OpenWHO: A Document-Level Parallel Corpus for Health Translation in Low-Resource Languages

Raphaël Merx[λ] **Hanna Suominen**[ψ,ϕ] **Trevor Cohn**[λ] **Ekaterina Vylomova**[λ]
[λ] School of Computing and Information Systems, The University of Melbourne
[ψ] School of Computing, The Australian National University
[ϕ] School of Medicine and Psychology, The Australian National University

Abstract

In machine translation (MT), health is a high-stakes domain characterised by widespread deployment and domain-specific vocabulary. However, there is a lack of MT evaluation datasets for low-resource languages in this domain. To address this gap, we introduce OpenWHO, a document-level parallel corpus of 2,978 documents and 26,824 sentences from the World Health Organization's e-learning platform. Sourced from expert-authored, professionally translated materials shielded from web-crawling, OpenWHO spans a diverse range of over 20 languages, of which nine are low-resource. Leveraging this new resource, we evaluate modern large language models (LLMs) against traditional MT models. Our findings reveal that LLMs consistently outperform traditional MT models, with Gemini 2.5 Flash achieving a +4.79 ChrF point improvement over NLLB-54B on our low-resource test set. Further, we investigate how LLM context utilisation affects accuracy, finding that the benefits of document-level translation are most pronounced in specialised domains like health. We release the OpenWHO corpus to encourage further research into low-resource MT in the health domain.

1 Introduction

Translation in the health domain combines clinical risks, widespread demand, and domain-specific complexity (Mehandru et al., 2022; Neves et al., 2024). By offering a timely and resource-efficient complement to human translation, machine translation (MT) can lower the barrier to disseminating health content, from education materials for local health workers (Hammond et al., 2024) to public safety information during crises (Federici et al., 2023; Utunen et al., 2023b). However, evaluation of MT in the health domain is hampered by a lack of datasets that cover a wide range of languages, particularly low-resource ones. The TICO-19 corpus (Anastasopoulos et al., 2020) stands as a notable exception, yet its focus on COVID-19 limits its utility on broader health topics, and its age raises the risk of training data contamination.

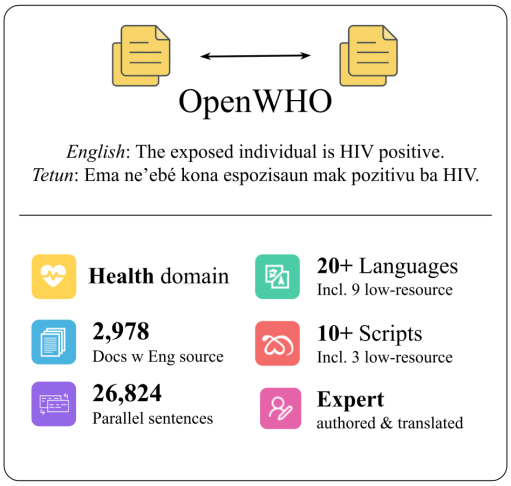

Figure 1: Overview of the OpenWHO parallel dataset, highlighting its depth across low-resource languages and scripts.

To address this gap, we introduce OpenWHO, a document-level parallel corpus designed for evaluating health MT. Sourced from the World Health Organization's multilingual e-learning platform, its content is expert-authored, professionally translated, and shielded from web-crawling, thus minimising contamination risk. The corpus covers over 20 languages, nine of which are low-resource, including some with low-resource scripts like Armenian, Georgian, and Sinhala. By focusing on health education, a domain fundamental to local quality of care (Merx et al., 2024b), OpenWHO provides a realistic benchmark for a high-impact MT use case.

Leveraging this new resource, we conduct a systematic evaluation comparing modern large language models (LLMs) against traditional NMT systems. For LLMs, we study different context strategies (document-level, sentence-level, etc) and to

Proceedings of the Tenth Conference on Machine Translation, pages 142–160
November 8-9, 2025

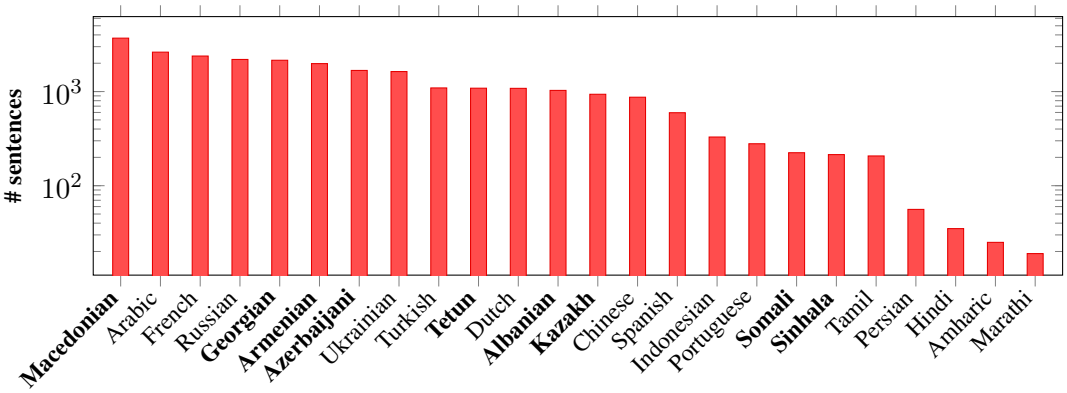

Figure 2: Number of parallel sentences per language in the OpenWHO dataset. The English source has 50,898 sentences. Low-resource languages covered in our experiments (Section 4) are in **bold**.

determine whether the benefits of document-level translation are specific to our dataset, we extend our evaluation to the news and literary subsets of the WMT24++ benchmark (Deutsch et al., 2025).

Our main contributions are [1]:

- **We introduce and release OpenWHO**, a parallel corpus for health MT that covers low-resource languages. It comprises 2,978 documents and 26,824 parallel sentences from expert-authored, professionally translated materials (§3).
- **We show that modern LLMs outperform traditional NMT models** for low-resource translation in the health domain. Our findings show Gemini 2.5 Flash with document-level context achieves a +4.79 ChrF point improvement over NLLB-54B on our test set (§4.2).
- **We find that the benefit of document-level context is model and domain-dependent** for low-resource MT. Accuracy gains are most pronounced when using best-performing models to translate specialised domains like health and literature, while the general (news) domain shows more modest improvements, highlighting that domain complexity drives context utility (§4.2).

2 Related Work

Document-level MT Document-level MT has long been recognised as desirable, as it allows models to leverage broader discourse for improved coherence and accuracy (Maruf et al., 2021). Early work with traditional NMT models has shown mixed results, with some studies demonstrating that document-level context can significantly improve translation quality (Miculicich et al., 2018; Wang et al., 2023; Post and Junczys-Dowmunt, 2024), while others questioned whether these improvements stemmed from true contextual understanding, arguing that the context encoder was not modeling discourse but acting as a "noise generator" that improves model robustness, rather than leveraging discourse information (Li et al., 2020; Appicharla et al., 2024).

LLMs for document-level MT LLMs, given their ability to process extended contexts, are well placed to benefit from document-level context. Koneru et al. (2024) explored contextual translation with Llama2-13B on English-German, finding mixed results, where document-level context sometimes providing no performance gains over sentence-level translation. Karpinska and Iyyer (2023) demonstrated that paragraph-level translation outperforms sentence-level approaches in literary fiction using GPT-3.5, though they have noted that their findings might not generalise to low-resource settings. Yang et al. (2024) fine-tuned LLaMA3-8B for context-aware translation, showing that increasing context window size yields gains, particularly when evaluated with neural metrics. Recent mechanistic analysis by Mohammed and Niculae (2025) revealed that LLMs can be surprisingly "context-insensitive," with smaller models showing limited ability to effectively utilise available context, a finding that may explain varying performance observed in earlier works.

