Benchmarking IsiXhosa Automatic Speech Recognition and Machine Translation for Digital Health Provision

Abby Blocker,[1*] Francois Meyer,[1] Ahmed Biyabani,[2] Joyce Mwangama,[1]
Mohammed Ishaaq Datay,[1] Bessie Malila[1]
[1]University of Cape Town, Cape Town, South Africa
[*]blcabb001@myuct.ac.za, {firstname.lastname}@uct.ac.za
[2]Carnegie Mellon University Africa, Kigali, Rwanda
ab3x@andrew.cmu.edu

Abstract

As digital health becomes more ubiquitous, people from different geographic regions are connected and there is thus a need for accurate language translation services. South Africa presents opportunity and need for digital health innovation, but implementing indigenous translation systems for digital health is difficult due to a lack of language resources. Understanding the accuracy of current models for use in medical translation of indigenous languages is crucial for designers looking to build quality digital health solutions. This paper presents a new dataset [1] with audio and text of primary health consultations for automatic speech recognition and machine translation in South African English and the indigenous South African language of isiXhosa. We then evaluate the performance of well-established pretrained models on this dataset. We found that isiXhosa had limited support in speech recognition models and showed high, variable character error rates for transcription (26-70%). For translation tasks, Google Cloud Translate and ChatGPT outperformed the other evaluated models, indicating large language models can have similar performance to dedicated machine translation models for low-resource language translation.

1 Introduction

Digital health has been recognized to improve access to healthcare services by decreasing wait times, improving care quality, and reducing cost (Erku et al., 2023; Caffery et al., 2016; Gentili et al., 2022). Many digital health initiatives have focused on improving access in under-resourced areas, which face some of the largest challenges in providing healthcare services (Maita et al., 2024). However, as patients in under-resourced areas are connected to healthcare providers in various locations, language barriers present a serious challenge to be considered.

In South Africa, 84.3% of the population is reliant public health facilities, many of which are under-resourced (Stats SA, 2023). There are 12 official languages of South Africa, with 9 of these being indigenous languages (Stats SA, 2022). Incorporating language translation services for the indigenous languages of South Africa within digital health solutions is not only helpful but necessary. However, there isn't a clear consensus on what the best available tools are for integrating translation services for digital health in South African languages.

The aim of this paper is to understand the performance of automatic speech recognition (ASR) and machine translation (MT) services by assessing currently available pretrained models on South African English and isiXhosa, a South African indigenous language. Our contributions include a new dataset, consisting of audio and text in South African English and isiXhosa to support further development and evaluation of ASR and MT models.[1] The results indicate that for ASR, error rates for South African English are comparable to human transcription; but, for isiXhosa, error rates are above an acceptable range, particularly for use in the medical field. For MT, large language models (LLMs) showed

[1] https://github.com/blocker-abby/xh-en-health-data/

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 169–179

comparable results to dedicated MT models, and the commercially available models outperformed the open-source models evaluated.

2 Background and Related Works

2.1 ASR

A widely used open-source ASR model is Whisper, developed by OpenAI (Radford et al., 2022). Whisper supports ASR for South African English, but not isiXhosa. Whisper cites a 9.3% error rate for English, but English spoken with African accents showed lower accuracy rates (Afonja et al., 2024). Therefore, assessment of South African English accents specifically is necessary to verify these results, and particularly on health-domain-specific data. In addition to Whisper, the Massive Multilingual Speech (MMS) model is an open-source ASR model developed by Meta, which supports South African English and isiXhosa. Pratap et al. (2023) demonstrated that MMS had higher accuracy when compared to Google and Whisper when using the FLEURS dataset (which includes isiXhosa data).

In addition to open-source models, there are several successful commercial models for ASR. Particularly, the leaders in commercial cloud computing offer ASR APIs, these being Google Cloud Platform (GCP), Microsoft Azure, and Amazon Web Services (AWS) (Borra, 2024). Out of these, only GCP offers ASR for isiXhosa. These commercially available models have been cited to have better performance for ASR when compared to open-source models (Ferraro et al., 2023).

2.2 Translation

In the translation domain, the development of massive multilingual neural machine translation (NMT) models has contributed to improved translation of low-resource languages like isiXhosa. Meta's No Language Left Behind (NLLB) is an open-source NMT model which provides translation for 200 languages, many of which are low-resource (Costa-jussà et al., 2022). In the commercial translation space, GCP, Azure, and AWS all offer translation APIs. Two of these (GCP and Azure) offer services for isiXhosa translation. Open source and commercial models have been cited to have similar performance in the translation domain (Licht et al., 2024).

Current research has investigated the use of LLMs such as ChatGPT for translation tasks. Some research has found that they have high accuracy in comparison to NMTs (Wang et al., 2023). However, experiments with low-resource and African languages (of which isiXhosa is both) have shown results that still lag behind dedicated MT models like NLLB (Robinson et al., 2023; Ojo et al., 2024).

