A Single Model Ensemble Framework for Neural Machine Translation using Pivot Translation

Seokjin Oh[1†], **Keonwoong Noh**[2], **Woohwan Jung**[2*]
[1]SK Siltron, [2]Korea University
seokjin.oh@sk.com, {nohkw011, woohwan}@korea.ac.kr

Abstract

Despite the recent remarkable advances in neural machine translation, translation quality for low-resource language pairs remains subpar. Ensembling multiple systems is a widely adopted technique to enhance performance, often accomplished by combining probability distributions. However, previous approaches face the challenge of high computational costs for training multiple models. Furthermore, for black-box models, averaging token-level probabilities at each decoding step is not feasible. To address the problems of multi-model ensemble methods, we present a pivot-based single model ensemble. The proposed strategy consists of two steps: pivot-based candidate generation and post-hoc aggregation. In the first step, we generate candidates through pivot translation. This can be achieved with only a single model and facilitates knowledge transfer from high-resource pivot languages, resulting in candidates that are not only diverse but also more accurate. Next, in the aggregation step, we select *k* high-quality candidates from the generated candidates and merge them to generate a final translation that outperforms the existing candidates. Our experimental results show that our method produces translations of superior quality by leveraging candidates from pivot translation to capture the subtle nuances of the source sentence.

Keywords: Neural Machine Translation, Pivot Translation, Ensemble Method, Low-resource Languages

1. Introduction

Neural machine translation (NMT) models exhibit outstanding capabilities when a large volume of the parallel corpus is available (e.g., translating from and to English). However, their performance still falls short in cases involving low-resource languages (e.g., Basque) and translating between non-English languages from different language families (e.g., German-Russian) (Artetxe et al., 2018). Top-performing large language models (LLMs), such as GPT models (Ouyang et al., 2022), also demonstrate suboptimal translation performance in low-resource language pairs (Robinson et al., 2023; Moslem et al., 2023). The scarcity of parallel data, primarily due to limited cultural interaction, makes the low-resource translation task more challenging.

In many generation tasks, ensembling multiple systems has proven to be a successful strategy for performance enhancement. In NMT, traditional ensemble methods average probability distributions over output tokens from multiple models during decoding. However, the high cost of training multiple models is the primary shortcoming of ensemble decoding. Additionally, computing token-level probabilities at each decoding step is not feasible with recent black-box models such as GPT-4o and Gemini (OpenAI, 2024; Team, 2023).

Ensemble methods that can be utilized even when token-level probabilities cannot be computed have also been proposed. A selection-based ensemble method involves generating candidates from multiple models and then selecting the best candidate among them (Wang et al., 2022; Hendy et al., 2023). However, in this ensemble fashion, the final output space is limited to the existing candidate pool. In contrast, the generation-based ensemble, such as LLM-Blender (Jiang et al., 2023), creates improved outputs using candidates obtained from multiple models. This approach aims to generate a final output superior to the existing candidates. Nonetheless, the main drawback of the notably high cost of generating candidates through multiple models remains, inducing computational overhead. As the size of the models used in the ensemble increases, the cost proportionally escalates, becoming more burdensome. In addition, due to the varying performance of MT systems, the quality of some candidates can be significantly lower than that of others, leading to a degradation in the overall performance.

To alleviate the problems above, we propose **Pivot**-based single model **E**nsemble (PIVOTE), a novel generation-based approach. Our intuition of a single model ensemble primarily stems from pivot translation, which can produce diverse and more accurate translations. Pivot translation (Wu and Wang, 2007; Utiyama and Isahara, 2007) is a method that splits the end task into two sequential steps: source→pivot and pivot→target. Pivoting has been employed to enhance low-resource translation by transferring knowledge from high-resource pairs. In many cases, English, being a

[†] Work done while at Hanyang University.
[*] Corresponding author.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8520–8534
11–16 May 2026.

resource-rich language, serves as the intermediate language. However, we employ not only English but various pivot languages for candidate generation, thereby producing diverse hypotheses using a single model.

In the next aggregation step, we select the top candidates for the ensemble and merge them to generate the final output. Since the quality of candidates directly impacts the results of the ensemble, it is important to select high-quality candidates. Given that the best pivot language for translation varies with each source sentence, we select the top-k candidates for each source sentence via quality estimation (QE). By leveraging diverse candidates from pivot translation and knowledge of the merging module, PIVOTE generates final translations that accurately convey the meaning and subtle nuances of the source sentence, superior to selecting from pre-existing candidates. Our contributions can be summarized as follows:

- We propose a simple but effective pivot-based single model ensemble method, PIVOTE, to improve low-resource MT.
- We show that a single model can effectively generate diverse and accurate hypotheses and that leveraging these candidates in an ensemble process can enhance translation quality while reducing computational overhead.
- The empirical results on various language pairs demonstrate that we consistently outperform state-of-the-art methods, validating the effectiveness of the pivot-based ensemble.

2. Related Work

Pivot-based approaches. Pivot translation is an approach that decomposes the translation task into two sequential steps (Wu and Wang, 2007; Utiyama and Isahara, 2007). By transferring knowledge from high-resource pivot languages, pivoting is especially effective in translation between low-resource languages (Zoph et al., 2016; Aji et al., 2020; He et al., 2022). In this study, pivot translation enables us to obtain high-quality candidates for the ensemble. Kim et al. (2019) discusses a pivot-based transfer learning technique where source→pivot and pivot→target models are first trained separately, then use pre-trained models to initialize the source→target model, allowing effective training of a single, direct NMT model. Zhang et al. (2022) further investigates the transfer learning approach by utilizing auxiliary monolingual data.

Pivot translation typically employs English as the bridge language. Nonetheless, previous studies have explored the use of diverse pivot languages, taking into account factors such as data size and the relationships between languages (Paul et al., 2009; Dabre et al., 2015). By leveraging the ability of pivot translation to produce diverse outputs, several studies have focused on generating paraphrases (Mallinson et al., 2017; Guo et al., 2019). More recently, Mohammadshahi et al. (2024) uses pivot translation for an ensemble, but it requires computing token-level probabilities and fails to improve translation. Our work shares motivation with these studies, generating translations depending on the pivot path to obtain a variety of candidates.

