Leveraging AI to Bridge Classical Arabic and Modern Standard Arabic for Text Simplification

Shatha Altammami
King Saud University
Riyadh, Saudi Arabia
shaltammami@ksu.edu.sa

Abstract

This paper introduces the Hadith Simplification Dataset, a novel resource comprising 250 pairs of Classical Arabic (CA) Hadith texts and their simplified Modern Standard Arabic (MSA) equivalents. Addressing the lack of resources for simplifying culturally and religiously significant texts, this dataset bridges linguistic and accessibility gaps while preserving theological integrity. The simplifications were generated using a large language model and rigorously verified by an Islamic Studies expert to ensure precision and cultural sensitivity. By tackling the unique lexical, syntactic, and cultural challenges of CA-to-MSA transformation, this resource advances Arabic text simplification research. Beyond religious texts, the methodology developed is adaptable to other domains, such as poetry and historical literature. This work underscores the importance of ethical AI applications in preserving the integrity of religious texts while enhancing their accessibility to modern audiences.

1 Introduction

Automatic text simplification (TS) is the process of reducing the linguistic complexity of a text to improve its readability and accessibility while preserving its original meaning and core information (Saggion and Hirst, 2017). TS serves a wide range of purposes, including aiding readers with low literacy levels, second-language learners, and children (Alhafni et al., 2024). It is particularly valuable for facilitating the understanding of religious texts, which often feature intricate linguistic structures, symbolic expressions, and culturally specific content that can pose challenges for contemporary audiences (Brown, 2017).

When simplifying religious texts, it is crucial to address the unique characteristics of the source material. Strategies such as elaborative modification or adjusting linguistic form to enhance clarity are essential (Siddharthan, 2014). This need is particularly pronounced for Classical Arabic (CA), the language of foundational Islamic texts, including the Quran and Hadith. CA is renowned for its rich linguistic structures, characterized by complex syntax and vocabulary that is often symbolic or archaic (Kadaoui et al., 2023). In contrast, Modern Standard Arabic (MSA), which evolved from CA, offers a simpler and more standardized linguistic framework, making it more accessible for modern communications (Habash, 2010).

Simplifying Hadith texts requires addressing not only linguistic and syntactic differences but also sensitivity to cultural and theological nuances. These texts carry profound religious significance, necessitating methods that safeguard their authenticity while enhancing accessibility. Moreover, many words and grammatical structures in CA have no direct equivalents in MSA or modern Arabic dialects, posing an additional layer of complexity (Kadaoui et al., 2023).

Recent advancements in natural language processing (NLP) have facilitated the use of machine learning and large language models (LLMs) for tasks like text simplification (Al-Thanyyan and Azmi, 2023; Alhafni et al., 2024; Khallaf and Sharoff, 2022; Nassiri et al., 2022; Elneima et al., 2024). However, LLMs trained on modern data often lack exposure to the archaic vocabulary and syntax of CA, as many of its linguistic features are absent in MSA or contemporary dialects (Kadaoui et al., 2023). To address this gap, this study introduces the Hadith Simplification Dataset, consisting of 250 Hadith texts simplified from CA to MSA[1]. Although the dataset is relatively small, it aligns well with the capabilities of current LLMs,

---

[1]https://github.com/ShathaTm/CA_MSA

which can learn effectively from a few high-quality examples, as demonstrated by the promise of few-shot in-context learning. This approach leverages semantic alignment and contextual examples to enable LLMs to perform effectively on low-resource languages (Cahyawijaya et al., 2024).

The primary contribution of this work is not only the introduction of the Hadith Simplification Dataset but also the development of a versatile framework for producing similar datasets. Although the focus here is on religious texts, the framework is adaptable to other domains, such as poetry, historical documents, legal texts, and classical literature, where simplifying Classical Arabic (CA) to Modern Standard Arabic (MSA) can enhance accessibility and comprehension. For example, simplifying pre-Islamic Arabic poetry could make its rich imagery and cultural insights more understandable to modern readers, while converting legal texts written in CA into MSA could improve their usability for contemporary legal practice.

