Reconnaissance automatique de manuscrits arabes scientifiques médiévaux : Création d'un jeu de données et évaluation de grands modèles de langues

Soumission anonyme

RÉSUMÉ

La numérisation du patrimoine scientifique arabe constitue un défi important pour l'analyse de ces documents, en raison de la complexité spatiale de l'écriture (cursivité, diacritiques, ligatures). Malgré les progrès récents, ces manuscrits demeurent souvent indéchiffrables pour les modèles d'intelligence artificielle actuels. Dans cet article, nous présentons une évaluation de méthodes de l'état de l'art sur un nouveau corpus édité et annoté manuellement : *al-Qānūn al-Masʿūdī d'Al-Bīrūnī*, un traité majeur d'astronomie mathématique, comparable à l'Almageste de Ptoléméeaux et écrit aux alentours de 1030. Nous présentons la méthodologie de construction du corpus, puis détaillons l'approche mise en place pour la numérisation de ce manuscrit. Cette dernière prend la forme d'un pipeline unifié à base de grands modèles de langue dont l'entrée est une image brute et la sortie un texte numérisé. La comparaison des résultats obtenus par différents modèles allant de modèles généralistes multilingues à arabocentrés, en passant par des modèles OCR spécialisés met en lumière les limites des systèmes actuels face aux spécificités des textes scientifiques anciens et permet d'identifier des lignes directrices pour développer des systèmes de reconnaissance automatique de ces manuscrits.

ABSTRACT

Automatic Recognition of Medieval Scientific Arabic Manuscripts : Dataset Creation and Large Language Models Evaluation

The digitization of Arab scientific heritage presents a significant challenge for the analysis of these documents due to the spatial complexity of the writing (cursive script, diacritics, ligatures). Despite recent advances, these manuscripts often remain indecipherable for current artificial intelligence models. In this article, we present an evaluation of state-of-the-art methods on a new manually edited and annotated corpus : *Al-Bīrūnī's al-Qānūn al-Masʿūdī*, a major treatise on mathematical astronomy, comparable to Ptolemy's Almagest and written around 1030. We present the methodology used to construct the corpus, then detail the approach implemented for digitizing this manuscript. The latter takes the form of a unified pipeline based on large language models whose input is a raw image and whose output is a digitized text. Comparing the results obtained by different models, ranging from generalist multilingual models to Arabic-centered models and specialized OCR models, highlights the limitations of current systems when faced with the specificities of ancient scientific texts and allows us to identify guidelines for developing automatic recognition systems for these manuscripts.

MOTS-CLÉS : Reconnaissance de texte manuscrit, Documents historiques arabes, Grand modèle de langue.

KEYWORDS: Handwritten Text Recognition, Arabic Historical Documents, Large Language Models.

1 Motivations

La préservation et l'analyse des documents historiques représentent un enjeu majeur pour la compréhension de l'évolution des sociétés humaines. À l'échelle mondiale, des millions de manuscrits, rédigés dans des langues variées telles que le chinois ancien, le latin ou le français médiéval demeurent inexploités dans les archives et les bibliothèques. Ces documents constituent des sources primaires inestimables renfermant des données cruciales sur la philosophie, les sciences et la littérature. Cependant, l'accès à ce savoir reste limité car, si la numérisation massive permet d'en produire des images, leur contenu textuel demeure souvent inaccessible aux méthodes modernes d'indexation et de fouille de données.

Pour répondre à ce besoin, la communauté scientifique a développé des outils de reconnaissance automatique de l'écriture manuscrite (HTR). Des avancées ont été réalisées pour certaines langues comme le chinois ancien (Zhao *et al.*, 2025), le latin médiéval (Koch *et al.*, 2023) ou le français ancien (Clérice *et al.*, 2026), démontrant la capacité de l'intelligence artificielle à transcrire et structurer des corpus historiques complexes.

Pour d'autres langues comme l'arabe, la reconnaissance automatique du patrimoine manuscrit connaît un certain retard. Estimé à plus de trois millions de volumes, ce corpus représente l'une des traditions écrites les plus vastes et les plus riches de l'histoire. Si des travaux récents ont permis des progrès sur la reconnaissance des documents imprimés de la littérature arabe (Bhatia *et al.*, 2024) et sur des manuscrits littéraires et administratifs (Chan *et al.*, 2025), les manuscrits scientifiques médiévaux restent, comme le soulignent Sommerschield *et al.* (2023), un « angle mort » dans les recherches actuelles en humanités numériques. Ceci est particulièrement saillant lorsqu'il s'agit de manuscrits scientifiques de l'âge d'or islamique au cours duquel des figures intellectuelles majeures ont émergé. Leurs travaux ont fondé une grande partie de la science moderne. Parmi ces scientifiques, Abū Rayhān al-Bīrūnī (973-1048) occupe une place prépondérante. Savant polymathe, il a produit une œuvre encyclopédique couvrant des domaines aussi variés que les mathématiques, l'astronomie, la géographie, l'indologie et l'histoire. Nous nous intéressons plus spécifiquement à son traité magistral d'astronomie, *al-Qānūn al-Masᶜūdī*, rédigé vers 1030.

Cet travail s'inscrit dans un projet dont l'ambition dépasse la simple transcription d'une œuvre. Nous visons à développer une méthode robuste pour la reconnaissance automatique de manuscrits scientifiques arabes médiévaux, adaptée à leurs spécificités (multi-polices, chiffres, etc.) et généralisable à des corpus similaires. Dans cette perspective, notre article apporte trois contributions principales :

1. **La constitution d'un premier corpus de référence**, qui compile et aligne six manuscrits du chapitre 6 de *al-Qānūn al-Masᶜūdī*. Cette vérité terrain est établie à partir de l'édition critique de Loizelet (2021). À notre connaissance, ce corpus manuellement annoté composé de **2 431 lignes** constitue la première ressource dans le domaine de la reconnaissances de manuscrits scientifiques médiévaux arabes. Ce dernier sera mis à disposition à la communauté en cas d'acceptation.

2. **Une évaluation** des performances de différentes types de modèles vision-langage (VLM) pour la numérisation de ce jeu de données.

3. **Un analyse d'erreurs illustrant la difficultés de la tâche** permettant de guider les développements futurs.

Table 1 – Comparaison des datasets HTR arabes historiques et modernes. Les volumes sont donnés en nombre de pages (p) et de lignes (l).

