TULUN: Transparent and Adaptable Low-resource Machine Translation

Raphaël Merx[λ] **Hanna Suominen**[ψ,ϕ] **Lois Hong**[ζ]
Nick Thieberger[λ] **Trevor Cohn**[λ] **Ekaterina Vylomova**[λ]
[λ]The University of Melbourne [ψ]The Australian National University
[ϕ]University of Turku [ζ]Maluk Timor

Abstract

Machine translation (MT) systems that support low-resource languages often struggle on specialized domains. While researchers have proposed various techniques for domain adaptation, these approaches typically require model fine-tuning, making them impractical for non-technical users and small organizations. To address this gap, we propose TULUN,[1] a versatile solution for terminology-aware translation, combining neural MT with large language model (LLM)-based post-editing guided by existing glossaries and translation memories. Our open-source web-based platform enables users to easily create, edit, and leverage terminology resources, fostering a collaborative human-machine translation process that respects and incorporates domain expertise while increasing MT accuracy. Evaluations show effectiveness in both real-world and benchmark scenarios: on medical and disaster relief translation tasks for Tetun and Bislama, our system achieves improvements of 16.90–22.41 ChrF++ points over baseline MT systems. Across six low-resource languages on the FLORES dataset, TULUN outperforms both standalone MT and LLM approaches, achieving an average improvement of 2.8 ChrF++ points over NLLB-54B. TULUN is publicly accessible at bislama-trans.rapha.dev.

1 Introduction

Machine translation (MT) systems have transformed how organizations manage their translation needs (Stefaniak, 2022; Utunen et al., 2023), yet domain accuracy and consistency remain a significant challenge, particularly for low-resource languages (Haddow et al., 2022; Khiu et al., 2024; Marashian et al., 2025). For instance, a health organization we work with in Timor-Leste struggled to leverage MT to accurately translate medical education materials from English to Tetun, despite having a glossary and a corpus of past translations that could inform MT output. Tetun, a low-resource language that is the lingua franca in Timor-Leste, lacks available corpora in the health domain (Merx et al., 2024), making in-domain resources particularly valuable to improve MT accuracy. This case exemplifies a broader challenge: model adaptation and deployment requires technical expertise, and commercial MT providers rarely offer low-resource language support, let alone terminology customization, leaving no practical option for a small organization to rely on MT for low-resource in-domain translation.

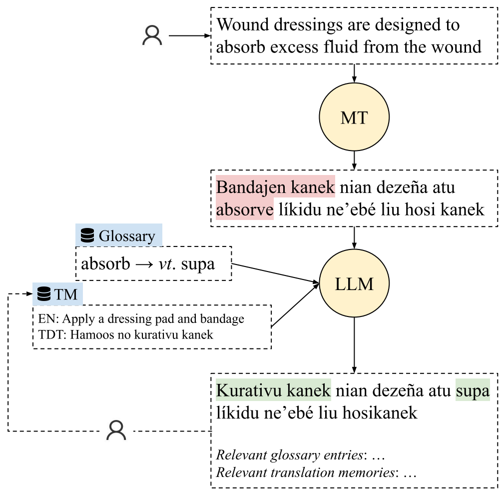

Figure 1: System overview with example translation from English to Tetun (en-tdt). The system components and data are configurable by end-users.

Translation memories and terminology management are well-established tools in professional translation software, improving translation accuracy while reducing cognitive load (Dillon and Fraser, 2006; Drugan et al., 2023). Research has demonstrated that lexicons can bring substantial accuracy gains, particularly for low-resource MT (Jones et al., 2023). However, existing approaches to incorporate terminology constraints into neural

[1]Tulun means "assistance" in Tetun, highlighting a philosophy of augmenting rather than replacing human expertise.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 129–139
July 27 - August 1, 2025

MT systems typically require model fine-tuning (Niehues, 2021; Reid and Artetxe, 2022), making them inaccessible to small organizations (Bane et al., 2023). Recent advances in large language models (LLMs) offer a promising alternative: while LLMs may underperform specialized MT systems for low-resource languages (Robinson et al., 2023), their ability to adapt to new contexts at inference time (Brown et al., 2020) makes them particularly suitable for terminology-aware post-editing (Raunak et al., 2023).