Ensemble in NLG tasks. Ensemble learning is a widely adopted strategy to obtain more accurate predictions by employing multiple systems (Sagi and Rokach, 2018). In NMT, the traditional approach involves averaging the probability distributions of the next target token, which is predicted at each decoding step by multiple models (Bojar et al., 2014) or by different snapshots (Huang et al., 2017). When multiple sources are available, an ensemble can be conducted with predictions obtained from different sources (Firat et al., 2016). A token-level ensemble through vocabulary alignment across LLMs has also been proposed (Xu et al., 2024). However, these methods are not applicable to recent black-box models as they cannot compute token-level probabilities at decoding time.

A selection-based ensemble has also been explored, which chooses the final output among the existing candidates. This can be achieved through majority voting by selecting the most frequent one (Wang et al., 2022) or selecting the best candidate with QE (Fernandes et al., 2022; Hendy et al., 2023). Recently, MBR decoding (Goel and Byrne, 2000; Farinhas et al., 2023), which aims to find the hypothesis with the highest expected utility, has gained attention. However, this approach limits the final output space to the existing candidate pool.

On the other hand, the generation-based ensemble method involves generating a new final prediction. Fusion-in-Decoder (Izacard and Grave, 2021) proposes an architecture that aggregates additional information with a given input. More recently, within the context of LLMs, Jiang et al. (2023) and Yin et al. (2023) investigate a method of using LLMs to generate multiple outputs and aggregate them. Generating new outputs through LLMs offers the benefit of explicitly harnessing their pre-trained knowledge within the ensemble process.

3. Pivot-based Single Model Ensemble

In this section, we first introduce a overview of PIVOTE framework (§3.1). Then, we describe the candidate generation process through pivot translation (§3.2) and the aggregation process (§3.3).

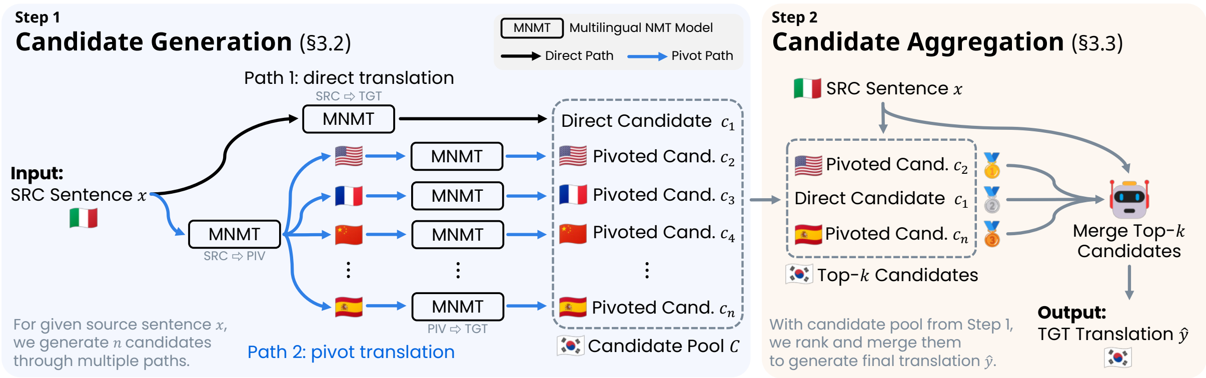


Figure 1: Overview of PIVOTE framework.

3.1. Overview

Our objective is the same as that of conventional translation tasks: converting the given source language sentence x into a target language sentence $\hat{y}$. PIVOTE consists of two steps: candidate generation and candidate aggregation. Figure 1 illustrates an overview of the proposed ensemble framework.

As the first step, we input x to generate candidates through a single multilingual NMT model. One translation path could be directly translating from the source to the target through the source→target path. Alternatively, pivot translations can be achieved by employing high-resource pivot languages, enabling translation paths of source→pivot and pivot→target. During the pivot process, leveraging abundant parallel data enables knowledge transfer from high-resource pivot languages, thereby facilitating the generation of diverse and more accurate translations. Through these n paths, we can obtain a candidate pool $C = \{c_1, ..., c_n\}$ composed of n candidates in the target language, employing only a single model.

As the second step, a ranking process is first conducted within the candidate pool C since not all candidates contribute to the ensemble. Using the estimated quality of each candidate, we select the top-*k* candidates. We then generate the final output $\hat{y}$ using the selected high-quality candidates. This generation-based approach facilitates the production of outputs superior to the existing candidates.

3.2. Pivot-based Candidate Generation

In the first step, PIVOTE takes a source sentence x as input and generates n candidates. Direct translation yields only one candidate, whereas pivot translation enables the generation of multiple candidates from a single source sentence using a single model. Generating candidates through pivot translation has two major advantages: diversity and quality.

First, we can obtain diverse candidates that can act complementarily. One of the key principles of the ensemble is that the participants must be sufficiently diverse to provide various inductive biases. In PIVOTE, each source sentence is translated diversely by passing through multiple translation paths. Diverse translation paths enhance the likelihood of providing expressions that convey the accurate meaning of the source sentence. Pivot-based candidate generation shares a similar goal with a previous study that generates paraphrases through round-trip translation, aiming to generate diverse translations (Thompson and Post, 2020).

Second, by utilizing a parallel corpus of high-resource pivot languages, pivoting enables more accurate translations. For low-resource language pairs, more appropriate translations can be achieved through two-step decoding via a pivot language (He et al., 2022). Moreover, leveraging pivot languages with abundant parallel data, not limited to English, allows us to obtain better translations (Paul et al., 2009; Dabre et al., 2015).

In addition, pivot translation with a single model offers practical benefits over employing multiple models. Firstly, it can reduce the costs of operating multiple models, including LLMs. Secondly, the substantial performance disparities among models mean that using the top-performing single model for candidate generation often leads to higher-quality outcomes. Lastly, it reduces inference latency by using a single model for batched inferences, while multi-model ensembles require up to 11 models, causing significant overhead and limiting real-time response capability. Given that pivot translation with a single model allows for the creation of diverse and more accurate translations, we utilize an MNMT model to generate the candidates.

