QAMAR: A New Fully Verified and Accurate Quranic Arabic Morphological Analysis Resource

Sara Faqihi[1], Karim Bouzoubaa[1], Rachida Tajmout[1], Driss Namly[2]
[1]Mohammadia School of Engineers, Mohammed V University in Rabat, Morocco
[2]Institute of African, Euro-Mediterranean and Ibero-American Studies,
Mohammed V University in Rabat, Morocco

sara_faqihi@um5.ac.ma, karim.bouzoubaa@um5r.ac.ma,
rachida.tajmout@emi.um5.ac.ma, d.namly@um5r.ac.ma

Abstract

Several Quranic morphological corpora have been developed to support Arabic linguistic analysis and NLP applications, yet they often lack full coverage, consistency, or manual verification. We present QAMAR, a morphologically oriented, multi-task corpus derived from the Qur'an. This comprehensive, manually verified resource provides a detailed linguistic layer for every Quranic word, including the Modern Standard Arabic (MSA) equivalent, the stem, the lemma, the root, and the part of speech (POS). QAMAR supports multiple NLP tasks, such as normalization, lemmatization, root extraction, and POS tagging, and serves as a gold-standard reference for Quranic and Arabic NLP research, including corpus-to-corpus evaluation and morphological analyzer benchmarking. The paper details QAMAR's annotation framework, verification process, and resource structure, and reports comparative analyses with existing Quranic morphological resources and outputs produced by current large language models (LLMs).

1 Introduction

Arabic Natural Language Processing (ANLP) is a specialized domain within computational linguistics and artificial intelligence. It addresses challenges specific to the Arabic language which features highly inflectional and derivational morphology. The language also has three main textual forms: Classical Arabic (CA), Modern Standard Arabic (MSA), and dialects. Each form presents distinct linguistic features requiring specialized NLP approaches. (Shaalan et al., 2003).

We focus on NLP applications in the religious domain, specifically examining the Quranic text (Bashir et al., 2023). The Holy Quran, revered as the ultimate source of guidance in Islam, is written in CA (Atwell et al., 2010) and exhibits specific morphological characteristics (Al-Sughaiyer and Al-Kharashi, 2004). In addition, the Quran employs the Uthmani script rather than the contemporary Arabic one. For example, the MSA word 'تَبَرَّؤُوا' (tabara~&uwA[1] (here), 'they have abandoned') is written in the Uthmani script as 'تَبَرَّءُوا۟' (tabar~a'uwA@).

Research on Quranic analysis began in the early 2000s with the development of several annotated resources, each employing different approaches to produce specific annotations of the Quran. Some studies focused on Tajwid annotation (Brierley et al., 2019), while others examined prosodic boundaries (Brierley et al., 2012; Sawalha et al., 2014). This work examines Quranic morphological analysis.

Despite these early efforts, current Quranic morphological resources still exhibit several key limitations, as will be detailed later in the paper. First, many words have non-unique analyses, meaning a single word may receive multiple conflicting annotations. Second, some resources lack thorough manual verification, which leads to errors in tags such as POS, lemma, or stem. Third, in several cases the stem and lemma do not correspond correctly, reducing the reliability of morphological information. Fourth, certain annotations are specific to the Uthmani script and do not generalize to Modern Standard Arabic forms. Finally, many resources omit diacritic marks, which are essential for accurate reading, pronunciation, and morphological disambiguation. Addressing these challenges motivates the development of a fully accurate and comprehensive Quranic morphological re-

[1]We use the Buckwalter transliteration system. http://www.qamus.org/transliteration.htm

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 301–312

source.

This paper introduces a morphological resource for the Quran, including the MSA, the stem, the lemma, the root, and the POS classification. Our goal is to provide a reliable reference for researchers and broader audiences interested in accurate Quranic analysis.

To assess the quality of QAMAR as a reference resource, we compare it with all existing Quranic morphological resources, highlighting both the gaps in coverage and the contributions of previous works. Additionally, given the success of LLMs in many Arabic NLP tasks, we also evaluate their performance on Quranic morphological analysis by comparing their outputs against QAMAR.

In the adopted approach, we employ the stem–lemma–root concept, which aligns with traditional Arabic linguistics (سيبويه 1898) and differs from definitions commonly used in other languages. The stem is identified as the word form without proclitics and enclitics. The lemma, also known as the dictionary entry, is defined as the word form without suffixes and prefixes[2]. Finally, the root is defined as the basic form from which word families are derived, typically consisting of three consonants. For example, for the MSA word 'فَجَعَلْناهُم' (fajaEalonAhum, 'So We made them'), the stem is 'جَعَلْنا' (jaEalonA, 'We made'), the lemma is 'جَعَلَ' (jaEala, 'he made'), and the root is 'جعل' (jEl), sometimes also displayed as 'ل ع ج'.

The remainder of the paper is organized as follows. Section 2 reviews previous work on Quranic text analysis. Section 3 outlines a systematic approach for selecting suitable LLMs for Quranic morphological analysis. Section 4 describes the methodology and development of the proposed resource. Section 5 compares QAMAR with related corpora and selected generative LLMs. Finally, Section 6 concludes with a summary of findings and a discussion of their implications.

2 Previous Annotations of The Quran

This section presents the most relevant Quranic annotation efforts, focusing on their methodologies, scope, and limitations to contextualize the development of QAMAR.

The first study by Dror et al. (2004), titled **Morphological Analysis of the Qur'an,** employs a rule-based finite-state toolbox to automatically generate morphological analyses using a tagger with 50 noun rules and 300 verb rules. However, the results lack manual verification, and the system produces a single morphological interpretation for approximately 70% of Quranic words, while the remaining words yield multiple possible analyses. Based on a random sample, the final dataset achieves an estimated F-measure accuracy of 86%.

