) est principalement utilisée.

— Pour le nettoyage et le prétraitement partiel des images, les fichiers vides ne présentant aucune image et les images non associées à une question-réponse dans le jeu de données ont d'abord été supprimés. L'ensemble des images restantes est ensuite redimensionné au format 224 × 224 pixels.

3.3 Environnement expérimental

3.3.1 Infrastructure matérielle

Les expérimentations ont été conduites sur la plateforme Google Colab, utilisant alternativement deux configurations GPU selon la disponibilité :

- Configuration 1 - NVIDIA A100 : GPU basé sur l'architecture Ampere avec 40 Go de VRAM et 83 Go de RAM système, offrant 19.5 TFLOPS (FP32) et 312 TFLOPS en précision mixte (FP16 Tensor Cores).
- Configuration 2 - NVIDIA L4 : GPU basé sur l'architecture Ada Lovelace avec 24 Go de VRAM GDDR6 et 51 Go de RAM système, offrant 30.3 TFLOPS (FP32) et 242 TFLOPS en précision mixte (FP16 Tensor Cores).

Les deux configurations supportent les opérations en précision mixte (bfloat16), permettant d'accélérer l'entraînement tout en réduisant l'utilisation de la mémoire GPU.

3.3.2 Environnement logiciel

L'environnement logiciel comprend les composants suivants :

- Plateforme : Google Colab (Ubuntu 22.04 LTS)
- Langage : Python 3.12
- Framework : PyTorch 2.0+
- CUDA : 12.2 (A100) ou 12.0 (L4)
- Transformateurs : Transformers 4.35+ (Hugging Face)
- Vision : Torchvision 0.15+, Pillow 10.0+
- Données : NumPy 1.24+, Pandas 2.0+
- Visualisation : Matplotlib 3.7+, Seaborn 0.12+

3.3.3 Outils de développement

Les outils employés pour le développement et la reproductibilité incluent :

— Développement : Google Colab Notebooks avec intégration Google Drive

— Stockage : Google Drive

Le code source est disponible publiquement (Mijiyawa, 2025).

3.4 Métriques employées

3.4.1 La métrique Wu-Palmer

La mesure de similitude Wu-Palmer évalue la ressemblance sémantique entre deux notions, concept_1 et concept_2, en se basant sur leur emplacement hiérarchique (Z. Wu & Palmer, 1994). Elle s'appuie sur la profondeur d'un concept dans la hiérarchie de la taxonomie ($\text{depth}(\cdot)$) de chaque concept et sur la profondeur du LCS.

$$Wu - Palmer = \frac{2 \times \text{depth}(\cdot)(\text{LCS})}{\text{depth}(\cdot)(\text{concept}_1) + \text{depth}(\cdot)(\text{concept}_2)}$$

Cette mesure fournit un résultat compris entre 0 et 1, qui reflète le niveau de similarité sémantique entre deux entités.

— Plus la valeur se rapproche de 1, plus la similarité est élevée ;

— Une valeur proche de 0 indique une faible similarité.

3.4.2 La métrique Accuracy

L' exactitude, également connue sous le nom de Accuracy, évalue les performances d'un classificateur. Elle représente le pourcentage de prédictions correctes sur l'ensemble des données :

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Où :

— TP : Vrais résultats positifs (*True Positives*)

— TN : Vrais résultats négatifs (*True Negatives*)

— FP : Faux résultats positifs (*False Positives*)

— FN : Faux résultats négatifs (*False Negatives*)

Cette mesure se révèle pratique pour évaluer des ensembles de données dont les classes sont équilibrées. Toutefois, elle peut produire des résultats trompeurs si les classes ne le sont pas (Fawcett, 2006).

3.4.3 La métrique F1-score

La métrique F1-score est une mesure harmonique combinant laPrecision et leRecall d'un modèle de classification binaire. Elle est particulièrement utile pour des jeux de données déséquilibrés, car elle fournit une évaluation équilibrée entre les FP et les FN.

$$F1 - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

où :

— Precision = $\frac{TP}{TP+FP}$

— Recall = $\frac{TP}{TP+FN}$

Le F1-score est élevé uniquement lorsque la Precision et le Recall sont tous deux élevés, ce qui en fait une mesure robuste pour les tâches de classification avec déséquilibre de classes (Sasaki, 2007).

3.4.4 La métrique perte / Loss

En apprentissage automatique, la loss (ou *loss function*) quantifie l'erreur entre les prédictions et les valeurs réelles attendues. Par exemple, on définit la $\mathcal{L}_{\text{cross-entropy}}$ comme suit :

$$\mathcal{L}_{\text{cross-entropy}} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

où :

— y_i est la vraie étiquette (valeur binaire ou catégorielle),

— $\hat{y}_i$ est la probabilité prédite pour la classe correcte,

— N est le nombre d'exemples.

Une fonction de perte bien choisie est essentielle pour la convergence du modèle et la qualité finale des prédictions (Heaton, 2017).

3.5 Architecture et modèles utilisés

3.5.1 Les grands modèles de langue (GML) utilisés

Le tableau TABLE 3.2 présente les différents GML ainsi que le nombre de leurs paramètres qui ont été utilisés dans ce travail. Les performances en haoussa ont été évaluées sur un ensemble de modèles comprenant mT0-base, mT0-large, BLOOMZ-560M, BLOOMZ-1B7, AfriBERTa, Afro-XLMR, Llama-3-8B, DeepSeek-R1 et Gemini (D. I. Adelani, Liu et al., 2024b ; DeepSeek-AI, 2025 ; Gemini Team et al., 2025 ; Meta AI, 2024 ; Muennighoff, Wang, Sutawika, Roberts, Biderman, Scao et al., 2023a, 2023b ; Muennighoff et al., 2022a, 2022b ; Ogueji et al., 2021a). D'après ce tableau, on voit que certains GML utilisés sont de base préentraînés sur le haoussa, mais aucun de ces GML n'est de base affiné sur le haoussa.

GML	Nombre de paramètres	Préentraîné sur le haoussa	Affiné sur le haoussa
mt0-base	580 millions	Oui	Non
mt0-large	1,2 milliard	Oui	Non
afriberta_large	126 millions	Oui	Non
afro-xlmr-large-76L	560 millions	Oui	Non
gemini	770 millions	Oui	Non
bloomz560	560 millions	Non	Non
bloomz1b7	1,7 milliard	Non	Non
Llama-3.2-1B	1,23 milliard	Non	Non
deepseek-R1-1.5B	1,5 milliard	Non	Non

Table 3.2 – Nombre de paramètres, préentraînement et fine-tuning en haoussa des différents GML utilisés pour le système de QRV.

3.5.2 Les transformateurs d'images (TI) utilisés

Le tableau TABLE 3.3 montre les diverses options d'images ainsi que le nombre correspondant de paramètres employés lors de nos recherches. Les TI suivants ont été intégrés à notre chaîne de traitement :

- ViT : un transformateur de vision standard pré-entraîné sur ImageNet-21k ;
- CLIP : un modèle alignant représentations visuelles et textuelles ;
- MAE : un auto-encodeur masqué optimisé pour la reconstruction d'images ;
- DeiT : une architecture conçue pour un entraînement efficace avec moins de données.

Toutes ces architectures sont documentées dans les travaux de Dosovitskiy et al. (2021b), K. He et al. (2022), Radford et al. (2021) et Touvron et al. (2021b).

TI	Nombre de paramètres
vit-base-patch16-224-in21k	86,4 millions
clip-vit-base-patch32	149 millions
mae-base	86 millions
deit-base-patch16-224	86 millions

Table 3.3 – Nombre de paramètres des différents encodeurs (transformateurs) d'images utilisés pour le système QRV-Haoussa

3.5.3 Stratégies de fusion multimodale

Le système QRV appliqué à n'importe quelle langue, la fusion des modalités visuelle et textuelle est un défi crucial pour assurer l'efficacité du modèle. On distingue plusieurs techniques d'intégration multimodale.

3.5.3.1 Fusion précoce (early fusion)

La fusion précoce est une technique qui consiste à associer la représentation de l'information textuelle et visuelle dès les premiers niveaux du modèle, généralement après avoir projeté ces informations dans un espace latent partagé. Cette méthode offre une interconnexion détaillée entre les deux formes de données, mais elle peut s'avérer coûteuse en termes de calculs et sensible aux perturbations, notamment lorsqu'elle est appliquée à des langues peu dotées en ressources, telles que le haoussa (Ilharco et al., 2019).

3.5.3.2 Fusion tardive (late fusion)

La fusion tardive consiste à traiter séparément les deux modalités (texte et image) par des encodeurs spécialisés (par exemple, GML pour le texte et TI pour l'image), avant de combiner leurs représentations à un niveau décisionnel (par concaténation ou moyenne pondérée). Cette approche est plus modulaire, mais elle peut limiter les interactions croisées entre le texte et l'image (Cho et al., 2021).

3.5.3.3 Fusion croisée (cross-modal attention)

La fusion par attention croisée, plus récente et plus performante, permet à chaque modalité (texte ou image) d'influencer dynamiquement la représentation de l'autre. Les modèles LXMERT et ViLBERT intègrent des unités d'attention multi-têtes entre l'image et le texte, ce qui leur permet de comprendre en profondeur le contexte visuel influencé par la question (L. H. Li et al., 2019 ; Tan & Bansal, 2019). Cette stratégie fonctionne très bien pour capturer les relations sémantiques complexes, mais elle nécessite une quantité considérable de données annotées. Cela constitue un défi dans le cas des langues à faibles ressources, comme le haoussa.

3.5.3.4 Approche adoptée

Nous avons choisi une méthode de fusion tardive, qui consiste à combiner les représentations contextuelles (embeddings) obtenues à partir d'un TI et d'un GML adapté au haoussa. Cette fusion se fait par l'intermédiaire d'une couche de projection, suivie d'une concaténation, avant qu'une tête de classification ne soit appliquée. Cette architecture offre un équilibre satisfaisant entre expressivité et réalisabilité dans des contextes à faible ressource, comme dans le cas de la langue haoussa.

3.6 Entraînement et raffinement des modèles par classification

3.6.1 Configuration des hyperparamètres

L'ensemble des hyperparamètres utilisés pour l'entraînement de nos modèles est détaillé ci-dessous. Ces hyperparamètres sont classés en trois catégories : les hyperparamètres fixes qui restent constants durant l'entraînement, les hyperparamètres dynamiques qui évoluent selon des stratégies d'optimisation ou définissent la durée totale d'entraînement, et les critères de sélection du modèle optimal.

3.6.1.1 Hyperparamètres fixes

Les hyperparamètres suivants restent constants tout au long de l'entraînement et définissent la configuration de base du processus d'apprentissage.

- ☐ Reproductibilité et stockage : La graine aléatoire (`seed`) a été fixée à 12345 pour garantir la reproductibilité des expérimentations. Cette valeur permet des comparaisons fiables entre différentes configurations de modèles en éliminant les variations dues à l'initialisation aléatoire. Les checkpoints sont stockés dans le répertoire `checkpoint`.
- ☐ Taille des lots et gestion des données : Une taille de lot a été fixée à 32 échantillons par GPU (`per_device_train_batch_size = 32`, `per_device_eval_batch_size = 32`). Cette valeur se situe dans la plage recommandée (16–64) pour le fine-tuning de modèles vision-langage et reste compatible avec les contraintes mémorielles de nos environnements d'expérimentation, notamment le GPU NVIDIA L4 (24 Go de VRAM) et le NVIDIA A100 (40 Go de VRAM) disponibles sur Google Colab. Cette configuration permet un bon compromis entre stabilité d'entraînement et efficacité de l'utilisation mémoire. Le chargeur de données (`DataLoader`) exploite 8 processus parallèles (`dataloader_num_workers = 8`), afin d'accélérer le chargement des données et d'éviter que le GPU ne reste inactif en attente d'E/S, tout en maintenant une charge CPU raisonnable dans l'environnement Colab. Nous conservons l'ensemble des colonnes du jeu de données (`remove_unused_columns = False`) afin de préserver l'accès aux métadonnées nécessaires aux différentes étapes de traitement multimodal.
- ☐ Régularisation : Une régularisation L2 a été appliquée avec un coefficient de 1×10^{-4} (`weight_decay = 1e-4`), valeur standard dans la littérature pour le fine-tuning de transformateurs (Devlin et al., 2019b). Des valeurs plus élevées (10^{-3}) contraignent excessivement les poids, tandis que des valeurs plus faibles (10^{-5}) offrent une régularisation insuffisante pour notre ensemble d'entraînement de taille limitée.
- ☐ Précision numérique : La précision mixte bfloat16 a été activée (`bf16 = True`), permettant une réduction de 50 % de l'utilisation mémoire GPU et des gains de vitesse jusqu'à 2× par rapport à la précision mixte float32. Contrairement à la précision mixte float16, bfloat16 conserve la même plage d'exposant que float32, éliminant les problèmes de débordement numérique.
- ☐ Fréquence d'évaluation et de sauvegarde : L'évaluation est effectuée toutes les 100 étapes (`evaluation_strategy = "steps"`), correspondant approximativement à 0.8 époque avec notre configuration. Cette fréquence représente un compromis entre surveillance des per-

formances et temps d'évaluation. Les métriques sont enregistrées (`logging_steps = 100`) et les checkpoints sauvegardés (`save_steps = 100`) à la même fréquence, avec une limite de trois checkpoints conservés (`save_total_limit = 3`) pour gérer l'espace disque. Le format PyTorch est utilisé (`save_safetensors = False`).

- ☐ Phase de préchauffage (warmup) : Une phase de préchauffage couvre les 1 % premières étapes (`warmup_ratio = 0.01`), durant laquelle le taux d'apprentissage augmente linéairement de zéro à 1×10^{-5}. Cette technique stabilise l'entraînement initial en évitant des mises à jour brutales causées par les gradients initiaux de grande magnitude. Le ratio de 1 % (environ 120 étapes) offre une stabilisation suffisante sans consommer une portion excessive du budget d'entraînement.

3.6.1.2 Hyperparamètres dynamiques

Les hyperparamètres suivants évoluent au cours de l'entraînement ou définissent sa durée totale, permettant une adaptation progressive du processus d'apprentissage.

- ☐ Durée d'entraînement : Les modèles sont entraînés pendant 60 à 80 époques selon la configuration ($\texttt{num\_train\_epochs} \in [60, 80]$). Ce nombre a été déterminé empiriquement : des expérimentations préliminaires ont révélé une convergence stable autour de 60–80 époques. Au-delà de 80 époques, un surapprentissage a été observé, tandis qu'en deçà de 60 époques, le modèle n'atteignait pas son plein potentiel (sousapprentissage). Cette durée évolue conceptuellement durant l'entraînement (progression de 0 à 80 époques) et détermine la durée totale du processus d'optimisation, justifiant son classement parmi les hyperparamètres dynamiques.
- ☐ Taux d'apprentissage avec ordonnanceur : Le taux d'apprentissage initial est fixé à 1×10^{-5} (`learning_rate = 1e-5`), valeur adaptée au raffinement (fine-tuning) pour éviter l'*Oubli catastrophique* (*catastrophic forgetting*). Des expérimentations préliminaires avec 10^{-4} ont montré une convergence instable, tandis que 10^{-6} convergeait trop lentement. Un ordonnanceur linéaire (`lr_scheduler_type = "linear"`) diminue progressivement ce taux de sa valeur initiale jusqu'à zéro, permettant une exploration large en début d'entraînement et

une convergence fine en fin d'entraînement.

- Évolution temporelle du taux d'apprentissage : Le taux d'apprentissage suit en moyenne la trajectoire suivante :
  - Phase 1 (0–1 % des étapes) : Augmentation linéaire de 0 à 1×10^{-5}
  - Phase 2 (1–100 % des étapes) : Décroissance linéaire de 1×10^{-5} à 0

3.6.1.3 Critères de sélection du modèle optimal

Nous définissons des critères de sélection pour identifier le meilleur modèle parmi tous les checkpoints générés durant l'entraînement.

- Métrique de performance : Le meilleur modèle est identifié en surveillant la perte d'évaluation (`metric_for_best_model = "eval_loss"`) que nous minimisons (`greater_is_better = False`). La perte a été préférée à la précision car elle offre une sensibilité plus fine aux variations de performance, capturant des différences subtiles dans les probabilités prédites plutôt que des changements discrets de classification. Elle est également directement liée à l'objectif d'optimisation, assurant une cohérence entre entraînement et sélection.
- Chargement du modèle optimal : Le checkpoint ayant obtenu la meilleure perte d'évaluation est automatiquement chargé à la fin de l'entraînement (`load_best_model_at_end = True`). Cette stratégie garantit que le modèle final correspond aux meilleures performances observées plutôt qu'au dernier checkpoint, qui pourrait présenter un surapprentissage. Cette approche est particulièrement pertinente pour notre contexte de données limitées ($HaVQA$).

3.6.1.4 Récapitulatif

Le tableau 3.4 résume l'ensemble de ces hyperparamètres avec leurs valeurs et leur nature.

