When Flores Bloomz Wrong: Cross-Direction Contamination in Machine Translation Evaluation

David Tan[1], Pinzhen Chen[2,4], Josef van Genabith[1,3], Koel Dutta Chowdhury[1]
[1]Saarland University, Saarland Informatics Campus [2]Queen's University Belfast
[3]German Research Center for Artificial Intelligence (DFKI) [4]University of Edinburgh
data00003@stud.uni-saarland.de

Abstract

Large language models (LLMs) can be benchmark-contaminated, resulting in inflated scores that mask memorization as generalization, and in multilingual settings, this memorization can even transfer to "uncontaminated" languages. Using the FLORES-200 translation benchmark as a diagnostic, we study two 7–8B instruction-tuned multilingual LLMs: Bloomz, which was trained on FLORES, and Llama as an uncontaminated control. We confirm Bloomz's FLORES contamination and demonstrate that machine translation contamination can be cross-directional, artificially boosting performance in unseen translation directions due to target-side memorization. Further analysis shows that recall of memorized references often persists despite various source-side perturbation efforts like paraphrasing and named entity replacement. However, replacing named entities leads to a consistent decrease in BLEU, suggesting an effective probing method for memorization in contaminated models.

1 Introduction

Large language models (LLMs) typically go through pre-training and fine-tuning stages to become applicable to downstream tasks, where data is a fundamental building block. Despite this, details about training data are often too vague to infer useful information, both for open- and closed-source models. This has negative implications for LLM research, one of which is the reliability of using public benchmarks to evaluate model performance. LLM training data is typically in the magnitude of billions to trillions of tokens (Gao et al., 2020; Hoffmann et al., 2022). Consequently, public test sets may inadvertently be incorporated into training data (Sainz et al., 2024) and artificially boost scores—a phenomenon referred to as *data contamination* (Magar and Schwartz, 2022). The problem is even more pressing in multilingual evaluation, where contamination can cross-lingually transfer into languages not seen before (Yao et al., 2024).

This work uses machine translation as a diagnostic task to investigate cross-direction data contamination: the artificial boost of unseen language pair scores, resulting from training on language pairs with the same target. We use FLORES200 (NLLB et al., 2022), a test suite that supports hundreds of languages with multiway parallelism, yet this wide coverage increases the likelihood of contamination. We explore whether its multiway parallel structure may result in cross-direction transfer of train-test leakage. We benchmark Bloomz-7B1 (Muennighoff et al., 2023), a multilingual LLM with FLORES documented as part of its fine-tuning data, and compare it to Llama-3.1-8B-Instruct (Grattafiori et al., 2024) which is reportedly clean. Compared to Kocyigit et al. (2025)'s focus on the contribution of source or target-side text during training, we probe cross-direction contamination in machine translation at inference, specifically whether altered source inputs can still trigger memorized targets. We analyze patterns of contamination through controlled studies and design perturbation-based tests to probe how contamination manifests. We present three findings:

1. We show that contamination happens cross-directionally for machine translation, where contaminated LLMs can perform moderately well in unseen translation directions, mainly due to target-side language memorization.
2. While it is assumed that train-test source input similarity leads to patterns of contamination, we demonstrate that a not-so-similar input may still lead to a model calling memorized text.
3. We discover that replacing named entities consistently decreases BLEU, suggesting an effective method for probing memorization.

Code at github.com/Mr-Ao-25/cross-ling-contamination.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 2: Short Papers, pages 345–358
March 24-29, 2026 ©2026 Association for Computational Linguistics