The next work, **Morphological Annotation of Quranic Arabic** (Dukes and Habash, 2010), introduces a morphologically annotated Quranic resource, hereafter referred to as the Quranic Arabic Corpus (QAC). They first employed an automatic algorithm to generate diacritized morphological features for each word. In a second step, two human annotators reviewed the output, followed by online corrections. The authors reported that approximately 75% of words were correctly analyzed by the automatic algorithm. Nevertheless, QAC differs from QAMAR due to variations in linguistic layer selection and the direct analysis of the Uthmani script.

First, the authors define the stem as the core part of an Arabic word that conveys its fundamental meaning and serves as the base for affix attachment. Compared to our definitions, their stem corresponds to what we define as the lemma, resulting in discrepancies in stem and lemma values in QAC. For example, the MSA word 'فَجَعَلْناهُم' (fajaEalonAhum, 'So We made them') has the stem 'جَعَل' (jaEal, 'made') with prefix 'فَ' (fa) and suffix 'ناهُم' (nAhum). According to our stem definition, the stem is 'جَعَلْنا' (jaEalonA, 'We made'), excluding the proclitic 'فَ' (fa) and the enclitic 'هُم' (hum).

Second, these differences affect lemma definitions as well. Words such as 'عَٰلَمِينَ' (Ea'lamiyna, 'Worlds'), 'ظُلُمَٰتٍ' (Zuluma'tK, 'Darkness'), and 'كَٰفِرُونَ' (ka'firuwna, 'Disbelievers') are treated as lemmas, whereas ac-

[2]The complete list of prefixes and suffixes can be found at: https://arabic.emi.ac.ma/murabaa/

cording to our definitions, they are stems. Their correct lemmas are 'عَالَم' (EaAlam, 'World'), 'ظُلْمَة' (Zulomap, "Dark"), and 'كَافِر' (kaAfir, 'Disbeliever'), respectively. Moreover, the QAC retains several Uthmani-specific annotations. Approximately 10.97% of stems (8588 items) and 11.28% of lemmas (8,829 items) contain the elongation mark 'ـٰ', as in 'قَٰنِتُونَ' (qānituwna, 'obedient') instead of 'قَانِتُونَ' (qānituwna, 'obedient'), and include forms that should be split into two words, such as 'يَٰٓأَرْضُ' (yāʾarḍu) corresponding to 'يَا' (yā, 'O') and 'أَرْضُ' (ʾarḍu, 'land'). More than 2,463 stems (3.14%) and 2,279 lemmas (2.91%) contain the mark 'آ', as in 'سَوَآءٌ' (sawāʾun, 'the same') instead of 'سَوَاءٌ' (sawāʾun, 'the same'). Around 510 stems include the annotation 'ٱ', as in 'ٱلَيِمٌ' (alīmunٱ) instead of 'أَلِيمٌ' (alīmun, 'painful').

Third, about 284 stems contain the annotation 'ـُو', as in 'إِنَّهُۥ' (innahuـُو, 'he') instead of 'إِنَّهُ' (innahu, 'he').

Finally, some roots are incorrectly assigned. For instance, 'شَيْئًا' ($y~'a, 'a thing') and 'سُوءَ' (suw'a, 'wronged') are assigned the roots 'شيا' ($yA) and 'سوا' (swA), instead of the correct roots 'شيء' ($y') and 'سوء' (sw').

The third work, **A New Quranic Corpus Rich in Morphological Information** (Zeroual and Lakhouaja, 2016), introduces a morphologically annotated Quranic resource, hereafter referred to as NQC. The authors adopt a semi-automatic approach: in the first stage, the second version of AlKhalil Morpho Sys (Boudchiche et al., 2017) is used to extract morphological features, including stem, lemma, root, and POS; in the second stage, these features are manually verified.

However, the resource still contains incorrect entries. For example, the MSA form of 'الرَّحْمَٰنِ' was recorded as 'الرَّحْمَنِ', whereas the correct form is 'الرَّحْمَان' (Alr~aHomaAni, 'The Most Merciful'). Another example concerns stem and lemma annotation of the word 'عَلَيْهِمْ' (Ealayohimo, 'those') which are identified as 'علي' and 'عَلَيْ', respectively, while the correct form for both features should be 'عَلَى'.

Furthermore, 99.95% of stem values appear without diacritic marks. For instance, the word 'تَجْرِي' (tajoriy, 'sailed' or 'flow') is assigned the root 'تجر' (tjr), instead of the correct root 'جري' (jry).

Finally, 'فَزَادَهُمْ' (fazaAdahumo, 'adds to their') is annotated as a noun 'اسم', whereas it is a verb 'فعل'.

The fourth resource, **QuranMorph Morphologically Annotated Quranic Corpus** (QMAC), developed by Akra et al. (2025), uses a fully automated approach to create a Quranic resource that assigns each Uthmani word its MSA form, lemma, and POS, along with a detailed classification of 39 POS tags.

In QMAC, some MSA values retain elements of the Uthmani script. For example, the word 'يَٰٓأَيُّهَا' should be separated into two words, 'يَا أَيُّهَا'. Approximately 13.28% of Quranic words have multiple lemma values. The word 'أَنْعَمْتَ' (>anoEamota, 'You have blessed') has four different lemma forms, which are 'أَنْعَمَ', 'أَنْعَمَ عَلَى', 'أَنْعَمَ فِي', and 'أَنْعَمَ لِ'. These forms differ by the addition of prepositions to the base word 'أَنْعَمَ' (>anoEama, 'he have blessed').