Moreover, it is undeniable that large language models (LLMs) are increasingly being used to answer critical questions (Dam et al., 2024), including those on sensitive religious and cultural topics (Alan et al., 2024; Benkler et al., 2023). Given the widespread adoption of these technologies, it is no longer feasible to prevent people from using LLMs for such purposes. Therefore, it is essential to take a proactive approach to ensure these models are guided toward accuracy and authenticity. Training LLMs with carefully curated datasets, including the one introduced in this paper, is not aimed at replacing scholars but at mitigating the risks of these models inadvertently disseminating misinformation. This effort is especially important as many users, regardless of their age or background, may lack the expertise to identify inaccuracies in the responses provided (Chen and Shu, 2023). By equipping LLMs with reliable data, the goal is to enable them to serve as complementary tools to Islamic scholarship, fostering a well-informed and responsible digital discourse while preserving the integrity of religious knowledge in an era increasingly influenced by artificial intelligence.

2 Related Work

The field of Text Simplification (TS) has seen a significant advancement through the creation of various corpora aimed at simplifying texts across different languages. One of the most prominent datasets is the English Wikipedia (EW) and Simple English Wikipedia (SEW) corpus, which provides 137,000 aligned sentence pairs between English Wikipedia articles and their simplified counterparts in Simple English (Coster and Kauchak, 2011). Another English-based resources are Zhu et al. (2010); Kajiwara and Komachi (2016) with more than 50K sentence pairs. These dataset are widely used due to its scale and the lexical and syntactic diversity it provides.

Beyond English, TS research in other languages has also benefited from various datasets. For example, the Simplext corpus (Saggion et al., 2015) is a Spanish dataset with 200 simplified news texts, covering domains such as national and international news, culture, and society. In Italian, the SIMPITIKI corpus (Brunato et al., 2016) provides 1,166 aligned sentences and combines semi-automatically and manually simplified texts, making it a useful resource for rule-based syntactic simplification research. In French, the Alector corpus (Gala et al., 2020) comprising 79 texts with simplified versions adapted for young readers, Alector is designed to improve readability in primary school education and is simplified at morpho-syntactic, lexical, and discourse levels.

Although recent resources mark progress in text simplification (TS) across various languages, research and resources for Arabic TS remain comparatively limited, especially given the language's range, including Classical Arabic (CA), Modern Standard Arabic (MSA), and various regional dialects. One of the newer datasets in this area is the Saaq al-Bambuu Corpus (Khallaf and Sharoff, 2022), which includes 2,980 parallel sentences from the Arabic novel Saaq al-Bambuu, aligned between complex and simplified versions. This corpus represents MSA rather than CA, focusing on narrative simplifications and lacking the religious and traditional stylistic elements found in Hadith literature. Consequently, while it offers valuable data for syntactic and lexical simplification, its structure, vocabulary, and themes differ considerably from those in CA

Hadith, which embodies distinct stylistic and lexical patterns.

Another resource is the Arabic EW-SEW (English Wikipedia–Simple English Wikipedia) and Arabic WikiLarge (Al-Thanyyan and Azmi, 2023) datasets are machine-translated Arabic adaptations of popular English text simplification resources (Coster and Kauchak, 2011). The Arabic EW-SEW contains 82,585 sentence pairs and is primarily based on general encyclopedic content, aligning complex and simplified forms on a wide variety of topics. Both datasets primarily address contemporary MSA content, thus differing significantly from CA Hadith in both thematic focus and linguistic style. The Hadith texts involve highly specialized religious content and formal syntax, which require preserving theological and moral nuances, a complexity less emphasized in the encyclopedic datasets.

Each of these resources, while valuable for general Arabic text simplification, does not address the unique complexities involved in simplifying Classical Arabic (CA) Hadith to Modern Standard Arabic (MSA). CA contains many words and grammatical structures that no longer appear in MSA or Arabic dialects, and large language models (LLMs) trained primarily on modern data have limited exposure to these structures, which impacts their performance on such data (Kadaoui et al., 2023).

Simplifying religious texts like Hadith requires precision to preserve theological accuracy, religious sensitivity, and the formal, traditional style specific to Hadith literature. The proposed Hadith simplification dataset aligns CA Hadith with their simplified MSA versions, carefully maintaining the semantic integrity of each Hadith while enhancing readability. This alignment contributes uniquely to Arabic text simplification, filling a critical gap for resources that address the specific demands of religious and CA with high fidelity.

