The Arabic Bible as an Evaluation Tool: The Case Study of the Khalīlī Arabic Dialect

Jakub Zbrzeżny[1] **Ehud Reiter**[2] **Wei Zhao**[2]

Department of Divinity[1] Department of Computing Science[2]

University of Aberdeen

jakub.zbrzezny@abdn.ac.uk e.reiter@abdn.ac.uk wei.zhao@abdn.ac.uk

Abstract

The paper presents a fully documented case study of how high-quality data combined with evaluators' expertise can be utilised for conducting basic NLP experiments in the realm of low-resource languages such as local varieties of Colloquial Arabic, and how the Arabic Bible, hitherto underutilised in NLP, can serve as an evaluation tool. Our experiments on one of the rural Palestinian Arabic dialects of al-Khalīl / Hebron illustrate two points. On the one hand, popular models are clearly limited in their ability to produce outputs of a high level of dialectal specificity (here: rural area surrounding a major urban centre). On the other hand, they are capable to generate accurate translations from such dialects into Modern Standard Arabic. Thus, the models appear better at understanding dialects than at producing dialects.

1 Introduction

Whether in its Jewish or one of several Christian forms, the Bible is a remarkable linguistic data set of extraordinary comparability, synchronic and diachronic alike. That Natural Language Processing (NLP) can benefit from the Bible both as a training data set and as an evaluation tool has been well noted (Alastruey et al., 2026). Still, even if Bible translations have often been the first written literary productions in many languages, there seems to be remarkable little literature on their usefulness for NLP in low-resource languages.

Such is the case of written expressions of Colloquial Arabic, the actual mother tongue of Arabic native speakers, which must be distinguished from formally learnt types of Arabic like Classical Arabic, the language of literature, or Modern Standard Arabic (MSA), the language of official media[1]. Unlike highly standardised forms of formally learnt Arabic, Colloquial Arabic represents a spectrum of varieties, which are commonly associated with geographical regions, countries, and smaller areas, from country districts to cities and their surroundings to individual towns or villages. Even such local varieties are often further subdivided into nomadic and sedentary types, of which the latter have distinct urban and rural forms[2].

The issue of the under-representation of Colloquial Arabic in NLP has been emphatically noted (e.g. Fakhraddin et al. (2025); Nacar et al. (2025)). Related Machine Translation challenges started to be addressed long before the advent of the Large Language Models (LLMs) and continue to be discussed (e.g. Zbib et al. (2012); Baniata et al. (2018); Harrat et al. (2019); Zakraoui et al. (2021); Alabdullah et al. (2025)). There is also no lack of experiments in testing LLMs capability to deal with it (e.g. Kadaoui et al. (2023); Khondaker et al. (2023))[3]. Nonetheless, in this context it is very rare to see the use of Colloquial Arabic Bible translations, and when such use is mentioned, the quality of data is not discussed (Sajjad et al., 2020). Admittedly, there is no comprehensive research on their largely undocumented nature in terms of their underlying texts, the sociolinguistic profile of translators, or translation techniques[4].

This paper provides an exploratory case study showing how a specific sample of Colloquial Arabic representing a sub-city variety can be employed in NLP in the context of collaboration between lo-

[1]For a more nuanced overview of the use of Classical Arabic and Modern Standard Arabic see reference works such as *Semitic Languages: an international Handbook* (Weninger et al., 2012), or the *Encyclopedia of Arabic Language and Linguistics* (Versteegh, 2005–2007).

[2]The aforesaid reference works will provide further introductory guidance into Arabic dialects.

[3]Note, however, that the research concerns high level (region / country) varieties of Colloquial Arabic (e.g. Elmadany et al. (2023); Al-Haff et al. (2022)). It is still rare to see work concerning varieties at a city-level (see Bouamor et al. (2018) on the MADAR corpus, or Mekki et al. (2026) on the Alexandria corpus.

[4]Such texts appear on confessional websites like `https://www.bible.com/` (currently including partial translations labelled as representing several country-level dialects.

Proceedings of the 1st Symposium on Natural Language Generation Evaluations, pages 24–32

cal experts, biblical scholars, and computer scientists. We discuss experiments in challenging selected LLMs to create dialect outputs meant to represent translations from English and MSA into this specific dialect, and to translate from samples of this dialect into MSA. Our case study is meant to be replicable and adaptable beyond the scholarship concerning the Arabic Bible, and even beyond Colloquial Arabic, extending towards other low-resource Semitic languages.