POS values are categorized into 39 distinct tags. Some tags include multiple values, with up to four values per tag. For instance, the MSA word 'مَا' (maA, 'do not') has the following POS values 'أداة نفي', 'أداة استفهام', 'حرف', 'اسم موصول', and 'تعجب'.

Despite advances in Quranic morphological analysis, existing resources face notable limitations. Dror et al.'s (2004) rule-based system produces non-unique analyses for roughly 30% of words and lacks manual verification. QAC shows mismatches between stem and lemma definitions, contains Uthmani-specific annotations, and exhibits errors in root identification. NQC preserves incorrect lemmas, roots, and POS tags, and most stem values lack diacritics. QMAC retains Uthmani annotations in MSA forms, assigns multiple lemmas to over 13% of words, and employs complex multi-valued POS tags, while defining only the MSA form, the lemma, and the POS classification for each word.

Together, these limitations underscore the

need for a precise, linguistically consistent, and manually verified Quranic morphological resource.

In summary, among the four presented resources, only QAC and NQC are retained. The Dror resource is excluded due to clear inconsistencies, while QMAC is discarded because it does not incorporate the four target morphological features required for comparison.

As previously discussed, our work must also be compared against the best-performing LLMs for Arabic morphology. The following section is therefore dedicated to identifying the most suitable model for this comparison.

3 Generative LLM Models

Our objective is to evaluate whether LLMs can achieve performance comparable to expert-level Quranic morphological analysis. To this end, we applied a systematic filtering process to select the most suitable model for generating a complete morphological corpus.

The first step selects candidate LLMs based on several criteria, including Arabic language support, architectural diversity, training data sources, model size, and design strategies. This process yields 34 candidate models.

The second step identifies which of these 34 models can perform morphological analysis using an initial benchmark of ten Quranic words.

The third step applies the selected models to a representative Quranic sample to identify the best-performing model.

Finally, the selected LLM is used to generate a complete morphological corpus for the entire Quran.

3.1 LLMs Selection

We selected 34 LLMs based on Arabic support, architecture, and model size, as these factors influence Arabic text processing performance. The selection includes Arabic-specific, bilingual, and multilingual models trained on diverse datasets and architectures. Notable examples include BLOOM (Huber and Niklaus, 2025), FANAR (Abbas et al., 2025), GPT-4o (Hurst et al., 2024) and (Filipovska et al., 2024), Jais (Sengupta et al., 2023), LLaMA 3 and 4 (Meta AI), Qwen2 (Team et al., 2024), SILMA (Al-Rasheed et al., 2025), and StableLM (Alyafeai et al., 2024). These models range from 816M to 13B parameters, covering diverse computational scales.

3.2 First Filtering Process

To identify LLMs capable of morphological analysis, we followed the procedure shown in Figure 1.

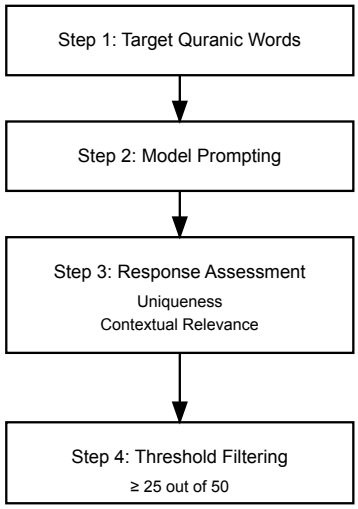

Figure 1: LLM Prompting and Filtering.

In the first step, we used a set of ten target Quranic words representing three categories. The first category includes combined words in their classical forms, such as ‘وَيْكَأَنَّه’ (wayoka>an~ah, ‘Surely’), ‘فَإِلَّم’ (fa<il~am, ‘If + Subject + do not’), ‘وَأَلَّو’ (wa>al~aw, ‘If’), and ‘أَوَآبَاؤُنَا’ (>awa|baA&unaA, ‘And our fathers’). For example, ‘فَإِلَّم’ appears in a combined form, while its corresponding MSA form is split into ‘فَإِن’ and ‘لَم’. The second category comprises the longest Quranic words, including ‘فَأَسْقَيْنَاكُمُوهُ’ (fa>asoqayonaAkumuwhu, ‘which you drink’), ‘فَسَيَكْفِيكَهُم’ (fasayakofiykahum, ‘he will suffice you against them’), and ‘أَنُلْزِمُكُمُوهَا’ (>anulozimukumuwhaA, ‘can we force it upon you’). The third category contains words appearing only once in the Quran, such as ‘أَغْطَش’ (>agoTa$, ‘Gave darkness’), ‘ضِيزَى’ (DiyzaY, ‘an unjust’), and ‘زَنِيم’ (zaniym, ‘mean and infamous’). This selection ensures balanced coverage of three nouns, four verbs, and three particles.

In step 2, candidate LLMs were prompted to perform the morphological analysis for the selected words using a fixed prompting configuration obtained through a preliminary prompt optimization phase. Each prompt was parameterized by four components: a task context, explicit definitions of the targeted morphological features, an optional example, and a constrained task instruction, together with a language parameter. In a first test, we evaluated eight prompt configurations corresponding to different combinations of these components, ranging from zero-shot prompting (Kong et al., 2024) to one-shot prompting (Yoon, 2023). In a second test, we examined the impact of the prompt language by comparing English-only, Arabic-only, and mixed Arabic–English prompts. Based on a binary comparison of the generated outputs against QAMAR values, the final prompt was selected as a one-shot configuration including context, a single example, and explicit task instructions, formulated in English while preserving Arabic for Quranic words.

