Exploring Zoonotic Disease Knowledge Through AI for Enhanced Risk, Prevention, and Response Awareness in Low-Resource Languages

Armand De Wet[1], Reolin Govender[1], Tedson Nkoana[2], Wanda Markotter[2], Abiodun Modupe[1,3], Vukosi Marivate[1,3,4,5]

[1]Department of Computer Science, University of Pretoria,
[2]Future Africa Institute, University of Pretoria,
[3]Data Science for Social Impact,
[4]AfriDSAI, University of Pretoria, [5]Lelapa AI,

Correspondence: reolin.govender@tuks.co.za

Abstract

Limited linguistic inclusivity in public health communication leaves many South African communities underserved, particularly regarding critical information on zoonotic diseases such as rabies. This pilot study addresses this gap by developing and evaluating AI-driven methods for delivering reliable rabies information to Sepedi speakers, a low-resource language group. The study presents a novel, curated Sepedi dataset of 60 question–answer pairs, created through a systematic pipeline: thematic analysis of authoritative English sources guided the synthetic generation of QA pairs, which were then translated and manually verified by a native-speaking expert. This dataset was used to compare two large language models, GPT-4o and Gemini-1.5 Flash, under both base and fine-tuned conditions. Evaluation used a human-centred rubric assessing fluency, accuracy, and cultural appropriateness. The findings reveal a key nuance in applying LLMs to low-resource domains. The base GPT-4o model, with strong foundational multilingual capabilities, outperformed all other configurations, including its own fine-tuned variant. In contrast, fine-tuning provided a marked improvement for the less capable base Gemini model. This result indicates that fine-tuning can enhance weaker models; its benefits are not universal and may be outweighed by the strong zero-shot performance of state-of-the-art architectures when training data is scarce. The curated Sepedi rabies QA dataset will be released under an open licence to support future work in low-resource public health communication. The data can be found at this GitHub link.

1 Introduction

Linguistic barriers persistently hinder effective public health communication in South Africa. While critical health information is widely available in English, it often fails to reach communities where indigenous languages are predominantly spoken. This gap is perilous in the context of high-stakes diseases like rabies, a preventable yet fatal zoonotic illness that continues to pose a threat in rural provinces (World Health Organization, 2024; National Institute for Communicable Diseases, 2025). The scarcity of accessible, culturally relevant health resources in languages such as Sepedi, spoken by over 4.6 million people, has been linked to reduced disease awareness and delayed treatment, with severe consequences (Makungo et al., 2025; Cotoc and Notaro, 2022).

The emergence of large language models (LLMs) presents a novel opportunity to bridge this linguistic divide. This study investigates that potential by developing and evaluating AI-driven methods for delivering reliable rabies information in Sepedi. At the core of the study is the creation of a novel, thematically structured question–answer (QA) dataset in Sepedi. To build this resource, authoritative English-language public health sources were first analysed to identify key thematic areas. These themes were then used to guide the generation of 60 QA pairs, which were translated into Sepedi and rigorously validated by a native-speaking expert to ensure linguistic and cultural accuracy. This dataset served as the basis for fine-tuning and evaluating two leading LLMs: OpenAI's `GPT-4o` and Google's `Gemini-1.5 Flash`.

This paper makes two primary contributions. First, it introduces what, to the authors' knowledge, is the first curated Sepedi QA dataset for a specific public health domain. The dataset is intended for public release under an open licence to enable external inspection and reuse, and can be found at this GitHub link. Second, it presents a comparative evaluation of base and fine-tuned LLMs for low-resource health communication. The experiments yielded a significant, counterintuitive result: the base `GPT-4o` model, without any fine-tuning, outperformed all other configurations, including

its own fine-tuned variant. This finding challenges the common assumption that fine-tuning is always optimal and highlights the impact of strong foundational multilingual capabilities when training data is scarce. By contrast, fine-tuning provided a marked improvement for the Gemini model, indicating that targeted adaptation remains valuable for models with weaker baseline performance in this setting.

The remainder of this paper is organised as follows. Section 2 surveys related work on low-resource NLP and health communication. Section 3 details our data curation process and the analysis that guided it. Section 4 describes our modelling approach and results. Section 6 discusses the study's limitations and outlines future research directions. Finally, Section 5 summarises our key findings and broader implications.