2 Methodology

2.1 Input texts

The source of our dialect data is a new paraphrastic rendition of selected biblical books from the pre-modern Arabic Bible into a particular dialect of the Levantine Arabic (al-Shāmī)[5]. It is one of the statistically most common Palestinian dialects, which is spoken in the southern West Bank (al-Khalīl / Hebron Governorate), and hence is called Khalīlī (hereafter Khalili[6]). Whether in its rural or urban form, it is locally easily identifiable and it bears significant cultural associations. Nonetheless, the dialect remains largely unexplored[7]. The rendition used as the source of data for our experiments is being prepared by a team of socio-linguistically aligned collaborators, who speak the rural variety of the Khalili dialect. The whole process is academically documented within research projects conducted at the University of Aberdeen[8]. At this moment, it encompasses drafts of the complete Book of Genesis (hereafter Genesis), and the complete Gospel according to Matthew (hereafter Matthew), constituting a corpus of approximately 40,000 words.

Our experiments investigate the following linguistic pairs or triplets:

- English to Khalili Dialect
- English to MSA to Khalili Dialect
- MSA to Khalili Dialect
- Khalili Dialect to MSA

The English text of Genesis was taken from *The Holy Scriptures According to the Masoretic Text* (1917)[9]. The English text of Matthew was taken from the OpenEnglishBible (2020)[10]. The MSA translation of both books comes from the STEP-Bible edition of the popular Van Dyck's translation (1865)[11] available under the licence CC BY-SA 4.0. The selection of English and MSA translations required that they are modern (but not necessarily the most recent), documented (but not necessarily meeting the current disciplinary standards), and available digitally in public domain.

Some of our experiments included among the input texts an example of how the dialect is translated into English or MSA. The dialect text was taken from selected fragments from the team's rendition of Genesis and Matthew, and these were accompanied by English and MSA translations created by our team for the purpose of the experiments.

The input texts were organised into twenty units, ten from Genesis, and ten from Matthew. On average, English units had 500 words, MSA units had 250 words, and Dialect units had 300 words (numbers rounded). The average word count was determined through initial checks, which tested the capacity of selected models to return meaningful and complete results. The examples consisted of 2,000 words in the dialect, 4,000 words in English, and 1,400 words in MSA (numbers rounded).

2.2 Models

The experiments involved three models. The selection condition was the presence of a privacy policy protecting input data from being incorporated into training data sets of a given model. Further, two models were meant to represent Generative AI tools that were publicly available at the time of experiments. The following were selected: Gemini 2 Flash (hereafter Gem2F) and ChatGPT 4o mini (2024.07.18) (hereafter GPT4om). One model was meant to represent a Generative AI tool used by more advanced non-expert users at the time of experiments. The following was selected: ChatGPT 4o (2024-08-06) (hereafter GPT4o). All three models were accessed through API calls facilitated by one of the commercial platforms (AI/ML API). This was to ensure that the texts from the hitherto unpublished Khalili dialect rendition of Genesis and Matthew remain outwith LLM training data sets.

[5] Eastern Mediterranean. Note that commonly used English categorizations based on wider geographical regions or modern states occasionally reflect the legacy of European colonialism rather than the actual dialect distribution.

[6] Simplified Romanization.

[7] For the most comprehensive bibliography of literature on Palestinian Arabic dialects, see Ulrich Seeger's list at https://arab.useeger.de/lit/Seeger-Biblio-Pal-Arabic.pdf (updated regularly; note that the list includes few positions in Modern Hebrew).

[8] Under the overarching title Hexapla Arabica.

[9] See https://en.wikisource.org/wiki/Bible_(Jewish_Publication_Society_1917.

[10] See https://openenglishbible.org/oeb/2022.1/OEB-2022.1-Cth.txt.

[11] See https://www.stepbible.org/version.jsp?version=AraSVD.

2.3 Prompts

All prompts for dialect outputs included the task of translating the input text from the stated language ("English" or "Modern Standard Arabic") into "the rural Arabic dialect of al-Khalīl in the West Bank.", or vice versa, from the dialect to MSA. The prompt default structure was: "Translate Text 1 from X into Z. Text 1 is as follows." There were three subsets of prompts:

- plain prompts, which were equal to the default (applied to all three models)
- more advanced versions of plain prompts to translate from English but with a mid-translation into MSA[12] (applied to all three models)
- prompts with an example among the input texts[13] (applied to GPT4o only).