Step 3 filters LLM responses using two qualitative criteria: *uniqueness* and *contextual relevance*. We define an output as *non-unique* if it fails to commit to a single value per morphological feature. For example, a non-unique output may analyze the word 'يُؤْمِنُونَ' (*yuʾminūna*, "they believe") by proposing two competing root values, such as 'أمن' (*ʾ-m-n*) and 'ؤمن' (*ʾ-m-n*), instead of normalizing and selecting a single canonical root representation. We define an output as *context-irrelevant* if the proposed analysis is inconsistent with the lexical or morphological context in which the word appears within the processed batch. For instance, a context-irrelevant output may assign an incorrect POS, such as analyzing 'نُور' (*nūr*, "light") as a verb where it is consistently treated as a noun. Models failing either criterion are excluded.

For example, when analyzing 'فَأَسْقَيْنَاكُمُوهُ' (fa>asoqayonaAkumuwhu, 'which you drink'), the SILMA[3] model produced an incorrect, non-diacritized stem 'ف-سق-ي' (f-sq-y) and omitted the lemma, root, and POS values (Figure 2). The correct analysis should include the MSA form 'فَأَسْقَيْنَاكُمُوهُ', the stem 'أَسْقَيْنَا' ('we gave water to drink'), the lemma 'أَسْقَى' (>asoqaY, 'he gave water to drink'), the root 'سقي' (sqy), and POS 'فعل' (verb).

Figure 2: SILMA model response.

The pre-selected LLMs are ChatGPT (GPT-4o), MetaAI (LLaMA-4), FANAR, and BardAI (Gemini).

In step 4, a quantitative evaluation applies a binary scoring system, assigning 1 point for correct and 0 for incorrect responses. The acceptance threshold was 25 out of 50, based on 10 words annotated with five morphological features: MSA form, stem, lemma, root, and POS, yielding a maximum score of 50 points.

Results indicate that current LLMs show limited reliability in converting Classical Arabic into corresponding MSA forms. Only ChatGPT (GPT-4o) and MetaAI (LLaMA-4) met the selection criteria, as summarized in Table 1.

Table 1: Platforms Final Score.

Platform	Score (/50)
ChatGPT	**37**
MetaAI	26
BardAI	22
FANAR	12

3.3 Second Filtering Process

Next, we perform a second evaluation using the two selected LLMs on the first 10,000 words of the Quran, representing ~13% of the text. Each word's MSA form was used as input.

Table 2 shows the accuracy of each model for four morphological features: stem, lemma, root, and POS tagging.

[3] https://huggingface.co/silma-ai/SILMA-9B-Instruct-v1.0

Table 2: LLM accuracy on Quranic morphology using QAMAR resource (percentage).

Corpora	Stem	Lemma	Root	POS	Mean
MetaAI (LLaMA-4)	12.01	**25.22**	**44.09**	**94.42**	**43.93**
ChatGPT (GPT-4o)	**25.01**	20.65	21.58	90.65	39.47

Table 2 reports the morphological tagging performance of two LLMs. Rows correspond to models, while columns present accuracy for each morphological feature. MetaAI (LLaMA-4) achieves the highest mean accuracy and is therefore selected to generate the LLM-based corpus (LGC) applied to the entire Quran.

4 Quranic Arabic Morphological Analysis Resource Building

QAMAR (Quranic Arabic Morphological Analysis Resource) is developed as a comprehensive, fully verified resource for the Quran. The construction process involves three main steps. First, we focus on gathering all Quranic words from reliable electronic sources. Next, we perform an annotation by systematically assigning morphological features, ensuring consistency and linguistic accuracy. Finally, we define a representation scheme to structure these annotations, enabling easy access, querying, and computational use.

4.1 Source

We downloaded both the Uthmani and the MSA versions of the Quran from the Tanzil website[4]. The primary text used is the Uthmani script in the Hafs recitation, as it is the most widely adopted version in the Islamic world. The Uthmani script preserves orthographic features such as tatweel (elongation marks) beneath superscript alefs (e.g., الرَّحْمَٰن) and merged forms of words, such as 'يَٰٓأَيُّهَا', which correspond to two separate MSA words: 'يَا' (yaA, 'O') and 'أَيُّهَا' (>ay~uhaA, 'you'). After conversion, the corpus consists of 78,252 MSA words compared to 77,881 Uthmani words, including the Basmalah (البسملة), referring to the Quranic expression 'بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ' (bisomi All~ahi Alr~aHoma~ni Alr~aHiymi, 'In the name of Allah, most benevolent, ever merciful') that opens most chapters.

Table 3 summarizes the transformations applied to derive the MSA morphological feature values from Uthmani script words, along with their corresponding statistics and examples.

A total of **23,211** transformations were applied to derive the MSA feature values from the Quranic text in the Uthmani script.

4.2 QAMAR Annotation: Process and Feature Definitions

The QAMAR corpus is developed using a semi-automatic approach. The SAFAR platform (Software Architecture For ARabic; Jaafar and Bouzoubaa, 2015; Bouzoubaa et al., 2021) provides the core tools for generating the initial version of the corpus, including a stemmer (Namly and Bouzoubaa, 2025), a lemmatizer (Namly et al., 2019), a root tagger, and a light POS classifier (Tnaji et al., 2021) based on a tripartite classification into Noun, Verb, and Particle. This automatically generated resource is explicitly referred to as the SAFAR-generated corpus (SGC) and is later used for comparison with existing resources as well as with the reference QAMAR corpus.

