QURAN-MD: A Fine-Grained Multilingual Multimodal Dataset of the Quran

Muhammad Umar Salman
MBZUAI
Abu Dhabi, UAE
umar.salman@mbzuai.ac.ae

Mohammad Areeb Qazi
MBZUAI
Abu Dhabi, UAE
mohammad.qazi@mbzuai.ac.ae

Mohammed Talha Alam
MBZUAI
Abu Dhabi, UAE
mohammed.alam@mbzuai.ac.ae

Abstract

We present QURAN-MD, a comprehensive multimodal dataset of the Quran that integrates textual, linguistic, and audio dimensions at the verse and word levels. For each verse (ayah), the dataset provides its original Arabic text, English translation, and phonetic transliteration. To capture the rich oral tradition of Quranic recitation, we include verse-level audio from 32 distinct reciters, reflecting diverse recitation styles and dialectical nuances. At the word level, each token is paired with its corresponding Arabic script, English translation, transliteration, and an aligned audio recording, allowing fine-grained analysis of pronunciation, phonology, and semantic context. This dataset supports various applications, including natural language processing, speech recognition, text-to-speech synthesis, linguistic analysis, and digital Islamic studies. Bridging text and audio modalities across multiple reciters, this dataset provides a unique resource to advance computational approaches to Quranic recitation and study. Beyond enabling tasks such as ASR, tajweed detection, and Quranic TTS, it lays the foundation for multimodal embeddings, semantic retrieval, style transfer, and personalized tutoring systems that can support both research and community applications. The dataset is available at https://huggingface.co/datasets/Buraaq/quran-audio-text-dataset

1 Introduction

The Quran, the holy book of Islam, occupies a central role in the spiritual, intellectual, and cultural lives of nearly two billion Muslims worldwide. Revealed in Classical Arabic (Fusha), it is not only revered as sacred scripture but also preserved through its oral tradition of recitation. Mastery of recitation is guided by tajweed which is a set of phonological and prosodic rules that ensure the Quran is read with precision, beauty, and correctness. For Muslims, engaging with the Quran involves more than simply reading the text but is an act of devotion that requires both accurate pronunciation and a deep understanding of its meaning.

Despite its origins in Arabic, the Quran is recited and studied across every corner of the globe, including by millions of non-Arabic speakers from diverse linguistic and cultural backgrounds. This global engagement highlights the challenges faced by learners who must navigate not only the Arabic script but also its sounds, phonology, and interpretive meanings. While Islamic scholars, including Ulama (religious scholars), Fuqaha (jurists), and researchers, have made significant contribu-

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MusIML

Table 1: Feature comparison across prominent Quranic datasets. The proposed QURAN-MD is the first to holistically integrate all listed modalities across multiple reciters at both verse and word levels, filling a critical gap in the available resources.

Feature	Dataset				
	Quranic Arabic Corpus [1]	Tanzil Project [4]	Tarteel [5]	QDAT [6]	**QURAN-MD (Ours)**
Text (Arabic)	✓	✓	✓	✓	✓
Morpho-syntactic Data	✓	✗	✗	✗	✓
Multi-Language Translation	✗	✓	✗	✗	✓
Word-level Audio	✗	✗	✗	✓	✓
Verse-level Audio	✗	✗	✓	✓	✓
Multiple Reciters	✗	✗	✓	✗	✓

tions to Quranic studies, the integration of modern computational approaches, particularly artificial intelligence, remains very limited compared to the rapid advances in AI. AI has the potential to illuminate linguistic patterns, aid recitation training, and support both Arabic and non-Arabic learners in mastering tajweed and understanding the Qurans message.

Existing resources, however, often fall short of capturing the Qurans full multimodal nature. Some datasets provide the Arabic text alone, others include translations or transliterations, and a few extend to audio recordings at either the verse or word level. Rarely are these modalities combined in a way that allows fine-grained analysis of both the written and spoken Quran. This fragmentation restricts their utility for advancing AI-driven research in areas such as natural language processing, speech recognition, and text-to-speech synthesis. Table 1 summarizes key features of prior Quranic datasets, where QURAN-MD offers the most comprehensive multimodal coverage.

