

# BridG MT: Enhancing LLMs’ Machine Translation Capabilities with Sentence Bridging and Gradual MT

Seungwoo Choi, Gahyun Yoo, Jay-Yoon Lee\*

Seoul National University

{rhdn520, padme0421, lee.jayyoon}@snu.ac.kr

## Abstract

Recent Large Language Models (LLMs) have demonstrated impressive translation performance without requiring fine-tuning on additional parallel corpora. However, they still face significant challenges in certain scenarios, particularly when translating low-resource languages. A common approach to address this issue is to provide external knowledge, such as few-shot examples, to assist LLMs in translating specific source sentences. However, this method is fundamentally limited by the quality or quantity of relevant sources, which cannot always be guaranteed. To reduce LLMs’ reliance on external sources, we propose *BridG MT*, a method that combines *Sentence Bridging*, which generates a sequence of sentences as a bridge that gradually transition from easy-to-translate to more difficult, and *Gradual MT*, which sequentially translates these sentences using earlier translations as few-shot examples for subsequent ones. Experiments conducted on four LLMs across seven languages demonstrate that our method effectively enhances translation performance, even outperforming translation methods that rely on a large number of few-shot examples.

## 1 Introduction

Recent Large Language Models (LLMs) have shown strong performance in translation tasks without the need for fine-tuning on specific parallel datasets. Previous studies have demonstrated that LLMs’ translation capabilities are reliable in most use cases, particularly when the source and target language are high-resource languages (Zhu et al., 2024; Robinson et al., 2023; Jiao et al., 2023). However, because LLMs require training on large corpora, they still face challenges when translating low-resource languages that are not sufficiently represented in the training corpora. (Stap and Araabi, 2023; Robinson et al., 2023; Enis and Hopkins, 2024).

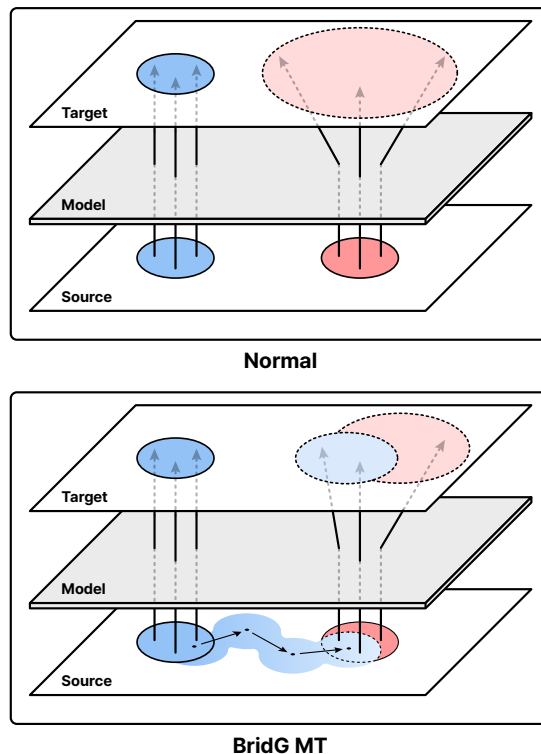


Figure 1: Figurative illustration of BridG MT. Machine translation maps between source and target language spaces. BridG MT expands well-performing areas (blue) by leveraging neighboring examples, reaching poorly performing areas (red). Circle sizes indicate output noisiness.

Previous research has attempted to address these challenges by leveraging the in-context-learning capabilities of large language models (LLMs), particularly through the use of external knowledge such as few-shot examples or dictionaries during inference. However, relevant examples are not always guaranteed to be available, and constructing such external knowledge sources can be costly. A potential solution is to reduce reliance on external sources altogether.

In this paper, we investigate whether LLMs’ translation capabilities can be improved without

relying on external knowledge. Our approach is grounded in a fundamental idea: machine translation can be understood as a mapping between two (sub)spaces representing the source and target languages. Within the source language space, some regions yield accurate translations, while others do not. The key question is whether we can expand the model’s confident region by refeeding examples from it, ultimately extending into weaker regions and improving performance in those areas.

We propose *BridG MT* to enhance machine translation by bridging regions where the LLM performs well with those where it struggles. BridG MT integrates two key techniques: *Sentence Bridging* and *Gradual MT*. Sentence Bridging is a prompting method that generates a sequence of sentences, progressively transitioning between them. Gradual MT iteratively translates a list of sentences, using the model’s previous translations as few-shot examples for subsequent ones. Sentence Bridging helps bridge high- and low-performance regions, while Gradual MT gradually expands the model’s strong performance areas by leveraging in-context learning. These concepts are visually illustrated in Figure 1.

We evaluate the effectiveness of BridG MT by applying it to four different LLMs: GPT-3.5, Mistral-Nemo-Instruct, Llama-3.1-70B-Instruct, and Llama-3.1-8B-Instruct. The experiments cover seven target languages: German (De), Chinese (Zh), Hindi (Hi), Korean (Ko), Swahili (Sw), Marathi (Mr), and Bengali (Bn). The results demonstrate that BridG MT significantly enhances the translation capabilities of LLMs, particularly in low-resource languages. To the best of our knowledge, we are the first to propose the concept of Sentence Bridging.

## 2 Related Work

### 2.1 Enhancing LLMs’ Translation Capabilities without Fine-tuning

Modern LLMs demonstrate strong translation capabilities in high-resource languages but struggle with low-resource languages (Jiao et al., 2023; Stap and Araabi, 2023; Zhu et al., 2024; Enis and Hopkins, 2024). Several studies have focused on enhancing LLMs’ translation performance without additional fine-tuning. A primary approach involves leveraging LLMs’ ability to learn from demonstrations or descriptions through in-context learning (Brown et al., 2020; Wei et al., 2022). Researchers

have explored methods such as selecting appropriate exemplars for few-shot learning and demonstrating linguistic knowledge (Agrawal et al., 2022; Vilar et al., 2023; Zhang et al., 2024). Beyond simply providing examples, some approaches conduct a prior analysis of the sentence to be translated, using LLMs to extract relevant information. These approaches then supplement the translation process with resources such as chains of multilingual dictionaries (Lu et al., 2024), or adopt human-like strategies by providing the LLM with keywords, topics, and generated demonstrations (He et al., 2024). Other studies have instead focused on generating multiple translation candidates and selecting the optimal output through ranking mechanisms such as Minimum Bayes Risk decoding (Fernandes et al., 2022).

### 2.2 Self-Demonstration

Manually generating appropriate exemplars for in-context learning can be resource-intensive. To address this challenge, previous studies have explored enabling models to generate their own few-shot examples for tasks such as classification (Lyu et al., 2023; Kim et al., 2022) or other reasoning tasks (Zhang et al., 2023; Li et al., 2024). Our work is aligned with these efforts, as it also focuses on generating the model’s own few-shot examples. However, these techniques have yet to be applied to translation, and BridG MT’s novel approach of intentionally bridging confident regions to current predictions remains unexplored.

## 3 Methodology

We introduce two major components of BridG MT—Sentence Bridging and Gradual MT—in Sections 3.1 and 3.2, respectively, and then describe how we combine them in Section 3.3.

### 3.1 Sentence Bridging

Sentence bridging is a prompting technique that asks the model to generate a list of sentences that gradually transition from a *start sentence* to an *end sentence*. The objective of this technique is to generate a list of sentences where each sentence is distinct, yet not excessively different from its adjacent sentences. We call these sentences the *bridge*. In our experiments, we utilized three bridging examples from GPT-4 (OpenAI et al., 2024) as a few-shot to control the output format. See Appendices A and B for these examples and sample

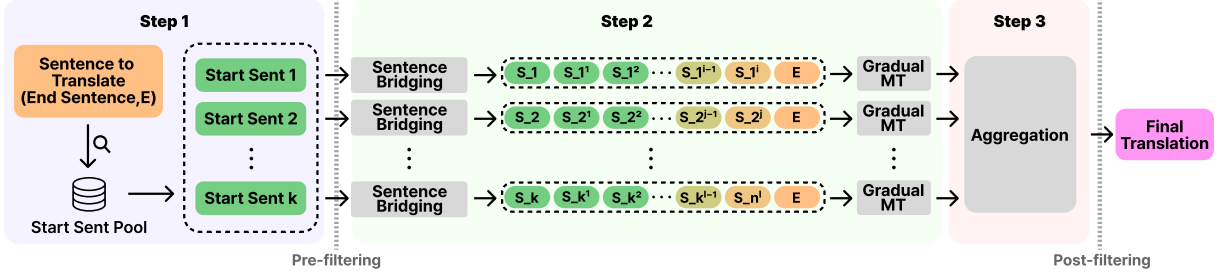


Figure 2: Illustration of the BridG MT algorithm. BridG MT integrates Sentence Bridging with Gradual MT. In Step 1,  $k$  start sentences are selected from a predefined start-sentence pool. In Step 2, these start sentences are bridged to the end sentence, creating  $k$  individual bridges. Each bridge is then processed through Gradual MT, generating translation results for every sentence along the path. In Step 3, the MT results from all bridges are aggregated into a single output translation. Optional pre- and post-filtering steps can be applied between Steps 1 and 2, and again after aggregation, to refine the sentences on which BridG MT is applied.

---

**Algorithm 1** Pseudo Code for Gradual MT

---

**Input:** Bridge  $X = \{x_1, x_2, \dots, x_n\}$ ,  $n = |X|$ ,  
Translation model  $M$   
Set  $fewshot = []$   
**for**  $x_i \in X$  **do**  
     $\hat{y}_i = M(x_i, fewshot)$   
    Append  $\{x_i, \hat{y}_i\}$  to  $fewshot$   
**end for**  
**Output:**  $\hat{y}_n$

---

bridges. Following is the prompt that we use in our experiments:

I will give you two sentences. Can you gradually change the first sentence to make it exactly the same as the second sentence? Just give me the sentences and don't provide additional comments.  
Sentence1:  $\langle Sentence1 \rangle$   
Sentence2:  $\langle Sentence2 \rangle$

### 3.2 Gradual MT

Gradual MT is a prompting technique that enables an LLM to leverage its previous translations as prompts. This approach sequentially processes a bridge, translating each source sentence while using the preceding translation results as few-shot examples for the current sentence. An illustration of the gradual MT algorithm is provided in Algorithm 1. By applying gradual MT, we construct an expanding set of few-shot examples, which become increasingly useful for translating the final sentence that the model must process.

### 3.3 Overall Method

BridG MT combines sentence bridging and gradual MT. An illustration of the algorithm is shown in

Figure 2.

**Step 0: Start Sentence Pool Creation** Before applying BridG MT, the *start sentence pool* must first be created. This pool consists of sentences that the zero-shot LLM can translate with high accuracy. Since we are considering a scenario where there is no reliable gold translation, we utilize a reference-free quality estimation (QE) model such as CometKiwi (Rei et al., 2023) to construct the start sentence pool. We refer to the source sentences in the start sentence pool as *start sentences*.

**Step 1: Start Sentence Selection** BridG MT begins with selecting  $k$  start sentences from the start sentence pool by calculating the similarity with the source sentence that the LLM is trying to translate, which we call an *end sentence*. Measuring similarity between sentences can be done in various ways. In this paper, we utilized SBERT similarity (Reimers and Gurevych, 2019) as a primary metric. Details about start sentence selection strategy can be found in Section 6.1.

**Step 2: Sentence Bridging & Gradual MT** After selecting start sentences, sentence bridging is performed between each start sentence and the corresponding end sentence, creating  $k$  separate *bridges*. The bridges are then processed through the Gradual MT, generating translation results for each sentence.