The second phase of the methodology entails manual validation performed by human linguistic experts. The first and second validation iterations are conducted in chunks, each comprising 1,000 Quranic words annotated with their corresponding morphological features, while the third iteration is dedicated to in-context verification. The entire verification process spans three iterations over a period exceeding 1 year and 10 months. Throughout this process, disagreements regarding specific annotation values were resolved through expert discussions during dedicated meetings, in which annotators exchanged analyses and justifications before reaching consensus on a final selected value. The features we consider in our resource are defined as follows:

- **Uthmani:** The Uthmani script is the writing style chosen during the early days of Islam by the third Khalifa (companion of the Prophet PBUH), Uthman Ibn Affan, for writing the words of the Holy Quran and depicting its letters.

[4] https://tanzil.net/

Table 3: Transformations applied to obtain MSA values and their statistics.

Transformation	Count	Example (Uthmani → MSA)
ٱ → ا	183	ٱلصَّلَوٰةَ → الصَّلَاةَ
Deletion of ـٔ	553	وَلَكِنْ → وَلَكِن
ـٰ → ا	5636	ٱلرَّحْمَٰنِ → الرَّحْمَنِ
ٱ → ا	4751	ٱللَّهِ → اللَّهِ
Deletion of ىْ	2	أَفَإِيْن → أَفَإِن
Splitting (+يَا)	360	يَٰٓـَٔادَمُ → يَا + آدَمُ
Splitting (+هَا)	4	هَٰٓأَنتُمْ → هَا + أَنتُمْ
Complex split (example 1)	2	فَمَا لِ + لَّذِينَ → فَمَالِ الَّذِينَ
Complex split (example 2)	2	أَوَ + ءَابَآؤُنَا → أَوَءَابَآؤُنَا
Splitting into 3 parts	1	يَا + ابْنَ + أُمَّ → يَبْنَؤُمَّ
Compound splitting	1	وَأَنْ + لَو → وَأَلَّو
Other splits	2	وَيْ + كَأَنَّ → وَيْكَأَنَّ
Add sukun (ْ)	6803	يُنفِقُونَ → يُنفِقُونَ
Delete shaddah (ّ)	3727	لِّلْمُتَّقِينَ → لِلْمُتَّقِينَ
ة → ت	44	رَحْمَةَ → رَحْمَتَ
وًا → وٍ	10	هُزُوًا → هُزُوا
Short forms expansion	17	تَكُٰنْ → تَكُ
Kasrah (ِ) → sukun (ْ)	281	لَمَنِ → لَمَنْ
Add ى	42	تَرَىٰ → تَرَ
Dammah (ُ) → sukun (ْ)	669	بِكُمْ → بِكُمُ
Add ي	48	دَعَانِ → دَعَانِي
Add ل	73	ٱلَّيْلِ → اللَّيْلِ

- **MSA:** The Arabic words used by contemporary Arabic readers.

- **Stem:** The stem is defined as the main part of an Arabic word after removing any syntactic proclitics or enclitics.

- **Lemma:** Lemmatization involves reducing inflected words to their basic form. For nouns, the lemma corresponds to their masculine singular form (when it exists) without clitics. For instance, the respective lemmas of the words 'ٱلرِّجَال' (Alr~ijaAl, 'men') is 'رَجُل' (rajul, 'a man'). For verbs, the lemma refers to the form of the verb conjugated in the past tense in the third person singular masculine, without clitics. For example, the lemma of 'فَجَعَلْنَاهُم' (fajaEalonAhum, 'So We made them') is 'جَعَلَ' (jaEala, 'he made'). For particles, the lemma is simply the particle itself devoid of clitics, so the lemma of 'عَلَيْهِن' (Ealayohin, 'on them' or 'upon them') is 'عَلَى' (EalaY, 'on' or 'upon'). Moreover, since a lemma has complete meaning only when all diacritics are specified, we have decided to retain all these markings in the lemma, except for the final letter, which is determined by the word's context or its placement within the sentence.

- **Root:** The root consists of the core letters of a trilateral or quadrilateral verb from which the word is derived. For the same examples, the roots are respectively 'رجل' and 'جعل'. Since particles in Arabic lack roots, we represent their root with the '#' symbol.

- **Part of Speech (POS):** In general, a POS tag defines the grammatical category of a specific word. It is well known that tag sets for the Arabic language vary significantly from one project to another. Therefore, to facilitate the comparison of our resource with other existing ones, we

decided to simplify the tag set by dividing it into two levels.

The first level named POS1 tag is selected as the foundation because it represents the basic classification of Arabic words and simplifies comparisons with other teams' resources. The first POS1 level classifies words into Noun (اسم), Verb (فعل), and Particle (حرف).

The second level named as POS2 tag is chosen to address the specificities of the Quran by further categorizing Quranic words according to specific cases within each of the three main categories. For particles, the following tag specifications are used: 'حروف قرآنية' (Quranic particles), which refer to unique particles such as the word 'الٓمٓ' in Surah Al-Baqarah, verse 1. All other particles are tagged as 'حرف'. For nouns, additional tag values include: 'اسم جلالة' for the specific noun 'اللَّه', 'اسم حسن' for names and attributes of Allah such as 'الْقَادِرُ' (AloqaAdiru, 'He has power'); 'اسم إشارة' (demonstrative noun) such as the word 'ذَلِكَ' (*alika, 'This is'), and 'اسم موصول' (relative noun) such as the word 'الَّذِي' (Al~a*iy, 'who'), since such words can be listed in an information retrieval task as stop words but are not categorized in Arabic morphology as particles. The tag 'اسم علم' (proper noun) is used to identify names of people such as 'مُوسَى' (muwsaY, 'Moses') and places such as 'بَكَّةَ' (bak~apa, 'Bakkah'), while 'اسم' (noun) is used for all other types of nouns. For verbs, the second layer of classification includes the tags 'فعل' (verb) and 'اسم فعل' (nominal verb). For example, in Surah Al-Mu'minun 'الْمُؤْمِنُونَ' (Alomu&ominuwna, 'The Believers'), verse 36, the word 'هَيْهَاتَ' is classified as 'اسم فعل' and not 'فعل' (verb) in Arabic linguistics.

