DOWNSTREAM EFFECTS OF TRANSLATION SCALE WITH LANGUAGE DIFFICULTY

Aditya V. Kulkarni **Dharmam Savani** **Ammar Ahmed Pallikonda Latheef**

Pritam Mukherjee **Jacob M. Luber*** **Paul H. Yi*,†**

*Co-senior authors †Corresponding author: paul.yi@stjude.org

Intelligent Imaging Informatics (I3), Department of Radiology
St. Jude Childrens Research Hospital, Memphis, TN, USA

ABSTRACT

Large language models (LLMs) perform best in high resource languages, motivating Machine Translation (MT) as a preprocessing step for multilingual inference. However, translation may alter task-relevant linguistic cues, degrading downstream models. It remains unclear whether such degradation is arbitrary or systematic across languages. We quantify translation-induced downstream drift using round-trip translation (English to pivot language to English) across eight pivot languages from Europe (German, Spanish, French, Italian, Portuguese) and Asia (Chinese, Hindi, Thai) while holding the source texts and downstream models fixed. Across two downstream tasks (radiology finding extraction from clinical reports and text retrieval), translation introduces performance drops that increase with language difficulty (US State Dept. categories; Spearman $|\rho| \geq 0.83$), suggesting systematic rather than random drift and providing an external, pre-translation diagnostic. Over repeated round-trip translations, performance drops early, and then stabilizes in subsequent round-trips. Semantic similarity metrics (COMET) can track this drift, providing a lightweight post-translation diagnostic for downstream drift. Our findings suggest that preprocessing non-English texts using MT may introduce systematic biases that could degrade downstream trained models and tasks, highlighting an important pitfall in the equitable multilingual use of LLMs at scale.

1 INTRODUCTION

Text data available for natural language processing (NLP) is skewed towards English and a few other high-resource languages (Ranathunga & de Silva, 2022). For example, only ≈5% of Llama 3's pretraining data is non-English (Meta AI, 2024). There is also a disparity in the availability of task-specific models, evaluation benchmarks and data resources between English and other languages (Poria & Huang, 2025; Ranathunga & de Silva, 2022). Furthermore, even multilingual models often perform better in high resource languages like English (Ahuja et al., 2023; Xuan et al., 2025). As a result, an attractive workaround for multilingual inference is to *translate to English* and then reuse English-centric foundation models and downstream tools. (Artetxe et al., 2023; Etxaniz et al., 2024). Large Language Models (LLMs) can produce high quality translations (Sizov et al., 2024), but are also known to suffer from hallucinations (Ji et al., 2023) and omissions (Asgari et al., 2025), especially in low resource languages (Guerreiro et al., 2023). In high-risk domains like healthcare, translation accuracy is critical: even subtle shifts in meaning like uncertainty or phrasing can change downstream model behavior, which can have grave consequences in the setting of medical decision-making. We therefore use clinical radiology as a high-risk application testbed for such translation-related pitfalls, and ask the question: *How much downstream model drift does translation induce, and does that drift follow systematic patterns across languages or vary arbitrarily?*

Recent work uses *round-trip translations (RTT)* (English $\rightarrow$ Pivot Language $\rightarrow$ English) as a reference-free evaluation metric for Machine Translation (MT) systems (Zhuo et al., 2023; Crone et al., 2021; Moon et al., 2020). In healthcare, RTT is less extensively studied, often using reference-based (bilingual expert back translations) methods to evaluate the quality of MT reports (Kong et al., 2025; Khoong et al., 2019). However, Mehandru et al. (2023) evaluates the effectiveness of RTT, showing that it can help clinicians surface critical translation errors associated with potential clinical harm. While these works provide valuable insights about using RTT for quality estimatation for specific language pairs, they do not study whether translation errors follow *systematic, cross-language trends* (e.g., increasing with language difficulty) or propagate to downstream AI systems such as weak labelers or foundation model inference.