To address these challenges, we propose TULUN, a versatile solution that combines neural MT with LLM-based post-editing, guided by existing glossaries and translation memories (Figure 1). TULUN continuously adapts as new entries are added to the translation memory and glossary. Packaged as an open-source[2] web platform, it relies on a modular architecture that allows users to configure their choice of MT system, LLM, and retrieval options, as well as create, edit, and rely on terminology resources. To see a demo video of TULUN, visit https://youtu.be/fQFwOxzR4MI.

Our system has the following characteristics:

- **Accurate**: On real-world medical and disaster relief translation tasks for Tetun and Bislama (national language of Vanuatu), our system shows impressive improvements of 16.90–22.41 ChrF++ points over baseline MT systems (§4.1).[3] A broader evaluation across six low-resource languages on the FLORES dataset shows TULUN outperforms both standalone MT and LLM approaches (§4.2).
- **User-friendly**: Our usability study, based on the system usability scale (SUS), averages an excellent score of 81.25, with users rating the system's overall usefulness at 5/5 for their translation tasks (§4.1.2).
- **Adaptable**: Target language, MT model, and prompt are all configurable from the user interface (UI). Glossary and translation memories can be bulk-imported and managed through the UI (§3.1).
- **Transparent**: Users can verify how their glossary entries and past translations inform the current translation.
- **Lightweight**: Easy to deploy (§3.2), does not require model training.

[2]Code: github.com/raphaelmerx/tulun/, MIT license

[3]ChrF and ChrF++ refer to evaluation metrics for MT that both apply the ***F***-score for evaluating ***char***acter n-gram matches, but the latter metric also includes word n-grams.

Fundamentally, TULUN represents a shift in MT philosophy, moving away from the paradigm of users as passive consumers of opaque systems (Liebling et al., 2022), toward one where users' expertise and preferences actively shape the translation process (Liu et al., 2025). By making glossary and translation memory matches explicit to users, and by allowing configuration of the underlying data and systems, TULUN aims to foster a transparent, collaborative process that respects and leverages users' domain knowledge. This approach benefits low-resource in-domain translation, where local expertise is often the most valuable resource for producing accurate, culturally appropriate translations (Nekoto et al., 2020).

2 Related Work

Glossary and translation memory integration in MT The integration of custom terminology and translation memories into MT systems can deliver more consistent, domain-adapted translations (Scansani and Dugast, 2021). Recent research has demonstrated that such lexical customization brings substantial accuracy gains, particularly for low-resource MT (Jones et al., 2023). Approaches to incorporate terminology constraints into neural MT models include replacing source words with their target translation in the source (Reid and Artetxe, 2022), and prepending dictionary entries to the source text (Niehues, 2021). However, these approaches typically require either custom models, or model fine-tuning, which can be resource-intensive for smaller organizations (Bane et al., 2023), and prone to catastrophic forgetting (Saunders, 2022). TULUN addresses this gap by providing a deployable solution that requires no model training while delivering terminology-consistent translations.

LLMs and Automated Post-Editing (APE) The ability for LLMs to adapt to new tasks at inference time (Brown et al., 2020) makes them of interest to both MT (Moslem et al., 2023a) and related tasks, such as synthetic data generation and automated post-editing (Moslem et al., 2023b). While their MT accuracy can lag behind that of specialized MT models when translating into low-resource languages (Robinson et al., 2023) or in specialized domains (Uguet et al., 2024), Raunak et al. (2023) find that combining specialized MT with an LLM for APE results in more accurate translations than each module used in isolation (measured using COMET on high-resource language pairs). Con-

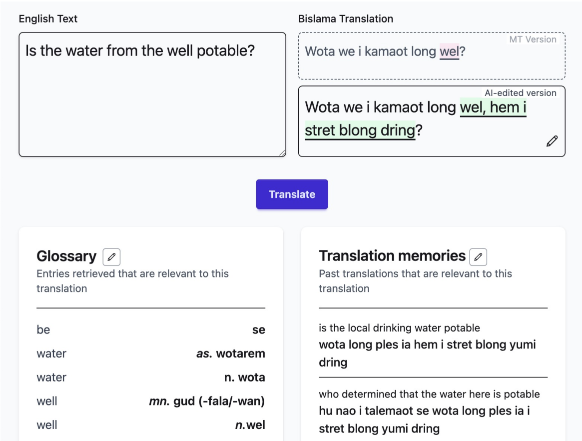

Figure 2: Translation View with the MT text, post-edited text, and the glossary entries and past translations relevant to this translation