The QAMAR resource provides fully diacritized MSA forms, while the stem and lemma forms are diacritized except for the final character, since the diacritic on the last letter depends on the syntactic context, whereas stems and lemmas are intended to be context-independent. For example, 'كِتَابٌ' (kitābun) and 'كِتَابٍ' (kitābin) differ only in their case endings but correspond to the same lexical item; therefore, the final letter of stem and lemma forms is intentionally left unvowelized.

4.3 TEI Encoding for the QAMAR Corpus

We offer the QAMAR resource in TEI[5] (Text Encoding Initiative) format. As an XML-based standard, TEI effectively organizes linguistic corpora, making it well-suited for computational applications. The corpus is organized to capture the morphological features of Quranic words. Figure 3 shows the TEI file structure used in the QAMAR corpus for defining word-level morphological features.

Figure 3: QAMAR corpus XML tree structure illustrating the TEI-based representation of Quranic morphological features.

The provided TEI file structure captures the transformation of Quranic words from Uthmani script to MSA while preserving detailed morphological information. It begins with a metadata section (teiHeader) containing essential bibliographic details, including the title "QAMAR Corpus", publication information, and a source description referencing the original Quranic text in both Uthmani and MSA forms.

[5] https://tei-c.org/

The main content is structured within the <text> and <body> tags, where elements represent Surahs. Each Surah is identified by its name (<head> element) and number (n attribute). Within each Surah, elements further divide the text into Ayas, each marked with an Aya number (n attribute) and containing the full Aya text (<phr> element).

In each Aya, words are encapsulated within <w> elements. The Uthmani form appears as the content of the <w> element, marked with type="Uthmani", while its corresponding MSA version is nested inside a <seg> element labeled with type="msa". The <seg> element contains detailed linguistic features within an <fs> (Feature Structure) element, which organizes morphological attributes in a structured way. The <fs> element includes the stem, lemma, root, and part of speech. The altPos attribute provides an alternative POS for POS2.

This structure provides a comprehensive morphological analysis of every word in the Quranic text, offering insights into both the Uthmani and MSA forms. The structure effectively captures all possible transformation scenarios, including:

Words that remain unchanged between the Uthmani and MSA versions, such as the word 'ٱللَّهِ' in Uthmani, which remains 'اللَّهِ' in MSA.

Words that undergo morphological transformations, such as the replacement of the elongation mark (ـٰ) with Alef (ا), for example, 'ٱلْعَٰلَمِينَ' in Uthmani script becomes 'الْعَالَمِينَ' in MSA.

Cases where a word is split into multiple segments during the conversion process, such as the word 'يَٰلَيْتَنِى' in Uthmani, which is split into 'يَا' and 'لَيْتَنِي' in MSA.

5 Comparison Process & Results Discussion

This section presents a comparative evaluation of existing Quranic morphological resources with respect to the QAMAR reference corpus. We compare QAMAR with four resources: QAC, NQC, SGC, and LGC. Two evaluation strategies are applied: a strict comparison (S) and a broad comparison (B). Results are reported for both the full corpus (Table 4) and a restricted sample (Table 5) to support evaluation at different granularities. Table 4 reports the accuracy of the evaluated corpora over all Quranic words, while Table 5 focuses on a subset of 24,202 words (~30% of the Quran) whose MSA and Uthmani forms match exactly. This restriction is necessary because QAC uses Uthmani orthography as input.

In the strict comparison, a value is considered correct only if it exactly matches the corresponding QAMAR value, after normalization of combined diacritics (e.g., shadda 'ّ' and fatḥa 'َ' order). Any other discrepancy is counted as incorrect. The broad comparison relaxes this condition in specific cases, including differences in final diacritics or undiacritized forms, to account for systematic variations between QAMAR and the other corpora.

Table 4: Full-corpus accuracy of Quranic morphological corpora under strict (S) and broad (B) comparison with QAMAR.

Corpora		MSA	Stem	Lemma	Root	POS	Mean
SGC	S	**100**	**77.44**	**68.70**	57.17	85.75	**77.81**
	B	**100**	85.88	75.15	57.17	85.75	80.79
NQC	S	98.07	0.05	49.57	**93.81**	87.38	65.78
	B	98.07	**87.79**	**85.94**	**93.81**	87.38	**90.60**
QAC	S	45.75	9.99	45.81	84.57	**94.64**	56.15
	B	55.17	41.27	60.38	84.57	**94.64**	67.20
LGC	S	**100**	19.30	31.20	40.84	85.85	55.44
	B	**100**	44.20	59.90	40.85	85.85	66.16

Table 4 summarizes full-corpus accuracy across the 4 evaluated resources. Under strict comparison, the SGC achieves the most accurate stem and lemma annotations, with scores of 77.44% and 68.70%, respectively. NQC achieves the highest root extraction performance at 93.81%, while QAC excels in POS tagging, reaching 94.64%. Overall, SGC remains the closest to QAMAR, with a mean accuracy of 77.81%, followed by NQC (65.78%), QAC (56.15%), and LGC (55.44%).