**Step 3: MT Results Aggregation** When the number of start sentences is more than one, we proceed to aggregate the translation results from each bridge. This step is skipped if we choose to work with single start sentence in Step 1. There could be several methods for aggregating translations; in our approach, we input all the translation

results into the LLM once again as few-shot examples to generate the final translation. A detailed ablation of this aggregation strategy can be found in Section 6.2. After aggregation, we obtain the final translation, which we refer to as the *output*.

**Pre- & Post-filtering** BridG MT can be applied to any sentence; however, it is often more effective to use it selectively for two reasons. First, determining which sentences will undergo BridG MT before its execution can significantly reduce computational costs. Second, even after applying BridG MT, it is better to discard the BridG MT result if its quality is worse than the zero-shot translation. We explored these two possibilities—referred to as Pre-filtering and Post-filtering—in Section 6.3 and show that these methods can improve quality with smaller computation than full application of BridG MT.

## 4 Experiment

### 4.1 Setup

**Models** For **translation**, we use four different LLMs: GPT-3.5 (GPT-3.5-Turbo-0125)<sup>1</sup>, Mistral-Nemo (Mistral-Nemo-Instruct-2407)<sup>2</sup>, and two Llama models (Llama-3.1-70B-Instruct, Llama-3.1-8B-Instruct) (Dubey et al., 2024). ChatGPT is accessed via OpenAI’s API, while the other models ran on local GPU. For **sentence bridging**, we employ Qwen2-72B-Instruct (Yang et al., 2024) with 4-bit quantization as our primary bridging model. We also tested Llama-3.1-8B-Instruct and Llama-3.1-3B-Instruct to test the generalizability of the method in more cost-efficient settings. See Appendices D and E for settings for translation and bridging. For **pre- and post-filtering**, we utilize a reference-free QE model CometKiwi (Rei et al., 2023) to avoid peeking at the gold translations. CometKiwi predicts a DA score, which rates translation quality on a scale from 0 to 100, normalized to a range of 0 to 1. Lastly, we used all-mpnet-base-v2<sup>3</sup> for **SBERT similarity** calculation in retrieving start sentences and few-shot examples for baseline experiments. All experiments were conducted once.

<sup>1</sup><https://platform.openai.com/docs/models/gpt-3-5-turbo>

<sup>2</sup><https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407/>

<sup>3</sup><https://huggingface.co/sentence-transformers/all-mpnet-base-v2>

**Target Languages** We fixed English as the source language. The target languages tested in the experiments are German (De), Chinese (Zh), Korean (Ko), Hindi (Hi), Swahili (Sw), Bengali (Bn), and Marathi (Mr). Based on Joshi et al. (2020)’s 6 scale resource level, we classify German and Chinese as high resource, Korean and Hindi as mid resource, and the rest as low resource. For GPT 3.5, we experiment on every languages. For the Llama 3.1 models, we only experiment on German and Hindi, as they do not support other languages. Mistral-Nemo does not officially support Ko, Hi, Sw, Bn, and Mr, but we conduct experiments on those languages nevertheless, as it has some capability to generate them.

**Dataset** We use the FLORES-200 benchmark dataset (Team et al., 2022) for validation and evaluation. The development split is used to construct the start sentence pool. For the test set, we sample 90% of the data and reserve the remaining 10% to determine the QE score threshold for pre-filtering. We also use NTREX-128 (Federmann et al., 2022) to evaluate whether BridG MT remains effective when the distribution of the start sentence pool differs from that of the test set. The results for NTREX-128 are presented in Section 7.

**Start Sentence Pool Creation** The start sentence pool is created by translating source sentences from the dev split of the FLORES-200 dataset using a zero-shot approach. Each source sentence is translated five times and evaluated with CometKiwi—a reference-free model. The most frequently occurring translation is selected as the *representative translation*. If no translation is repeated, the one with a score closest to the average is chosen. After selecting each representative translation, the top 100 translation pairs with the highest DA scores are selected.

**Baselines** Since our focus is on scenarios where providing relevant few-shot examples is not feasible, we first compare BridG MT with zero-shot baselines. We then extend the comparison to conventional few-shot translation settings, where gold translations sampled from the start sentence pool are used as demonstrations (15, 30, and 50 examples). This enables us to evaluate how well BridG MT—designed specifically for settings without accessible few-shot examples—performs relative to standard few-shot approaches. We further compare BridG MT with related methods: MAPS (He et al.,



Table 1: xCOMET scores of BridG MT across different translation models and target languages. Sentence bridging was performed using Qwen2-72B-Instruct. Scores are multiplied by 100 for readability. For each MT model and language, the highest score is shown in bold and the second-highest is underlined. Cells are left blank when a specific model or method does not support the corresponding language.

MT Model	Method	High Resource		Mid Resource		Low Resource		
		DE	ZH	HI	KO	SW	BN	MR
GPT 3.5	Zero-shot	97.63	91.30	71.89	89.48	81.23	68.73	44.53
	MAPS	97.68	92.06	-	-	-	-	-
	QA Decode	97.45	91.63	75.45	90.19	78.26	73.90	45.82
	BridG <sub>Post</sub>	<b>98.04</b>	<b>92.63</b>	<b>77.90</b>	<b>92.57</b>	<b>83.65</b>	<b>75.41</b>	<b>53.11</b>
	BridG <sub>Pre &amp; Post</sub>	<u>97.86</u>	<u>92.35</u>	<u>77.63</u>	<u>92.07</u>	<u>83.34</u>	<u>75.14</u>	<u>51.51</u>
Llama 3.1 70B	Zero-shot	97.33	-	79.41	-	-	-	-
	MAPS	96.53	-	-	-	-	-	-
	QA Decode	97.37	-	-	-	-	-	-
	BridG <sub>Post</sub>	<b>97.93</b>	-	<b>84.45</b>	-	-	-	-
	BridG <sub>Pre &amp; Post</sub>	<u>97.73</u>	-	<u>84.31</u>	-	-	-	-
Llama 3.1 8B	Zero-shot	94.99	-	69.93	-	-	-	-
	MAPS	79.10	-	-	-	-	-	-
	QA Decode	88.08	-	<u>77.67</u>	-	-	-	-
	BridG <sub>Post</sub>	<b>97.25</b>	-	<b>78.30</b>	-	-	-	-
	BridG <sub>Pre &amp; Post</sub>	<u>96.99</u>	-	77.37	-	-	-	-
Mistral Nemo	Zero-shot	96.70	88.01	66.78	81.43	38.97	71.89	43.99
	MAPS	85.75	88.71	-	-	-	-	-
	QA Decode	96.33	87.18	64.42	80.34	28.86	68.59	35.37
	BridG <sub>Post</sub>	<b>97.70</b>	<b>90.99</b>	<b>73.95</b>	<b>89.45</b>	<b>45.67</b>	<b>78.35</b>	<b>57.24</b>
	BridG <sub>Pre &amp; Post</sub>	<b>97.70</b>	<b>90.99</b>	<u>73.82</u>	89.34	<u>45.18</u>	<u>77.83</u>	<u>54.55</u>
TowerInstruct		97.69	89.89	-	91.29	-	-	-
NLLB		96.21	67.88	81.00	82.20	77.17	82.70	71.83

2024)<sup>4</sup> and Quality Aware(QA) Decode (Fernandes et al., 2022)<sup>5</sup>. Among the target languages in our experiments, MAPS provides few-shot examples only for De and Zh in its published codebase; therefore, it was evaluated only on these languages. For QA Decode, we used MBR reranking, as it showed the best performance in their paper. Finally, to provide a broader context for comparison, we benchmark BridG MT against strong baseline models, including TowerInstruct 13B (Alves et al., 2024) and NLLB-200-3.3B<sup>6</sup>. Among the languages supported by TowerInstruct, only De, Zh, and Ko overlapped with our target languages; thus, experiments with TowerInstruct were conducted only on these languages.

**Evaluation** We employ four different QE models for evaluation; xCOMET, MetricX(Juraska et al., 2023), CometKiwi, and BLEURT(Pu et al., 2021). xCOMET, CometKiwi and BLEURT predict DA score, and MetricX predicts MQM score. MQM score assesses translation errors on a scale from 0 to 25, where lower scores indicate higher quality. Lastly, we report BLEU and ChrF.

<sup>4</sup><https://github.com/zwhe99/MAPS-mt>

<sup>5</sup><https://github.com/deep-spin/qaware-decode>

<sup>6</sup><https://huggingface.co/facebook/nllb-200-3.3B>

## 5 Results

We first conducted a series of ablation experiments for En-Ko translation using GPT 3.5 to identify the optimal configuration, including start selection (Step 1) and result aggregation (Step 3). This optimal configuration was then applied in the main experiments across different translation models, sentence bridging models, and target languages. A detailed explanation of the ablation study is provided in Section 6. In this section, we present the xCOMET results of BridG MT across various LLMs and target languages. The full results are provided in Appendix F.

### 5.1 Comparing with Zeroshot Baselines

As shown in Table 1, BridG MT consistently improves the translation performance of LLMs, particularly for low-resource languages. For example, the performance gaps of GPT 3.5 between BridG MT and the baseline are 0.4, 1.3, 6.0, 3.1, 2.4, 6.7, and 8.6 for De, Zh, Hi, Ko, Sw, Bn, and Mr, respectively. This trend remains consistent across other evaluation metrics (Tables 12, 13, 14, 15, and 16). BridG MT also outperforms alternative methods—MAPS and QA Decode. Notably, similar patterns are observed even when using QE models

Table 2: xCOMET scores of BridG MT across different translation models, sentence bridging models, and target languages. Scores are multiplied by 100 for readability, and the two highest scores for each MT model and language are highlighted in bold and underlined. For languages not supported by each model, the corresponding cells were left blank.

MT Model	Method	Bridging Model	DE	ZH	HI	KO	SW	BN	MR
GPT 3.5	BridG MT	Qwen2 72B Instruct	<b>98.04</b>	<b>92.63</b>	<b>77.90</b>	<b>92.57</b>	<b>83.65</b>	<b>75.41</b>	<b>53.11</b>
		Llama 3.1 8B	<u>97.97</u>	<u>92.38</u>	<u>77.24</u>	<u>91.85</u>	<u>83.62</u>	<u>74.24</u>	<u>52.27</u>
		Llama 3.2 3B	97.97	92.35	76.44	91.48	83.49	<u>74.35</u>	52.13
	Zero-shot	N/A	97.63	91.30	71.89	89.48	81.23	68.73	44.53
Llama 3.1 70B	BridG MT	Qwen2 72B Instruct	<b>97.93</b>	-	<b>84.45</b>	-	-	-	-
		Llama 3.1 8B	<u>97.87</u>	-	<u>83.97</u>	-	-	-	-
		Llama 3.2 3B	97.85	-	83.65	-	-	-	-
	Zero-shot	N/A	97.33	-	79.41	-	-	-	-
Llama 3.1 8B	BridG MT	Qwen2 72B Instruct	<b>97.25</b>	-	<b>78.30</b>	-	-	-	-
		Llama 3.1 8B	96.93	-	<u>77.83</u>	-	-	-	-
		Llama 3.2 3B	97.14	-	77.65	-	-	-	-
	Zero-shot	N/A	94.99	-	69.93	-	-	-	-
Mistral-Nemo	BridG MT	Qwen2 72B Instruct	<b>97.70</b>	<b>90.99</b>	<b>73.95</b>	<b>89.45</b>	<b>45.67</b>	<b>78.35</b>	<b>57.24</b>
		Llama 3.1 8B	<u>97.52</u>	<u>90.40</u>	<u>72.34</u>	<u>87.83</u>	<u>43.33</u>	<u>76.73</u>	<u>52.93</u>
		Llama 3.2 3B	97.42	90.04	71.62	86.55	42.30	75.83	51.66
	Zero-shot	N/A	96.70	88.01	66.78	81.43	38.97	71.89	43.99

other than COMET and when varying the choice of bridging model. As shown in Table 2, using smaller bridging models leads to a slight decline in performance, but the overall improvement over the baseline remains substantial.