Selecting pivot languages. For each language pair, we carefully select pivot languages based on the assumption that pivot languages with abundant mutual knowledge would allow us to obtain higher-quality candidates. We select n top-performing paths for our study based on BLEU scores on the FLORES-200 benchmark (Costa-jussà et al., 2022). We evaluate the outputs for each path, including direct translation and through various pivot translations. NLLB (Costa-jussà et al., 2022) is used to generate candidates, and results on the FLORES-200 for selecting translation paths are in Appendix A. If pivot languages are selected based on BLEU scores, high-resource languages are predominantly chosen, rather than low-resource ones. The experiments detailed in Appendix B demonstrate that overly prioritizing diversity by employing low-resource pivot languages, at the expense of candidate quality, does not result in improvements in the final translation. The experiments comparing metrics for selecting translation paths are in Appendix C. As a result, we compose the candidate pool using the 4 paths.

3.3. Candidate Aggregation

In the aggregation step, we take the candidate pool C as input and output the merged final translation $\hat{y}$. The post-hoc aggregation process encompasses two stages: selecting and merging. In the first stage, we select candidates by a ranking method. There are two approaches to selecting candidates. One approach evaluates each translation path and selects the best paths for all source sentences. The other approach involves selecting the best top-*k* candidates for each source sentence. After selecting *k* candidates, we generate the final translation $\hat{y}$ using the merging module. This process enables the creation of better outputs beyond the quality of existing candidates.

Selecting the top-*k* candidates. The pivot language that generates the highest-quality candidate varies for each source sentence. The best output is not guaranteed from one translation path alone, as it can vary depending on factors such as the size of the parallel corpus and the relationship between languages. First, PIVOTE uses QE to rank all n candidates from candidate pool $C = \{c_1, ..., c_n\}$. Afterward, we select top-*k* candidates from n candidate pool. Selecting the top-*k* candidates ensures the quality of the output by filtering out low-quality candidates while also efficiently reducing the cost during the merging process. We use the reference-free COMETkiwi (*wmt22-COMETkiwi-da*) (Rei et al., 2022b) for ranking candidates.

Generating the final translation. We generate the final translation $\hat{y}$ by merging the top-*k* candidates using two methods: LLM-based approaches and encoder-decoder ensemble architectures.

LLM-based approaches offer the advantage of implicitly leveraging their translation capabilities during the ensemble, as the source sentence is also provided. We conduct experiments with GENFUSER (Jiang et al., 2023), Llama-3 (AI@Meta, 2024), and GPT models (Ouyang et al., 2022; OpenAI, 2023, 2024). When employing GENFUSER, we construct the input by concatenating the top-k candidates to the prompt, as presented in Jiang et al. (2023). For merging with Llama-3 and GPT, we use the ensemble prompt template in Appendix D.

On the other hand, encoder-decoder architectures employ smaller models dedicated to the ensemble, effectively reducing both training and inference costs. We conduct experiments using Fusion-in-Decoder (FiD) (Izacard and Grave, 2021) and TRICE (Huang et al., 2021). For FiD, the instruction and the source sentence are concatenated with each candidate and independently encoded. The decoder then takes the concatenated representations and generates the final translation. For TRICE, the model is trained with a two-stage fine-tuning method. In the first stage, the model is trained on two different inputs and a single target: Source→Target and Candidate→Target. In the second stage, the source and the candidate are concatenated and provided as a single input. The overall input formats are: `Translate source into <target language> referring <target language> candidate. source:` $<x>$ `candidate:` $<c_k>$ for FiD, and $<x></s><l_s>;<c_1></s><l_t>;\dots;<c_k></s><l_t>$ with the language token $<l_{lang}>$ for TRICE. We provide architectural illustrations of these encoder-decoder ensemble architectures in Appendix E.

By leveraging various candidates, each with different strengths, the aggregation process can effectively mitigate errors in a complementary manner.

4. Experiments

We use NVIDIA RTX 3090 or 4090 GPUs for experiments.

4.1. Datasets

Table 1: Dataset statistics.

Lang-pair	Dataset	# Sentences		
		Train	Dev	Test
KO ↔ IT	TED 2020 v1 (Reimers and Gurevych, 2020)	357,733	2,000	2,000
AR ↔ PT	WikiMatrix v1 (Schwenk et al., 2019)	153,441	2,000	2,000

We conduct experiments on the linguistically distant languages within pairs: not in the same language family and using different scripts. We select 2 language pairs, resulting in 4 translation directions in total, Korean (*Koreanic*)↔Italian (*Romance*) and Arabic (*Arabic*)↔Portuguese (*Romance*). The language family grouping is defined by the criteria presented in Fan et al. (2020).

We validate our approach across various domains. For Korean↔Italian pair, we run experiments on TED2020 (Reimers and Gurevych, 2020). For Arabic↔Portuguese pair, we use WikiMatrix (Schwenk et al., 2019). All the datasets are obtained from the OPUS[1] (Tiedemann, 2012) project. The statistics for the datasets are listed in Table 1.

4.2. Evaluation Metrics

We assess the translation quality using BLEU (Papineni et al., 2002), chrF++ (Popović, 2017), and reference-based COMET (*wmt22-COMET-da*) (Rei et al., 2022a). For reporting BLEU, *SacreBLEU* (Post, 2018) is used with the `ko-mecab` tokenizer for Korean and `13a` tokenizer for the others.

4.3. Baselines

As an encoder-decoder NMT model, we use NLLB-200-distilled-600M (Costa-jussà et al., 2022). When training NLLB, we use the Transformers library from HuggingFace (Wolf et al., 2020). AdamW optimizer (Loshchilov and Hutter, 2019) is used with a learning rate of $2e{-}5$, batch size of 2, and dropout with a probability of 0.1. When validation BLEU did not improve for 3 checkpoints, with 30k steps between them, we stopped training.