3 Data Collection

The dataset of Classical Arabic (CA) hadith was collected using a structured approach to ensure the selection of relevant, authentic, and concise hadith texts for effective simplification into Modern Standard Arabic (MSA). Below are the detailed steps taken during data collection:

3.1 Source of Hadith

Hadith were sourced from the LK Hadith Corpus[2] (Altammami et al., 2020). In this corpus, Matn (actual Hadith teaching) is separated from Isnad (chain of narrators), allowing for the extraction of only the Matn, which is the primary focus for simplification.

3.2 Selection Criteria

Hadith were selected exclusively from the Sahih Al-Bukhari and Sahih Muslim collection to ensure authenticity, given that other collections within the corpus include hadith with varying levels of authenticity. Random sampling was then applied to achieve a diverse and representative dataset, prioritizing hadith with 100 words or fewer to maintain manageable text lengths for analysis and simplification (Al-Shameri and Al-Khalifa, 2024). Furthermore, duplicate hadith were identified and removed, ensuring that each entry in the dataset is unique and distinct.

3.3 Preprocessing

Diacritics were systematically removed from the Matn text to ensure consistency and focus on the core linguistic structure, minimizing variations that arise from diacritical marks. This approach aligns with findings from previous research, which indicated that removing diacritics from CA texts generally enhances translation accuracy, as models tend to interpret meanings inaccurately when diacritics are present (Kadaoui et al., 2023)

4 Methodology

4.1 Using a large Language Model

Closed-source language models, while limiting access to their underlying architecture and training data, continue to be attractive due to their exceptional performance on a wide range of natural language processing tasks (Bang et al., 2023). Particularly, GPT-4 (OpenAI, 2023), a powerful language model known for its ability to generate human-quality text, was chosen as the primary tool for this task.

4.2 Prompt Engineering

To enhance the quality of the simplifications, task-specific instructions and domain-specific knowl-

---

[2]https://github.com/ShathaTm/LK-Hadith-Corpus

edge were incorporated into the prompts provided to the model. This approach, inspired by previous research (Peng et al., 2023; Gao et al., 2023), guided the LLM towards more accurate and contextually relevant outputs. The following lines show the prompt given to the LLM:

```
    Please simplify the following
    hadith from Classical Arabic to
  clear formal Modern Standard Arabic.
  Start directly with the main content,
    omitting any chain of narrators.
    Use straightforward language,
    replacing any archaic terms
  with widely understood MSA vocabulary.
  For religious or specialized terms,
    provide simple explanations or
  add clarifying words in parentheses
    where needed. Keep the original
    dialogue structure and retain all
        essential details.
```

4.3 Self-correction

A self-correction mechanism was employed to mitigate potential errors introduced by the LLM itself. By asking the language model to review and refine its own responses, the overall quality of the generated text was improved. This is particularly aligned with other studies (Lu et al., 2023)

4.4 Using English prompt

To optimize the model's performance, a concise and direct English prompt was adopted for the experiments. This approach, supported by previous research (Khondaker et al., 2023), provided clear instructions to the LLM and maximized its capabilities.

4.5 Few-shot learning

Inspired by recent research on enhancing the performance of large language models (LLMs) (Kadaoui et al., 2023), a few-shot learning approach was employed to guide the LLM in generating simplified versions of Classical Arabic (CA) Hadith texts into Modern Standard Arabic (MSA). Few-shot learning involves presenting the model with a small, carefully selected set of annotated examples to demonstrate the desired task and output structure. In this study, five examples of CA Hadith, paired with their human-generated MSA simplifications, were incorporated into the prompts to clearly define the task. This approach leverages the model's

in-context learning capability, enabling it to adapt its responses by recognizing and replicating the patterns, structures, and stylistic nuances present in the provided examples.

5 Human Verification

To ensure the accuracy and reliability of the Hadith simplification process from Classical Arabic (CA) to Modern Standard Arabic (MSA) performed by the large language model (LLM), a human verification step was incorporated. This verification was conducted by an Islamic Studies graduate with a strong background in interpreting religious texts. The primary responsibility of the annotator was to carefully review the simplified MSA version of each Hadith and compare it to its original CA counterpart. The annotator was asked to refer to external sources of exegesis if necessary. The annotation was done through a Google Form interface.