Nom	Manuscripts	Siècles	Volume	Format	Référence
Muharaf	historiques et modernes	XIX-XXI	1 644 p. 36 311 l.	PAGE-XML JSON	(Saeed *et al.*, 2024)
Agapet	arabes chrétiens	IX-XVIII	831 p.	PAGE-XML	(Ibrahim, 2024)
Historical Arabic HTR	islamiques	Variés	40 p. 640 l.	docx	(Najam & Faizullah, 2024)
RASM	historiques variés	IX-XIX	120 p. 2 613 l.	PAGE-XML	(Clausner *et al.*, 2018)
RASAM	en écriture maghrébine	X-XVIII	300 p. 7 540 l.	PAGE-XML	(Vidal-Gorène *et al.*, 2021)
Iskandar	Roman d'Alexandre	XVIII-XIX	297 p.	PAGE-XML	(LiPA & Huma-Num, 2024)

2 Reconnaissance automatique des manuscrits arabes

2.1 Corpus existants

La performance des systèmes de reconnaissance automatique de l'écriture manuscrite (*Handwritten Text Recognition* ou HTR) repose avant tout sur la qualité et la pertinence des données d'entraînement. Pour l'arabe, plusieurs corpus ont été constitués à partir de sources textuelles variées (journaux, lettres, manuscrits) comme Muharaf (Saeed *et al.*, 2024), ou via la collecte d'écritures manuscrites modernes produites par des volontaires, comme KHATT (Mahmoud *et al.*, 2012).
Notre étude porte sur les manuscrits historiques. Comme l'illustre le Table 1, la plupart des datasets existants recoupent notre période d'intérêt (entre le X^{e} et le XVII^{e} siècle) et adoptent le format standard PAGE-XML (*Page Analysis and Ground-truth Elements*), introduit par Pletschacher & Antonacopoulos (2010) (voir Annexe 7), facilitant ainsi l'analyse de la structure des pages. Néanmoins, une analyse plus fine de leur contenu révèle des limitations pour notre cas d'usage :

Domaine spécifique. Les corpus se composent majoritairement de textes religieux (chrétiens pour Agapet, islamiques pour *Historical Arabic HTR*), textes littéraires (comme Iskandar ou RASAM), ou proposent un mélange hétérogène de domaines (RASM). Aucun n'est dédié aux manuscrits scientifiques.

Variabilité multi-copies. La caractéristique « multi-police » ou la variation d'écriture pour une même œuvre n'est pas prise en compte dans la constitution de ces corpus.

Ce constat met en évidence deux lacunes. La plupart de ces datasets sont conçus pour des tâches génériques et ne sont pas adaptés à des traitements de textes complexes comme ceux du domaine des sciences exactes (astronomie, mathématiques). D'autre part, la prise en compte de la variabilité des styles d'écriture pour un même manuscrit est souvent absente. En effet, pour garantir la stabilité de l'apprentissage et éviter les biais (comme l'*overfitting* sur un seul copiste), les modèles doivent à être exposés à une diversité d'écritures pour un contenu identique.
Ces caractéristiques, spécialisation scientifique et la variabilité intra-œuvre, ont guidé la création de notre corpus que nous détaillons dans la section suivante.

3 Présentation de la tâche

Notre étude vise à évaluer la capacité des systèmes d'intelligence artificielle actuels à transcrire des manuscrits scientifiques arabes médiévaux. Notre approche repose sur la constitution d'un jeu de données, alignant six versions manuscrites différente du chapitre 6 d'*al-Qānūn al-Mas^{c}ūdī* avec leur édition critique de référence. L'objectif est de mesurer le taux d'erreur brut, de comparer les performances des VLMs, et d'analyser les erreurs spécifiques à ce type de document.

3.1 Corpus

La transcription de ce corpus comporte plusieurs difficultés techniques.

Multigraphie et variabilité des copistes. Le corpus couvre huit siècles de transmission manuscrite. Chaque copie possède son propre style calligraphique (du *naskh* au *nastaliq* [1]), ses propres conventions graphiques et son niveau de lisibilité. La difficulté pour les modèles est de généraliser à partir de cette diversité sans sur-apprentissage de l'écriture de scribes particuliers.

Terminologie scientifique. Contrairement aux textes littéraires, le *Qānūn* utilise la terminologie de l'astronomie et des mathématiques ainsi que des tournures syntaxiques rares. Cette spécificité peut poser problème pour des modèles de langue entraînés sur des corpus en arabe moderne standard.

Mise en page. Les manuscrits originaux contiennent des annotations dans les marges, des tables astronomiques et des diagrammes. Nous les avons exclus de notre corpus qui contient uniquement le **texte principal** [2]. Les modèles doivent isoler ce flux textuel central et ignorer les éléments structurels environnant.

Paléographie. L'écriture du corpus diffère radicalement de celle en arabe moderne (Figure 1b) et de l'arabe imprimée (Figure 1a). On note l'usage fréquent de l'élongation (*kashida* / *tatweel* ; cf. Figure 1c), des ligatures complexes, et une variation dans la représentation des chiffres (cf. Figure 1c).

État des manuscrits. La dégradation physique des supports (taches d'humidité, encre effacée, déchirures) introduit un bruit visuel important que les modèles doivent apprendre à filtrer.

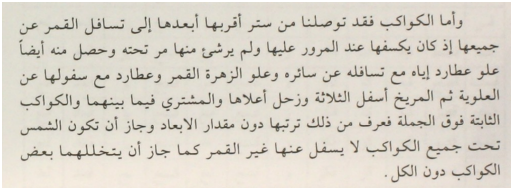
3.2 Hypothèses

Notre évaluation est guidée par trois hypothèses.

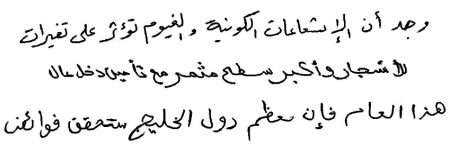
Effets de la taille des modèles et des contextes. Notre première hypothèse est que les modèles vision-langage (*vison language models* ou VLMs) généralistes de très grande taille comme Gemini 3 Flash et Qwen3.5-VL sont capables de transcrire une part de texte significativement plus importante de chaque manuscrit. Le nombre de leurs paramètres et la taille de leurs fenêtres de contexte leur permettent de traiter des images à très haute résolution sans perte d'information globale.