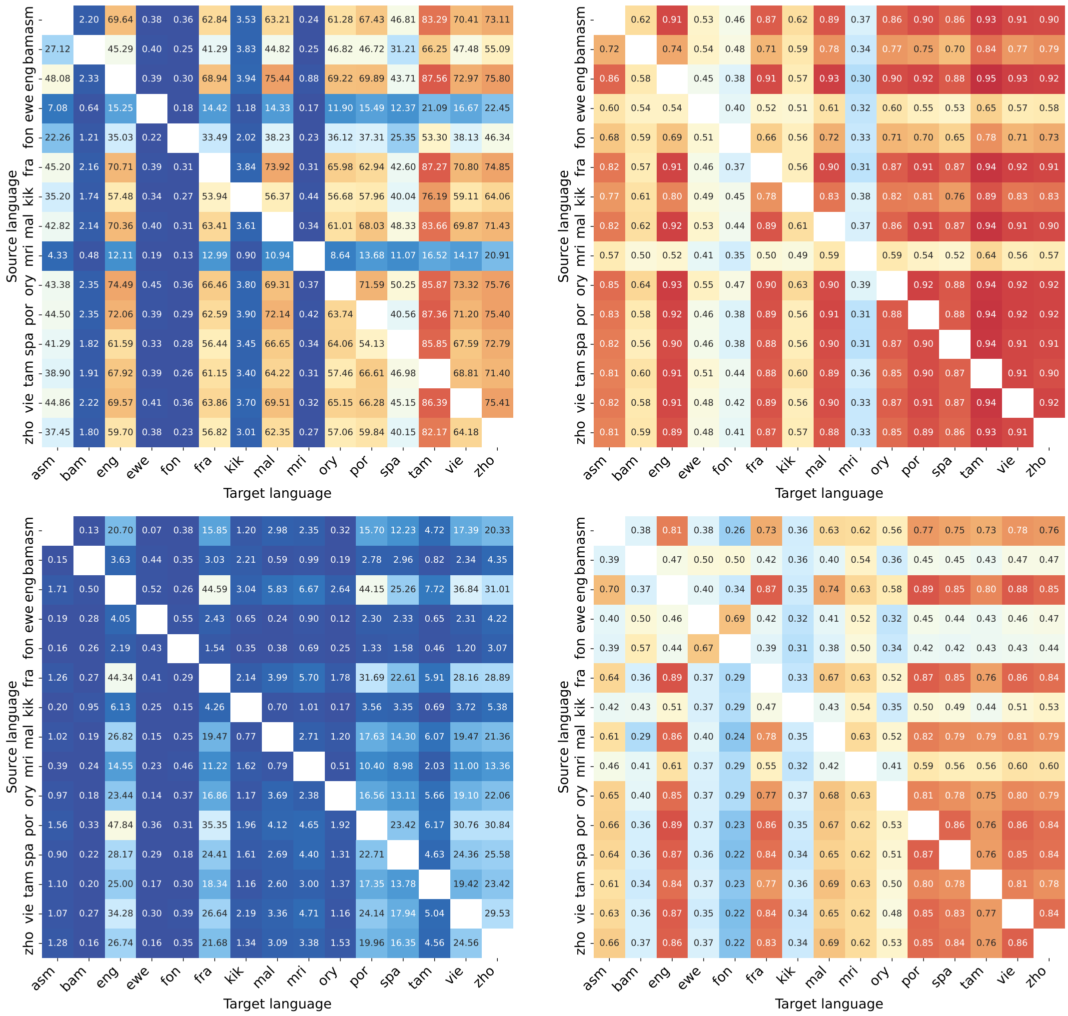

Figure 1: BLEU for Bloomz (upper left), Llama (lower left) and COMET for Bloomz (upper right), Llama (lower right). All plots have the same sources, targets, and color bar with a blue-red gradient (low-to-high).

2 Experimental Setup

Test set and Models We use FLORES200 (NLLB et al., 2022, or *FLORES*) as our test set for its wide language coverage and multi-way parallelism. We use the dev set of 997 samples and evaluate the model's initial generation for each sample. We select 15 languages spanning different resource levels—high-resourced: English (`eng`), Simplified Chinese (`zho`), Spanish (`spa`), Portuguese (`por`), French (`fra`), and Vietnamese (`vie`); medium-resourced: Malayalam (`mal`), Tamil (`tam`), Assamese (`asm`), and Odia (`ory`); low-resourced: Bambara (`bam`), Fon (`fon`), Ewe (`ewe`), Kikuyu (`kik`), and Maori (`mri`). Notably, Bloomz has not been trained on `ewe` and `mri`.

We investigate various aspects of contamination using `bloomz-7b1` (Muennighoff et al., 2023, *Bloomz*) and `Llama-3.1-8B-Instruct` (Grattafiori et al., 2024, *Llama*). Both models are instruction-tuned, similarly sized (7–8B parameters), and designed for multilingual text generation, making them broadly comparable in their cross-lingual capabilities. While Bloomz was explicitly trained on FLORES with unspecified translation directions, there is no reported use of FLORES in the training of Llama. For entity replacement experiments, we use `aya-expanse-8b` (Dang et al., 2024, *Aya*).

3 Contamination in Machine Translation

3.1 Is there contamination?

We first investigate whether Bloomz and Llama have memorized FLORES. We run an evaluation on FLORES for every source-target pair formed by our 15 chosen languages. We use BLEU implemented in sacrebleu (Post, 2018; Papineni et al., 2002) as a measure of surface form overlap w.r.t. the reference, rather than translation quality, and

COMET-22-DA (Rei et al., 2022) for semantic similarity. As our diagnostic, we take that an unreasonably high BLEU coupled with high COMET (e.g. 80 BLEU, 0.9 COMET) would imply contamination, while a reasonable BLEU with a high COMET score (e.g. 40 BLEU, 0.9 COMET) would indicate a good performance without contamination (or with less likelihood of it). Low BLEU and COMET indicate divergence in both form and meaning; moderate BLEU with high COMET reflects semantic similarity despite surface variation.

Figure 1 presents BLEU and COMET heatmaps for Bloomz and Llama. For Bloomz, we see oddly high BLEU (40–90) in `xxx`→{`eng`, `fra`, `mal`, `ory`, `por`, `tam`, `vie`, `zho`}, hinting at contamination; medium BLEU scores (≤ 50) in `xxx`→{`asm`, `spa`}; and very low scores (≈ 0) in `xxx`→{`bam`, `ewe`, `fon`, `kik`, `mri`}. COMET scores generally follow trends seen in BLEU; language pairs with medium to high BLEU achieve > 0.8 COMET; low BLEU language pairs achieve $\leq 0.5 – 0.65$ COMET.

In contrast, Llama shows little evidence of contamination: reasonably high BLEU scores (30–40) only appear in a few high-resource languages or directions, and COMET scores are only high (≥ 0.8) when BLEU scores are >15. For Bloomz, high COMET (≥ 0.8) results from ≥ 35 BLEU.

Memorization not generalization To rule out the possibility that Bloomz simply has superb performance in those directions with high BLEU, we test the relatively low-resource `eng`→{`mal`, `ory`, `tam`} available in PMIndia (Haddow and Kirefu, 2020) and Mann-ki-Baat (Jain et al., 2024). Section B Table 2 shows near-zero BLEU scores for both datasets, indicating that Bloomz cannot translate those directions, adding further evidence of FLORES contamination.

Disentangling source and target We study the pattern of contamination when a language is placed on the source or target side. Figure 1 shows that when placed on the target side, a handful of languages have a "clean" column: `bam`, `ewe`, `fon`, `kik`, and `mri`, with BLEU ≈ 0. In contrast, when placed on the source side, a few languages that did not exhibit contamination signals as the target (`bam`, `fon`, `kik`) now have medium-range BLEU scores across their rows. Overall, no single language as a source can escape contamination: even `ewe` and `mri` that are not supported by Bloomz achieve >10 BLEU into a few target languages, higher than Llama. This suggests that contamination manifests itself

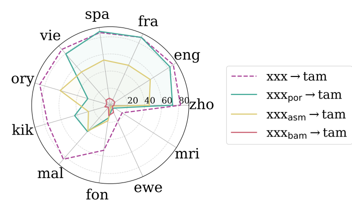

Figure 2: Bloomz's BLEU for {$\texttt{xxx}_{\texttt{por}}$, $\texttt{xxx}_{\texttt{bam}}$, $\texttt{xxx}_{\texttt{asm}}$}→`tam`. Llama back-translated {`por`, `bam`, `asm`} into $\texttt{xxx}_{\{\texttt{por, bam, asm}\}}$.

asymmetrically in terms of whether a language is at the source or target side: inflated scores are due to the memorization of the target language text.

