Sudanese-Flores: Extending Flores+ to Sudanese Arabic Dialect

Hadia Mohmmedosman Ahmed Samil
Mila - Quebec AI Institute, Montreal, Canada
hadia.samil2@mila.quebec

David Ifeoluwa Adelani
Mila - Quebec AI Institute, McGill University & Canada CIFAR AI Chair
david.adelani@mila.quebec

Abstract

In this work, we introduce Sudanese-Flores, an extension of the popular Flores+ machine translation (MT) benchmark to the Sudanese Arabic dialect. We translate both the DEV and DEVTEST splits of the Modern Standard Arabic dataset into the corresponding Sudanese dialect, resulting in a total of 2,009 sentences. While the dialect was recently introduced in Google Translate, there are no available benchmark in this dialect despite it being spoken by over 40 million people. Our evaluation on GPT-4.1, Gemini 2.5 Flash, and open-weight models (LLaMA-3.3 70B and Qwen-3 80B) showed that while the performance from English to Arabic is impressive (more than 23 BLEU), they struggle on Sudanese dialect (less than 11 BLEU) in zero-shot settings. In the few-shot scenario, we achieved only a slight improvement in performance.

1 Introduction

Sudanese Arabic is a widely spoken variety in Sudan, yet it remains underrepresented in natural language processing (NLP) resources, limiting research on machine translation and multilingual modeling for this language. Despite the growing interest in low resource languages and Arabic dialects, most NLP benchmarks focus on Modern Standard Arabic (MSA) or other widely spoken dialects, leaving Sudanese Arabic largely unexplored. This gap motivates the need for parallel datasets that enable systematic evaluation of machine translation models for Sudanese Arabic.

In this work, we introduce Sudanese-Flores, an extension of the Flores+ evaluation benchmark that builds on Flores-200 (NLLB-Team et al., 2022) with Sudanese Arabic translations for all sentences in the original DEV & DEVTEST dataset. Our new data is primarily translated from Modern Standard Arabic (MSA), with additional verification from the original English split. The Sudanese-Flores enables translation between Sudanese Arabic and all other >200 languages in FLORES including the six directions: MSA ↔ Sudanese Arabic, Sudanese Arabic ↔ English and MSA↔ English. Our new dataset preserves the diversity of domains and sentence types present in FLORES+, ensuring a broad linguistic and cultural representation.

We perform zero-shot MT experiments using proprietary models (GPT-4.1, Gemini 2.5 Flash) as well as open-weight models (LLaMA-3.3, Qwen-3) to balance performance evaluation and reproducibility. Translation quality is evaluated using BLEU and ChrF metrics, highlighting both strengths and weaknesses across different directions. Our experiments show that LLMs struggle to translate from English to Sudanese Arabic compared to MSA, with over -13.0 BLEU drop in performance. Our results highlight the need to expand machine translation capabilities to more low-resource languages and dialects. By releasing Sudanese-Flores, we provide a benchmark that enables reproducible research, supports the development of inclusive multilingual NLP models, and contributes to expanding NLP coverage to underrepresented languages and dialects.

2 Related Work

Research on Arabic NLP has traditionally focused on MSA, often overlooking dialectal varieties. While more African Arabic dialects are represented in large-scale datasets such as Flores-200 (NLLB-Team et al., 2022) and SIB-200 (Adelani et al., 2024) (Tunisian Arabic, Moroccan Arabic, and Egyptian Arabic), Sudanese Arabic is not covered. In general, Sudanese Arabic is particularly underrepresented in NLP research such as MT, sentiment analysis & speech recognition. Existing dialectal corpora that include Sudanese Arabic (Jarrar et al., 2022) are relatively small and primarily designed for morphological