Supériorité des VLMs arabocentrés. Les modèles spécifiquement pré-entraînés ou ajustés sur des corpus de textes et d'images en langue arabe devraient démontrer une reconnaissance

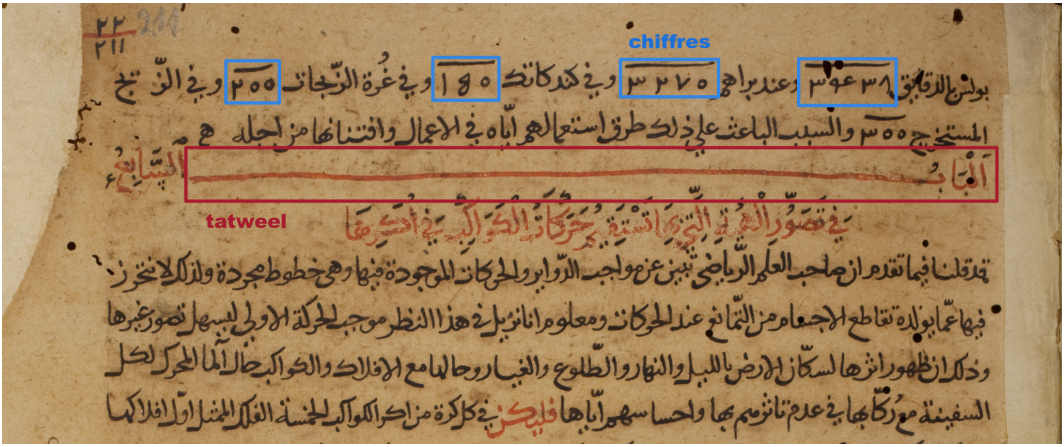
1. https://en.wikipedia.org/wiki/Arabic_script
2. les éléments non-textuelles tel que les diagrammes et les tables ne sont pas pris en compte dans cette phase préliminaire d'évaluation.


وأما الكواكب فقد توصلنا من ستر أقربها أبعدها إلى تسافل القمر عن جميعها إذ كان يكسفها عند المرور عليها ولم يرشئ منها مر تحته وحصل منه أيضاً علو عطارد إياه مع تسافله عن سائره وعلو الزهرة القمر وعطارد مع سفولها عن العلوية ثم المريخ أسفل الثلاثة وزحل أعلاها والمشتري فيما بينهما والكواكب الثابتة فوق الجملة فعرف من ذلك ترتيبها دون مقدار الابعاد وجاز أن تكون الشمس تحت جميع الكواكب لا يسفل عنها غير القمر كما جاز أن يتخللهما بعض الكواكب دون الكل.

(a) Arabe imprimé moderne *Qānūn* (Ed02)



(b) Manuscrit arabe moderne standard (Mahmoud *et al.*, 2012)



(c) Extrait du corpus *Qānūn* (Ms01).

Figure 1 – Exemples de difficultés présentes dans les manuscrits (cursivité, chiffres, *tatweel*).

supérieure des manuscrits. Cette spécialisation linguistique devrait réduire les hallucinations par rapport aux autres modèles.

Hallucination structurelle. Les pages où la densité de texte est extrême induisent un taux d'hallucination plus élevé, notamment pour les tâches *zero-shot*. Dans ces zones de confusion visuelle, les modèles pourraient s'appuyer de façon disproportionnée sur ses connaissances internes et privilégier la génération d'un texte grammaticalement correct à la retranscription exacte de l'image.

4 Jeu de données

La qualité des données d'entraînement et d'évaluation est un facteur déterminant pour la réussite d'un système HTR sur des écritures anciennes. Notre objectif est donc de constituer un corpus aligné, permettant une évaluation robuste des différent systèmes HTR et préparant des expériences de fine-tuning.

Table 2 – Détails des six manuscrits du corpus.

Sigle	Référence	Date	Style	Statistiques			
				Pages	Lignes	Mots	Caractères
Ed01	(Al-Bīrūnī, 1954 1956)	1956	Naskh moderne	10	189	2 602	13 237
Ed02	(Al-Bīrūnī, 2002)	2002	Naskh moderne	8	212	2 594	13 176
Ms01	(Al-Bīrūnī, 1167)	1167	Naskh	5	159	2 598	13 185
Ms02	(Al-Bīrūnī, 1834)	1834	Nasta'liq	7	180	2 584	13 124
Ms13	(Al-Bīrūnī, 1174)	1174	Naskh	5	146	2 562	13 003
Ms17	(Al-Bīrūnī, 1108)	1108	Naskh	4	145	2 581	13 104

4.1 Présentation du corpus

Notre étude repose sur un ensemble de six manuscrits du chapitre 6 d'*al-Qānūn al-Masᶜūdī*, sélectionnés pour leur représentativité historique et la diversité de leurs provenances (Table 2). Ces copies, conservées dans des institutions internationales prestigieuses (BNF, British Library...), couvrent une période allant du XII^{e} au XX^{e} siècle, offrant ainsi un panorama de l'évolution des styles calligraphiques scientifiques.
Le socle de ce dataset est l'édition critique établie par Loizelet (2021). Ce travail philologique est une transcription du manuscrit qui synthétise les six versions pour reconstruire un texte « idéal », corrigé et amendé selon le contexte scientifique. Cette édition critique sert de pivot pour la génération de la vérité terrain de chaque copie.

4.2 Création du dataset

Le jeu de données a été constitué à partir de scans haute résolution et du fichier source LaTeX de l'édition critique. Nous avons utilisé un pipeline qui comporte deux étapes : (1) Génération d'un flux textuel propre à chaque copie à partir du fichier source LaTeX de l'édition critique (dépliage automatique des variantes) ; (2) Alignement de ce texte aux scans haute résolution via l'annotation des documents manuscrits en utilisant la plate-forme d'annotation de manuscrits historiques *Calfa Vision*[3] (Figure 2), qui permet d'exporter les annotations et transcriptions au format PAGE-XML. Les annotations sont ensuite converties en paires exploitables pour l'apprentissage et l'évaluation, sous forme de couples `(Image_Ligne, Transcription)` et `(Image_Page, Transcription)` extraits automatiquement des fichiers PAGE-XML.

5 Reconnaissance automatique

Les récents progrès en apprentissage profond ont démontré que les VLMs entraînés sur des données multimodales massives, constituent des candidats prometteurs pour des tâches de reconnaissance visuelle complexes (Heakl *et al.*, 2025b). Contrairement aux systèmes HTR classiques entraînés spécifiquement sur des données manuscrites, ces modèles réalisent la tâche de transcription en se fondant sur leurs compétences linguistiques. Nous évaluons ici leur capacité à transcrire les manuscrits scientifiques arabes médiévaux de notre corpus dans un cadre *zero-shot* afin de tester leur aptitude à transférer leurs connaissances acquises sur des corpus d'entraînement massifs vers notre domaine cible sans exemples de départ.