3.2 Is the recall due to the exact source input?

We investigate whether Bloomz' recall of the memorized target is due to seeing the *exact* source string paired with the target. For all `xxx`→`yyy` translation directions, we utilize a back-translated source, $\texttt{xxx}_{\texttt{zzz}}$, originating from a third language `zzz`; this is essentially done by Llama in `zzz`→`xxx` directions in Section 3.1. FLORES' multiway parallelism ensures that the back-translated source remains parallel to the original target reference, while staying different from the original source.

We try three originating languages `zzz`={`asm`, `bam`, `por`}, for which Llama had distinct translation performance (into any `xxx`): $\texttt{xxx}_{\texttt{bam}}$ are near-zero BLEU, $\texttt{xxx}_{\texttt{asm}}$ are in the 10–20 range, and $\texttt{xxx}_{\texttt{por}}$ are in the 20–50 range. These scores are shown in Llama's heatmap in Figure 1. Specifically, we experiment with all chosen languages, except {`por`, `bam`, `asm`}, as source `xxx`. The target language is fixed to `tam`, which exhibited the highest degree of target-side memorization. This forms three test directions: {$\texttt{xxx}_{\texttt{por}}$, $\texttt{xxx}_{\texttt{asm}}$, $\texttt{xxx}_{\texttt{bam}}$}→`tam`.

We also utilize a paraphrased source generated by Llama denoted as $\texttt{xxx}_{\text{pp}}$ to compare against the Llama back-translated $\texttt{xxx}_{\text{zzz}}$. We test high-resource source languages into `tam`: $\texttt{xxx}_{\text{pp}} \rightarrow \texttt{tam}$ and compare it against $\texttt{xxx}_{\text{por}} \rightarrow \texttt{tam}$. We also assess the similarity between the altered sources and the original source denoted as $\texttt{xxx}_{\text{og}}$: $\{\texttt{xxx}_{\text{por}}, \texttt{xxx}_{\text{pp}}\} \rightarrow \texttt{xxx}_{\text{og}}$.

Results Figure 2 presents Bloomz' performance in translating back-translated source. Generally, the BLEU scores of $\texttt{xxx}_{\text{src}}$→`tam` align with the performance of Llama when producing the back-translations ($\texttt{xxx}_{\text{por}}$, $\texttt{xxx}_{\text{bam}}$, $\texttt{xxx}_{\text{asm}}$). When Llama's back-translation is in the 20–50 BLEU range, Bloomz' translation has a high

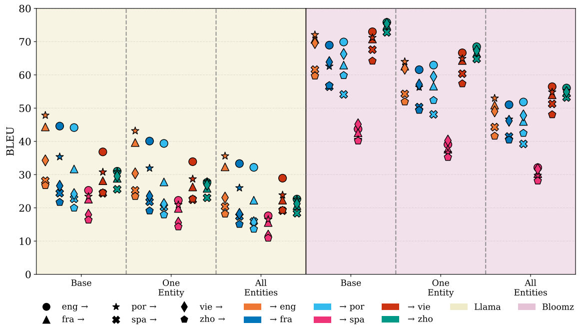

Figure 3: BLEU scores for Llama (left) and Bloomz (right) across language pairs under different entity replacement settings (Base, One Entity, All Entities).

	eng	fra	spa	vie	zho
$\texttt{xxx}_{\text{por}} \rightarrow \texttt{tam}$	82.92	87.28	87.24	79.38	72.81
$\texttt{xxx}_{\text{pp}} \rightarrow \texttt{tam}$	64.60	66.55	61.15	62.18	53.67
$\texttt{xxx}_{\text{por}} \rightarrow \texttt{xxx}_{\text{og}}$	47.83	35.35	23.41	30.75	30.84
$\texttt{xxx}_{\text{pp}} \rightarrow \texttt{xxx}_{\text{og}}$	53.03	60.92	54.38	62.60	61.11

Table 1: $\{\texttt{xxx}_{\text{por}}, \texttt{xxx}_{\text{pp}}\} \rightarrow \texttt{tam}$ evaluated by Bloomz and the perturbed sources compared against the original source: $\{\texttt{xxx}_{\text{por}}, \texttt{xxx}_{\text{pp}}\} \rightarrow \texttt{xxx}_{\text{og}}$

>60 BLEU, as exhibited in $\{$vie, spa, fra, eng, zho$\}_{\text{por}} \rightarrow$tam. When the string overlap between the back-translated and original sources is lower (10–20 BLEU), Bloomz gets a moderate 45–50 BLEU as exhibited in $\{$vie, spa, fra, eng, zho$\}_{\text{asm}} \rightarrow$tam.

Unexpected results are observed in a few back-translated sources with close to 0 BLEU (i.e. $\{$ory, kik, mal$\}_{\{\text{por, asm}\}}$). Given these back-translated sources, Bloomz still yield 20–60 BLEU, indicating moderate to high recall of memorized targets. In other cases, where the back-translated source has poor BLEU w.r.t. the original source, memorized target recall is minimal, as shown in other low-resource sources from por or asm, as well as in all $\texttt{xxx}_{\text{bam}} \rightarrow$tam.

Table 1 presents the paraphrased scores compared against the back-translated scores. Despite the paraphrased sources having higher surface form overlap with the original, the back-translated sources score 20–30 BLEU higher for all pairs. Our results suggest that the exact source is the culprit, but even a not-so-similar back-translated or paraphrased input does not guarantee a low recall, indicating that some aspect of the source functions as a "trigger" for recall.

3.3 What features contribute to the recall?

We investigate the curious case above, where back-translated sources still lead to recall of memorized target. First, we manually inspect Llama's back-translations (refer to Table 4 in section G) and observe that even when they diverge from the original sources significantly, named entities are still retained. To investigate whether named entities contribute to recall, we create new entities to replace those in the original using six high resource languages (eng, fra, por, spa, vie, zho) and observe how the recall of memorized targets varies, using Bloomz' and Llama's BLEU scores.