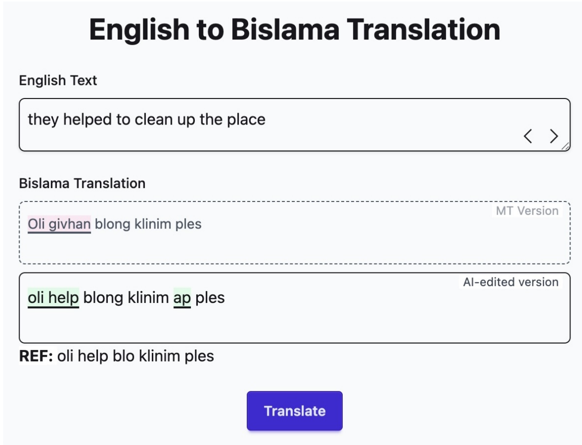

Figure 3: Eval mode: users can browse evaluation results, and see the reference translation

firming the potential of LLMs at the post-editing stage, Ki and Carpuat (2024) give external feedback to an LLM to improve MT outputs, and Lu et al. (2025) find that LLMs can identify and correct translation mistakes across high and low-resource language pairs. These findings suggest that LLMs can be valuable for terminology-aware post-editing, where their adaptation capabilities are combined with the robustness of specialized MT systems.

Our contribution Building on research showing the potential of terminology-aware translation and LLM-based post-editing, TULUN extends the applicability of these techniques through a modular user-friendly interface, and demonstrates their effectiveness across diverse scenarios, from applied use cases (§4.1) to systematic evaluation (§4.2). Our system serves as both a practical tool for immediate use, and as a research platform that demonstrates how these models can be effectively combined. To our knowledge, it is the only open-source terminology-aware MT tool that supports low-resource languages like Tetun and Bislama.

3 System Design & Implementation

3.1 System Design

Translation View When users open TULUN, they are presented with the Translation View, where they can enter a sentence or paragraph, and have it first machine translated, then post-edited using an LLM (Figure 2). Post-editing changes are highlighted in the machine translated text (in red) and in the post-edited final translation (in green). In addition, users are presented with the relevant glossary entries and similar sentences retrieved to guide the LLM for post-editing. For example, in Figure 2, "potable" is translated incorrectly by the MT model, but the LLM identifies the correct translation ("stret blong dring") from the translation memory, and applies this change at the post-editing phase.

Glossary and Translation Memory View Both the glossary entries and the translation memories are editable by end-users (if they are given permission to do so, a setting configured through the admin). This allows users to iteratively improve the translation quality, by adding or correcting entries as missing or incorrect entries are found. In addition, data can be bulk-imported from a CSV, and a new translation memory can be added from the current (source, final translation) pair directly from the Translation View in a dedicated modal.

System Configuration Admin users can set through the web UI: (1) Site metadata, including target language and site title, (2) MT model, with a choice between Google Translate or any model available on HuggingFace through its "translation" pipeline,[4] and (3) LLM configuration for post-editing, including the choice of LLM among the hundreds of providers supported by LiteLLM,[5] the system prompt, and the number of translation memories retrieved for in-context learning.

Evaluation Mode TULUN includes a dedicated evaluation feature that allows users to assess translation quality against reference translations. After uploading an evaluation dataset through the admin UI, users can navigate through these test translations within the Translation View (Figure 3). When

[4] huggingface.co/models?pipeline_tag=translation
[5] docs.litellm.ai

a source sentence from the evaluation set is entered, the system automatically displays both the system-generated translation and the human reference for comparison. This helps users identify areas where improvements to the glossary, translation memory, LLM system prompt, or MT model selection might be beneficial.

3.2 System Architecture

Backend We implement TULUN as a configurable Django project, with data models for the glossary and translation memory, a `Translator` class that implements compatibility with either Google Translate (through the Cloud Translation API)[6] or the HuggingFace Translation pipeline, and an LLM post-editing layer that supports hundreds of providers via LiteLLM, with the choice of model and prompt configurable through the web UI.

Glossary and Translation Memory Retrieval At the post-edition stage, relevant glossary entries are retrieved using {1,2}-gram overlap with the input text tokens (tokenization is handled by spaCy's `en_core_web_sm` model). Relevant translation memories are the top N (where N is configurable) BM25 matches between the input text and the source side of the memory, implemented using the Tantivy library.[7] We select BM25 for retrieval because of its high performance on retrieving translation memories for MT through in-context learning (Bouthors et al., 2024).