3. https://vision.calfa.fr

Figure 2 – Interface d'annotation sur *Calfa Vision*.

5.1 Familles de modèles étudiées

VLMs généralistes multilingues. Ces modèles sont conçus pour des tâches multimodales variées sans spécialisation linguistique ou documentaire particulière. (1) **Gemini 3 Flash** (Google DeepMind, 2025) : modèle multimodal généraliste optimisé pour la rapidité et le traitement de longs contextes ; (2) **Qwen-3.5-VL** (Bai *et al.*, 2025) : modèle *open-weights* reconnu pour ses capacités à extraire le texte contenu dans des images complexes.

VLMs arabocentrés. Ces modèles sont orientés vers l'arabe tout en restant généralistes (HTR non spécialisés). (1) **AIN-7B** (Heakl *et al.*, 2025a) : Modèle multimodal *Arabic INclusive* pré-entraîné sur 3,6 millions d'exemples bilingues (arabe-anglais). Le corpus d'entraînement contient 35% de données arabes authentiques. Le modèle couvre une grande variété de tâches, dont la compréhension de documents complexes ; (2) **Peacock-8B** (Alwajih *et al.*, 2024) : Famille de modèles multimodaux de grande taille spécifiquement conçus pour l'arabe. Basé sur l'architecture InstructBLIP, il utilise un encodeur textuel AraLLaMA. Peacock est entraîné sur environ un million de paires image-texte.

OCR-VLMs spécialisés Ces modèles sont fine-tunés pour la reconnaissance de textes. (1) **Mistral OCR** (Mistral AI, 2025) : Système OCR orienté vers la compréhension de documents multilingues, capable de traiter des mises en page complexes (tables, équations, textes multi-colonnes) ; (2) **Qari-OCR-2B** (Wasfy *et al.*, 2025) : VLM finetuné basée sur Qwen2-VL destiné à la reconnaissance de textes arabes ; (3) **CHURRO-3B** (Semnani *et al.*, 2025) : VLM *open-weights* dédié à la reconnaissance de texte historique, entraîné sur des données multilingues de manuscrits anciens dont 2 367 pages des manuscrits arabes.

5.2 Protocole expérimental

Étant donnée la complexité de traitement automatique au niveau ligne, nous proposons d'évaluer en premier le **niveau page** : chaque modèle reçoit l'image complète d'une page sans pré-traitement

et produit la transcription intégrale de son contenu textuel. Les prompts sont rédigés en anglais. Moudjari & Benamara (2025) ont en effet observé que les instructions en anglais tendent à améliorer les performances des VLMs non fine-tunés sur les tâches de classification de textes arabes. Les sorties de chaque modèle sont ensuite alignées avec la vérité terrain correspondante et évaluées par le biais de deux métriques standards du domaine à savoir le **taux d'erreur caractère** (*character error rate* ou CER) et le **taux d'erreur mot** (*word error rate* ou WER). Analyse qualitative des types d'erreurs observées pour le modèle le plus performant est ensuite proposée. Nous avons poses ces paramètres pour l'ensemble des modèles (température $T = 0$, $\texttt{max\_tokens} = 2000$). L'ensemble des expérimentations a été conduit sur les clusters de calcul de la plateforme *OcciData*[4] à l'aide d'un unique GPU NVIDIA RTX 8000 (48 Go VRAM), sous Python 3.12 et CUDA 12.8.

6 Résultats

La Figure 3 ci-dessous offre une visualisation comparative détaillée des résultats exposés dans la Table 3 (voir Annexe). On se limite à Ms01, Ms02, Ms13 et Ms17 les seuls manuscrits anciens de notre corpus (les résultats des autres documents figurent à la Table 3). Afin de faciliter la lecture de ces graphiques (où des barres ascendantes illustrent une meilleure performance), nous représentons la précision au niveau des caractères et des mots sous forme de *Character Accuracy Rate* ($CAR = 1 - CER$) et de *Word Accuracy Rate* ($WAR = 1 - WER$).
Pour contextualiser ces résultats d'un point de vue applicatif, nous nous appuyons sur les travaux de Muehlberger *et al.* (2019), qui soulignent qu'une transcription présentant jusqu'à 20 % de CER (soit un $CAR \geq 80\,\%$) conserve une réelle utilité pratique, notamment pour des tâches automatisées de recherche par mots-clés ou de fouille de textes en humanités numériques. Par conséquent, dans un contexte d'évaluation *zero-shot* sur un corpus historique complexe, nous considérons qu'un modèle atteignant ou dépassant ce seuil de 80 % de CAR témoigne d'une capacité de généralisation remarquable. Les valeurs de CAR négatives, quant à elles, illustrent un taux d'erreur (CER) supérieur à 100 %, caractéristique d'un phénomène d'hallucination sévère.
Àu vu de nos hypothèses de travail (cf. Section 3.2) et la lecture des résultats de la Table 3 et la Figure 3, nous pouvons tiré les conclusions suivantes :

Taille des modèles et des contextes (VLMs généralistes). Les résultats confirment partiellement notre première hypothèse concernant l'avantage conféré par la taille des modèles. Globalement, les VLMs généralistes Gemini 3 Flash et Qwen-3.5-VL obtiennent les meilleures performances globales, avec des CER très bas sur certains documents (1,1% pour Gemini sur Ed02 et 7,2% sur Ed01). Les performances des ces modèles s'observent aussi dans la longueur de leurs prédictions en caractère (L_{Pred}) : le ratio L_{Pred}/L_{GT} [5] reste systématiquement très proche de 1,0 sur les textes imprimés et oscille entre 0,64 et 1,0 pour les manuscrits de difficulté moyenne (Ms01, Ms02). Ces résultats suggèrent que leur fenêtre de contexte large et le nombre de leurs paramètres leur permettent de traiter l'intégralité d'une page haute résolution sans omettre de texte, validant ainsi leur pertinence pour le traitement de documents anciens complexes.

Limites de la spécialisation linguistique (VLMs arabocentrés). À l'inverse, nos résultats infirment l'hypothèse d'une supériorité des modèles spécialisés pour l'arabe. Les VLMs AIN et Peacock affichent globalement les performances les plus faibles sur les manuscrits (Ms01 à Ms17), Peacock

4. https://occidata.irit.fr
5. L_{GT} : longueur de vérité terrain en caractères.