To ensure that the same named entities are identified and labeled for every language, we use spaCy (Honnibal et al., 2020) to first identify named entities for every sentence and manually edit any missing or incorrectly labeled entities. Given the English source and its corresponding list of entities, we use Aya to create new entities to replace the original and translate the new entities into every high resource language. For entity types, descriptions, and examples of entity-replaced sentences, refer to Table 5 and Table 6. In total, FLORES contains around 640 sentences with entities for our chosen languages.

We experiment with two replacement settings, one-entity and all-entities, and compare against the original for each language pair. For the one-entity setting, sentences with multiple entities result in multiple variants, from which one is chosen randomly. The all-entities setting replaces every entity in a sentence. We then score BLEU for each language pair against the original reference to measure recall of memorized text.

Results Figure 3 shows the changes in BLEU for both models when entity replacement is applied. Generally, results indicate that replacing entities in source texts are associated with a reduced level of recall. Bloomz results in a reduction of 5-10 BLEU for single entity replacement and 10-20 BLEU for all entity replacement. Llama experiences an average decrease of about 5 BLEU for the one-entity setting and about 10 for the all-entities setting, with certain pairs decreasing by up to 15 BLEU. For each target language, the relative ordering of scores across language pairs is largely preserved, confirming that target-side memorization remains the primary factor. However, the consistent BLEU drop makes entity replacement a potentially effective method for probing memorization in parallel datasets like FLORES.

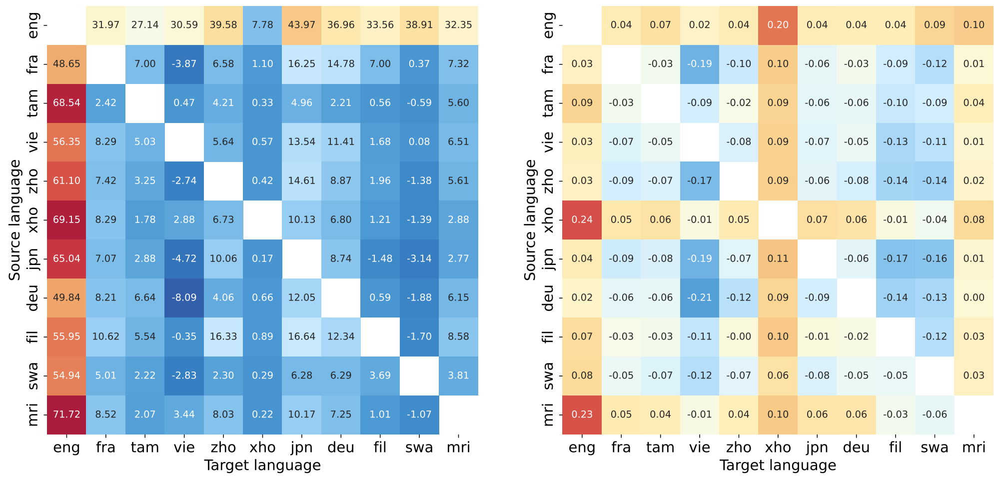

Figure 4: The BLEU (left) and COMET (right) difference between the finetuned and base Llama model. Positive/negative indicates increases/decreases in the fine-tuned model compared to the base model.

4 Is contamination occurring cross-directionally?

To investigate whether the multilingual nature of FLORES inadvertently causes cross-direction contamination during training, we perform fine-tuning experiments with FLORES `eng` $\leftrightarrow$ `xxx` data in 11 languages {`eng`, `fra`, `tam`, `vie`, `mri`, `zho`, Xhosa (`xho`), Japanese (`jpn`), German (`deu`), Tagalog (`fil`), Swahili (`swa`)} using Axolotl.[1] For the given languages, `xxx`, `yyy`, and `zzz`, we define cross-direction contamination as the artificial boost in scores of unseen translation directions (`zzz` $\rightarrow$ `yyy`) due to the memorization of `yyy` in seen translation directions (`xxx` $\rightarrow$ `yyy`). Although exact translation directions used for model training are rarely specified, it is reasonable to assume English is used as both source and target.

We fine-tune Llama until it severely memorizes the `eng` $\leftrightarrow$ `xxx` data to mirror the level of memorization observed in Bloomz (refer to Table 3 for fine-tuning parameters). Using the fine-tuned model, we translate every source-target pair for the 11 languages and evaluate using BLEU and COMET.

Results Figure 4 presents the BLEU and COMET differences between the fine-tuned and base models (refer to Figure 5 for full scores). In general, BLEU scores for most unseen language pairs increased up to 16 points, the exceptions being `xxx` $\rightarrow$ {`vie`, `swa`}. As for `eng` $\leftrightarrow$ `xxx`, the gains in `xxx` $\rightarrow$ `eng` overshadow the gains in `eng` $\rightarrow$ `xxx` reinforcing target side memorization.

Concerning source and target memorization, we observe trends mirroring Bloomz. Although `xho` yielded low BLEU scores as a target, its effectiveness as a source was similar to other languages. Likewise, `xxx` $\rightarrow$ {`vie`, `swa`} showed decreases with most sources, but performed similarly to other languages when used as sources, reinforcing that contamination is due to target-side memorization.

Conversely, the trends in COMET contradict the overall increases in BLEU. Outside of `xxx` $\rightarrow$ {`mri`, `xho`}, COMET scores decreased for all pairs. We conjecture this is due to the extreme memorization of some samples, leading to increased overall BLEU scores at the cost of the non-memorized samples. Regarding the exceptions {`mri`, `xho`}, we surmise that they could not deteriorate further, as the BLEU and COMET in the base model already indicated little similarity with the reference.

5 Conclusion

We verified that Bloomz is FLORES-contaminated, using Llama as an uncontaminated control, and demonstrate through fine-tuning Llama with FLORES `eng`$\leftrightarrow$ `xxx` pairs that contamination is cross-directional, driven by target-side memorization. We further show that back-translated and paraphrased sources do not necessarily reduce the recall of memorized references. We found that replacing named entities leads to a consistent decrease in BLEU, suggesting an effective method for probing memorization in potentially contaminated models. We recommend that practitioners verify contamination across translation directions when evaluating on multiway parallel benchmarks.