Prompt Design The glossary and translation memories are injected in the LLM prompt to inform post-editing (see an example prompt in Appendix A). For all evaluations in Section 4, we rely on a system prompt that includes few-shot examples (Brown et al., 2020) with chain of thought reasoning (Wei et al., 2022). The prompt can be manually adjusted in the admin UI.

Deployment We package TULUN using Docker and Docker Compose, allowing organizations to run the system on their infrastructure with minimal setup. The Docker configuration handles dependencies and environment configuration, while Docker Compose simplifies the orchestration process. This packaging approach ensures that the system can be deployed consistently across different environments.

[6]pypi.org/project/google-cloud-translate/
[7]pypi.org/project/tantivy/

4 Evaluation

4.1 Applications: Tetun Medical Translation and Bislama Disaster Relief Translation

We evaluate TULUN in two real-world low-resource language settings with distinct domain needs. For Tetun medical translation, we collaborate with Maluk Timor,[8] a health organization in Timor-Leste that regularly translates health education materials from English to Tetun. This translation work is needed as health workers (particularly nurses and community health workers) are most comfortable learning in Tetun rather than English or Portuguese (Greksakova, 2018). Maluk Timor reports that professional translation costs represent a significant organizational expense, and while they utilize machine translation, MT outputs typically require substantial post-editing to ensure accuracy and domain-appropriateness. For Bislama disaster relief translation, we partner with researchers working on a Pacific Creoles project[9] who need to translate transcripts while maintaining consistent terminology. Both scenarios provide practical test cases for TULUN's ability to support organizations working with specialized domains in low-resource languages.

4.1.1 MT Accuracy Evaluation

Problem Statement From both organizations, we get a glossary (Tetun medical glossary: 2,698 entries; Bislama dictionary: 5,769 entries) and a translation memory (1,018 sentences for Tetun, 3,353 utterances for Bislama). We reserve some of the translation memory for evaluation (451 sentences for Tetun, 841 utterances for Bislama). Both datasets belong to their respective organizations, but are available upon request for research purposes with appropriate data sharing agreements.

Choice of Baseline and Prompt Given that neither Tetun nor Bislama are part of NLLB, we initially use MADLAD-400 10B (Kudugunta et al., 2023) as baseline. We find that it performs poorly on Bislama, often copying the English source, and choose to also evaluate OPUS-MT models as an alternative baseline[10] (Tiedemann et al., 2024). For post-editing, we use Gemini 2.0 Flash (Gemini Team et al., 2024b,a), with a prompt that describes the post-editing task and gives a few examples (see an example in Appendix A).

[8]maluktimor.org
[9]anu.edu.au/projects/modelling-pacific-creole-languages
[10]opus-mt-en-tdt; opus-mt-en-bi

Method	TDT	BIS	AVG
NMT only			
MADLAD-400-10B	35.78	15.22	24.37
opus-mt-en-**	16.01	32.60	24.31
LLM only			
Gemini, 0-shot	44.05	37.78	39.63
Gemini, 10-shot	44.59	45.37	44.98
Ours: NMT + LLM APE			
MADLAD + Gemini	47.87	47.94	47.91
Δ vs MADLAD	+12.09	+32.72	+22.41
opus-mt + Gemini	34.27	48.14	41.21
Δ vs opus-mt	+18.26	+15.54	+16.90

Table 1: ChrF++ score comparison on test sets for Tetun (tdt) and Bislama (bis). LLM only uses the system prompt from (Caswell et al., 2025), and examples from the Tetun/Bislama corpus.

Results Our approach demonstrates substantial translation quality improvements over baseline MT systems for both settings, with LLM post-editing yielding ChrF++ gains of 16.90–22.41 points (Table 1). Qualitatively, we observe that for both settings, LLM post-editing helps (1) improve in-domain terminology translation (see an example in Figure 2) and (2) repair hallucinations that are frequent for out-of-domain MT inference (Raunak et al., 2021), with the latter particularly relevant for the speech domain covered in our Bislama experiment.

4.1.2 Usability and Usefulness Study

Usability We perform a usability study of the TULUN interface using the System Usability Scale (SUS, Brooke, 1996). We collect two responses, one from the clinical director at Maluk Timor, and the other from a linguist working with Bislama. We get an average SUS score of 81.25, corresponding to an excellent perceived usability (Bangor et al., 2008).