This paper addressed that gap in two ways. First, it constructed a curated Sepedi question–answer dataset on rabies, derived from authoritative South African public health material and verified by a native Sepedi speaker. Second, it evaluated base and fine-tuned variants of two large language models, GPT-4o and Gemini-1.5 Flash, to test whether they can generate reliable, culturally appropriate health guidance in Sepedi. This is not a general study of "AI in health," but a targeted examination of Sepedi rabies communication using large language models.

2 Literature Review

Developing effective AI tools for public health communication in under-resourced languages, such as Sepedi, requires navigating the intersection of natural language processing (NLP), multilingual AI, and health informatics. This review synthesises existing research in these areas to contextualise the study's challenges and contributions. The discussion focuses on three aspects: the scarcity of linguistic resources, the use of large language models in specialised and low-resource settings, and the methods guiding data curation and evaluation.

This scarcity is even sharper in specialised domains like public health. Institutions such as the National Institute for Communicable Diseases (NICD) produce many English-language resources on topics like rabies. However, similar materials in Sepedi are almost nonexistent. This linguistic gap hinders effective health communication and increases the need for domain-specific datasets that are both medically relevant and culturally sound. This study directly addresses this by creating a curated Sepedi QA dataset for rabies. It provides a foundational resource for developing and evaluating specialised conversational AI.

2.1 The Data Scarcity Challenge in Sepedi

A lack of digital resources has long hampered the development of NLP for African languages. Sepedi, in particular, is consistently classified as a low-resource language (Marivate et al., 2020). Foundational work by the National Centre for Human Language Technology (NCHLT) and projects like Atshumato have highlighted the scarcity of structured corpora and annotated data. While researchers have demonstrated the feasibility of training models for Sepedi, they consistently note severe constraints related to corpus size and quality (Makgatho et al., 2021). The Pedi corpus remains one of the few publicly available benchmarks for the language, underscoring the resource gap (Marivate et al., 2020).

2.2 Large Language Models for Low-Resource Health Communication

Recent advancements in multilingual LLMs offer a promising pathway for bridging information gaps in healthcare. Research has shown that fine-tuning models on specialised, domain-specific data can significantly improve performance. For instance, studies using LLM derivatives in Vietnamese health communication demonstrated that targeted fine-tuning enhanced factual accuracy and user trust (Bui et al., 2025). Similarly, work adapting models for biomedical and ophthalmology question-answering has confirmed that domain-specific fine-tuning improves both fluency and factual correctness (Tan et al., 2024; Christophe et al., 2024).

These findings directly motivated our decision to experiment with fine-tuning for Sepedi, rather than relying solely on zero-shot prompting. Furthermore, this body of work highlights the inadequacy of standard automated metrics (e.g., BLEU, ROUGE) for evaluating performance in specialised, low-resource settings, thus justifying the human-centred evaluation strategy employed in our study.

2.3 Methodological Considerations for Analysis and Evaluation

The dataset was constructed using established exploratory NLP techniques. Topic modelling, specifically Latent Dirichlet Allocation (LDA) and Non-

negative Matrix Factorisation (NMF), is effective for discovering latent themes in public health corpora (Murel and Kavlakoglu, 2024; Murakami et al., 2017). This approach was used to identify the core informational categories in the source documents. Similarly, Named Entity Recognition (NER) is a standard method for extracting structured information, such as organisational and geographic references, from text (Li et al., 2022). These methods provided an empirical foundation for the thematic structure of the generated QA pairs.

For evaluation, our human-centric approach aligns with a growing consensus that qualitative, expert-led assessment is essential in high-stakes domains. Unlike automated metrics, a human-led rubric can assess dimensions such as cultural appropriateness, contextual relevance, and the subtle nuances of factual accuracy, all of which are paramount for building trustworthy health information systems. This qualitative approach ensures that our assessment of model performance is grounded in real-world communicative effectiveness rather than abstract statistical similarity.