In this work, we curate and harmonize resources from three data sources to construct QURAN-MD, a comprehensive multimodal dataset of the Quran. Our dataset unifies Arabic text, English translation, and phonetic transliteration at both the verse and word levels, while pairing each word with an aligned audio recording of its pronunciation. We provide high-quality recitations from 32 distinct reciters at the verse level, representing diverse styles and dialectical nuances (see Table 2 in the Appendix). Through careful validation and consistency checks, we ensure the reliability of the textual and audio components. By making this dataset available on Hugging Face, we aim to provide scholars, linguists, and AI researchers with a standardized resource that bridges textual, phonological, and semantic dimensions. This contribution supports the study of Quranic language and recitation traditions. It enables and promotes future work in cutting-edge computational applications at the intersection of speech, language, and Quranic studies.

2 Related Work

There has been growing interest in developing resources for Quranic text and recitation, yet existing efforts typically focus on either textual analysis or audio recordings, rather than providing a fully integrated multimodal resource. Early structured corpora, such as the *Quranic Arabic Corpus* [1], provide full text with detailed morphological and syntactic annotations, supporting parsing, part-of-speech tagging, and grammatical analysis. More recent contributions, including the *MASAQ corpus* [2] and the corpus introduced by [3], expand these annotations and incorporate orthographic variants (Uthmani and Imlaai scripts), Buckwalter and phonetic transliterations, as well as English translations at the verse level. Similarly, the *Tanzil project* [4] offers the Quran aligned with translations in more than 40 languages, supporting multilingual NLP research, though without detailed linguistic or recitational information.

Parallel to these textual resources, several initiatives have sought to capture Quranic recitation across multiple readers. Large-scale audio collections, such as *Tarteel AI EveryAyah*[5] and the *Quran-Recitations Dataset* [7], provide verse-level recordings across numerous reciters, while the *Quranic Audio Dataset* [8] and *RetaSy* [9] focus on non-native reciters with labeled correctness annotations. The *OpenSLR Quran Speech-to-Text corpus* [10] offers complete verse coverage for multiple reciters, and smaller targeted collections, such as *QDAT* [11] and its extensions [12], provide labeled data for Tajweed error detection using deep neural models. Recent Kaggle contributions, including the

Quran Ayat Speech-to-Text dataset [13], *Quran Reciters dataset* [14], *Quran.com Audio dataset* [15], *Quran Recitations for Audio Classification dataset* [16], and the *Comprehensive Quranic Dataset v1 (CQDV1)* [17], offer large-scale verse-level audio recordings from dozens of reciters, but generally provide only Arabic text and audio without consistent transliteration or translation.

Taken together, these resources demonstrate substantial progress in both linguistic annotation and audio collection, yet they remain fragmented: either focusing on text or audio, verse- or word-level data. None combine verse and word-level text, transliteration, English translation, and audio across multiple reciters with consistent cross-modal alignment. QURAN-MD addresses these gaps by integrating verse and word-level text, English translation, transliteration, word-level audio, and verse audio from 32 reciters, providing a unified multimodal resource that supports research across natural language processing, speech technologies, linguistic analysis, and digital Islamic studies.

3 Dataset (QURAN-MD)

The dataset was constructed by aggregating and harmonizing three publicly available sources of Quranic data: (1) a Kaggle dataset of verse-level speech-to-text recordings by 32 reciters [18], (2) the `quranwbw` repository containing aligned word-by-word Arabic text, English translations, and transliterations [19], and (3) the Internet Archive collection of word-by-word Quran audio recordings [20]. To unify these heterogeneous resources, we designed a hierarchical JSON template. At the top level, each `surah` (chapter) is represented by its numerical index (e.g., `"112"` for the 114 surah's) and annotated with its Arabic name, English translation, transliteration, and total verse count. Each surah contains a nested set of `verses` (ayahs), where each verse entry includes the Arabic text, its English translation, phonetic transliteration, paths to reciter-specific verse-level audio, and a list of word-level objects. Word objects store the word Arabic script, English translation, transliteration, and the corresponding aligned audio file. This structure allows seamless integration of textual, transliterated, and auditory modalities across both verse- and word-level granularities.

The preparation pipeline proceeded in four main steps. First, surah-level metadata (surah number, names, and verse counts) was populated into the template. Second, from the Kaggle dataset, verse-level recitations by 32 reciters were aligned with their corresponding verse texts and added under the `audio_ayah_path` field. Third, word-level Arabic text, translations, and transliterations from the `quranwbw` repository were inserted into the template. Fourth, word-level audio files from the Internet Archive collection were matched to individual tokens and linked under the `audio_word_path` field. After populating all layers, automated validation scripts were applied to ensure correctness, checking that every word and verse was paired with its corresponding audio, and identifying and correcting missing or misaligned entries. This pipeline resulted in a complete, consistent, and multimodal representation of the Quran, ready for release as a standardized dataset.