Note that prompts for Gem2F and GPT4om had the parameters "temp" and "top_p" unstated. For GPT4o, these were always given explicit values: 0.0 and 0.1, respectively.

2.4 Output texts

The output texts were meant to be produced in the dialect and in MSA. The models were prompted to produce 200 units in the dialect. These included:

- 60 units from English directly to the dialect (all three models),
- 60 units from English through MSA to the dialect (all three models),
- 60 units from MSA to the dialect (all three models),
- 20 units from English to the dialect with an example (GPT4o).

On average, the dialect units had 250 words each, resulting in a corpus of approximately 50,000 words. There were also 80 units meant to represent MSA translations from the dialect. These included:

- 60 units created with prompts without an example (all three models),
- 20 units created with an example (GPT4o).

On average, the MSA units had 275 words, providing a corpus of approximately 22,000 words.

[12]These were formulated as follows: "Translate Text 1 from English into Modern Standard Arabic and then translate the Modern Standard Arabic translation into the rural Arabic dialect etc."

[13]These were formulated as follows: "Use the example of how Text 2 (the dialect) has been translated into Text 3 (English / MSA). […] Text 2 is as follows: […]. Text 3 is as follows: […]."

2.5 Evaluation

The evaluation was conducted independently by three team members acting as evaluators, who are native speakers of rural Khalili Arabic with solid knowledge of Classical Arabic (religious education) and MSA (secular education)[14]. The evaluators were closely aligned in terms of socio-linguistic features. They were highly familiar with the content of Genesis and Matthew through their earlier work on transcribing relevant texts from manuscripts and creating their Khalili dialect rendition. Before starting the evaluation, the evaluators discussed the scoring matrix among themselves and agreed on principles of scoring within the given parameters.

The related dialect outputs were randomly sorted in two batches: English to the dialect (140 units, grouped into 20 files, each with 7 units) and MSA to the dialect (60 units, grouped into 20 files, each with 3 units). The related MSA outputs were similarly arranged in one batch (80 units, grouped into 20 files, each with 4 units). The numerical identifiers of all units were anonymised through randomly generated numbers.

The evaluators were instructed to score each dialect output unit in terms of its dialect specificity according to the following metric with a 0-100 range:

100: rural Khalili dialect

75: Khalili dialect

50: Palestinian dialect

25: Levantine dialect

0: Colloquial Arabic

The dialect scoring matrix was formulated in writing in Colloquial Arabic in a descriptive way as follows:

100: العاميه الخليليه الفلاحيه

["rural Khalili Colloquial Arabic"]

75: العاميه الخليليه

["Khalili Colloquial Arabic"],

with the gloss: يعني الخليلي بس مش معروف مدني او فلاحي ["i.e. Khalili Arabic, but impossible to see whether urban or rural"]

50: العاميه الفلسطينيه

["Palestinian Colloquial"],

[14]The experiments were so designed that the evaluators' knowledge of English was irrelevant.

with the gloss: يعني الفلسطيني بس مش معروف من وين في البلد ["i.e. Palestinian Arabic, impossible to see from where in the country"]

25: العاميه الشاميه

["Levantine dialect"],

with the gloss: ممكن الفلسطيني ممكن الاردوني ممكن اللبناني ممكن السوري يعني بلاد الشام بس مش واضح من وين ["perhaps Palestinian Arabic, perhaps Jordanian Arabic, perhaps Lebanese Arabic, perhaps Syrian Arabic, that is, from the Levant, but unclear from where"[15]]

0: العاميه

["Colloquial Arabic"],

with the gloss: يعني ببين العاميه بس مش معروف من وين ["i.e., it seems to be Colloquial Arabic but it is impossible to see from where"]

The evaluation task formulated for MSA units was to score each MSA output unit on the scale from 0 to 100 in terms of its accuracy in translating the underlying dialect input. Again, the matrix was formulated in writing in Colloquial Arabic:

100: الترجمه الدقيقه

["accurate translation"]

50: مكس

["mixture"]

0: النص اللي مش ترجمه من اللي انتو كتبتو في لهجتكو بس مكس من ترجمات التوراه او الانجيل للرسميه اللي موجوده اونلاين

["The text does not constitute a translation from what you wrote in your dialect, but it is a mixture of MSA translations of the Hebrew Bible and the New Testament that are available online."].