The verification process focused on ensuring that the meaning, context, and theological accuracy of the Hadith were preserved in the MSA version. In cases where discrepancies were identified, the annotator would rewrite the Hadith text to ensure alignment with the original meaning. The involvement of an annotator was crucial to maintaining the integrity and authenticity of the Hadith, particularly given the sensitive nature of religious texts. As shown in Figure 1, a significant number of Hadith simplifications required modification by the human annotator. These modifications ranged from minor word choice adjustments to more substantial rephrasing to accurately convey the original meaning. For example, the simplification of the word فتمعكت was not correct and needed to be revised. In other cases, the AI model struggled with complex grammatical structures or nuanced meanings, leading to inaccuracies in the simplified text. The human annotator's role was essential in identifying and correcting these errors, ensuring that the simplified Hadith remained faithful to the original text.

6 Evaluation

Traditional readability metrics for Arabic, as described in the literature, rely heavily on surface-level linguistic features. For example, Al-Dawsari (2004) developed a formula based on

CA Hadith	AI Generated MSA Hadith	Correct?	Human modified MSA Hadith	Type of mistake
1 جاء رجل إلى عمر بن الخطاب، فقال: إني أجنبت فلم أصب الماء. فقال عمار بن ياسر لعمر بن الخطاب: أما تذكر أنا كنا في سفر، أنا وأنت، فأما أنت فلم تصل، وأما أنا فتمعكت فصليت، فذكرت للنبي صلى الله عليه وسلم، فقال النبي صلى الله عليه وسلم: إنما يكفيك هكذا. فضرب النبي صلى الله عليه وسلم بكفيه الأرض، ونفخ فيهما، ثم مسح بهما وجهه وكفيه.	هذا ليس النص الأصلي للحديث، بل تفسير مبسط له: جاء رجل إلى عمر بن الخطاب وقال: "أصابتني جنابة ولم أجد ماء. " فقال عمار بن ياسر: هل تذكر أننا كنا في سفر، ولم تصل أنت، أما أنا فتوسخت أو اتسخت ثم صليت؟ وعندما أخبرت النبي ﷺ، قال: كان يكفيك أن تفعل هكذا . ثم ضرب النبي ﷺ كفيه على الأرض، ونفخ فيهما ، ومسح بهما وجهه وكفيه.	No	هذا ليس النص الأصلي للحديث، بل تفسير مبسط له: جاء رجل إلى عمر بن الخطاب وقال: "أصابتني جنابة (أي حالة تتطلب الغسل) ولم أجد ماء". عندها تذكر عمار بن ياسر وقال لعمر: "هل تذكر أننا كنا في سفر، وأنت لم تصل، بينما أنا تمرغت (تقلبت في التراب لتطهير نفسي) ثم صليت؟ وعندما أخبرت النبي صلى الله عليه وسلم بذلك، قال: كان يكفيك أن تفعل هكذا". ثم ضرب النبي كفيه على الأرض، ونفخ فيهما، ومسح بهما وجهه وكفيه.	تفسير معنى كلمة تمعكت
2 عن عائشة، أن هند بنت عتبة، قالت يا رسول الله إن أبا سفيان رجل شحيح، وليس يعطيني ما يكفيني وولدي، إلا ما أخذت منه وهو لا يعلم فقال خذي ما يكفيك وولدك بالمعروف .	هذا ليس النص الأصلي للحديث، بل تفسير مبسط له: جاءت هند بنت عتبة إلى النبي وقالت: يا رسول الله، إن أبا سفيان رجل بخيل، فهل يجوز لي أن آخذ من ماله لإطعام عائلتنا؟" فأجابها النبي ﷺ: "نعم، لكن بما هو معروف ومعتدل بما يكفي للحاجة فقط."	No	هذا ليس النص الأصلي للحديث، بل تفسير مبسط له: جاءت هند بنت عتبة إلى النبي وقالت، إن أبا سفيان رجل بخيل، ولايعطيني مايكفيني أنا وولدي فهل يجوز لي أن آخذ من ماله وهو لايعلم " فأجابها النبي ﷺ: "نعم، لكن بما هو معروف ومعتدل بما يكفي للحاجة فقط). "	عدم اتمام المعنى المقصود