## 5.2 Comparing with Baselines Using Labelled Data

We further compared our results to a few-shot setup. To provide few-shot examples under conditions similar to our approach, we built a few-shot pool by selecting 100 sentences from the FLORES-200 dev set that the model translates well—similar to how we created the start sents pool. However, in this case, we employed xCOMET, a reference-based QE model. Figure 3 presents the average translation scores of GPT-3.5 and Mistral-Nemo. The results indicate that our method is comparable to, and often outperforms, few-shot translations. This demonstrates that even without high-quality gold translations, BridG MT can surpass translations generated by an LLM equipped with gold few-shot references. Additionally, we observe a performance drop when the model is given 50-shot examples, suggesting that BridG MT achieves performance levels that cannot be matched simply by increasing the number of few-shot examples. These findings remain consistent when using Llama models and evaluating with other metrics. See Appendix F for detailed results.

## 6 Ablation Study

To find the optimal configuration for the BridG MT, we conducted an ablation study on En-Ko transla-

tion task using GPT-3.5 in terms of four different dimensions: Start sentence selection criteria, Number of start sentences, MT result aggregation strategy, and Filtering strategy. We utilized xCOMET with gold translations in the ablation study to construct the highest-quality start pool possible. Full results with all different combination of ablation settings can be found in Appendix G.

### 6.1 Start Sentence Selection Strategy

Based on the intuition that start sentences similar to the source sentence will be helpful, we employed three different metrics—SBERT similarity (Reimers and Gurevych, 2019), Levenshtein distance (Levenshtein and others, 1966), and tree edit distance (Zhang and Shasha, 1989)—to calculate the similarity between sentences, and we combined them in three different ways. The first approach, *Sort*, sorts sentences using multiple metrics with varying priorities. The second approach, *Filter*, initially selects the top 10 sentences based on SBERT similarity and then sorts that selection using the other metrics. The third approach, *Tops*, picks the top sentence based on the highest similarity scores from each metric. As shown in Table 3, the selection strategies that produced the highest scores differed across aggregation methods. We chose to sort by SBERT similarity and then by tree edit distance (Sort(S-T)), as this approach yielded the highest average scores.

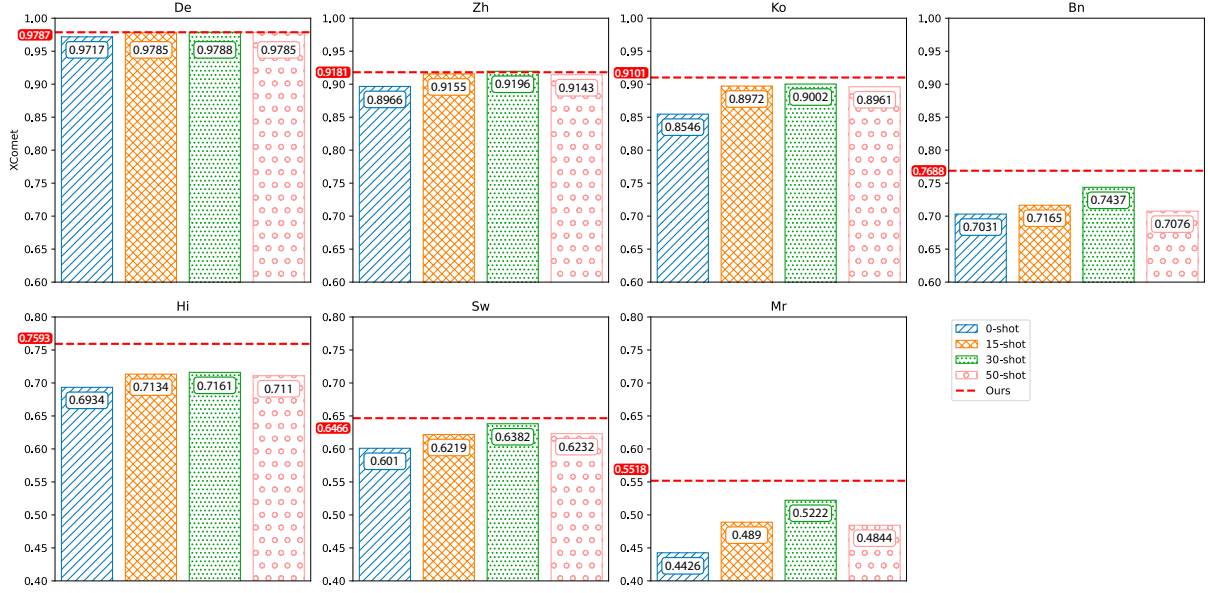


Figure 3: xCOMET scores for few-shot translations (bars) and BridG MT translations (horizontal line). Both BridG MT and few-shot scores represent the average xCOMET scores of GPT-3.5 and Mistral-Nemo.

Table 3: Averaged scores with different start selection strategies, start sentence numbers and aggregation strategies applied to EN-KO translation task. ‘Sort’, ‘Filter’, and ‘Tops’ denote the start sentence selection methods. Each letter in parentheses represents a similarity metric, with the order indicating the priority of these metrics. ‘L’, ‘T’ and ‘S’ stands for Levenshtein distance, Tree edit distance, and SBERT similarity, respectively. The highest average values for each axis are highlighted in bold. The highest values for each start sentence number and aggregation strategy are highlighted in underline.

Start Selection Strategy	Start Sents.Num & Aggregation Strategy			Average
	1 (n/a)	3 Poll	3 Prompt	
Filter(T-L)	91.29	91.22	91.72	91.41
Filter(L-T)	<u>91.67</u>	91.23	91.59	91.50
Sort(L-S)	91.35	91.22	91.57	91.38
Sort(T-S)	90.95	91.09	91.65	91.23
Sort(L-T-S)	91.19	91.21	91.50	91.30
Sort(T-L-S)	91.19	91.00	<u>91.93</u>	91.37
Sort(S-T)	91.50	<u>91.39</u>	91.83	<b>91.57</b>
Tops	-	91.23	91.69	91.46
<b>Average</b>	91.31	91.20	<b>91.69</b>	

## 6.2 Number of Start Sentences & MT Aggregation Strategy

When selecting the start sentences, we need to decide whether to use more than one start sentence. If we choose more than one, we must aggregate each translation result generated by Gradual MT. We investigated two distinct strategies for aggregating the results of Gradual MT. The first method, referred to as *Polling*, selects the MT result with

the highest number of duplicates, drawing inspiration from prior research on self-consistency (Wang et al., 2023). If no duplicates are found, a result is selected randomly. The second method, *Prompting*, involves feeding all Gradual MT results into the LLM as few-shot examples to generate the final MT output. As shown in Table 3, the prompting strategy outperforms polling by 0.49 points. Polling is often worse than using a single start sentence.

## 6.3 Filtering Strategy

We tested three strategies for filtering. The first strategy, ‘Pre-filtering’, aims to minimize costs by evaluating zero-shot translation results with a QE model and applying BridG MT only when necessary (i.e., when the QE score falls below a certain threshold). The second strategy, ‘Post-filtering’, prioritizes maximizing performance by applying BridG MT first and using its output only if the QE score exceeds that of the zero-shot translation. The third strategy, ‘Pre- & Post-filtering’ combines the first two: applying BridG MT when the zero-shot translation’s QE score is below a threshold, and only if BridG MT’s score is higher. We employed CometKiwi, a reference-free QE model, to implement these strategies. To compare them, we analyzed xCOMET scores and score changes of selected outputs, applying the optimal strategies for start selection (Sort(S-T)) and aggregation (Prompting). As shown in Table 4, all strategies improved overall performance, with ‘Post-filtering’ achiev-

Table 4: Average xCOMET scores and score changes of selected outputs when applying the optimal strategies for start selection and aggregation in En-Ko translation. ‘Score change’ is calculated only for the adopted outputs. We also report the number of end sentences for which bridging and Gradual MT is executed (‘# Bridging’) and the number of end sentences for which the BridG MT output is selected over zero-shot translation (‘# Selected Outputs’). The results show that ‘Pre & Post’ reduces the number of bridging by more than half while maintaining nearly the same translation performance.

Filtering Strategy	Avg. Score	Avg. Score Change of Selected Outputs	# Bridging	# Selected Outputs
Zeroshot	89.48	-	-	-
All	91.22	1.74	911 (100%)	911 (100%)
Pre	91.50	4.49	410 (45%)	410 (45%)
Post	<b>92.54</b>	5.61	911 (100%)	497 (54%)
Pre & Post	92.08	<b>8.50</b>	410 (45%)	279 (31%)

ing a notable gain of over 1 point in QE scores compared to zero-shot MT. The results also indicate that ‘Pre- & Post-filtering’ reduces bridging by more than half while maintaining comparable performance to ‘Post-filtering’, offering an effective compromise between computational efficiency and translation quality. Results with every combination of ablation strategies are shown in Table 18.

#### 6.4 Optimal Configuration

Based on our ablation study, we found the optimal strategy to be selecting start sentences using SBERT similarity and tree edit distance (Sort(S-T)), using three start sentences, aggregating Gradual MT results by feeding them as few-shot examples for the final translation, and applying post-filtering. However, given the computational cost, using both pre- and post-filtering can also be a good compromise between efficiency and accuracy.

### 7 Discussion

#### 7.1 Generalizability of Start Sentence Pool

To test the robustness of the start sentence pool, we evaluated BridG MT on the NTREX-128 dataset, which has a different distribution from FLORES-200—the dataset used to construct the start pool. We used Llama-3.1-8B for sentence bridging. The results show that even when the same start pool is used to translate a different dataset, BridG MT still achieves a significant performance improvement. Detailed results are provided in Table 5.

#### 7.2 Balancing Performance and Speed in BridG MT

BridG MT is useful in that it can significantly enhance the translation performance of LLMs. However, the additional inference time introduced by

sentence bridging and Gradual MT can be a drawback in scenarios where fast translation is required. In such cases, it may be more important to prioritize speed, even at the cost of a slight drop in performance. We analyzed the cost-performance trade-offs of three approaches to reduce cost: using a smaller bridging model, applying bridge sampling and pre-filtering. Assuming a setting where sentences are translated one at a time, we measured the time spent on each step of BridG MT and analyzed the inference time and performance across different configurations. A detailed breakdown of the inference time and cost analysis results can be found in Appendix C.

**Smaller Model for Sentence Bridging** Using smaller models can significantly reduce the time required for sentence bridging. As shown in Table 8, even with these smaller models, our approach achieves performance comparable to that of Qwen2-72B-Instruct while reducing the computation time by more than half.

**Bridge Sampling** Bridge sampling is a method to reduce the computational cost by applying Gradual MT only to the first(start), middle, and last(end) sentences along the interpolation path. We tested the bridge sampling using only one start sentence. Bridge sampling reduces the time for Gradual MT by 78% while showing higher performance than baselines. Results are shown in Table 9.

**Pre-Filtering** Pre-filtering reduces the overall computational cost by applying BridG MT only to sentences that the model struggles to translate well. While pre-filtering introduces an additional cost for computing QE scores, this is outweighed by the overall reduction in computational cost. According to our statistics, pre-filtering reduced the overall



Table 5: xCOMET results of BridG MT on the NTREX-128 dataset for various target languages.