Building on this work, we analyze translation through the lens of downstream robustness and structured cross-language patterns in translation-induced degradation. We use N-round RTT as a controlled perturbation of English reports and measure its propagation in weak labeling and embedding-based retrieval. Our observations suggest that translation-induced drift is systematic rather than arbitrary. First, we find that language difficulty is correlated with downstream drift (Section 3), providing an interpretable external and pre-translation predictor of downstream drift. Second, we find that drift does not compound under repeated RTT (Section 4), stagnating after an initial drop in performance. Finally, traditional MT metrics like COMET (Section 5) correlate with downstream drift, serving as a post-translation diagnostic.

2 Evaluation Framework

Pipeline schematic We study the following information flow:

$$\underbrace{x}_{\text{source text (S)}} \xrightarrow{T_{S\to P}} \underbrace{x^P}_{\text{pivot text (P)}} \xrightarrow{T_{P\to S}} \underbrace{RTT(x)}_{\text{round-trip back to S}} \xrightarrow{\text{downstream model}} \underbrace{\hat{y}}_{\text{outcome}}$$

The source report x (in source language S) is translated into a pivot language P, then translated back to S before being passed to a fixed downstream model to obtain the task output $\hat{y}$.

Round-trip translation (RTT) We define one round trip as the composition of the forward and backward translation operators:

$$RTT(x) \;=\; T_{P\to S}\big(T_{S\to P}(x)\big).$$

Starting from the original source text x_0 (in S), we apply this operator iteratively to produce a sequence of perturbed texts:

$$x_{n+1} = RTT(x_n), \qquad n = 0, 1, \ldots, N-1.$$

Here $T_{S\to P}$ and $T_{P\to S}$ denote the chosen machine-translation systems for the two directions, P is the pivot language, and N is the total number of round trips.

Reference baseline and comparison We evaluate downstream behavior on both the original report and its round-trip variants. Concretely, we treat x_0 as the reference text and compute downstream outputs

$$\hat{y}_0 = f(x_0), \qquad \hat{y}_n = f(x_n),\ n = 1, \ldots, N$$

where $f(\cdot)$ denotes the fixed downstream model for a given task. We then quantify translation-induced drift by comparing $\hat{y}_n$ against $\hat{y}_0$ (e.g., label agreement for weak labeling or retrieval consistency for embedding-based search).

Isolating the translation layer Isolating the impact of translation on downstream performance is difficult because changing languages may introduce confounders like downstream model sensitivity. For example, a retrieval model may perform better on English than on other languages and language-specific factors like style, verbosity, or clinical conventions can change task difficulty independent of translation errors. To attribute downstream changes specifically to the translation layer, we (i) keep the *original text* fixed, and (ii) evaluate all downstream tasks in a single language (English) using *one fixed downstream model* per task.

2.1 EXPERIMENTAL SETUP

We use Llama-3.3-70B-Instruct (Grattafiori et al., 2024) and Qwen2.5-72B-Instruct (Qwen et al., 2025) as MT systems, setting temperature to zero to make the translations as deterministic as possible. Model inference was served using vLLM (Kwon et al., 2023). We consider eight pivot languages: Spanish, French, Portuguese, Italian, German, Hindi, Thai, and Chinese. The English source text is taken from the MIMIC-CXR dataset (Appendix for details). We measure translation-induced effects on two downstream tasks: (i) weak labeling via the CheXbert labeler (Smit et al., 2020), and (ii) representational sensitivity via retrieval using MedGemma (Sellergren et al., 2025) embeddings. All downstream evaluations compare model outputs on x_n against the original x_0. We evaluate downstream models every 5 round trips.

Weak Label Extraction We use the CheXbert labeler to extract radiology findings from reports, a widely adopted step in radiology dataset curation and labeling. Landmark datasets such as MIMIC-CXR (Johnson et al., 2024) distribute NLP-derived labels from report text. CheXbert (Smit et al., 2020) reports near–expert-level agreement between CheXbert outputs and radiologist annotations on several common findings. We use Cohen's Kappa (κ) (Cohen, 1968) to evaluate agreement betweeen original and RTT after n rounds: $\kappa(f(x_0), f(x_n))$ (Appendix A.3.1).