English	MSA	Sudanese Arabic
Such children are called "feral" or wild. Some feral children have been confined by people (usually their own parents); in some cases this child abandonment was due to the parents' rejection of a child's severe intellectual or physical impairment.	تمّ حبسُ بعضٍ الأطفال الضّالين من قبل أشخاصٍ (عادةً والديهم)؛ في بعضٍ الحالات، كان هذا التّخلي عن الطّفل بسبب رفض الوالدين لإعاقة الطّفل العقليّة أو الجسديّة الشديدة، ويطلق على هؤلاء الأطفال اسم ""الوحشي"" أو البريّ.	حبسوا شوية **شفع** ضالين من ناس (عادة والديهم)؛ في بعض الحالاتٍ، كان سبب التخلي هو رفض الوالدين لأعاقة الشافع العقليّة او الجسدية الشديدة، وبسموا الشفع ديل ""الوحشي"" او البري
Lion prides act much like packs of wolves or dogs, animals surprisingly similar to lions (but not other big cats) in behavior, and also very deadly to their prey.	قطيع الأسود يعمل مثل قطيع الذئاب أو الكلاب، والمفاجئ أن الحيوانات مشابهة للأسد (لكن ليس القطط الكبيرة الأخرى) في السلوك ومميتة جداً لفرائسها.	قطيع الاسود بشتغل زي قطيع الذئاب او الكلاب، و المفاجئ انو الحيوانات بتشبه الاسد (لكن ما **الكدايس** الكبيرة التانية) في السلوك و مميتة خالص لفرايسها.
It warns that no one can guarantee that any course of action in Iraq at this point will stop sectarian warfare, growing violence, or a slide toward chaos.	يحذر من أنه لا أحد يمكنه ضمان أن أي مسار عمل في هذه اللحظة في العراق سيؤدي إلى وقف الحرب الطائفية أو العنف المتزايد أو الانحدار نحو الفوضى.	يحذر من انو مافي **زول** بقدر يضمن انو اي مسار عمل في اللحظة دي في العراق حيادي لي وقف الحرب الطائفية او العنف المتزايد او الانحدار ناحية الفوضى.

Table 1: **Example sentences from Sudanese-FLORES.** In the Sudanese Arabic column, **dialect-specific words are highlighted in bold** to show the linguistic differences from MSA. These words reflect local lexical choices (e.g., **زول ، الكدايس ، الشفع**) and demonstrate authentic Sudanese Arabic expressions captured in the dataset. To the best of our knowledge, these words are used only in Sudan.

analysis rather than MT, highlighting the need for larger, translation-focused datasets. Similarly, other large-scale pre-training data covering over 400 languages, such as FineWeb2 (Penedo et al., 2025) and MADLAD-400 (Kudugunta et al., 2023), do not include the Sudanese dialect.

3 Dataset and Experimental Setup

3.1 Dataset creation

Sudanese-Flores was created by a single native Sudanese speaker, who is trilingual (speaks English, Arabic and Sudanese dialect). The translations represent the Khartoum/Central Sudanese dialect variety, which is widely understood across Sudan and neighboring regions. The data set contains 1,012 sentences in the DEV set and 997 sentences in the DEVTEST set. Although she is fluent in the three languages, for ease of translation, she translated Arabic into Sudanese since they are closely related. In cases where the MSA text was ambiguous, the English source was consulted to ensure the meaning was preserved. This translation took between three to four weeks, with feedback from other native speakers during the process. Table 1 provides examples that highlight some differences between the MSA and Sudanese Arabic dialect.

Additionally, 100 sentences were randomly selected for verification by a second native speaker, who confirmed that each sentence preserved the original meaning and used natural Sudanese Arabic. We note that there is no standardized way to write Sudanese Arabic and some letters have the same pronunciation. These differences in writing do not cause any issues in understanding, and readers can clearly interpret the meaning of all sentences.

3.2 Experiments Setup

We prompted two leading LLMs in both zero-shot and few-shot (5-shots): GPT-4.1 and Gemini 2.5 Flash. In addition, we also evaluate open-weight large language models, including Llama-3.3 and Qwen-3, to improve reproducibility and accessibility. We make use of a single prompt. The same prompt template is used for all models and all translation directions. The prompt we used is very simple, obtained from AfroBench paper (Ojo et al., 2025):

```
You are a translation expert. Translate
    the following {{source_lang}}
    sentences to {{target_lang}}

{{source_lang}} sentence: {{source_text
    }}
{{target_lang}} sentence:
```

For few shots, we added five examples in the DEV set into the prompt. We note that while few-shot prompting can improve performance in some cases, it can also negatively affect results, particularly for dialectal translation, suggesting that few-shot prompting does not consistently generalize across