3 Data Curation and Exploratory Analysis

A significant challenge in developing NLP tools for Sepedi is the scarcity of structured, domain-specific digital text. To address this in the context of rabies, the study adopted a two-stage approach. First, an exploratory data analysis (EDA) was conducted on a curated English-language corpus to identify key thematic and terminological patterns. Second, these insights were used to guide the creation of a synthetic, high-quality Sepedi question–answer (QA) dataset for fine-tuning.

3.1 Source Corpus and Analysis

Our initial corpus was compiled from authoritative English-language sources, including public awareness materials from the National Institute for Communicable Diseases (NICD), regional surveillance reports, and relevant peer-reviewed studies. This corpus was subjected to a suite of NLP techniques, including word frequency analysis, n-gram extraction, topic modelling, and Named Entity Recognition (NER), to systematically map the rabies information landscape in the South African context. The primary goal was to establish an evidence-based blueprint for our Sepedi dataset.

3.2 Key Findings from the English Corpus

The analysis revealed a strong focus on public health and clinical response. Word frequency analysis (Figure 2) and n-gram extraction (Figure 3) highlighted the centrality of terms such as *rabies*, *animal*, *vaccination*, and phrases like "post exposure prophylaxis" and "rabies south africa" (Table 3).

Table 1: Topic descriptions and key themes using NMF

Topic	Keywords
Dog Health	dog, africa, south, province, species, wildlife, domestic, disease, table, areas
Specimen Collection	case, specimen, biopsy, mortem, csf, nicd, saliva, skin, post
Post-Exposure Prophylaxis	rig, exposure, animal, vaccine, dose, pep, vaccination, day, wound, doses
Transmission Risk	animals, animal, gov, infected, contact, state, humans, prevent, symptoms
Disease Reporting	conditions, notifiable, fever, medical, laboratories, list, public, private

Table 2: Topic descriptions and key themes using LDA

Topic	Keywords
Notifiable Diseases	medical, disease, conditions, notifiable, fever, category, list, laboratories, public, private
NICD & Animal Exposure	nicd, exposure, case, post, animals, csf, saliva, prophylaxis, africa, south
Reporting Animal Contact	animal, contact, gov, state, wildlife, com, must, veterinarian, suspect
Animal Vaccination	animal, exposure, vaccine, vaccination, may, south, africa, bat, dogs
Dog Surveillance	africa, south, dog, species, disease, province, domestic, wildlife, surveillance

Topic modelling, using both Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorisation (NMF), further distilled these patterns into distinct thematic clusters (Tables 2 and 1). These clusters consistently aligned with six core public health domains:

1. Epidemiology and surveillance
2. Prevention and control
3. Virology and diagnosis
4. Clinical presentation

5. Geographical risk areas

6. Public awareness and education

These six domains were not chosen ad hoc. They were derived directly from authoritative South African public health sources, including National Institute for Communicable Diseases (NICD) rabies guidance, provincial surveillance reports, and clinician-facing bite management and post-exposure protocols. Topic modelling, frequency analysis, and named entity recognition over these materials confirmed that these domains captured core informational needs around exposure, vaccination, reporting, symptom escalation, and local geographic risk.

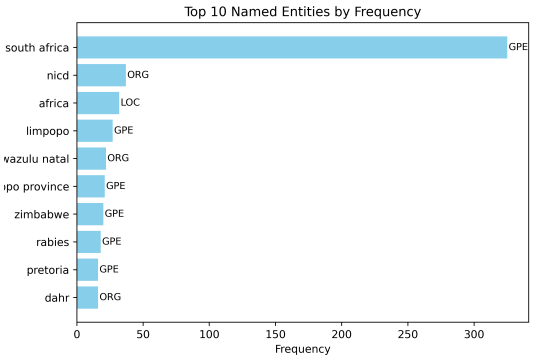

Figure 1: Top named entities identified using NER. Note: Some geographic entities were misclassified due to model limitations (e.g., "KwaZulu-Natal" was classified as ORG rather than GPE).

Additionally, NER confirmed the corpus's regional relevance, identifying key entities such as *South Africa*, *Limpopo*, *NICD*, and specific animal vectors (Figure 1). While the NER tool occasionally misclassified some entities (e.g., treating provinces as organisations), it provided a useful high-level confirmation of the corpus's focus.