MT Model	Setting	De	Ch	Hi	Ko	Sw	Be	Mr
GPT-3.5	Zeroshot	95.10	83.71	67.54	81.85	78.98	62.62	41.55
	BridG MT	<b>95.74</b>	<b>85.48</b>	<b>71.95</b>	<b>84.90</b>	<b>81.49</b>	<b>69.04</b>	<b>48.40</b>
Llama-70B	Zeroshot	94.39	-	75.58	-	-	-	-
	BridG MT	<b>95.50</b>	-	<b>80.03</b>	-	-	-	-
Llama-8B	Zeroshot	91.33	-	64.94	-	-	-	-
	BridG MT	<b>93.79</b>	-	<b>73.47</b>	-	-	-	-
Mistral Nemo	Zeroshot	91.02	74.36	61.13	65.10	32.99	63.66	35.19
	BridG MT	<b>94.74</b>	<b>80.76</b>	<b>68.45</b>	<b>78.33</b>	<b>39.61</b>	<b>72.25</b>	<b>47.19</b>

computation time by 28%. Results are shown in Table 10.

**Results from Cost-Efficient Configuration** We further tested the cost-efficient configuration, in which we use Llama-3.1-8B for both sentence bridging and translation, and applied bridge sampling. We also applied pre-filtering, as we had already confirmed its effectiveness in our ablation study (Section 6.3). The test result on Hindi shows that applying these strategies can reduce inference time by 63%, while compromising performance by only 2.4 points.

Table 6: Comparison between the results of the main experiment (‘Main’) and a more efficient configuration (‘Efficient’). The translation model used is Llama-3.1-8B, and the target language is Hindi.

		Main	Efficient	Zeroshot
Components	Bridging Model	Qwen2 72B	Llama 8B	-
	Pre-filtering	X	O	-
	Path Sampling	X	O	-
	Post-filtering	O	X	-
Avg. Inference Time		46.59	17.19	1.74
xCOMET		78.30	75.88	69.93

### 7.3 Sentence Bridging Analysis

We conducted an analysis of the sentences within the sentence bridges. In particular, we examined whether the sentences within a bridge gradually become more similar to the end sentence on both the source and target sides as illustrated in Figure 1. The analysis revealed that the embedding distance between the bridge sentences and the end sentence progressively decreases on the source side, and a similar trend is observed on the target side, where the embedding distance between the gold translation and the translations generated at each step of Gradual MT also decreases. A detailed analysis is provided in Appendix I.

### 7.4 Quality of a Start Pool

We analyzed the impact of the start pool quality using GPT-3.5 and Mistral-Nemo across four languages: Korean (Ko), Hindi (Hi), Swahili (Sw), and Marathi (Mr). For each language, we selected three distinct start pools from the FLORES-200 dev set, sorted by their translation xCOMET scores: *High*, *Mid*, and *Low*. The *High* start pool consists of the 100 sentences with the highest xCOMET scores. The *Low* start pool includes the 100 sentences with the lowest scores, while the *Mid* start pool is constructed using the 100 sentences located in the middle of the sorted dev set. We conducted BridG MT with a single start sentence, without applying any filtering method. The results show that the quality of the start pool had a degrading effect on the final translation when using Mistral-Nemo, whereas GPT-3.5 demonstrated robustness. Full results can be found in Appendix H.

## 8 Conclusion

In this paper, we proposed BridG MT, a novel method to enhance the machine translation capabilities of various LLMs. BridG MT leverages sentence bridging and gradual MT to guide models, eliciting stronger translation performance. Experimental results across various models and languages demonstrate that our approach consistently improves translation quality, particularly in low-resource languages. Our approach is practical in that it does not require extra training and does not conflict with previous methods that utilize other kinds of prompting techniques.

### Limitations

BridG MT has a limitation in that it introduces computational overhead due to Sentence Bridging and the recursive nature of Gradual MT. While we have tested various methods to minimize the cost

of BridG MT, such as pre-filtering and bridge sampling, future work could focus on further optimizing computational efficiency. Additionally, since the model used for sentence bridging did not perform well in languages other than English, we had to limit our study to cases where English was the source language. Exploring better prompting techniques to interpolate non-English sentences presents a promising direction for future research.

## Acknowledgments

This work was supported in part by the National Research Foundation of Korea (NRF) grant (RS-2023-00280883, RS-2023-00222663); by the National Super computing Center with super computing resources including technical support (KSC-2023-CRE-0176, KSC-2024-CRE-0065); by the Korea Institute of Science and Technology Information (KISTI) in 2025 (No.(KISTI) K25L1M1C1), aimed at developing KONI (KISTI Open Neural Intelligence), a large language model specialized in science and technology; and by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2025-02263754); by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety, 2025의약안003); partially supported by New Faculty Startup Fund from Seoul National University.

## References

- Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. [In-context Examples Selection for Machine Translation](#). *arXiv preprint: 2212.02437*.
- Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. [Tower: An open multilingual large language model for translation-related tasks](#). *Preprint*, arXiv:2402.17733.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Al-lonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yearly, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shao-liang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyan Fu, Whitney Meers,

Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuwei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwon Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damla, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelen, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsim-poukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks,

Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratan-chandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vitor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiao Cheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. [The Llama 3 Herd of Models](#). *arXiv preprint*. ArXiv:2407.21783 [cs].

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

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. [NTREX-128 – news test references for MT evaluation of 128 languages](#). In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. [Quality-aware decoding for neural machine translation](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujia Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. [Exploring human-](#)



- like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.
- Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. *Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine*. *arXiv preprint*. ArXiv:2301.08745 [cs].
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. *The state and fate of linguistic diversity and inclusion in the NLP world*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. *MetricX-23: The Google Submission to the WMT 2023 Metrics Shared Task*. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.
- Hyuhng Joon Kim, Hyunsoo Cho, Junyeob Kim, Taek Kim, Kang Min Yoo, and Sang-goo Lee. 2022. *Self-Generated In-Context Learning: Leveraging Auto-regressive Language Models as a Demonstration Generator*. *arXiv preprint*. arXiv:2206.08082.
- Vladimir I Levenshtein and others. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.
- Junlong Li, Jinyuan Wang, Zhuosheng Zhang, and Hai Zhao. 2024. *Self-Prompting Large Language Models for Zero-Shot Open-Domain QA*. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 296–310, Mexico City, Mexico. Association for Computational Linguistics.
- Hongyuan Lu, Haoran Yang, Haoyang Huang, Dongdong Zhang, Wai Lam, and Furu Wei. 2024. *Chain-of-dictionary prompting elicits translation in large language models*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 958–976, Miami, Florida, USA. Association for Computational Linguistics.
- Xinxi Lyu, Sewon Min, Iz Beltagy, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. *Z-ICL: Zero-Shot In-Context Learning with Pseudo-Demonstrations*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2304–2317, Toronto, Canada. Association for Computational Linguistics.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Curry, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Felipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such,



- Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. [GPT-4 Technical Report](#). [\\_eprint: 2303.08774](#).
- Amy Pu, Hyung Won Chung, Ankur P Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for mt. In *Proceedings of EMNLP*.
- Ricardo Rei, Nuno M. Guerreiro, JosÃ\copyright Pomal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. [Scaling up CometKiw: Unbabel-IST 2023 Submission for the Quality Estimation Shared Task](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 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.
- David Stap and Ali Araabi. 2023. [ChatGPT is not a good indigenous translator](#). In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 163–167, Toronto, Canada. Association for Computational Linguistics.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, 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, Searle 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, and Jeff Wang. 2022. [No Language Left Behind: Scaling Human-Centered Machine Translation](#). *arXiv preprint. ArXiv:2207.04672 [cs]*.
- David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. [Prompting PaLM for Translation: Assessing Strategies and Performance](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). In *The Eleventh International Conference on Learning Representations*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. [Qwen2 Technical Report](#). [\\_eprint: 2407.10671](#).
- Kaizhong Zhang and Dennis Shasha. 1989. Simple fast algorithms for the editing distance between trees and related problems. *SIAM journal on computing*, 18(6):1245–1262. Publisher: SIAM.
- Kexun Zhang, Yee Man Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. [Hire a Linguist!: Learning Endangered Languages with In-Context Linguistic Descriptions](#). [\\_eprint: 2402.18025](#).
- Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. [Automatic chain of thought prompting in large language models](#). In *The Eleventh International Conference on Learning Representations*.
- Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. [Multilingual Machine Translation with](#)

Large Language Models: Empirical Results and Analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

## A Few-shot Examples for Sentence Bridging

In this study, we gave three examples to the bridging model. We arbitrarily picked start and end sentences from the New York Times news articles.

### A.1 Example 1

**Sentence 1:** Surprisingly often, animals show up uninvited at sporting events. Sometimes, it gets a little weird.

**Sentence 2:** D. Cohen tried to parry attacks on his honesty as Todd Blanche, Donald J. Trump’s lawyer, tried to destroy his credibility with jurors.

#### Bridge

1. Surprisingly often, animals show up uninvited at sporting events. Sometimes, it gets a little weird.
2. Often, unexpected events occur during live events, and sometimes these can get quite weird.
3. At public events, such as sports or trials, unexpected guests often cause disturbances, leading to weird situations.
4. In public spectacles, like sports or courtrooms, unexpected participants can disrupt the normal proceedings in strange ways.
5. During public hearings, like in court, surprising accusations and defenses can emerge, often causing odd disruptions.
6. In courtroom battles, unexpected arguments and witnesses can often turn the proceedings weird as each side tries to undermine the other.
7. In legal battles, lawyers frequently confront each other with surprising tactics to challenge credibility, which can make the proceedings seem strange.
8. Michael D. Cohen, during his trial, encountered surprising tactics by Todd Blanche, Donald J. Trump’s

lawyer, aiming to disrupt his credibility in unusual ways.

9. Michael D. Cohen tried to parry attacks on his honesty as Todd Blanche, Donald J. Trump’s lawyer, tried to destroy his credibility with jurors.

### A.2 Example 2

**Sentence 1:** Georgia’s Parliament overrode a presidential veto to give final approval on Tuesday evening to a contentious bill that has plunged the country into a political crisis and threatened to derail the pro-Western aspirations of many Georgians in favor of closer ties with Russia.

**Sentence 2:** At that time, the buildings were abandoned and inaccessible to the public, and the famous ferry that shuttled immigrants from the island to Manhattan sank in a storm in 1968.

#### Bridge

1. Georgia’s Parliament overrode a presidential veto to give final approval on Tuesday evening to a contentious bill that has plunged the country into a political crisis and threatened to derail the pro-Western aspirations of many Georgians in favor of closer ties with Russia.
2. The decision made by Georgia’s Parliament on Tuesday evening, to override a presidential veto and approve a controversial bill, has sparked a political crisis in the country and raised concerns about the impact on the country’s pro-Western stance and relations with Russia.
3. The controversial bill approved by Georgia’s Parliament on Tuesday, despite a presidential veto, has caused a political crisis and stirred debates about the country’s Western alignment and its relation with Russia.
4. The unilateral move by Georgia’s Parliament on Tuesday to push

through a divisive bill despite opposition from the president has led to a political crisis and raised concerns about the nation's Western affiliations and ties with Russia.

5. The recent decision by Georgia's Parliament to defy the presidential veto and pass a controversial bill has created a political turmoil, questioning the country's Western orientation and relationship with Russia.
6. Georgia's Parliament's decision to override the presidential veto and approve the contentious bill has caused a political crisis in the country, prompting discussions on the impact on its Western ties and relations with Russia.
7. Georgia's Parliament's act of rejecting the presidential veto and endorsing the controversial bill on Tuesday has resulted in a political crisis, with implications for the country's pro-Western stance and connections with Russia.
8. Georgia's Parliament moved forward with finalizing a contentious bill on Tuesday, overriding the presidential veto, which has triggered a political crisis and raised concerns about the nation's Western aspirations and relationship with Russia.
9. At that time, the buildings were abandoned and inaccessible to the public, and the famous ferry that shuttled immigrants from the island to Manhattan sank in a storm in 1968.