Low-resource MT with LLMs While LLMs have shown promising results on low-resource MT

[1] Code: github.com/raphaelmerx/openwho-code
Dataset: huggingface.co/datasets/raphaelmerx/openwho

Work	Low-resource	LLMs	Document-level	Specialised domains
Ours	✓	✓	✓	✓
Post and Junczys-Dowmunt (2024)	✗	✗	✓	✗
Wang et al. (2023)	✗	✗	✓	✗
Koneru et al. (2024)	✗	✓	✓	✗
Karpinska and Iyyer (2023)	✗	✓	✓	✓
Enis and Hopkins (2024)	✓	✓	✓	✗
Zebaze et al. (2025)	✓	✓	✗	✓
Yang et al. (2024)	✗	✓	✗	✓
Mohammed and Niculae (2025)	✗	✓	✓	✗
Pang et al. (2025)	✗	✓	✓	✓

Table 1: Comparison with prior work on context utilisation for MT.

(Guo et al., 2024; Merx et al., 2024a), evaluation has been predominantly limited to sentence-level translation. Enis and Hopkins (2024) demonstrated that Claude significantly outperforms NLLB on Yoruba-English translation, finding substantial improvements from document-level over sentence-level translation, though their evaluation focused solely on the low-resource-to-English direction. Zebaze et al. (2025) explored low-resource translation with LLMs using compositional approaches on datasets including FLORES, NTREX, and TICO-19, but operated at the sentence level with few-shot learning rather than document-level context. This sentence-level focus in low-resource settings represents a significant gap, as the dynamics of context utilisation may be fundamentally different for low-resource languages where models have seen limited training data and where translation errors could compound across sentences within a document.

Gaps remain in our understanding of document-level translation for low-resource languages in specialised domains. First, there is a shortage of evaluation data for document-level low-resource machine translation in specialised domains, such as healthcare. Second, there has been no systematic analysis of how LLMs utilise document-level context when translating low-resource languages, particularly in specialised domains where coherence and terminological consistency are required.

3 The OpenWHO dataset

3.1 Source and Motivation

The OpenWHO platform. Our corpus is drawn from OpenWHO.org, the World Health Organization's (WHO) former e-learning platform for public health education. Active from 2017 to 2024, the platform's primary goal was to disseminate health knowledge to healthcare professionals, frontline responders, and the public, particularly during health emergencies (George et al., 2022; Utunen et al., 2023a). The content was authored and vetted by WHO experts and its global network of partner institutions, ensuring that the information and its translations were authoritative, accurate, and reflected up-to-date scientific guidance (George et al., 2022). The topics covered a wide range of public health issues, including specific disease responses (e.g., COVID-19, Ebola), vaccination protocols, infection prevention, and emergency preparedness (Utunen et al., 2020, 2023a).

Multilingual focus. A key tenet of the OpenWHO initiative was to ensure equitable access to information, which included a deliberate strategy of multilingual dissemination. Course materials were translated from English into a range of languages, with a focus on providing resources for low- and middle-income countries (George et al., 2022; Utunen et al., 2023a). The course-based format ceased operations in December 2024, transitioning to a static resource library. The data for our corpus was collected prior to this change and exclusively comprises materials from the course-based period (2017–2024). This commitment to creating expert-authored, multilingual content made the OpenWHO platform a high-quality source for extracting a document-level parallel corpus in the health domain, covering several low-resource languages.

Because all course material was hosted behind a login screen, it was shielded from the large-scale web crawling that constitutes the training data for most LLMs, mitigating risk of pre-training contamination. To confirm this, we conducted searches across publicly available web-scraped corpora (C4, MADLAD), and performed targeted web searches (via Google Search) using sentences found in OpenWHO course content. These searches revealed only publicly accessible OpenWHO course descriptions

(which are not part of our corpus), with no course content found within these data sources.

3.2 Data Curation Pipeline

3.2.1 Document Extraction

Scraping While we secured authorization from the WHO to collect and release this data, a direct database export was not available. Therefore, in consultation with the WHO, we developed a web scraping pipeline to gather the course materials. Using the Scrapy framework,[2] we developed a web scraper to navigate the OpenWHO site, enrol in each individual course, and extract the raw HTML content of every course page. Each page was uniquely identified by its course ID and language, as well as its position within the course structure (section and subsection numbers).

Content filtering. A significant portion of the OpenWHO curriculum relies on video-based learning. As our focus is on creating a parallel text corpus, we filtered out pages where video was the primary medium. To further ensure the quality of the extracted documents, we applied a series of heuristic filters to remove low-value content: we discarded pages that primarily consisted of a list of references, contained fewer than ten words, or featured boilerplate text used to introduce a course or section.

3.2.2 Document Pairing

The structured nature of the OpenWHO platform facilitated document alignment. For any given course page, the quadruplet (`language code`, `course id`, `section index`, `subsection index`) serves as a unique identifier. By varying the `language code`, we could accurately identify and group parallel course pages that are direct translations of one another.

After applying the quality filters described in the previous section, this pairing process yielded $2,978$ parallel documents. This set includes significant coverage for several low- and mid-resource languages, with Tetun, Albanian, Macedonian, Azerbaijani, Kazakh, Georgian, and Armenian each having more than 50 parallel documents. A breakdown of document counts per language is presented in Table 5.

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

3.2.3 Sentence Mining

Traditional NMT models rely on sentence-level translation. To release a dataset that can be used for NMT evaluation, and potentially fine-tuning, we mined parallel sentences from parallel documents.

Annotation To evaluate our sentence mining pipeline, we manually annotated 10 parallel documents for 8 target low-resource languages (Macedonian, Georgian, Armenian, Azerbaijani, Tetun, Albanian, Kazakh, Tamil) that are of interest to our experiment. For each document pair, we manually segmented the source and target texts into aligned sentences (relying on back-translation for the languages we are not familiar with), ensuring one-to-one correspondence. This process yielded a reference corpus totalling $2,645$ parallel sentences. This annotated set serves as the ground truth for the subsequent sentence-splitting and alignment evaluation.

Sentence splitting Using the target-language sentences from our manually annotated corpus as a reference, we evaluated the performance of three sentence-splitters: `NLTK` (Bird and Loper, 2004), `Stanza` (Qi et al., 2020), and `pysbd` (Sadvilkar and Neumann, 2020). We measured sentence splitting performance as accuracy of sentence boundaries against our ground truth segmentation. The results (shown in Appendix B) indicated that `pysbd` achieves the highest accuracy overall with accuracy ranging from 82.0% for Kazakh to 94.0% for Tetun, but `stanza` performs better for Kazakh (89.1%), Tetun (94.6%) and Georgian (93.2%). NLTK's performance was generally lower than the other two. Based on these findings, we selected the best-performing tool (either `pysbd` or `stanza`) on a per-language basis to segment the entire corpus.

Sentence alignment After sentence splitting, we aligned sentences to create parallel pairs. Here we rely on sentence semantic similarity, using LaBSE (*Language-agnostic BERT Sentence Embedding* (Feng et al., 2022)), which supports all our languages of interest except Tetun (as a consequence, for Tetun, we first translated target sentences back to English before encoding). Because the OpenWHO documents are relatively short, this approach is highly effective: when evaluated against our manually annotated ground truth, the method yielded F1 scores ranging from 98.6% (for Tetun) to 100% (for Kazakh and Georgian).