Text Retrieval We use Med-Gemma to extract text embeddings from radiology reports. To test whether translation affects the embedding space, we run a retrieval experiment between the original source reports and their RTT versions across N rounds. For each translated report, we retrieve from a pool of original reports and treat its own original version as the relevant target. We quantify retrieval quality using mean reciprocal rank (MRR) (Appendix A.3.2).

2.2 DEFINING LANGUAGE DIFFICULTY

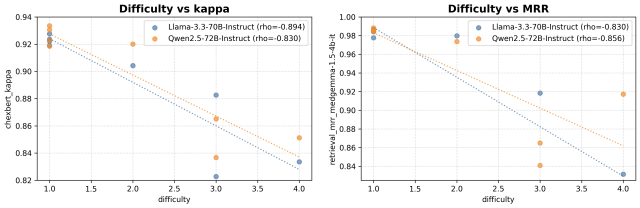
To quantify difficulty of a pivot language relative to English, we use the United States Department of State's Foreign Service's language difficulty categories (U.S Department of State), referred to here as *language difficulty (LD)*. This ordinal scale ranges from 1 (closest to English) to 4 (most difficult for native English speakers). Our pivot languages span this spectrum: Spanish, French, Portuguese, and Italian fall into LD1; German into LD2; Hindi and Thai into LD3; and Chinese into LD4.

3 DOWNSTREAM DEGRADATION IS CORRELATED WITH LANGUAGE DIFFICULTY

We test whether LD (Section 2.2) predicts translation-induced downstream degradation. We measure degradation as the performance change after RTT through a given pivot language, keeping the source language fixed to English. We assess association using Spearman's rank correlation (Spearman, 1961). We find LD is negatively correlated with downstream degradation across both tasks and models (Spearman $\rho = -0.83$ to -0.89, permutation $p < 0.03$; Fig. 1), indicating that higher difficulty languages suffer from larger translation-related degradation. Averaged across pivot languages after the first round trip, LD1–LD2 languages retain high downstream performance for both models (mean MRR: 0.97 for Llama, 0.98 for Qwen; mean κ: 0.92 and 0.93, respectively). In contrast, LD3–LD4 languages exhibit substantially lower performance (mean MRR: 0.85 for Llama, 0.87 for Qwen; mean κ: 0.85 and 0.85). These results suggest translation-induced degradation is systematic and language-dependent: downstream drift shows a strong monotonic association with LD, rather than random variation across languages. However, LD is an ordinal human-learning taxonomy and should not be interpreted as a causal driver of LLM translation behavior, rather an external, pre-translation, language-level predictor of translation-induced downstream drift.

3.1 ROLE OF LANGUAGE PROFICIENCY

In Section 3, we found that higher LD is associated with larger downstream performance drops (i.e., greater degradation), independent of the translator. However, translation quality may be both language- and translator-dependent: a translator may be more proficient in some languages than others, and this proficiency may also correlate with LD. We refer to proficiency of an LLM in a language as its performance when operating directly in that language (e.g., on benchmark tasks), independent



(a) (b)

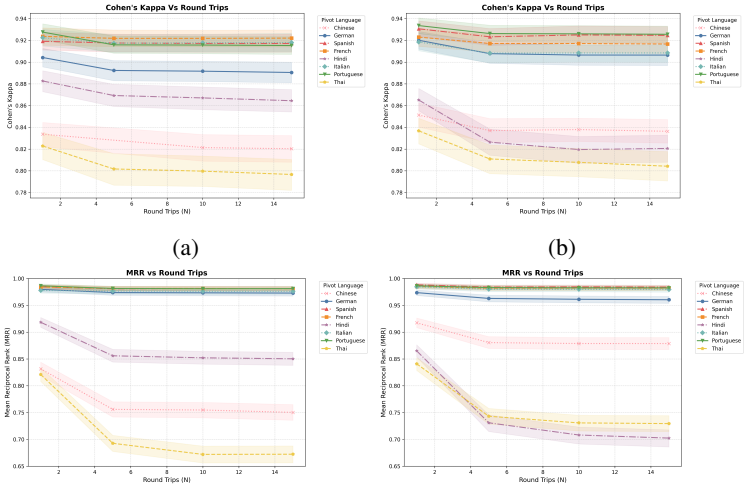
Figure 1: Language difficulty (LD) is correlated with downstream performance: LD Vs (a) Agreement on CheXbert labels (κ) and (b) Text-retrieval (MRR). Points are pivot languages.