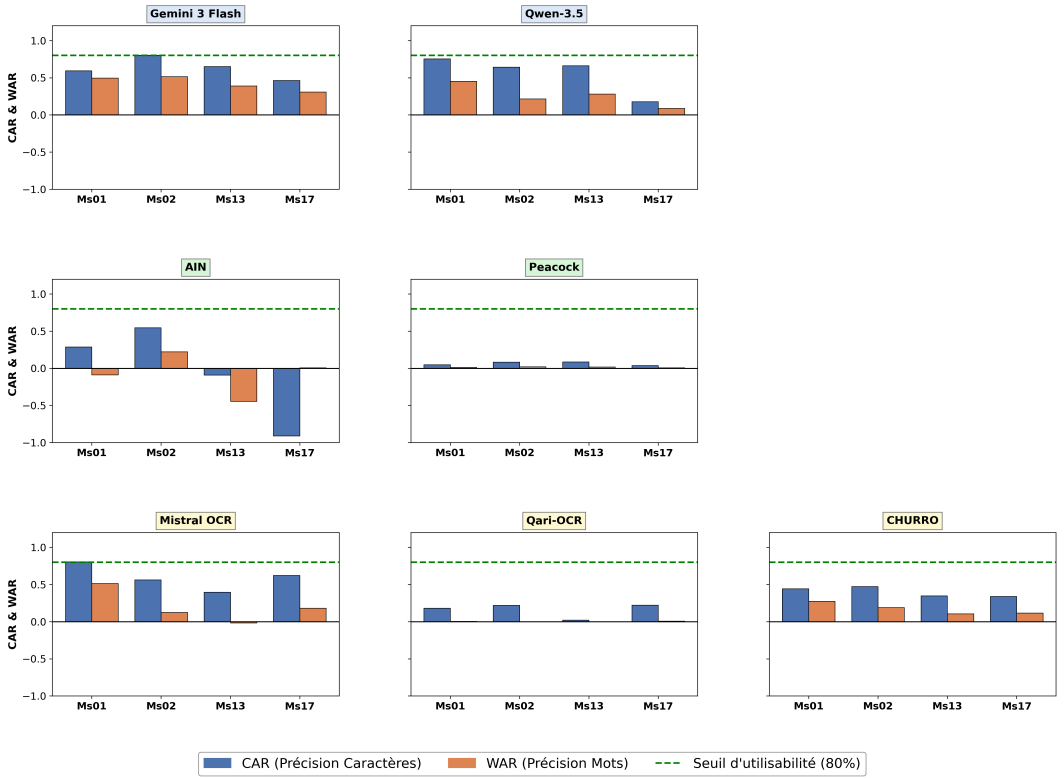


Figure 3 – Évaluation de la précision (CAR et WAR) de chaque modèle sur les manuscrits (Ms01, Ms02, Ms13 et Ms17).

présente un effondrement complet avec des CER compris entre 91 % et 196 %. Sa longueur de texte prédit est drastiquement inférieure à celle de la vérité terrain (générant régulièrement moins de 300 caractères pour des pages qui en contiennent plus de 2 000, soit un ratio $L_{Pred}/L_{GT} < 0,10$). Ces modèles tendent à produire des descriptions génériques de l'image au lieu de la transcription du document. Le pré-entraînement sur des donnés arabes multimodales variées semble insuffisant pour l'acquisition d'une compétence en lecture OCR *zero-shot* continue.

Hallucinations et instabilité des modèles spécialisés. Les OCR-VLMs présentent une efficacité hétérogène. Si Mistral OCR s'impose comme une solution compétitive supérieure aux modèles généralistes avec un CER de 19,4% sur Ms01 et 37,7% sur Ms17, les modèles *open-weights* spécialisés comme Qari-OCR s'avère incapable de maintenir une cohérence globale d'après son comportement sur le Ms17 (sous-génération extrême, prédiction de 92 caractères pour un texte source de 2 686 caractères) . D'autre part, le comportement de AIN sur le manuscrit Ms17 illustre d'ailleurs notre troisième hypothèse sur les risques d'hallucination : le modèle a généré 7 957 caractères pour un texte source de seulement 4 131 caractères, résultant en un CER catastrophique de 190,9%. Cela se traduit par une mauvaise interprétation du *tatweel*, due à la complexité de cette caractéristique face aux capacités actuelles du modèle (voir Annexe, Figure 5).

Impact des caractéristiques calligraphiques des manuscrits. Les fortes variations de performances d'un document à l'autre confirment la complexité inhérente de la reconnaissance automatique des manuscrits scientifiques arabes. Pour les documents les plus denses comme Ms13 et Ms17, la précision chute pour l'ensemble des systèmes évalués. Par exemple, Qwen-3.5 interrompre précocement la transcription de Ms17 ($L_{Pred} = 900$ contre $L_{GT} = 4\,131$). Ces résultats vont dans le sens de notre hypothèse que la confusion visuelle induite par un interligne resserré ou une forte densité d'encre, pousse les modèles à abandonner le tracé au profit de leurs connaissances linguistiques internes, ou à stopper la transcription.

7 Conclusion

Ce travail avait pour objectif principal d'évaluer, dans un contexte *zero-shot*, les performances de différentes catégories de VLMs de l'état de l'art pour la transcription de manuscrits scientifiques arabes. À cette fin, nous avons constitué un jeu de données représentatif de ce vaste corpus. L'évaluation systématique menée à l'aide de métriques standards (CER, WER, L_{Pred}/L_{GT}) a ainsi permis de mettre en lumière les forces et les limites de ces systèmes.

Les résultats confirment que les grands modèles généralistes (tels que Gemini, Qwen) et les modèles spécialisés comme Mistral OCR dominent l'état de l'art. Toutefois, leurs performances déclinent significativement lorsque le degré de complexité augmente (textes denses, présence de *tatweel*, etc.). Par ailleurs, l'échec relatif des modèles fine-tunés existants met en évidence qu'un entraînement sur des données très variées en terme des styles d'écriture et des caractéristiques paléographiques, est indispensable pour garantir une bonne capacité de généralisation.