Strategy	Description
Sentence level	Our baseline. Each sentence is translated independently without any additional context.
Sentence window (batched sliding window in Koneru et al. (2024))	A constrained-context approach. The model receives only the immediately preceding and succeeding sentences as context, aiming to capture local discourse phenomena without overwhelming the model.
Sentence + doc context	The model is provided with the full source document as context within the prompt but is instructed to translate only the single, target sentence.
Document level	The model is given the entire source document and instructed to translate the whole text. As per Enis and Hopkins (2024), we use one sentence per line, and evaluate at the sentence level after translation.
Doc-level + self-correct, as per (Wu et al., 2025)	A two-step approach: (1) Document-level translation then (2) feed the generated translation back to the model with a new prompt asking it to review and improve its own output, testing its self-correction and refinement capabilities.

Table 2: The five translation strategies evaluated in our experiments. Each strategy represents a different approach to leveraging context for machine translation. Associated prompts are in Appendix F.

Quality Control and Filtering Finally, to ensure the quality of the mined sentence pairs, we implemented an additional filtering stage based on empirical rules. We removed sentence pairs that were likely to be misaligned or uninformative for translation tasks. This included removing (1) pairs where the source English sentence contained fewer than five words, as these are often section headers or fragments; (2) pairs where the target-language side was in English; and (3) sentence pairs that were exact duplicates across different course pages, which often correspond to repeated instructions or boilerplate phrases.

Starting from an initial pool of $43,732$ candidate sentence pairs, we arrived at a final, clean set of **26,824** parallel sentences. This includes nine low-resource languages with over 200 parallel sentences (Macedonian, Georgian, Armenian, Azerbaijani, Tetun, Albanian, Kazakh, Somali, Sinhala). The count of sentence pairs per language is detailed in Table 5.

3.3 Dataset Statistics

The resulting OpenWHO corpus comprises **2,978** parallel documents and **26,824** aligned parallel sentences between English and over 20 other languages. The corpus contains a mix of high-resource and low-resource languages, with significant depth in the latter, including six with over 1,000 parallel sentences (Macedonian, Georgian, Armenian, Albanian, Kazakh, and Tetun). A key feature of this dataset is its origin: all content is expert-authored and professionally translated, providing high-fidelity, domain-specific text that is a level above standard web-crawled corpora in terms of quality and consistency. The data is structured at both the document and sentence level, enabling experiments in document-level machine translation, terminology extraction, and domain adaptation. However, a potential weakness of this dataset is its unbalanced language distribution, as not all courses were translated into all languages.

3.4 Data Availability

With permission from the WHO, we release this dataset under a Creative Commons NonCommercial license (CC BY-NC 4.0), allowing re-use, modification and distribution for non-commercial use, while requiring attribution. Data will be available both at the document level and at the sentence level.

4 Experiments

Having established an evaluation corpus for document-level low-resource MT in the health domain, we now turn to investigating what models perform best on this dataset, and how context utilisation strategies affect LLM performance on this dataset. Our experimental design addresses a fundamental tension in document-level translation: while broader context can improve coherence and terminological consistency, it may also introduce noise or lead to error propagation.

We work with the following research questions:

- **RQ1**: How do state-of-the-art LLMs compare to traditional NMT models for health low-resource translation?
- **RQ2**: What is the most effective context strategy for LLM-based translation into

low-resource languages? (sentence-level, document-level, sliding sentence window, etc)

- **RQ3**: How does model capability interact with these context strategies?

4.1 Experimental Setup

Datasets We evaluate on two datasets, always in the EN-XX direction. The first is our newly introduced OpenWHO corpus. To ensure a controlled comparison across languages, we focus our experiments on a single, extensively translated course: "Infection Prevention and Control through Hand Hygiene (IPC-HH)". We select the nine low- to mid-resource languages available for this course for our evaluation: Albanian (sqi), Armenian (hye), Azerbaijani (aze), Georgian (kat), Kazakh (kaz), Macedonian (mkd), Sinhala (sin), Somali (som), and Tetun (tet). For a comparison with high-resource languages, we separately evaluate on French (fra), Russian (rus) and Spanish (spa).

To test the generalisability of our findings beyond the health domain, we also evaluate on the WMT24++ benchmark (Deutsch et al., 2025), an expansion of the WMT24 dataset to 55 languages. To align with our research focus, we select a sample of five low- to mid-resource languages present in this dataset: Bulgarian (bul), Serbian (srp), Swahili (swh), Tamil (tam), and Zulu (zul). Because this dataset is available at the paragraph level, for our sentence-level analysis, we split paragraphs into aligned sentences using Gemini 2.5 Flash.

Models Our model selection is designed to compare modern LLMs (both open and closed weights) against conventional NMT baselines. For NMT baselines, we select NLLB-200 (3.3B & 54B, Costa-jussà et al., 2024) and MADLAD-400 10B (Kudugunta et al., 2023), both of which cover languages covered in our evaluation (except Tetun for NLLB). For LLMs, we select Gemini 2.5 Flash (Gemini Team, 2025), a powerful closed-weight model, DeepSeek-V3 671B (DeepSeek-AI, 2024), which represents the state-of-the-art in open-weight models at the time of our experiments, and Gemma 3 27B (Team, 2025), a smaller LLM with broad multilingual support. We run all model calls through OpenRouter.[3]

Metrics We primarily evaluate with ChrF++ (Popović, 2017), an n-gram based metric which has been shown to correlate better with human judgement than BLEU (Papineni et al., 2002) particularly for morphologically rich languages like Kazakh or Georgian. To validate results found with ChrF++, we also evaluate with MetricX-24[4] (Juraska et al., 2024) and AutoMQM (Fernandes et al., 2023). MetricX is a neural metric which correlates better with human judgement than ChrF++ for high-resource languages. While it has not been evaluated on low-resource languages, it is based on mT5 (Xue et al., 2021), which has been pretrained on all languages in our study, aside from Tetun. AutoMQM uses a large language model to characterise translation errors using MQM (Lommel et al., 2013). To avoid self-preference bias that may arise from using the same LLM for AutoMQM as that used for translation (Wataoka et al., 2025), we run AutoMQM with Kimi K2 (Kimi-AI, 2025).

	ChrF ↑	MetricX ↓
OpenWHO (9 low-res langs)		
NLLB 54B	50.52	3.45
Gemini	55.32 ↑ **4.79**	3.10 ↓ **-0.43**
DeepSeek-v3	49.38 ↓ **-1.14**	3.92 ↑ **0.39**
Gemma 3	48.01 ↓ **-2.51**	4.24 ↑ **0.71**
WMT24++ literary (5 low-res langs)		
NLLB 54B	43.00	5.83
Gemini	50.66 ↑ **7.66**	3.76 ↓ **-2.07**
DeepSeek-v3	46.88 ↑ **3.88**	4.57 ↓ **-1.26**
Gemma 3	44.45 ↑ **1.45**	5.26 ↓ **-0.57**
WMT24++ news (5 low-res langs)		
NLLB 54B	53.58	3.45
Gemini	54.83 ↑ **1.24**	2.69 ↓ **-0.76**
DeepSeek-v3	51.40 ↓ **-2.18**	3.42 ↓ **-0.04**
Gemma 3	50.71 ↓ **-2.87**	3.61 ↑ **0.16**

Table 3: Average performance per model, with score difference from NLLB 54B. Modern LLMs like Gemini outperform NLLB on specialised domain low-resource MT, like health or literary fiction. See Tables 7 and 10 for scores per language, which vary from 37 to 63 ChrF.