For open-source LLMs, we use Vicuna 13B (Chiang et al., 2023), Baize 13B (Xu et al., 2023a), and Llama-3-8B-Instruct (AI@Meta, 2024) as baselines. We fine-tuned these LLMs with QLoRA (Dettmers et al., 2024); r=16, α=64, dropout=0.1 for all linear layers. For black-box LLMs, we use GPT-4 (OpenAI, 2023) and GPT-4o (OpenAI, 2024). The versions `gpt-4-1106-preview` and `gpt-4o-2024-08-06` are employed for GPT-4 and GPT-4o, respectively. For GPT models, *temperature* is set to 0.0 for stable responses (Peng et al., 2023) and *top_ p* is set to 0.1 to ensure reproducibility. For LLMs, we use the zero-shot prompt template of Hendy et al. (2023), as presented in Appendix D.

As state-of-the-art ensemble baselines, we employ LLM-Blender (Jiang et al., 2023), EVA (Xu et al., 2024), and MBR (Farinhas et al., 2023). For LLM-Blender and EVA, we fine-tuned the same open-source LLMs used in each study with the train data in Table 1. The list of the LLMs is in Appendix F. *temperature* is set to 0.1 to mitigate hallucination for low-resource pairs (Guerreiro et al., 2023). For MBR, we generate a set of 5 hypotheses using GPT-4. When generating hypotheses, *temperature* was set to 0.0 for its optimal performance, based on the results of our pilot experiments in Appendix G and prior study (Peng et al., 2023). Other configurations are the same as in the original work (Farinhas et al., 2023).

[1] `https://opus.nlpl.eu`

4.4. Implementation Details

In the candidate generation step of PIVOTE, we employ NLLB. For each source-target language pair, we use an NLLB fine-tuned for the language pair in Table 1 to generate the directly translated candidates. For the merging module, we use Llama-3, GPT-4, and GPT-4o. For all models used in PIVOTE, including NLLB, Llama-3, GPT-4, and GPT-4o, we apply the same settings in §4.3.

As detailed in §3.3, we explore two approaches in the ensemble process: one dynamically selects the top-*k* (*k*=3) candidates, and another uses candidates obtained from fixed paths. To select the top-*k* candidates for each source sentence, we use the reference-free COMETkiwi. When selecting candidates from fixed paths, we use directly translated candidates and English-pivoted candidates, which were the top-performing paths on the FLORES-200 benchmark.

4.5. Main Results

Table 2 reports the overall performance of PIVOTE and other methods. The results show that PIVOTE consistently outperforms baselines across all language pairs. While standalone NMT systems rely solely on pre-trained knowledge, PIVOTE explicitly leverages candidates during ensemble. Even when training an open-source LLM, Llama-3, translation quality improves by utilizing candidates obtained via pivoting. Compared to LLM-based translation, performance improves at minimal cost with a small 0.6B model. Table 3 presents the quality of candidates used in the ensemble.

Comparison with multi-model ensemble. We compare PIVOTE with LLM-Blender (Jiang et al., 2023) and EVA (Xu et al., 2024), state-of-the-art ensemble methods utilizing multiple models. LLM-Blender employs N (N=11) LLMs for candidate generation, picks top-3 candidates with PAIRRANKER, and fuses them with GENFUSER. EVA is a token-level ensemble method that leverages vocabulary alignment across multiple models.

Results in Table 2 show that PIVOTE outperforms multi-model ensemble baselines by a considerable margin. LLM-Blender was unable to improve outputs compared to its candidate LLMs in non-English translation tasks. Additionally, LLMs used for generating candidates in LLM-Blender, such as Vicuna and Baize, exhibit subpar performance on given tasks. These results align with recent work (Xu et al., 2023b); open-source LLMs often struggle when not translating into English.

EVA is not only ineffective on the given tasks but also has several limitations inherent to its design as a token-level ensemble. First, EVA is unable to use black-box models such as GPT-4. Second, it is memory-intensive, as it requires loading mul-

Table 2: Main results. The best scores in each pair are marked **bold**. Within parentheses in the proposed method, the parts separated by semicolons denote the merging module and the candidates used. *D* and *E* represent candidates obtained from direct translation and English pivot, respectively.

Model	Korean→Italian			Italian→Korean			Arabic→Portuguese			Portuguese→Arabic		
	BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET	BLEU	chrF++	COMET
Standalone NMT System												
NLLB (Costa-jussà et al., 2022)	16.27	41.14	84.60	17.40	23.39	87.33	27.25	50.35	84.21	13.50	40.90	84.24
Vicuna (Chiang et al., 2023)	10.11	31.15	70.29	10.60	16.51	72.29	17.64	38.44	76.01	8.40	27.38	79.18
Baize (Xu et al., 2023a)	10.62	31.87	73.62	10.38	16.44	76.63	16.56	36.67	76.87	8.50	27.28	79.18
Llama-3 (AI@Meta, 2024)	11.79	34.82	77.37	13.82	18.95	85.80	18.78	40.20	78.73	12.25	35.16	82.79
GPT-4 (OpenAI, 2023)	14.07	42.22	86.80	17.23	22.96	86.94	25.82	51.89	85.46	15.11	41.39	83.99
GPT-4o (OpenAI, 2024)	15.11	42.59	85.93	17.20	22.82	85.31	27.28	52.57	85.90	16.28	42.40	83.82
Prior Ensemble Method												
LLM-Blender (Jiang et al., 2023)	8.77	28.74	82.80	0.03	0.85	42.77	11.80	29.85	67.95	0.94	2.69	46.49
EVA (Xu et al., 2024)	2.53	15.26	39.00	1.51	3.57	37.17	9.77	28.40	68.75	7.99	27.00	73.15
MBR (Farinhas et al., 2023)	14.10	42.24	86.70	17.14	23.00	87.53	25.45	51.78	85.55	14.66	41.11	83.93
Proposed Method												
PIVOTE (Llama-3; top3)	15.60	39.86	84.10	14.56	19.92	87.34	23.41	45.95	81.66	14.27	38.25	81.80
PIVOTE (Llama-3; *D*, *E*)	13.85	37.36	69.96	14.97	20.21	85.42	21.35	43.75	79.71	12.37	36.51	82.09
PIVOTE (GPT-4; top3)	16.66	42.85	**86.82**	17.95	23.84	87.50	27.22	51.73	85.65	16.53	42.41	84.46
PIVOTE (GPT-4; *D*, *E*)	17.10	43.29	85.92	18.18	24.05	**88.74**	27.98	52.41	85.27	17.02	43.02	**84.82**
PIVOTE (GPT-4o; top3)	17.77	43.38	85.46	18.08	23.98	88.15	28.62	52.53	85.87	16.92	42.93	84.52
PIVOTE (GPT-4o; *D*, *E*)	**18.02**	**43.46**	86.19	**18.31**	**24.32**	88.33	**29.50**	**53.16**	**86.03**	**17.66**	**43.73**	84.27

Table 3: Quality of ensemble candidates.