Table 3: Top 10 Trigrams with Frequency Counts from the source corpus.

Freq.	Word 1	Word 2	Word 3
51	post	exposure	prophylaxis
42	rabies	south	africa
42	significant	disease	clusters
34	human	rabies	cases
34	trop	med	infect
24	pre	exposure	vaccination
23	world	health	organization
22	bat	eared	foxes
21	black	backed	jackals
21	positive	rabies	cases

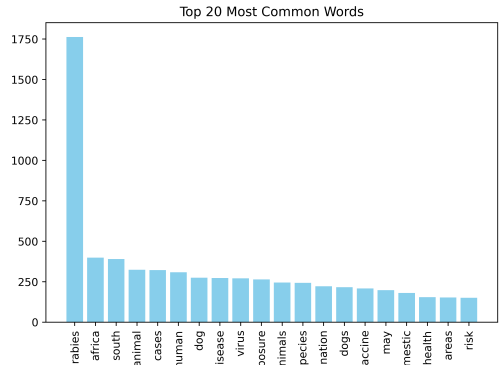

Figure 2: Top 20 most frequently occurring words in the English source corpus.

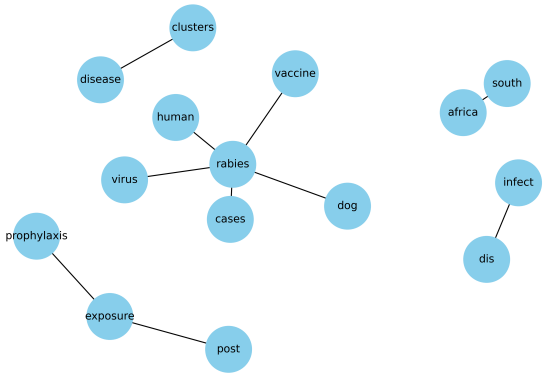

Figure 3: Bigram network visualising core word pairings in the source corpus.

3.3 Guided Creation of the Sepedi Dataset

The six themes identified in the EDA served as a structured framework for generating our fine-tuning dataset. This process involved two key steps:

1. **Synthetic QA Generation:** The initial English question–answer pairs were produced using OpenAI's GPT-4o in an interactive browser session. The model was instructed to act as a public health explainer and to generate clear, non-technical answers about rabies for a general South African audience. The instructions emphasised three constraints: (i) provide medically responsible guidance consistent with recognised South African practice, (ii) describe immediate actions after potential exposure (for example wound cleaning and urgent reporting) without giving alarmist or shaming language, and (iii) avoid inventing unavailable services or phone numbers. The prompting was iterative rather than fully scripted: the model was asked to generate

question–answer pairs for each high-priority theme, the output was inspected, and prompts were refined until all required themes were covered. No custom decoding parameters (for example, temperature or maximum tokens) were manually set in that interface, and exact provider defaults from that session are not recoverable. "Comprehensive coverage", therefore, meant enforced inclusion of each of the six domains identified in Section 3.2, not blind acceptance of the model's first attempt.

The 60 English question–answer pairs were generated so that all six public health domains identified in Section 3.2 (epidemiology and surveillance, prevention and control, virology and diagnosis, clinical presentation, geographical risk areas, and public awareness and education) were represented. At least one QA pair was created for each subtopic within these domains (e.g., vaccination, bite response, and reporting requirements), ensuring no domain was left empty. Every core domain defined from authoritative sources was present in the draft set before translation, and no clinically meaningful theme was omitted.

2. **Translation and Expert Verification:** The English pairs were then translated into Sepedi using GPT-4o and subjected to a manual review by a native Sepedi speaker with practical familiarity in health communication. The reviewer examined each question–answer pair for (i) grammatical correctness and natural phrasing in Sepedi, (ii) preservation of factual meaning from the English source, (iii) cultural and situational relevance (for example, how care-seeking or animal exposure would realistically be described), and (iv) medically safe advice. Corrections were applied at the sentence level, and terminology for key rabies concepts (for example, vaccination, post-exposure treatment, symptom onset, and reporting urgency) was standardised across all pairs for internal consistency. This procedure relied on a single expert reviewer and did not include a second independent rater, so no inter-rater agreement was calculated.