[1] github.com/axolotl-ai-cloud/axolotl

Acknowledgment

We thank Barry Haddow for his comment on paraphrasing the Flores input. The work received funding from Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1102 Information Density and Linguistic Encoding.

6 Limitations

Although we gave Aya the source, source entities, and target during entity generation, it is possible that the new entities have different inflection compared to the original entities, changing the sentence more than intended. The drop in BLEU for entity replacement may differ depending on the number of sentences that contain entities. Also, our work yielded empirical conclusions but did not investigate the internal mechanisms of cross-direction contamination or the activation of memorized text. Our experiments are limited to 7-8B parameter models; results may vary for different model sizes. Additionally, our fine-tuning experiments simulate contamination but may not fully replicate contamination that occurs during pre-training.

7 Ethical Considerations

This study sought to uncover various forms of contamination in machine translation tasks. Our work poses no risk to readers, practitioners, or the wider community. All model and data artifacts were used in full compliance with their respective licenses.

References

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Koel Dutta Chowdhury, Cristina España-Bonet, and Josef van Genabith. 2020. Understanding translationese in multi-view embedding spaces. In *Proceedings of the 28th international conference on computational linguistics*, pages 6056–6062.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, and others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*.

Koel Dutta Chowdhury, Cristina España-Bonet, and Josef van Genabith. 2021. Tracing source language interference in translation with graph-isomorphism measures. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 375–385, Held Online. INCOMA Ltd.

Aparna Elangovan, Jiayuan He, and Karin Verspoor. 2021. Memorization vs. generalization : Quantifying data leakage in NLP performance evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, and others. 2020. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

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

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*.

Barry Haddow and Faheem Kirefu. 2020. Pmindia–a collection of parallel corpora of languages of india. *arXiv preprint arXiv:2001.09907*.

Valentin Hartmann, Anshuman Suri, Vincent Bindschaedler, David Evans, Shruti Tople, and Robert West. 2023. Sok: Memorization in general-purpose large language models. *arXiv preprint arXiv:2310.18362*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and others. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Sparsh Jain, Ashwin Sankar, Devilal Choudhary, Dhairya Suman, Nikhil Narasimhan, Mohammed Safi Ur Rahman Khan, Anoop Kunchukuttan, Mitesh M Khapra, and Raj Dabre. 2024. Bhasaanuvaad: A speech translation dataset for 14 indian languages. *arXiv preprint arXiv: 2411.04699*.

Minhao Jiang, Ken Ziyu Liu, Ming Zhong, Rylan Schaeffer, Siru Ouyang, Jiawei Han, and Sanmi Koyejo. 2024. Investigating data contamination for pre-training language models. *arXiv preprint arXiv:2401.06059*.

Muhammed Yusuf Kocyigit, Eleftheria Briakou, Daniel Deutsch, Jiaming Luo, Colin Cherry, and Markus Freitag. 2025. Overestimation in LLM evaluation: A controlled large-scale study on data contamination's impact on machine translation. In *Forty-second International Conference on Machine Learning*.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. *arXiv preprint arXiv:2203.08242*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, and others. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Team NLLB, Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, and others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Yonatan Oren, Nicole Meister, Niladri S Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. 2023. Proving test set contamination in black-box language models. In *The Twelfth International Conference on Learning Representations*.

Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. 2020. Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy (SP)*.

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*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*.

Vikas Raunak and Arul Menezes. 2022. Finding memo: Extractive memorization in constrained sequence generation tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2022*.

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*.

Vikas Raunak, Matt Post, and Arul Menezes. 2022. Salted: A framework for salient long-tail translation error detection. *arXiv preprint arXiv:2205.09988*.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation*.

Iker Sainz, Oscar aFnd García-Ferrero, Alon Jacovi, Jon Ander Campos, Yanai Elazar, Eneko Agirre, Yoav Goldberg, Wei-Lin Chen, Jenny Chim, Leshem Choshen, Luca D'Amico-Wong, and others. 2024. Data contamination report from the 2024 CONDA shared task. In *Proceedings of the 1st Workshop on Data Contamination (CONDA)*.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E Gonzalez, and Ion Stoica. 2023. Rethinking benchmark and contamination for language models with rephrased samples. *arXiv preprint arXiv:2311.04850*.

Feng Yao, Yufan Zhuang, Zihao Sun, Sunan Xu, Animesh Kumar, and Jingbo Shang. 2024. Data contamination can cross language barriers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*.

A Related Work

Recent studies highlighted the severe implications of data contamination and called for increased scrutiny of this issue (Pan et al., 2020; Zhou et al., 2023; Jacovi et al., 2023; Dodge et al., 2021). LLMs exhibit substantial memorization capabilities (Elangovan et al., 2021; Hartmann et al., 2023; Carlini et al., 2022), raising concerns that benchmark gains may reflect leakage rather than generalization. To address this, prior work proposed both post hoc detection (Shi et al., 2024; Oren et al., 2023) and controlled experiments where models are deliberately trained with test data to measure their effects (Jiang et al., 2024; Yang et al., 2023; Magar and Schwartz, 2022).

In machine translation (MT), memorization and contamination manifest differently. Raunak et al. (2021) documented that hallucinations often arise from memorized artifacts. Raunak and Menezes (2022) showed the phenomenon of extractive memorization, where models reproduce target segments after partial source exposure. Raunak et al. (2022) linked memorizing rare patterns to long-tail translation errors. Guerreiro et al. (2023) provided a systematic analysis of hallucinations, showing how memorization failures scale across multilingual settings. Chowdhury et al. (2020); Dutta Chowdhury et al. (2021) demonstrated that translation artifacts manifest as detectable patterns in embedding spaces. More recently, Kocyigit et al. (2025) conducted controlled pre-training and found that accidental inclusion of evaluation examples in pre-training corpora critically undermines validity. Taken together, prior research has focused on contamination and memorization in monolingual or bilingual MT. To our knowledge, our work is the first to systematically examine cross-lingual contamination in MT, demonstrating that memorization in the target language can artificially boost performance in translating many source languages.