Under broad comparison, NQC shows strong base-letter consistency, with stem and lemma accuracies of 87.79% and 85.94%, respectively. For example, the form 'يوم' (yawm, 'day') is considered a correct match for the QAMAR stem 'يَوْم' (yawm, 'day'). It also maintains high performance in root extraction (93.81%), while QAC retains the top score for POS tagging (94.6%). Root and POS accura-

cies remain unchanged, as these features are independent of diacritic marks. Both SAFAR and LLaMA-4 achieve 100% in the MSA column, as they were given QAMAR MSA forms as input. The NQC ranks second in MSA accuracy (98.07%). In summary, under broad comparison, NQC is the resource closest to QAMAR across the entire Quran.

Table 5: Accuracy of evaluated corpora on the restricted sample (identical MSA and Uthmani forms) under strict (S) and broad (B) comparison.

Corpora		MSA	Stem	Lemma	Root	POS	Mean
SGC	S	**100**	**86.52**	**77.99**	56.84	85.03	**81.28**
	B	**100**	**91.24**	81.41	56.84	85.03	82.91
NQC	S	99.91	0.00	44.86	**97.61**	76.25	63.73
	B	99.91	89.48	**89.21**	**97.61**	76.25	**90.49**
QAC	S	99.84	19.41	54.30	91.93	**92.11**	71.52
	B	99.84	76.68	78.53	91.93	**92.11**	87.82
LGC	S	**100**	80.87	28.87	41.27	87.77	67.76
	B	**100**	52.25	65.09	41.28	87.77	69.28

Table 5 illustrates the results of the restricted sample. Under strict comparison, the SGC aligns most closely with QAMAR, achieving 81.28% mean accuracy. In the broad comparison, the NQC performs best, reaching 90.49%, reflecting strong base-letter consistency but incomplete diacritization. QAC attains 71.52% under strict comparison and improves by 16.30% under broad comparison, indicating correct base letters but unstable diacritic assignment. For instance, 'قَبْلَ' (qabola, 'before') and 'قَبْلُ' (qabolu) are different stem values that differ in their case endings, although both refer to the same lexical item. The LGC shows the lowest performance, confirming earlier observations that current LLMs still misanalyse nearly half of the Quranic text.

In summary, SGC provides the most accurate fully diacritized outputs; NQC preserves base letters most reliably; QAC shows moderate performance; and LGC remains the least dependable.

6 Conclusion

We introduced QAMAR, a new resource grounded in traditional Arabic linguistic principles, designed to support precise morphological annotation. Our findings underscore the importance of combining automated methods with linguistic validation. QAMAR provides a fully annotated Quranic text with detailed morphological features.

This work also evaluated existing Quranic corpora and LLM-generated outputs, identifying inconsistencies and varying levels of coverage across core morphological features, including the stem, the lemma, the root, and the POS classification.

By offering a linguistically informed and manually validated corpus, QAMAR serves as a reference for the development and evaluation of Arabic NLP tools and supports the training and optimization of current and future LLMs.

7 Limitations

Despite the contributions of this work, several methodological challenges and limitations should be acknowledged.

First, defining a consistent and sufficiently expressive POS label set, particularly at finer levels of granularity, posed a non-trivial challenge and required multiple iterations before convergence.

Second, when LLMs were employed to assist morphological annotation, some outputs were not fully aligned with the target linguistic framework, necessitating prompt refinement and systematic output filtering.

Third, Quranic Arabic morphological analysis is inherently context-dependent. Identical surface forms sharing the same letter sequences and diacritic patterns may receive different morphological analyses depending on verse context. For example, the form أَهْلَكَ (*ʾahlaka*) may function as a noun meaning "your family" or as a verb meaning "he destroyed" when derived from the root هلك (hlk). Such cases require contextual disambiguation and affect multiple morphological feature values.

Finally, as a current limitation rather than a methodological shortcoming, the present release of the corpus focuses on morphological annotation and does not yet incorporate explicit semantic information, such as word-level semantic fields or structured synonym sets. In addition, the translation layer remains limited and does not provide systematic English translations for each Quranic word.

8 Ethics Statement

The authors confirm that this research was conducted in accordance with established ethical standards. No conflicts of interest are declared. All data used in this study were obtained from publicly available sources or generated during the research process.

The QAMAR corpus will be released as an open-source resource for the academic community following the completion of the peer-review process. It will be made publicly available for research and educational purposes.

References

F. T. Abbas, M. S. Ahmad, F. Alam, E. Altinisik, E. Asgari, Y. Boshmaf, S. Boughorbel, S. Chawla, S. A. Chowdhury, F. Dalvi, K. Darwish, N. Durrani, M. G. Elfeky, A. K. Elmagarmid, M. Y. Eltabakh, M. Fatehkia, A. Fragkopoulos, M. Hasanain, M. Hawasly, and 22 others. 2025. Fanar: An arabic-centric multimodal generative AI platform. *arXiv*. ArXiv preprint arXiv:2501.13944.

D. Akra, T. Hammouda, and M. Jarrar. 2025. Quranmorph: Morphologically annotated quranic corpus. *arXiv*. ArXiv preprint arXiv:2506.18148.

R. Al-Rasheed, A. Al Muaddi, H. Aljasim, R. Al-Matham, M. Alhoshan, A. Al Wazrah, and A. AlOsaimy. 2025. Evaluating RAG pipelines for arabic lexical information retrieval: A comparative study of embedding and generation models. In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, pages 155–164.

I. A. Al-Sughaiyer and I. A. Al-Kharashi. 2004. Arabic morphological analysis techniques: A comprehensive survey. *Journal of the American Society for Information Science and Technology*, 55(3):189–213.