of English. Here, we explore how translator proficiency modulates the magnitude of degradation for particular languages. For example, in our experiments, Chinese exhibits smaller degradation under Qwen than under Llama (Table 1), consistent with prior work showing Qwen is more proficient than Llama on Chinese benchmarks, including higher accuracy relative to human evaluation (Zhao et al., 2025). This matches vendor-reported language support: Qwen2.5 explicitly supports Chinese, whereas Llama3.3 does not. Considering the opposing effects of LD and proficiency: (i) language-level factors predict translation-induced drift for a fixed model, and (ii) translator proficiency can attenuate (high proficiency) or amplify (low proficiency) this drift; disentangling these mechanisms causally is left to future work.

Model	$\kappa \uparrow$ [95% CI]	MRR $\uparrow$ [95% CI]
Llama-3.3-70B-Instruct	0.83 [0.82, 0.84]	0.83 [0.82, 0.84]
Qwen2.5-72B-Instruct	**0.85** [0.84, 0.86]	**0.92** [0.91, 0.93]

Table 1: Downstream performance across translators with Chinese as the pivot language.

4 Translation-induced degradation saturates under repeated perturbation



(a) (b)

(c) (d)

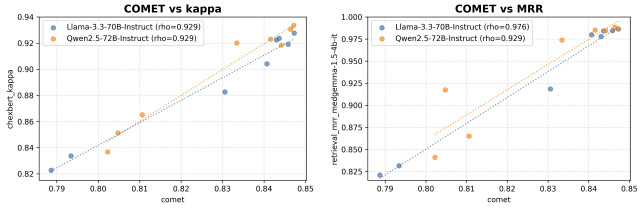
Figure 2: Downstream performance across RTT: (a) Agreement on CheXbert labels (κ) using Llama-3.3-70B-Instruct as translator, (b) Agreement on CheXbert labels (κ) using Qwen2.5-72B-Instruct, (c) Text-retrieval (MRR) using Llama-3.3-70B-Instruct, (d) Text-retrieval (MRR) using Qwen2.5-72B-Instruct. Each line represents a pivot language; shaded regions indicate 95% confidence bands.

We stress-test the translation layer by applying repeated RTT for $N = 15$ iterations, evaluating downstream tasks every five rounds. Our aim is to determine whether translation-induced drift compounds across iterations or stabilizes under sustained perturbation. Figure 2 shows that after an initial drop in performance, downstream metrics remain approximately stable across subsequent iterations.

One plausible explanation is that repeated translation acts like a standardization operator: early round trips resolve ambiguous, idiosyncratic phrasing into more verbose, canonical forms, and once text enters this less ambiguous regime, further translations introduce only limited change. The magnitude of the initial drop appears to depend on the quality of the standardized form, which in turn relates to language difficulty and translator/model proficiency (Section 3). We support this standardization hypothesis with two additional observations reported in the Appendix. First, COMET scores drop after the initial round trip and then plateau across subsequent rounds (Figure A.1), mirroring the stabilization pattern in downstream performance, which suggests that translations do not change much after the initial drop. Second, report length increases over early iterations (both in token counts and character counts) and then stabilizes (Figure A.2), consistent with translations converging to more verbose, canonical phrasing.