### A.3 Example 3

**Sentence 1:** The cake has become one of Jamaica's most popular desserts, often found in various stores.

**Sentence 2:** Some plans might cover only generic versions of certain contraceptives, but patients are still entitled to coverage of a specific product that their providers deem medically necessary.

### Bridge

1. The cake has become one of Jamaica's most popular desserts, often found in various stores.
2. Some plans might only cover basic versions of popular desserts like the cake, while others are more comprehensive.
3. In many cases, plans might cover only generic versions of popular products, but specific varieties might still be available.
4. Some plans might cover only generic versions of certain products, with exceptions allowed for specific needs.
5. Some plans might cover only generic versions of certain health-care products, but specific options are still accessible if necessary.
6. Some plans might cover only generic versions of certain contraceptives, but patients can still get specific types if deemed necessary.
7. Some plans might cover only generic versions of certain contraceptives, but patients are still entitled to coverage of a specific product that their providers deem medically necessary.

## B Sample Sentence Bridge

Below is a sample of the bridge generated by Qwen2-72b-Intstruct during the experiment.

### B.1 Sample 1

#### Prompt

Sentence1: I am happy that there are people willing to support me.

Sentence2: "We now have 4-month-old mice that are non-diabetic that used to be diabetic," he added.

#### Output

1. I am happy that there are people willing to support me.
2. It's good to know that support is available when needed.
3. It's reassuring to see support being provided in various situations.

4. He mentioned seeing positive changes in different scenarios.
5. "We have seen positive transformations in our subjects," he added.
6. "We now have improved conditions in our test subjects," he added.
7. "We now have 4-month-old mice that are non-diabetic that used to be diabetic," he added.

## B.2 Sample 2

### Prompt

Sentence1: The death toll is at least 15, a figure which is expected to rise.

Sentence2: The moisture on your hands will react with the outer layers, which will feel funny and form a sort of shell.

### Output

1. The death toll is at least 15, a figure which is expected to rise.
2. The number affected is at least 15, an amount which could increase due to underlying factors.
3. An initial count shows at least 15 impacted, with this figure potentially growing as outer layers are peeled back.
4. A preliminary assessment reveals at least 15 affected individuals, with the outer layers revealing more as they are examined.
5. The moisture on your hands will interact with the outer layers, which may reveal more than initially thought.
6. The moisture on your hands will react with the outer layers, which will feel peculiar and begin to peel away.
7. The moisture on your hands will react with the outer layers, which will feel funny and form a sort of shell.

## C Computational Cost

Tables 7, 8, 9 show the breakdown of the computational cost of BridG MT and the costs of possible scenarios.

## D Prompts and Settings for Translation

### D.1 ChatGPT

ChatGPT(gpt-3.5-turbo-0125) was used via API for translation with the same prompt from OpenAI's official documentation.<sup>7</sup> Temperature and top\_p were set to 0.3 and 1, respectively. The actual prompt is as follows:

**System:** You will be provided with a sentence in English, and your task is to translate it into  $\langle$  Target Language  $\rangle$ .

**User:**  $\langle$  Sentence  $\rangle$

### D.2 Llama-3.1 70B & 8B

Llama-3.1 Instruct models were run on one A6000 GPU, using transformers library. 70B model were 4-bit quantized. Temperature and top\_p were set to 0.6 and 0.9, respectively. The actual prompt is as follows:

**System:** You will be provided with a sentence in English, and your task is to translate it into  $\langle$  Target Language  $\rangle$ .

**User:**  $\langle$  Sentence  $\rangle$

### D.3 Mistral-Nemo-Instruct-2407

Mistral-Nemo-Instruct-2407 was run on one A6000 GPU, using transformers library. Temperature and top\_p were set to 0.6 and 0.9, respectively. The actual prompt is as follows:

**User:** You will be provided with a sentence in English, and your task is to translate it into  $\langle$  Target Language  $\rangle$ .

Sentence:  $\langle$  Sentence  $\rangle$

## E Settings for Sentence Bridging

Qwen2-72B-Instruct model was used for main experiments. It was run on one A6000 GPU with 4-bit quantization using transformers library. Temperature and top\_p were set to 0.6 and 0.9, respectively.

Llama-3.1-8B-Instruct model was used for additional experiments. It was run on one A6000 GPU using transformers library. Temperature and top\_p were set to 0.6 and 0.9, respectively.

## F Full Results with Different Metrics

Table 11, 12, 13, 14, 15, 16 shows full results of our main experiment with different metrics; xCOMET, CometKiwi, MetricX, BLEURT, BLEU and ChrF.

<sup>7</sup><https://platform.openai.com/docs/examples>



Table 7: Breakdown of inference time per component. The statistics are calculated from the FLORES test set into Hindi using Llama-3.1-8b.

Step	Average Computation Time per Sentence
<b>Start Sentence Selection (FAISS)</b>	0.09s (for whole dataset)
<b>Pre-Filtering / Post-Filtering (batch size = 8)</b>	6.22s
<b>Sentence Bridging</b>	
w/ Llama-3.2-3B-Instruct	6.03s
w/ LLaMa-3.1-8B-Instruct	12.94s
w/ Qwen2-72b-Instruct	26.17s
<b>Translation (Llama-3.1-8B-Instruct)</b>	
Zero-shot	1.74s
Gradual MT	10.28s
Gradual MT (w/ bridge sampling)	2.29s
3-shot (for the final output)	2.18s
<b>Baselines</b>	
0 shot	1.74s
30 shot	2.33s
50 shot	2.65s

Table 8: Comparison of sentence bridging models in terms of inference time and average xCOMET score across different MT models and languages.

Bridging Model	Time for Sentence Bridging	Time for Overall Process	Average Score
Qwen2-72B-Instruct	26.17s	44.41s	81.37
Llama-3-1-8B-Instruct	12.94s (49.45%)	31.18s (70.21%)	80.40
Llama-3-2-3B-Instruct	6.03s (23.04%)	24.27s (54.65%)	80.00
0-shot Baseline	1.74s	1.74s	76.35
50-shot Baseline	2.65s	2.65s	78.71

## G Results from Ablation

Table 17 shows the results for every combination of strategies that we explored in the ablation study(6) with En-Ko translation. Table 18 presents the average results and changes in xCOMET scores for each combination of start selection strategies, the number of start sentences, and MT aggregation strategies.

Table 9: Comparison of using the full bridge and bridge sampling in terms of inference time and xCOMET score across different MT models and languages.

Bridge Sampling	Time for Gradual MT	Time for Overall Process	Average Score (xCOMET)
Full Path 3 start sentence	10.28s	44.41s	81.37
Bridge Sampling 1 start sentence	2.29s (22.28%)	36.42s (82.01%)	79.96
0-shot Baseline	1.74s	–	76.35
50-shot Baseline	2.65s	–	78.71

Table 10: Comparison of filtering strategies in terms of time and xCOMET score across different MT models and languages.

Filtering	Time for Overall Process	Average Score (xComet)
post-filtering	46.59s	81.37
pre-filtering	33.46s	80.25
0-shot Baseline	1.74s	76.35
50-shot Baseline	2.65s	78.71

Table 11: Full xCOMET results across different sentence bridging models, translation models and target languages.

	GPT-3.5-turbo-0125							Llama-3.1-70B		Llama-3.1-8B		Mistral-Nemo						
SETTING	DE	ZH	HI	KO	SW	BN	MR	DE	HI	DE	HI	DE	ZH	HI	KO	SW	BN	MR
<b>BASELINE 0 Shot</b>	97.63	91.30	71.89	89.48	81.23	68.73	44.53	97.33	79.41	94.99	69.93	96.70	88.01	66.78	81.43	38.97	71.89	43.99
<b>BASELINE 15 Shot</b>	98.01	92.16	73.13	90.73	81.59	69.70	45.54	97.10	77.95	96.26	73.70	97.68	90.94	69.54	88.71	42.79	73.60	52.26
<b>BASELINE 30 Shot</b>	97.92	91.87	72.98	90.48	81.62	68.58	45.33	96.73	74.67	96.55	73.74	97.72	90.79	69.84	88.06	42.59	74.10	52.52
<b>BASELINE 50 Shot</b>	97.99	91.95	72.85	90.93	82.10	67.72	44.84	96.56	72.68	96.42	73.36	97.72	90.92	69.35	88.29	42.53	73.80	52.04
<b>BridG MT w/ Qwen2-72b</b>																		
Post Filtering	98.04	92.63	77.90	92.57	83.65	75.41	53.11	97.93	84.45	97.25	78.30	97.70	90.99	73.95	89.45	45.67	78.35	57.24
Pre & Post Filtering	97.86	92.35	77.63	92.07	83.34	75.14	51.51	97.73	84.31	96.99	77.37	97.70	90.99	73.82	89.34	45.18	77.83	54.55
<b>BridG MT w/ Llama-3.1-8b</b>																		
Post Filtering	97.97	92.38	77.24	91.85	83.62	74.24	52.27	97.87	83.97	96.93	77.83	97.52	90.40	72.34	87.83	43.33	76.73	52.93
Pre & Post Filtering	97.95	91.54	76.96	91.42	83.52	74.06	50.86	97.73	83.90	96.82	73.09	97.41	90.40	71.19	87.42	43.09	76.23	51.29
<b>BridG MT w/ Llama-3.2-3b</b>																		
Post Filtering	97.97	92.35	76.44	91.48	83.49	74.35	52.13	97.85	83.65	97.14	77.65	97.42	90.04	71.62	86.55	42.30	75.83	51.66
Pre & Post Filtering	97.92	92.29	76.39	91.21	82.74	74.10	52.04	97.61	83.28	96.86	76.84	97.42	90.04	71.46	86.33	42.14	75.75	51.05
<b>QA Decode</b>	97.45	91.63	75.45	90.19	78.26	73.90	45.82	97.37	82.32	88.08	77.67	96.33	87.18	64.42	80.34	28.86	68.59	35.37
<b>MAPS</b>	97.68	92.07	-	-	-	-	-	96.53	-	79.10	-	85.75	88.71	-	-	-	-	-

Table 12: Full CometKiwi results across different sentence bridging models, translation models and target languages.