Translation Strategies For LLM translation, we rely on a fixed one-shot prompt (Appendix F), and we systematically evaluate five translation strategies that introduce contextual information in different ways. Detailed in Table 2, these include translating sentences one at a time, translating sentences with some surrounding context, and translating entire documents at once. For NMT models (NLLB

[3] https://openrouter.ai/

[4] google/metricx-24-hybrid-large-v2p6-bfloat16

and MADLAD), as they were trained at the sentence level, we evaluate only at the sentence level. To ensure a fair comparison across models and strategies, all outputs, including those generated at the document level, are segmented and evaluated at the sentence level against the reference translations.

4.2 Results

LLMs outperform NMT on health low-resource translation (RQ1). On OpenWHO, Gemini 2.5 Flash, when translating at the document level, outperforms NLLB 54B across all languages,[5] by an average of +4.79 ChrF points (Table 3). MetricX and AutoMQM results confirm this overall trend. However, other LLMs evaluated (DeepSeek-v3 and Gemma 3) are still outperformed by NLLB-54B, albeit by a small margin for DeepSeek. This means that among open weight models, NLLB-54B is still the preferred choice. Further, at equivalent performance before fine-tuning, LLMs require far more computation, with around one order of magnitude more parameters for the same performance (e.g. DeepSeek-v3 671B roughly equivalent to NLLB 54B; Gemma 3 27B equivalent to NLLB 3.3B).

Error analysis: Gemini vs NLLB Error analysis using AutoMQM (Table 12) shows that Gemini translations contain substantially fewer critical errors than NLLB, with less mistranslations (where target text does not accurately represent the source meaning) and less incorrect terminology, at the cost however of more omissions (where target text is missing information present in the source) and overtranslations (target text more specific than the source).

On high-resource languages, NLLB and LLMs are very close to each other. On our sample of high-resource OpenWHO languages (French, Russian, Spanish), the average scores for NLLB 54B, Gemini, DeepSeek, and Gemma 3 are all remarkably close to each other, as measured by both ChrF (averages in the 59-62 range for all 4 models) and MetricX (averages in the 2.3-2.4 range). This result indicates that in the health domain, the advantage of LLMs over NMT is more pronounced on low-resource languages compared to high-resource. Unsurprisingly, performance on high-resource languages is notably higher than on low-resource ones, with a gap of 7-12 ChrF points between high-resource and low-resource across all models.

LLMs tend to work best at the document level, for specialised domains (RQ2). On OpenWHO, both Gemini and DeepSeek translate best at the document level, with +3.62 and +2.00 ChrF points over sentence-level translation respectively (Table 4), but no measurable improvements in MetricX scores. On WMT24++ literary, the advantage of document-level over sentence-level is even clearer, with +6.37 Chrf points for Gemini, +3.34 for DeepSeek, and similar improvements in MetricX scores (Table 11). For Gemma 3 27B however, additional context from document-level only marginally improves translation accuracy, on both OpenWHO and WMT24++ literary. Overall, we observe a trend where **the larger the LLM, the more it benefits from document-level translation (RQ3)** over sentence-level translation.

In the general domain, the advantage of modern LLMs and document-level translation are less clear. On the WMT24++ news set, we do not see meaningful accuracy improvements for document-level over sentence-level translation, using either metric (ChrF and MetricX). We also see less variation in scores between models on this domain, both when comparing NLLB to LLMs, and when comparing LLMs with each other. Overall, the advantage of document-level translation over sentence-level translation for low-resource MT is not uniform across domains and models.

5 Discussion

Our experiments present **three key findings**: First, modern LLMs tend to outperform NMT (e.g. Gemini outperforms NLLB 54B) on low-resource translation in specialised domains (health with OpenWHO, literary text with WMT24++). Second, modern LLMs translate best at the document-level in specialised domains (health and literary), but the advantage of document-level translation is less clear for smaller models and for the general domain. Third, other context-utilisation strategies (e.g. sentence window, document context with one sentence at a time) tend to perform less well than whole-document translation.

Why Gemini outperforms NLLB in low-resource specialised domain MT (RQ1) Our investigation into performance differences between

[5] We exclude Tetun from this comparison, as NLLB does not support it.

Doc vs sent	ChrF Δ	MetricX Δ
OpenWHO (9 low-res langs)		
Gemini	↑ 3.62	→ 0.00
DeepSeek	↑ 2.00	→ 0.02
Gemma3	↓ -0.21	↑ 0.24
WMT literary (5 low-res langs)		
Gemini	↑ 6.37	↓ -1.18
DeepSeek	↑ 3.34	↓ -0.79
Gemma3	↑ 2.06	↓ -0.14
WMT news (5 low-res langs)		
Gemini	↑ 1.24	↓ -0.08
DeepSeek	↓ -0.82	↓ -0.11
Gemma3	↓ -0.14	↑ 0.13

Table 4: Performance difference for document-level vs sentence-level translation, averaged across languages. In specialised domains (health, literary fiction), the larger the LLM, the more it benefits from doc-level translation. See Tables 8 and 11 for scores per language.

LLMs and NMT models reveals that Gemini's advantage over NLLB 54B in specialised domains stems directly from its ability to leverage document-level context. When both models are constrained to sentence-level translation, their performance is very similar across all three datasets evaluated (OpenWHO, WMT24++ literary, and WMT24++ news, all within a narrow 1.5 ChrF point margin). It is only when Gemini is provided with the full document that it establishes a clear performance lead.

The role of context strategy across domains Our findings show that the optimal context strategy depends on both text domain (RQ2) and model capability (RQ3). The benefit of document-level translation is most pronounced in specialised domains like health and literature, potentially because their discourse structure requires a high degree of linguistic coherence for both accuracy (e.g. correct health terminology) and stylistic integrity (e.g. sustained narrative tone). In contrast, the news domain may rely more on self-contained sentences that allow skimming and quoting, reducing the benefit of context. Further, our results indicate that smaller models only gain marginal benefits from document context, potentially lacking the capacity to maintain coherence without introducing noise.

Our findings, particularly the dependence of context utility on model capability and domain specificity, offer a nuanced picture for where document-level context is most useful, which may explain past work that either did not (Li et al., 2020; Appicharla et al., 2024; Koneru et al., 2024) or did (Wang et al., 2023; Post and Junczys-Dowmunt, 2024; Wu et al., 2024) find added benefits from contextual level translation.

Recommendations Based on our results, we offer three recommendations for researchers working on low-resource MT:

1. **Evaluate LLMs at the document level for specialised domains.** Sentence-level evaluation can mask the advantage of modern LLMs, which lies in their ability to use context.

2. **Utilise the most capable LLMs to maximise the benefit of document context.** The performance gains from document-level translation are most significant with the largest models.

3. **Analyse performance on a per-language basis.** Average model rankings do not always reflect performance on individual languages, making granular analysis essential for model selection.

Future directions Several avenues for future work emerge from our findings. First, the development of reliable evaluation metrics tailored to low-resource MT in the health domain. Second, further exploration of strategies to optimise LLM-based translation for low-resource health contexts, such as fine-tuning on domain-specific data or different prompting techniques. Third, the creation of evaluation benchmarks for low-resource health on other tasks, such as question answering, which OpenWHO could be leveraged for.