Table 3.4 – Récapitulatif des hyperparamètres d'entraînement

Type	Hyperparamètre	Valeur / Valeur moyenne
Fixes	Graine aléatoire (`seed`)	12345
	Taille de lot par GPU (train)	32
	Taille de lot par GPU (eval)	32
	Régularisation L2 (`weight_decay`)	1×10^{-4}
	Précision mixte (`bf16`)	True
	Fréquence d'évaluation	100 étapes
	Fréquence de sauvegarde	100 étapes
	Limite de checkpoints	3
	Processus de chargement	8
	Conservation colonnes	False
	Format de sauvegarde	PyTorch
Dynamiques	Nombre d'époques (`num_train_epochs`)	entre 60 et 80
	Taux d'apprentissage (initial)	1×10^{-5}
	Ordonnanceur	linear ($\rightarrow$ 0)
	Phase de préchauffage	1 % (0 $\rightarrow 10^{-5}$)
Sélection	Métrique de sélection	eval_loss
	Direction d'optimisation	minimisation
	Chargement meilleur modèle	Oui

3.6.1.5 Résumé du processus d'entraînement des modèles de question-réponse visuelle (QRV)

Un pipeline d'entraînement crée un modèle multimodal QRV composé. Une étape intermédiaire, appelée `collator`, prend soin de traiter chaque ensemble de données. Cette phase consiste à analyser la structure des phrases (tokenisation) ainsi qu'à effectuer certaines opérations de transformation sur les images. Le texte et l'image sont encodés avec la classe `AutoModel` de la librairie `transformers` (*mean-pooling* pour le texte et `pooler_output` ou *mean-pooling* pour l'image), puis fusionnés avant une couche de classification. L'entraînement via l'objet `Trainer` utilise des `TrainingArguments` avec évaluation, journalisation et sauvegarde toutes les 100 itérations (`save_total_limit=3`, rechargement automatique du meilleur modèle). Le plan d'entraînement est entre 60 et 80 époques ($\texttt{num\_train\_epochs} \in [60, 80]$), avec un scheduler linéaire et un échauffement proportionnel (`warmup_ratio=0.01`), l'activation du `bf16` (`bf16=True`), un taux d'apprentissage de `1e-5`, une pénalisation `weight_decay=1e-4` et `dataloader_num_workers=8`. Les métriques calculées sont Wu-Palmer, Accuracy et F1-score. Deux callbacks assurent un affi-

chage lisible des logs et un arrêt précoce `Exact20` qui compare la perte d'entraînement cumulée à la perte de validation et interrompt si l'écart relatif atteint une cible (20 % ou 5 %) dans une tolérance donnée. Enfin, l'exécution suit les fonctions `train()` puis `evaluate()`.

3.6.2 Explication du fine-tuning du système question-réponse visuelle (QRV) par classification en langue haoussa

Dans le contexte des langues à faibles ressources, comme le haoussa, le système QRV est formulé ici comme un système de classification multi-classe. Il est question d' identifier la meilleure option de réponse à partir d'un groupe restreint de choix, étant donné qu'une image et une question lui correspondent. Ce cadre restrictif, appelé *classification à vocabulaire prédéfini*, permet de contourner les limites des générateurs de texte dans les contextes où les données annotées sont rares ou peu diversifiées (Agrawal et al., 2016 ; Tan & Bansal, 2019).

3.6.2.1 Architecture générale du système question-réponse visuelle QRV

Ce système fonctionne grâce à deux encodeurs spécialisés :

— Un encodeur (transformateur) visuel (comme ViT ou CLIP-Vision), qui génère des représentations vectorielles à partir d'une image.

— Un encodeur (transformateur) textuel (GML multilingue ou spécifique au haoussa) qui convertit la question en vecteurs sémantiques.

Ces deux représentations sont combinées (soit par concaténation, soit par projection dans un espace commun), puis soumises à une tête de classification (généralement une ou plusieurs couches linéaires) qui prédit la classe (résultat) la plus probable.

3.6.2.2 Stratégie d'ajustement fin (fine-tuning) des modèles

L' ajustement fin consiste à adapter simultanément le système aux particularités du jeu de données $HaVQA$ (Parida et al., 2023). Cela exige :

— De limiter ou d'interdire complètement l'accès aux couches des transformateurs préentraînés (GML, ViT) pour prévenir le surapprentissage (Chronopoulou et al., 2019 ; Howard &

Ruder, 2018).

— Entraînement de la tête de classification sur les triplets (image, question, réponse) du corpus annoté. Dans notre cas, les triplets sont sous la forme *(image*, *question en haoussa*, *réponse en haoussa)* (Parida et al., 2023). Une fonction de coût de type entropie croisée (*cross-entropy*), spécialement développée pour la classification supervisée, est utilisée (Goodfellow et al., 2016).

3.7 Techniques d'augmentation des données

3.7.1 Technique d'augmentation de données en ligne

L' augmentation de données en ligne consiste à générer, dynamiquement, des variantes des données d'entrée pour chaque itération ou sous-ensemble durant l'entraînement du modèle. Contrairement à l'augmentation hors ligne, qui consiste à créer un jeu de données élargi préalablement stocké sur disque, l'augmentation en ligne s'effectue à la volée, ce qui réduit l'espace mémoire nécessaire et introduit de la diversité à chaque passe dans le modèle (Shorten & Khoshgoftaar, 2019a).

3.7.1.1 Objectifs et utilité pour la langue haoussa

Dans le contexte du système QRV-haoussa, une langue à faible ressource, l'ajout en ligne vise à minimiser le surapprentissage et à améliorer l'efficacité du modèle face à la variabilité linguistique et visuelle. Cela s'avère particulièrement bénéfique lorsque les données étiquetées se font rares ou ne reflètent pas adéquatement la richesse des expressions courantes.

3.7.1.2 Méthodes textuelles pour l'expansion

Plusieurs méthodes permettent d'accroître la composante textuelle (les questions) : remplacement de termes par leurs synonymes à l'aide de WordNet ou de modèles multilingues comme mBERT, insertion, suppression ou recombinaison de mots sans modification de la structure grammaticale. En raison de l'insuffisance des ressources pour la langue haoussa, le choix a été fait de ne pas utiliser de techniques d'augmentation textuelle afin d'éviter des biais potentiels susceptibles

d'influencer défavorablement l'efficacité de l'apprentissage du système de question-réponse visuelle, comme illustré à la figure 3.5. Des expérimentations préliminaires visant à accroître les données en ligne ont en effet révélé que certains remplacements par synonymie produisent des mots absents du vocabulaire haousa.

3.7.1.3 Méthodes visuelles d'augmentation

En ce qui concerne la modalité image, des transformations telles que la rotation aléatoire, le recadrage, le zoom et l'inversion horizontale sont appliquées grâce à des bibliothèques telles que `torchvision` :

- Rotations aléatoires ;
- Recadrages ;
- Agrandissement ;
- Inversions horizontales ;
- Ajustement du contraste et de l'éclairage, avec ajout de bruit aléatoire.
- Masquage aléatoire pour simuler des zones manquantes.

En ce qui concerne ce travail de recherche, la méthode de rotation aléatoire et la méthode de retournement horizontal aléatoire ont été utilisées, comme le montre la figure 3.5.

3.7.1.4 Avantages de l'approche en ligne

L'approche en ligne présente plusieurs avantages :

- Elle offre une grande diversité d'exemples, sans augmenter la taille physique du jeu de données.
- Grâce à cette technique, les algorithmes peuvent mieux se prémunir contre diverses formes de bruit et ainsi améliorer leur taux de réussite.
- L'intégration de cette stratégie au processus d'apprentissage est relativement simple, ne nécessitant pas de phases de prétraitement complexes.

3.7.2 Technique d'augmentation de données hors ligne

La méthode de l'augmentation de données hors ligne consiste à produire de nouvelles données transformées avant l'entraînement du modèle, puis à les enregistrer sur disque pour une utilisation ultérieure. Contrairement aux exemples additionnels, qui ne sont présents que temporairement pendant la phase d'entraînement, ces exemples sont intégrés de manière permanente, ce qui offre un meilleur contrôle sur les données présentées au modèle à chaque étape (Shorten & Khoshgoftaar, 2019a).

3.7.2.1 Objectifs et utilité pour le haoussa

Dans les situations de pénurie de ressources, comme c'est le cas pour la langue haoussa, l'apprentissage hors ligne permet d'étendre le corpus d'entraînement en multipliant les paires de questions et d'images annotées de manière artificielle. Cette stratégie s'avère appropriée quand on possède une capacité de stockage adéquate et qu'on désire maintenir un certain niveau de stabilité quant aux données perçues par le modèle. Elle permet d'assurer une diversité, sans avoir à procéder constamment au recalcul des variations visuelles.

3.7.2.2 Méthodes textuelles d'augmentation

En mode hors ligne, la formulation des questions peut être améliorée en utilisant diverses techniques. Voici quelques-unes de ces techniques :

— La traduction bidirectionnelle avec des modèles linguistiques multiples permet de reformuler une question sans en modifier le sens initial (Sennrich et al., 2016a).

— Le remplacement contextuel de mots est possible grâce à des modèles tels que mBERT et AfriBERTa, assurant ainsi une correspondance linguistique plus précise pour les langues africaines (Alabi et al., 2022b).

— La génération paraphrastique supervisée, qui utilise des modèles génératifs préentraînés, est limitée pour le haoussa en raison de la rareté de corpus parallèles.

Toutefois, comme indiqué précédemment, les techniques d'augmentation textuelle hors ligne pour le haoussa ont été exclues. Cela permet d'éviter les biais découlant des erreurs de traduction, des

synonymes inappropriés, etc., qui sont fréquents dans les contextes multilingues à faibles ressources (Nekoto, Marivate, Matsila, Fasubaa, Fagbohungbe, Akinola et al., 2020).

3.7.2.3 Méthodes visuelles d'augmentation

Les images du jeu de données peuvent être transformées hors ligne par des scripts automatisés en appliquant :

- Transformations géométriques : rotation et retournement horizontal
- Dégradations simulées : flou, bruit, occlusion
- Ajustements photométriques : couleur, contraste et saturation

Dans ce travail, le jeu de données $HaVQA$ a été augmenté en appliquant des rotations ($\pm$15°) et des retournements horizontaux. Ces transformations permettent de générer 2 à 3 versions distinctes de chaque image initiale. Les images augmentées ont été sauvegardées et réutilisées pour tous les entraînements ultérieurs, comme illustré à la figure 3.6.

3.7.2.4 Avantages et limites

L'approche hors ligne présente plusieurs avantages :

- Meilleure traçabilité et reproductibilité des transformations appliquées
- Réduction des coûts de calcul durant l'entraînement
- Adaptation optimale aux environnements disposant d'un espace de stockage suffisant

Toutefois, cette méthode est moins flexible que l'approche en ligne. Une fois les données augmentées enregistrées, il devient difficile de varier dynamiquement les transformations d'une époque à l'autre. Cette limitation peut réduire la robustesse du modèle face à des perturbations imprévues (Shorten & Khoshgoftaar, 2019a).

3.8 Architectures expérimentales du système QRV

3.8.1 Architecture sans augmentation de données

3.8.1.1 Configuration de référence

Dans cette configuration, le modèle est entraîné sur la version originale du jeu de données $HaVQA$, présentée au tableau 3.5. Chaque instance d'entraînement est un triplet composé de : (i) une question en haoussa Qha, (ii) son image associée I, et (iii) la réponse Rha correspondante.

3.8.1.2 Architecture du modèle

La figure 3.4 illustre le flux de données et l'architecture d'entraînement du modèle sans augmentation. Le traitement s'effectue en plusieurs étapes :

1. Encodage textuel : La question est tokenisée et encodée par un GML, produisant une représentation vectorielle du texte
2. Encodage visuel : L'image est divisée en patches et encodée par un TI, générant une représentation vectorielle de l'image
3. Fusion multimodale : Les représentations textuelles et visuelles sont combinées par un mécanisme d'attention conjointe en une représentation unifiée
4. Classification : Une tête de classification projette ce vecteur fusionné dans l'espace des réponses prédéfinies
5. Prédiction : La classe ayant le score le plus élevé est sélectionnée comme réponse finale

Les résultats de cette configuration de référence sont rapportés dans le tableau 4.1.

3.8.1.3 Motivation pour l'augmentation des données

Bien que l'augmentation des données soit couramment utilisée dans les systèmes unimodaux du TALN et de la vision par ordinateur, elle reste peu exploitée dans les systèmes QRV. Or, dans le contexte du haoussa, langue à faibles ressources, les transformations synthétiques offrent plu-

sieurs avantages : extension du nombre d'exemples d'entraînement, amélioration de la robustesse du modèle, et réduction du surapprentissage (Feng et al., 2021; Shorten & Khoshgoftaar, 2019a).

	Train	Test	Images
$HaVQA$	4 816	1 204	1 555

Table 3.5 – Détails sur le dataset $HaVQA$ et ses partitions pour l'entrainement et l'évaluation. Source : (Parida et al., 2023)

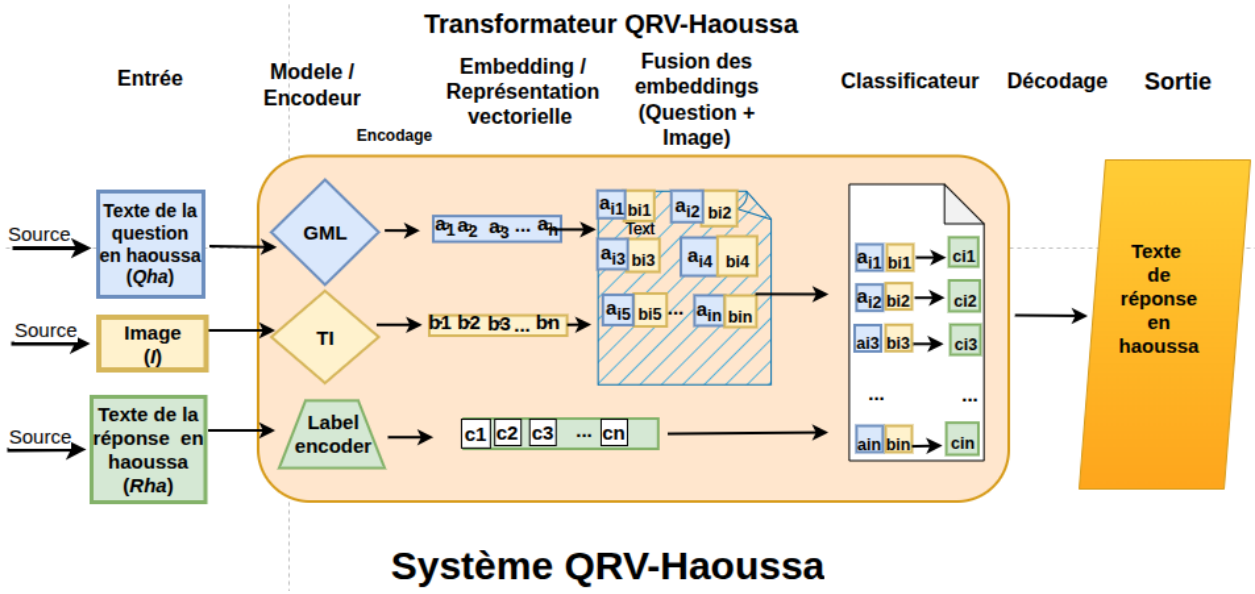

Figure 3.4 – Architecture du système de question-réponse en langue haoussa combinant un grand modèle de langue (GML) et un transformateur d'image (TI), inspirée des travaux de (Parida et al., 2023)

3.8.2 Architecture du système par classification avec augmentation en ligne

3.8.2.1 Principe de l'augmentation en ligne

L'approche d'augmentation en ligne conserve la répartition originale du jeu de données $HaVQA$ (tableau 3.5) et applique des transformations stochastiques à chaque propagation avant (figure 3.5). Chaque instance d'entraînement est un triplet (I, Qha, Rha) composé de l'image I, de la question en haoussa Qha et de la réponse de référence Rha.

3.8.2.2 Transformations appliquées

Seule la composante visuelle est modifiée durant l'entraînement. Les transformations suivantes sont appliquées séquentiellement :

1. Retournement horizontal avec une probabilité de 0.5
2. Rotation aléatoire légère, selon la méthode proposée par Krizhevsky et al. (2012)

Cette stratégie permet de générer une variété pratiquement infinie de vues tout en préservant la sémantique des objets représentés.

3.8.2.3 Architecture de traitement

La question non modifiée est tokenisée puis encodée par l'encodeur GML, tandis que l'image augmentée est traitée par un TI. Les représentations textuelles et visuelles sont fusionnées par un mécanisme d'attention croisée, puis transmises à une tête de classification qui attribue un score à chaque étiquette de réponse. La classe ayant le score le plus élevé est sélectionnée comme prédiction finale.

3.8.2.4 Avantages de l'approche en ligne

L'augmentation en ligne présente plusieurs avantages :

— Aucun stockage supplémentaire requis

— Maintien d'un entraînement rapide

— Introduction d'une diversité visuelle continue

— Réduction du surapprentissage sans introduction de biais linguistique

Ces caractéristiques sont particulièrement précieuses dans le contexte du haoussa, langue à faibles ressources. Les résultats obtenus avec cette méthode sont présentés dans le tableau 4.4.

3.8.2.5 Vers l'augmentation hors ligne

Bien que les transformations en temps réel renforcent la robustesse du modèle, une stratégie d'augmentation hors ligne est également explorée. Cette approche consiste à créer, avant l'en-

traînement, un jeu de données élargi et fixe dans lequel les modalités visuelle et textuelle sont augmentées. Ce pool précalculé de paires texte-image permet d'évaluer les bénéfices d'une approche hors ligne par rapport à l'augmentation en ligne, conformément aux recommandations issues des travaux sur l'enrichissement de données multimodales (Kafle et al., 2017 ; Shorten & Khoshgoftaar, 2019a). La section suivante fournit une description détaillée de ce protocole hors ligne.