3 اوصاني خليلي بثلاث لا ادعهن حتى اموت: صوم ثلاثة ايام من كل شهر، وصلاة الضحى، ونوم على وتر.	هذا ليس النص الأصلي للحديث، بل تفسير مبسط له: قال النبي صلى الله عليه وسلم: أوصاني خليلي بثلاثة أمور لن أتركها حتى أموت: صيام ثلاثة أيام من كل شهر، وصلاة الضحى، وختم الليل بصلاة الوتر.	No	هذا ليس النص الأصلي للحديث، بل تفسير مبسط له: قال أبو هريرة رضى الله عنه أوصاني خليلي وهو النبي صلى الله عليه وسلم لن أتركها حتى أموت: صيام ثلاثة أيام من كل شهر، وصلاة الضحى، وختم الليل بصلاة الوتر.	قائل هذا الحديث أبو هريرة
4 من اعتق شقصا له في عبد، اعتق كله، ان كان له مال، والا يستسع غير مشقوق عليه.	هذا ليس النص الأصلي للحديث، بل تفسير مبسط له: من حرر نصيبًا له في عبد، يُعتق العبد بالكامل إذا كان لديه مال لدفع باقي قيمته. وإن لم يكن لديه مال، يعمل العبد لسداد باقي قيمته دون تحميله مشقة.	No	هذا ليس النص الأصلي للحديث، بل تفسير مبسط له: إذا اشترك رجلين في ثمن عبد فدفع كل منهم نصف ثمنه ثم بعد ذلك أراد واحد منهم أن يحرر العبد في نصيبه فقط هنا من النبي صلى الله عليه وسلم أن يُحرر العبد بالكامل إذا كانت هناك مقدرة مالية للعاتق فوق احتياجاته الأساسية، وإن لم يكن لديه المقدرة يعمل العبد لسداد باقي قيمته دون تحميله مشقة	عدم التوضيح
5 اذا كان يوم الجمعة، كان على كل باب من ابواب المسجد الملائكة، يكتبون الاول فالاول فاذا جلس الامام طووا الصحف، وجاؤوا يستمعون الذكر.	هذا ليس النص الأصلي للحديث، بل تفسير مبسط له: عندما يأتي يوم الجمعة، تقف الملائكة عند أبواب المسجد، يكتبون أسماء القادمين بالترتيب. فإذا جلس الإمام لبدء الخطبة، طووا الصحف وجلسوا يستمعون للذكر.	No	هذا ليس النص الأصلي للحديث، بل تفسير مبسط له: عندما يأتي يوم الجمعة، تقف الملائكة عند أبواب المسجد، يكتبون أسماء القادمين بالترتيب. فإذا جلس الإمام لبدء الخطبة، تركت كتابة من يأتي بعد ذلك ، وجلسوا ليستمعوا للخطبة.	عدم تبسيط بعض الكلمات
6 قال النبي صلى الله عليه وسلم: الجنة اقرب الى احدكم من شراك نعله	هذا ليس النص الأصلي للحديث، بل تفسير مبسط له: قال النبي صلى الله عليه وسلم: "الجنة قريبة من أحدكم بقدر قرب شراك نعله" (وهو الحزام الذي يثبت النعل على القدم).	Yes	-	-
7 عن رسول الله ﷺ قال يتبع الميت ثلاثة، فيرجع اثنان ويبقى معه واحد: يتبعه أهله وماله وعمله، فيرجع أهله وماله، ويبقى عمله.	هذا ليس النص الأصلي للحديث، بل تفسير مبسط له: قال النبي صلى الله عليه وسلم: يتبع الميت ثلاثة أشياء: أهله، وماله، وعمله. فيرجع أهله وماله، ويبقى معه عمله.	Yes	-	-
8 كنت عند النبي صلى الله عليه وسلم إذ جاءه رسول إحدى بناته، وعنده سعد وأبي بن كعب ومعاذ، أن ابنها يجود بنفسه، فبعث إليها: لله ما أخذ وله ما أعطى، كل بأجل، فلتصبر ولتحتسب.	هذا ليس النص الأصلي للحديث، بل تفسير مبسط له: جاء رسول إحدى بنات النبي صلى الله عليه وسلم يخبره أن ابنها يحتضر. فقال النبي صلى الله عليه وسلم: "لله ما أخذ، وله ما أعطى، وكل شيء عنده بأجل محدد. فلتصبر ولتحتسب."	Yes	-	-