6 Conclusion

In this work, we introduced OpenWHO, a high-quality parallel corpus for health MT, with a focus on low-resource languages. Sourced from the World Health Organization's expert-authored materials, it addresses a gap in evaluation resources and provides a benchmark for future research at the intersection of health and low-resource languages. The dataset strengths include (1) the grounding of its source English text in evidence-based WHO guidance (2) its professional translation into various languages and (3) its availability at both the document and sentence level. However, OpenWHO is language imbalanced (not all courses were translated into all languages), which can limit its comparative value.

Our experiments demonstrate that modern LLMs, when provided with full document-level context, outperform traditional NMT models on low-resource translation in specialised domains like health and literature. We found that this advantage is most pronounced for the largest models and diminishes in the general (news) domain, highlighting that the utility of context depends on both model capability and domain complexity. Our work underscores the potential of document-aware LLMs to improve translation quality in high-impact settings, while also revealing the critical need for domain-specific evaluation benchmarks and context-aware translation strategies.

7 Limitations

Metrics Our findings rely exclusively on automated metrics (ChrF, MetricX, AutoMQM). While these metrics give a useful signal when they all agree, we have limited ability to resolve differences when they arise. ChrF is a recognised standard for low-resource MT but may not always correlate well with human judgement (Wang et al., 2024); MetricX and AutoMQM have not been evaluated on low-resource languages, let alone in the health domain. Overall, more work is needed to determine what is the right metric for low-resource health MT, including a comprehensive human evaluation to validate our findings and gain a more nuanced understanding of translation quality.

Generalisability across other domains In our experiments on context utilisation, we rely on two specialised domains: health (OpenWHO) and literary fiction (WMT24++). While we find similar trends, our findings may not generalise to other specialised domains, such as legal, financial, or technical texts. The specific characteristics of each domain may influence the utility of document-level context, and a broader, structured evaluation across multiple domains would be needed to draw more general conclusions.

Caveats of a direct comparison between LLMs and NMT While document-level LLM translation beats sentence-level NMT translation for the languages and specialised domains we evaluate on, this comparison might be unfair to NMT models, which could be adapted to benefit from document-level context for a more equivalent comparison, and have far fewer model parameters at equivalent performance levels. In practice, LLM outputs could be leveraged for knowledge distillation, creating smaller, domain-specific models that retain much of the performance advantage while being more efficient (Gibert et al., 2025).

Dataset language imbalance Finally, the OpenWHO dataset itself has limitations. Its language distribution is imbalanced, as not all source materials were translated into every target language. This can constrain its utility for direct cross-language comparisons.

8 Ethics Statement

Consent This work adheres to ethical guidelines for data collection and research in natural language processing. The OpenWHO corpus was compiled from the WHO's e-learning platform with explicit authorization from the WHO for both data collection and public release. Our work aligns with the WHO's mission to disseminate health information globally and respects their ownership of the content.

Dual use and societal impact We have carefully considered the potential for dual use of the OpenWHO dataset and our research findings. Our primary objective is to enhance access to health education material by improving MT for low-resource languages in the high-stakes health domain. The dataset comprises expert-authored, professionally translated public health materials, limiting risks of misuse. The humanitarian and public health benefits of facilitating information access in underserved languages significantly outweigh dual-use concerns.

Acknowledgements

We are deeply grateful to the World Health Organization (WHO) for their collaboration and for granting us permission to collect and publicly release the OpenWHO dataset. In particular, we would like to express our sincere gratitude to Heini Utunen, Corentin Piroux, and Melissa Attias for their support and guidance on this project.

This research was supported by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative.

References

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann,

Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Ramakrishna Appicharla, Baban Gain, Santanu Pal, Asif Ekbal, and Pushpak Bhattacharyya. 2024. A case study on context-aware neural machine translation with multi-task learning. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 246–257, Sheffield, UK. European Association for Machine Translation (EAMT).

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

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, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846. Publisher: Nature Publishing Group.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the Language Coverage of WMT24 to 55 Languages & Dialects. *arXiv preprint*. ArXiv:2502.12404 [cs].

Maxim Enis and Mark Hopkins. 2024. From LLM to NMT: Advancing Low-Resource Machine Translation with Claude. *arXiv preprint*. ArXiv:2404.13813 [cs].

Federico M. Federici, Christophe Declercq, Jorge Díaz Cintas, and Rocío Baños Piñero. 2023. Ethics, Automated Processes, Machine Translation, and Crises. In Helena Moniz and Carla Parra Escartín, editors, *Towards Responsible Machine Translation: Ethical and Legal Considerations in Machine Translation*, pages 135–156. Springer International Publishing, Cham.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The Devil Is in the Errors: Leveraging Large Language Models for Fine-grained Machine Translation Evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics.

Google Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Richelle George, Heini Utunen, Ngouille Ndiaye, Anna Tokar, Lama Mattar, Corentin Piroux, and Gaya Gamhewage. 2022. Ensuring equity in access to online courses: Perspectives from the WHO health emergency learning response. *World Medical & Health Policy*, 14(2):413–427. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wmh3.492.

Ona de Gibert, Joseph Attieh, Teemu Vahtola, Mikko Aulamo, Zihao Li, Raúl Vázquez, Tiancheng Hu, and Jörg Tiedemann. 2025. Scaling Low-Resource MT via Synthetic Data Generation with LLMs. *arXiv preprint*. ArXiv:2505.14423 [cs].

Ping Guo, Yubing Ren, Yue Hu, Yunpeng Li, Jiarui Zhang, Xingsheng Zhang, and Heyan Huang. 2024. Teaching large language models to translate on low-resource languages with textbook prompting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15685–15697, Torino, Italia. ELRA and ICCL.

Robert Hammond, Antonito Hornay Cabral, Jeremy Beckett, Xhian Meng Quah, Natarajan Rajaraman, Sanjay Mathew, Amrutha Gopalakrishnan, Mariano Pereira, Manuel Natercio Noronha, Bernardo Pinto, João de Jesus Arcanjo, Celia Gusmao dos Santos, Telma Joana Corte-Real de Oliveira, Ingrid Bucens, and Charlotte Hall. 2024. Lessons Learnt Delivering a Novel Infectious Diseases National Training Programme to Timor-Leste’s Primary Care Workforce. *Annals of Global Health*, 90(1).

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google Submission to the WMT 2024 Metrics Shared Task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Marzena Karpinska and Mohit Iyyer. 2023. Large Language Models Effectively Leverage Document-level Context for Literary Translation, but Critical Errors Persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.

Kimi-AI. 2025. Kimi k2: Open agentic intelligence. *Preprint*, arXiv:2507.20534.

Sai Koneru, Miriam Exel, Matthias Huck, and Jan Niehues. 2024. Contextual Refinement of Translations: Large Language Models for Sentence and Document-Level Post-Editing. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2711–2725, Mexico City, Mexico. Association for Computational Linguistics.

Sneha Kudugunta, Isaac 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. *Advances in Neural Information Processing Systems*, 36:67284–67296.

Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 2020. Does multi-encoder help? a case study on context-aware neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3512–3518, Online. Association for Computational Linguistics.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A Survey on Document-level Neural Machine Translation: Methods and Evaluation. *ACM Comput. Surv.*, 54(2):45:1–45:36.

Nikita Mehandru, Samantha Robertson, and Niloufar Salehi. 2022. Reliable and Safe Use of Machine Translation in Medical Settings. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 2016–2025, New York, NY, USA. Association for Computing Machinery.