Afin de surmonter ces obstacles, les futurs travaux devront s'orienter vers une extraction au niveau de la ligne plutôt qu'au niveau de la page entière, ce qui permettrait d'augmenter la résolution spatiale des caractéristiques d'écriture. De plus, l'intégration d'étapes de pré-traitement des images, ainsi que le *fine-tuning* de VLMs *open-weights* sur des corpus historiques arabes spécifiques, constitueront des pistes privilégiées pour rendre ces architectures véritablement robustes.

Références

AL-BĪRŪNĪ A. R. (1108). Al-qānūn al-mas'ūdī fī al-hay'a wa-al-nujūm. Manuscrit, Cote : Arabe 6840. Bibliothèque Nationale de France, Paris.

AL-BĪRŪNĪ A. R. (1167). Al-qānūn al-mas'ūdī. Manuscrit, Cote : Or. quart. 1613. Staatsbibliothek, Berlin.

AL-BĪRŪNĪ A. R. (1174). Al-qānūn al-mas'ūdī. Manuscrit, Cote : Or. 1997. British Library, Londres.

AL-BĪRŪNĪ A. R. (1834). Al-qānūn al-mas'ūdī fī al-hay'a wa-al-nujūm. Manuscrit, Cote : Or. oct. 275. Staatsbibliothek, Berlin.

AL-BĪRŪNĪ A. R. (1954–1956). Al-qānūn al-mas'ūdī. Nizamū d-Dīn éditeur. Hyderabad : Osmania Oriental Publications Bureau.

AL-BĪRŪNĪ A. R. (2002). Al-qānūn al-mas'ūdī. Édition par 'Abd al-Karīm Sāmī Jindī. Beyrouth : Dār al-kutub al-'ilmīyah.

ALWAJIH F., NAGOUDI E. M. B., BHATIA G., MOHAMED A. & ABDUL-MAGEED M. (2024). Peacock : A Family of Arabic Multimodal Large Language Models and Benchmarks. arXiv :2403.01031 [cs], DOI : 10.48550/arXiv.2403.01031.

BAI S., CHEN K., LIU X., WANG J., GE W., SONG S., DANG K., WANG P., WANG S., TANG J., ZHONG H., ZHU Y., YANG M., LI Z., WAN J., WANG P., DING W., FU Z., XU Y., YE J., ZHANG X., XIE T., CHENG Z., ZHANG H., YANG Z., XU H. & LIN J. (2025). Qwen2.5-vl technical report.

BHATIA G., NAGOUDI E. M. B., ALWAJIH F. & ABDUL-MAGEED M. (2024). Qalam : A Multimodal LLM for Arabic Optical Character and Handwriting Recognition. arXiv :2407.13559 [cs], DOI : 10.48550/arXiv.2407.13559.

CHAGUÉ A. (2022). eScriptorium : une application libre pour la transcription automatique des manuscrits. *Arabesques*, **107**, 25. DOI : 10.35562/arabesques.3100.

CHAN A., MIJAR A., SAEED M., WONG C.-W. & KHATER A. (2025). HATFormer : Historic Handwritten Arabic Text Recognition with Transformers. arXiv :2410.02179 [cs], DOI : 10.48550/arXiv.2410.02179.

CLAUSNER C., ANTONACOPOULOS A., MCGREGOR N. & WILSON-NUNN D. (2018). Icfhr 2018 competition on recognition of historical arabic scientific manuscripts – rasm2018. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, p. 471–476. DOI : 10.1109/ICFHR-2018.2018.00088.

CLÉRICE T., BAWDEN R., GLAISE A., PINCHE A. & SMITH D. (2026). Pre-Editorial Normalization for Automatically Transcribed Medieval Manuscripts in Old French and Latin. arXiv :2602.13905 [cs] version : 1, DOI : 10.48550/arXiv.2602.13905.

COLUTTO S., KAHLE P., HACKL G. & MÜHLBERGER G. (2019). Transkribus. a platform for automated text recognition and searching of historical documents. In *2019 15th International Conference on eScience (eScience)*, p. 463–466. DOI : 10.1109/eScience.2019.00060.

GOOGLE DEEPMIND (2025). Gemini 3 flash : Best for frontier intelligence at speed. https://deepmind.google/models/gemini/flash/.

HEAKL A., GHABOURA S., THAWKAR O., KHAN F. S., CHOLAKKAL H., ANWER R. M. & KHAN S. (2025a). AIN : The Arabic INclusive Large Multimodal Model. arXiv :2502.00094 [cs], DOI : 10.48550/arXiv.2502.00094.

HEAKL A., SOHAIL A., RANJAN M., HOSSAM R., AHMAD G. S., EL-GEISH M., MAHER O., SHEN Z., KHAN F. & KHAN S. (2025b). KITAB-Bench : A Comprehensive Multi-Domain Benchmark for Arabic OCR and Document Understanding. arXiv :2502.14949 [cs], DOI : 10.48550/arXiv.2502.14949.

IBRAHIM H. (2024). Agapet : Advanced HTR for Christian Arabic Manuscripts. Accessed : 2026-02-18, DOI : 10.5281/zenodo.14382310.

KOCH P., NUÑEZ G. V., GARCES ARIAS E., HEUMANN C., SCHÖFFEL M., HÄBERLIN A. & ASSENMACHER M. (2023). A tailored handwritten-text-recognition system for medieval Latin. In A. ANDERSON, S. GORDIN, B. LI, Y. LIU & M. C. PASSAROTTI, Éds., *Proceedings of the Ancient Language Processing Workshop*, p. 103–110, Varna, Bulgaria : INCOMA Ltd., Shoumen, Bulgaria.

LIPA & HUMA-NUM (2024). Iskandar dataset. https://gitlab.huma-num.fr/lipa/iskandar. Accessed : 2026-02-18.

LOIZELET G. (2021). *Mesurer et ordonner les astres d'al-Farghānī à al-Bīrūnī : la tradition arabe du Livre des Hypothèses de Ptolémée (IXe-XIe s.). Avec une édition et une traduction française du chapitre X.6 d'al-Qānūn al-Mas'ūdī d'al-Bīrūnī.* Thèse de doctorat, Université Paris Cité.

MAHMOUD S. A., AHMAD I., ALSHAYEB M., AL-KHATIB W. G., PARVEZ M. T., FINK G. A., MÄRGNER V. & ABED H. E. (2012). KHATT : Arabic Offline Handwritten Text Database. In *2012 International Conference on Frontiers in Handwriting Recognition*, p. 449–454. DOI : 10.1109/ICFHR.2012.224.

MISTRAL AI (2025). Mistral OCR. https://mistral.ai/news/mistral-ocr. Accessed : 2026-02-23.