Model	Korean→Italian		
	BLEU	chrF++	COMET
Candidate			
NLLB (direct)	16.27	41.14	84.60
NLLB (Portuguese pivot)	13.13	37.57	83.21
NLLB (Spanish pivot)	13.87	38.47	83.71
NLLB (English pivot)	14.77	39.39	81.48

Table 4: Results on all language pairs.

Model	BLEU	chrF++	COMET	BLEU	chrF++	COMET
	Distant Language Pairs					
	Portuguese→Russian			Russian→Portuguese		
NLLB	25.17	51.77	90.12	29.69	55.81	86.01
GPT-4	26.50	52.76	91.11	25.51	54.05	86.69
PIVOTE	**27.48**	**53.49**	**91.74**	**30.82**	**56.73**	**88.37**
	Dutch→Russian			Russian→Dutch		
NLLB	22.95	50.21	89.92	25.56	53.60	88.18
GPT-4	24.37	51.32	91.28	24.46	53.85	88.58
PIVOTE	**25.45**	**52.16**	**91.47**	**28.05**	**55.80**	**89.35**
	French→Ukrainian			Ukrainian→French		
NLLB	14.58	37.11	82.99	20.69	44.04	80.61
GPT-4	13.84	39.03	84.12	23.30	47.13	83.43
PIVOTE	**17.20**	**39.82**	**86.55**	**24.35**	**47.17**	**84.36**
	Similar Language Pair (Romance)					
	Spanish→Portuguese			Portuguese→Spanish		
NLLB	32.38	56.97	86.88	33.63	57.61	85.13
GPT-4	29.94	55.26	84.84	34.70	58.63	86.75
PIVOTE	**34.06**	**58.11**	**87.70**	**36.03**	**59.32**	**86.92**
	Similar Language Pair (Slavic)					
	Ukrainian→Russian			Russian→Ukrainian		
NLLB	22.16	45.41	89.82	19.67	43.35	89.87
GPT-4	24.41	**47.59**	89.43	**22.42**	**45.61**	90.39
PIVOTE	**24.64**	47.51	**90.78**	22.09	45.40	**90.70**

tiple models into memory simultaneously. While multi-model ensemble methods generate candidates using up to 11 LLMs (with sizes up to 13B), PIVOTE generates candidates with a significantly smaller single model (0.6B), thereby greatly reducing computational overhead.

Results on all language pairs. To validate generalizability, we report the results for all language pairs we experimented with, including those within the same language family. Distant pairs refer to languages that belong to different families and use different scripts, while similar pairs belong to the same family and share the same script. The statistics for each language pair are in Appendix H. Language pairs used in the experiments are as follows:

- Distant language pairs: Portuguese↔Russian, Dutch↔Russian, and French↔Ukrainian
- Similar language pairs: Spanish↔Portuguese and Ukrainian↔Russian

Table 4 shows the results with the top-performing baselines, NLLB (Costa-jussà et al., 2022) and GPT-4 (OpenAI, 2023). PIVOTE consistently exhibits superior performance compared to strong baselines on distant language pairs. Surprisingly, it also showed improvements in similar language pairs, such as Spanish↔Portuguese.

Case study. We conduct a qualitative analysis to verify the impact of candidates on the final translation. We compare the output of GPT-4, used as the merging module, with PIVOTE, which utilizes candidates for the ensemble process. In Table 5, we provide two examples along with the source and target sentences, as well as the top-3 candidates.

In the first example, we observe that PIVOTE appropriately translates homonyms in context. In Korean, "자문" has the meaning of both "consultation" and "asking oneself". Considering the context, the expression should convey the meaning of "asking ourselves", as also shown in the English translation. However, GPT-4 mistranslated the source sentence, converting the phrase "자문해왔습니다" to "abbiamo cercato consigli" ("seeking consulta-

Table 5: Case study. Parts equivalent to the source sentence and mistranslated parts are highlighted in blue and red, respectively. English translation of the source comes from another pair within the dataset.

#	Type	Example
1	Source Sentence	그래서 그동안 자문해왔습니다. 왜 우리는 질병들과 싸우기에 더 현명하고, 정확하며 더욱 적합한 ...
		(English Translation: So we've been asking ourselves, why should we limit this smarter, more precise, more appropriate ...)
	Target Reference	Quindi mi sono chiesta: perché dovremmo limitare questo modo intelligente, preciso, migliore ...
	Top-1 Candidate	Quindi ci siamo chiesti, perché dovremmo limitare questo modo più intelligente, più preciso e più appropriato ...
	Top-2 Candidate	Quindi ci siamo chiesti: perché dovremmo limitare questo metodo più intelligente, più preciso e più adatto ...
	Top-3 Candidate	Quindi nel corso di questo tempo, abbiamo chiesto: perché dovremmo limitare questo metodo più intelligente, più preciso e più adeguato ...
	GPT-4	Quindi, abbiamo cercato consigli fino ad ora. Perché dobbiamo limitare questo metodo, che è più intelligente, preciso e più adatto ...
	PIVOTE	Quindi ci siamo chiesti: perché dovremmo limitare questo metodo più intelligente, più preciso e più adatto ...
2	Source Sentence	많은 사람들이 헤드폰을 사용하는데 이것의 문제점은 3가지 큰 건강 이슈를 가져온다는 것입니다.
		(English Translation: The trouble with widespread headphone use is it brings three really big health issues.)
	Target Reference	Il problema dell'utilizzo diffuso degli auricolari è che scatenano tre grandi problemi di salute.
	Top-1 Candidate	Il problema è che molte persone usano le cuffie, e questo Porta a tre grandi problemi di salute.
	Top-2 Candidate	Il problema è che molte persone usano le cuffie, e questo Porta a tre grandi problemi di salute.
	Top-3 Candidate	Il problema è che molte persone usano cuffie, e questo è ciò che causa tre problemi di salute principali.
	GPT-4	Molte persone utilizzano le cuffie, ma il problema è che ciò comporta tre importanti questioni di salute.
	PIVOTE	Il problema è che molte persone usano le cuffie, e questo porta a tre grandi problemi di salute.