Raphaël Merx, Aso Mahmudi, Katrina Langford, Leo Alberto de Araujo, and Ekaterina Vylomova. 2024a. Low-resource machine translation through retrieval-augmented LLM prompting: A study on the Mambai language. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024*, pages 1–11, Torino, Italia. ELRA and ICCL.

Raphaël Merx, Christine Phillips, and Hanna Suominen. 2024b. Machine Translation Technology in Health: A Scoping Review. In *Health. Innovation. Community: It Starts With Us*, pages 78–83. IOS Press.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Wafaa Mohammed and Vlad Niculae. 2025. Context-Aware or Context-Insensitive? Assessing LLMs' Performance in Document-Level Translation. *arXiv preprint*. ArXiv:2410.14391 [cs].

Mariana Neves, Cristian Grozea, Philippe Thomas, Roland Roller, Rachel Bawden, Aurélie Névéol, Steffen Castle, Vanessa Bonato, Giorgio Maria Di Nunzio, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, and Antonio Jimeno Yepes. 2024. Findings of the WMT 2024 biomedical translation shared task: Test sets on abstract level. In *Proceedings of the Ninth Conference on Machine Translation*, pages 124–138, Miami, Florida, USA. Association for Computational Linguistics.

Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. 2025. Salute the Classic: Revisiting Challenges of Machine Translation in the Age of Large Language Models. *Transactions of the Association for Computational Linguistics*, 13:73–95.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post and Marcin Junczys-Dowmunt. 2024. Escaping the sentence-level paradigm in machine translation. *arXiv preprint*. ArXiv:2304.12959 [cs].

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Nipun Sadvilkar and Mark Neumann. 2020. PySBD: Pragmatic sentence boundary disambiguation. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 110–114, Online. Association for Computational Linguistics.

Gemma Team. 2025. Gemma 3.

Heini Utunen, Ranil Appuhamy, Melissa Attias, Ngouille Ndiaye, Richelle George, Elham Arabi, and Anna Tokar. 2023a. Observations from three years of online pandemic learning response on OpenWHO.

The International Journal of Information and Learning Technology, 40(5):527–540. Publisher: Emerald Publishing Limited.

Heini Utunen, Ngouille Ndiaye, Corentin Piroux, Richelle George, Melissa Attias, and Gaya Gamhewage. 2020. Global Reach of an Online COVID-19 Course in Multiple Languages on OpenWHO in the First Quarter of 2020: Analysis of Platform Use Data. *Journal of Medical Internet Research*, 22(4):e19076. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

Heini Utunen, Thomas Staubitz, Richelle George, Yu Ursula Zhao, Sebastian Serth, and Anna Tokar. 2023b. Scale Up Multilingualism in Health Emergency Learning: Developing an Automated Transcription and Translation Tool. *Studies in Health Technology and Informatics*, 302:408–412.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, and 39 others. 2024. AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced African languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5997–6023, Mexico City, Mexico. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-Level Machine Translation with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2025. Self-Preference Bias in LLM-as-a-Judge. *arXiv preprint*. ArXiv:2410.21819 [cs].

Di Wu, Seth Aycock, and Christof Monz. 2025. Please Translate Again: Two Simple Experiments on Whether Human-Like Reasoning Helps Translation. *arXiv preprint*. ArXiv:2506.04521 [cs].

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting Large Language Models for Document-Level Machine Translation.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Xinye Yang, Yida Mu, Kalina Bontcheva, and Xingyi Song. 2024. Optimising LLM-Driven Machine Translation with Context-Aware Sliding Windows. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1004–1010, Miami, Florida, USA. Association for Computational Linguistics.

Armel Zebaze, Benoît Sagot, and Rachel Bawden. 2025. Compositional Translation: A Novel LLM-based Approach for Low-resource Machine Translation. *arXiv preprint*. ArXiv:2503.04554 [cs].

A Documents and sentences per language

Language	Script	Number of documents	Number of sentences
Russian (rus)	Cyrillic	315	2194
French (fra)	Latin	301	2385
Arabic (ara)	Arabic	293	2623
Macedonian (mkd)	Cyrillic	254	3695
Ukrainian (ukr)	Cyrillic	204	1632
Chinese (zho)	Chinese	203	871
Spanish (spa)	Latin	149	596
Georgian (kat)	Georgian	131	2151
Armenian (hye)	Armenian	125	1982
Kazakh (kaz)	Cyrillic	103	936
Azerbaijani (aze)	Latin	98	1677
Turkish (tur)	Latin	81	1093
Indonesian (ind)	Latin	80	329
Dutch (nld)	Latin	77	1082
Albanian (sqi)	Latin	74	1029
Tetun (tet)	Latin	67	1086
Portuguese (por)	Latin	62	279
Hindi (hin)	Devanagari	28	35
Tamil (tam)	Tamil	26	207
Sinhala (sin)	Sinhala	25	214
Persian (fas)	Perso-Arabic	25	56
Amharic (amh)	Ethiopic	22	25
Marathi (mar)	Devanagari	21	19
Somali (som)	Latin	20	224
Italian (ita)	Latin	20	60
Lao (lao)	Lao	18	29
Yoruba (yor)	Latin	17	14
Burmese (mya)	Burmese	15	32
Swahili (swa)	Latin	14	107
Vietnamese (vie)	Latin	11	13
Catalan (cat)	Latin	10	8
Pushto (pus)	Perso-Arabic	8	22
Hausa (hau)	Latin	8	28
Thai (tha)	Thai	7	8
Shan (shn)	Shan	7	3
S'gaw Karen (ksw)	Karen	7	2
Japanese (jpn)	Japanese	6	9
Bulgarian (bul)	Cyrillic	6	9
Bengali (ben)	Bengali	6	8
Urdu (urd)	Perso-Arabic	4	10
Telugu (tel)	Telugu	4	4
Greek (ell)	Greek	4	4
Serbian (srp)	Latin	3	4
Polish (pol)	Latin	3	5
Panjabi (pan)	Gurmukhi	3	4
Oriya (ori)	Odia	3	4
Kurdish (kur)	Latin/Arabic	3	3
Tajik (tgk)	Cyrillic	2	1
Romanian (ron)	Latin	2	6
Nigerian Pidgin (pcm)	Latin	2	–
Lingala (lin)	Latin	1	7

Table 5: Number of OpenWHO documents and sentences per language. Low-resource languages are in **bold**.

B Sentence splitting performance

Method	Tamil	Armenian	Azerbaijani	Macedonian	Kazakh	Tetun	Georgian	Albanian
pysbd	**86.8**	**87.8**	**90.9**	**89.6**	82.0	94.0	92.1	**91.3**
nltk	80.7	35.6	85.6	82.9	82.0	85.5	91.8	88.0
stanza	80.3	83.4	88.9	84.0	**89.1**	**94.6**	**93.2**	76.1

Table 6: Sentence splitting performance (Accuracy %) per language. The best score for each language is highlighted in bold.