B PMIndia and Maan-ki-Baat BLEU Scores

Test set	eng→mal	eng→ory	eng→tam
FLORES	75.44	69.21	87.55
PMIndia	0.12	1.85	0.11
Mann-ki-Baat	0.26	1.01	0.82

Table 2: Bloomz's BLEU for eng→{mal, ory, tam}.

C Fine-tuning Parameters

```
base_model: meta-llama/Meta-Llama-3.1-8B-Instruct
model_config:
    attention_dropout: 0.1
sequence_len: 512
sample_packing: true
pad_to_sequence_len: true
num_epochs: 3
micro_batch_size: 4
gradient_accumulation_steps: 8
learning_rate: 5e-6
weight_decay: 0.05
warmup_ratio: 0.02
lr_scheduler: cosine
max_grad_norm: 0.25
optim: adamw_torch
deepspeed: deepspeed_configs/zero3_bf16.json
torch_distributed_type: DEEPSPEED
```

Table 3: Fine-tuning Parameters

D Computing Infrastructure

All experiments were conducted on a single NVIDIA A100 GPU (80GB). Total compute time: approximately 30 GPU-hours. Translation inference with Bloomz and Llama, as well as fine-tuning Llama were performed on this setup.

E Use of AI Assistants

AI assistants were used for editing prose and debugging code. All scientific claims, experimental design, and analysis were conducted by the authors.

F Fine-tuned and Base Llama BLEU and COMET Heatmaps

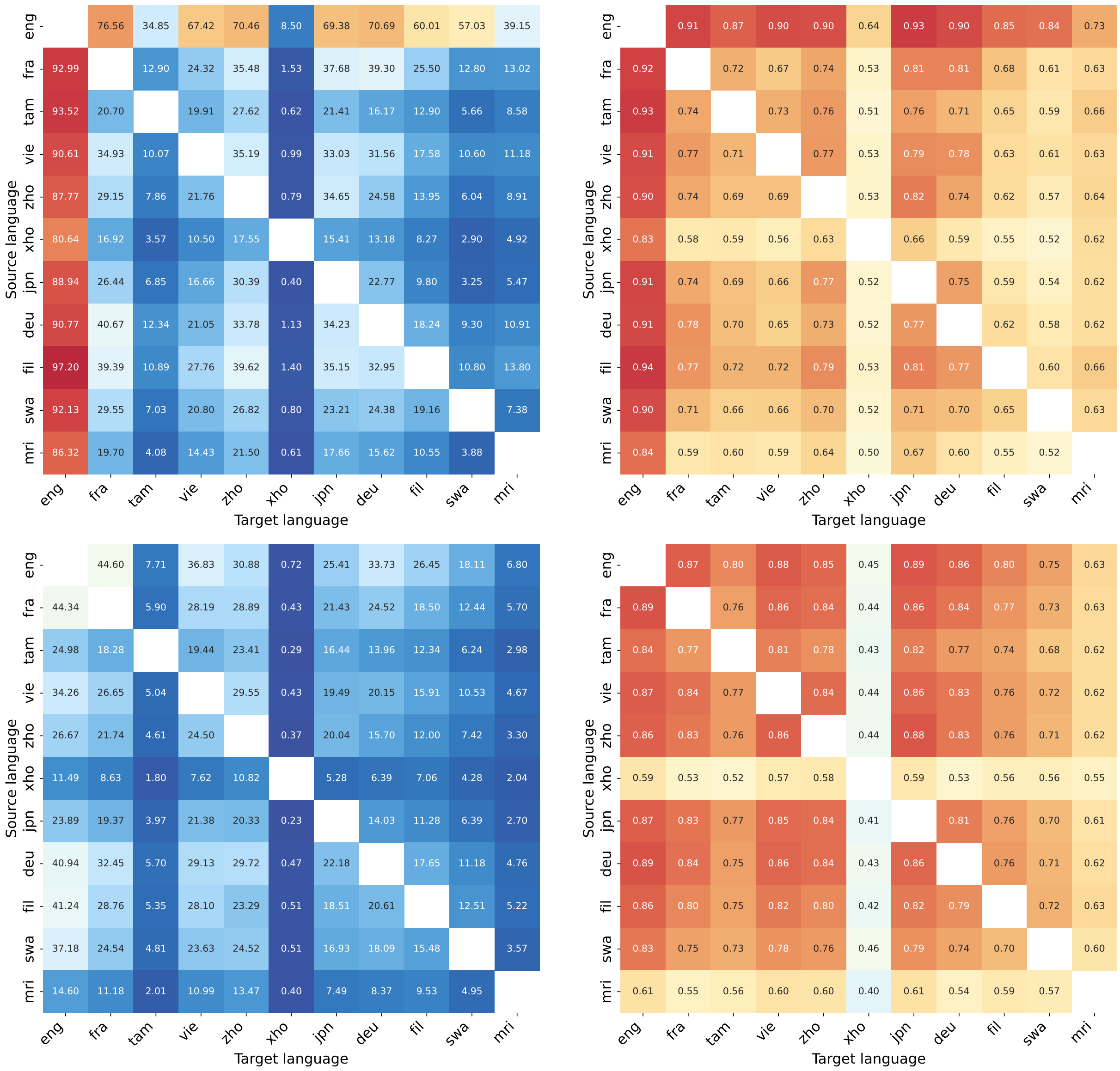

Figure 5: BLEU scores for the fine-tuned Llama (upper left), base Llama (lower left) and COMET scores for the fine-tuned Llama (upper right) and base Llama (lower right).