The nature of the gloss to the score 0 stemmed from our initial experiments. The preliminary test outputs for translations from the dialect into English seemed to represent not the underlying paraphrastic dialect text in English, but rather a text appearing to be a mixture of modern English translations of corresponding passages in the Bible. This apparent tendency to align the supposed English translation of the Arabic paraphrase with the standard English text requires further investigation, which would to measure it and establish whether it occurs in relation to other textual corpora beyond the Bible in English.

Note that the entire work communication between the evaluators and the team members based in the UK was conducted in their dialect and without the use of English.

3 Results and Analysis

3.1 English / MSA to Dialect

The Figures 1 to 4 present the results of the evaluation. The horizontal axis represents the frequency of occurrences of a particular score given on the vertical axis:

75-99: Khalili dialect

50-74: Palestinian dialect

25-49: Levantine dialect

0-24: Colloquial dialect

Note that no unit was evaluated at **100** (rural Khalili dialect). The column 'error' indicates the number of instances where the model did not create an output.

The scoring of outputs created with prompts to translate from English directly into the dialect are shown in Figure 1. It will be seen that the Gem2F outputs received much higher scores than those created by the two other models. There is also a very high number of cases in which GPTo refused to perform the task where the prompt included an example). Figure 2 shows the results of prompts to translate from English into the dialect with an MSA mid-translation. Again, the Gem2F outputs were scored higher. When the scoring is compared across the outputs created with or without an MSA mid-translation, it will be noticed that Gem2F and GPTo performed better with MSA, but the opposite is true for GPT4om. This is shown in Figure 3. Finally, in direct translation from MSA to the dialect, the outputs created by Gem2F again scored much better than the two other models. This is shown in Figure 4.

[15] Due to the influence of Modern Hebrew and its distinctive features, the Israeli Arabic has not been enumerated here.