Figure 1: Examples of CA Hadith Simplified to MSA using LLM, then verified and corrected if necessary by a Human annotator

five features, such as average word length, sentence length, and word frequency. Similarly, Al Tamimi et al. (2014) introduced AARI, which calculates readability using seven features, including the number of characters, words, sentences, and difficult words. On the other hand, El-Haj and Rayson (2016) proposed the OSMAN metric, which leverages Modern Standard Arabic (MSA) script markers and syllable counts derived from automatic diacritization. While these methods provide valuable insights into general readability, they are less likely to account for the nuanced linguistic simplifications involved in simplifying religious texts like Hadith from CA to MSA.
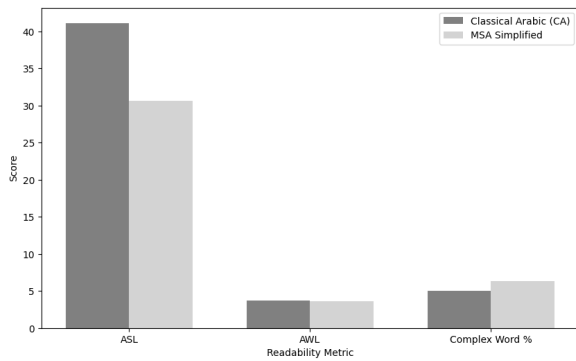

Figure 2: Comparison of traditional features used in readability metrics

This limitation is evident in the data presented in Figure 2, which shows metrics such as Average Word Length (AWL) and the percentage of complex words. Although AWL remains relatively stable, the increase in the percentage of complex words highlights the inadequacy of these traditional measures for this task. This is because the process of simplifying Hadith often involves explaining archaic terms, which can lead to an increase in the overall word count and a higher number of words exceeding six characters. Consequently, traditional metrics, which rely on surface-level features like sentence and word length, are less likely to reflect deeper shifts in linguistic style and register. Hence, these methods do not fully encapsulate the readability improvements achieved through the simplification of CA Hadith to MSA texts.

In contrast, the SAMER readability metric is specifically designed to address the unique linguistic features of Arabic, making it particularly
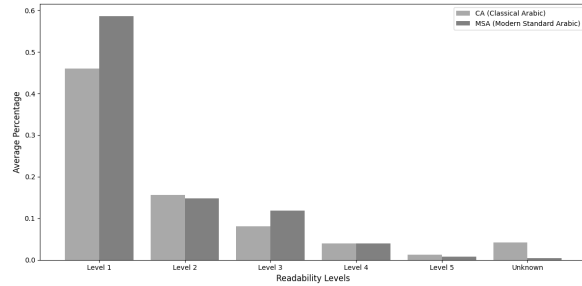

Figure 3: SAMER readability scores of Hadith texts in their Classical Arabic (CA) and Modern Standard Arabic (MSA)

suitable for evaluating the simplification process of Hadith texts. Unlike traditional metrics, SAMER goes beyond surface-level features by incorporating the morphological richness and syntactic complexity inherent in Arabic. SAMER evaluates readability by analyzing lexical, syntactic, and morphological features, mapping them to predefined readability levels.

The SAMER tool is underpinned by the SAMER Arabic Text Simplification Corpus, a resource created from 15 publicly available Arabic fiction novels. These texts were manually simplified into parallel versions representing specific readability levels. This enables SAMER to capture the linguistic shifts involved in simplifying CA to MSA, such as replacing archaic terms and restructuring sentences to enhance readability. Therefore, it is used to measure the readability of the CA hadiths and and their simplified meaning to MSA. An example of how the SAMER tool analyzes text is shown in Figures 4 and Figure 5 where words are color coded to show which level of readability it falls into.

SAMER has developed a five-level readability scale. This scale categorizes words based on their frequency and complexity, facilitating the adaptation of texts to appropriate reading levels. The levels are defined as follows:

- Level 1: Contains the most basic and frequently used words in MSA, suitable for beginners.