2.3 Healthcare Applications

ASR and MT in the healthcare sector is a debated topic. Accuracy in healthcare communication is vital, as miscommunication has the potential to drastically affect medical decisions and could lead to negative outcomes. Some healthcare bodies recommend against these techniques because of the risk (Vieira et al., 2019). However, when used responsibly, ASR and MT services can provide benefits in environments where human translation services cannot be provided, either due to resource constraints or lack of expertise. Recommendations for healthcare providers using these services include being aware of the potential errors, being alert to non-verbal communication from the patient, and for translation, back-translating (inputting translated materials into the MT model for translation back into the source language) to analyze where errors may have occurred (Randhawa et al., 2013). Therefore, it is important to understand the current state of ASR and MT, in order to apply it to digital health solutions safely.

Understanding the development context is also important in determining the best-fit ASR and MT models for digital health applications. While accuracy is extremely important, there are other additional factors which can influence the uptake of solutions. The mobile application AwezaMed provides an example of this. The app provides translation of medical text for all South African languages using a list of predefined phrases (Marais et al., 2020). While there are benefits to the accuracy of using static translations, including the ability for human validation, there are also difficulties in that real-time and customized translation is not possible. In a real-time digital health application such as telemedicine, this solution may not address the needs of users; therefore, it is important to consider other factors along with accuracy to select the most appropriate translation models for digital health solutions.

3 Method

3.1 Data

Conversations between primary health care providers and patients were used as evaluation data. Conversation data was adapted from the PriMock57 dataset (Korfiatis et al., 2022), which provides audio and transcribed conversational data from mock telemedicine consultations. Ten random consultations were chosen from the available dataset of 57. The text data from each consult was then translated by a professional human translator with experience in English-isiXhosa medical translation.

While audio files of the consultations were available in the PriMock57 dataset, the speakers were not South African. As spoken accents can affect the accuracy of ASR models, it was important to utilize authentic audio of South African English speakers. Therefore, the conversations were re-enacted between South African paid actors. A total of 5 actors (3 male, 2 female) were used, with two actors (one acting as the doctor, and one acting as the patient) per consultation. Two of the three male actors were included only in South African English recordings. The other male actor was included only in the isiXhosa recordings. The two female actors were included in both South African English and isiXhosa recordings. Each of the actors were fluent in the languages they recorded in. The actors read the consultation dialogue exactly as it was stated in the written text. Where speaking errors were made, this was cleaned in post-processing of the audio file using Audacity.[2] Audio was saved as a stereo, 48kHz sampled FLAC file. Azure speech-to-text and MMS required a 16kHz sample WAV file input, so the audio was also converted to this format during evaluation of both models.

Text data of the conversations was subdivided based on conversational dialogues. Each time the speaker changed, the text data was separated into a new text for evaluation. This resulted in a total of 580 English texts and 580 isiXhosa texts. Each text was input into each model once and the first output result was used for evaluation.

3.2 Selected ASR Models

The chosen models for evaluation are highlighted in Table 1. The chosen models for ASR of South African English were Google Cloud Speech-to-Text v1,[3] Microsoft Azure AI Speech's speech-to-text,[4] Whisper base model (Radford et al., 2022), and MMS speech-to-text (Pratap et al., 2023). Not all of the four chosen models offered isiXhosa services; those that did were Google Cloud Speech-to-Text v1 and MMS speech-to-text.

Whisper allowed for prompting capabilities, while the other ASR models did not. When providing an audio file input to Whisper, it is recommended to also provide a list of expected words to improve accuracy. The model was evaluated both with and without using this prompting feature. The expected terms used for

Automatic Speech Recognition			
Model	**Developer**	**Availability**	**Supported Language**
Google Cloud Speech-to-Text v1	Google	Commercial	en, xh
Azure AI Speech speech-to-text	Microsoft	Commercial	en
Whisper base	OpenAI	Open Source	en
MMS	Facebook	Open Source	en, xh

Machine Translation			
Model	**Developer**	**Availability**	**Type**
Google Cloud Translate v2	Google	Commercial	Dedicated MT
Azure Translator	Microsoft	Commercial	Dedicated MT
NLLB 200M distilled 600M	Facebook	Open Source	Dedicated MT
ChatGPT GPT-4o	OpenAI	Commercial	LLM
Gemini Flash 1.5	Google	Commercial	LLM

Table 1: Selected Models for Evaluation

[2] https://www.audacityteam.org/

[3] https://cloud.google.com/speech-to-text?hl=en

[4] https://learn.microsoft.com/en-us/azure/ai-services/speech-service/speech-to-text

prompting were selected from the South African Department of Sport, Arts, and Culture's medical terms list.[5] This document translates medical terms into 10 of the 12 South African languages. The list was reduced to include only terms contained within the dataset, which totaled 39 unique English terms and 56 unique isiXhosa terms (given that some of the terms had multiple translations). Then, for each transcription, only the terms from the list included within the ground truth were included in the prompt input.