G Examples of Llama Back-Translations

Completely unrelated

Llama's por→ewe:	Amevuwo aɖekele mekpɛkpɛkpɛwo ɖe le ɖekele mekpɛkpɛkpɛwo ɖekele mekpɛkpɛkpɛwo ɖekele mekpɛkpɛkpɛwo ... ɖekele mekpɛkpɛkpɛwo ɖekele mekp (*There are no cats in the one cats one cats one cats ... one cats snowflakes one snowflakes one snowflakes*)
Reference:	Nugɔmekula siwo do ŋgɔ gblɔ be esia ateŋu aɖe kansa, yɔmekpe, HIV dɔlékuia kple asrãdɔ fia kaba le du siwo me fetu mede naneke o la me. Dukɔ siwo me amewo tsia agbe le dɔlélewo abe no me kansa ene la anɔ abe dukɔ deŋgɔwo ƒe afã tɔ ene. (*Previous researchers have suggested that this could lead to early detection of cancer, tuberculosis, HIV and malaria in low-income cities. Countries have about half the survival rates of developed countries.*)
Llama's bam→vie:	Người đàn ông đã nhìn thấy một người phụ nữ đang đi trên đường, nhưng người phụ nữ đó đã đi mất rồi. (*The man saw a woman walking down the street, but she had already left.*)
Reference:	Truyền thông địa phương đưa tin một phương tiện chữa cháy sân bay đã tới khi trả lời. (*Local media reported that an airport fire engine responded.*)

Sharing entities and numbers

Llama's bam→mal:	ആകാ ഒ സൂ ഇൻ സലെു ഫിലനാൻ ടൂൺ യെ ആകാ സസെോൺ സലെു 60 നാൻ ഡെ യെ, മിൻ ബെ അലെ കെ ന്റോളോന്റില ഫോളോ യെ കാ സലെു 60 ഡോ വലിമ കാ സയ നിൻ സസെോൺ കലേൻ സലെു ഡോനൻ യെ കബ (*If you are in the same season as the 60th, then you will be able to follow the 60th season of the 60th season.*)
Reference:	അദ്ദേഹം നടിയ രണ്ടാമത്തെ ഗോൾ സീസണിലെ അറുപതാമത്തേത് ആയിരുന്നു, 1995-96 കാലഘട്ടത്തിൽ ജറോമിർ ജാഗറും മരിയോ ലമ്യൂക്സും ആ നേട്ടം കൈവരിച്ച ശേഷം ഒരു സീസണിൽ 60-ഓ അതിൽ കൂടുതലോ ഗോളുകൾ നേടുന്ന ആദ്യ കളിക്കാരനായി അതോടെ അദ്ദേഹം മാറി. (*His second goal was his 60th of the season, making him the first player to score 60 or more goals in a season since Jaromir Jáger and Mario Lemieux achieved the feat in 1995–96.*)
Llama's por→fon:	Afon: De kɛra ɲa Ansa, "a polisi ɲa ɲɛra kɛra ɲa ɲɛra kɛra ɲa ɲɛra ... (*Afon: De becomes eyes Ansa, "a police eyes eyes eyes ...*)
Reference:	Xó e Ansa ɖɔ e wɛ nyi ɖɔ "adò hu kponɔ lɛ do gǎn ɖaxó wè ye hu sixu zon bɔ ahwan nan bɛ dó gǎn yɔyɔ̌ ye jló na sɔ́ wutu. (*According to Ansa, "the police are worried that the killing of two senior leaders could lead to a war over the new leader they want to elect.*)

Sharing structures and phrases

Llama's asm→zho:	茶叶和茶四叶在心态和心态协调方面与心灵感应者保持着安全的距离。 (*Tea and Tea Four Leaves maintain a safe distance from telepaths in terms of mindset and mental coordination.*)
Reference:	据报告，焦立中（Chiao）和沙里波夫（Sharipov）与调姿推进器的距离是安全的 (*Chiao and Sharipov were reportedly at a safe distance from the attitude thrusters.*)
Llama's asm→spa:	Los factores naturales pueden combinarse y generar condiciones óptimas, lo que permite que el alga crezca enormemente en cantidad. (*Natural factors can be cut to produce an ideal cartogram, and these trays are approved to fictitiously increase the quantity.*)
Reference:	Se pueden permitir cortar los factores naturales para producir un ideal cartograma, y se aprueban estas bandejas para aumentar la cantidad ficticiamente. (*Natural factors can combine to generate optimal conditions, allowing the algae to grow enormously in quantity.*)

Table 4: Examples of Llama's back-translated source (*with English translations from Google Translate*).

H Examples of Entity Replaced Sentences

Entity Replaced Sentences	
Source	**Since then, the Brazilian has featured in 53 matches for the club in all competitions and has scored 24 goals.**
Labeled Entities:	{'label': 'NORP', 'text': 'Brazilian', 'new_ent': 'Argentine'}, {'label': 'CARDINAL', 'text': '53', 'new_ent': '78'}, {'label': 'CARDINAL', 'text': '24', 'new_ent': '19'}
New Sentences:	Since then, the **Argentine** has featured in 53 matches for the club in all competitions and has scored 24 goals. Since then, the Brazilian has featured in **78** matches for the club in all competitions and has scored 24 goals. Since then, the Brazilian has featured in 53 matches for the club in all competitions and has scored **19** goals.
Source	**1970 年，医学博士兼研究科学家雷蒙德·达马迪安 (Raymond Damadian) 发现了使用磁共振成像作为医学诊断工具的基础原理。** *(In the year 1970, Raymond Damadian, a medical doctor and research scientist, discovered the basis for using magnetic resonance imaging as a tool for medical diagnosis.)*
Labeled Entities:	{'label': 'DATE', 'text': '1970 年', 'eng': 'the year 1970', 'new_eng_ent': 'the year 2022', 'new_zho_ent': '2022 年'}, {'label': 'PERSON', 'text': '雷蒙德·达马迪安 (Raymond Damadian)', 'eng': 'Raymond Damadian', 'new_eng_ent': 'Elon Musk', 'new_zho_ent': '埃隆·马斯克'}
New Sentences:	**2022 年**，医学博士兼研究科学家雷蒙德·达马迪安 (Raymond Damadian) 发现了使用磁共振成像作为医学诊断工具的基础原理。 1970 年，医学博士兼研究科学家**埃隆·马斯克**发现了使用磁共振成像作为医学诊断工具的基础原理。
Source	**La Guerre de Succession d'Espagne marqua la Première Guerre dont l'issue principale fut l'équilibre des pouvoirs.** (*The War of Spanish Succession marked the first war whose central issue was the balance of power.*)
Labeled Entities:	{'label' : 'EVENT', 'text' : 'La Guerre de Succession d'Espagne', 'eng' : 'The War of Spanish Succession', 'new_eng_ent' : 'The Napoleonic Wars', 'new_fra_ent' : 'Les guerres napoléoniennes'}, {'label' : 'ORDINAL', 'text' : 'Première', 'eng' : 'first', 'new_eng_ent' : 'second', 'new_fra_ent' : 'deuxième'}
New Sentences:	**Les guerres napoléoniennes** marqua la Première Guerre dont l'issue principale fut l'équilibre des pouvoirs. La Guerre de Succession d'Espagne marqua la **deuxième** Guerre dont l'issue principale fut l'équilibre des pouvoirs.