Dataset Format: The dataset is organized in a hierarchical JSON structure, where each surah (chapter) of the Quran is represented as a nested object. This design ensures that verse-level and word-level information is consistently accessible and easily parsable for downstream applications. An example of Surah 112 (*Al-Ikhlas*) is shown in Figure 1 (Appendix), highlighting metadata and audio/text linkage. This structure enables researchers to work seamlessly at both the verse level (for example, analyzing prosody between reciters) and the word level (for example, studying phoneme-grapheme alignment).

```
{
    "surah_name_ar": "SURAH_NAME_IN_ARABIC",
  "surah_name_en": "SURAH_NAME_IN_ENGLISH",
  "surah_name_tr": "SURAH_NAME_PHONETIC",
  "ayah_count": TOTAL_VERSES,
  "ayahs": {
    "001": {
      "ayah_ar": "VERSE_TEXT_IN_ARABIC",
      "ayah_en": "VERSE_TEXT_IN_ENGLISH",
      "ayah_tr": "VERSE_PHONETIC",
      "audio_ayah_path": {
        "reciter_name": "PATH_TO_VERSE_AUDIO"
      },
      "words": [
        {
          "id": "001",
          "word_ar": "WORD_1_IN_ARABIC",
          "word_en": "WORD_1_IN_ENGLISH",
```

```
          "word_tr": "WORD_1_PHONETIC",
          "audio_word_path": "PATH_TO_WORD_1_AUDIO"
        },
        // Additional words continue here...
      ]
    }
    // Additional ayahs continue here...
  }
}
```

4 Future Work

QURAN-MD opens multiple avenues for future research at the intersection of speech processing, natural language processing, and digital Islamic studies. We outline three directions: (A) Speech-only tasks, (B) Multimodal approaches integrating speech and text, and (C) Tools and applications that can support both research and community engagement.

Speech-only Research: A major direction lies in automatic speech recognition (ASR) for Quranic recitation, both at the verse and word level. The audio and verse-level transcripts aligned in the dataset with 32 reciters provide a unique testbed to develop end-to-end ASR models adapted to classical Arabic with the unique prosody and melodic contours in recitations. Evaluations can measure word error rate and segmentation accuracy across reciters, while challenges include modeling tajweed-driven prosody (madd, pauses, emphatics) and script-pronunciation mismatches.

Multimodal Speech-Text Research: The full potential emerges when the modalities are combined. The dataset can drive progress in forced alignment and phoneme-level segmentation, producing high-quality timestamps for words and phonemes that benefit pronunciation modeling. It enables pronunciation tutoring systems tailored to Quranic recitation, which automatically detect tajweed issues and mispronounced words, providing corrective audio feedback and guidance on proper melodic and rhythmic patterns.

Another promising direction is Quranic TTS that respects tajweed rules and recitation styles. Leveraging parallel recordings from 32 reciters, the dataset enables the development of personalized TTS systems that provide corrective feedback for tajweed and Quranic melody in a users own voice, facilitating accurate imitation and learning. In addition, style transfer approaches can allow users to reproduce the prosody and melodic patterns of renowned reciters while preserving tajweed correctness, supporting both educational applications and expressive recitation synthesis.

Tools, Applications, and Community Resources: QURAN-MD also enables the development of advanced tools and applications that bridge research and community engagement. Multimodal embeddings of Quranic text and audio can support semantic retrieval of specific verses, style-aware search, and clustering of reciters, while facilitating the creation of robust tutoring systems for pronunciation and tajweed. Beyond machine learning applications, the dataset can be integrated into retrieval platforms, APIs, and interactive visualization tools for scholars, educators, and students, providing aligned verse- and word-level corpora that support linguistic, phonological, and prosodical research. Such resources can foster broader accessibility and engagement with the Qur'an and its studies.