MOUDJARI L. & BENAMARA F. (2025). Are dialects better prompters ? a case study on Arabic subjective text classification. In W. CHE, J. NABENDE, E. SHUTOVA & M. T. PILEHVAR, Éds., *Findings of the Association for Computational Linguistics : ACL 2025*, p. 17356–17371, Vienna, Austria : Association for Computational Linguistics. DOI : 10.18653/v1/2025.findings-acl.892.

MUEHLBERGER G., SEAWARD L., TERRAS M., ARES OLIVEIRA S., BOSCH V., BRYAN M., COLUTTO S., DÉJEAN H., DIEM M., FIEL S., GATOS B., GREINOECKER A., GRÜNING T., HACKL G., HAUKKOVAARA V., HEYER G., HIRVONEN L., HODEL T., JOKINEN M., KAHLE P., KALLIO M., KAPLAN F., KLEBER F., LABAHN R., LANG E. M., LAUBE S., LEIFERT G., LOULOUDIS G., MCNICHOLL R., MEUNIER J.-L., MICHAEL J., MÜHLBAUER E., PHILIPP N., PRATIKAKIS I., PUIGCERVER PÉREZ J., PUTZ H., RETSINAS G., ROMERO V., SABLATNIG R., SÁNCHEZ J. A., SCHOFIELD P., SFIKAS G., SIEBER C., STAMATOPOULOS N., STRAUSS T., TERBUL T., TOSELLI A. H., ULREICH B., VILLEGAS M., VIDAL E., WALCHER J., WEIDEMANN M., WURSTER H. & ZAGORIS K. (2019). Transforming scholarship in the archives through handwritten text recognition : Transkribus as a case study. *Journal of Documentation*, **75**(5), 954–976. DOI : 10.1108/JD-07-2018-0114.

NAJAM R. & FAIZULLAH S. (2024). A scarce dataset for ancient arabic handwritten text recognition. *Data in Brief*, **56**, 110813. DOI : https ://doi.org/10.1016/j.dib.2024.110813.

PLETSCHACHER S. & ANTONACOPOULOS A. (2010). The PAGE (Page Analysis and Ground-Truth Elements) Format Framework. In *2010 20th International Conference on Pattern Recognition*, p. 257–260, Istanbul, Turkey : IEEE. DOI : 10.1109/ICPR.2010.72.

SAEED M., CHAN A., MIJAR A., MOUKARZEL J., HABCHI G., YOUNES C., ELIAS A., WONG C.-W. & KHATER A. (2024). Muharaf : Manuscripts of Handwritten Arabic Dataset for Cursive

Text Recognition. *Advances in Neural Information Processing Systems*, **37**, 58525–58538. DOI : 10.52202/079017-1865.

SEMNANI S. J., ZHANG H., HE X., TEKGÜRLER M. & LAM M. S. (2025). CHURRO : Making History Readable with an Open-Weight Large Vision-Language Model for High-Accuracy, Low-Cost Historical Text Recognition. arXiv :2509.19768 [cs], DOI : 10.48550/arXiv.2509.19768.

SOMMERSCHIELD T., ASSAEL Y., PAVLOPOULOS J., STEFANAK V., SENIOR A., DYER C., BODEL J., PRAG J., ANDROUTSOPOULOS I. & DE FREITAS N. (2023). Machine Learning for Ancient Languages : A Survey. *Computational Linguistics*, **49**(3), 703–747. DOI : 10.1162/coli_a_00481.

VIDAL-GORÈNE C., LUCAS N., SALAH C., DECOURS-PEREZ A. & DUPIN B. (2021). RASAM – A Dataset for the Recognition and Analysis of Scripts in Arabic Maghrebi. In E. H. BARNEY SMITH & U. PAL, Éds., *Document Analysis and Recognition – ICDAR 2021 Workshops*, volume 12916, p. 265–281. Cham : Springer International Publishing. Series Title : Lecture Notes in Computer Science, DOI : 10.1007/978-3-030-86198-8_19.

WASFY A., NACAR O., ELKHATEB A., REDA M., ELSHEHY O., AMMAR A. & BOULILA W. (2025). QARI-OCR : High-Fidelity Arabic Text Recognition through Multimodal Large Language Model Adaptation.

ZHAO S., ZHOU Y., REN Y., CHEN Z., JIA C., ZHE F., LONG Z., LIU S. & LAN M. (2025). Fùxì : A Benchmark for Evaluating Language Models on Ancient Chinese Text Understanding and Generation. arXiv :2503.15837 [cs] version : 1, DOI : 10.48550/arXiv.2503.15837.

Annexe

Format standard de données

La reconnaissance de texte se divise traditionnellement en deux paradigmes : le mode "online", pour la lecture des chèques bancaires, vérification des signatures, traitement des factures de services publics et des demandes dans le secteur des assurances, et le mode "offline", dédié au traitement d'images statiques de documents numérisés, imprimés et écrit par la main. Notre étude s'inscrit exclusivement dans ce second cadre.
Dans l'état de l'art actuel de l'HTR "offline", un standard de représentation s'est conventionné pour l'entraînement des modèles supervisés : le format **PAGE-XML** (Page Analysis and Ground-truth Elements), introduit par le laboratoire PRImA (Pletschacher & Antonacopoulos, 2010). Contrairement aux formats simples (comme TXT), le PAGE-XML offre une description hiérarchique et géométrique précise de la page, indispensable pour traiter des mises en page complexes.
Une structure PAGE-XML typique se décompose en deux niveaux essentiels pour notre tâche :

- **TextRegion :** Ce conteneur de haut niveau délimite les blocs sémantiques de la page (paragraphes, notes marginales, titres). Il est crucial pour distinguer le texte principal des commentaires, fréquents dans les manuscrits scientifiques.
- **TextLine :** Imbriqué dans les régions, cet élément définit la géométrie exacte de chaque ligne via un polygone de coordonnées. Il associe à chaque ligne sa transcription ("Ground Truth") encodée en Unicode (UTF-8). Ce niveau de précision est particulièrement nécessaire pour l'écriture arabe, où les jambages et les diacritiques de lignes adjacentes se chevauchent souvent, rendant les simples boîtes englobantes (bounding boxes) insuffisantes.

L'adoption de ce format standard garantit non seulement la compatibilité avec les outils majeurs (comme Transkribus (Colutto *et al.*, 2019) ou eScriptorium (Chagué, 2022)), mais permet surtout d'entraîner des modèles de segmentation robustes capables de démêler les lignes d'écriture cursive dense.