3.3 Selected MT Models

The chosen models for translation, featured in Table 1, were Google Cloud Translate v2, [6] Microsoft Azure Translator,[7] NLLB-200 distilled 600M (Costa-jussà et al., 2022), ChatGPT GPT-4o mini, and Gemini Flash 1.5.[8]

Given that ChatGPT and Gemini are LLMs, they require a prompt input to provide instructions rather than only the text to be evaluated. A modified prompt used by Robinson et al. (2023) was used for LLM translation, which was the following: **"This is an [*source language*] to [*target language*] translation, please provide the [*target language*] translation for this sentence. Do not provide any explanations or text apart from the translation. [*Translation text*]."** In addition to this prompt, a modified prompt was also tested by providing language pairs in English and isiXhosa. The language pairs were selected from the medical terms list translations, with only the terms in the input text being included in the prompt. This modified prompt added the following text before supplying the text to be translated: **"In this context, [*source language term*] translates to [*target language term*]."**

3.4 Evaluation Metrics

Two metrics were employed for evaluating ASR, as English and isiXhosa languages have different characteristics which are better explained by different methodologies. The standard measure for ASR evaluation is word error rate (WER). However, WER does not fully characterize ASR results for agglutinative languages such as isiXhosa (Thennal et al., 2024). This is because words in isiXhosa have prefixes and suffixes that often correspond to individual words in English. Therefore, WER may incorrectly inflate the error rate of ASR for isiXhosa in comparison to English. To address this, both WER and character error rate (CER) were calculated for isiXhosa transcriptions. Both metrics were calculated using the HuggingFace evaluate library (Von Werra et al., 2022).

For translation, character level F-score (CHRF++) and bilingual evaluation understudy (BLEU) were used to evaluate model performance (Callison-Burch et al., 2007). The original and human-translated texts were used as ground truth comparisons. To address the agglutinative structure of isiXhosa, CHRF++ was chosen as it accounts for both character and word accuracy (Popović, 2015).

Because the analysis aimed to understand model performance on health domain data, an analysis was also conducted on the error rate of models in transcribing and translating health terms. It is critical that this terminology be transcribed and translated correctly, as it has the potential to affect medical decision-making. Results were analyzed based on the list of health terms used for modified prompting. Error rate was calculated by dividing the occurrences of each health term in the resultant text by the occurrences in the ground truth text and subtracting from 100. Furthermore, because some health terms had multiple translations from English to isiXhosa, any of the isiXhosa translations were accepted for the accuracy measure. The type of isiXhosa translation used in the result was also noted and categorized into one of three types: an isiXhosa term; a borrowed English word with isiXhosa spelling; or a borrowed English word with English spelling. Additionally, all health terms in both languages were classified into the following three categories: anatomy; condition; or treatment.

The average costs for transcription and translation were calculated using available pricing for commercial MT models. For LLMs, tokens used per character were calculated for each prompt and then converted to price per character based on the model pricing. Open source models were not

[5] https://www.dsac.gov.za/sites/default/files/2023-11/Multilingual%20Pharmaceutical%20Terminology%20List_0.pdf

[6] https://cloud.google.com/dotnet/docs/reference/Google.Cloud.Translation.V2/latest

[7] https://azure.microsoft.com/en-us/products/ai-services/ai-translator

[8] https://ai.google.dev/gemini-api/docs/models/gemini#gemini-1.5-flash

included in the cost analysis, although it is acknowledged that running open-source models on local machines does incur associated costs.

From the results, models which showed appropriate accuracies for digital health systems were implemented into an existing virtual clinic system web application (Blocker et al., 2024). The system involved ASR for South African English and translation of English and isiXhosa text. The system takes user input (either audio or text) and sends a request to the backend with the data. The backend then either processes the data (in the case of open-source models) or creates an additional request and sends the data to the cloud computing service (for commercial models). When the response is received, it is returned to the front-end and displayed for the user. The time taken for each model to return a text response to the front end was measured in milliseconds for each model. For commercial implementations (Azure and GCP), real-time translation methods were utilized instead of batch translations.

4 Results

4.1 ASR Model Error Rates

Figure 1 presents the WER of South African English. Lower WER indicates higher accuracy of the ASR model. The lowest WERs for South African English were achieved by Whisper with prompting (7.1%) and Azure (7.6%). "Quick" human transcription of conversational speech has been cited with a WER of 9.6% (Stolcke & Droppo, 2017), indicating that the results from these models concur with human transcription. There was a 4.5% difference in WER between using Whisper with and without prompts. Similar prompting techniques were attempted with GCP and Azure ASR models using phrase lists; however, both models produced identical transcriptions regardless of whether phrase lists were employed. Results for South African English ASR by GCP ranged from 17.33-25.34%, which agrees with the literature range of approximately 15-25% WER (Filippidou & Moussiades, 2020). Results for Whisper (without prompting) were slightly higher than the cited metric of 9.3% for English (Radford et al., 2022); however, this reported value was for general English (en), not South African English (en-za).