Usefulness To measure usefulness, we adapt the technology acceptance model (TAM, Venkatesh et al., 2003) questions on general usefulness to our translation context. We get average scores between 4 and 5 for all questions (out of 5), with a 5/5 score for overall usefulness ("Overall, I find this system useful for my translation tasks"), a 4.5/5 score for the system impact on translation quality ("Using this system improves the quality of my translations"), and a 4.5/5 score for the system's helpfulness to translate technical content ("Using this system makes it easier to translate technical/specialized content").

We report all questions, with scores for each annotator, in Appendix B.

4.2 Generalizable Evaluation: FLORES-200

Languages To measure the broader efficacy of our solution, we work with six low-resource languages (Tok Pisin TPI, Dzongkha DZO, Quechua QUY, Rundi RUN, Lingala LIN, Assamese ASM), spanning four continents and three different scripts. We select these languages because they are all (1) low-resource (2) institutionalized, which makes them more likely to be standardized and in demand for MT (Bird, 2024), (3) part of the FLORES-200 evaluation benchmark (Costa-jussà et al., 2024) and (4) represented in the GATITOS glossary project (Jones et al., 2023).

Data We evaluate on all 1,012 sentences from the FLORES-200 "devtest" split, using NLLB-54B as a baseline MT model.[11] For populating the post-editing prompt, we rely on glossary entries from GATITOS, and on parallel sentences from allenai/nllb, which is based on the NLLB data mining strategy.

Models We compare MT performance using ChrF++ (given the lack of neural metrics available for the languages we work with) on the following setups: (1) MT only, using NLLB-54B (2) LLM only with 10 fixed examples (3) MT + LLM APE (our solution). We use Gemini 2.0 Flash (Gemini Team et al., 2024a) as LLM throughout our experiments.

Results Our system achieves higher average accuracy than both baselines, by 2.83 and 2.15 ChrF++ points for NLLB and Gemini respectively (Table 2). Interestingly, we find that Gemini often beats NLLB-54, but that our system tends to improve on NLLB or Gemini, whichever is higher. One exception, Rundi (-1.34 points), is discussed in Section 5.

This evaluation shows the effectiveness of our approach, even on general domain benchmarks like FLORES-200. The sharp difference in accuracy gains between this experiment and the specialized domain evaluation in Section 4.1 shows that our system is most useful for specialized domains, where adaptation to new terminology and translation style is needed most.

[11]We get FLORES-200 translations by NLLB-54B from tinyurl.com/nllbflorestranslations

Method	TPI	DZO	QUY	RUN	LIN	ASM	**AVERAGE**
Baselines: NMT / LLM only							
NLLB 54B	41.61	34.67	26.87	42.51	47.99	35.91	38.26
Gemini, 10-shot	44.07	30.74	31.28	39.83	49.42	38.29	38.94
TULUN: *NMT + LLM APE*							
NLLB + Gemini APE	46.80	35.76	32.40	41.17	50.58	39.80	41.09
Δ vs NLLB 54B	+5.19	+1.09	+5.53	-1.34	+2.59	+3.89	+2.83

Table 2: ChrF++ score comparison on FLORES for 6 low-resource languages, using the Gatitos glossary and sentences from the NLLB training set in the translation memory

5 Discussion

Accuracy Across Languages While our solution is effective in both applied and theoretical scenarios, the impact of LLM post-editing on MT accuracy varies (Table 2), including a negative effect for Rundi (-1.34 ChrF++ points). Through qualitative analysis and evaluation without injecting the glossary in the prompt for Rundi (resulting in +0.25 points compared to NLLB), we find this is due to incorrect word changes by the LLM using the glossary, highlighting the need for prompt tuning, and for glossary adjustments. We further discuss this error mode, and give an example, in Appendix C.

User-friendliness and Adaptability Our usability study (§4.1.2) confirms TULUN's ease of use, but the system's configurability (§3.1) presents a potential trade-off: while it allows users to adapt the system to their needs, it also requires some understanding of MT and LLM options. Future work could explore intelligent defaults and guidance to improve accessibility, including a system module for prompt tuning (see also §6).

Explainability The transparency provided by displaying glossary matches and translation memory hits helps users understand how the system post-edited translations (see responses to question 5 in Appendix B), but relies on the LLM's capability to use these resources effectively. For extremely low-resource languages with complex morphology or rare scripts, where LLMs have minimal prior language exposure, this assumption might not hold, resulting in higher rates of hallucination.

6 Conclusion & Future Work

In this work, we present TULUN, an open-source translation system that combines MT with LLM-based post-editing for a more accurate and adaptable low-resource translation. By leveraging existing glossaries and translation memories to guide the post-editing process, our approach achieves significant improvements over standalone MT, without requiring model fine-tuning or technical expertise. It also introduces a change of paradigm in MT, where end-users are given the opportunity to constantly improve the translation process, fostering a transparent, collaborative process that respects local expertise.