5 Conclusion

We curated QURAN-MD, a comprehensive multimodal dataset of the Quran from three different sources that integrates textual and audio information with linguistic annotations at the verse- and word-level, encompassing 32 distinct reciters to capture the rich diversity of Quranic recitation. By aligning Arabic text, English translations, phonetic transliterations, and fine-grained audio recordings at both verse- and word-levels, the dataset enables a wide range of research applications, including automatic speech recognition, tajweed detection, pronunciation tutoring, personalized Quranic TTS, style transfer, and prosody modeling. Furthermore, it supports the development of multimodal embeddings for semantic retrieval, style-aware search, and reciter clustering, as well as interactive tools for scholars, educators, and learners. This resource bridges text and audio modalities at scale, providing a unique platform to advance computational modeling and applications in Quranic research.

6 Appendix

Beyond ASR, the dataset can enable automatic tajweed detection and error classification, an area currently limited to small corpora like QDAT. In addition, it provides an opportunity to investigate the explainability of tajweed errors or mispronounced words, allowing models to highlight specific phonological or tajweed-related deviations and provide interpretable feedback for learners and researchers. Leveraging recitation at a word level, models could detect rules such as *ghunnah, idgham, or madd*, with precision/recall measured against expert annotations. Similarly, the dataset supports reciter identification and style analysis, using embeddings to group readers by dialect or melodic tendencies, and prosody modeling, which quantifies melodic contours and pause structures for downstream use in Quranic TTS.

Example: Surah 112 (Al-Ikhlas): Figure 1 shows a simplified representation of the dataset format for the 112th surah (Al-Ikhlas). This example illustrates how each surah contains its metadata, how verses are indexed, and how both verse- and word-level audio/textual information are linked.

Figure 1: Example of format of Surah 112 (Al-Ikhlas) in the Dataset.

Table 2: Overview of the QURAN-MD dataset.

Category	Attribute	Statistics / Details
Corpus Size	Surahs	114
	Ayahs	6,236
	Words	∼77.8k
Audio	Reciters	32 (diverse styles)
	Verse-level Audio	∼665 hours
	Word-level Audio	∼22 hours
Modalities	Text	Arabic, English, Transliteration
	Audio	Verse- and Word-level recordings

References

[1] A. Dukhovny and E. Atwell. The quranic arabic corpus: An annotated linguistic resource for the holy quran. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, 2009.

[2] Majdi Sawalha, Fahad Alshammari, Salam Khalifa, and Nizar Habash. Masaq: Morphologically-analyzed and syntactically-annotated quran corpus. *Language Resources and Evaluation*, 2024.

[3] A. Nashir and colleagues. Quranic corpus with orthographic, morphological, and syntactic layers. Mendeley Data, V1, 2025.

[4] Z. Zarrabi-Zadeh. Tanzil: Verified quran text and translations. `https://tanzil.net`, 2008.

[5] Taha Khan et al. Tarteel: Crowdsourcing a large dataset of quran recitations for speech recognition. In *Proceedings of Interspeech*, 2021.

[6] Hanaa Mohammed Osman, Ban Sharief Mustafa, and Yusra Faisal. Qdat: A data set for reciting the quran. *International Journal on Islamic Applications in Computer Science And Technology*, 9(1), 2021. Provides audio annotated for Tajweed rules (Al Mad, Ghunnah, Ikhfaa).

[7] MohamedRashad. Quran-recitations dataset. Hugging Face Dataset, 2024. Verse-level audio + fully diacritized Arabic text, multiple reciters.

[8] Raghad Salameh, Mohamad Al Mdfaa, Nursultan Askarbekuly, and Manuel Mazzara. Quranic audio dataset: Crowdsourced and labeled recitation from non-arabic speakers. In *arXiv preprint*, 2024. Crowdsourced labeling of correct/incorrect recitations from non-native reciters.

[9] Ahmad Salameh and colleagues. Retasy: A crowdsourced quran recitation dataset from non-arabic speakers. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, 2024.

[10] OpenSLR. Quran speech-to-text (slr132). `https://www.openslr.org/132`, 2017.

[11] A. Osman and colleagues. Qdat: Quranic dataset for automatic tajweed error detection. In *Proceedings of the IEEE International Conference on Arabic Language Processing (ICALP)*, 2021.

[12] Dim Shaiakhmetov, Gulnaz Gimaletdinova, Selcuk Cankurt, and Kadyrmamat Momunov. Evaluation of the pronunciation of tajweed rules based on dnn as a step towards interactive recitation learning. 2025. Uses QDAT dataset for classifying Tajweed rules: Al-Mad, Ghunnah, Ikhfaa.