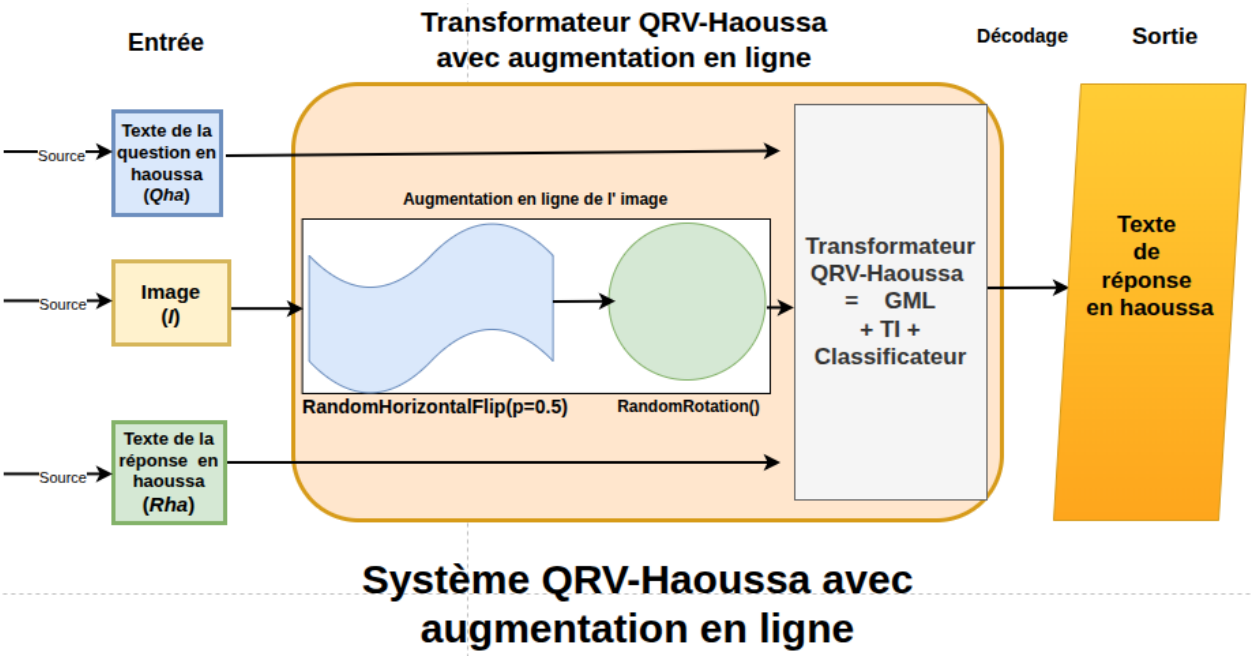

Figure 3.5 – Architecture du système question-réponse visuelle (QRV) de la langue haoussa combinant un grand modèle de langue (GML) et un transformateur d'image (TI) avec une augmentation des données en ligne basé sur les travaux de (Parida et al., 2023 ; Wei & Zou, 2019)

3.8.3 Technique et architecture du système de question-réponse visuelle par classification avec augmentation des données hors ligne

3.8.3.1 Pipeline d'augmentation hors ligne

La figure 3.6 illustre le pipeline d'augmentation hors ligne, qui étend le jeu de données original (tableau 3.5) en appliquant simultanément des transformations textuelles et visuelles, suivies d'une traduction de l'anglais vers le haoussa. Cette approche vise à pallier la rareté des données et à réduire les biais potentiels (Gong et al., 2022 ; W. He et al., 2023 ; Z. Liu et al., 2023).

3.8.3.2 Processus d'augmentation

Concrètement, le processus s'effectue en plusieurs étapes pour chaque triplet (question, réponse, image) du jeu de données anglais :

1. Duplication du triplet original
2. Substitution de synonymes dans les textes dupliqués (Wei & Zou, 2019)
3. Traduction automatique des textes de l'anglais vers le haoussa (Parida et al., 2023)
4. Application de transformations visuelles : retournement horizontal aléatoire et rotation (Shorten & Khoshgoftaar, 2019a)

L'ensemble des données originales est préservé, et le processus génère un jeu de données augmenté, $HaVQAaug$, de taille fixe équivalente au double du jeu de données initial. La distribution des sous-ensembles d'entraînement et de test est présentée dans le tableau 4.7.

3.8.3.3 Jeu de données résultant

Le jeu de données $HaVQAaug$ ainsi créé est accessible publiquement sur le lien : (Mijiyawa, 2025). Ce corpus enrichi permet d'entraîner le système QRV composé d'un encodeur GML, d'un transformateur visuel et d'un classificateur de fusion, conformément à l'architecture de base, afin de produire des réponses précises en haoussa.

	Train	Test	Images
$HaVQAaug$	9 625	2 407	3 110

Table 3.6 – Détails sur les partitions du jeu de données $HaVQAaug$ après augmentation hors ligne du jeu de données $HaVQA$ pour l'entraînement et l'évaluation. Source : (Parida et al., 2023)

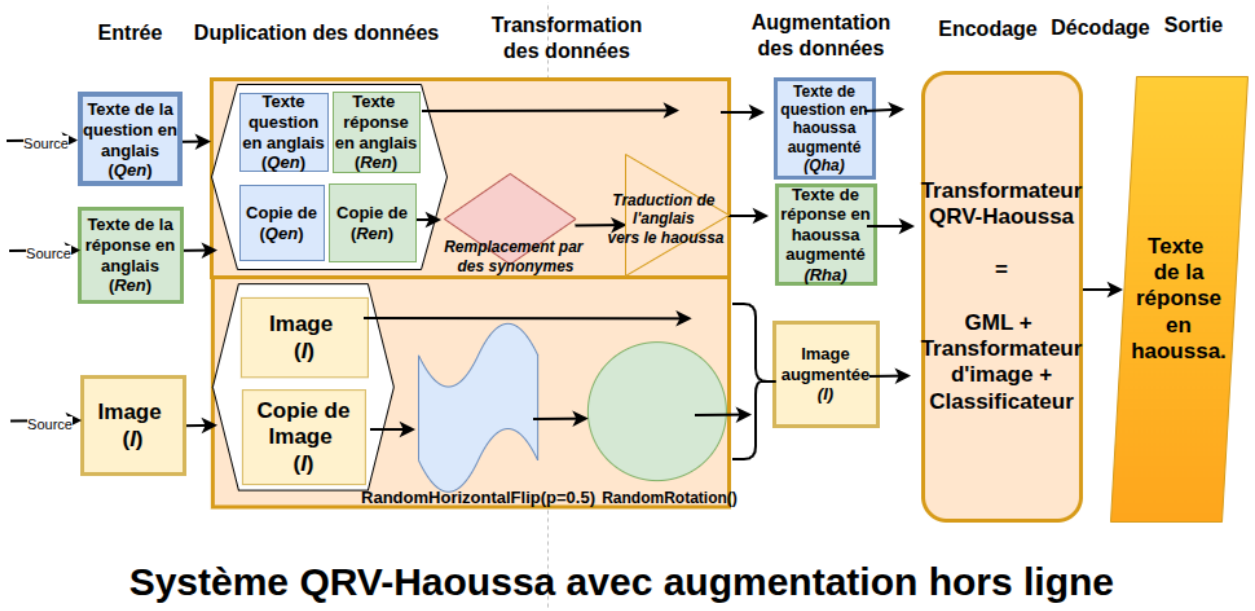

Figure 3.6 – Architecture du système question-réponse visuelle (QRV) pour le haoussa avec augmentation hors ligne inspirée des travaux de (Parida et al., 2023)

3.9 Conclusion

Ce chapitre a présenté la méthodologie complète pour développer un système de question-réponse visuelle pour le haoussa. L'analyse du jeu de données $HaVQA$ a révélé des caractéristiques linguistiques importantes : questions plus courtes en haoussa et distribution déséquilibrée des termes interrogatifs. Les métriques d'évaluation (Accuracy, F1-score, Wu-Palmer) et les modèles sélectionnés (9 GML et 4 TI) ont été justifiés par leur pertinence pour les langues à faibles ressources. L'architecture adoptée repose sur une fusion tardive, offrant un compromis efficace entre modularité et performance. Trois stratégies d'entraînement ont été expérimentées : sans augmentation (configuration de référence), avec augmentation en ligne (transformations visuelles dynamiques), et avec augmentation hors ligne (corpus $HaVQAaug$ obtenu par augmentation hors ligne à partir de $HaVQA$). Ces approches visent à pallier la rareté des données, accroître la robustesse et améliorer les performances. Les hyperparamètres ont été optimisés pour ce contexte à faibles ressources. Le cadre méthodologique établi permet d'évaluer 36 variantes de modèles selon trois régimes distincts. Le chapitre suivant présente l'analyse comparative des résultats, identifiant les configurations les plus performantes pour le système de question-réponse visuelle en haoussa.

CHAPITRE 4
RÉSULTATS EXPÉRIMENTAUX ET ANALYSES

4.1 Introduction

Après avoir défini le cadre méthodologique et les choix d'architectures multimodales, ce chapitre se consacre à l'exposition et à l'interprétation des résultats obtenus. L'objectif principal est de mesurer l'efficacité des différentes combinaisons entre grands modèles de langue (GML) et transformateurs d'images (TI) pour le système de Question-Réponse Visuelle (QRV) de la langue haoussa. La performance du système est évaluée à travers trois protocoles d'entraînement distincts, permettant d'isoler l'impact des stratégies d'augmentation de données : une approche de base sans augmentation de données ; une approche avec augmentation en ligne (dynamique) ; une approche avec augmentation hors ligne (prétraîtée). Pour chacune de ces configurations, 36 duos de modèles ont été testés et comparés selon trois indicateurs clés : l'Accuracy, le F1-score (macro) et la métrique de Wu-Palmer. L'analyse qui suit s'articule autour de la comparaison des performances globales, de l'étude de la robustesse des transformateurs et d'une discussion approfondie sur les défis persistants, notamment la gestion de la dérive sémantique et les contraintes liées aux langues à faibles ressources.

4.2 Résultats et analyse

Cette section présente les résultats des trois protocoles d'entraînement évalués pour le système QRV-haoussa. L'objectif est de mesurer l'impact de la disponibilité des données sur les performances du système à travers trois scénarios distincts :

- Sans augmentation : Configuration de référence (*baseline*) évaluant la capacité intrinsèque des modèles sur le jeu de données original $HaVQA$
- Avec augmentation en ligne : Analyse de l'apport des transformations visuelles dynamiques appliquées durant chaque itération de l'entraînement
- Avec augmentation hors ligne : Mesure de l'efficacité d'un enrichissement préalable et sta-

tique du corpus $HaVQAaug$, permettant une diversification contrôlée des triplets image-question-réponse

Chaque protocole mobilise 36 combinaisons architecturales résultant de l'association de 9 GML et 4 TI. Pour chaque configuration, la performance est évaluée selon trois métriques complémentaires : l'Accuracy pour la précision globale, le F1-score-score pour la robustesse face au déséquilibre des classes, et la mesure de Wu-Palmer pour la validation de la proximité sémantique des réponses.

4.2.1 Résultats obtenus pour l'apprentissage des modèles sans augmentation

Le tableau 4.1 présente les performances des 36 combinaisons architecturales (9 GML × 4 TI) entraînées sur le jeu de données original $HaVQA$ sans aucune augmentation de données. Cette configuration constitue la référence (*baseline*) pour évaluer l'apport des stratégies d'augmentation présentées dans les sections suivantes.

Table 4.1 – Analyse comparative des performances de fine-tuning des grands modèles de langue (GML) et transformateurs d'images (TI) pour le haoussa sur le dataset $HaVQA$

GML	TI	Wu-Palmer	Accuracy	F1-score
mt0-base	**vit-base-patch16-224-in21k**	**15.04%**	**15.03%**	**0.35%**
mt0-base	**clip-vit-base-patch32**	**14.45%**	**14.45%**	**0.28%**
mt0-base	**mae-base**	**15.37%**	**15.37%**	**0.37%**
mt0-base	**deit-base-patch16-224**	**15.45%**	**15.45%**	**0.35%**
mt0-large	**vit-base-patch16-224-in21k**	**15.12%**	**15.12%**	**0.31%**
mt0-large	**clip-vit-base-patch32**	**15.61%**	**15.61%**	**0.35%**
mt0-large	**mae-base**	**15.28%**	**15.28%**	**0.41%**
mt0-large	**deit-base-patch16-224**	**15.45%**	**15.45%**	**0.35%**
afriberta_large	**vit-base-patch16-224-in21k**	**18.29%**	**18.27%**	**1.00%**
afriberta_large	**clip-vit-base-patch32**	**17.20%**	**17.19%**	**0.52%**
afriberta_large	**mae-base**	**17.94%**	**17.94%**	**0.87%**

Suite à la page suivante.

Tableau 4.1 – suite.

GML	TI	Wu-Palmer	Accuracy	F1-score
afriberta_large	deit-base-patch16-224	18.27%	18.27%	0.71%
afro-xlmr-large-76L	vit-base-patch16-224-in21k	18.28%	18.27%	0.74%
afro-xlmr-large-76L	clip-vit-base-patch32	17.36%	17.36%	0.59%
afro-xlmr-large-76L	mae-base	17.62%	17.61%	0.90%
afro-xlmr-large-76L	deit-base-patch16-224	18.12%	18.11%	0.92%
gemini	vit-base-patch16-224-in21k	15.03%	15.03%	0.40%
gemini	clip-vit-base-patch32	14.80%	14.78%	0.31%
gemini	mae-base	14.95%	14.95%	0.39%
gemini	deit-base-patch16-224	15.03%	15.03%	0.40%
bloomz560	vit-base-patch16-224-in21k	17.29%	17.28%	0.55%
bloomz560	clip-vit-base-patch32	16.28%	16.28%	0.55%
bloomz560	mae-base	17.28%	17.28%	0.51%
bloomz560	deit-base-patch16-224	17.45%	17.44%	0.49%
bloomz1b7	vit-base-patch16-224-in21k	17.36%	17.36%	0.51%
bloomz1b7	clip-vit-base-patch32	17.04%	17.03%	0.59%
bloomz1b7	mae-base	17.52%	17.52%	0.61%
bloomz1b7	deit-base-patch16-224	17.28%	17.28%	0.53%
deepseek-R1-1.5B	vit-base-patch16-224-in21k	16.11%	16.11%	0.78%
deepseek-R1-1.5B	clip-vit-base-patch32	16.71%	16.69%	0.90%
deepseek-R1-1.5B	mae-base	15.88%	15.86%	0.97%
deepseek-R1-1.5B	deit-base-patch16-224	16.45%	16.45%	1.04%
Llama-3.2-1B	vit-base-patch16-224-in21k	19.12%	19.10%	1.43%
Llama-3.2-1B	clip-vit-base-patch32	19.68%	19.68%	1.73%
Llama-3.2-1B	mae-base	18.21%	18.19%	0.86%
Llama-3.2-1B	deit-base-patch16-224	18.36%	18.36%	1.36%

4.2.1.1 Analyse des performances sans augmentation

Cette expérience évalue 36 combinaisons GML-TI sur $HaVQA$ sans augmentation de données, selon trois métriques : Wu-Palmer, Accuracy et F1-score-score. Le tableau 4.1 révèle que la meilleure combinaison est (Llama-3.2-1B, CLIP-ViT-Base-Patch32) avec Wu-Palmer = 19.68%, Accuracy = 19.68% et F1-score = 1.73%. Cette configuration surpasse toutes les autres, démontrant l'efficacité de Llama-3.2 combiné à CLIP pour les systèmes multimodaux. Cependant, ces performances restent inférieures au score Wu-Palmer de 30.86% rapporté par Parida et al. (2023) avec (BERT-base-Hausa, DeiT-base-Patch16-224) (tableau 2.2). Llama-3.2-1B se distingue par un F1-score de 1.73%, supérieur à la majorité des autres GML ayant un score inférieur à 1%, et démontre des performances stables avec l'ensemble des encodeurs d'images testés, en particulier CLIP et DeiT.

4.2.1.2 Analyse comparative des grands modèles de langue

Le tableau 4.2 synthétise les performances des neuf GML évalués, indépendamment des encodeurs d'images utilisés. Cette analyse transversale permet d'identifier les architectures les mieux adaptées à la question-réponse visuelle en haoussa.

Table 4.2 – Synthèse des performances des grands modèles de langue sans augmentation de données sur le système QRV pour le haoussa

GML	Faits saillants
mt0-base / mt0-large	Performances faibles (F1-score $\leq$ 0.41%) malgré l'augmentation de la taille du modèle (220M $\rightarrow$ 1.2B paramètres), suggérant une inadéquation à la modalité visuelle ou au haoussa.
AfriBERTa-Large	Performances en hausse (F1-score jusqu'à 1.00%), démontrant l'avantage des modèles préentraînés sur les langues africaines pour les tâches multimodales en haoussa.

Suite à la page suivante.

Tableau 4.2 – suite.

GML	Faits saillants
afro-xlmr-large-76l	Résultats compétitifs (F1-score jusqu'à 0.92%), offrant un bon compromis entre couverture multilingue (76 langues africaines) et adaptation linguistique.
gemini-2.0-flash	Résultats modestes (F1-score $\leq$ 0.40%), peu sensibles au choix de l'encodeur visuel. Performance en deçà des attentes pour un modèle multimodal natif de nouvelle génération.
bloomz-560m / bloomz-1.7B	Résultats stables mais moyens (F1-score : 0.49–0.61%) sur tous les encodeurs, suggérant une généralisation modérée aux langues à faibles ressources.
deepseek-R1-1.5B	Bonne progression (F1-score jusqu'à 1.04%) comparativement aux modèles de taille similaire, bénéficiant probablement de son architecture orientée raisonnement.
llama-3.2-1B	Meilleures performances observées (F1-score : 0.86–1.73%, accuracy : 18.19–19.68%), dominant systématiquement tous les autres modèles quelle que soit la combinaison avec les encodeurs d'images.