Figure 2 presents the measured WER and CER for isiXhosa transcription. WER for both GCP and MMS were greater than human WER. CER as an evaluation metric for ASR is less common than WER, therefore there is not a generally accepted human error rate for comparison. However, ASR models in literature for isiXhosa transcription report CER values ranging from 13.8-40.7% (Reitmaier et al., 2022; Jacobs et al., 2025; Baas & Kamper, 2022). GCP and MMS had averages of 43.7% and 51.4% CER respectively. These results are higher than those reported in literature, which highlights the challenge encountered in translating health-domain-specific conversations. The range of results is much wider than that seen for South African English, with a 45% difference between the lowest and highest error rate for isiXhosa. Given that the human WER is 9.3%, and CER

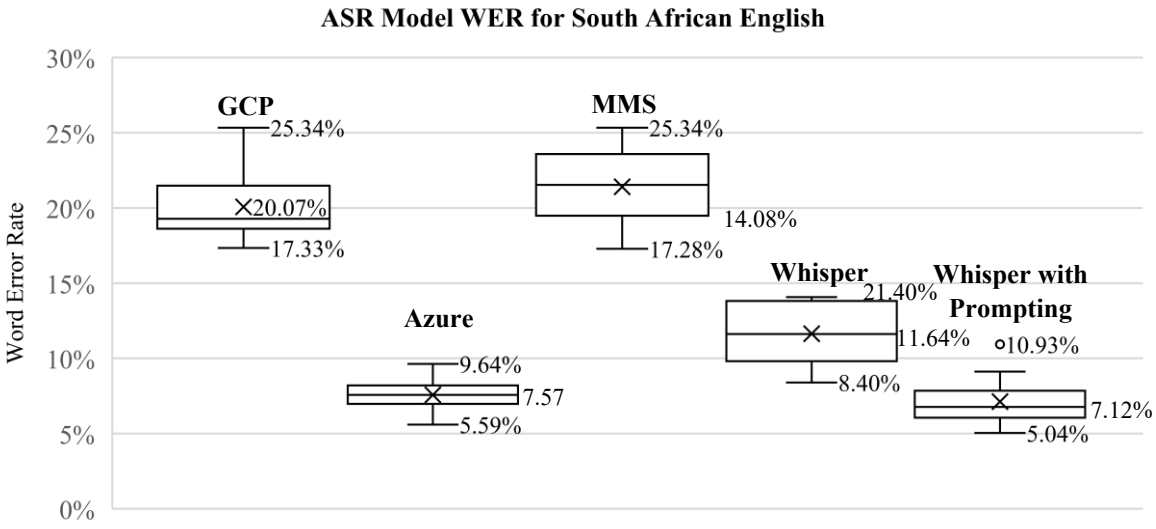

Figure 1: Measured WER for transcription of South African English

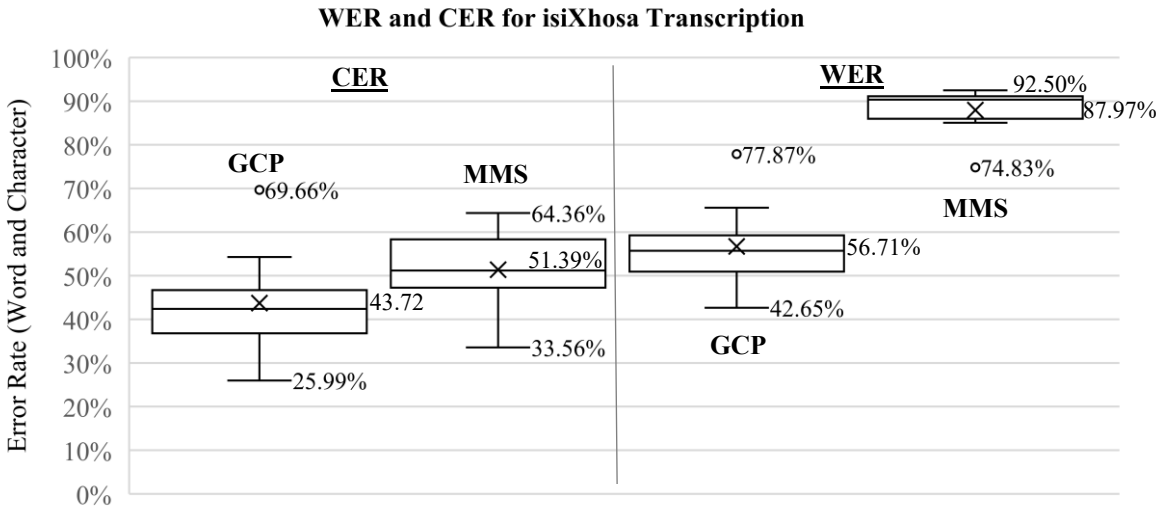

Figure 2: Measured WER and CER for transcription of isiXhosa.

tends to be lower than WER (Ravanelli et al., 2024), this indicates that neither model performed adequately for isiXhosa ASR.

Performance of commercial versus open-source models did not follow a clear trend. Both Whisper (open-source) and Azure (commercial) achieved low WERs for South African English, while GCP (commercial) and MMS (open-source) had higher error rates for both South African English and isiXhosa.