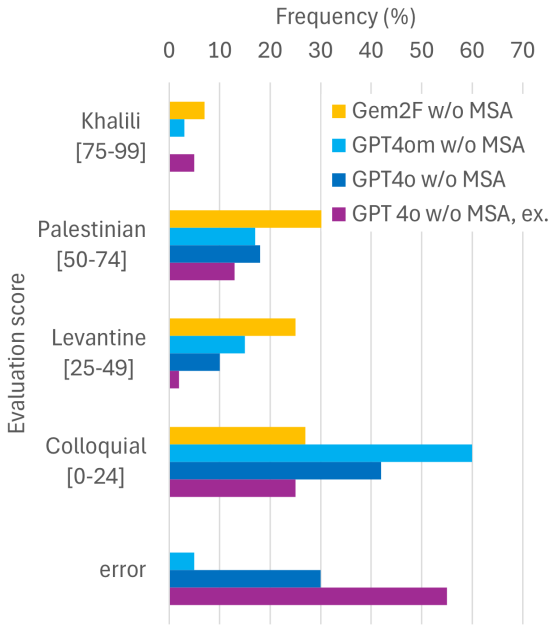

Figure 1: English to Dialect

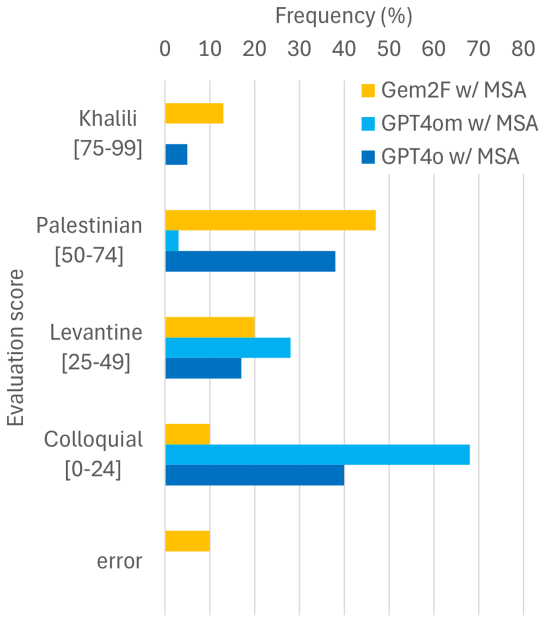

Figure 2: English to MSA to Dialect

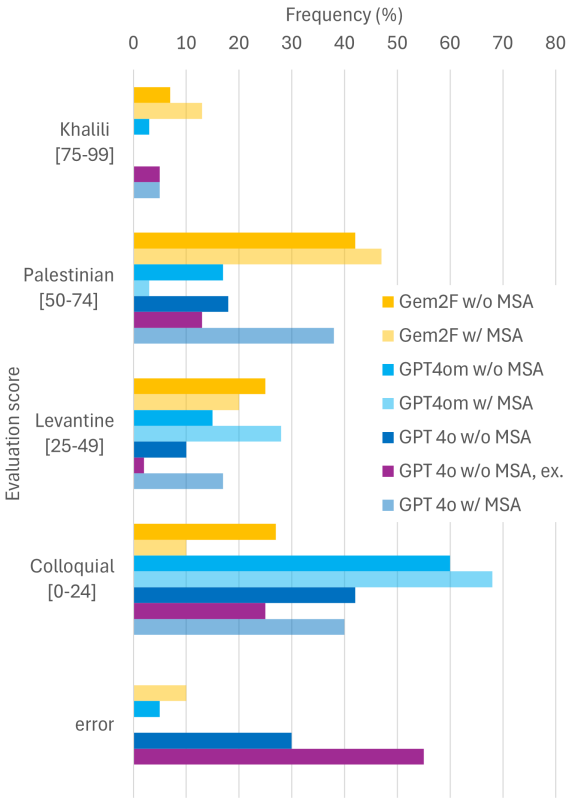

Figure 3: English to Dialect with/without MSA

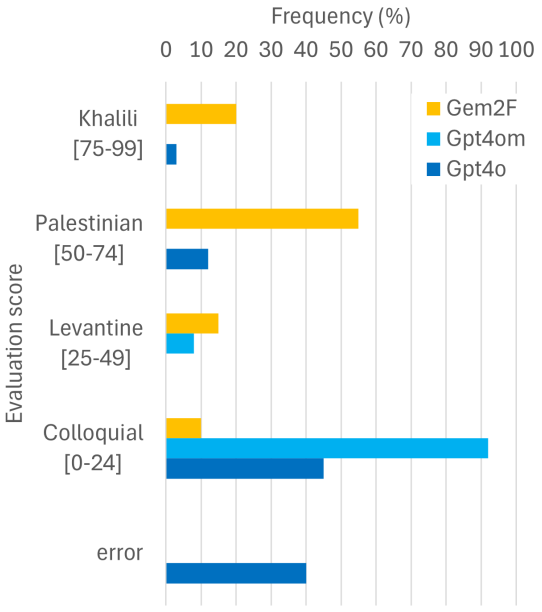

Figure 4: MSA to Dialect

Whereas the evaluation of particular models that were in use at the time of conducting our experiments provides only a snapshot into their capability at a point of time, the results give an insight into the evaluation process that remains independent of technical developments. This pertains to consistency in scoring among socio-linguistically aligned expert evaluators. It is shown in Figure 5. The figure gives the frequency of disagreements among the evaluators (the horizontal axis) of a particular value (the vertical axis), calculated as the difference between the highest and lowest scoring for each unit.

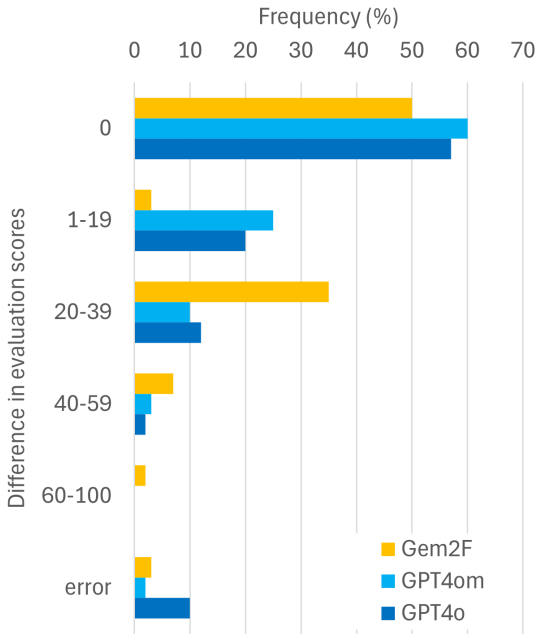

Figure 5: Disagreement scale in evaluation

This consistency is further illustrated by the average scores across the three evaluators given in Table 1[16].