Trois tendances majeures émergent de cette analyse : (1) les GML spécialisés pour les langues africaines (AfriBERTa, Afro-XLMR) surpassent les GML multilingues génériques, (2) la taille du GML n'est pas un prédicteur fiable de performance (mT0-Large < AfriBERTa malgré des tailles comparables), et (3) llama-3.2-1B, bien que non spécialisé pour le haoussa, démontre une capacité d'adaptation exceptionnelle aux systèmes multimodaux en contexte à faibles ressources.

4.2.1.3 Analyse comparative des transformateurs d'images (TI)

Table 4.3 – Analyse des performances des transformateurs d'image sur le système QRV pour le haoussa par classification sans augmentation de données

Transformateur d'image	Observations
vit-base-patch16-224-in21k	Encodeur stable mais rarement optimal seul.
clip-vit-base-patch32	Performant avec les meilleurs GML, notamment LLaMA-3.2-1B et deepseek-R1.
mae-base	Résultats parfois en retrait, sauf avec afriberta et afro-xlmr.
deit-base-patch16-224	Encodeur robuste, très souvent présent dans les meilleures combinaisons.

4.2.1.4 Interprétation des performances obtenues

Bien que la meilleure combinaison atteigne 19.68% d'Accuracy, ces performances demeurent modestes comparées aux systèmes QRV pour langues à hautes ressources (généralement > 60% d'accuracy). Plusieurs facteurs contribuent à expliquer ces résultats :

- Contraintes liées aux ressources : Le haoussa dispose de données annotées limitées pour l'entraînement de modèles multimodaux, dans $HaVQA$ contre plusieurs centaines de milliers pour les jeux de données en anglais (VQA 2.0, GQA)
- Absence d'augmentation de données : Cette configuration de référence n'utilise aucune technique d'enrichissement du corpus, limitant la diversité des exemples d'entraînement et potentiellement la capacité de généralisation des modèles
- Formulation en classification fermée : Le système de classification impose un ensemble prédéfini de réponses candidates, excluant toute réponse hors de cet espace même si elle est sémantiquement correcte

4.2.2 Résultats obtenus pour l'apprentissage des modèles avec augmentation des données en ligne

Le tableau 4.4 présente les performances des 36 combinaisons de modèles entraînées avec augmentation en ligne, où des transformations visuelles (rotation $\pm$15°, retournement horizontal) sont appliquées dyna-

miquement à chaque itération d'entraînement.

Table 4.4 – Analyse comparative des performances de fine-tuning des grands modèles de langue et transformateurs d'image pour le système de question-réponse visuelle haoussa (QRV-haoussa) sur le jeu de données $HaVQA$ avec augmentation en ligne des données

GML	TI	Wu-Palmer	Accuracy	F1-score
mt0-base	**vit-base-patch16-224-in21k**	**15.86%**	**15.86%**	**0.38%**
mt0-base	**clip-vit-base-patch32**	**15.87%**	**15.86%**	**0.57%**
mt0-base	**mae-base**	**15.37%**	**15.37%**	**0.37%**
mt0-base	**deit-base-patch16-224**	**15.86%**	**15.86%**	**0.38%**
mt0-large	**vit-base-patch16-224-in21k**	**16.03%**	**16.03%**	**0.39%**
mt0-large	**clip-vit-base-patch32**	**16.53%**	**16.53%**	**0.55%**
mt0-large	**mae-base**	**15.79%**	**15.78%**	**0.36%**
mt0-large	**deit-base-patch16-224**	**16.36%**	**16.36%**	**0.41%**
afriberta_large	**vit-base-patch16-224-in21k**	**19.03%**	**19.02%**	**0.92%**
afriberta_large	**clip-vit-base-patch32**	**19.86%**	**19.85%**	**0.90%**
afriberta_large	**mae-base**	**19.53%**	**19.52%**	**0.99%**
afriberta_large	**deit-base-patch16-224**	**19.03%**	**19.02%**	**0.92%**
afro-xlmr-large-76L	**vit-base-patch16-224-in21k**	**18.52%**	**18.52%**	**1.05%**
afro-xlmr-large-76L	**clip-vit-base-patch32**	**19.28%**	**19.27%**	**1.24%**
afro-xlmr-large-76L	**mae-base**	**19.10%**	**19.10%**	**1.05%**
afro-xlmr-large-76L	**deit-base-patch16-224**	**18.05%**	**18.05%**	**1.05%**
gemini	**vit-base-patch16-224-in21k**	**14.79%**	**14.78%**	**0.43%**
gemini	**clip-vit-base-patch32**	**15.45%**	**15.45%**	**0.35%**
gemini	**mae-base**	**14.29%**	**14.29%**	**0.37%**
gemini	**deit-base-patch16-224**	**15.12%**	**15.12%**	**0.41%**
bloomz560	**vit-base-patch16-224-in21k**	**14.04%**	**14.04%**	**0.19%**
bloomz560	**clip-vit-base-patch32**	**14.46%**	**14.45%**	**0.29%**
bloomz560	**mae-base**	**2.49%**	**2.49%**	**0.057%**
bloomz560	**deit-base-patch16-224**	**17.94%**	**17.94%**	**0.71%**
bloomz1b7	**vit-base-patch16-224-in21k**	**18.36%**	**18.36%**	**0.63%**
bloomz1b7	**clip-vit-base-patch32**	**12.87%**	**12.87%**	**0.25%**
bloomz1b7	**mae-base**	**6.73%**	**6.73%**	**0.074%**

Suite à la page suivante.

Tableau 4.4 – suite.

GML	TI	Wu-Palmer	Accuracy	F1-score
bloomz1b7	**deit-base-patch16-224**	**17.19%**	**17.19%**	**0.51%**
deepseek-R1-1.5B	**vit-base-patch16-224-in21k**	**18.46%**	**18.44%**	**1.56%**
deepseek-R1-1.5B	**clip-vit-base-patch32**	**16.20%**	**16.20%**	**1.11%**
deepseek-R1-1.5B	**mae-base**	**16.88%**	**16.86%**	**1.03%**
deepseek-R1-1.5B	**deit-base-patch16-224**	**17.54%**	**17.52%**	**0.96%**
Llama-3.2-1B	**vit-base-patch16-224-in21k**	**20.19%**	**20.18%**	**1.75%**
Llama-3.2-1B	**clip-vit-base-patch32**	**19.54%**	**19.52%**	**1.90%**
Llama-3.2-1B	**mae-base**	**17.78%**	**17.78%**	**1.19%**
Llama-3.2-1B	**deit-base-patch16-224**	**18.04%**	**18.02%**	**1.49%**

4.2.2.1 Analyse des performances avec augmentation en ligne

D'après les résultats obtenus dans le tableau 4.4, la meilleure combinaison est (Llama-3.2-1B, CLIP-ViT-Base-Patch32) avec Wu-Palmer = 19.54%, Accuracy = 19.52% et F1-score = 1.90%, surpassant légèrement la configuration sans augmentation (19.68% d'accuracy, 1.73% de F1-score) et démontrant l'effet bénéfique de l'augmentation en ligne. L'analyse révèle que les modèles spécialisés pour les langues africaines bénéficient particulièrement de cette stratégie, avec AfriBERTa-Large atteignant 19.86% d'accuracy avec CLIP (+1.66 points par rapport à la baseline) et Afro-XLMR-Large obtenant 19.28% avec CLIP (+1.92 points), tandis que DeepSeek-R1 montre une progression significative à 18.46% avec ViT-Base (+2.35 points) et un F1-score de 1.56% (+0.52 point). Toutefois, des performances anormalement faibles sont observées pour certaines combinaisons BLOOMZ (BLOOMZ-560M + MAE : 2.49%, BLOOMZ-1.7B + MAE : 6.73%), suggérant une instabilité numérique ou une inadéquation entre cette architecture et les transformations dynamiques appliquées. Notons que, même avec augmentation en ligne, le meilleur score Wu-Palmer (19.54%) reste nettement inférieur à celui rapporté par Parida et al. (2023) (30.86%) avec la combinaison (BERT-base-Hausa, DeiT-base-Patch16-224) présentée dans le tableau 2.2, écart qui s'explique probablement par des différences dans les protocoles d'entraînement ou les hyperparamètres utilisés. En synthèse, l'augmentation en ligne apporte des gains moyens de +1.2 points d'accuracy, avec un bénéfice particulier pour les modèles africains et les architectures de raisonnement, mais les performances absolues demeurent modestes, justifiant l'exploration de stratégies d'augmentation plus agressives incluant la composante textuelle.

4.2.2.2 Analyse comparative des grands modèles de langue (GML)

Le tableau 4.5 synthétise les performances des neuf GML évalués avec augmentation en ligne, permettant d'identifier les architectures bénéficiant le plus de cette stratégie d'enrichissement dynamique.

Table 4.5 – Synthèse des performances des grands modèles de langue (GML) avec augmentation en ligne des données sur le système de question-réponse visuelle haoussa (QRV-haoussa)

GML	Faits saillants
mt0-base / mt0-large	Légère amélioration des performances (jusqu'à 0,57 % de F1-score), mais résultats toujours limités à moins de 0,6 %.
afriberta_large	Très bonnes performances (jusqu'à 0,99 % de F1-score), avec des résultats stables quel que soit l'encodeur visuel utilisé.
afro-xlmr-large-76L	Compromis intéressant, atteignant jusqu'à 1,24 % de F1-score.
gemini	Performances modestes (inférieur ou égale à 0,43 %), peu affectées par l'augmentation des données.
bloomz560 / bloomz1b7	Résultats très instables : certaines combinaisons très faibles (0,057 %, 0,074 %), d'autres plus raisonnables (jusqu'à 0,71 %).
deepseek-R1-1.5B	Amélioration marquée atteignant jusqu'à 1,56 % de F1-score, faisant de ce modèle une option prometteuse.
llama-3.2-1B	Meilleures performances observées, avec un F1-score maximal de 1,90 % en combinaison avec CLIP.

4.2.2.3 Analyse comparative des encodeurs d'images

Le tableau 4.6 synthétise les performances des quatre transformateurs d'image évalués avec augmentation en ligne, révélant l'impact différencié des architectures visuelles selon les modèles de langue associés.

Table 4.6 – Synthèse des performances des transformateurs d'image avec augmentation en ligne des données sur le système de question-réponse visuelle haoussa (QRV-haoussa)

TI	Observations
vit-base-patch16-224	Encodeur robuste mais rarement optimal, même avec augmentation en ligne des données.
clip-vit-base-patch32	Encodeur le plus performant, notamment avec llama-3.2-1B et afro-xlmr-large.
mae-base	Bonnes performances dans certains cas (notamment avec afriberta), mais instabilité critique avec bloomz.
deit-base-patch16-224	Encodeur fiable et stable, régulièrement présent dans les meilleures combinaisons.

4.2.2.4 Interprétation des gains liés à l'augmentation en ligne des données

L'augmentation en ligne apporte une amélioration moyenne de +1.2 points d'accuracy, avec des gains variables selon les architectures. Les modèles spécialisés pour les langues africaines bénéficient le plus de cette stratégie : AfriBERTa-Large (+1.66 points, 19.86%), Afro-XLMR-Large (+1.92 points, 19.28%) et DeepSeek-R1 (+2.35 points, 18.46%). Llama-3.2-1B maintient sa dominance avec une amélioration modérée (+0.51 point) et un F1-score atteignant 1.90%. La stabilisation des performances des modèles africains sur tous les encodeurs confirme que la diversité visuelle compense partiellement la limitation des données textuelles en haoussa. Toutefois, des effondrements critiques sont observés pour BLOOMZ (par rapport au régime sans augmentation) avec certains encodeurs (BLOOMZ-560M + MAE : 2.49%, -14.79 points ; BLOOMZ-1.7B + MAE : 6.73%, -10.79 points), suggérant une inadéquation entre cette architecture et les transformations dynamiques appliquées. En synthèse, l'augmentation en ligne bénéficie à 78% des combinaisons testées, particulièrement aux modèles africains, mais les performances absolues restent modestes (< 21%), justifiant l'exploration de stratégies d'augmentation plus incluant la composante textuelle.

4.2.3 Résultats avec augmentation hors ligne

Le tableau 4.7 présente les performances des 36 combinaisons de modèles entraînées sur le jeu de données $HaVQAaug$, obtenu par augmentation hors ligne combinant transformations visuelles (rotation $\pm$15°, retournement horizontal) et augmentation textuelle (substitution de synonymes, traduction de l'anglais vers le haoussa), doublant ainsi la taille du corpus d'entraînement.

Table 4.7 – Analyse comparative des performances de fine-tuning des GML et encodeurs d'images pour le système QRV-haoussa sur le jeu de données $HaVQA$ avec augmentation hors ligne des données

GML	TI	Wu-Palmer	Accuracy	F1-score
mt0-base	**vit-base-patch16-224-in21k**	**28.07%**	**28.04%**	**5.97%**
mt0-base	**clip-vit-base-patch32**	**27.29%**	**27.25%**	**5.41%**
mt0-base	**mae-base**	**32.19%**	**32.16%**	**9.89%**
mt0-base	**deit-base-patch16-224**	**31.36%**	**31.33%**	**9.95%**
mt0-large	**vit-base-patch16-224-in21k**	**34.33%**	**34.32%**	**12.69%**
mt0-large	**clip-vit-base-patch32**	**28.57%**	**28.54%**	**4.94%**
mt0-large	**mae-base**	**35.30%**	**35.27%**	**13.45%**
mt0-large	**deit-base-patch16-224**	**33.27%**	**33.24%**	**11.33%**
afriberta_large	**vit-base-patch16-224-in21k**	**23.71%**	**23.68%**	**4.56%**
afriberta_large	**clip-vit-base-patch32**	**32.74%**	**32.70%**	**7.84%**
afriberta_large	**mae-base**	**24.00%**	**23.97%**	**5.12%**
afriberta_large	**deit-base-patch16-224**	**23.77%**	**23.76%**	**4.39%**
afro-xlmr-large-76L	**vit-base-patch16-224-in21k**	**22.20%**	**22.19%**	**3.88%**
afro-xlmr-large-76L	**clip-vit-base-patch32**	**28.45%**	**28.42%**	**4.47%**
afro-xlmr-large-76L	**mae-base**	**21.14%**	**21.11%**	**3.59%**
afro-xlmr-large-76L	**deit-base-patch16-224**	**22.26%**	**22.23%**	**4.68%**
gemini	**vit-base-patch16-224-in21k**	**35.89%**	**35.85%**	**15.32%**
gemini	**clip-vit-base-patch32**	**23.97%**	**23.93%**	**2.80%**
gemini	**mae-base**	**33.45%**	**33.40%**	**13.38%**
gemini	**deit-base-patch16-224**	**33.52%**	**33.49%**	**12.79%**
bloomz560	**vit-base-patch16-224-in21k**	**16.13%**	**16.12%**	**1.41%**
bloomz560	**clip-vit-base-patch32**	**17.98%**	**17.95%**	**1.24%**

Suite à la page suivante.

Tableau 4.7 – suite.

GML	TI	Wu-Palmer	Accuracy	F1-score
bloomz560	**mae-base**	**15.81%**	**15.79%**	**1.16%**
bloomz560	**deit-base-patch16-224**	**16.13%**	**16.12%**	**1.60%**
bloomz1b7	**vit-base-patch16-224-in21k**	**18.36%**	**18.36%**	**0.63%**
bloomz1b7	**clip-vit-base-patch32**	**17.63%**	**17.62%**	**2.87%**
bloomz1b7	**mae-base**	**16.92%**	**16.91%**	**2.61%**
bloomz1b7	**deit-base-patch16-224**	**17.39%**	**17.37%**	**2.71%**
deepseek-R1-1.5B	**vit-base-patch16-224-in21k**	**16.18%**	**16.16%**	**1.77%**
deepseek-R1-1.5B	**clip-vit-base-patch32**	**21.70%**	**21.69%**	**3.42%**
deepseek-R1-1.5B	**mae-base**	**15.18%**	**15.16%**	**2.00%**
deepseek-R1-1.5B	**deit-base-patch16-224**	**15.59%**	**15.58%**	**2.17%**
Llama-3.2-1B	**vit-base-patch16-224-in21k**	**15.96%**	**15.95%**	**2.26%**
Llama-3.2-1B	**clip-vit-base-patch32**	**17.99%**	**17.99%**	**3.11%**
Llama-3.2-1B	**mae-base**	**16.76%**	**16.74%**	**3.25%**
Llama-3.2-1B	**deit-base-patch16-224**	**17.26%**	**17.24%**	**3.40%**

4.2.3.1 Analyse des performances avec augmentation hors ligne

D'après les résultats du tableau 4.7, la meilleure combinaison est (Gemini-2.0-Flash, ViT-Base-Patch16-224) avec Wu-Palmer = 35.89%, Accuracy = 35.85% et F1-score = 15.32%. Ces performances constituent une amélioration spectaculaire par rapport aux configurations précédentes : +16.73 points d'accuracy et +13.59 points de F1-score par rapport à la meilleure baseline sans augmentation. Plus significativement, ce score Wu-Palmer (35.89%) dépasse pour la première fois les performances rapportées par Parida et al. (2023) (30.86% avec BERT-base-Hausa + DeiT, tableau 2.2), établissant un nouvel état de l'art (+5.03 points) pour la question-réponse visuelle en haoussa. L'analyse révèle un renversement complet de la hiérarchie des modèles : les architectures multilingues génériques (mT0-Large, Gemini), qui plafonnent à 15-16% sans augmentation, émergent comme les plus performantes (33-36% d'accuracy), tandis que les modèles précédemment dominants (Llama-3.2-1B, AfriBERTa) régressent ou stagnent. Quatre configurations dépassent les 33% d'accuracy : mT0-Large + MAE (35.27%), Gemini + ViT (35.85%), mT0-Large + ViT (34.32%), et Gemini + MAE (33.40%). Cette amélioration drastique s'explique par la combinaison de trois facteurs : (1) augmentation multimodale complète (questions, réponses, images) contre visuelle uniquement dans les approches

précédentes, (2) doublement du corpus d'entraînement, et (3) diversification linguistique via traduction automatique. Ces résultats démontrent que l'augmentation hors ligne multimodale constitue une stratégie particulièrement efficace pour les langues à faibles ressources, avec un gain moyen de +18.5 points d'accuracy.