- Level 2: Includes slightly more complex words that are still common, appropriate for early intermediate learners.

- Level 3: Comprises words of moderate complexity, intended for intermediate learners.

- Level 4: Features less common and more complex words, suitable for advanced learners.

- Level 5: Consists of the most complex and least frequent words, appropriate for proficient readers.

Figure 3 illustrates the SAMER readability scores of the Hadith Simplification Dataset by comparing their CA and MSA scores, showcasing the effectiveness of simplification. The distribution highlights a significant shift in readability levels, with MSA texts displaying a higher proportion of words categorized in Level 1, indicating simpler and more accessible content. Meanwhile, the percentage of words in higher complexity levels, such as Level 4 and Level 5, is notably reduced in the MSA version compared to CA. This transformation reflects the effort to clarify and simplify archaic terms and dense sentence structures in the original CA Hadith.

7 Conclusion and Future Work

This paper introduces the framework for developing the Hadith Simplification Dataset, which comprises 250 pairs of Classical Arabic (CA) Hadith texts and their simplified Modern Standard Arabic (MSA) equivalents. Unlike general-purpose Arabic corpora, this specialized dataset preserves the semantic integrity of culturally and religiously significant Hadith texts while enhancing accessibility for MSA readers. By addressing the intricate lexical, syntactic, and cultural challenges inherent in transforming CA into MSA, this resource makes a substantial contribution to Arabic text simplification research, particularly in domains requiring precision, cultural sensitivity, and semantic fidelity, such as religious texts.

This research represents an ongoing effort to explore how computational tools can responsibly contribute to addressing sensitive topics in religious studies. With the growing reliance on AI systems for religious and ethical guidance, it is crucial to address their limitations and refine methodologies in this space. Avoiding this domain entirely could exacerbate the issue, as individuals are unlikely to refrain from using AI and chatbots for religious inquiries despite their current shortcomings. By proactively developing

approaches to manage large language models (LLMs) within this context, the potential for misinformation can be mitigated, ensuring that AI systems handle religious topics with greater accuracy, responsibility, and cultural respect.

Future directions include expanding the dataset by incorporating more annotators and a wider selection of text pairs to increase its robustness and applicability. Additionally, feedback from the Coling-Rel workshop's community will be instrumental in refining and extending this initiative. A collaborative and iterative approach will help tackle the challenges posed by AI in religious contexts, creating solutions that harmonize technological innovation with ethical and cultural responsibilities.

Ethical Note

The creation of a dataset for Hadith simplification from Classical Arabic (CA) to Modern Standard Arabic (MSA) entails significant ethical considerations due to the theological and cultural importance of Hadith. Any simplification effort must maintain theological accuracy, respect the integrity of the original texts, and avoid altering their essential meaning or spiritual significance. To address these concerns, the simplifications generated by GPT-4 have been rigorously reviewed by an Islamic studies graduate to ensure their fidelity to the original texts.

This work not only aims to make Hadith more accessible to non-specialists and modern readers who may find CA challenging but also contributes to the field of Natural Language Processing (NLP) by providing a carefully validated dataset. This dataset serves as a resource for training models dedicated to Arabic text simplification, emphasizing the ethical responsibility to preserve the meaning and sanctity of religious texts while advancing AI applications.

By bridging the gap between traditional Islamic texts and contemporary understanding, this research exemplifies a commitment to ethical rigor in AI-based religious text processing. It highlights the importance of ensuring that technological advancements respect cultural and theological values, while also addressing the under-representation of

Figure 4: Original CA Hadith analyzed using SAMER readability metric

Figure 5: The simplified Hadith analyzed using SAMER readability metric

CA in NLP applications.

Acknowledgments

References

M Al-Dawsari. 2004. The assessment of readability books content (boys-girls) of the first grade of intermediate school according to readability standards. *Sultan Qaboos University, Muscat.*

Noora Al-Shameri and Hend Al-Khalifa. 2024. Arabic paraphrased parallel synthetic dataset. *Data in Brief*, page 111004.

Abdel Karim Al Tamimi, Manar Jaradat, Nuha Al-Jarrah, and Sahar Ghanem. 2014. Aari: automatic arabic readability index. *Int. Arab J. Inf. Technol.*, 11(4):370–378.