ASR model results were also assessed for health term error rate, both generally and within the three categories – anatomy, condition, or treatment. Results for health term error rate are provided in Table 2. For transcription of South African English, Whisper with prompting had the lowest health term error rate at 5.39%, followed by Azure with 9.47%. Whisper had a 10.21% decrease in error rate when health terms were introduced into the prompt. For isiXhosa transcription, GCP had lower error rate than MMS. Treatment terms, which mainly consisted of medication names (Paracetamol, Ibuprofen, Metformin, etc.) had high error rates when transcribed from isiXhosa audio. For both South African English and isiXhosa, MMS had high error rate in transcribing treatment terms; this could be due to the nature of the training dataset used for MMS, which was domain-specific and not general. Overall, the models evaluated had acceptable performance for transcribing medical conversations in South African English, but struggled in transcribing isiXhosa medical conversations.

4.2 Translation Model Results

Table 3 provides CHRF++ and BLEU results for MT. Higher scores for both metrics indicate that predicted translations are closer to the ground truth translations. For English to isiXhosa, Google Cloud Translate reported the highest scores,

South African English				
Model	**Overall**	**Anatomy**	**Conditions**	**Treatment**
GCP	26.02%	20.9%	39.18%	19.01%
Azure	9.47%	4.70%	18.71%	9.63%
MMS	50.34%	26.07%	77.00%	96.30%
Whisper	15.6%	8.12%	19.21%	40.00%
Whisper with prompting	**5.39%**	**2.99%**	**12.50%**	**0.00%**

isiXhosa				
Model	**Overall**	**Anatomy**	**Conditions**	**Treatment**
GCP	**58.84%**	**38.98%**	**78.65%**	**96.67%**
MMS	76.27%	67.73%	81.92%	100.00%

Table 2: Error rate for ASR of health terms.

Model	CHRF++ Score		BLEU Score	
	English to isiXhosa	isiXhosa to English	English to isiXhosa	isiXhosa to English
Google Cloud Translate	**63.79**	57.23	**0.284**	**0.286**
Azure	56.31	53.56	0.168	0.233
NLLB	48.39	50.84	0.081	0.213
ChatGPT	51.91	**57.64**	0.115	0.270
ChatGPT (mod)	52.38	57.59	0.114	0.267
Gemini	48.50	54.64	0.074	0.245
Gemini (mod)	48.75	54.93	0.075	0.248

Table 3: CHRF++ and BLEU scores for translation between English and isiXhosa.

comfortably outperforming all other models. NLLB, Gemini, and Gemini with modified prompting had the lowest scores, with a difference of 16.15 between highest and lowest average score. For isiXhosa to English, the performance was less distributed, with the difference between highest and lowest scores at 6.8. ChatGPT and Google Cloud Translate were the highest scoring models, and NLLB the lowest scoring model.

The only open-source translation-dedicated model tested, NLLB, had generally lower scores than the commercial models evaluated. In comparing translation-dedicated models to LLMs, ChatGPT had higher scores when compared to Azure and NLLB, for translation of isiXhosa to English, but this did not carry over to English to isiXhosa translation. Between LLMs, ChatGPT had higher scores than Gemini. Modified prompting did not have a significant effect on the overall score.

Health term error rate was also calculated for translation results, with lower error rates indicating more accurate translations of health terms. Health term error rate decreased when using modified prompts with both ChatGPT and Gemini LLMs, as shown in Table 4. Google Cloud Translate had the lowest error rate of all evaluated models for English to isiXhosa translation, with a 10% difference in error rates between the next best performing model, Azure. This is in contrast to isiXhosa to English translation, where the top 4 performing models in terms of health term accuracy (ChatGPT, ChatGPT with modified prompts, Gemini with modified prompts, and Google Cloud Translate) were within 5% error rate of one another. Generally, health term accuracy

isiXhosa to English				
Model	**Overall**	**Anatomy**	**Condition**	**Treatment**
ChatGPT	16.61%	14.13%	32.14%	**0.00%**
ChatGPT with modified prompt	**11.82%**	**10.87%**	20.71%	**0.00%**
Gemini	25.51%	26.09%	37.86%	**0.00%**
Gemini with modified prompt	13.18%	14.13%	**15.00%**	5.26%
Google Cloud Translate	14.44%	11.23%	30.71%	**0.00%**
Azure	20.55%	11.96%	54.29%	**0.00%**
NLLB	36.47%	23.91%	71.43%	32.89%

English to isiXhosa				
Model	**Overall**	**Anatomy**	**Condition**	**Treatment**
ChatGPT	49.14%	43.50%	78.86%	18.18%
ChatGPT with modified prompt	38.14%	32.55%	64.13%	17.39%
Gemini	62.32%	59.67%	93.48%	13.64%
Gemini with modified prompt	55.17%	51.17%	82.98%	18.18%
Google Cloud Translate	**18.99%**	**14.91%**	**38.04%**	**4.55%**
Azure	29.25%	23.04%	58.70%	**4.55%**
NLLB	58.91%	50.60%	87.68%	46.36%

Table 4: Health term error rate for translations.

was lower for translations from English to isiXhosa compared to isiXhosa to English.