	GPT-3.5-turbo-0125							Llama-3.1-70B		Llama-3.1-8B		Mistral-Nemo						
SETTING	DE	ZH	HI	KO	SW	BN	MR	DE	HI	DE	HI	DE	ZH	HI	KO	SW	BN	MR
<b>BASELINE 0 Shot</b>	86.12	85.84	69.12	87.73	83.24	67.93	56.23	84.97	74.55	80.06	66.71	83.32	82.90	59.72	78.41	42.72	62.03	52.53
<b>BASELINE 15 Shot</b>	86.71	86.71	69.72	88.85	82.99	69.65	58.37	84.11	72.94	81.64	69.58	85.46	85.73	66.77	87.48	47.51	72.81	63.73
<b>BASELINE 30 Shot</b>	86.44	86.74	69.51	88.86	83.12	68.39	57.71	83.22	70.19	82.17	69.40	85.38	85.74	67.00	86.88	48.08	73.00	63.35
<b>BASELINE 50 Shot</b>	86.61	86.81	69.30	89.18	83.42	67.98	57.32	82.70	68.28	81.95	69.05	85.30	86.04	66.78	87.08	47.61	72.83	63.09
<b>BridG MT w/ Qwen2-72b</b>																		
Post Filtering	87.42	87.64	74.30	90.88	86.09	75.34	66.01	87.25	78.04	85.40	74.00	86.63	86.89	70.96	89.51	54.71	77.65	68.16
Pre & Post Filtering	86.80	87.23	74.12	90.65	85.92	75.09	64.94	86.33	77.96	84.44	73.52	86.63	86.89	70.85	89.48	54.40	77.26	66.82
<b>BridG MT w/ Llama-3.1-8b</b>																		
Post Filtering	87.28	87.42	73.83	90.36	85.76	74.30	64.86	87.16	77.91	84.86	73.25	86.48	86.39	69.42	88.17	51.82	75.79	64.70
Pre & Post Filtering	87.20	86.27	73.66	90.12	85.73	74.07	64.01	86.28	77.82	84.53	70.18	86.11	86.39	68.75	87.98	51.49	75.42	63.82
<b>BridG MT w/ Llama-3.2-3b</b>																		
Post Filtering	87.13	87.38	73.24	90.29	85.58	74.42	64.74	87.19	77.94	85.31	73.30	86.20	85.97	68.62	87.63	50.19	74.83	6371
Pre & Post Filtering	86.91	87.26	73.21	90.08	84.95	74.22	64.67	85.98	77.64	84.41	72.84	86.20	85.97	68.54	87.47	50.00	7480	6349
<b>QA Decode</b>	85.07	85.84	68.52	88.16	82.98	72.29	63.61	84.85	73.31	82.52	69.60	81.93	82.20	55.94	77.56	39.83	60.05	50.49
<b>MAPS</b>	85.61	86.39	-	-	-	-	-	83.48	-	71.19	-	71.06	84.61	-	-	-	-	-

Table 13: Full BLEURT results across different sentence bridging models, translation models and target languages.

SETTING	GPT-3.5-turbo-0125							Llama-3.1-70B		Llama-3.1-8B		Mistral-Nemo						
	DE	ZH	HI	KO	SW	BN	MR	DE	HI	DE	HI	DE	ZH	HI	KO	SW	BN	MR
<b>BASELINE 0 Shot</b>	78.58	73.70	68.31	68.77	75.80	67.95	68.03	77.57	71.34	74.69	66.96	76.27	69.87	60.95	59.13	51.51	62.29	64.20
<b>BASELINE 15 Shot</b>	79.27	74.16	68.49	69.41	75.84	68.46	69.29	77.12	70.61	75.69	68.54	77.95	72.05	66.52	66.94	54.71	70.14	70.75
<b>BASELINE 30 Shot</b>	79.06	74.19	68.75	69.41	76.01	67.87	69.00	76.35	69.32	75.99	68.62	78.07	72.30	66.64	66.56	55.00	69.79	70.91
<b>BASELINE 50 Shot</b>	79.10	74.20	68.74	69.54	76.13	67.82	68.56	76.10	68.68	75.80	68.41	77.89	72.40	66.46	66.49	55.19	70.15	70.46
<b>BridG MT w/ Qwen2-72b</b>																		
Post Filtering	79.11	74.11	69.91	70.28	76.51	70.85	71.22	78.56	72.39	76.98	69.20	77.91	71.38	66.85	67.17	56.50	71.71	71.40
Pre & Post Filtering	78.88	73.99	69.80	70.05	76.40	70.70	70.84	78.17	72.35	76.49	68.92	77.91	71.38	66.82	67.15	56.39	71.41	70.66
<b>BridG MT w/ Llama-3.1-8b</b>																		
Post Filtering	79.13	74.19	69.75	69.99	76.31	70.54	70.65	78.61	72.28	76.75	68.88	77.67	71.04	65.94	66.00	55.16	70.71	70.04
Pre & Post Filtering	79.11	73.85	69.69	69.59	76.24	70.41	70.32	78.19	72.33	76.61	67.85	77.48	71.04	65.58	65.79	55.14	70.42	69.56
<b>BridG MT w/ Llama-3.2-3b</b>																		
Post Filtering	78.94	74.16	69.51	69.62	76.31	70.10	70.61	78.55	72.30	76.93	68.84	77.65	70.94	65.59	65.03	54.61	70.02	69.45
Pre & Post Filtering	78.89	74.10	69.51	69.52	76.08	70.01	70.59	78.07	72.28	76.42	68.69	77.65	70.94	65.53	6492	54.56	70.00	69.33
<b>QA Decode</b>	78.72	74.07	69.83	69.45	76.60	71.63	72.04	78.41	72.67	76.78	70.00	76.11	69.81	60.85	59.18	51.42	61.93	64.32
<b>MAPS</b>	79.08	74.26	-	-	-	-	-	77.40	-	70.81	-	63.18	71.69	-	-	-	-	-

Table 14: Full MetricX results across different sentence bridging models, translation models and target languages.

	GPT-3.5-turbo-0125							Llama-3.1-70B		Llama-3.1-8B		Mistral-Nemo							
SETTING	DE	ZH	HI	KO	SW	BN	MR	DE	HI	DE	HI	DE	ZH	HI	KO	SW	BN	MR	
<b>BASELINE 0 Shot</b>	0.5864	1.1045	1.2189	0.5998	1.3240	2.2832	2.4860	0.6576	0.9091	1.0922	1.3274	0.6904	1.2751	1.6646	1.2064	7.2095	1.9082	2.6305	
<b>BASELINE 15 Shot</b>	0.5428	1.0382	1.1840	0.5156	1.3116	2.0955	2.1551	0.7084	0.9696	0.7956	1.1012	0.5839	1.0667	1.2366	0.5462	6.3982	1.6427	1.6782	
<b>BASELINE 30 Shot</b>	0.5477	1.0515	1.1629	0.5000	1.3188	2.1823	2.2366	0.7742	1.1978	0.7188	1.0863	0.5884	1.0718	1.2482	0.5842	6.3487	1.6230	1.8343	
<b>BASELINE 50 Shot</b>	0.5532	1.0531	1.1319	0.5038	1.2999	2.2167	2.3578	0.7932	1.2621	0.7150	1.0899	0.5820	1.0725	1.2728	0.5568	6.5473	1.6864	1.8806	
<b>BridG MT w/ Qwen2-72b</b>																			
Post Filtering	0.5432	1.0154	0.9179	0.4343	1.1661	1.5156	1.4508	0.5701	0.6854	0.6825	0.8481	0.5639	1.0808	1.0749	0.5559	5.3817	1.1181	1.2351	
Pre & Post Filtering	0.5609	1.0370	0.9264	0.4460	1.1716	1.5285	1.5104	0.6058	0.6895	0.7356	0.8846	0.5639	1.0808	1.0809	0.5606	5.4107	1.1342	1.2881	
<b>BridG MT w/ Llama-3.1-8b</b>																			
Post Filtering	0.5494	1.0534	0.9157	0.4822	1.1922	1.6673	1.5956	0.5868	0.7184	0.7214	0.9050	0.5784	1.1299	1.2340	0.6131	5.8600	1.3074	1.5470	
Pre & Post Filtering	0.5497	1.0896	0.9205	0.4948	1.1984	1.6793	1.6469	0.6022	0.7205	0.7467	1.0581	0.6008	1.1299	1.2725	0.6371	5.8815	1.3304	1.5953	
<b>BridG MT w/ Llama-3.2-3b</b>																			
Post Filtering	0.5471	1.0415	0.9824	0.4673	1.2055	1.6778	1.6031	0.5681	0.7210	0.6756	0.8896	0.6017	1.1462	1.2412	0.7434	6.1683	1.3213	1.6552	
Pre & Post Filtering	0.5532	1.0404	0.9827	0.4759	1.2334	1.6827	1.6037	0.6146	0.7326	0.7168	0.9233	0.6017	1.1462	1.2521	0.7620	6.1826	1.3214	1.6684	
<b>QA Decode</b>	0.5764	1.0660	0.9071	0.4992	1.2059	1.3878	1.2131	0.5948	0.7066	0.7348	0.7422	0.6983	1.2705	1.6804	1.1841	7.2072	1.9488	2.5831	
<b>MAPS</b>	0.5516	1.0451	-	-	-	-	-	0.6939	-	5.9713	-	3.6419	1.1743	-	-	-	-	-	

Table 15: Full BLEU results across different sentence bridging models, translation models and target languages.

	GPT-3.5-turbo-0125							Llama-3.1-70B		Llama-3.1-8B		Mistral-Nemo						
SETTING	DE	ZH	HI	KO	SW	BN	MR	DE	HI	DE	HI	DE	ZH	HI	KO	SW	BN	MR
<b>BASELINE 0 Shot</b>	40.45	45.58	23.06	27.86	32.93	9.99	5.94	38.71	29.09	30.81	21.54	35.70	38.83	17.46	20.17	12.13	8.35	5.31
<b>BASELINE 15 Shot</b>	40.86	46.02	22.98	28.81	33.80	9.67	7.54	37.33	25.54	32.13	22.14	36.89	40.65	20.88	25.08	14.40	11.03	7.30
<b>BASELINE 30 Shot</b>	40.86	46.00	23.08	29.08	33.57	9.10	6.94	36.13	23.29	32.75	22.36	36.96	41.03	21.05	25.45	14.44	11.13	7.45
<b>BASELINE 50 Shot</b>	40.96	45.98	23.51	28.90	33.95	9.75	6.08	35.39	22.84	32.61	22.37	36.72	41.14	21.05	25.28	13.75	11.16	7.69
<b>BridG MT w/ Qwen2-72b</b>																		
Post Filtering	40.87	46.35	23.72	29.01	33.30	11.06	8.00	38.42	28.08	32.43	22.12	35.79	37.76	18.53	23.96	14.09	10.06	7.19
Pre & Post Filtering	40.65	46.05	23.60	28.65	33.18	10.94	7.96	38.68	28.13	32.29	22.19	35.79	37.76	18.52	23.92	13.96	10.13	6.85
<b>BridG MT w/ Llama-3.1-8b</b>																		
Post Filtering	40.86	46.27	23.97	28.87	33.57	10.93	7.94	39.11	28.11	32.23	22.09	35.65	37.77	18.71	23.57	13.39	10.11	6.80
Pre & Post Filtering	40.82	45.65	23.87	28.22	33.46	10.81	7.79	39.08	28.25	32.15	21.97	36.23	37.77	18.51	23.83	13.35	10.07	6.47
<b>BridG MT w/ Llama-3.2-3b</b>																		
Post Filtering	40.79	46.18	23.53	28.66	33.30	10.86	8.42	38.95	28.36	32.37	22.09	36.08	37.49	18.82	22.89	12.96	9.87	6.07
Pre & Post Filtering	40.68	46.01	23.55	28.33	33.19	10.70	8.47	38.79	28.43	32.07	22.20	36.08	37.49	18.73	23.07	12.87	9.84	6.10
<b>QA Decode</b>	41.07	46.02	23.65	28.54	33.59	11.16	8.57	39.56	30.06	34.51	23.66	35.81	38.67	17.46	20.39	12.11	7.98	5.40
<b>MAPS</b>	40.83	46.15	-	-	-	-	-	38.45	-	11.04	-	17.74	40.63	-	-	-	-	-

Table 16: Full ChrF results across different sentence bridging models, translation models and target languages.