Table 6: Candidate generation method comparison.

Cand. Generation	# Cand.	Korean→Italian		
		BLEU	chrF++	COMET
LLMs (LLM-Blender)	11	14.75	41.29	86.20
LLMs + NLLB (direct)	12	16.08	42.38	86.22
NLLB (pivot, ours)	4	**16.66**	**42.85**	**86.82**

Table 7: Evaluation of merging module variants.

Model	Korean→Italian		
	BLEU	chrF++	COMET
Standalone NMT System			
NLLB (Costa-jussà et al., 2022)	16.27	41.14	84.60
Encoder-Decoder			
PIVOTE (FiD)	13.74	36.78	78.98
PIVOTE (TRICE)	15.89	41.98	84.06
LLM-based			
PIVOTE (GENFUSER)	14.56	39.32	80.07
PIVOTE (GPT-4)	**16.66**	**42.85**	**86.82**

tion from others"). On the other hand, PIVOTE accurately translates "ci sono chiesti", meaning "asking ourselves", aligning well with the context by leveraging information from candidates.

In the second example, GPT-4 translates the source sentence by translating the noun "이슈" into "questioni". However, given the topic of discussing potential health risks, this translation does not fit well with the overall context. By contrast, the ensemble result of PIVOTE, generated using the identical model, improves translation quality by using a more accurate expression "problemi", despite having access to the same pre-trained knowledge. Additionally, when more suitable expressions (e.g., "ne vale la pena") appear in candidates, PIVOTE utilizes them to refine the final translation.

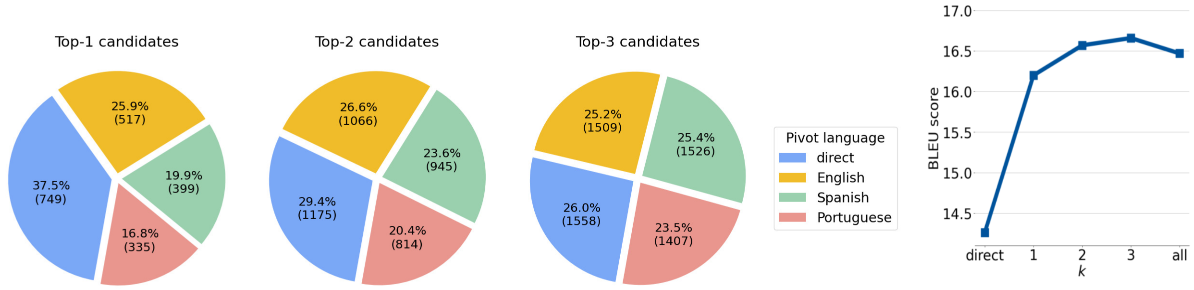
4.6. Analysis

Candidate generation. To validate the effectiveness of PIVOTE, we conduct experiments only varying the candidate generation method, while using the same merging module. We compare a candidate pool of size 4 obtained through pivot translation (PIVOTE) with a candidate pool of size 11 obtained using 11 LLMs in LLM-Blender.

As shown in Table 6, the proposed method of generating candidates through pivot translation achieves the highest performance, despite using the smallest candidate pool. From the perspective of direct translation in NLLB, leveraging 3 candidates obtained through pivot translation yields higher scores than incorporating candidates generated by 11 LLMs. These results demonstrate that using stable-quality candidates generated by a single model via pivot translation outperforms the use of multiple models with performance disparities.

Candidate aggregation. We first investigate whether PIVOTE shows improvement when utilizing other merging modules. As detailed in §3.3, we run experiments with three architectures: FiD (Izacard and Grave, 2021), TRICE (Huang et al., 2021), and GENFUSER (Jiang et al., 2023). When implementing FiD, we replace the backbone of FiD with mT5_{BASE} (Xue et al., 2021). TRICE is a method proposed for multi-source translation. Since TRICE was not originally intended for ensemble use, we repurposed it by training on the following two tasks: The first task is the original translation, which converts source sentences into target sentences. The second task is refining candidates that are paired with target references. In the case of TRICE, only the highest quality candidates, which are the directly translated ones, are used due to its architecture. FiD and GENFUSER use top-3 candidates.

Table 7 shows that the ensemble methods using encoder-decoder architectures and GENFUSER do not yield improved results. These methods struggle to leverage additional information from the candidates and, consequently, do not enhance performance. In contrast, using GPT-4 as the merging module leads to better performance compared to the standalone NMT system.



(a) Proportion of pivot languages (Korean→Italian) comprising the top-*k* candidates.

(b) Impact of top-*k* values on performance.

Figure 2: Analysis of pivot language distribution and top-k performance.

Table 8: Impact of candidate ranking strategies.

Method	Korean→Italian		
	BLEU	chrF++	COMET
PAIRRANKER (Jiang et al., 2023)	**16.74**	42.82	85.92
COMETkiwi (Rei et al., 2022b)	16.66	**42.85**	**86.82**

We also compare ranking methods COMETkiwi (Rei et al., 2022b) and PAIRRANKER (Jiang et al., 2023). Table 8 compares the results after selecting the top-3 using PAIRRANKER and COMETkiwi. As shown in the results, the difference is not significant. We believe this is because the candidates selected by both ranking methods are similar. There are 979 out of 2000 test sentences (48.95%) where the top-3 candidates selected by both ranking methods are the same. In cases with 2 out of 3 matches, there are 1533 instances (76.65%). Given the similarity in predictions by both ranking methods, the final scores exhibit comparable performance, except in the case of COMET. From the cost perspective, PAIRRANKER requires comparisons for $O(N^2)$ unique pair combinations depending on the number of candidates N. However, COMETkiwi only needs to sort the scores of N candidates, resulting in a time complexity of $O(N \log N)$. Therefore, due to its computational efficiency, we use COMETkiwi to rank candidates.