Reflecting on our experiences in designing and developing TULUN, we lay out the following future research directions:

Prompt Engineering and Optimization While our current prompt design yields promising results, future work could explore systematic prompt engineering approaches to maximize post-editing accuracy. This includes automatically generating language-specific prompts using techniques like DSPy's MIPRO (Khattab et al., 2024), optimizing few-shot examples based on error patterns, and developing prompts that better handle linguistic nuances in different target languages.

Offline Deployment Option To better serve users with limited internet connectivity, we plan to explore lightweight LLM options that can run locally. This likely would involve specialized small models fine-tuned specifically for the post-editing task, enabling organizations to maintain terminology consistency without relying on cloud-based LLM providers.

Extended Usability Study Future work will include a larger comprehensive usability evaluation, with a more diverse set of users across different language communities. This would enable us to better understand how different users (translators, subject matter experts, and community members) interact with the system, helping refine the interface and process.

Broader LLM Evaluation While the current study utilized Gemini 2.0 Flash due to its cost-effectiveness, future work will extend our evaluations to other state-of-the-art LLMs, including open models such as DeepSeek-R1 (DeepSeek-AI et al., 2025).

Ethics and Broader Impact Statement

TULUN is designed to augment human translation expertise rather than replace it, particularly for low-resource languages where professional translation resources are limited. The Tetun medical glossary and Bislama dictionary used in our evaluations belong to their respective organizations and were used with explicit permission for research purposes. Usability study participants engaged voluntarily in this research and have been actively using the system since its creation. Two of the participants are co-authors of this paper, ensuring their contributions are properly acknowledged and used directly to inform our system design and evaluation.

We recognize that translation technologies can impact professional translators' workflows, and TULUN's interface aims to give users control over the translation process while maintaining human oversight, especially for sensitive domains like health. We acknowledge that the system's effectiveness will vary across languages and domains, and plan to further research language-specific limitations that warrant refinement.

Acknowledgments

We thank Maluk Timor, which was instrumental in coming up with an initial problem statement that TULUN responds to, and helped test and refine the system. We are also grateful for Gabriela Leite Soares' proofreading of this paper.

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

References

Fred Bane, Anna Zaretskaya, Tània Blanch Miró, Celia Soler Uguet, and João Torres. 2023. Coming to terms with glossary enforcement: A study of three approaches to enforcing terminology in NMT. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 345–353, Tampere, Finland. European Association for Machine Translation.

Aaron Bangor, Philip T. Kortum, and James T. Miller. 2008. An Empirical Evaluation of the System Usability Scale. *International Journal of Human–Computer Interaction*, 24(6):574–594. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447310802205776.

Steven Bird. 2024. Must NLP be Extractive? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14915–14929, Bangkok, Thailand. Association for Computational Linguistics.

Maxime Bouthors, Josep Crego, and François Yvon. 2024. Retrieving Examples from Memory for Retrieval Augmented Neural Machine Translation: A Systematic Comparison. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3022–3039, Mexico City, Mexico. Association for Computational Linguistics.

john Brooke. 1996. SUS: A 'Quick and Dirty' Usability Scale. In *Usability Evaluation In Industry*. CRC Press. Num Pages: 6.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diane, Koulako Moussa Doumbouya, Djibrila Diane, and Solo Farabado Cissé. 2025. SMOL: Professionally translated parallel data for 115 under-represented languages. *arXiv preprint*. ArXiv:2502.12301 [cs].

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Jeff Wang, and

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

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.

Sarah Dillon and Janet Fraser. 2006. Translators and TM: An investigation of translators' perceptions of translation memory adoption. *Machine Translation*, 20(2):67–79.

Joanna Drugan, Joss Moorkens, María Fernández-Parra, Andrew Rothwell, and Frank Austermuehl. 2023. *Translation Tools and Technologies*, 1 edition. Routledge, London.

Gemini Team, Demis Hassabis, and Koray Kavukcuoglu. 2024a. Introducing Gemini 2.0: our new AI model for the agentic era.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and et al. 2024b. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint*.

Zuzana Greksakova. 2018. *Tetun in Timor-Leste: the role of language contact in its development*. doctoralThesis, 00500::Universidade de Coimbra. Accepted: 2018-09-03T09:37:10Z.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of Low-Resource Machine Translation. *Computational Linguistics*, 48(3):673–732. Place: Cambridge, MA Publisher: MIT Press.