Model	Ev.1	Ev.2	Ev.3	Diff. range
E>D / E>MSA>D				
Gem2F	52	48	47	5
GPT4o	39	36	37	3
GPT4om	27	28	25	3
MSA>D				
Gem2F	62	58	51	11
GPT4o	31	30	28	3
GPT4om	16	15	15	1

Table 1: Average scores across evaluators

[16]Cases where the models refused to produce an output were excluded from calculations.

3.2 Dialect to MSA

The evaluation of MSA units in terms of their accuracy towards the corresponding dialect input texts brought an unexpected lack of variation. With very few exceptions, the scoring was almost uniformly 100. This means that outputs were seen as representing an accurate translation of the underlying dialect texts. The experiments provide measurable evidence for this important finding.

This result underwent a deeper albeit preliminary investigation by a team member with expertise in Biblical Studies. Firstly, a unique set of four units that scored consistently not 100 but 50 across all three evaluators was investigated along the lines of Qualitative Error Analysis. The unit is made largely of the passage with the so-called genealogy of Jesus (Matthew 1:1-17), which is a list of personal names. A sample of Matthew 1:2-7 contains 30 names in the dialect text. In many cases, their orthography differs from that found in published MSA translations of Matthew[17]. Indeed, out of 30 names, only 8 were uniform across the dialect input text and the generated MSA output texts[18], and only 3 were in partial agreement. The other 17 cases had the names 'corrected' in the MSA outputs into forms identical with or closer to those found in published MSA translations (e.g. the 'incorrect' "Būdh" amended into "Boaz", or the 'incorrect' "Dūth" amended into "Ruth"). These 'corrections' were detected by our evaluators and interpreted as signs of inaccuracy in translating the dialect version into MSA.

Secondly, it was investigated in detail how a clearly paraphrastic dialect passage was translated into MSA while scoring 100 in terms of accuracy. The selected passage was Matthew 1:18-19, which reads in one of the standard English translations as follows:

> Now the birth of Jesus the Messiah took place in this way. When his mother Mary had been engaged to Joseph, but before they lived together, she was found to be pregnant from the Holy Spirit. Her husband Joseph, being a righteous man and unwilling to expose her to public disgrace, planned to divorce her quietly. (NRSV2021)

[17]This is a result of pre-modern and modern scribal mistakes in copying the list of largely unfamiliar and non-Arabic names.

[18]These were mostly well-known names such as, in their English form, "Abraham", "Isaac", "Jacob", "David", or "Judah".

The dialect rendition gives a significantly different text, with several omissions, additions, and changes. It can be translated into English as follows, with omissions indicated by the underscore, additions by the underline, and changes by italics:

> _____ Mary, the mother of Christ, was engaged to Joseph. Before they got married (that is, before the nuptial night occurred), *Joseph noticed* that Mary, his fiancée, is pregnant _____. Just note: it was before the nuptial night. Now, Joseph was a good person. When he learnt about the matter, he did not want to put her to shame. To the contrary, he wanted to protect her. Thus, he said to himself that he should divorce her in secret, leave her, and stay away from her[19].

All the outputs closely followed the dialect input text. Significantly, they did not add the 'missing' fragments such as the introductory sentence or the mention of the Holy Spirit. Further, they replicated explanatory additions either in full alignment with the dialect input (e.g. on the 'nuptial night'), or in partial alignment (e.g. the phrase "to protect her" occurs in 2 out of 4 outputs, and the phrase "leave her, and be away from her" occurs in 3 out of 4 outputs). They also followed the changed constructions (the active "Joseph noticed" instead of the passive "she was found"). Thus, the evaluation was correct in terms of assessing the surprising accuracy of the translation of a passage that could have been aligned with standard MSA translations.