		Arabic - English		English - Arabic		MSA - Sudanese	
Models	**Setting**	**arb-eng**	**apd-eng**	**eng-arb**	**eng-apd**	**apd-arb**	**arb-apd**
GPT-4.1	0-shot	39.9 / 68.1	32.3 / 63.0	**27.9 / 59.7**	10.4 / 46.6	**54.4 / 78.7**	24.2 / 64.1
	5-shots	41.7 / **69.0**	**36.5 / 65.8**	24.8 / 58.6	**12.0 / 47.9**	46.9 / 75.2	24.1 / 63.1
Gemini 2.5 Flash	0-shot	24.4 / 60.1	21.5 / 57.2	23.5 / 56.4	9.7 / 46.2	47.0 / 74.3	27.0 / **66.4**
	5-shots	40.3 / 67.9	35.1 / 64.3	24.7 / 58.0	11.9 / 47.3	49.2 / 76.2	25.2 / 63.2
Llama-3.3 70B	0-shot	42.6 / 66.5	32.9 / 59.9	23.5 / 51.7	8.1 / 37.9	44.1 / 69.0	17.7 / 48.8
	5-shots	**43.2** / 66.2	36.0 / 61.2	23.6 / 51.6	10.7 / 40.3	48.1 / 71.6	**29.3** / 61.2
Qwen-3-Next-80B-A3B	0-shot	40.7 / 64.5	33.5 / 59.1	24.1 / 52.4	8.7 / 39.2	41.2 / 66.8	16.9 / 50.1
	5-shots	40.6 / 64.3	34.7 / 59.6	24.1 / 52.4	8.2 / 38.9	45.2 / 69.4	18.8 / 53.0

Table 2: **MT translation BLEU / ChrF++ performance on Sudanese-Flores** across six translation directions: Arabic dialects → English (arb-eng & apd-eng), English → Arabic dialects (eng-arb & eng-apd), and MSA↔Sudanese Arabic (apd-arb & arb-apd). The results reported on DEVTEST set. The best result are in **Bold** and the second best are underlined.

translation directions for Sudanese Arabic.

We evaluate six translation directions to comprehensively assess model performance: MSA↔English, Sudanese Arabic↔English, and MSA ↔ Sudanese Arabic. The language directions reported in Table 1 are ordered by source–target pairs and include all combinations of arb, apd, and eng.

3.3 Evaluation Metrics

We measure the translation quality using BLEU (SacreBLEU) (Post, 2018) and ChrF++ (Popović, 2017).

4 Results and Discussion

Table 2 shows the result of GPT-4.1 and Gemini 2.5 Flash as well as open-weight models (Llama-3.3 and Qwen-3) on Sudanese-Flores in a zero-shot and few-shot settings across the six translation directions. Results are reported on the DEVTEST split using BLEU and ChrF metrics.

GPT-4.1 generally outperforms Gemini 2.5 Flash In zero-shot settings, we find GPT-4.1 is able to translate better than Gemini-2.5 Flash in almost all directions except arb-apd. In the apd-eng direction, the performance gap in BLEU points is more than +10, however, in eng-apd, it is lower. This finding is similar to the observation of AfroBench-Lite (Ojo et al., 2025)

Proprietary LLMs are consistently better than Open LLMs In almost all settings, we find the proprietary LLMs such as GPT-4.1 and Gemini 2.5 Flash having better performance than the open LLMs such as Llama 3.3 70B and Qwen 3 80B. However, with few shots, there is significant boost in performance for Llama 3.3 70B with at least +2.0 increase for direction into Sudanese dialect, especially for the arb-apd direction. On the other hand, Qwen 3 80B gains after few shots are not consistent and often smaller.

Translation from English into Sudanese dialect is challenging We find a big gap in performance between translating from eng-arb and eng-apd with almost -17 BLEU point drop with GPT-4.1 and -13 points with Gemini-2.5 Flash in zero-shot settings. A similar degradation is observed for open-weight models, indicating that translation into Sudanese Arabic remains challenging across different model families. Surprisingly, few-shot did not seem to help much although there are some slight improvements.

Translation from Sudanese dialect to English seems easier Given the similarity of MSA and Sudanese dialect, it is unsurprising that LLMs are able to easily translate into English. Although, the performance is lower than from MSA, it is noticeably high (i.e. over 32 BLEU with GPT-4.1 and over 21 BLEU with Gemini).This trend is consistent across models, though performance remains lower than translation from MSA.

Few-shot is not so helpful, and may even hurt performance The easiest direction we find is from MSA to Sudanese, and vice versa. However, with few-shots, sometimes it helps in the direction of apd-arb while in some other cases it severely hurt the performance when generating Sudanese. This further suggests that few-shot prompting does