Alexander Jones, Isaac Caswell, Orhan Firat, and Ishank Saxena. 2023. GATITOS: Using a New Multilingual Lexicon for Low-resource Machine Translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 371–405, Singapore. Association for Computational Linguistics.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan A, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. DSPy: Compiling Declarative Language Model Calls into State-of-the-Art Pipelines. In *Proceedings of the Twelfth International Conference on Learning Representations*.

Eric Khiu, Hasti Toossi, David Anugraha, Jinyu Liu, Jiaxu Li, Juan Flores, Leandro Roman, A. Seza Doğruöz, and En-Shiun Lee. 2024. Predicting Machine Translation Performance on Low-Resource Languages: The Role of Domain Similarity. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1474–1486, St. Julian's, Malta. Association for Computational Linguistics.

Dayeon Ki and Marine Carpuat. 2024. Guiding large language models to post-edit machine translation with error annotations. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4253–4273, 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.

Daniel Liebling, Katherine Heller, Samantha Robertson, and Wesley Deng. 2022. Opportunities for human-centered evaluation of machine translation systems. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 229–240, Seattle, United States. Association for Computational Linguistics.

Sinuo Liu, Chenyang Lyu, Minghao Wu, Longyue Wang, Weihua Luo, Kaifu Zhang, and Zifu Shang. 2025. New Trends for Modern Machine Translation with Large Reasoning Models. *arXiv preprint*. ArXiv:2503.10351 [cs].

Qingyu Lu, Liang Ding, Kanjian Zhang, Jinxia Zhang, and Dacheng Tao. 2025. MQM-APE: Toward high-quality error annotation predictors with automatic post-editing in LLM translation evaluators. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5570–5587, Abu Dhabi, UAE. Association for Computational Linguistics.

Ali Marashian, Enora Rice, Luke Gessler, Alexis Palmer, and Katharina von der Wense. 2025. From Priest to Doctor: Domain Adaptation for Low-Resource Neural Machine Translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7087–7098, Abu Dhabi, UAE. Association for Computational Linguistics.

Raphaël Merx, Aso Mahmudi, Katrina Langford, Leo Alberto de Araujo, and Ekaterina Vylomova. 2024. 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) at LREC-COLING 2024*, pages 1–11, Torino, Italia. ELRA and ICCL.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023a. Adaptive Machine Translation with Large Language Models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. 2023b. Domain Terminology Integration into Machine Translation: Leveraging Large Language Models. In *Proceedings of the Eighth Conference on Machine Translation*, pages 902–911, Singapore. Association for Computational Linguistics.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Jan Niehues. 2021. Continuous learning in neural machine translation using bilingual dictionaries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 830–840, Online. Association for Computational Linguistics.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Awadalla, and Arul Menezes. 2023. Leveraging GPT-4 for Automatic Translation Post-Editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12009–12024, Singapore. Association for Computational Linguistics.

Machel Reid and Mikel Artetxe. 2022. PARADISE: Exploiting parallel data for multilingual sequence-to-sequence pretraining. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 800–810, Seattle, United States. Association for Computational Linguistics.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Danielle Saunders. 2022. Domain Adaptation and Multi-Domain Adaptation for Neural Machine Translation: A Survey. *Journal of Artificial Intelligence Research*, 75:351–424.

Randy Scansani and Loïc Dugast. 2021. Glossary functionality in commercial machine translation: does it help? a first step to identify best practices for a language service provider. In *Proceedings of Machine Translation Summit XVIII: Users and Providers*

Track, pages 78–88, Virtual. Association for Machine Translation in the Americas.

Karolina Stefaniak. 2022. Machine Translation and Terminology: The Experience of the European Commission. In *Proceedings of the New Trends in Transaltion and Technology Conference – NeTTT 2022*, Rhodes Island, Greece.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raúl Vázquez, and Sami Virpioja. 2024. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, 58(2):713–755.

Celia Uguet, Fred Bane, Mahmoud Aymo, João Torres, Anna Zaretskaya, and Tània Blanch Miró Blanch Miró. 2024. LLMs in post-translation workflows: Comparing performance in post-editing and error analysis. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 373–386, Sheffield, UK. European Association for Machine Translation (EAMT).