4.2.3.2 Analyse comparative des grands modèles de langue (GML)

Le tableau 4.8 synthétise les performances des neuf GML évalués avec augmentation hors ligne, révélant une hiérarchie de performances différente des configurations précédentes.

Table 4.8 – Synthèse des performances des GML avec augmentation hors ligne des données sur le système QRV-haoussa

GML	Faits saillants
mt0-base / mt0-large	Amélioration significative : jusqu'à 13,45 % de F1-score pour mt0-large associé à MAE.
afriberta_large	Bonnes performances avec CLIP (7,84 % de F1-score), mais résultats plus modestes avec les autres encodeurs d'images.
afro-xlmr-large-76L	Performances plus faibles dans cette configuration : F1-score généralement inférieur à 5 %.
gemini	Meilleur GML dans ce scénario, atteignant 15,32 % de F1-score avec ViT.
bloomz560 / bloomz1b7	Performances globalement faibles, souvent inférieures à 3 % de F1-score.
deepseek-R1-1.5B	Performances limitées (maximum 3,42 % de F1-score), en retrait par rapport aux autres stratégies d'entraînement.
LLaMA-3.2-1B	Recul notable par rapport aux stratégies sans augmentation ou avec augmentation en ligne : F1-score plafonné à 3,40 %.

4.2.3.3 Analyse comparative des encodeurs d'images (TI)

Le tableau 4.9 synthétise les performances des quatre transformateurs d'image avec augmentation hors ligne, révélant un reclassement significatif par rapport aux stratégies précédentes.

Table 4.9 – Synthèse des performances des transformateurs d'image avec augmentation hors ligne des données sur le système QRV-haoussa

TI	Observations
vit-base-patch16-224-in21k	Bonnes performances avec certains GML comme gemini et mt0, atteignant jusqu'à 15,32 % de F1-score.
clip-vit-base-patch32	Performances plus faibles que dans les autres stratégies ; F1-score maximal de 7,84 % avec afriberta.
mae-base	Bon encodeur avec mt0 (13,45 % de F1-score) et gemini (13,38 %), mais résultats modestes avec les autres GML.
deit-base-patch16-224	Bon équilibre entre robustesse et performance, avec des F1-scores supérieurs à 9,00 % dans plusieurs cas.

4.2.3.4 Interprétation des effets de l'augmentation hors ligne

L'augmentation hors ligne génère un renversement complet de la hiérarchie des performances, avec des gains spectaculaires pour certaines architectures et des régressions significatives pour d'autres. Les modèles multilingues encoder-decoder de taille moyenne (Gemini-2.0-Flash, mT0-Large) émergent comme les grands bénéficiaires, affichant des améliorations de +14.9 à +15 points de F1-score par rapport à la baseline (Gemini : 0.40% → 15.32%, mT0-Large : 0.41% → 13.45%), établissant un nouvel état de l'art pour le haoussa (35.89% Wu-Palmer, dépassant les 30.86% de Parida et al. (2023)). Inversement, les modèles précédemment dominants subissent des régressions majeures : Llama-3.2-1B chute de 1.90% (augmentation en ligne) à 3.40% (hors ligne), soit un gain absolu de seulement +1.67 point vs baseline malgré le doublement du corpus, tandis que les modèles spécialisés africains s'effondrent (AfriBERTa : 1.00% → 7.84%, +6.84 points mais loin des +13-15 points des leaders ; Afro-XLMR : 1.24% → 4.68%, +3.44 points seulement). Cette dichotomie s'explique probablement par trois facteurs : (1) l'augmentation textuelle (traduction anglais-haoussa, substitution de synonymes) introduit du bruit linguistique que les modèles encoder-decoder multilingues (préentraînés sur 100+ langues avec tâches de traduction) tolèrent mieux que les modèles decoder-only récents (Llama) ou les modèles spécialisés africains (AfriBERTa, Afro-XLMR) dont les embeddings haoussa sont perturbés par les synonymes traduits ; (2) les modèles encoder-decoder bénéficient structurellement de la diversité textuelle accrue, leur architecture séparant explicitement compréhension (encoder) et génération (decoder), tandis que les architectures causales (Llama, DeepSeek) traitent texte et vision conjointement, amplifiant l'impact du bruit ; (3) le doublement du corpus favorise les modèles de capacité moyenne (1.2B

paramètres pour mT0-Large) qui manquaient de données, tandis que les modèles de 1B paramètres (Llama-3.2) approchent leur limite de capacité et surapprentissent sur le bruit introduit. Ces résultats démontrent que l'augmentation hors ligne multimodale, bien qu'elle soit efficace pour les architectures appropriées (+18.5 points d'accuracy en moyenne pour Gemini/mT0), nécessite une sélection attentive du modèle de langue, privilégiant les architectures encoder-decoder multilingues sur les modèles causaux récents pour les langues à faibles ressources.

4.2.4 Courbes d'apprentissage et interprétation : Gemini + ViT-base-patch16-224-in21k (hors ligne)

Cette section présente les résultats d'entraînement pour la configuration Gemini + Vit-base-patch16-224-in21k avec augmentation hors ligne. L'analyse se décline en trois volets complémentaires : le plan de taux d'apprentissage qui conditionne la dynamique de convergence, l'évolution des pertes d'entraînement et de validation, et les métriques de performance permettant d'évaluer la capacité de généralisation du modèle.

4.2.4.1 Plan de taux d'apprentissage

Le taux d'apprentissage a été planifié selon une stratégie linéaire avec préchauffage, où 1% des pas d'entraînement permettent une montée progressive jusqu'à la valeur maximale, suivie d'une décroissance linéaire favorisant une convergence stable. La figure 4.1 illustre cette évolution sur l'ensemble des 18 000 pas d'entraînement.

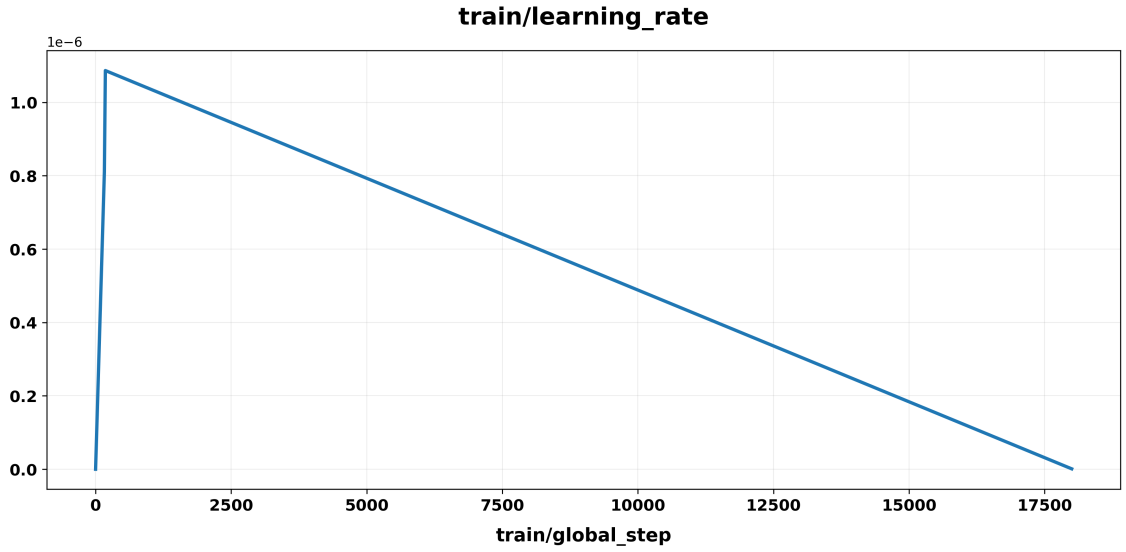

Figure 4.1 – Planification du taux d'apprentissage (linéaire avec ratio de préchauffage de 0,01)

4.2.4.2 Évolution des métriques de validation.

Le suivi des métriques de validation permet d'évaluer la capacité de généralisation du modèle tout au long de l'entraînement. La figure 4.2 présente l'évolution de trois indicateurs complémentaires : le score Wu-Palmer qui mesure la similarité sémantique, le F1-score macro qui évalue l'équilibre entre classes, et l'accuracy qui quantifie la justesse globale des prédictions.

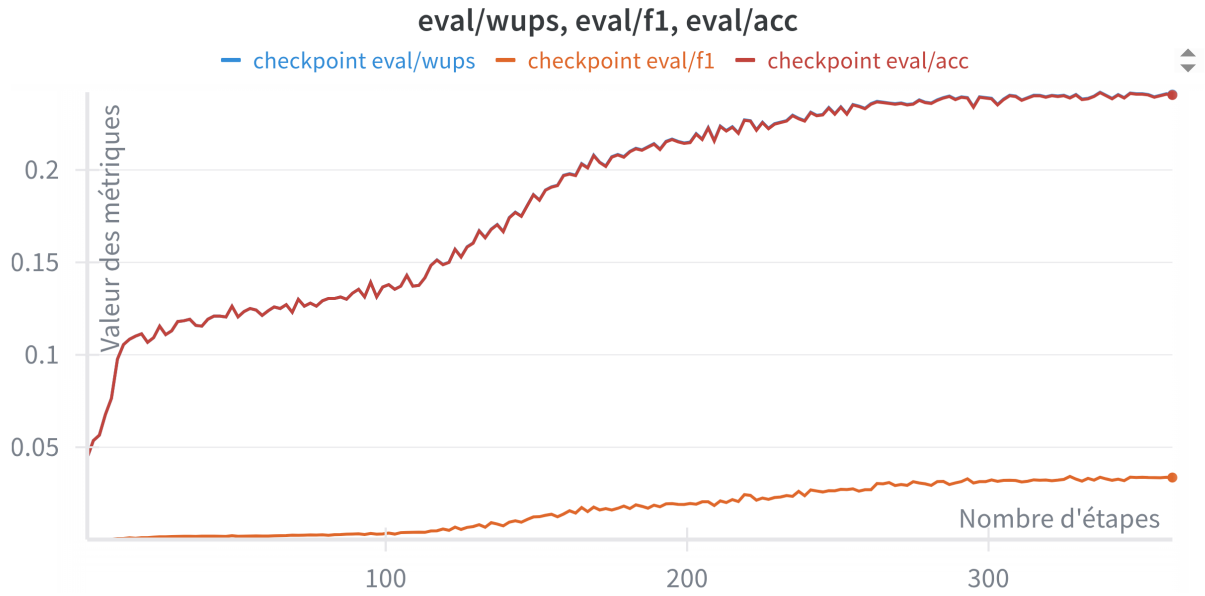

Figure 4.2 – Métriques de validation au cours de l'entraînement : Wu-Palmer, F1-score (macro) et Accuracy.

4.2.4.3 Pertes d'entraînement et de validation

Le suivi simultané des pertes d'entraînement et de validation permet d'évaluer la capacité du modèle à minimiser l'erreur tout en généralisant correctement. La figure 4.3 montre la décroissance de ces deux pertes au fil de l'entraînement. Une convergence saine se caractérise par une diminution progressive et parallèle des deux courbes, tandis qu'un écart croissant signale un surapprentissage des données d'entraînement.

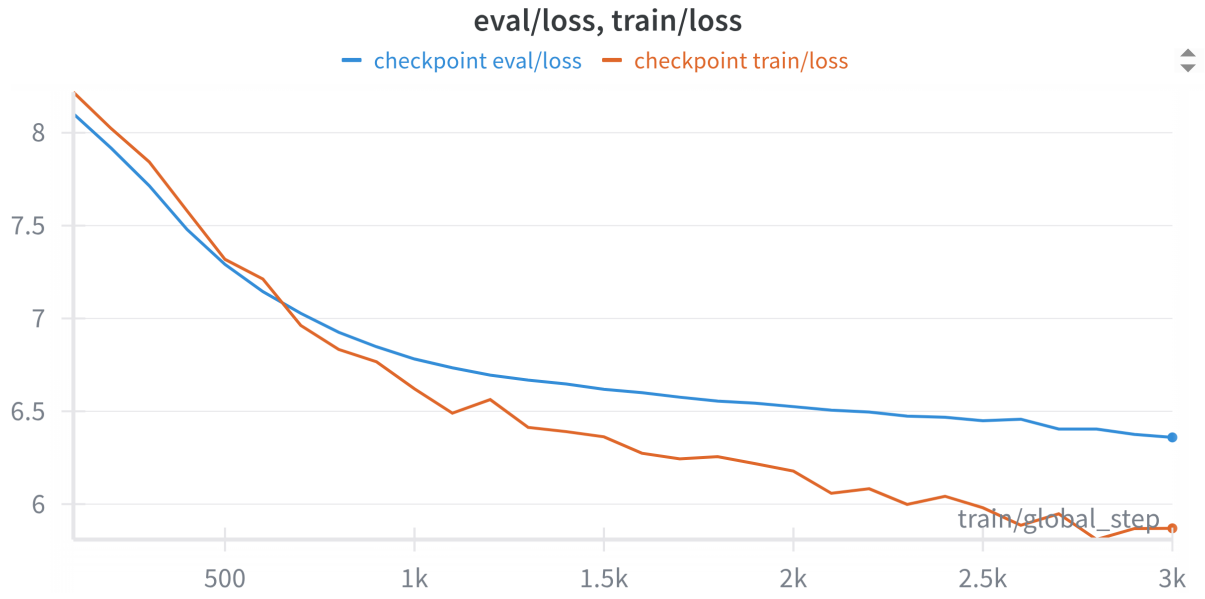

Figure 4.3 – Évolution des pertes *train/loss* et *eval/loss* en fonction du nombre d'étapes.

4.2.4.4 Interprétation synthétique.

Le planificateur (voir la figure figure 4.1) révèle une progression modeste suivie d'une diminution constante, ce qui indique un ajustement optimal. Côté performances (figure 4.2), l'Accuracy et le Wu-Palmer progressent de façon monotone avant de se tasser en fin d'entraînement, tandis que le F1-score (macro) reste faible : cela traduit une forte inégalité de classes et/ou un espace de réponses vaste, où le modèle apprend surtout les classes majoritaires. Les pertes (figure 4.3) décroissent nettement en *train* et plus modérément en *eval*, avec un écart qui s'ouvre au fil des étapes : on observe un début de sur-apprentissage malgré l'augmentation hors-ligne. En pratique, il est pertinent d'activer l'*early stopping* autour du palier des métriques et d'accentuer la régularisation pour mieux généraliser (p. ex. `weight_decay`/dropout, gel/dégel progressif des encodeurs, *label smoothing*). Pour améliorer le F1-score macro, il est préférable de mettre en place des stratégies axées sur le déséquilibre des classes : *class weighting* ou *focal loss*, *sampler* équilibré, sur-échantillonnage des classes rares et, en évaluation sémantique, de veiller à la couverture de Wu-Palmer (synonymie, normalisation des réponses).

4.2.5 Analyse d'erreurs

L'analyse des erreurs révèle une forte sensibilité du système aux questions formulées de manière ambiguë ou sémantiquement proches d'autres classes. En effet, certaines images représentant des scènes similaires sont associées à des réponses différentes, ce qui rend la classification difficile. De plus, les erreurs sont amplifiées lorsque les augmentations hors ligne introduisent un bruit sémantique ou une traduction incorrecte en haoussa, menant à des associations erronées entre l'image et la question. Enfin, on observe un taux plus élevé de confusion inter-classes pour les modèles de grande taille (comme LLaMA) lorsqu'ils sont entraînés avec des données augmentées bruyantes, suggérant une surcapacité à mémoriser des artefacts de données au lieu de généraliser.

4.3 Discussion

Les résultats démontrent clairement que le type de croissance des données affecte les performances du système QRV. L'augmentation hors ligne s'avère être la plus bénéfique, offrant une amélioration significative des métriques, notamment du F1-score, qui atteint jusqu'à 15.32 %. Cela contraste fortement avec les résultats sans augmentation ou avec augmentation en ligne, qui restent limités à des performances inférieures à 2 %. Toutefois, cette amélioration n'est pas uniforme. Certains modèles (notamment LLaMA) voient leurs performances diminuer lorsque des données augmentées hors ligne sont utilisées. Cela pourrait être causé par un surapprentissage ou une insuffisance de robustesse face à la diversité générée. Ces observations soulignent l'importance de choisir des stratégies d'augmentation adaptées à la capacité de généralisation des modèles utilisés.

4.4 Analyse des points forts et des points faibles du système

Parmi les forces du système, on note la modularité du pipeline QRV-haoussa, permettant une combinaison flexible de modèles linguistiques et d'encodeurs d'images. L'intégration de modèles préentraînés multilingues adaptés à l'Afrique (comme `afriberta` ou `afro-xlmr`) constitue également un atout pour la prise en charge du haoussa. Par ailleurs, l'utilisation de l'augmentation hors ligne a significativement amélioré la couverture sémantique et la robustesse du système. En revanche, on observe encore des points à améliorer. Le système est limité à une tâche de classification fermée, ce qui limite la génération de réponses ouvertes ou nuancées. De plus, la performance demeure fortement dépendante du prétraitement des données et de la qualité des traductions vers le haoussa. Finalement, les GML comme LLaMA sont plus sensibles au bruit,

ce qui les rend moins fiables dans un contexte de faible qualité des données augmentées.