Source → Target	Example Source	Model Output	Reference	Type of Error	Observations
Sudanese → Arabic	المصطلح دا مشتق من المعرفة القديمة بيق الفراش	المصطلح مشتق من المعرفة القديمة بيق الفراش	هذا المصطلح مشتق من المعرفة القديمة بيق الفراش	MSA hallucination	Insertion of standard determiner; meaning preserved
Sudanese → English	كيرتس كوبر قام باكتشاف اكبر عدد اولى معروف	Curtis Cooper… discovered the largest known prime number to date…	Curtis Cooper… has discovered the largest known prime number to date…	Tense mismatch	Present perfect reduced to past tense
Arabic → English	قام كيرتس كوبر باكتشاف أكبر عدد أولي معروف	Curtis Cooper… discovered the largest known prime number to date…	Curtis Cooper… discovered the largest known prime number to date…	Minimal	Stylistic variation only
English → Arabic	There, they found the body of Saroja Balasubramanian…	هناك، وجدوا جثة **ساروجا** بالاسوبرامانيان	وهناك وجدوا جثة ساروجا بالاسوبرامانيان	Punctuation / style	Comma insertion; meaning preserved
English → Sudanese	The find also grants insight into the evolution of feathers in birds.	الاكتشاف **ده** كمان بيعطي نظرة على تطور الريش في الطيور	الاكتشاف بدي برضو نظرة على تطور الريش في الطيور.	Dialectal variation	Alternative Sudanese lexical realization
Arabic → Sudanese	جميع سكان مدينة الفاتيكان من الروم الكاثوليك.	كل سكان مدينة الفاتيكان من الروم الكاثوليك	جميع سكان مدينة الفاتيكان من الروم الكاثوليك.	Minor	Dropped period; meaning preserved

Table 3: **Qualitative error analysis across translation directions.** Each entry shows the source sentence, model output, reference, type of error, and observations describing the deviation. Dialect-specific or stylistic deviations are highlighted in the observations column; Sudanese-specific words are bolded in the model output.

not consistently generalize across translation directions for Sudanese Arabic. We leave further investigation of prompting strategies for future work.

5 Qualitative Error Analysis

To better understand model behavior on dialectal data, we perform a qualitative error analysis with representative examples shown in Table 3. The model exhibits occasional **MSA hallucination** when translating from Sudanese Arabic, inserting standard forms not present in the source. For Sudanese outputs, errors mainly involve **dialectal realization**, including inconsistent lexical choices and weakening of Sudanese-specific idioms or morphological markers, while core meaning is preserved. Overall, errors are largely stylistic rather than semantic and are most pronounced in Sudanese-related directions due to dialectal variability and limited training data.

6 Conclusion

In this paper, we introduce, Sudanese-Flores, a new benchmark that extends Flores-200 to Sudanese Arabic dialect. Our evaluations clearly shows the need for such a benchmark, since current LLMs struggle to generate Sudanese dialect, especially when translating from English which is also an official language in Sudan.

7 Acknowledgement

This research was supported by Mila-Quebec AI Institute P2v5 grant. We are grateful for the support of Siva Reddy throughout the project.

Limitations

While Sudanese-Flores offers a valuable benchmark for Sudanese Arabic MT, it has some inherent constraints. The dataset was primarily created by a single native speaker, which may limit dialectal variation, though we included verification from a second native speaker to ensure naturalness. The dataset size (2,009 sentences) is modest, reflecting the resource-scarce nature of this dialect, but it is sufficient for benchmarking and evaluation. Our focus on the Khartoum/Central Sudanese dialect ensures broad intelligibility across Sudan, though regional variations may exist. Finally, the orthography of Sudanese Arabic is not fully standardized, which may affect surface-level evaluation but does not hinder semantic understanding.

Ethical Considerations

We prioritize the linguistic and cultural integrity of Sudanese Arabic. Translations were performed by a native speaker and verified by a second native speaker to ensure accuracy and naturalness. The dataset contains no personally identifiable information or sensitive content. Users should be aware that MT models may reproduce biases present in their training data; careful evaluation is advised when applying models in real-world contexts involving this dialect.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Mustafa Jarrar, Fadi A. Zaraket, Tymaa Hammouda, Daanish Masood Alavi, and Martin Wählisch. 2022. Lîsan: Yemeni, iraqi, libyan, and sudanese arabic dialect corpora with morphological annotations. *Preprint*, arXiv:2212.06468. Accessed: 2025-12-20.

Sneha Kudugunta, Isaac Rayburn Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and document-level large audited dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

NLLB-Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. AfroBench: How good are large language models on African languages? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19048–19095, Vienna, Austria. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all — adapting pre-training data processing to every language. In *Second Conference on Language Modeling*.

Maja Popović. 2017. chrF++: Towards character and word n-gram F-score for automatic mt evaluation. In *Proceedings of the Second Conference on Machine Translation*, pages 568–575. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. Association for Computational Linguistics.