This systematic methodology allowed us to overcome the initial data scarcity by creating a bespoke, thematically robust Sepedi dataset. This dataset, while modest in size, is grounded in the empirical findings of our EDA and explicitly tailored for fine-tuning LLMs for rabies-related health communication, forming the foundation for the experiments detailed in Section 4.

4 Modelling & Experimentation

This section details the experimental methodology, including the selection of large language models, the fine-tuning process, the evaluation strategy, and the results of our comparative analysis.

4.1 Model Selection

Two models were selected for this study: OpenAI's `gpt-4o-2024-08-06` and Google's `gemini-1.5-flash-001`. These models were chosen for both practical and strategic reasons. Both support parameter adaptation via accessible APIs and offer mature Python SDKs, which facilitate integration, experimentation, and inference orchestration. Their underlying transformer architectures and token efficiency also made them suitable candidates for our low-resource context. The primary goal was to compare two leading models under both base and fine-tuning conditions to assess their ability to generate accurate Sepedi-language responses.

4.2 Fine-Tuning Methodology

The fine-tuning stage used the validated Sepedi question–answer dataset described in Section 3.3. Each of the 60 reviewed pairs was converted into the provider-specific JSONL structure required by the OpenAI and Google fine-tuning APIs (OpenAI, 2024; Google DeepMind, 2024). Fine-tuning jobs were then initiated via the respective SDKs, and the adapted GPT-4o and Gemini-1.5 Flash models were exposed through their inference endpoints. This subsection, therefore, focuses on how the curated dataset was consumed by the models and deployed for inference, rather than re-describing the data creation and verification pipeline already presented in Section 3.3.

4.3 Evaluation Strategy

Standard automated NLP metrics, such as BLEU and ROUGE, are known to be unreliable in morphologically rich, low-resource languages like Sepedi, as they do not capture culturally inappropriate phrasing or medically unsafe guidance. For this reason, the evaluation was designed around human judgment rather than relying solely on text-similarity scores.

A held-out evaluation set of questions was created that did not appear in the fine-tuning data but followed the same six core public health themes defined in Section 3 (epidemiology and surveillance, prevention and control, virology and diagnosis, clinical presentation, geographical risk areas, and public awareness and education). Each question in this held-out set was posed to multiple model variants, and all model responses were anonymised and randomised before scoring. This blind setup ensured the reviewer could not see which system generated which answer, reducing bias toward any specific vendor model.

The reviewer was instructed to assess each answer for factual accuracy, clarity, and cultural appropriateness in Sepedi, with specific attention to safety-relevant content, including advice on exposure, reporting, and treatment timing. This procedure treated the task as health communication, rather than generic translation, and aimed to approximate real-world usefulness for a Sepedi-speaking end-user.

A structured set of held-out evaluation questions, distinct from the training set but aligned with the same six public health themes, was developed. Each question was posed to four model configurations:

- Base `GPT-4o` with Context
- Fine-tuned `GPT-4o`
- Base `Gemini-1.5 Flash` with Context
- Fine-tuned `Gemini-1.5 Flash`

The "with Context" configurations refer to a simple prompt engineering strategy in which each query was prepended with a static, manually crafted summary of key rabies information. This approach ensured that the base models had access to relevant domain knowledge without the complexity of a whole Retrieval-Augmented Generation (RAG) pipeline. All model outputs were anonymised and their order randomised to facilitate a blind evaluation by our Sepedi-speaking expert, thereby minimising potential bias.

4.4 Evaluation Rubric

Model responses were scored against a five-dimension rubric, detailed in Table 4. The criteria, fluency, factual accuracy, relevance, clarity, and cultural appropriateness, were rated on a five-point scale from 1 (Poor) to 5 (Excellent). This structured framework enabled a detailed, qualitative analysis of output quality, capturing nuances that automated metrics would miss.

The rubric operationalised "quality" as it matters in practice for health guidance: Can a Sepedi speaker understand the answer? Is the advice factually safe? Is the tone locally appropriate rather than culturally alien or condescending?

Each criterion was scored on a five-point scale. This allowed us to compare model variants on clinically and socially relevant dimensions in high-stakes public health messaging, rather than only on text overlap with a reference.