[13] BigGuyUbuntu. Quran ayat speech-to-text dataset. Kaggle, `https://www.kaggle.com/datasets/bigguyubuntu/quran-ayat-speech-to-text`, 2023.

[14] Omar Tariq. Quran reciters dataset. Kaggle, `https://www.kaggle.com/datasets/omartariq612/quran-reciters`, 2025.

[15] Abdo3id. Quran.com audio dataset. Kaggle, `https://www.kaggle.com/datasets/abdo3id/quran-com-audio-files`, 2022.

[16] Mohammed Alrajeh. Quran recitations for audio classification. Kaggle, `https://www.kaggle.com/datasets/mohammedalrajeh/quran-recitations-for-audio-classification`, 2021.

[17] QuranicDataset. Comprehensive quranic dataset v1 (cqdv1). Kaggle, `https://www.kaggle.com/datasets/quranicdataset/quranic-dataset-v1`, 2024.

[18] Bigguyubuntu. Quran ayat speech-to-text dataset. `https://www.kaggle.com/datasets/bigguyubuntu/quran-ayat-speech-to-text`, 2021. Accessed: 2025-09-14.

[19] Qazasaz. Quran word-by-word repository. `https://github.com/qazasaz/quranwbw`, 2018. Accessed: 2025-09-14.

[20] Internet Archive. Quran word-by-word audio collection. `https://archive.org/details/quran-wordbyword`, 2017. Accessed: 2025-09-14.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly describe the papers primary contribution: the construction of a multimodal Quran dataset that integrates text, translation, transliteration, and aligned audio across multiple speakers at both the verse and word levels. The claims are descriptive rather than exaggerated, and they align with what the dataset actually provides. The scope is well defined as a resource paper, with potential applications (e.g., ASR, TTS, NLP, and tajweed analysis) mentioned as opportunities for future work rather than as current achievements. This ensures that the abstract and introduction accurately reflect the contributions without overstating results or generalization.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [NA]

   Justification: The paper does not explicitly include a discussion of limitations. While some scope boundaries can be inferred from the description of the dataset, a dedicated limitations section is not provided. As such, the work is presented without a clear reflection on assumptions, potential shortcomings, or constraints.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper's contribution is the creation and description of a new dataset. It does not introduce any theoretical results, theorems, or mathematical proofs that would require such justification.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper's main contribution is a dataset, not experimental results in a traditional sense. It clearly discloses the methodology for the dataset's creation by identifying the three public sources used and detailing the four-step pipeline for data aggregation, harmonization, and validation. This is sufficient to understand how the dataset was constructed.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper explicitly states that "The dataset curation code will be released, and the dataset itself will be made publicly available". It further specifies that the dataset will be released on Hugging Face, a standard platform for sharing research assets, ensuring community access.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: The paper does not conduct experiments that involve model training or evaluation. Its focus is on the curation of a dataset, so details regarding experimental settings like hyperparameters or data splits are not applicable.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: This is a resource paper that describes a dataset. It does not report quantitative experimental results that would require statistical significance testing or error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The paper's contribution is a dataset and does not involve computational experiments like model training. Therefore, information about compute resources is not applicable.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

Justification: The research involves the curation of publicly available data to create a resource for academic and community use. The work is handled with respect for the cultural and religious significance of the source material (the Quran) and aims to provide a tool for educational and research advancement, aligning with ethical guidelines.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper extensively discusses positive societal impacts, such as developing educational tools for recitation , supporting linguistic analysis, and fostering broader community engagement with Quranic studies. While it does not explicitly detail negative impacts, the work, a structured dataset of a public religious text, does not present obvious direct paths to misuse.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The dataset is compiled from publicly available sources of a religious text and its recitation. It does not contain sensitive personal information or generative models with a high potential for misuse, so special safeguards are not applicable.

Guidelines:

- The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly credits the original owners of the three source datasets by citing them directly and providing URLs in the references. This allows users to access the original assets and view their respective licenses and terms of use.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new asset, the Quran-MD dataset, is well-documented within the paper. The methodology for its creation is described in Section 3 , its hierarchical JSON structure is detailed in Section 4 , and a clear example is provided in the appendix.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve any new crowdsourcing or direct research with human subjects. It curates the dataset from existing public sources.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: As the research did not involve new experiments with human subjects, IRB approval was not applicable.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core methodology involved data aggregation, alignment, and structuring from existing public sources. LLMs were not a component of the dataset curation process.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.