Z. Alyafeai, M. Pieler, H. B. Teufel, J. Tow, M. Bellagente, D. Phung, N. Pinnaparaju, R. Adithyan, P. Rocha, M. Zhuravinskyi, and C. Riquelme. 2024. Arabic stable LM: Adapting stable LM 2 1.6b to arabic. *arXiv*. ArXiv preprint arXiv:2412.04277.

E. Atwell, N. Habash, B. Louw, B. A. Shawar, T. McEnery, W. Zaghouani, and M. El-Haj. 2010. Understanding the Quran: a new grand challenge for computer science and artificial intelligence. In *Proceedings of the GCCR'2010 Grand Challenges in Computing Research*. UKCRC.

M. H. Bashir, A. M. Azmi, H. Nawaz, W. Zaghouani, M. Diab, A. Al-Fuqaha, and J. Qadir. 2023. Arabic natural language processing for Qur'anic research: a systematic review. *Artificial Intelligence Review*, 56(7):6801–6854.

M. Boudchiche, A. Mazroui, M. O. A. O. Bebah, A. Lakhouaja, and A. Boudlal. 2017. Alkhalil morpho sys 2: A robust arabic morpho-syntactic analyzer. *Journal of King Saud University-Computer and Information Sciences*, 29(2):141–146.

K. Bouzoubaa, Y. Jaafar, D. Namly, R. Tachicart, R. Tajmout, H. Khamar, H. Jaafar, S. L. Aouragh, and A. Yousfi. 2021. A description and demonstration of SAFAR framework. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 127–134.

C. Brierley, M. Sawalha, and E. Atwell. 2012. Open-source boundary-annotated corpus for arabic speech and language processing. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 1011–1016.

C. Brierley, M. Sawalha, and H. El-Farahaty. 2019. Translating sacred sounds: Encoding tajwīd rules in automatically generated IPA transcriptions of Quranic arabic. In *The Routledge Handbook of Arabic Translation*, pages 46–64. Routledge.

J. Dror, D. Shaharabani, R. Talmon, and S. Wintner. 2004. Morphological analysis of the Qur'an. *Literary and Linguistic Computing*, 19(4):431–452.

K. Dukes and N. Habash. 2010. Morphological annotation of Quranic arabic. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 2530–2536.

E. Filipovska, A. Mladenovska, M. Bajrami, J. Dobreva, V. Hillman, P. Lameski, and E. Zdravevski. 2024. Benchmarking OpenAI's apis and other large language models for repeatable and efficient question answering across multiple documents. In *2024 19th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pages 107–117. IEEE.

T. Huber and C. Niklaus. 2025. LLMs meet Bloom's taxonomy: A cognitive view on large language model evaluations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5211–5246.

A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. J. Ostrow, A. Welihinda, A. Hayes, A. Radford, and 1 others. 2024. Gpt-4o system card. *arXiv*. ArXiv preprint arXiv:2410.21276.

Y. Jaafar and K. Bouzoubaa. 2015. Arabic natural language processing from software engineering to complex pipeline. In *2015 First International Conference on Arabic Computational Linguistics (ACLing)*, pages 29–36. IEEE.

A. Kong, S. Zhao, H. Chen, Q. Li, Y. Qin, R. Sun, X. Zhou, E. Wang, and X. Dong. 2024. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4099–4113, Mexico City, Mexico. Association for Computational Linguistics.

D. Namly and K. Bouzoubaa. 2025. An innovative arabic light stemmer developed using a hybrid approach. *International Journal of Electrical and Computer Engineering (IJECE)*, 15(2):2356–2363.

D. Namly, K. Bouzoubaa, A. El Jihad, and S. L. Aouragh. 2019. Improving arabic lemmatization through a lemmas database and a machine-learning technique. In *Recent Advances in NLP: The Case of Arabic Language*, pages 81–100. Springer.

M. Sawalha, C. Brierley, and E. Atwell. 2014. Automatically generated, phonemic arabic-IPA pronunciation tiers for the boundary annotated Qur'an dataset for machine learning (version 2.0). In *Proceedings of LRE-Rel 2: 2nd Workshop on Language Resource and Evaluation for Religious Texts, LREC 2014 post-conference workshop*, pages 42–47, Reykjavik, Iceland. The University of Leeds.

N. Sengupta, S. K. Sahu, B. Jia, S. Katipomu, H. Li, F. Koto, O. M. Afzal, S. Kamboj, O. A. Pandit, R. Pal, L. Pradhan, Z. M. Mujahid, M. Baali, X. Han, A. Aji, Z. Liu, A. Hock, A. Feldman, J. Lee, and 4 others. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv*. ArXiv preprint arXiv:2308.16149.

K. Shaalan, A. Allam, and A. Gomah. 2003. Towards automatic spell checking for arabic. In *Proceedings of the 4th Conference on Language Engineering, Egyptian Society of Language Engineering (ELSE)*, pages 21–22, Cairo, Egypt.

Sibawayh. 1898. *Kitab Sibawayh*, volume 1. Al-Matba'ah al-Kubra al-Amiriyyah. Arabic edition.

Qwen Team and 1 others. 2024. Qwen2 technical report. *arXiv*. ArXiv preprint arXiv:2407.10671.

K. Tnaji, K. Bouzoubaa, and S. L. Aouragh. 2021. A light arabic POS tagger using a hybrid approach. In *International Conference on Digital Technologies and Applications*, pages 199–208, Cham. Springer International Publishing.

Su-Youn Yoon. 2023. Short answer grading using one-shot prompting and text similarity scoring model. *arXiv preprint arXiv:2305.18638*.

I. Zeroual and A. Lakhouaja. 2016. A new quranic corpus rich in morphosyntactical information. *International Journal of Speech Technology*, 19(2):339–346.