Finally, one of the key New Testament passages was assessed, that is, the so-called "Lord's Prayer", also known as "Our Father", in Matthew 6:9-13. This central text of Christian tradition would have been expected to be particularly prone to being aligned with its standard translations, especially given the fact that the Khalili dialect rendition departs from the well-known wording and provides a highly poetic rendition of the text. Among its most striking features is shift in the possessive pronoun attached to the word 'Father' from 'our' to 'your' (plural), and a paraphrastic translation of 'your will be done' into 'may what God has decreed come to being (o Lord, according to what you wish!)'. These highly unusual features are retained in all four MSA outputs, with just one exception with a case of the standard 'Our Father'.

[19]Probably meant as to refrain from violence against her.

This preliminary investigation into a handful of notable cases exemplifies what the Evaluators detected in their assessment of MSA outputs: accuracy and capacity to morph dialect phrases into MSA. Its potential advantages notwithstanding, this raises a concern related to speeding up the process of dialect levelling, especially if predictive text or autocorrections are powered by LLMs.

4 Conclusions

The case study presented in this paper shows NLP experiments with Colloquial Arabic related to the Arabic Bible bringing meaningful results. This is shown in Table 2 with average scores[20] across the three models. One model (Gem2F) performed clearly better than others in creating dialect outputs, especially when translating from MSA or with an MSA mid-translation. However, even in these two cases, the model did not reach the level of Khalili specificity, and, on average, produced outputs categorised as representing only more broadly Palestinian dialects. The majority of outputs from other models were categorised as representing Levantine dialects. This inability of the models to produce outputs meant to have a high level of dialect specificity contrasts with the fact that all the models were assessed as highly accurate in translating dialect inputs into MSA.

The fact that the results were meaningful lies in the expertise of the evaluators and in the collaboration of local experts, biblical scholars, and computer scientists. This ensured the quality of data as well as culturally and socially appropriate evaluation process. Using Colloquial Arabic as a medium of work communication should also be noted.

It remains to be explored how applying NLP to Colloquial Arabic can contribute to the investigation of some complex linguistic phenomena such as Arabic diglossia. This potential extends beyond Arabic and is applicable to other low-resource Semitic languages such as ancient Hebrew, ancient Aramaic, or endangered Neo-Aramaic dialects. Such exploration can be successfully conducted not only within the digital humanities, but also — perhaps even more effectively — by means of multidisciplinary collaboration between the humanities and computer science.

[20]Cases where the models refused to produce an output were excluded from calculations.

Model	Direction	Avg. score	Category
Dialect Specificity			
Gem2F	MSA>D	57	Palestinian dialect
Gem2F	E>MSA>D	54	Palestinian dialect
Gem2F	E>D	45	Levantine dialect
GPT4o	E>MSA>D	41	Levantine dialect
GPT4o	E>D+ex	38	Levantine dialect
GPT4o	E>D	32	Levantine dialect
GPT4om	E>D	31	Levantine dialect
Gpt4o	MSA>D	30	Levantine dialect
GPT4om	E>MSA>D	24	Colloquial Arabic
Gpt4om	MSA>D	15	Colloquial Arabic
Accuracy			
GPT 4o	D>MSA+ex	97	Accurate translation
GPT 4o	D>MSA	96	Accurate translation
Gem2F	D>MSA	96	Accurate translation
GPT4om	D>MSA	95	Accurate translation

Table 2: Average scores across models

Limitations

It would be valuable to conduct follow-up experiments with evaluators who are not immediately familiar with Genesis and Matthew to see how they score human-created outputs in relation to AI-created counterparts, when both types of texts are randomly mixed with anonymised labels. High scores for the actual Khalili texts would provide further validation for the method presented in this paper.

Ethical Statement

All due considerations (i.a. religious, social, cultural, political, ethical, economic) were undertaken prior to the commencement of the research to ensure that the safety of the evaluators was not put at risk, and that their work was appropriately remunerated.

Acknowledgments

The research presented in this paper was conducted as part of the project 'Al-ᶜĀmmīyah (Colloquial Arabic) and Generative AI — A Snapshot of its Emerging Text-to-Text Abilities' funded by the Royal Society of Edinburgh (Collaboration Grant 4423). The underlying dialect texts were created within projects co-funded by the British Academy, the University of Aberdeen, the Department of Near Eastern Studies at Cornell University, and by private donors. This complex work involves nearly twenty colleagues from Professional Services at the University of Aberdeen, to whom we are grateful for their essential support as part of the wider team.