4.5 Results and Discussion

The human evaluation results, summarised in Figure 4, reveal essential differences in model performance. Contrary to our initial hypothesis, the base `GPT-4o` model achieved the highest scores across all five criteria, marginally outperforming its own fine-tuned variant. This suggests that for a model with exceptionally strong pre-existing multilingual capabilities, fine-tuning on a small dataset provides limited, if any, additional benefit.

In contrast, fine-tuning yielded a marked improvement for the Gemini model. The fine-tuned version demonstrated significantly better factual accuracy and relevance compared to its base counterpart, which scored lowest across all metrics. This highlights the value of domain-specific adaptation for models with weaker baseline performance in a given low-resource language.

These findings challenge the assumption that fine-tuning is universally superior. Instead, they indicate that adaptation effectiveness is highly dependent on both the strength of the base model and the scale of the available training data. It is important to note that a single expert reviewer conducted this evaluation. While this ensured consistency, it also introduces subjectivity; the results should therefore be interpreted as indicative of performance within this pilot study.

5 Conclusion

This pilot study investigated the feasibility of using multilingual LLMs to deliver critical public health information on rabies to Sepedi-speaking communities. By focusing on a low-resource language, this work addressed a significant gap in the accessibility of digital health resources in South Africa. It explored a pathway toward more linguistically

Table 4: Scoring rubric used for human evaluation of model-generated Sepedi responses.

Criterion	1 (Poor)	2 (Fair)	3 (Good)	4 (Very Good)	5 (Excellent)
Fluency / Grammar	Unnatural, many errors	Frequent grammar or phrasing issues	Understandable but some awkward parts	Smooth with minor issues	Native-like, fluent and natural
Factual Accuracy	Incorrect or misleading	Several factual errors	Mostly correct but minor inaccuracies	Almost entirely accurate	Completely accurate
Relevance to Prompt	Off-topic or unrelated	Partially addresses the question	Mostly relevant	Very relevant	Directly and completely answers the prompt
Clarity / Understandability	Confusing or incoherent	Difficult to follow	Understandable with effort	Clear and mostly easy to follow	Very clear and easily understood
Cultural & Contextual Appropriateness	Offensive or inappropriate	Culturally out of touch or awkward	Mostly appropriate	Suitable and respectful	Culturally aligned, respectful, and sensitive

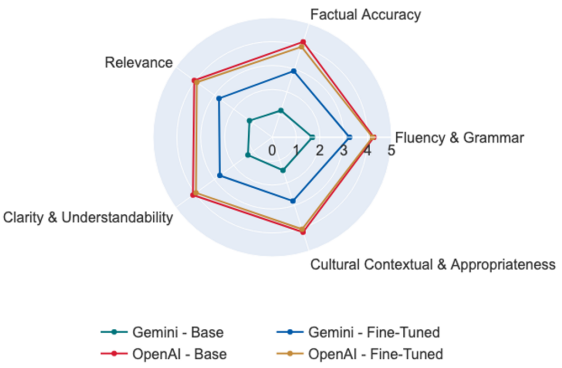

Figure 4: Radar plot comparing human evaluation scores across five criteria. While fine-tuning improved Gemini's performance, the base OpenAI model outperformed all other configurations, including its fine-tuned version. Scores reflect the assessment of a single Sepedi-speaking expert reviewer.

equitable AI.

The methodology followed a multi-stage process, beginning with an exploratory analysis of authoritative materials to identify core public health themes. This analysis informed the creation of a curated dataset of 60 Sepedi question–answer pairs, which was subsequently used to fine-tune and evaluate OpenAI's `GPT-4o` and Google's `Gemini-1.5 Flash`. Using a human-centred rubric that assessed linguistic fluency, cultural appropriateness, and factual accuracy, the study compared the performance of base and fine-tuned model variants.

The evaluation yielded a key, and somewhat counterintuitive finding: the base `GPT-4o` model outperformed all other configurations, including its own fine-tuned version. This result suggests that for models with powerful foundational multilingual capabilities, the benefits of fine-tuning on a small, specialised dataset may be marginal. In contrast, fine-tuning significantly improved the Gemini model's performance, confirming the value of task-specific adaptation for LLMs that are less capable at baseline in low-resource languages.