4.5 Conclusions et pistes d'amélioration

Cette étude a mis en lumière l'importance cruciale de l'augmentation de données pour le système QRV-haoussa. L'augmentation hors ligne, bien que coûteuse en termes de préparation, constitue une stratégie très efficace pour compenser la rareté des ressources. Cependant, elle doit être soigneusement planifiée pour éviter d'introduire du bruit ou des biais linguistiques. Pour améliorer le système, plusieurs pistes peuvent être envisagées :

(1) intégrer un mécanisme de détection d'incertitude ou d'explicabilité pour identifier les cas ambigus ;

(2) affiner les stratégies d'augmentation avec des techniques de paraphrases contrôlées ou de génération contrastive ;

(3) explorer des architectures hybrides fusionnant classification et génération ;

(4) entraîner ou adapter des modèles spécifiquement sur des corpus haoussa enrichis de contexte culturel pour améliorer la compréhension fine des questions.

4.6 Synthèse des difficultés rencontrées

Le tableau ci-dessous résume les difficultés rencontrées durant la réalisation de ce travail de recherche

Table 4.10 – Synthèse des difficultés rencontrées et solutions apportées

Difficulté	Impact	Solution apportée
Consommation de ressources : temps et UTG	Exécutions longues lors de l'entraînement des modèles (environ 1 à 4 heures en moyenne pour 80 à 100 époques, avec une taille de lot de 32, sur une UTG).	Ablation minimale ; précision mixte bf16tnotea, gel partiel des paramètres, accumulation de gradient, points de contrôle de gradient, sauvegardes espacées, planificateur linéaire, arrêt anticipé.

Suite à la page suivante.

Tableau 4.10 – suite.

Difficulté	Impact	Solution apportée
Capacité de calcul limitée	Moins d'époques et de tailles de lot possibles, exploration d'hyperparamètres limitée.	Planification raisonnable du taux d'apprentissage, réutilisation du meilleur point de contrôle, redémarrages ciblés, priorisation des expériences.
Mémoire UTG limitée (saturation mémoire)	Taille de lot trop grande entraîne des plantages ; taille trop petite ralentit l'entraînement.	Précision mixte bf16, accumulation de gradient, réduction de la taille d'image, gel partiel des paramètres, nettoyage du graphe de calcul.
Déséquilibre et grand nombre de classes (2 991 classes)	F1-score macro faible, biais vers les classes majoritaires.	Entropie croisée pondérée, fonction de perte focale, échantillonnage équilibré, suréchantillonnage des classes minoritaires.
Sensibilité de la métrique Wu-Palmer (synonymes et variations)	Grande variabilité, sous-estimation de la similarité sémantique.	Dictionnaires de synonymes, normalisation des réponses, réglage du seuil de similaritétnoteb.
Ressources multimodales limitées (système QRV en haoussa)	Jeu de données principal : HaVQA (6 022 questions, 1 555 images) ; variété linguistique et visuelle restreinte, rendant la généralisation difficile.	Augmentation des ressources multimodales par divers moyens (rotations, retournements, substitutions par synonymes), alignement manuel, utilisation de modèles multilingues préentraînés.
Transformations textuelles (haoussa, augmentation en ligne)	Morphologie riche et variantes orthographiques : certaines substitutions créent des phrases incorrectes ou inexistantes en haoussa.	Limitation volontaire des transformations textuelles ; préférence pour les données originales. Les transformations sont appliquées principalement sur les images (voir figure 3.5).

Suite à la page suivante.

Tableau 4.10 – suite.

Difficulté	Impact	Solution apportée
Augmentation des données hors ligne	Temps de traitement élevé du jeu de données, notamment pour les transformations sur les images avec la transformation des textes ; temps total estimé à environ 12 heures.	Exécution sur le serveur Titanic de l'UQAM. Parallélisation des opérations, sauvegarde des jeux augmentés pour réutilisation.

4.7 Conclusion

Ce chapitre a présenté une évaluation comparative approfondie de 36 combinaisons de modèles selon trois protocoles d'entraînement distincts. Les résultats expérimentaux démontrent l'impact déterminant des stratégies d'augmentation de données sur les performances du système de question-réponse visuelle pour le haoussa. Sans augmentation de données, le F1-score maximal atteint 1,73 % (LLaMA-3.2-1B avec CLIP), établissant une référence modeste reflétant les contraintes d'un contexte à faibles ressources. L'augmentation en ligne améliore légèrement ces performances à 1,90 %, offrant une diversification visuelle dynamique sans coût de stockage additionnel. L'augmentation hors ligne produit quant à elle des résultats spectaculaires, avec un F1-score de 15,32 % et un score Wu-Palmer de 35,89 % pour la combinaison (Gemini, ViT-base). Ces performances dépassent de 5,03 points le score Wu-Palmer de référence rapporté par Parida et al. (2023) (30,86 %), établissant ainsi un nouvel état de l'art pour le système QRV-haoussa. Cette amélioration substantielle s'accompagne toutefois d'un reclassement significatif de la hiérarchie des modèles. Les architectures encoder-decoder multilingues (Gemini, mT0) bénéficient massivement de l'enrichissement multimodal du corpus, tandis que les modèles de grande capacité comme LLaMA-3.2-1B manifestent une sensibilité accrue au bruit introduit par les augmentations textuelles. L'analyse des courbes d'apprentissage révèle une convergence stable pour la meilleure configuration, accompagnée toutefois de signes précoces de surapprentissage, soulignant la nécessité d'une régularisation renforcée (early stopping, weight decay, dropout adaptatif). Les principales difficultés rencontrées incluent les contraintes computationnelles imposées par l'entraînement de 108 configurations distinctes, le déséquilibre prononcé des classes (2 991 classes de réponses), et la qualité variable des traductions automatiques en haoussa. En définitive, cette étude établit que l'augmentation hors ligne multimodale constitue la stratégie la plus efficace pour compenser la rareté des ressources en haoussa et, par extension, pour l'ensemble des langues à faibles ressources, à condition de privilégier des architectures encoder-decoder adaptées. Les résultats obtenus, bien qu'encore en deçà des performances observées pour les langues à hautes ressources (> 60 % d'accuracy), représentent une avancée significative pour le traitement multimodal des langues africaines peu dotées et ouvrent des perspectives prometteuses pour des travaux futurs. Le chapitre suivant présentera les conclusions générales de cette recherche, ses limites intrinsèques et les perspectives d'amélioration du système QRV-haoussa.

CHAPITRE 5
CONCLUSION ET PERSPECTIVES

5.1 Résumé des apports

Cette recherche a proposé une évaluation comparative approfondie d'un système de question-réponse visuelle (QRV) par classification pour la langue haoussa, une langue africaine à faibles ressources. L'étude a examiné 36 combinaisons architecturales résultant du croisement de 9 grands modèles de langue (GML) et de 4 transformateurs d'image (TI), évaluées selon trois protocoles d'entraînement distincts adaptés au contexte de langue à faibles ressources : sans augmentation de données (baseline), avec augmentation en ligne (transformations dynamiques), et avec augmentation hors ligne (corpus enrichi $HaVQAaug$ obtenu par extension de $HaVQA$.

5.1.1 Résultats expérimentaux

Les expérimentations ont démontré que l'intégration méthodique de modèles linguistiques avancés, de transformateurs d'images robustes et de stratégies éprouvées d'enrichissement des données permet d'améliorer significativement les performances des systèmes QRV, et ce, malgré un contexte marqué par la rareté des ressources. Ces résultats confirment que les techniques multimodales et les approches d'augmentation de données constituent un levier efficace pour renforcer la qualité des modèles destinés aux langues à faibles ressources. L'approche standard sans augmentation a établi une référence avec un F1-score maximal de 1,73 % pour la combinaison (LLaMA-3.2-1B, CLIP). L'augmentation en ligne a apporté des gains modestes, atteignant 1,90 % de F1-score, offrant une diversification visuelle dynamique sans coût de stockage additionnel. En revanche, l'augmentation hors ligne s'est révélée particulièrement efficace, produisant des résultats spectaculaires sur le jeu de données $HaVQAaug$ qui est une extension du corpus $HaVQA$ obtenue par augmentation hors ligne multimodal durant .

5.1.2 Performance optimale

La meilleure configuration identifiée combine le modèle Gemini-2.0-Flash avec le transformateur ViT-base-patch16-224-in21k, atteignant un score d'exactitude (Accuracy) de 35,85 %, un score Wu-Palmer de 35,89 % et un score F1-score de 15,32 %. Ces performances dépassent de 5,03 points le score Wu-Palmer de référence rapporté par Parida et al. (2023) (30,86 %), établissant ainsi un nouvel état de l'art pour le système QRV-haoussa. Il convient de souligner que le modèle Gemini est pré-entraîné sur le haoussa, ce qui explique en partie ses excellentes performances. Le nouveau jeu de données $HaVQAaug$ est accessible publiquement (Mijiyawa, 2025).

5.1.3 Enseignements méthodologiques

L'analyse comparative a révélé plusieurs enseignements clés. Premièrement, le préentraînement spécifique à la langue cible amplifie de manière systématique les bénéfices de l'augmentation hors ligne, mettant en évidence l'importance de concevoir conjointement le modèle et les données en fonction de la langue étudiée. Deuxièmement, les architectures encodeur-décodeur multilingues (Gemini, mT0) bénéficient massivement de l'enrichissement multimodal du corpus, tandis que les modèles de type décodeur de grande capacité (LLaMA-3.2-1B) manifestent une sensibilité accrue au bruit introduit par les augmentations textuelles. Troisièmement, la capacité remarquable de LLaMA-3.2-1B à catégoriser correctement les réponses en l'absence de tout préentraînement sur le haoussa démontre le potentiel du transfert interlangue pour les langues à faibles ressources.

5.2 Limites de la recherche

Bien que cette étude ait apporté une contribution significative à la compréhension du problème étudié, elle présente néanmoins certaines limitations qui constituent autant de pistes pour des travaux futurs.

5.2.1 Contraintes computationnelles

Les contraintes matérielles et de calcul ont considérablement restreint l'étendue des expérimentations. La mémoire GPU limitée et la consommation élevée en ressources ont entravé l'exploration systématique des hyperparamètres. Des ajustements techniques ont été nécessaires (réduction du *batch size*, utilisation de la précision mixte `bf16`, accumulation des gradients, gel partiel des paramètres) afin de rendre les expériences réalisables. Ces compromis ont potentiellement limité l'optimisation des performances et la profondeur de l'analyse comparative.

5.2.2 Limitations inhérentes au corpus

Le jeu de données $HaVQA$, bien qu'il constitue une ressource novatrice pour le haoussa, présente des limitations intrinsèques. Composé de 6 022 paires question-réponse et de 1 555 images, il demeure restreint sur les plans linguistique et visuel, rendant la généralisation difficile. Le déséquilibre prononcé des classes (2 991 catégories de réponses) se traduit par des performances plus faibles sur les classes minoritaires et un F1-score macro modeste, même après augmentation. De plus, l'absence d'autres corpus de référence pour le QRV-haoussa rend impossible toute validation croisée ou comparaison inter-corpus, ce qui aurait pu renforcer la robustesse de l'évaluation.

5.2.3 Écart de performance persistant

Malgré les améliorations apportées, les performances absolues (35,85 % d'accuracy) demeurent nettement inférieures à celles observées pour les langues à hautes ressources (généralement supérieures à 60 %, voire 75 % et plus). Cet écart s'explique par la complexité intrinsèque des systèmes QRV, la nature morphologiquement riche du haoussa, la taille limitée du corpus d'entraînement, et possiblement par des biais ou incohérences dans les annotations originales. Une expansion substantielle du corpus ou l'utilisation d'un jeu de données mieux équilibré permettrait probablement d'atteindre des performances supérieures.

5.2.4 Défis spécifiques à l'augmentation de données

Les transformations textuelles pour l'augmentation en ligne se sont avérées difficiles à réaliser pour le haoussa, en raison de sa grande variété de formes grammaticales et d'orthographes. Ces contraintes morphologiques ont limité la diversité des modifications possibles, afin d'éviter toute erreur linguistique. Pour l'augmentation hors ligne, le traitement multimodal (images, questions et réponses) a demandé un temps considérable (environ 12 heures), ralentissant les itérations expérimentales. De plus, la qualité variable des traductions automatiques et des substitutions synonymiques a introduit du bruit sémantique, particulièrement préjudiciable pour certaines architectures.

5.2.5 Limites des métriques d'évaluation

Les métriques utilisées, notamment la similarité Wu-Palmer, présentent leurs propres limitations. Cette métrique peut sous-estimer la proximité sémantique réelle entre certaines réponses, en raison de sa sensibilité aux variations lexicales et synonymiques. Bien que des mécanismes de normalisation et d'ajustement de seuils aient été introduits, cette contrainte a parfois réduit la fiabilité de l'évaluation quantitative. En synthèse, ces limitations, qu'elles soient dues aux ressources computationnelles, à la disponibilité et la qualité des données, ou aux indicateurs d'évaluation qui soulignent la nécessité de renforcer les ressources multimodales en haoussa, de développer des stratégies d'augmentation adaptées aux spécificités linguistiques africaines, et d'investir dans des environnements de calcul plus puissants pour permettre une exploration plus approfondie.

5.3 Travaux futurs

Une feuille de route structurée en deux volets est proposée afin de soutenir l'évolution du QRV pour le haoussa et, plus largement, celle des systèmes de question-réponse multimodale pour les langues à faibles ressources.

- *Objectifs à court terme :*
  - Explorer la combinaison synergique de l'augmentation hors ligne et de l'augmentation en ligne avec

des techniques d'adaptation efficaces en paramètres (*parameter-efficient fine-tuning*), telles que les *adapters*, LoRA (*Low-Rank Adaptation*) et le *prompt-tuning*, afin de maximiser les performances sous contraintes computationnelles tout en préservant les connaissances multilingues des modèles préentraînés.

- Poursuivre l'enrichissement du jeu de données $HaVQA$ pour les systèmes QRV, et constituer un nouveau corpus dédié à la traduction automatique de question-réponse visuelle (TAQRV) anglais–haoussa. Cette ressource viserait à supporter la tâche de traduction multimodale, en combinant des exemples annotés manuellement et des données pseudo-étiquetées générées par des modèles de vision-langage récents, afin de mieux capturer les spécificités culturelles et multimodales propres au haoussa.
- Implémenter des stratégies avancées de rééquilibrage des classes (sur-échantillonnage, perte focale, échantillonnage stratifié) et développer des métriques d'évaluation complémentaires tenant compte de la diversité sémantique et de la couverture culturelle des réponses.

■ *Objectifs à long terme :*

- Mettre en place des visualisations d'attention intermodale et des modules explicatifs (XAI) afin d'analyser l'interaction entre indices linguistiques et visuels, garantissant que les systèmes QRV fournissent des justifications transparentes et compréhensibles pour les locuteurs haoussa. Cette dimension d'interprétabilité est essentielle pour identifier les biais potentiels et améliorer la confiance des utilisateurs.
- Étendre et valider ce pipeline QRV sur d'autres langues africaines sous-représentées (yoruba, igbo, swahili, amharique), en recourant au préentraînement multilingue massif et à l'apprentissage par transfert cross-lingue. L'objectif est de constituer une infrastructure mutualisée pour les systèmes multimodaux africains.
- Explorer la compression et la distillation de modèles pour le déploiement embarqué sur des dispositifs à ressources limitées, permettant ainsi une diffusion plus large des technologies QRV dans les régions à faible connectivité. Parallèlement, élargir l'approche aux entrées audio et vidéo, au-delà des images statiques, pour couvrir des scénarios d'usage plus diversifiés.
- Intégrer des dispositifs socio-culturels de protection et des audits de biais dès la phase de constitution des données et tout au long du cycle de vie des modèles. Cela inclut la validation communautaire des annotations, l'évaluation de l'équité inter-classes, et la documentation exhaustive des sources et des processus de collecte.
- Concevoir des composants QRV quantiques exploratoires en étudiant l'encodage quantique des caractéristiques pour les images et questions, en expérimentant des architectures hybrides quantique–classique pour l'inférence, et en appliquant l'optimisation quantique à l'ajustement des paramètres dans des conditions de données limitées. Un prototype de sous-système QRV quantique reposant sur des Circuits Quantiques Variationnels (*Variational Quantum Circuits*) et des couches d'attention quantique sera élaboré pour évaluer les gains potentiels en rapidité, robustesse et efficacité énergétique sur $HaVQA$.

Ces perspectives, fondées sur l'augmentation ciblée des données, des techniques d'adaptation efficaces, la création de corpus spécialisés (tels que le TAQRV), l'interprétabilité (XAI), l'extension multilingue et l'exploration du calcul quantique, visent à réduire l'écart entre les systèmes de question-réponse multimodale à ressources abondantes et ceux déployés dans des contextes à faibles ressources, tout en promouvant une intelligence artificielle inclusive, éthique et culturellement pertinente.