	GPT-3.5-turbo-0125							Llama-3.1-70B		Llama-3.1-8B		Mistral-Nemo							
SETTING	DE	ZH	HI	KO	SW	BN	MR	DE	HI	DE	HI	DE	ZH	HI	KO	SW	BN	MR	
<b>BASELINE 0 Shot</b>	66.91	39.28	50.42	34.72	62.20	42.82	37.01	65.77	55.83	60.95	48.38	63.49	33.81	42.18	27.60	41.66	38.18	34.12	
<b>BASELINE 15 Shot</b>	67.16	39.90	50.15	35.58	62.88	42.71	39.71	64.91	52.34	61.64	49.09	64.26	35.68	46.90	32.24	43.19	44.58	39.71	
<b>BASELINE 30 Shot</b>	67.15	39.82	50.71	35.66	62.67	42.09	38.62	64.25	50.53	61.73	49.00	64.31	36.25	47.16	32.39	43.27	44.67	39.77	
<b>BASELINE 50 Shot</b>	67.05	39.78	50.60	35.76	62.80	42.41	37.68	63.73	50.00	61.81	49.12	64.01	36.26	47.13	32.34	42.94	44.50	39.67	
<b>BridG MT w/ Qwen2-72b</b>																			
Post Filtering	67.35	39.76	51.40	35.85	62.65	44.52	40.70	65.72	55.27	61.33	49.45	63.48	32.96	45.63	31.26	43.71	44.01	39.52	
Pre & Post Filtering	67.11	39.51	51.31	35.48	62.56	44.42	40.48	65.76	55.28	61.52	49.30	63.48	32.96	45.61	31.26	43.68	43.96	39.04	
<b>BridG MT w/ Llama-3.1-8b</b>																			
Post Filtering	67.24	39.69	51.53	35.61	62.62	44.31	40.34	66.16	55.26	61.63	49.39	63.49	32.90	45.24	30.70	42.99	43.52	38.10	
Pre & Post Filtering	67.20	39.31	51.44	35.04	62.56	44.24	40.18	66.01	55.32	61.61	48.99	63.81	32.90	45.05	30.97	43.05	43.52	37.87	
<b>BridG MT w/ Llama-3.2-3b</b>																			
Post Filtering	67.18	39.64	51.32	35.37	62.50	44.20	40.83	66.06	55.44	61.54	49.33	63.65	32.83	45.05	30.33	42.87	43.05	38.10	
Pre & Post Filtering	67.17	39.47	51.33	35.09	62.44	44.13	40.82	65.92	55.57	61.45	49.34	63.65	32.83	44.94	30.51	42.88	43.04	38.06	
<b>QA Decode</b>	67.26	39.65	51.19	35.31	62.68	44.28	41.06	66.47	56.82	62.92	50.60	63.47	33.81	41.99	27.75	41.76	38.07	34.05	
<b>MAPS</b>	67.29	39.69	-	-	-	-	-	65.17	-	47.66	-	54.36	34.68	-	-	-	-	-	

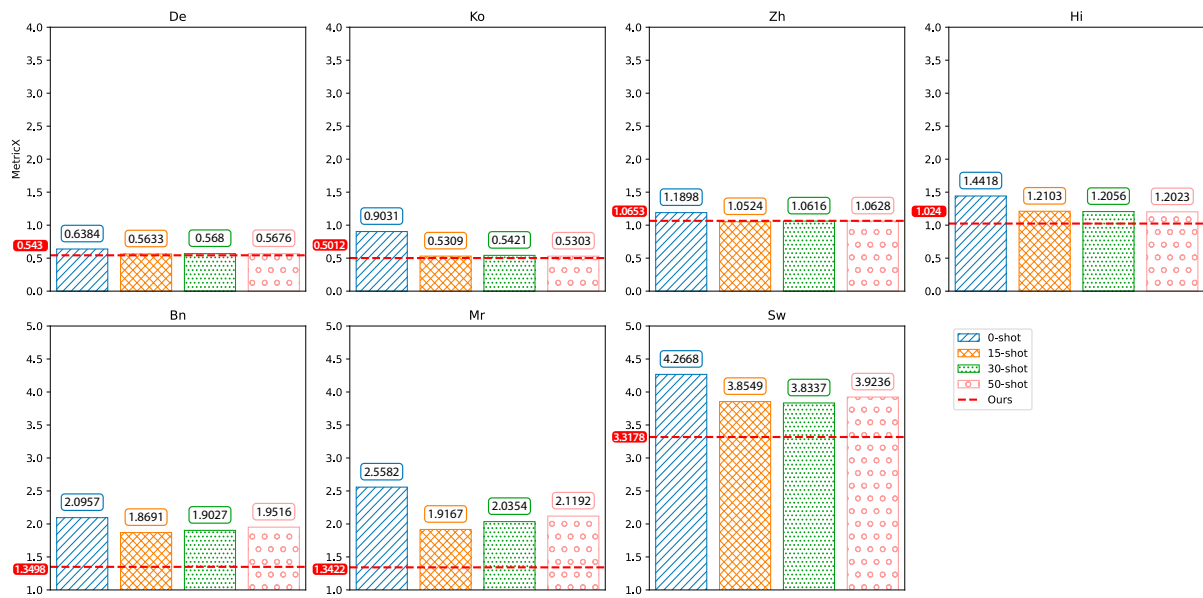


Figure 4: MetricX scores for few-shot translations (bars) and BridG MT translations (horizontal line). Lower score indicates better performance. Both BridG MT and few-shot scores represent the average MetricX scores of GPT-3.5 and Mistral-Nemo.



Table 17: Full results on ablation study with En-Ko translation task. All scores are measured with DA score by xCOMET. ‘Sort’, ‘Filter’, and ‘Tops’ denote the start sentence selection methods. Each letter in parentheses denotes a similarity metric, with the order indicating the priority of the metrics. ‘L’ stands for Levenshtein distance, ‘T’ stands for tree edit distance, and ‘S’ stands for SBERT similarity. The highest results for each start selection strategy are highlighted in bold, while the second-highest results are underlined.

Start Selection	Filtering	Aggregation			Baseline (3shot)
		None	Poll	Prompt	
Sort (S-T)	All	90.69	90.42	91.22	90.26
	Pre	91.31	91.14	91.50	
	Post	<u>92.29</u>	92.18	<b>92.54</b>	
	Pre&Post	91.70	91.84	92.08	
Sort (T-S)	All	89.93	90.02	90.93	90.57
	Pre	90.31	90.71	91.24	
	Post	91.81	91.97	<b>92.47</b>	
	Pre&Post	91.74	91.64	<u>91.98</u>	
Sort (L-S)	All	90.38	90.27	90.86	90.27
	Pre	91.10	90.95	91.12	
	Post	<u>92.08</u>	91.96	<b>92.41</b>	
	Pre&Post	91.83	91.68	91.91	
Sort (T-L-S)	All	90.17	89.99	91.32	90.44
	Pre	90.54	90.29	91.53	
	Post	92.06	91.90	<b>92.69</b>	
	Pre&Post	92.00	91.81	<u>92.20</u>	
Sort (L-T-S)	All	90.13	90.11	90.77	90.38
	Pre	90.69	90.85	91.13	
	Post	91.99	<u>92.05</u>	<b>92.24</b>	
	Pre&Post	91.93	91.82	91.86	
SBERT Filter + Sort (T-L)	All	90.36	90.23	91.07	90.35
	Pre	91.04	90.88	91.30	
	Post	92.07	<u>92.10</u>	<b>92.53</b>	
	Pre&Post	91.70	91.69	92.00	
SBERT Filter + Sort (L-T)	All	90.92	90.55	91.19	90.06
	Pre	91.46	90.98	91.23	
	Post	<u>92.37</u>	92.29	<b>92.61</b>	
	Pre&Post	91.95	91.12	91.34	
Tops	All	-	90.12	90.98	90.10
	Pre	-	90.65	91.30	
	Post	-	<u>92.10</u>	<b>92.46</b>	
	Pre&Post	-	92.06	92.01	

Table 18: Average xCOMET scores and score changes of selected outputs when applying *every combination* of strategies for start selection and aggregation in En-Ko translation. ‘Score change’ is calculated only for the adopted outputs. ‘All’ selects every output. ‘Pre-filtering’, denoted as ‘Pre’, applies zero-shot translation first and uses BridG MT only when the CometKiwi score is below a threshold. ‘Post-filtering’, denoted as ‘Post’, selects outputs only if they outperform zero-shot translations. ‘Pre- & Post- filtering’, denoted as ‘Pre & Post’ combines these two strategies. We also report the number of end sentences for which bridging and Gradual MT is executed (‘No. of Interpolated End Sents’) and the number of end sentences for which the BridG MT output is selected over zero-shot translation (‘No. of Selected Outputs’). The results show that ‘Pre & Post’ reduces the number of bridging by more than half while maintaining nearly the same translation performance.

Filtering Strategy	Avg. Score	Avg. Score Change of Selected Outputs	No. of Interpolated End Sents (%)	No. of Selected Outputs (%)
Zeroshot	89.48	-	-	-
All	90.55	1.06	21864 (100%)	21864 (100%)
Pre	91.01	3.70	10468 (48%)	10468 (48%)
Post	<b>92.21</b>	5.30	21864 (100%)	11270 (52%)
Pre&Post	91.81	<b>7.98</b>	10468 (48%)	6664 (30%)

## H Evaluation Results of BridG MT with Varying Start Pool Quality

Figure 5 shows xCOMET scores of BridG MT when using start pools of varying quality.

## I Analysis On Sentence Bridging

Below we present the analysis we conducted on sentence bridges. We analyzed sentence bridges from En-Ko translation results.

### I.1 Progresses of Bridging.

We examined whether LLMs genuinely bridge start and end sentences or simply generate random sentences. To assess this, we use SBERT to embed the sentences in bridge and calculate their Euclidean distances from the end sentence. If these distances generally decrease, it indicates successful bridging. To measure this, we defined *progress* as  $\text{progress} = d_{n-1,e} - d_{n,e}$  where  $d_{n,e}$  denotes the euclidian distance between  $n^{\text{th}}$  sentence in each bridge and end sentence.

As shown in Figure 6, the average progress of each bridge is generally positive on both the source(English) and target(Korean) sides, indicating that the sentence bridging effectively connects the two sentences. The mean and standard deviation of average progress across bridges were 0.12 and 0.27 for English, and 0.06 and 0.031 for Korean, respectively.

### I.2 Visualization of Bridges

In addition to the analysis on progress, we conducted visualizations on both the source and target sides to examine whether our experimental results align with the intuition illustrated in Figure 2. Specifically, we visualized the embeddings of the sentences within the sentence bridges and the Korean translations generated at each step of Gradual MT. For English sentences, we used all-mpnet-base-v2 to obtain embeddings, while for Korean sentences, we used intfloat/multilingual-e5 (Wang et al., 2024). On the Korean side, we additionally plotted 50-shot example translations to examine whether the intermediate translations produced at each Gradual MT step gradually move closer to the embedding of the gold translation, and how they compare to the few-shot examples. The results of this analysis show that sentence bridging effectively generates intermediate sentences that bridge the start and end sentences on the source side. Furthermore, we frequently observed that the

intermediate translations from Gradual MT also move progressively closer to the gold translation. These findings support the intuition underlying our approach. The visualizations are shown in Figure 7 and Figure 8.