5 Downstream drift aligns with semantic distortion

We test whether standard MT metrics like COMET (Rei et al., 2020) detect the translation-induced drift that matters for downstream models. Figure 3 shows a strong positive correlation between COMET and downstream performance across pivot languages, suggesting that COMET captures a shared "damage axis" that drives downstream degradation. Since a high COMET score does not guarantee clinical correctness, we treat COMET as a lightweight, post-translation diagnostic for anticipating downstream risk, rather than a safety guarantee.



(a) (b)

Figure 3: COMET score is correlated with downstream performance: COMET scores Vs (a) Agreement on CheXbert labels (κ) and (b) Text-retrieval (MRR). Points are pivot languages.

6 Conclusion

Machine translation is widely used as an upstream data-layer component, yet its downstream impact is rarely quantified in a controlled setting. We find that downstream performance degradation is systematic and strongly correlates with the target language difficulty. For example, the disparity in downstream performance between languages can be as high as $\approx 17\%$ MRR for text-retrieval and $\kappa \approx 11\%$ for weak labelling. The disparity may be modulated by proficiency of the translator in the target language. Together, these results suggest that translation-induced drift is often predictable from coarse language attributes and post-translation fidelity metrics (e.g., COMET), enabling lightweight risk estimation.

Our study does not capture additional challenges that arise in deployment settings, such as shorthand and abbreviations specific to a source language, which may introduce distinct failure modes beyond those revealed by RTT. Future work should systematically characterize such dataset-level failure cases and evaluate mitigation strategies, including targeted instruction tuning and language-specific adaptation. Another direction for future work could be to benchmark translation models directly in each target language, which would help disentangle language difficulty from translator proficiency.

Our findings suggest that preprocessing non-English texts using machine translation may introduce systematic biases that could degrade downstream trained models and tasks, highlighting an important, yet insidious pitfall in this approach. These findings are particularly important for high-stakes settings like healthcare, especially when considering deployment in settings both globally and in increasingly multilingual societies, such as the USA.

REFERENCES

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. MEGA: Multilingual evaluation of generative AI. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4232–4267, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.258. URL https://aclanthology.org/2023.emnlp-main.258/.

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. Revisiting machine translation for cross-lingual classification. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6489–6499, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.399. URL https://aclanthology.org/2023.emnlp-main.399/.

Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *npj Digital Medicine*, 8(1):274, 2025.

Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

Nathan Crone, Adam Power, and John Weldon. Quality estimation using round-trip translation with sentence embeddings. *arXiv preprint arXiv:2111.00554*, 2021.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. Do multilingual language models think better in English? In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 550–564, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.46. URL https://aclanthology.org/2024.naacl-short.46/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517, 2023. doi: 10.1162/tacl_a_00615. URL https://aclanthology.org/2023.tacl-1.85/.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), March 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Alistair Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. MIMIC-CXR Database. *PhysioNet*, July 2024. doi: 10.13026/4jqj-jw95. URL https://doi.org/10.13026/4jqj-jw95. Version 2.1.0.

Elaine C Khoong, Eric Steinbrook, Cortlyn Brown, and Alicia Fernandez. Assessing the use of google translate for spanish and chinese translations of emergency department discharge instructions. *JAMA internal medicine*, 179(4):580–582, 2019.

Marianna Kong, Alicia Fernandez, Jaskaran Bains, Ana Milisavljevic, Katherine C Brooks, Akash Shanmugam, Leslie Avilez, Junhong Li, Vladyslav Honcharov, Andersen Yang, et al. Evaluation of the accuracy and safety of machine translation of patient-specific discharge instructions: a comparative analysis. *BMJ quality & safety*, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Nikita Mehandru, Sweta Agrawal, Yimin Xiao, Ge Gao, Elaine Khoong, Marine Carpuat, and Niloufar Salehi. Physician detection of clinical harm in machine translation: Quality estimation aids in reliance and backtranslation identifies critical errors. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11633–11647, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.712. URL https://aclanthology.org/2023.emnlp-main.712/.