Finally, this research would not have been possible without the significant intellectual contributions of our five Khalili project partners in the West Bank, led by Ahmad Hroub. Their continued commitment is deeply appreciated — يعطيكو العافيه.

References

Karim Al-Haff, Mustafa Jarrar, Tymaa Hammouda, and Fadi Zaraket. 2022. Curras + baladi: Towards a levantine corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 769–778, Marseille, France. European Language Resources Association.

Abdullah Alabdullah, Lifeng Han, and Chenghua Lin. 2025. Advancing dialectal arabic to modern standard arabic machine translation. *Preprint*, arXiv:2507.20301.

Belen Alastruey, Niyati Bafna, Andrea Caciolai, Kevin Heffernan, Artyom Kozhevnikov, Christophe Ropers, Eduardo Sánchez, and 1 others. 2026. Omnilingual MT: Machine translation for 1,600 languages. *arXiv preprint arXiv:2603.16309*.

Laith H. Baniata, Seyoung Park, and Seong-Bae Park. 2018. A neural machine translation model for arabic dialects that utilises multitask learning (MTL). *Computational Intelligence and Neuroscience*, 2018:7534712.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani,

Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association.

AbdelRahim Elmadany, ElMoatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. ORCA: A challenging benchmark for arabic language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9559–9586, Toronto, Canada. Association for Computational Linguistics.

Alwajih Fakhraddin, Abdellah El Mekki, Samar Mohamed Magdy, AbdelRahim A. Elmadany, Omer Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, Rahaf Alhamouri, Hamzah A. Alsayadi, Hiba Zayed, Sara Shatnawi, Serry Sibaee, Yasir Ech-chammakhy, Walid Al-Dhabyani, Marwa Mohamed Ali, Imen Jarraya, and 25 others. 2025. Palm: A culturally inclusive and linguistically diverse dataset for arabic LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32871–32894, Vienna, Austria. Association for Computational Linguistics.

Salima Harrat, Karima Meftouh, and Kamel Smaili. 2019. Machine translation for arabic dialects (survey). *Information Processing Management*, 56(2):262–273. Advance Arabic Natural Language Processing (ANLP) and its Applications.

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. In *Proceedings of ArabicNLP 2023*, pages 52–75, Singapore (Hybrid). Association for Computational Linguistics.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. GPTAraEval: A comprehensive evaluation of ChatGPT on arabic NLP. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 220–247, Singapore. Association for Computational Linguistics.

Abdellah El Mekki, Samar M. Magdy, Houdaifa Atou, Ruwa AbuHweidi, Baraah Qawasmeh, Omer Nacar, Thikra Al-hibiri, Razan Saadie, Hamzah Alsayadi, Nadia Ghezaiel Hammouda, Alshima Alkhazimi, Aya Hamod, Al-Yas Al-Ghafri, Wesam El-Sayed, Asila Al sharji, Mohamad Ballout, Anas Belfathi, Karim Ghaddar, Serry Sibaee, and 28 others. 2026. Alexandria: A multi-domain dialectal arabic machine translation dataset for culturally inclusive and linguistically diverse llms. *Preprint*, arXiv:2601.13099.

Omer Nacar, Serry Taiseer Sibaee, Samar Ahmed, Safa Ben Atitallah, Adel Ammar, Yasser Alhabashi, Abdulrahman S. Al-Batati, Arwa Alsehibani, Nour Qandos, Omar Elshehy, Mohamed Abdelkader, and Anis Koubaa. 2025. Towards inclusive arabic LLMs: A culturally aligned benchmark in arabic large language model evaluation. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 387–401, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. 2020. AraBench: Benchmarking dialectal arabic-english machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5094–5107, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kees Versteegh, editor. 2005–2007. *Encyclopedia of Arabic Language and Linguistics*. Brill.

Stefan Weninger, Geoffrey Khan, Michael P. Streck, and Janet C. E. Watson, editors. 2012. *Semitic Languages: An International Handbook*. De Gruyter.

Jezia Zakraoui, Moutaz Saleh, Somaya Al-Maadeed, and Jihad M. Alja'am. 2021. Arabic machine translation: A survey with challenges and future directions. *IEEE Access*, 9:161445–161468.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine translation of arabic dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT '12)*, pages 49–59, USA. Association for Computational Linguistics.