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

Viswanath Venkatesh, Michael G. Morris, Gordon B. Davis, and Fred D. Davis. 2003. User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly*, 27(3):425–478. Publisher: Management Information Systems Research Center, University of Minnesota.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

A Example Prompt for Post-editing

SYSTEM

You are an expert translator. I am going to give you relevant glossary entries, and relevant past translations, where the first is the English source, the second is a machine translation of the English to Tetun, and the third is the Tetun reference translation. The sentences will be written
English: <sentence>
MT: <machine translated sentence>
Tetun: <translated sentence>.

After the example pairs, I am going to provide another sentence in English and its machine translation, and I want you to translate it into Tetun. Give only the translation, and no extra commentary, formatting, or chattiness. Translate the text from English to Tetun.

USER

<glossary entries>
no 0: check -> vt. kontrola.
no 1: burn -> n. keimadura (ahi-haan)
no 2: assessment -> n. avaliasaun.
</glossary entries>

<past translations>
English: Antibiotic prophylaxis for burns, wounds and bites, and treatment
MT: Profilaxia antibiótiku ba kanek, feridu no morde, no tratamentu
Tetun: Ba profilaxia antibiotiku kelmadura (ai-han), kanek, tata, tohar (tohar nakloke), no tratamentu.
...
</past translations>

Text to translate:
English: Always check burn again a couple of hours after first assessment, unless burn has been dressed.
MT: Sempre kontrola tan kanek rua oras hafoin avaliasaun dahuluk, la'ós kanek ne'ebé hetan tratamentu
Tetun:

ASSISTANT

Sempre kontrola fali keimadura (ahi-haan) iha oras balun nia laran depois de avaliasaun dahuluk, se karik keimadura falun ona.

B Usability and Usefulness Responses

Statement	R1	R2
1. I think that I would like to use this system frequently.	5	5
2. I found the system unnecessarily complex.	1	1
3. I thought the system was easy to use.	5	5
4. I think that I would need the support of a technical person to be able to use this system.	1	2
5. I found the various functions in this system were well integrated.	1	2
6. I thought there was too much inconsistency in this system.	1	2
7. I would imagine that most people would learn to use this system very quickly.	5	3
8. I found the system very cumbersome to use.	1	2
9. I felt very confident using the system.	5	4
10. I needed to learn a lot of things before I could get going with this system.	2	2

Table 3: Usability ratings (1-5 scale, 1 = strongly disagree, 5 = strongly agree)

Statement	R1	R2
1. Using this system improves the quality of my translations.	5	4
2. Using this system increases my productivity when translating documents.	5	4
3. Using this system enhances my effectiveness in maintaining terminology consistency.	4	4
4. Using this system makes it easier to translate technical/specialized content.	4	5
5. The glossary and translation memory features are useful for my translation work.	5	5
6. Overall, I find this system useful for my translation tasks.	5	5

Table 4: Usefulness ratings (1-5 scale, 1 = strongly disagree, 5 = strongly agree)

C Error Mode: Incorrect Glossary Applications by the LLM

Through qualitative analysis, we find that LLM post-editing can sometimes degrade MT accuracy, in particular when LLMs blindly apply glossary entries to the translation candidate. These errors fall into several categories:

1. **Glossary conflicts with the translation memory**: in our Bislama and Tetun setups, we observe that glossaries, put together by linguists, tend to rely more on native words, while translation memories, put together by professionals of the domains studied, tend to rely more on borrowed terms. This conflicting information given to the LLM can result in incorrect post-editing. It also demonstrates the fluidity of low-resource languages, and the usefulness of having translations that are grounded in individual preferences.

2. **Glossary entry is not a correct translation in the current context**: Words often have multiple meanings depending on context, but glossaries typically provide only one or a few translations per entry. For example, in Bislama, the English word "touch" in "This touches on a number of topics" was incorrectly post-edited from "tokbaot" (discuss/talk about) to "tajem" (physically touch) because the glossary contained the entry "touch $\rightarrow$ tajem" without contextual information. The LLM applied this glossary entry literally without recognizing the figurative meaning in this context, degrading translation quality. Similar issues occur with idioms and expressions where literal translations from glossary entries are inappropriate.

3. **Morphological adaptation failures**: For morphologically rich languages, the LLM needs an awareness of inflectional patterns to correctly adapt glossary entries to their proper grammatical form. Because glossaries often only contain base forms (e.g. verbs in infinitive form), the LLM must apply appropriate inflectional patterns to integrate the term correctly. This issue is particularly pronounced in agglutinative languages like Rundi.