Meta AI. Introducing meta llama 3: The most capable openly available llm to date. https://ai.meta.com/blog/meta-llama-3/, 2024.

Jihyung Moon, Hyunchang Cho, and Eunjeong L. Park. Revisiting round-trip translation for quality estimation. In André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada (eds.), *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pp. 91–104, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL https://aclanthology.org/2020.eamt-1.11/.

Sampoorna Poria and Xiaolei Huang. Bhaasha, bhāṣā, zaban: A survey for low-resourced languages in South Asia – current stage and challenges. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 1386–1406, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.73. URL https://aclanthology.org/2025.findings-emnlp.73/.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Surangika Ranathunga and Nisansa de Silva. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (eds.), *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 823–848, Online only, November 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.aacl-main.62. URL https://aclanthology.org/2022.aacl-main.62/.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213/.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Fedor Sizov, Cristina España-Bonet, Josef Van Genabith, Roy Xie, and Koel Dutta Chowdhury. Analysing translation artifacts: A comparative study of LLMs, NMTs, and human translations. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 1183–1199, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.116. URL https://aclanthology.org/2024.wmt-1.116/.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert, 2020. URL https://arxiv.org/abs/2004.09167.

Charles Spearman. The proof and measurement of association between two things. 1961.

Foreign Services U.S Department of State. Foreign language training. URL https://www.state.gov/national-foreign-affairs-training-center/foreign-language-training.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, et al. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. *arXiv preprint arXiv:2503.10497*, 2025.

Chenzhuo Zhao, Xinda Wang, Yue Huang, Junting Lu, and Ziqian Liu. Tase: Token awareness and structured evaluation for multilingual language models. *arXiv preprint arXiv:2508.05468*, 2025.

Terry Yue Zhuo, Qiongkai Xu, Xuanli He, and Trevor Cohn. Rethinking round-trip translation for machine translation evaluation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 319–337, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.22. URL https://aclanthology.org/2023.findings-acl.22/.

A Appendix

A.1 Data

We used radiology reports from the MIMIC-CXR dataset and restricted experiments to the test and validate splits. To focus on clinically informative reports, we excluded studies labeled as No Finding in the provided CheXpert-derived labels and removed reports with missing or empty impression sections. We further removed repeated report text throughout the dataset to ensure unique reports. The resulting dataset (N=2,683 radiology reports) consists of unique studies with non-empty impressions and at least one positive finding label.

A.2 Translation

A.2.1 Params

We served the LLM models with vLLM (Kwon et al., 2023) via an HTTP API and used deterministic decoding (temperature = 0.0), making outputs reproducible for a fixed prompt and input.

A.2.2 Translation Prompts

All translations use the following identical instruction template with language placeholders; outputs are restricted to the translated report content only, excluding any introductory text or explanation.

```
You are an expert AI radiologist fluent in {source_language}
and {target_language}.

Translate the following {source_language} Chest X-Ray report
into {target_language}.

Provide ONLY the {target_language} translation, with no
introductory text or explanation.
```

A.3 Evaluation Metrics

A.3.1 Cohen's Kappa

We used Cohen's Kappa score κ (`sklearn.metrics.cohen_kappa_score`) to evaluate agreement between original CheXbert predictions and round-trip CheXbert predictions. Confidence intervals were computed using non-parametric bootstrapping (2,000 samples) κ is defined as:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

Where Pr(a) represents the actual observed agreement, and Pr(e) represents chance agreement