C OpenWHO performance per language

Model	mkd	kaz	kat	hye	aze	sqi	tet	som	sin	AVG
MADLAD-400 10B	58.37 / 3.25	47.29 / 4.59	15.81 / 14.09	37.27 / 5.25	40.54 / 6.24	54.97 / 3.87	44.29 / 7.35	48.13 / 6.44	39.48 / 6.04	42.73 / 6.22
NLLB-200 3.3B	50.39 / 3.18	42.94 / 3.64	38.27 / 4.22	39.19 / 4.20	45.23 / 4.32	57.50 / 3.09	– / –	47.05 / 4.40	39.69 / 3.40	45.03 / 3.81
NLLB-200 54B	56.17 / 2.94	56.55 / **3.15**	43.90 / 4.21	42.91 / 3.62	48.78 / 3.92	59.01 / 2.84	– / –	48.23 / 4.50	48.64 / 3.04	50.52 / 3.53
Gemma-3 27B	58.52 / 2.95	48.90 / 3.99	43.76 / 4.66	43.37 / 4.09	46.09 / 4.36	58.12 / 3.11	36.85 / 8.82	46.01 / 5.62	39.32 / 5.11	48.01 / 4.24
DeepSeek-V3 671B	59.07 / 2.98	50.39 / 3.69	46.27 / 3.78	47.41 / 3.39	47.84 / 3.73	59.02 / 2.89	47.64 / 7.12	43.36 / 6.18	41.70 / 4.68	49.38 / 3.92
Gemini 2.5 Flash	**62.83 / 2.65**	**57.41** / 3.17	**50.28 / 3.00**	**49.40 / 3.00**	**52.62 / 3.51**	**60.33 / 2.72**	**51.86 / 6.22**	**55.09 / 4.15**	**54.58 / 2.58**	**55.32 / 3.10**

Table 7: Overall performance (ChrF++ / MetricX) on the OpenWHO test set. LLM scores represent their optimal strategy (max in Table 8). The best score in each column is in **bold**.

Model	Strategy	mkd	kaz	kat	hye	aze	sqi	tet	som	sin	AVG	Δ
Gemini 2.5 Flash	Sentence level	57.8	54.9	46.2	46.1	48.6	57.0	46.8	52.2	50.9	51.2	–
	Sentence window	58.5	54.8	47.2	45.6	48.9	58.5	46.7	48.9	50.1	51.0	−0.2
	Sentence + doc context	58.6	54.9	45.4	46.4	50.2	57.8	47.3	51.3	51.4	51.5	+0.3
	Document level	62.5	**57.4**	49.0	49.1	51.6	60.0	51.5	55.0	**54.6**	54.5	+3.3
	Doc-level + self-correct	**62.8**	56.3	**50.3**	**49.4**	**52.6**	**60.3**	**51.9**	**55.1**	54.5	**54.8**	+3.6
DeepSeek-V3	Sentence level	57.0	49.2	42.6	44.3	45.2	57.0	43.5	41.2	40.7	46.7	–
	Sentence window	57.8	49.5	44.9	45.4	47.6	58.1	44.5	42.8	**41.7**	48.0	+1.3
	Sentence + doc context	58.3	49.7	44.2	44.6	46.6	57.9	43.3	42.1	40.2	47.4	+0.7
	Document level	**59.1**	**50.4**	**46.3**	45.0	**47.8**	**59.0**	**47.6**	42.9	40.7	**48.7**	+2.0
	Doc-level + self-correct	55.7	46.3	43.3	**47.4**	47.4	57.1	45.7	**43.4**	41.3	47.5	+0.8
Gemma-3 27B	Sentence level	58.1	48.3	**43.8**	**43.4**	45.8	**58.1**	35.3	**46.0**	38.6	**46.4**	–
	Sentence window	58.3	48.5	43.1	43.0	45.7	58.0	32.7	45.4	37.7	45.8	−0.6
	Sentence + doc context	56.6	47.5	41.0	40.2	43.1	56.8	32.6	44.4	38.0	44.5	−1.9
	Document level	**58.5**	**48.9**	40.9	42.5	**46.1**	57.5	36.5	45.2	**39.3**	46.2	−0.2
	Doc-level + self-correct	57.4	46.2	42.6	41.5	46.0	57.3	**36.9**	45.4	38.3	45.7	−0.6

Table 8: Effect of different context strategies on LLM performance on the OpenWHO test set (ChrF++). The ‘Δ‘ column shows the change relative to the ‘sentence level‘ baseline.

D WMT24++ performance per language

Model	tam	zul	bul	srp	swh	AVG
NLLB-200 3.3B	43.85 / 3.52	**63.35 / 3.17**	58.39 / 3.09	51.59 / 3.16	52.17 / 4.62	53.87 / 3.51
NLLB-200 54B	45.52 / 3.38	58.39 / 3.23	59.80 / 2.78	53.18 / 2.8	51.02 / 5.07	53.58 / 3.45
MADLAD-400 10B	40.68 / 3.98	38.24 / 5.69	59.40 / 2.88	47.38 / 5.38	46.02 / 7.1	46.34 / 5.01
Gemma-3 27B	45.14 / 2.99	43.95 / 5.93	59.55 / 2.50	52.36 / 2.62	52.56 / 3.83	50.71 / 3.61
DeepSeek-V3 671B	43.95 / 3.50	49.41 / 4.55	58.27 / 2.66	52.53 / 2.60	52.84 / 3.72	51.40 / 3.42
Gemini 2.5 Flash	**45.84 / 2.61**	54.77 / 3.25	**61.40 / 2.33**	**56.83 / 2.29**	**55.29 / 2.99**	**54.83 / 2.69**

Table 9: Overall performance (ChrF++ / MetricX) on the **WMT24++ news** test set. LLM scores represent their optimal context strategy (see Table 11).

Model	tam	zul	bul	srp	swh	AVG
NLLB-200 3.3B	30.74 / 7.85	46.51 / 4.99	47.29 / 4.80	42.17 / 5.44	45.16 / 6.29	42.37 / 5.87
NLLB-200 54B	30.91 / 7.99	46.55 / 4.92	48.63 / 4.57	44.27 / 4.93	44.64 / 6.74	43.00 / 5.83
MADLAD-400 10B	27.44 / 9.21	35.42 / 6.76	47.38 / 4.75	37.73 / 7.02	38.67 / 7.94	37.33 / 7.14
Gemma-3 27B	38.29 / 4.59	38.26 / 6.71	54.98 / 3.45	47.94 / 3.89	47.93 / 5.24	45.48 / 5.26
DeepSeek-V3 671B	38.10 / 4.99	44.33 / 5.64	53.70 / 3.55	49.58 / 3.61	48.69 / 5.08	46.88 / 4.57
Gemini 2.5 Flash	**39.47 / 3.94**	**50.99 / 4.30**	**57.60 / 3.30**	**53.21 / 3.18**	**52.03 / 4.09**	**50.66 / 3.76**

Table 10: Overall performance (ChrF++ / MetricX) on the **WMT24++ literary** test set. LLM scores represent their optimal context strategy (see Table 11).