This asymmetry has direct practical consequences. It suggests that in a low-resource public health setting, a competent multilingual model may already produce usable Sepedi guidance when paired with careful prompt conditioning and contextual priming, even without task-specific fine-tuning. By contrast, a weaker baseline model can still be made viable through fine-tuning on a small but verified dataset. In practice, the question is not simply "should we fine-tune," but "which model type is being deployed." This distinction is crucial for any real-world rollout where both labelled data and expert review capacity are limited.

In summary, this research provides an early but tangible contribution towards equitable AI for health communication. Its primary output is not just a set of evaluated models, but a foundational workflow, from thematic data curation to human-centred evaluation, that can be adapted for other under-resourced languages and health domains. Future efforts should prioritise the development of larger, community-validated datasets and robust, infrastructure-aware deployment strategies to en-

hance the relevance, reliability, and public impact of these technologies.

A key barrier to deployment is the lack of infrastructure. The current approach assumes online access to inference endpoints, which is often not guaranteed in the rural settings where rabies risk is highest. Any real-world rollout will require offline-capable or low-bandwidth channels (e.g., SMS or USSD), along with human oversight for urgent escalations. Without this, technically strong models will not translate into practical access.

The workflow developed here is intentionally modular. Thematic extraction from authoritative sources, controlled generation of QA material, native-speaker correction for linguistic and cultural fit, and human-centred blind evaluation can, in principle, be repeated for other diseases and other South African languages. This is important for scalability, as it suggests that public health agencies and language communities can replicate the process for priorities such as maternal health or tuberculosis awareness without having to rebuild the entire methodology from scratch.

6 Limitations

While this pilot study demonstrates the potential of LLMs for Sepedi health communication, its findings are subject to several limitations that define a clear agenda for future work.

First, the primary constraint is the small, synthetic dataset of 60 question-answer pairs used for fine-tuning. Although curated for thematic relevance and linguistic accuracy, its limited size and synthetic origin may introduce translation artefacts and fail to capture the diversity of authentic, community-specific queries fully. This data scarcity likely explains why fine-tuning yielded only marginal improvements for a robust base model like `GPT-4o`, suggesting that the small-scale adaptation did not significantly enhance its strong foundational capabilities.

Second, the evaluation framework, while qualitatively rich, relied on a single expert reviewer. This approach, though consistent, introduces potential subjectivity and limits the generalisability of our findings. A more robust assessment would require a broader panel of reviewers and incorporate structured metrics for detecting factual inaccuracies or hallucinations, which are critical for safe deployment in a public health context.

Third, the project's scope was intentionally narrow, focusing on a single disease (rabies) and a single language (Sepedi). The direct applicability of these models and findings to other health domains or low-resource languages is not guaranteed without further adaptation and validation. Finally, the web-based prototype assumes consistent internet access, which remains a significant barrier in many of the rural communities this work aims to serve. By focusing on a low-resource language, this work addressed a substantial gap in the accessibility of digital health resources in South Africa. It explored a pathway toward more linguistically equitable AI.

However, practical deployment remains constrained by the availability of infrastructure. The current approach assumes access to internet-connected inference endpoints, which is not guaranteed in many rural communities with the highest risk of rabies exposure. This creates a gap between technical feasibility and actual delivery. Any real-world rollout would require offline-capable or low-bandwidth channels, such as SMS or USSD, or lightweight on-device responders, as well as a mechanism for human oversight in cases where medical escalation is time-sensitive.

The validated Sepedi rabies QA dataset described in this study will be made publicly available under an open licence and can be found at this [GitHub link. This is intended to support reproducibility and independent evaluation.

These limitations highlight several key directions for future research:

- **Dataset expansion through community participation:** Future work should prioritise the creation of larger and more diverse datasets to address the current data bottleneck. This can be achieved through participatory methods, including engagement with community health workers and local speakers to collect authentic questions and co-create culturally resonant answers.

- **Robust, multi-reviewer evaluation:** To improve reliability, future evaluation will adopt a multi-reviewer framework involving a panel of native speakers. Protocols for systematic hallucination detection and factual verification against trusted medical sources will complement this.