The error rate for each health term category is also depicted in Table 4. Treatments (which mainly consisted of medications) had low error rates, with <10% error rate for isiXhosa to English translation for all models excluding NLLB. Highest error rates were seen with the translation of conditions from English to isiXhosa. This included terms for both diseases (i.e. diabetes, asthma, stroke) and symptoms (i.e. cough, headache, pain). For all models, translation from isiXhosa to English had lower health term error rates (for all term classifications) than translation from English to isiXhosa. IsiXhosa health terms were categorized further into three types – borrowed English terms with English spelling (i.e., i-Paracetamol, meaning Paracetamol); borrowed English terms with isiXhosa spelling (i.e., ifiva, meaning fever) and isiXhosa terms (i.e., isisu, meaning stomach). Borrowed English words with isiXhosa spelling were not used frequently by any of the models; both borrowed English terms with English spelling and isiXhosa words were used more frequently.

4.3 Cost

There are other factors besides accuracy that one might consider when choosing systems for ASR and MT. Particularly when considering commercial solutions, cost is an important factor.

ASR (per minute)	
Model	**Cost**
GCP	Tiered pricing ranging from $0.016-$0.004 per minute
Azure	$0.01667 per minute with 5 hours per month free
Whisper	Associated computing costs
MMS	Associated computing costs

Translation (per million characters translated)	
Model	**Cost**
GCP	$20 (first 500k characters per month free)
Azure	$10 (first 2M characters per month free)
NLLB	Associated computing costs
ChatGPT 4o mini	$1.84
Gemini 1.5 Flash	$0

Table 5: Pricing of evaluated models.

Table 5 compares the cost of the various models evaluated. For ASR, GCP and Azure have similar costs, with GCP offering slightly lower rates for higher volumes of audio. Whisper is unique in that it is open source, so it can be run on a local machine or accessed through OpenAI's API. Running Whisper or MMS (open source) models on a local machine would incur costs for electricity and hardware. For MT, GCP and Azure can provide translation free of cost for low volumes of data (<500k and <2M characters, respectively). However, for larger volumes of translation, ChatGPT 4o mini provides a cheaper per-character rate at only $1.84 per million characters. Gemini 1.5 Flash is free to use, offering the cheapest commercial option for translation.

4.4 Latency

The South African English ASR models (excluding MMS) and the four commercial translation models were implemented in the system as part of a language translation feature. Figures 3 and 4 depict the measured latencies when using each model in the end-to-end translation system. Microsoft Azure offered the lowest latencies for both ASR and MT compared to the other evaluated models, though occasionally latency could be over 10 seconds for transcribing long audio clips. ASR latency was much higher than MT, but likely because there was some post-processing formatting that occurred before transcription. Additionally, requests with text data are smaller in size than their audio data counterparts, so sending a larger request over the network incurs greater time.

5 Discussion and Conclusion

Based on the evaluation performed, we found that Microsoft Azure provided the best performance for ASR of South African English, and Whisper provided a viable open source alternative. Whisper's performance can likely be attributed to its diverse training dataset, whereas the domain-specific nature of the MMS training dataset limited its performance in the health domain, and with varied speakers. For isiXhosa ASR, GCP and MMS did not provide low enough error rates to be considered reliable. IsiXhosa ASR models also demonstrated high error rates for health terms, particularly for treatment terms (medications). This highlights the existing inequality between high- and low-resource languages, which in the health context may exacerbate the gap between high- and

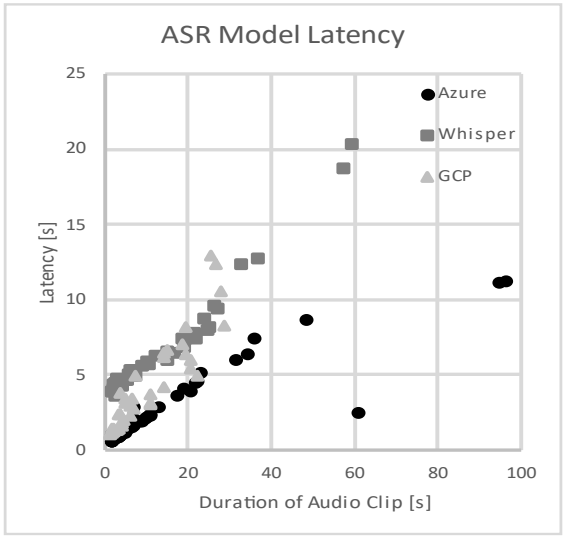

Figure 3: Measured latencies for Azure, GCP, and Whisper for South African English.

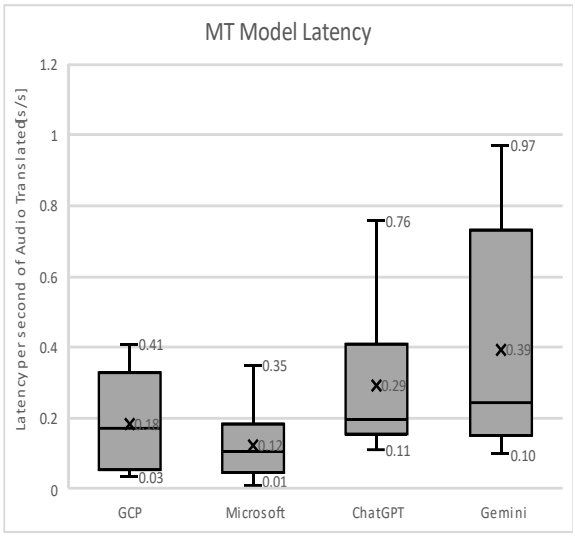

Figure 4: Measured latencies for GCP, Azure, ChatGPT, and Gemini for translating text.

low-resource medical care. If digital health developers must incorporate these models, then they should do so cautiously and with human input to validate results.