Comparison with selection-based ensemble. With a selection-based ensemble, we can choose one of the existing candidates as the final translation, rather than generating a new one. In this experiment, we compare our approach with a selection-based ensemble by selecting the top-1 translation using PAIRRANKER (Jiang et al., 2023) and COMETkiwi (Rei et al., 2022b). Additionally, we report results with an ideal case: selecting top-1 by considering references as well, which are not available in practice. The ideal top-1 is selected by reference-based COMET (Rei et al., 2022a).

As shown in Table 9, PIVOTE exhibits superior performance compared to the selection-based ensemble methods. Even when we leverage reference-based COMET, which is impossible in real-world scenarios due to the necessity for references, PIVOTE outperforms it in chrF++ and COMET. These results indicate that performing a generation-based ensemble with pivoting can effectively produce final translations that surpass those selected from the existing candidate pool.

Table 9: Compare with selection-based ensemble. COMET* is the ideal baseline, as it requires references. Best scores including COMET* are **bolded**, while best scores excluding it are <u>underlined</u>.

Category	Method	Korean→Italian		
		BLEU	chrF++	COMET
Selection-based (top-1)	PAIRRANKER	15.61	40.62	84.46
	COMETkiwi	15.61	40.71	84.10
	COMET* (ideal)	**17.77**	42.81	84.83
Generation-based	PIVOTE	<u>16.66</u>	<u>**42.85**</u>	<u>**86.82**</u>

Analysis of selected candidates. We conduct experiments to investigate the impact of the value of *k* in the top-*k* candidates and their composition. Figure 2a illustrates the proportion of pivot languages composing the top-*k* candidates. Top-*k* candidates, selected by the QE metric, are composed of diverse candidates obtained through various pivot languages. We also observe the same tendency in other datasets. This suggests that generating diverse candidates through multiple paths helps acquire higher-quality candidates.

Figure 2b presents BLEU for different values of the *k*. The highest BLEU is achieved when k is set to 3. These results demonstrate that more candidates in the aggregation process increase the likelihood of providing contextually appropriate information. However, it converges around top-3, which we attribute to the inclusion of lower-scoring candidates such as degenerated sentences. Hence, as k increases, the improvement reaches a plateau.

5. Conclusion

In this work, we introduced PivotE, a pivot-based single model ensemble framework, to enhance low-resource language translation. By transferring knowledge from diverse pivot languages, we were able to obtain not only diverse but also high-quality candidates. Since the optimal path to generating the best candidate varies per sentence, our study underscores the significance of exploiting a spectrum of pivot languages. Moreover, the single model generation process offers cost savings compared to multi-model ensemble approaches. Empirical results and qualitative analyses show that the proposed method can yield contextually suitable translations for the given source sentences by leveraging pivoted candidates.

6. Limitations

Although PivotE utilizes candidates obtained via pivoting, limitations arise from the nature of pivot translation. Constraining the pivot languages to high-resource languages can limit the number of candidates because pivoting through low-resource languages can lead to some loss of information due to error propagation inherent in the two-step translation. This semantic shift potentially causes a decrease in candidate quality. If the quality of candidates declines, improvements from the ensemble might not be significant, indicating a limitation in the number of pivot paths.

7. Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-16070246). This work was supported by the National R&D Program for Cancer Control through the National Cancer Center(NCC) funded by the Ministry of Health & Welfare, Republic of Korea (No.RS-2025-02264000).

8. Bibliographical References

AI@Meta. 2024. Llama 3 model card.

Alham Fikri Aji, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich. 2020. In neural machine translation, what does transfer learning transfer? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7701–7710, Online. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation.

Baichuan. 2023. Baichuan 2: Open large-scale language models.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ye Chen, Wei Cai, Liangmin Wu, Xiaowei Li, Zhanxuan Xin, and Cong Fu. 2023. Tigerbot: An open multilingual multitask llm.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv preprint*, abs/2210.11416.

Mike Conover, Matt Hayes, Ankit Mathur, Xiangrui Meng, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Raj Dabre, Fabien Cromieres, Sadao Kurohashi, and Pushpak Bhattacharyya. 2015. Leveraging small multilingual corpora for SMT using many pivot languages. In *Proceedings of the 2015 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1192–1202, Denver, Colorado. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

António Farinhas, José de Souza, and Andre Martins. 2023. An empirical study of translation hypothesis ensembling with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11956–11970, Singapore. 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.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multilingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. Koala: A dialogue model for academic research. Blog post.

Vaibhava Goel and William J Byrne. 2000. Minimum bayes-risk automatic speech recognition. *Computer Speech & Language*, 14(2):115–135.

Nuno M. Guerreiro, Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in large multilingual translation models.

Yinpeng Guo, Yi Liao, Xin Jiang, Qing Zhang, Yibo Zhang, and Qun Liu. 2019. Zero-shot paraphrase generation with multilingual language models.

Zhiwei He, Xing Wang, Zhaopeng Tu, Shuming Shi, and Rui Wang. 2022. Tencent AI lab - shanghai jiao tong university low-resource translation system for the WMT22 translation task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 260–267, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. 2017. Snapshot ensembles: Train 1, get m for free.

Xuancheng Huang, Jingfang Xu, Maosong Sun, and Yang Liu. 2021. Transfer learning for sequence generation: from single-source to multi-source. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5738–5750, Online. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st An-*

nual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14165–14178, Toronto, Canada. Association for Computational Linguistics.

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-English languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 866–876, Hong Kong, China. Association for Computational Linguistics.

LAION-AI. 2023. Open assistant. `https://github.com/LAION-AI/Open-Assistant`.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, Valencia, Spain. Association for Computational Linguistics.