5.4 Conclusion générale

Ce mémoire a abordé le défi ambitieux du développement d'un système de question-réponse visuelle pour le haoussa, une langue africaine comptant plus de 60 millions de locuteurs mais largement sous-représentée dans les technologies du traitement automatique du langage naturel. À travers une démarche méthodologique rigoureuse combinant modèles de langue avancés, transformateurs d'images et stratégies d'augmentation de données, cette recherche a démontré qu'il est possible d'obtenir des performances substantielles même dans un contexte de forte contrainte en ressources. Les contributions de ce travail s'articulent autour de trois axes principaux. Premièrement, l'établissement d'un nouveau référentiel de performance pour le QRV-haoussa, avec un score Wu-Palmer de 35,89 % dépassant l'état de l'art antérieur. Deuxièmement, la création et la mise à disposition publique du corpus enrichi $HaVQAaug$, doublant la taille du jeu de données original par augmentation multimodale. Troisièmement, une analyse comparative approfondie de 108 configurations expérimentales (36 combinaisons de modèles × 3 protocoles d'entraînement), révélant les architectures et stratégies les plus adaptées aux langues à faibles ressources. Au-delà des résultats quantitatifs, cette recherche souligne l'importance cruciale du préentraînement linguistiquement adapté et de l'augmentation multimodale pour compenser la rareté des données annotées. Elle met également en évidence les défis persistants entraînant le déséquilibre des classes, la variabilité morphologique, le coût computationnel. Ce qui nécessitent des solutions innovantes et adaptées au contexte africain. Les perspectives tracées ouvrent des voies prometteuses vers des systèmes multimodaux plus inclusifs, interprétables et culturellement ancrés. En contribuant à l'inclusion numérique des langues africaines dans l'écosystème de l'intelligence artificielle, ce travail s'inscrit dans une dynamique plus large visant à démocratiser l'accès aux technologies avancées et à valoriser la diversité linguistique mondiale comme une richesse plutôt qu'une contrainte. L'avenir du QRV pour les langues à faibles ressources dépendra de la mobilisation collective de la communauté scientifique, des locuteurs natifs, des décideurs et de l'industrie technologique pour investir dans la création de ressources de qualité, le développement d'infrastructures computationnelles accessibles, et la conception d'approches méthodologiques respectueuses des spécificités linguistiques et culturelles de chaque langue.

Abdulmumin, I., Dash, S. R., Dawud, M. A., Parida, S., Muhammad, S., Ahmad, I. S., Panda, S., Bojar, O., Galadanci, B. S., & Bello, B. S. (2022, juin). Hausa Visual Genome : A Dataset for Multi-Modal English to Hausa Machine Translation. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk & S. Piperidis (Éd.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (p. 6471-6479). European Language Resources Association. https://aclanthology.org/2022.lrec-1.694/

Adebara, I., Elmadany, A., & Abdul-Mageed, M. (2024a). Towards Equitable LLM Evaluation : A Framework for African Languages. *arXiv preprint arXiv :2406.12202*.

Adebara, I., Elmadany, A., & Abdul-Mageed, M. (2024b). Cheetah : Natural Language Generation for 517 African Languages. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, 12798-12823. https://doi.org/10.18653/v1/2024.acl-long.691

Adelani, D., Neubig, G., Ruder, S., Rijhwani, S., Beukman, M., Palen-Michel, C., Lignos, C., Alabi, J., Hassan Muhammad, S., Nabende, P., Dione, C., Bukula, A., Mabuya, R., Dossou, B., Sibanda, B., Buzaaba, H., Mukiibi, J., Kalipe, G., Mbaye, D., & Klakow, D. (2022). MasakhaNER 2.0 : Africa-centric Transfer Learning for Named Entity Recognition, 4488-4508. https://doi.org/10.18653/v1/2022.emnlp-main.298

Adelani, D. I. (2023). *bert-base-multilingual-cased-finetuned-hausa* (Version 1.0). Hugging Face. https://huggingface.co/Davlan/bert-base-multilingual-cased-finetuned-hausa

Adelani, D. I., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., Lignos, C., Palen-Michel, C., Buzaaba, H., Rijhwani, S., Ruder, S., et al. (2021). MasakhaNER : Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, *9*, 1116-1131.

Adelani, D. I., Alabi, J. O., Fan, A., Kreutzer, J., Shen, X., Reid, M., Ruiter, D., Klakow, D., Nabende, P., Chang, E., Gwadabe, T., Sackey, F., Dossou, B. F. P., Emezue, C., Leong, C., Beukman, M., Muhammad, S. H., Jarso, G. D., Yousuf, O., … Manthalu, S. (2022). A Few Thousand Translations Go a Long Way ! Leveraging Pre-trained Models for African News Translation. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Com-*

putational Linguistics : Human Language Technologies (NAACL 2022), 3053-3070. https://doi.org/10.18653/v1/2022.naacl-main.223

Adelani, D. I., Liu, H., Shen, X., Vassilyev, N., Alabi, J. O., Mao, Y., Gao, H., & Lee, A. E.-S. (2024a, mars). SIB-200 : A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects [arXiv :2309.07445 [cs]]. https://doi.org/10.48550/arXiv.2309.07445

Adelani, D. I., Liu, H., Shen, X., Vassilyev, N., Alabi, J. O., Mao, Y., Gao, H., & Lee, A. E.-S. (2024b, mars). SIB-200 : A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects [arXiv :2309.07445 [cs]]. https://doi.org/10.48550/arXiv.2309.07445

Adelani, D. I., Neubig, G., Ruder, S., Rijhwani, S., Beukman, M., Palen-Michel, C., Lignos, C., Alabi, J. O., Muhammad, S. H., Nabende, P., Dione, C. M. B., Bukula, A., Mabuya, R., Dossou, B. F. P., Sibanda, B., Buzaaba, H., Mukiibi, J., Kalipe, G., Mbaye, D., … Klakow, D. (2022, décembre). MasakhaNER 2.0 : Africa-centric Transfer Learning for Named Entity Recognition. In Y. Goldberg, Z. Kozareva & Y. Zhang (Éd.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (p. 4488-4508). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.298

Adelani, D. I., Ojo, J., Azime, I. A., Zhuang, J. Y., Alabi, J. O., He, X., Ochieng, M., Hooker, S., Bukula, A., Lee, E.-S. A., Chukwuneke, C., Buzaaba, H., Sibanda, B., Kalipe, G., Mukiibi, J., Kabongo, S., Yuehgoh, F., Setaka, M., Ndolela, L., … Stenetorp, P. (2024, juin). IrokoBench : A New Benchmark for African Languages in the Age of Large Language Models [arXiv :2406.03368 [cs]]. https://doi.org/10.48550/arXiv.2406.03368

Agić, Ž., & Vulić, I. (2019, juillet). JW300 : A Wide-Coverage Parallel Corpus for Low-Resource Languages. In A. Korhonen, D. Traum & L. Màrquez (Éd.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (p. 3204-3210). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1310

Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Batra, D., & Parikh, D. (2016, octobre). VQA : Visual Question Answering. https://doi.org/10.48550/arXiv.1505.00468

Ahia, O., et al. (2023). Do All Languages Cost the Same ? Tokenization in the Era of Commercial Language Models. *arXiv preprint arXiv :2305.13707*.

AI, M. (2024). The Llama 3 Herd of Models. *arXiv preprint arXiv :2407.21783*.

Alabi, J. O., Adelani, D. I., Mosbach, M., & Klakow, D. (2022a). Adapting Pre-trained Language Models to African Languages via Multilingual Adaptive Fine-Tuning. *Proceedings of the 29th International Conference on Computational Linguistics*, 4336-4349. https://aclanthology.org/2022.coling-1.382/

Alabi, J. O., Adelani, D. I., Mosbach, M., & Klakow, D. (2022b, octobre). Adapting Pre-trained Language Models to African Languages via Multilingual Adaptive Fine-Tuning. In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond & S.-H. Na (Éd.), *Proceedings of the 29th International Conference on Computational Linguistics* (p. 4336-4349). International Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.382/

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., … Simonyan, K. (2022). Flamingo : a Visual Language Model for Few-Shot Learning. *ArXiv*, *abs/2204.14198*. https://api.semanticscholar.org/CorpusID:248476411

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077-6086.

Ardila, R., et al. (2023). Common Voice : A Massively-Multilingual Speech Corpus. *arXiv preprint arXiv :2305.09823*.

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., & Weber, G. (2020, mai). Common Voice : A Massively-Multilingual Speech Corpus. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (Éd.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (p. 4218-4222). European Language Resources Association. https://aclanthology.org/2020.lrec-1.520/

Askarian, N., Abbasnejad, E., Zukerman, I., Buntine, W., & Haffari, G. (2022). Inductive Biases for Low Data VQA : A Data Augmentation Approach. *2022 IEEE/CVF Winter Conference on Ap-*

plications of Computer Vision Workshops (WACVW), 231-240. https://doi.org/10.1109/WACVW54805.2022.00029

Babu, A., et al. (2024). SeamlessM4T : Massively Multilingual & Multimodal Machine Translation. *Proceedings of NeurIPS*.

Black, A. W. (2019). CMU Wilderness Multilingual Speech Dataset. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5971-5975. https://doi.org/10.1109/ICASSP.2019.8683536

Blasi, D., Anastasopoulos, A., & Neubig, G. (2022, mai). Systematic Inequalities in Language Technology Performance across the World's Languages. In S. Muresan, P. Nakov & A. Villavicencio (Éd.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)* (p. 5486-5505). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.376

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Zhu, C. (2021). On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv :2108.07258*. https://doi.org/10.48550/arXiv.2108.07258

Bondarev, D. (2021). A Typology of West African Ajami Manuscripts : Languages, Layout and Research Perspectives. In J. B. Quenzer (Éd.), *Exploring Written Artefacts : Objects, Methods, and Concepts* (p. 707-728). De Gruyter. https://doi.org/10.1515/9783110753301-035

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877-1901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Bugliarello, E., et al. (2022). IGLUE : A Benchmark for Transfer Learning across Modalities, Tasks, and Languages. *Proceedings of ICML*.

Campbell, V., Lyle et Grondona. (2008). Ethnologue : Languages of the world. *Language*, *84*(3), 636-641. https://doi.org/10.1353/lan.0.0035

Chen, D., Zhuang, Y., Shen, Z., Yang, C., Wang, G., Tang, S., & Yang, Y. (2022). Cross-modal data augmentation for tasks of different modalities. *IEEE Transactions on Multimedia*, *25*, 7814-7824.

Chen, L., Zheng, Y., & Xiao, J. (2022). Rethinking Data Augmentation for Robust Visual Question Answering. *Computer Vision – ECCV 2022 : 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, 95-112. https://doi.org/10.1007/978-3-031-20059-5_6

Cho, J., Lei, J., Tan, H., & Bansal, M. (2021). Unifying Vision-and-Language Tasks via Text Generation [arXiv :2102.02779 [cs], Version 2]. https://doi.org/10.48550/arXiv.2102.02779

Chronopoulou, A., Baziotis, C., & Potamianos, A. (2019, juin). An Embarrassingly Simple Approach for Transfer Learning from Pretrained Language Models. In J. Burstein, C. Doran & T. Solorio (Éd.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)* (p. 2089-2095). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1213

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020, avril). Unsupervised Cross-lingual Representation Learning at Scale. https://doi.org/10.48550/arXiv.1911.02116

Dai, W., et al. (2023). InstructBLIP : Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv preprint arXiv :2305.06500*.

DeepSeek-AI. (2025, janvier). DeepSeek-R1 : Incentivizing Reasoning Capability in LLMs via Reinforcement Learning [arXiv :2501.12948 [cs]]. https://doi.org/10.48550/arXiv.2501.12948

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., … Zhang, Z. (2025, janvier). DeepSeek-R1 : Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. https://doi.org/10.48550/arXiv.2501.12948

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019a). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*, 1-16. https://doi.org/10.48550/arXiv.1810.04805

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019b, mai). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/arXiv.1810.04805

Ding, B., Qin, C., Zhao, R., Luo, T., Li, X., Chen, G., Xia, W., Hu, J., Luu, A. T., & Joty, S. (2024, juillet). Data Augmentation using Large Language Models : Data Perspectives, Learning Paradigms and Challenges. https://doi.org/10.48550/arXiv.2403.02990

Dione, C. M. B., Adelani, D. I., Nabende, P., Alabi, J., Sindane, T., Buzaaba, H., Muhammad, S. H., Emezue, C. C., Ogayo, P., Aremu, A., Gitau, C., Mbaye, D., Mukiibi, J., Sibanda, B., Dossou, B. F. P., Bukula, A., Mabuya, R., Tapo, A. A., Munkoh-Buabeng, E., … Klakow, D. (2023). MasakhaPOS : Part-of-Speech Tagging for Typologically Diverse African languages. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, 10883-10900. https://doi.org/10.18653/v1/2023.acl-long.609

Dong, L., Mallinson, J., Reddy, S., & Lapata, M. (2017, septembre). Learning to Paraphrase for Question Answering. In M. Palmer, R. Hwa & S. Riedel (Éd.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (p. 875-886). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1091

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021a). An Image is Worth 16x16 Words : Transformers for Image Recognition at Scale [ViT-Base-Patch16-224 model (86M parameters)]. *International Conference on Learning Representations (ICLR)*. https://huggingface.co/google/vit-base-patch16-224

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021b). An Image is Worth 16x16 Words : Transformers for Image Recognition at Scale [ViT-Base-Patch16-224 (86M parameters)]. *International Conference on Learning Representations (ICLR)*. Vision Transformer Base model. https://huggingface.co/google/vit-base-patch16-224

Dossou, B. F. P., Tonja, A. L., Yousuf, O., Osei, S., Oppong, A., Shode, I., Awoyomi, O. O., & Emezue, C. (2022, décembre). AfroLM : A Self-Active Learning-based Multilingual Pretrained Language Model for 23 African Languages. In A. Fan, I. Gurevych, Y. Hou, Z. Kozareva, S. Luccioni, N. Sadat Moosavi, S. Ravi, G. Kim, R. Schwartz & A. Rücklé (Éd.), *Proceedings of the Third*

Workshop on Simple and Efficient Natural Language Processing (SustaiNLP) (p. 52-64). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.sustainlp-1.11

Doumbouya, M., Einstein, L., & Piech, C. (2021, avril). Using Radio Archives for Low-Resource Speech Recognition : Towards an Intelligent Virtual Assistant for Illiterate Users [arXiv :2104.13083 [cs]]. https://doi.org/10.48550/arXiv.2104.13083

Elmadany, A., et al. (2024). Aya Model : An Instruction Finetuned Open-Access Multilingual Language Model. *arXiv preprint arXiv :2402.07827*.

Fawcett, T. (2006). Introduction to ROC analysis. *Pattern Recognition Letters*, *27*, 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. (2021, août). A Survey of Data Augmentation Approaches for NLP. In C. Zong, F. Xia, W. Li & R. Navigli (Éd.), *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021* (p. 968-988). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.84

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv :1606.01847*. https://doi.org/10.48550/arXiv.1606.01847

Furniss, G. (1995). The Power of Words and the Relationship between Hausa Genres. In G. Furniss & L. Gunner (Éd.), *Power, Marginality and African Oral Literature*. Cambridge University Press.

Furniss, G. (1996). *Poetry, Prose and Popular Culture in Hausa*. Edinburgh University Press for the International African Institute.

Furniss, G. (2019). *Poetry, prose and popular culture in Hausa*. Edinburgh University Press.

Gauthier, E., Besacier, L., Voisin, S., Melese, M., & Elingui, U. P. (2016, mai). Collecting Resources in Sub-Saharan African Languages for Automatic Speech Recognition : a Case Study of Wolof. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (Éd.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (p. 3863-3867). European Language Resources Association (ELRA). https://aclanthology.org/L16-1611/

Gemini Team, Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., Firat, O., et al. (2025, mai). Gemini : A Family of Highly Capable Multimodal Models [arXiv :2312.11805 [cs]. Access via Google AI Studio : https://ai.google.dev/gemini-api or Vertex AI]. https://doi.org/10.48550/arXiv.2312.11805

Gong, H., Chen, G., Mao, M., Li, Z., & Li, G. (2022). VQAMix : Conditional Triplet Mixup for Medical Visual Question Answering. *IEEE Transactions on Medical Imaging*, *41*(11), 3332-3343. https://doi.org/10.1109/TMI.2022.3185008

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT press.

Gupta, V., et al. (2024). IndicTTS : Text-to-Speech for Low-Resource Languages. *arXiv preprint arXiv :2401.09845*.

Hacheme, G., & Sayouti, N. (2021). Neural fashion image captioning : Accounting for data diversity. *arXiv preprint arXiv :2106.12154*. https://doi.org/10.31730/osf.io/hwtpq

HausaNLP. (2023). HausaVQA : Visual Question Answering Dataset for Hausa [Accessed : January 28, 2025].

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked Autoencoders Are Scalable Vision Learners [MAE ViT-Base (86M parameters)]. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16000-16009. Masked Autoencoder Vision Transformer. https://huggingface.co/facebook/vit-mae-base

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770-778. https://doi.org/10.1109/CVPR.2016.90

He, W., Ma, H., Li, S., Dong, H., Zhang, H., & Feng, J. (2023). Using Augmented Small Multimodal Models to Guide Large Language Models for Multimodal Relation Extraction. *Applied Sciences*, *13*, 12208. https://doi.org/10.3390/app132212208

Heaton, J. (2017). Ian Goodfellow, Yoshua Bengio, and Aaron Courville : Deep learning : The MIT Press, 2016, 800 pp, ISBN : 0262035618. *Genetic Programming and Evolvable Machines*, *19*. https://doi.org/10.1007/s10710-017-9314-z

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., & Klakow, D. (2021, juin). A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In K. Toutanova, A.

Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty & Y. Zhou (Éd.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies* (p. 2545-2568). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.201

Howard, J., & Ruder, S. (2018, mai). Universal Language Model Fine-tuning for Text Classification [arXiv :1801.06146 [cs]]. https://doi.org/10.48550/arXiv.1801.06146

Ilharco, G., Zhang, Y., & Baldridge, J. (2019, septembre). Large-scale representation learning from visually grounded untranscribed speech [arXiv :1909.08782 [cs]]. https://doi.org/10.48550/arXiv.1909.08782

Inuwa-Dutse, I. (2021). The first large scale collection of diverse Hausa language datasets. https://doi.org/10.48550/ARXIV.2102.06991

Jaggar, P. J. (2006a). Hausa. In K. Brown (Éd.), *Encyclopedia of Language & Linguistics (Second Edition)* (Second Edition, p. 222-225). Elsevier. https://doi.org/https://doi.org/10.1016/B0-08-044854-2/02071-X

Jaggar, P. J. (2006b). Hausa. In K. Brown (Éd.), *Encyclopedia of Language & Linguistics (Second Edition)* (Second Edition, p. 222-225). Elsevier. https://doi.org/https://doi.org/10.1016/B0-08-044854-2/02071-X

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020a). The state and fate of linguistic diversity and inclusion in the NLP world. *arXiv preprint arXiv :2004.09095*. https://doi.org/10.48550/arXiv.2004.09095

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020b). The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6282-6293. https://doi.org/10.18653/v1/2020.acl-main.560

Jude Ogundepo, O., Oladipo, A., Adeyemi, M., Ogueji, K., & Lin, J. (2022, juillet). AfriTeVA : Extending ?Small Data ? Pretraining Approaches to Sequence-to-Sequence Models. In C. Cherry, A. Fan, G. Foster, G. (Haffari, S. Khadivi, N. (Peng, X. Ren, E. Shareghi & S. Swayamdipta (Éd.), *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Lan-*

guage Processing (p. 126-135). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.deeplo-1.14

Kafle, K., Yousefhussien, M., & Kanan, C. (2017, septembre). Data Augmentation for Visual Question Answering. In J. M. Alonso, A. Bugarín & E. Reiter (Éd.), *Proceedings of the 10th International Conference on Natural Language Generation* (p. 198-202). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-3529

Kaushik, D., Hovy, E., & Lipton, Z. C. (2020, février). Learning the Difference that Makes a Difference with Counterfactually-Augmented Data [arXiv :1909.12434 [cs]]. https://doi.org/10.48550/arXiv.1909.12434

Comment : Published at ICLR 2020.

Kaye, A. S. (2002). The Hausa Language : An Encyclopedic Reference Grammar.(Reviews of Books). *The Journal of the American Oriental Society*, *122*(1), 97-99.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. Burges, L. Bottou & K. Q. Weinberger (Éd.), *Advances in Neural Information Processing Systems* (T. 25). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Kumar, G. K., Gehlot, A., Mullappilly, S. S., & Nandakumar, K. (2022, mai). MuCoT : Multilingual Contrastive Training for Question-Answering in Low-resource Languages. In B. R. Chakravarthi, R. Priyadharshini, A. K. Madasamy, P. Krishnamurthy, E. Sherly & S. Mahesan (Éd.), *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages* (p. 15-24). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.dravidianlangtech-1.3

Kumar, V., Choudhary, A., & Cho, E. (2020, décembre). Data Augmentation using Pre-trained Transformer Models. In W. M. Campbell, A. Waibel, D. Hakkani-Tur, T. J. Hazen, K. Kilgour, E. Cho, V. Kumar & H. Glaude (Éd.), *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems* (p. 18-26). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.lifelongnlp-1.3

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2 : Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *Proceedings of the 40th International*

Conference on Machine Learning, *202*, 19730-19742. https://proceedings.mlr.press/v202/li23q.html

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP : Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *Proceedings of ICML*, 12888-12900.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., & Chang, K.-W. (2019). VisualBERT : A Simple and Performant Baseline for Vision and Language. *Proceedings of NeurIPS Workshop on Visually Grounded Interaction and Language*.

Li, Y., et al. (2024). Synthetic Data Generation for Low-Resource Multimodal Tasks. *arXiv preprint arXiv :2405.12389*.

Liu, H., et al. (2023). Visual Instruction Tuning. *arXiv preprint arXiv :2304.08485*.

Liu, Z., Tang, Z., Shi, X., Zhang, A., Li, M., Shrivastava, A., & Wilson, A. G. (2023). Learning Multimodal Data Augmentation in Feature Space. *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=6SRDbbvU8s

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT : Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *Proceedings of NeurIPS*, 13-23.

Mallinson, J., Sennrich, R., & Lapata, M. (2017, avril). Paraphrasing Revisited with Neural Machine Translation. In M. Lapata, P. Blunsom & A. Koller (Éd.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers* (p. 881-893). Association for Computational Linguistics. https://aclanthology.org/E17-1083/

Mellouk, A., Dossou, B. F. P., Emezue, C. C., & Orife, I. (2021). AfroTTS : Building Multilingual Text-to-Speech Voices for African Languages. *Proceedings of the First Workshop on Multilingual Representation Learning (MRL)*, 96-106. https://doi.org/10.18653/v1/2021.mrl-1.10

Meta AI. (2024). The Llama 3 Herd of Models [Llama-3-8B model (requires access approval)]. https://doi.org/10.48550/arXiv.2407.21783

Mijiyawa, A. (2025). LLM_QRV_Hausa_HaVQA_aug : Visual Question Answering for Hausa using Large Language Models and Data Augmentation [Code and experiments for Visual Question Answering in Hausa].

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Le Scao, T., Bari, M. S., Shen, S., Yong, Z. X., Schoelkopf, H., Tang, X., Radev, D., Aji, A. F., Almubarak, K., Albanie, S., Alyafeai,

Z., Webson, A., Raff, E., & Raffel, C. (2023, juillet). Crosslingual Generalization through Multitask Finetuning. In A. Rogers, J. Boyd-Graber & N. Okazaki (Éd.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)* (p. 15991-16111). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.891

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T. L., Bari, M. S., Shen, S., Yong, Z.-X., Schoelkopf, H., Tang, X., Radev, D., Aji, A. F., Almubarak, K., Albanie, S., Alyafeai, Z., Webson, A., Raff, E., & Raffel, C. (2022a). Crosslingual Generalization through Multitask Finetuning [mT0-base model (300M parameters)]. https://huggingface.co/bigscience/mt0-base

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T. L., Bari, M. S., Shen, S., Yong, Z.-X., Schoelkopf, H., Tang, X., Radev, D., Aji, A. F., Almubarak, K., Albanie, S., Alyafeai, Z., Webson, A., Raff, E., & Raffel, C. (2022b). Crosslingual Generalization through Multitask Finetuning [mT0-large model (1.2B parameters)]. https://huggingface.co/bigscience/mt0-large

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T. L., Bari, M. S., Shen, S., Yong, Z.-X., Schoelkopf, H., Tang, X., Radev, D., Aji, A. F., Almubarak, K., Albanie, S., Alyafeai, Z., Webson, A., Raff, E., & Raffel, C. (2023a). Crosslingual Generalization through Multitask Finetuning [BLOOMZ-560m model]. https://huggingface.co/bigscience/bloomz-560m

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T. L., Bari, M. S., Shen, S., Yong, Z.-X., Schoelkopf, H., Tang, X., Radev, D., Aji, A. F., Almubarak, K., Albanie, S., Alyafeai, Z., Webson, A., Raff, E., & Raffel, C. (2023b). Crosslingual Generalization through Multitask Finetuning [BLOOMZ-1b7 model (1.7B parameters)]. https://huggingface.co/bigscience/bloomz-1b7

Muhammad, S. H., Abdulmumin, I., Ayele, A. A., Ousidhoum, N., Adelani, D. I., Yimam, S. M., Ahmad, I. S., Beloucif, M., Mohammad, S. M., Ruder, S., Hourrane, O., Brazdil, P., Jorge, A., Ali, F. D. M. A., David, D., Osei, S., Shehu Bello, B., Ibrahim, F., Gwadabe, T., … Arthur, S. (2023). AfriSenti : A Twitter Sentiment Analysis Benchmark for African Languages. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 13968-13981. https://doi.org/10.18653/v1/2023.emnlp-main.862

Muhammad, S. H., Adelani, D. I., Ruder, S., Ahmad, I. S., Abdulmumin, I., Bello, B. S., Choudhury, M., Emezue, C. C., Abdullahi, S. S., Aremu, A., Jeorge, A., & Brazdil, P. (2022). NaijaSenti : A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis. *arXiv preprint arXiv :2201.08277*. https://doi.org/10.48550/arXiv.2201.08277

Muhammad, S. H., Ahmad, I. S., Abdulmumin, I., Lawan, F. I., Sani, B., Imam, S. H., Aliyu, Y., Sani, S. A., Umar, A. U., Gwadabe, T., Church, K., & Marivate, V. (2025, juillet). HausaNLP : Current Status, Challenges and Future Directions for Hausa Natural Language Processing [arXiv :2505.14311 [cs]]. https://doi.org/10.48550/arXiv.2505.14311

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Kolawole, T., Adeyemi, M., Mokgonyane, T., Ahia, O., Osei, S., et al. (2020). Participatory research for low-resourced machine translation : A case study in African languages. *Findings of the Association for Computational Linguistics : EMNLP 2020*, 2144-2160. https://doi.org/10.18653/v1/2020.findings-emnlp.195

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S., Kabongo Kabenamualu, S., Osei, S., Sackey, F., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Berhe, M. M., Adeyemi, M., Mokgesi-Selinga, M., Okegbemi, L., Martinus, L., … Bashir, A. (2020, novembre). Participatory Research for Low-resourced Machine Translation : A Case Study in African Languages. In T. Cohn, Y. He & Y. Liu (Éd.), *Findings of the Association for Computational Linguistics : EMNLP 2020* (p. 2144-2160). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.195

Ogueji, K., Zhu, Y., & Lin, J. (2021a). Small Data ? No Problem ! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages [AfriBERTa-large model for 11 African languages]. https://huggingface.co/castorini/afriberta_large

Ogueji, K., Zhu, Y., & Lin, J. (2021b, novembre). Small Data ? No Problem ! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages. In D. Ataman, A. Birch, A. Conneau, O. Firat, S. Ruder & G. G. Sahin (Éd.), *Proceedings of the 1st Workshop on Multilingual Representation Learning* (p. 116-126). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.mrl-1.11

Ojo, J., Ogundepo, O., Oladipo, A., Ogueji, K., Lin, J., Stenetorp, P., & Adelani, D. I. (2025). AfroBench : How Good are Large Language Models on African Languages ? *Findings of the Asso-*

ciation for Computational Linguistics : ACL 2025, 19048-19095. https://doi.org/10.18653/v1/2025.findings-acl.976

OpenAI. (2023). *GPT-4V(ision) System Card* (rapp. tech.). OpenAI.

Otter, D. W., Medina, J. R., & Kalita, J. K. (2021). A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(2), 604-624. https://doi.org/10.1109/TNNLS.2020.2979670

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022, mars). Training language models to follow instructions with human feedback. https://doi.org/10.48550/arXiv.2203.02155

Parida, S., Abdulmumin, I., Muhammad, S. H., Bose, A., Kohli, G. S., Ahmad, I. S., Kotwal, K., Deb Sarkar, S., Bojar, O., & Kakudi, H. (2023, juillet). HaVQA : A Dataset for Visual Question Answering and Multimodal Research in Hausa Language. In A. Rogers, J. Boyd-Graber & N. Okazaki (Éd.), *Findings of the Association for Computational Linguistics : ACL 2023* (p. 10162-10183). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.646

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., & Collobert, R. (2020). MLS : A Large-Scale Multilingual Dataset for Speech Research. *Interspeech 2020*, 2757-2761. https://doi.org/10.21437/Interspeech.2020-2826

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision [CLIP ViT-B/16 (86M parameters for vision encoder)]. *International Conference on Machine Learning (ICML)*, 8748-8763. CLIP Vision Transformer model. https://huggingface.co/openai/clip-vit-base-patch16

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2023, septembre). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [arXiv :1910.10683 [cs]]. https://doi.org/10.48550/arXiv.1910.10683

Reid, M., et al. (2024). AfroLM : A Self-Active Learning-based Multilingual Pretrained Language Model for 23 African Languages. *Proceedings of EMNLP*.

Reid, M., Hu, J., Neubig, G., & Matsuo, Y. (2021, septembre). AfroMT : Pretraining Strategies and Reproducible Benchmarks for Translation of 8 African Languages [arXiv :2109.04715 [cs]]. https://doi.org/10.48550/arXiv.2109.04715
Comment : EMNLP 2021.

Romero, D., Lyu, C., Wibowo, H. A., Lynn, T., Hamed, I., Kishore, A. N., Mandal, A., Dragonetti, A., Abzaliev, A., Tonja, A. L., Balcha, B. F., Whitehouse, C., Salamea, C., Velasco, D. J., Adelani, D. I., Le Meur, D., Villa-Cueva, E., Koto, F., Farooqui, F., … Aji, A. F. (2024, novembre). CVQA : Culturally-diverse Multilingual Visual Question Answering Benchmark [arXiv :2406.05967 [cs]]. https://doi.org/10.48550/arXiv.2406.05967

Sasaki, Y. (2007). The truth of the F-measure. *Teach Tutor Mater*.

scikit-learn developers. (2024). sklearn.preprocessing.LabelEncoder [scikit-learn documentation. Accessed : 2025-01-28].

Sennrich, R., Haddow, B., & Birch, A. (2016a). Improving Neural Machine Translation Models with Monolingual Data. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, 86-96. https://doi.org/10.18653/v1/P16-1009

Sennrich, R., Haddow, B., & Birch, A. (2016b, août). Neural Machine Translation of Rare Words with Subword Units. In K. Erk & N. A. Smith (Éd.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)* (p. 1715-1725). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1162

Shorten, C., & Khoshgoftaar, T. M. (2019a). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, *6*(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Shorten, C., & Khoshgoftaar, T. M. (2019b). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, *6*(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Siaminwe, L., et al. (2024). Multimodal Learning for Bantu Languages : A Vision-Language Approach. *Proceedings of AfricaNLP*.

Simard, P. Y., Steinkraus, D., & Platt, J. C. (2003). Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis. *Proceedings of the Seventh International Conference*

on Document Analysis and Recognition (ICDAR 2003), 958-963. https://doi.org/10.1109/ICDAR.2003.1227801

Simonyan, K., & Zisserman, A. (2015, avril). Very Deep Convolutional Networks for Large-Scale Image Recognition [arXiv :1409.1556 [cs]]. https://doi.org/10.48550/arXiv.1409.1556

Singh, A., et al. (2024). XGen-MM : Scaling up Multimodal LLMs. *arXiv preprint arXiv :2408.08872*.

Tan, H., & Bansal, M. (2019). LXMERT : Learning Cross-Modality Encoder Representations from Transformers. *Proceedings of EMNLP-IJCNLP*, 5100-5111.

Tang, R., Ma, C., Zhang, W. E., Wu, Q., & Yang, X. (2020). Semantic Equivalent Adversarial Data Augmentation for Visual Question Answering. In A. Vedaldi, H. Bischof, T. Brox & J.-M. Frahm (Éd.), *Computer Vision – ECCV 2020* (p. 437-453). Springer International Publishing.

Team, G., Anil, R., Borgeaud, S., et al. (2025, mai). Gemini : A Family of Highly Capable Multimodal Models [arXiv :2312.11805 [cs]]. https://doi.org/10.48550/arXiv.2312.11805

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021a). Training Data-efficient Image Transformers & Distillation through Attention [DeiT-Base model]. *International Conference on Machine Learning (ICML)*, 10347-10357. https : / / huggingface . co / facebook/deit-base-patch16-224

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021b). Training Data-efficient Image Transformers & Distillation through Attention [DeiT-Base-Patch16-224 (86M parameters)]. *International Conference on Machine Learning (ICML)*, 10347-10357. Data-efficient Image Transformer. https://huggingface.co/facebook/deit-base-patch16-224

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023, août). Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762

Wang, J., Adelani, D. I., Agrawal, S., Masiak, M., Rei, R., Briakou, E., Carpuat, M., He, X., Bourhim, S., Bukula, A., Mohamed, M., Olatoye, T., Adewumi, T., Mokayed, H., Mwase, C., Kimotho, W., Yuehgoh, F., Aremu, A., Ojo, J., … Stenetorp, P. (2024, juin). AfriMTE and AfriCOMET : Enhancing COMET to Embrace Under-resourced African Languages. In K. Duh, H. Gomez & S. Bethard (Éd.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)* (p. 5997-6023). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.334

Wang, Z., Miao, Y., & Specia, L. (2021, octobre). Cross-Modal Generative Augmentation for Visual Question Answering [arXiv :2105.04780 [cs]]. https://doi.org/10.48550/arXiv.2105.04780 Comment : BMVC 2021.

Wei, J., & Zou, K. (2019, novembre). EDA : Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In K. Inui, J. Jiang, V. Ng & X. Wan (Éd.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (p. 6382-6388). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1670

Winata, G. I., et al. (2023). NollySenti : Leveraging Nollywood for Nigerian Movie Sentiment Classification. *Proceedings of AfricaNLP Workshop*.

Workshop, B., Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., et al. (2022). BLOOM : A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv :2211.05100*. https://doi.org/10.48550/arXiv.2211.05100

Wu, X., Lv, S., Zang, L., Han, J., & Hu, S. (2018, décembre). Conditional BERT Contextual Augmentation [arXiv :1812.06705 [cs]]. https://doi.org/10.48550/arXiv.1812.06705

Wu, Z., & Palmer, M. (1994). Verb Semantics and Lexical Selection. *32nd Annual Meeting of the Association for Computational Linguistics*, 133-138. https://doi.org/10.3115/981732.981751

Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., & Shen, F. (2023, novembre). Image Data Augmentation for Deep Learning : A Survey. https://doi.org/10.48550/arXiv.2204.08610

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, *13*(3), 55-75. https://doi.org/10.1109/MCI.2018.2840738

Zhang, X., Zhao, J., & LeCun, Y. (2016, avril). Character-level Convolutional Networks for Text Classification [arXiv :1509.01626 [cs]]. https://doi.org/10.48550/arXiv.1509.01626

Zheng, Y., Wang, Z., & Chen, L. (2024). Improving Data Augmentation for Robust Visual Question Answering with Effective Curriculum Learning. *Proceedings of the 2024 International Conference on Multimedia Retrieval*, 1084-1088. https://doi.org/10.1145/3652583.3657607