For MT, Google Cloud Translate provided the most accurate translations in both directions. However, ChatGPT provided a viable alternative for isiXhosa to English translation. When possible, dictionaries should be incorporated within prompts to further improve performance of LLMs, particularly verified dictionaries of health terms. Translation of health terms had low error rate, particularly for treatment terms as generally these words are kept the same throughout translation. Condition terms such as headache, nausea, and diabetes should be paid specific attention when translated to and from isiXhosa; these may not follow a typical "one-to-one" translation structure and therefore should be approached with caution and verified by humans during medical translation.

There are various advantages and disadvantages when comparing commercial and open-source models for ASR and MT. Open-source models provide a greater level of transparency, which provides greater opportunity for customization and development. Additionally, it allows developers to have more control over the privacy and security of their data. Given that medical transcriptions and translations may hold sensitive information about patients, this is an important factor to consider. However, not every digital health system has the capability to run large ASR or MT models. MMS and NLLB require high levels of computational power to run, which may not be feasible or necessary for small-scale applications. Latency should also be considered, especially in mission critical environments like trauma or emergency medicine. Open source models may experience latency depending on the hardware specifications used to run the models. Commercial options like GCP and Azure are susceptible to service outages an d slower response times depending on the traffic and conditions of their servers. Ultimately, one must consider the context of the digital health solution to select the best models for building a digital health translation system.

Future work may focus on expanding the dataset to incorporate more medical conversation audio and text. This would be beneficial to validate the results achieved here. Additionally, this data could be used to improve and customize models for isiXhosa and for healthcare contexts. Further research might also follow similar methods to the health term analysis described here, to evaluate for age- or gender-related terminology accuracy. Developers may also take this work forward to make evidence-based decisions on ASR and MT models for digital health applications.

Limitations

A limitation of this research is that results were not validated by human evaluators. An evaluation of how the meaning of each result correlates to the meaning of ground truth statements would provide further valuable insights into the accuracy of these models. Additionally, the data published with this work contributes to the resources available for

isiXhosa language applications, but is not enough standalone data to train a domain-specific ASR and MT for health. Finally, because commercial enterprises such as Google and Azure are constantly improving their services, the more recently released models may return different results than those reported on in this paper.

Ethical Considerations

This work provides an overview of the current capabilities of ASR and MT models for isiXhosa. The authors do not provide commentary on whether the results indicate a maturity level that is ready for deployment within the healthcare sector. Rather, we provide benchmarks so developers can make educated decisions regarding ASR and MT model incorporation within digital health systems.

Acknowledgements

This document has been produced with the financial assistance of the European Union (Grant no. DCIPANAF/2020/420-028), through the African Research Initiative for Scientific Excellence (ARISE), pilot programme. ARISE is implemented by the African Academy of Sciences with support from the European Commission and the African Union Commission. The contents of this document are the sole responsibility of the author(s) and can under no circumstances be regarded as reflecting the position of the European Union, the African Academy of Sciences, and the African Union Commission.

References

Tejumade Afonja, Tobi Olatunji, Sewade Ogun, Naome A. Etori, Abraham Owodunni, and Moshood Yekini. 2024. Performant ASR models for medical entities in accented speech. arXiv:2406.12387 [eess.AS]. Version 1.

Matthew Baas and Herman Kamper. 2021. Voice conversion can improve ASR in very low-resource settings. arXiv:2111.02674 [eess.AS]. Version 2.

Abby Blocker, Mohammed I. Datay, Joyce Mwangama, Bessie Malila. 2024. Development of a telemedicine virtual clinic system for remote, rural, and underserved areas using user-centered design methods. *Digital Health*, 10:20552076241256752.

Praveen Borra. 2024. Comparison and analysis of leading cloud service providers (AWS, Azure, and GCP). *International Journal of Advanced Research in Engineering & Technology*, 15(3):266-278.

Liam J. Caffery, Mutaz Farjian, and Anthony C. Smith. 2016. Telehealth interventions for reducing waiting lists and waiting times for specialist outpatient services: a scoping review. *Journal of Telemedicine and Telecare*, 22(8):504-512.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136-158, Prague, Czech Republic.

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, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: scaling human-centered machine translation. arXiv:2207.04672 [cs.CL]. Version 3.

Daniel Erku, Resham Khatri, Aklilu Endalamaw, Eskinder Wolka, Frehiwot Nigatu, Anteneh Zewdie, and Yibeltal Assefa. 2023. Digital health interventions to improve access to and quality of primary health care services: a scoping review. *International Journal of Environmental Research and Public Health,* 20(19):6854.

Antonino Ferraro, Antonio Galli, Valerio La Gatta, and Marco Postiglione. 2023. Benchmarking open source and paid services for speech to text: an analysis of quality and input variety. *Frontiers in Big Data*, 6:1210559.