Model	Strategy	tam	zul	bul	srp	swh	AVG
Gemini	*News*						
	Sent-level	**45.92** / 2.65	53.07 / 3.33	59.93 / 2.47	53.09 / 2.39	**55.90** / 3.03	53.58 / 2.77
	Doc-level	45.84 / **2.61**	**54.77 / 3.25**	**61.40 / 2.33**	**56.83 / 2.29**	55.29 / **2.99**	**54.83 / 2.69**
	Literary						
	Sent-level	34.41 / 5.71	44.34 / 5.12	49.80 / 4.41	44.91 / 4.66	48.00 / 4.82	44.29 / 4.94
	Doc-level	**39.47 / 3.94**	**50.99 / 4.30**	**57.60 / 3.30**	**53.21 / 3.18**	**52.03 / 4.09**	**50.66 / 3.76**
DeepSeek-V3	*News*						
	Sent-level	**43.95 / 3.50**	**49.41** / 4.60	**58.27** / 2.80	**52.53** / 2.73	**52.84** / 3.98	**51.40** / 3.52
	Doc-level	43.40 / 3.55	48.16 / **4.55**	56.96 / **2.66**	52.23 / **2.60**	52.14 / **3.72**	50.58 / **3.42**
	Literary						
	Sent-level	33.36 / 6.37	41.75 / 5.84	49.53 / 4.40	46.08 / 4.75	46.97 / 5.44	43.54 / 5.36
	Doc-level	**38.10 / 4.99**	**44.33 / 5.64**	**53.70 / 3.55**	**49.58 / 3.61**	**48.69 / 5.08**	**46.88 / 4.57**
Gemma-3 27B	*News*						
	Sent-level	**45.14 / 2.99**	**43.95 / 5.93**	59.55 / 2.62	52.36 / 2.69	**52.56 / 3.83**	**50.71 / 3.61**
	Doc-level	45.01 / 3.17	42.90 / 6.47	**60.28 / 2.50**	**52.43 / 2.62**	52.24 / 3.94	50.57 / 3.74
	Literary						
	Sent-level	34.53 / 5.84	**38.26 / 6.71**	50.29 / 4.27	42.66 / 4.84	46.21 / 5.33	42.39 / 5.40
	Doc-level	**38.29 / 4.59**	33.11 / 9.11	**54.98 / 3.45**	**47.94 / 3.89**	**47.93 / 5.24**	**44.45 / 5.26**

Table 11: Comparison of sentence-level and document-level strategies on the WMT24++ test set (ChrF++ / MetricX). The colored delta in the 'AVG' column shows the change relative to the 'sentence-level' baseline within each domain.

E AutoMQM results

Model	mkd	kaz	kat	hye	aze	sqi	som	sin	AVG
AutoMQM score (lower is better)									
NLLB-54B	-4.72	-3.35	-5.44	-4.34	-4.27	-3.11	-6.08	-4.54	-4.48
Gemini 2.5 Flash	**-2.80**	**-3.15**	**-3.12**	**-2.55**	**-3.01**	**-2.59**	**-4.88**	**-2.40**	**-3.06**
Difference in error counts (Gemini - NLLB)									
Error category	**mkd**	**kaz**	**kat**	**hye**	**aze**	**sqi**	**som**	**sin**	**AVG**
Accuracy									
Mistranslation	-4	-4	-33	-13	-5	-15	-25	-22	↓ **-15.1**
Overtranslation	-6	19	-7	13	-2	2	18	7	↑ **+5.5**
Undertranslation	0	-4	-9	0	2	-3	-7	-5	↓ **-3.3**
Addition	4	1	4	1	-5	-3	2	-4	→ **0.0**
Omission	5	13	1	3	1	4	-1	5	↑ **+3.9**
Untranslated	-10	0	-8	-1	-1	-4	-4	-5	↓ **-4.1**
Total Accuracy	**-11**	**25**	**-52**	**3**	**-10**	**-19**	**-17**	**-24**	↓ **-13.1**
Fluency									
Grammar	-1	-4	-8	-7	-1	5	-14	-6	↓ **-4.5**
Spelling	-7	-2	0	-10	-2	4	4	-81	↓ **-11.8**
Punctuation	-1	-15	-1	-9	-1	3	0	-5	↓ **-3.6**
Total Fluency	**-9**	**-21**	**-9**	**-26**	**-4**	**12**	**-10**	**-92**	↓ **-19.9**
Style									
Awkward	3	12	-11	-2	-7	11	10	0	↑ **+2.0**
Register	0	2	-2	-1	-1	-5	3	-2	↓ **-0.8**
Total Style	**3**	**14**	**-13**	**-3**	**-8**	**6**	**13**	**-2**	↑ **+1.3**
Terminology									
Inconsistent	-5	1	0	1	3	-1	0	4	↑ **+0.4**
Wrong	-8	1	-10	-7	-7	0	6	-12	↓ **-4.6**
Total Terminology	**-13**	**2**	**-10**	**-6**	**-4**	**-1**	**6**	**-8**	↓ **-4.3**
Non-translation	**-4**	**0**	**-10**	**-3**	**0**	**-1**	**-4**	**-1**	↓ **-2.9**

Table 12: AutoMQM analysis comparing NLLB-54B and Gemini 2.5 Flash (sentence-level) on the OpenWHO test set. **Top**: Overall MQM scores (higher is better). Gemini consistently outperforms NLLB. **Bottom**: Difference in error counts (Gemini errors minus NLLB errors) per category. Negative values indicate Gemini made fewer errors for that category. Gemini outputs less major errors like mistranslations and incorrect terminology, at the cost of a slight increase in over-translation and omissions.

F Prompts

System: Translate from English to [target lang name]. Give only the translation, and no extra commentary, or chattiness. Wrap the translated sentence in <result></result> tags.

User: <text to translate>She lives in Boston.</text to translate>

Assistant: <result>[Google Translate of "She lives in Boston." into target lang]</result>

User: <text to translate>[sentence to translate]</text to translate>

Prompt used for **Sentence level** translation. We ask the model to wrap the translation in <result> tags to avoid model commentary interfering with translation accuracy measurement.

System: Using the provided context, translate the "Sentence to translate" from English to [target lang name]. Give only the sentence translation, and no extra commentary, or chattiness. Wrap the translated sentence in <result></result> tags.

User: <context>
Her name is Mary. She lives in Boston. She is a doctor.
</context>
Sentence to translate:
She lives in Boston.

Assistant: <result>[Google Translate of "She lives in Boston." into target lang]</result>

User: <context>
[preceding sentence][sentence to translate][next sentence]
</context>
Sentence to translate:
[sentence to translate]

Prompt used for **Sentence window** translation.

System: Using the provided context, translate the "Sentence to translate" from English to [lang name]. Give only the sentence translation, and no extra commentary, or chattiness.

User: <context>
Her name is Mary. She lives in Boston. She is a doctor.
</context>
Sentence to translate:
She lives in Boston.

Assistant: <result>[Google Translate of "She lives in Boston." into target lang]</result>

User: <context>
[whole document for the sentence]
</context>
Sentence to translate:
[sentence to translate]

Prompt used for **Sentence + doc context** translation.

System: Translate from English to [lang name]. Give only the translation, and no extra commentary, or chattiness. Use the same formatting as the source text to translate, with one sentence per line. Enclose your translation in <result></result> tags.

User: <text to translate>
Her name is Mary.
She lives in Boston.
She is a doctor.
</text to translate>

Assistant: <result>
[Google Translate of "Her name is Mary." into target lang]
[Google Translate of "She lives in Boston." into target lang]
[Google Translate of "She is a doctor." into target lang]
</result>

User: <text to translate>
[document sentence 1]
[document sentence 2]
...
</text to translate>

Prompt used for **Document level** translation.

System: Translate from English to [lang name]. Give only the translation, and no extra commentary, or chattiness. Use the same formatting as the source text to translate, with one sentence per line. Enclose your translation in <result></result> tags.

User: <text to translate>
Her name is Mary.
She lives in Boston.
She is a doctor.
</text to translate>

Assistant: <result>
[Google Translate of "Her name is Mary." into target lang]
[Google Translate of "She lives in Boston." into target lang]
[Google Translate of "She is a doctor." into target lang]
</result>

User: <text to translate>
[document sentence 1]
[document sentence 2]
...
</text to translate>
Assistant: [assistant response from above]
User: Please translate again for a better version. Be particularly mindful of using the right script and tone, of adapting to context, and of translating each sentence faithfully.
<text to translate>[same as above]</text to translate>]

Prompt used for **Doc-level + self-correct** translation.