Table 5: Examples of entity replaced sentences.

I Entity Descriptions

CARDINAL	Numerals that do not fall under another type
DATE	Absolute or relative dates or periods
EVENT	Named hurricanes, battles, wars, sports events, etc.
FAC	Buildings, airports, highways, bridges, etc.
GPE	Countries, cities, states
LANGUAGE	Any named language
LAW	Named documents made into laws.
LOC	Non-GPE locations, mountain ranges, bodies of water
MONEY	Monetary values, including unit
NORP	Nationalities or religious or political groups
ORDINAL	”first”, ”second”, etc.
ORG	Companies, agencies, institutions, etc.
PERCENT	Percentage, including “%”
PERSON	People, including fictional
PRODUCT	Objects, vehicles, foods, etc. (not services)
QUANTITY	Measurements, as of weight or distance
TIME	Times smaller than a day
WORK_OF_ART	Titles of books, songs, etc.
MISC	Miscellaneous entities, e.g. events, nationalities, products or works of art

Table 6: Description of each Entity

J Prompts

```
f"{{{sent}}}\nCan you translate this to tgt_lang?"
```

Table 7: Prompt for translation with Bloomz.

```
{
    "role": "system",
    "content": "You are a helpful assistant that translates text. Your task is to translate a sentence that
    will be provided."
},
{
    "role": "user",
    "content": f"Translate the following {src_name} sentence to {tgt_name}. Your response should contain only
    the translation and be structured like this: {{{{Your response goes here}}}}\n{sent}"
}
```

Table 8: Prompt for translation with Llama.

```
f"You are a native {lang_name} speaker.\n"
f"Task: Rephrase the following {lang_name} sentence.\n"
f"Constraints:\n"
f"- Output only one sentence.\n"
f"- Do not include explanations, notes, or formatting.\n"
f"- Do not repeat the input.\n"
f"- Do not add any extra characters or line breaks.\n\n"
f"Input:\n{src}\n\n"
f"Output:"
```

Table 9: Prompt for paraphrasing with Llama.

```
{
                "role": "user",
                "content": (
                    f"You will be given a parallel English sentence and its corresponding "
                    f"{tgt_name} sentence. The labeled named entities of the English sentence "
                    "will be provided.\n\n"

                    "Your task:\n"
                    "- For each English entity, identify the exact surface form in the "
                    f"{tgt_name} sentence that refers to the same real-world entity.\n"
                    "- DO NOT add new entities.\n"
                    "- The number of returned entities and the labels MUST match the English list.\n"
                    "- Return each entity span exactly as it appears in the target sentence.\n"
                    "- If an English entity appears multiple times, include all of them.\n"
                    "- If the English entity list is empty, return an empty list: []\n\n"

                    "Your output MUST be valid JSON ONLY, with this exact structure:\n\n"
                    "{\n"
                    f'  "entities_{tgt_lang_code}": [\n'
                    '    {"label": "ENTITY_LABEL", "text": "ENTITY_IN_TARGET_LANGUAGE", "eng":
                    "ENTITY_IN_ENGLISH"}\n'
                    "  ]\n"
                    "}\n\n"

                    f"English Sentence:\n{src}\n\n"
                    f"English Entities:\n{ent_eng}\n\n"
                    f"{tgt_name} Sentence:\n{ref}\n\n"
                    "Output JSON:"
                )
            }
```

Table 10: Prompt for entity alignment with Aya.

```
{
                "role": "user",
                "content": (
                    "You are given a list of named entities in English.\n"
                    "For each entity, generate a NEW entity of the SAME label type.\n\n"

                    "RULES:\n"
                    "- Do NOT add or remove entities.\n"
                    "- 'new_ent' must be a completely different real entity from the original.\n"
                    "- Do NOT use entities found in the original 'text' value.\n"
                    "- Synonyms, paraphrases, abbreviations, variants, or anything derived from the original
                    are NOT allowed.\n"
                    "- The new entity must be real, valid for its label, and not share the same referent.\n"
                    "- Take the grammar structure of the entity into account if it is multiple words long.\n"
                    "- Output JSON only.\n\n"


                    "OUTPUT FORMAT:\n"
                    "[{\"label\": \"LABEL\", \"text\": \"ORIGINAL_TEXT_FROM_ENTITIES_LIST\", \"new_ent\":
                    \"NEW_ENTITY\"}]\n\n"

                    f"Entities:{ent_list}\n\n"
                    "Output JSON only:"
                )
            }
```

Table 11: Prompt for new entity generation with Aya.

```
{
                "role": "user",
                "content": (
                    f"Translate the following English *ENTITY* into {tgt_name}.\n\n"

                    "RULES:\n"
                  "- Your translation MUST be grammatically correct **for the position the entity occupies** "
                    "in the target-language sentence.\n"
                    "- Imagine the translated entity being inserted into the sentence below.\n\n"

                    "RESTRICTIONS:\n"
                    "- Translate ONLY the entity.\n"
                    "- Do NOT output the whole sentence.\n"
                    "- Do NOT rewrite or paraphrase anything.\n"
                    "- No explanations.\n"
                    "- No punctuation or quotes.\n"
                    "- Output ONLY the final translated entity.\n\n"

                    f"English Sentence (Original):\n{sources[i]}\n\n"
                    f"English Sentence After Replacement:\n{eng_replaced_list[i]}\n\n"
                    f"Target-Language Reference Sentence:\n{references[i]}\n\n"
                    f"ENTITY TO TRANSLATE:\n{new_ents[i]}\n\n"

                    "Output ONLY the translated entity:"
                )
            }
```

Table 12: Prompt for translating new entities with Aya.