Suha S Al-Thanyyan and Aqil M Azmi. 2023. Simplification of arabic text: A hybrid approach integrating machine translation and transformer-based lexical model. *Journal of King Saud University-Computer and Information Sciences*, 35(8):101662.

Ahmet Yusuf Alan, Enis Karaarslan, and Ömer Aydin. 2024. A rag-based question answering system proposal for understanding islam: Mufassirqas llm. *arXiv preprint arXiv:2401.15378.*

Bashar Alhafni, Reem Hazim, Juan Piñeros Liberato, Muhamed Al Khalil, and Nizar Habash. 2024. The samer arabic text simplification corpus. *arXiv preprint arXiv:2404.18615.*

Shatha Altammami, Eric Atwell, and Ammar Alsalka. 2020. The arabic-english parallel corpus of authentic hadith. *International Journal on Islamic Applications in Computer Science And Technology*, 8(2).

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. arxiv. *arXiv preprint arXiv:2302.04023.*

Noam Benkler, Drisana Mosaphir, Scott Friedman, Andrew Smart, and Sonja Schmer-Galunder. 2023. Assessing llms for moral value pluralism. *arXiv preprint arXiv:2312.10075.*

Jonathan AC Brown. 2017. *Hadith: Muhammad's legacy in the medieval and modern world.* Simon and Schuster.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, and Giulia Venturi. 2016. Paccss-it: A parallel corpus of complex-simple sentences for automatic text simplification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 351–361.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. Llms are few-shot in-context low-resource language learners. *arXiv preprint arXiv:2403.16512.*

Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788.*

William Coster and David Kauchak. 2011. Simple english wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669.

Sumit Kumar Dam, Choong Seon Hong, Yu Qiao, and Chaoning Zhang. 2024. A complete survey on llm-based ai chatbots. *arXiv preprint arXiv:2406.16937.*

Mahmoud El-Haj and Paul Rayson. 2016. Osman—a novel arabic readability metric. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 250–255.

Ashraf Hatim Elneima, AhmedElmogtaba Abdelmoniem Ali Abdelaziz, and Kareem Darwish. 2024. Osact6 dialect to msa translation shared task overview. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation@ LREC-COLING 2024*, pages 93–97.

Núria Gala, Anaïs Tack, Ludivine Javourey-Drevet, Thomas François, and Johannes C Ziegler. 2020. Alector: A parallel corpus of simplified french texts with alignments of misreadings by poor and dyslexic readers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1353–1361.

Yuan Gao, Ruili Wang, and Feng Hou. 2023. How to design translation prompts for chatgpt: An empirical study. *arXiv preprint arXiv:2304.02182.*

Nizar Y Habash. 2010. *Introduction to Arabic natural language processing.* Morgan & Claypool Publishers.

Karima Kadaoui, Samar M Magdy, Abdul Waheed, Md Tawkat Islam Khondaker, Ahmed Oumar El-Shangiti, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Tarjamat: Evaluation of bard and chatgpt on machine translation of ten arabic varieties. *arXiv preprint arXiv:2308.03051.*

Tomoyuki Kajiwara and Mamoru Komachi. 2016. Building a monolingual parallel corpus for text simplification using sentence similarity based on alignment between word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1147–1158.

Nouran Khallaf and Serge Sharoff. 2022. Towards arabic sentence simplification via classification and generative approaches. *arXiv preprint arXiv:2204.09292.*

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Gptaraeval: A comprehensive evaluation of chatgpt on arabic nlp. *arXiv preprint arXiv:2305.14976.*

Qingyu Lu, Baopu Qiu, Liang Ding, Liping Xie, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt.

Naoual Nassiri, Safae Berrichi, Abdelhak Lakhouaja, Violetta Cavalli-Sforza, and Azzeddine Mazroui. 2022. Lexical simplification of arabic educational texts through a classification approach. In *The International Conference on Artificial Intelligence and Smart Environment*, pages 581–587. Springer.

OpenAI. 2023. Gpt-4 technical report. https://openai.com/research/gpt-4.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of chatgpt for machine translation. *arXiv preprint arXiv:2303.13780*.

Horacio Saggion and Graeme Hirst. 2017. *Automatic text simplification*, volume 32. Springer.

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):1–36.

Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361.