Foteini Filippidou and Lefteris Moussiades. 2020. A benchmarking of IBM, Google, and Wit automatic speech recognition systems. *Artificial Intelligence Applications and Innovations AIAI 2020, IFIP Advances in Information and Communication Technology*, 583:73-82.

Andrea Gentili, Giovanna Failla, Andriy Melnyk, Valeria Puleo, Gian Luca Di Tanna, Walter Ricciardi, and Fidelia Cascini. 2022. The cost-effectiveness of digital health interventions: A systematic review of the literature. *Frontiers in Public Health*, 20:787135.

Christiaan Jacobs, Annelien Smith, Daleen Klop, Ondřej Klejch, Febe de Wet, and Herman Kamper. 2025. Speech recognition for automatically assessing Afrikaans and isiXhosa preschool oral narratives. arXiv:2501.06478 [eess.AS]. Version 1.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. PriMock57: a dataset of primary care mock consultations. arXiv:2204.00333[cs.CL]. Version 1.

Hauke Licht, Ronja Sczepanski, Moritz Laurer, and Ayjeren Bekmuratovna. 2024. No more cost in translation: validating open-source machine translation for quantitative text analysis. In *ECONtribute Discussion Papers, Reinhard Selten Institute (RSI)*, 276.

Karla C. Maita, Michael J. Maniaci, Clifton R. Haider, Francisco R. Avila, Ricardo A. Torres-Guzman, Sahar Borna, Julianne J. Lunde, Jordan D. Coffey, Bart M. Demaerschalk, and Antonio Jorge Forte. 2024. The impact of digital health solutions on bridging the health care gap in rural areas: a scoping review. *The Permanente Journal*, 28(3):130-143.

Laurette Marais, Johannes A. Louw, Jaco Badenhorst, Karen Calteaux, Ilana Wilken, and Nina van Niekerk. 2020. AwezaMed: A multilingual, multimodal speech-to-speech translation application for maternal health care. In *Proceedings of 2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, Rustenburg, South Africa.

Jessica Ojo, Kelechi Ogueji, Pontus Stenetorp, and David Ifeoluwa Adelani. 2024. How good are large language models for African languages? arXiv:2311.07978 [cs.CL]. Version 2.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392-395, Lisbon, Portugal.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1000+ languages. arXiv:2305.13516 [cs.CL]. Version 1.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. arXiv:2212.04356 [eess.AS]. Version 1.

Gurdeeshpal Randhawa, Mariella Ferreyra, Rukhsana Ahmed, Omar Ezzat and Kevin Pottie. 2013. Using machine translation in clinical practice. *Canadian Family Physician*, 59(4):382-383.

Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain de Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, Francesco Paissan, Davide Borra, Salah Zaiem, Zeyu Zhao, Shucong Zhang, Georgios Karakasidis, Sung-Lin Yeh, Pierre Champion, Aku Rouhe, Rudolf Braun, Florian Mai, Juan Zuluaga-Gomez, Seyed Mahed Mousavi, Andreas Nautsch, Ha Nguyen, Xuechen Liu, Sangeet Sagar, Jarod Duret, Salima Mdhaffar, Gaelle Laperriere, Mickael Rouvier, Renato De Mori, and Yannick Esteve. 2024. Open-source conversational AI with SpeechBrain 1.0. arXiv:2407.00463 [cs.LG]. Version 5.

Thomas Reitmaier, Electra Wallington, Dani Kalarikalayil Raju, Ondrej Klejch, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. 2022. Opportunities and challenges of automatic speech recognition systems for low-resource language speakers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 299:1-17.

Nathaniel R. Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: competitive for high- (but not low-) resource languages. arXiv:2309.07423 [cs.CL]. Version 1.

Stats SA. 2023. *General Household Survey 2023*. Statistics South Africa Department, Republic of South Africa, Private Bag X44, Pretoria, 0001, South Africa.

Stats SA. 2022. *Census 2022*. Statistics South Africa Department, Republic of South Africa, Private Bag X44, Pretoria, 0001, South Africa.

Andreas Stolcke and Jasha Droppo. 2017. Comparing human and machine errors in conversational speech transcription. arXiv:1708.08615 [cs.CL]. Version 1.

Thennal D. K., Jesin James, Deepa P. Gopinath, and Muhammed Ashraf K. (2024). Advocating character error rate for multilingual ASR evaluation. arXiv:2410.07400 [cs.CL]. Version 2.

Lucas Nunes Vieria, Minako O'Hagan, and Carol O'Sullivan. 2020. Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. *Information, Communication, and Society*, 24(11):1515-1532.

Leandro Von Werra, Lewis Tunstall, Abhishek Thakur, Sasha Luccioni, Tristan Thrush, Aleksandra Piktus, Felix Marty, Nazneen Rajani, Victor Mustar, and Helen Ngo. 2022. Evaluate & evaluation on the Hub: better best practices for data and model measurements. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 128-136, Abu Dhabi, UAE.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. arXiv:2304.02210 [cs.CL]. Version 2.