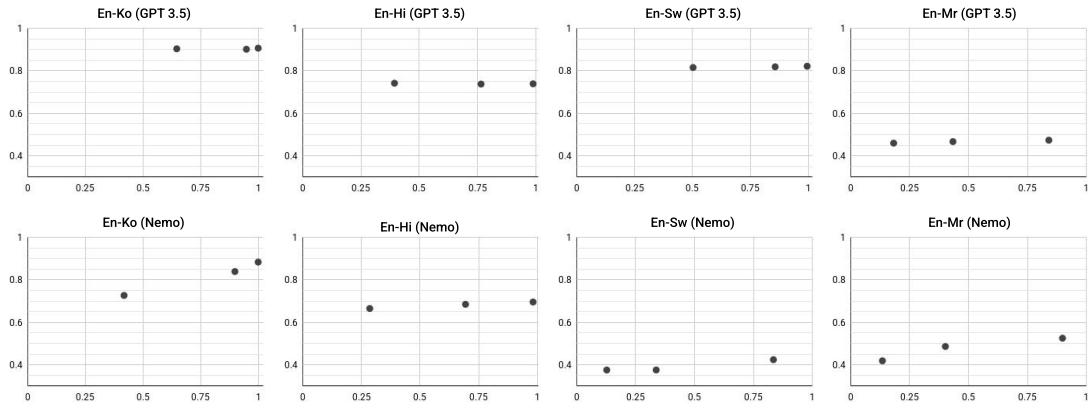


Figure 5: xCOMET scores of BridG MT when using start pools of varying quality. In each plot, the horizontal axis represents the xCOMET score of the start pool, while the vertical axis represents the xCOMET score of the final output. A single start sentence was used for evaluation, without the application of any filtering method.

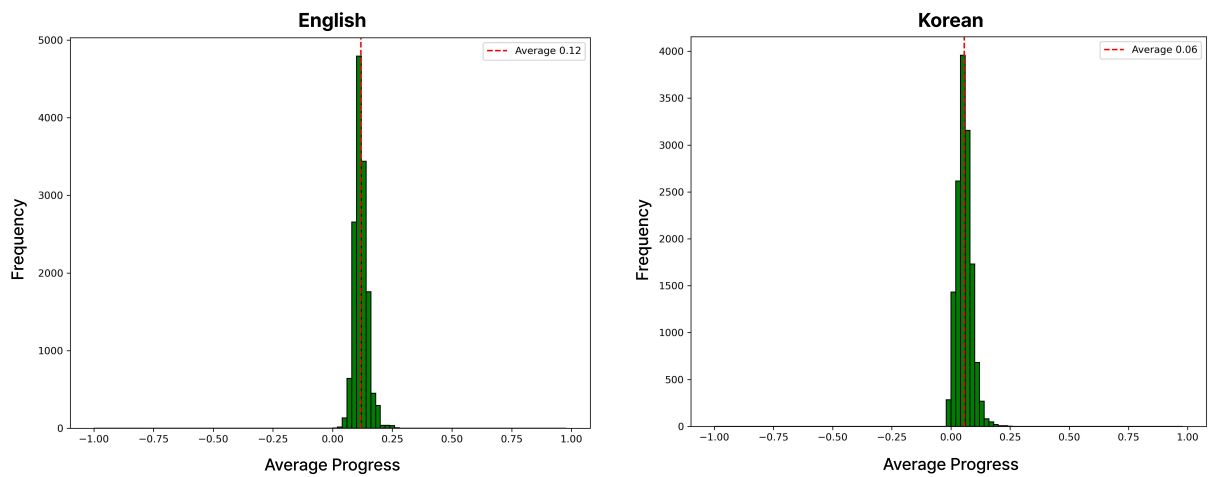


Figure 6: Distribution of average progress from each sentence bridge. The sentence bridges are drawn from the English-to-Korean BridG MT results. Progress indicates how much each sentence in the bridge moves closer to the target sentence in terms of Euclidean distance.

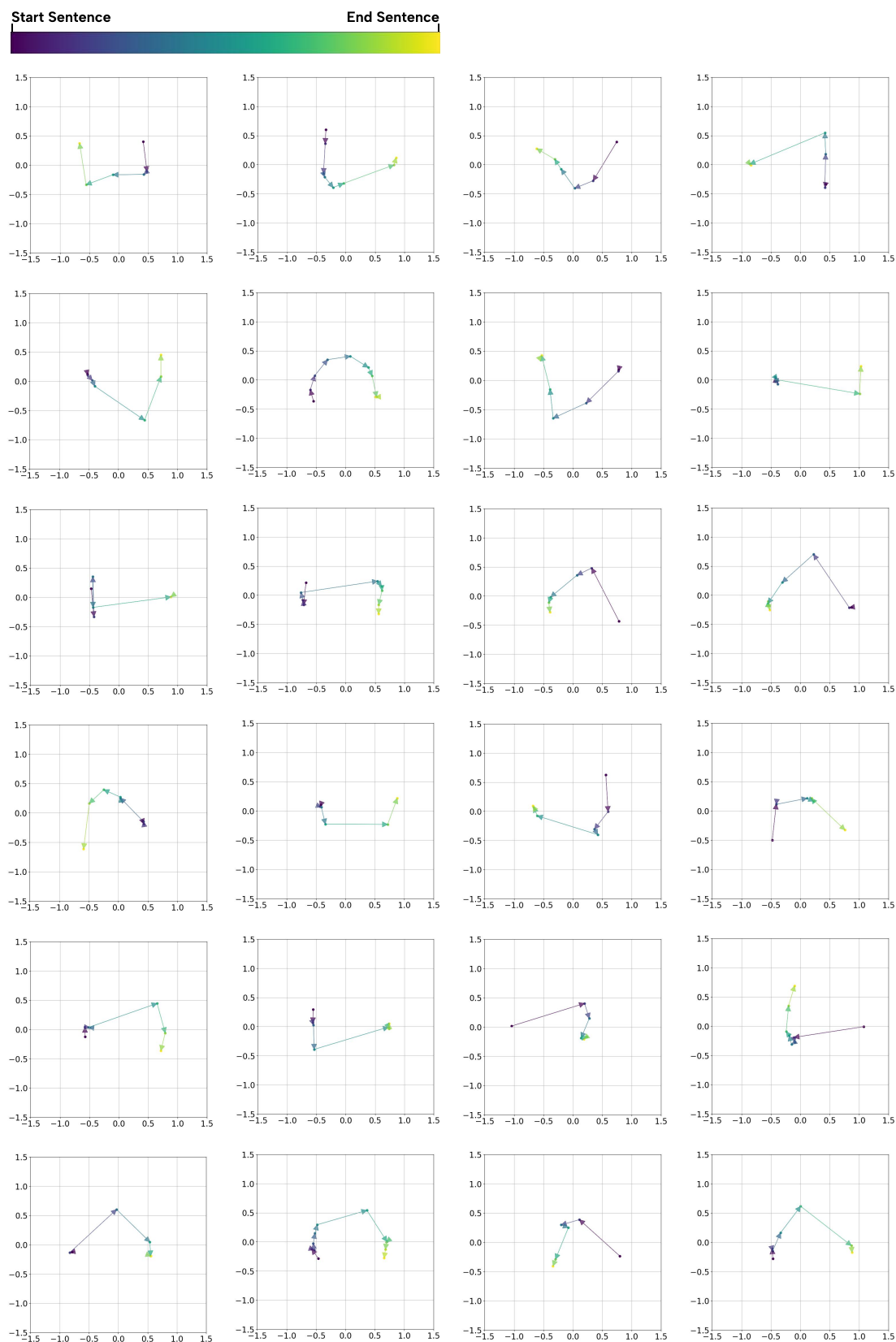


Figure 7: 24 samples of 2D scatter plots of embeddings from sentence bridges. The plots are projected from SBERT embeddings onto a 2D plane using PCA. The X and Y axes of each plot represent the first and second principal components, respectively. Arrows in each plot show the trajectory of sentence shifts from the start sentence (blue-colored dot) to the end sentence (yellow-colored dot).



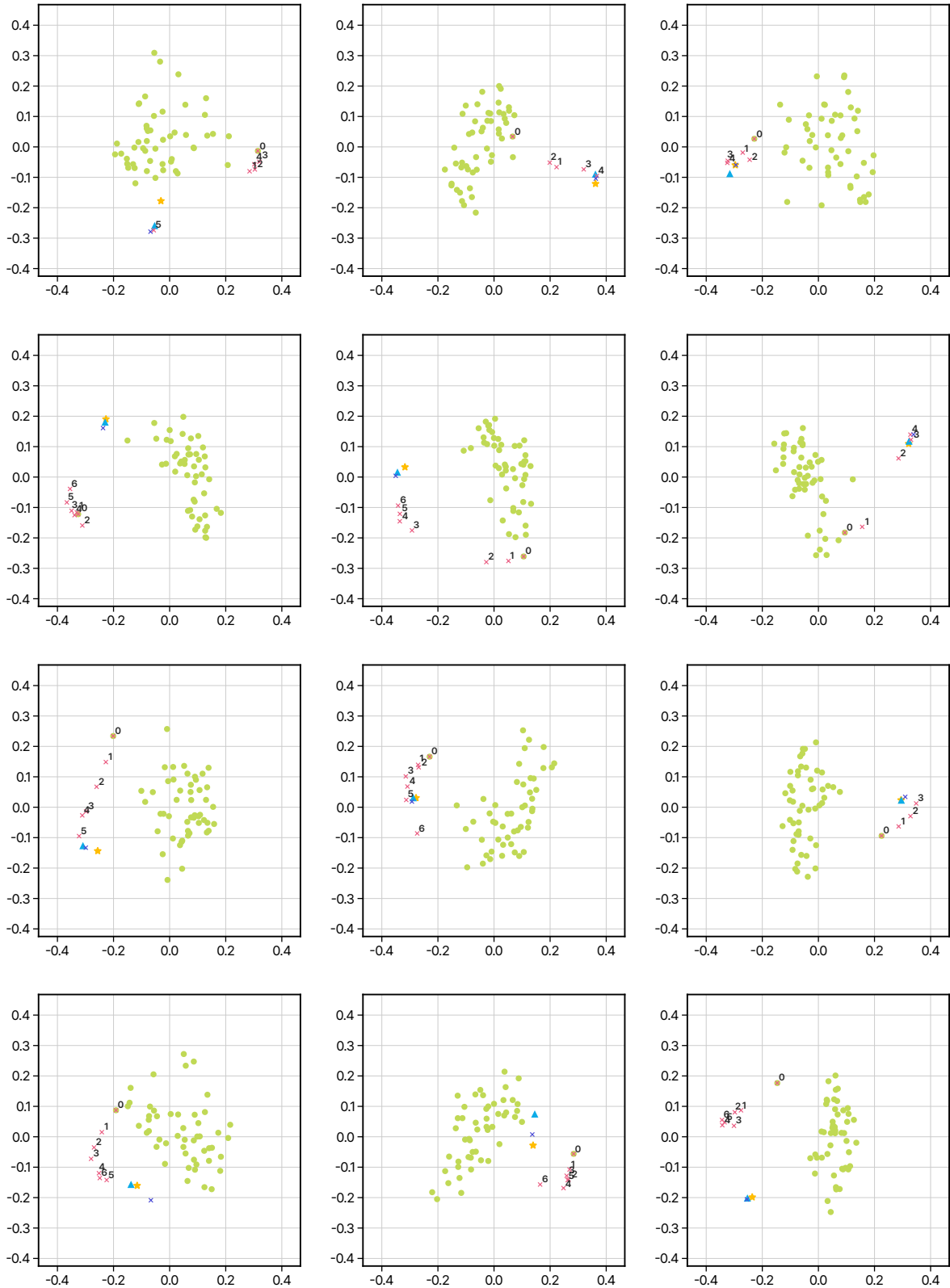


Figure 8: 12 2D scatter plots of embeddings from Gradual MT. Sentence embeddings from the intfloat/multilingual-e5-large model are projected onto a 2D plane using PCA. The X and Y axes represent the first and second principal components, respectively. Green dots and sky-blue triangles indicate the few-shot examples and 50-shot MT outputs. Purple and blue X marks denote the translations from each step and the final translation generated by Gradual MT.