A.3.2 Mean Reciprocal Rank

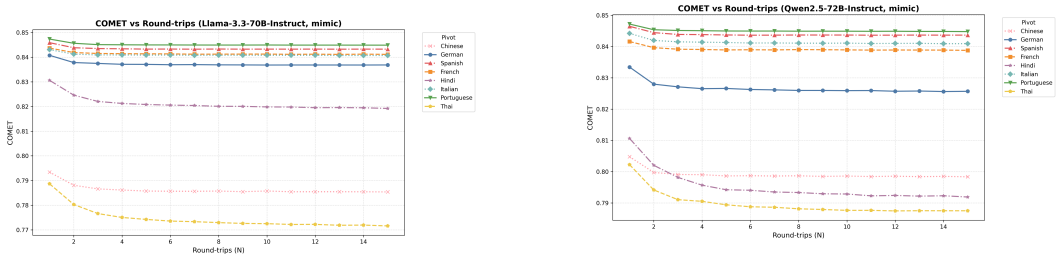
We used Mean Reciprocal Rank (MRR) as our retrieval metric. Confidence intervals were computed using non-parametric bootstrapping (2,000 samples). MRR is defined as:

$$MRR = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{rank_i}$$

A.3.3 Spearman's Rank Correlation

We used Spearman's Rank Correlation ρ (`scipy.stats.spearmanr`) to evaluate monotonic correlation between difficulty and downstream metrics (Cohen's Kappa, MRR). We assessed the significance of the Spearman rank correlations using a permutation test (`from scipy.stats.permutation_test`) to obtain p values and computed confidence intervals via nonparametric bootstrap resampling (2,000 resamples). We used the same setup for evaluating association between COMET scores and downstream metrics.

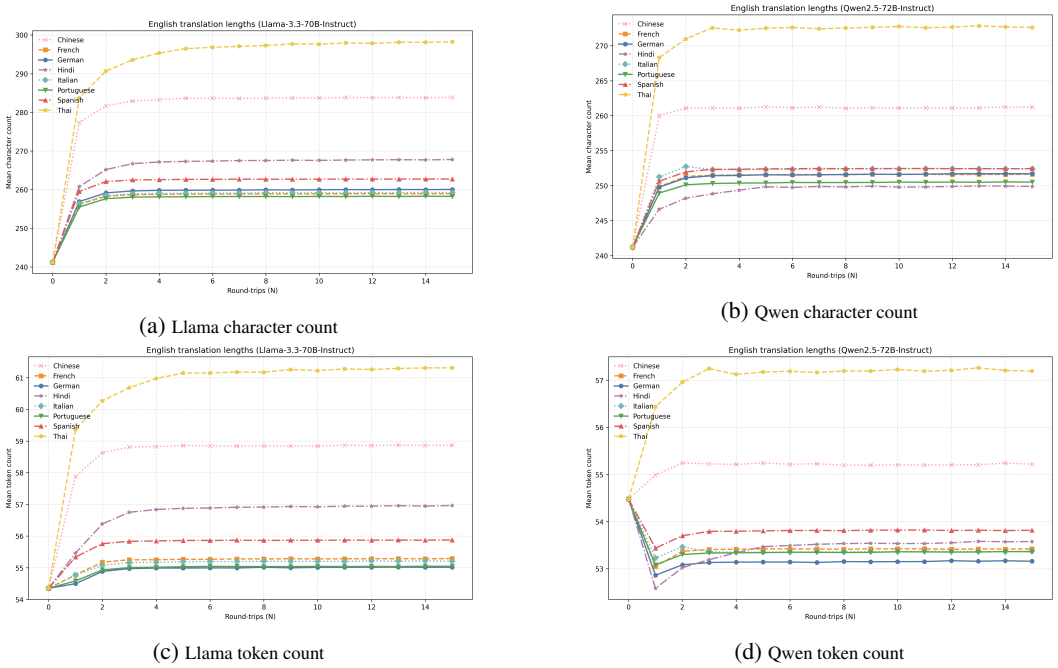
A.4 Character and Token Lengths across N rounds



(a) Llama COMET

(b) Qwen COMET

Figure A.1: COMET score across round-trip translation iterations: (a) using Llama as translator, (b) using Qwen. Each line represents a pivot language.



(a) Llama character count

(b) Qwen character count

(c) Llama token count

(d) Qwen token count

Figure A.2: Mean report length across round-trip translation iterations: (a) character count using Llama as translator, (b) character count using Qwen, (c) token count using Llama, (d) token count using Qwen. Each line represents a pivot language.