- **Accessible deployment models:** Infrastructure constraints are a practical barrier. A key

priority is to explore lightweight models suitable for offline or low-bandwidth deployment. This includes developing interfaces accessible via SMS or USSD to ensure equitable access for communities with limited internet connectivity.

- **Scaling to new domains and languages:** The established workflow was designed for reuse. Future projects will adapt this framework to other critical public health topics (e.g., maternal health, non-communicable diseases) and other under-resourced Southern African languages.

By addressing these areas, this work can be extended toward technically robust, culturally aligned, and practically accessible AI tools for public health communication across the African continent.

Acknowledgements

The authors gratefully acknowledge the support of the ABSA Chair of Data Science and the Data Science for Social Impact (DSFSI) Lab at the University of Pretoria. This work was supported by UK International Development and the International Development Research Centre (IDRC), Ottawa, Canada, under the AI4D Africa Program. DSFSI also acknowledges gifts from NVIDIA, Google.org, OpenAI, and Meta.

We also extend our thanks for the baseline knowledge support provided by the National Institute of Communicable Diseases of the National Health Laboratory Services, South African National Department of Health.

References

Nhat Bui, Giang Nguyen, Nguyen Nguyen, Bao Vo, Luan Vo, Tom Huynh, Arthur Tang, Van Nhiem Tran, Tuyen Huynh, Huy Quang Nguyen, and Minh Dinh. 2025. Fine-tuning large language models for improved health communication in low-resource languages. *Computer Methods and Programs in Biomedicine*, 263:108655.

Clément Christophe, Praveen K. Kanithi, Prateek Munjal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al-Mahrooqi, Avani Gupta, Muhammad Umar Salman, Gurpreet Gosal, Bhargav Kanakiya, Charles Chen, Natalia Vassilieva, Boulbaba Ben Amor, Marco A. F. Pimentel, and Shadab Khan. 2024. Med42: Evaluating fine-tuning strategies for medical LLMs: full-parameter vs. parameter-efficient approaches. *Preprint*, arXiv:2404.14779.

Crina Cotoc and Stephen Notaro. 2022. Race, zoonoses and animal-assisted interventions in paediatric cancer. *International Journal of Environmental Research and Public Health*, 19(13):7772.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Mack Makgatho, Vukosi Marivate, Tshephisho Sefara, and Valencia Wagner. 2021. Training cross-lingual embeddings for setswana and sepedi. *arXiv preprint arXiv:2111.06230*.

Unarine Makungo, Lehlohonolo Kumalo, Jacqueline Weyer, Tumiso Malatji, Lebitsi Ranoto, and Veerle Msimang. 2025. Epidemiological trends of animal bites and human rabies cases in limpopo, south africa, 2011–2023: A retrospective review. *Public Health Bulletin South Africa*, 6(3). Polokwane Mankweng Research Ethics Committee Ref: PMREC 30 October UL 2024/A.

Vukosi Marivate, Tshephisho Sefara, Vongani Chabalala, Keamogetswe Makhaya, Tumisho Mokgonyane, Rethabile Mokoena, and Abiodun Modupe. 2020. Low resource language dataset creation, curation and classification: Setswana and sepedi. *arXiv preprint arXiv:2004.13842*.

Akira Murakami, Paul Thompson, Susan Hunston, and Dominik Vajn. 2017. ‹what is this corpus about?›: using topic modelling to explore a specialised corpus. *Corpora*, 12(2):243–277.

Jacob Murel and Eda Kavlakoglu. 2024. What is topic modeling?

National Institute for Communicable Diseases. 2025. Rabies (disease index).

Ting Fang Tan, Kabilan Elangovan, Liyuan Jin, Jie Yao, Yong Li, Joshua Lim, Stanley Poh, Wei Yan Ng, Daniel Lim, Yuhe Ke, Nan Liu, and Daniel Shu Wei Ting. 2024. Fine-tuning large language model (LLM) artificial intelligence chatbots in ophthalmology and LLM-based evaluation using GPT-4. *Preprint*, arXiv:2402.10083.

World Health Organization. 2024. Rabies.