Détails expérimentaux et résultats complémentaires

Prompt pour l'extraction du texte

```
Below is an image of a full page of text from a historical
    Arabic manuscript.
- Task: Produce a verbatim transcription of all readable Arabic
    text on the image.
- Rules:
    - Preserve the original reading order.
    - Output Arabic text only. Do not translate, summarize,
        explain, or modernize spelling.
    - Preserve what you see: keep characters only if clearly
        present; otherwise don't invent them.
    - Do not hallucinate missing text.
```

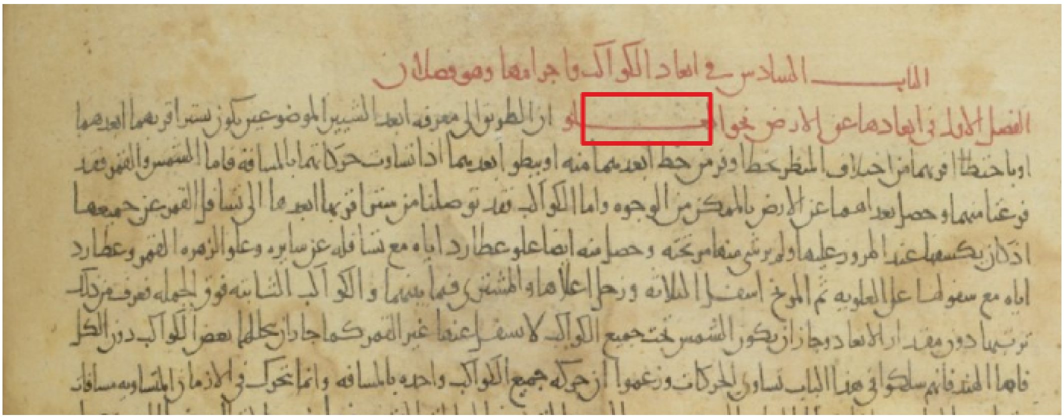
Figure 4 – Prompt rédigé en anglais pour instruire le modèle lors de l'extraction textuelle.

Table des résultats des VLMs

La Figure 3 représentent les performances des transcriptions en CER(Character Error Rate), WER(Word Error Rate) et longueur des prédictions en nombre de caractères (L_{Pred}) pour les six documents étudiés. La longueur de la vérité terrain de chaque document est également donnée en nombre de caractères (L_{GT}), et le rapport L_{Pred}/L_{GT} pour représenter le taux de couverture des documents.

Analyse qualitative de sortie de AIN-7B

La Figure 5 illustre une mauvaise interprétation de la part du modèle AIN face à la présence de *tatweel* (caractère d'élongation) dans le mot encadré sur l'image source. Cette confusion typographique perturbe la génération du modèle, entraînant une hallucination sous la forme d'une répétition anormale et continue du *tatweel* en sortie.



Impact du *tatweel* sur la prédiction

Sortie du modèle :

الماب المسادس ر اعداد الكواكب واجرامها وموفصلان الفصل الاول في اعداده عن الارض فا العـــ

Figure 5 – Illustration de l'erreur liée au *tatweel* (élongation). En haut : extrait de l'image d'entrée (*Ms17*). En bas : la transcription générée par le modèle, caractérisée par une hallucination sévère de caractères d'élongation.

Manuscrit (L_{GT})	Modèle	CER ↓	WER ↓	L_{Pred}	L_{Pred}/L_{GT}
Ed01 ($L_{GT} = 1026$)	Gemini 3 Flash	**0.072**	**0.146**	1085	1.058
	Qwen-3.5	0.077	0.173	1084	1.057
	AIN	0.077	0.249	1068	1.041
	Peacock	0.839	0.978	260	0.253
	Mistral OCR	0.077	0.184	1083	1.056
	Qari-OCR	0.075	0.222	1069	1.042
	CHURRO	0.119	0.346	1081	1.054
Ed02 ($L_{GT} = 1475$)	Gemini 3 Flash	**0.011**	**0.063**	1478	1.002
	Qwen-3.5	0.019	0.104	1473	0.999
	AIN	0.014	0.089	1475	1.000
	Peacock	0.899	0.974	167	0.113
	Mistral OCR	0.012	0.071	1479	1.003
	Qari-OCR	0.195	0.204	1557	1.056
	CHURRO	0.045	0.115	1439	0.976
Ms01 ($L_{GT} = 3506$)	Gemini 3 Flash	0.406	0.504	2243	0.640
	Qwen-3.5	0.244	0.548	3324	0.948
	AIN	0.714	1.089	1879	0.536
	Peacock	0.951	0.988	176	0.050
	Mistral OCR	**0.194**	**0.487**	3499	0.998
	Qari-OCR	0.817	0.995	1086	0.310
	CHURRO	0.556	0.725	1956	0.558
Ms02 ($L_{GT} = 1954$)	Gemini 3 Flash	**0.196**	**0.486**	1958	1.002
	Qwen-3.5	0.357	0.785	1898	0.971
	AIN	0.453	0.779	1481	0.758
	Peacock	0.918	0.980	176	0.090
	Mistral OCR	0.436	0.874	1774	0.908
	Qari-OCR	0.780	1.000	511	0.262
	CHURRO	0.526	0.810	1507	0.771
Ms13 ($L_{GT} = 2686$)	Gemini 3 Flash	0.349	**0.609**	2254	0.839
	Qwen-3.5	**0.339**	0.718	2517	0.937
	AIN	1.090	1.444	4379	1.630
	Peacock	0.913	0.982	270	0.101
	Mistral OCR	0.602	1.016	2379	0.886
	Qari-OCR	0.978	1.000	92	0.034
	CHURRO	0.652	0.894	1523	0.567
Ms17 ($L_{GT} = 4131$)	Gemini 3 Flash	0.537	**0.692**	2251	0.545
	Qwen-3.5	0.821	0.913	900	0.218
	AIN	1.909	0.992	7957	1.926
	Peacock	0.962	0.992	159	0.038
	Mistral OCR	**0.377**	0.817	3844	0.931
	Qari-OCR	0.777	0.993	1086	0.263
	CHURRO	0.660	0.883	1925	0.466

Légende des familles : VLMs Généralistes | VLMs Arabocentrés | OCR-VLMs Spécialisés

Table 3 – Métriques des VLMs exprimés en CER (plus bas = meilleur), WER et le rapport L_{Pred}/L_{GT}.