Alireza Mohammadshahi, Jannis Vamvas, and Rico Sennrich. 2024. Investigating multi-pivot ensembling with massively multilingual machine translation models. In *Proceedings of the Fifth Workshop on Insights from Negative Results in NLP*, pages 169–180, Mexico City, Mexico. Association for Computational Linguistics.

NLP Team MosaicML. 2023. Introducing mpt-7b: A new standard for open-source, ly usable llms. Accessed: 2023-05-23.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

OpenAI. 2023. Gpt-4 technical report.

OpenAI. 2024. Gpt-4o system card.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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

Michael Paul, Hirofumi Yamamoto, Eiichiro Sumita, and Satoshi Nakamura. 2009. On the importance of pivot language selection for statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 221–224.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of chatgpt for machine translation.

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

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

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. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte M. Alves, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022b. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task.

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.

Omer Sagi and Lior Rokach. 2018. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8.

Stability-AI. 2023. Stablelm: Stability ai language models. https://github.com/stability-AI/stableLM.

Tianxiang Sun and Xipeng Qiu. 2023. Moss. https://github.com/OpenLMLab/MOSS.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2023. Follow the wisdom of the crowd: Effective text generation via minimum Bayes risk decoding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4265–4293, Toronto, Canada. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models.

Brian Thompson and Matt Post. 2020. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation*, pages 561–570, Online. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491, Rochester, New York. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Rationale-augmented ensembles in language models.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 856–863, Prague, Czech Republic. Association for Computational Linguistics.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023a. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *ArXiv preprint*, abs/2304.01196.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023b. A paradigm shift in machine translation: Boosting translation performance of large language models.

Yangyifan Xu, Jinliang Lu, and Jiajun Zhang. 2024. Bridging the gap between different vocabularies for LLM ensemble. 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 7140–7152, Mexico City, Mexico. Association for Computational Linguistics.

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

Zhangyue Yin, Qiushi Sun, Cheng Chang, Qipeng Guo, Junqi Dai, Xuanjing Huang, and Xipeng Qiu. 2023. Exchange-of-thought: Enhancing large language model capabilities through cross-model communication. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15135–15153, Singapore. Association for Computational Linguistics.

Meng Zhang, Liangyou Li, and Qun Liu. 2022. Triangular transfer: Freezing the pivot for triangular machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 644–650, Dublin, Ireland. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

9. Language Resource References

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

Reimers, Nils and Gurevych, Iryna. 2020. *Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation*. Association for Computational Linguistics.

Holger Schwenk and Vishrav Chaudhary and Shuo Sun and Hongyu Gong and Francisco Guzmán. 2019. *WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia*.

Jörg Tiedemann. 2012. *Parallel Data, Tools and Interfaces in OPUS*. European Language Resources Association (ELRA).

10. Appendices

A. Pivot Language Selection

Based on the results from the FLORES-200 (Costa-jussà et al., 2022) benchmark, we select the top-4 pivot paths as presented in Table 10. We utilize the full 2009 sentences as our test set: 997 sentences from the *dev* and 1012 sentences from the *devtest*. The pivot language pool is chosen as the *bridge languages* in Fan et al. (2020).

Table 10: BLEU scores on FLORES-200 benchmark. Pivot languages are sorted in alphabetical order and top-4 pivot paths are marked **bold**.

Pivot Language	**Lang-pair**			
	KO→IT	IT→KO	AR→PT	PT→AR
direct	**14.02**	**18.63**	**27.15**	**15.22**
arb_Arab	11.03	15.82	-	-
ben_Beng	10.79	15.44	18.65	9.76
ces_Latn	11.48	16.08	21.23	11.55
deu_Latn	12.49	17.11	22.62	12.56
ell_Grek	11.96	16.54	22.53	12.54
eng_Latn	**14.82**	**19.34**	**28.40**	**15.92**
fin_Latn	9.62	14.31	17.27	9.48
fra_Latn	13.55	17.27	**24.96**	13.77
heb_Hebr	10.42	14.37	20.31	10.94
hin_Deva	11.54	17.12	21.79	11.72
hun_Latn	10.54	14.96	18.64	9.65
ind_Latn	12.41	17.03	22.47	11.97
ita_Latn	-	-	24.70	**14.09**
jpn_Jpan	10.60	14.73	14.29	7.31
kor_Hang	-	-	16.09	7.67
lit_Latn	10.46	14.96	18.14	9.47
nld_Latn	12.27	17.10	23.22	12.94
pes_Arab	11.09	15.86	20.88	11.50
pol_Latn	11.54	15.86	21.14	11.60
por_Latn	**13.80**	**18.01**	-	-
rus_Cyrl	12.25	16.57	22.77	12.39
spa_Latn	**13.89**	**18.39**	**26.60**	**14.91**
swe_Latn	11.93	16.54	22.34	12.25
swh_Latn	10.66	14.22	19.13	10.19
tam_Taml	9.90	14.92	18.09	9.48
tur_Latn	11.25	15.92	19.53	10.04
ukr_Cyrl	11.76	16.43	21.87	12.12
vie_Latn	12.00	16.32	21.39	11.49
zho_Hans	10.00	11.51	15.29	6.82

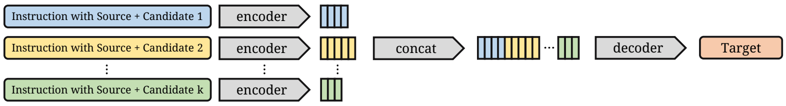


Figure 3: Illustration of the merging process using FiD (Izacard and Grave, 2021).

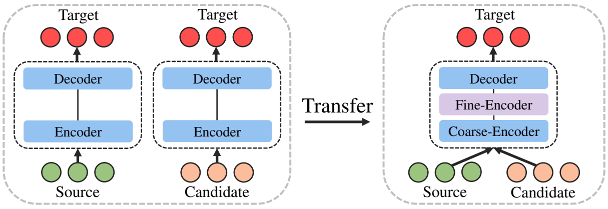


Figure 4: Illustration of the merging process using TRICE (Huang et al., 2021).

B. Impact of Resource-level of Pivot Languages

Under the assumption that high-quality candidates are more adept at conveying the meaning of the source sentence, we select the top-4 paths based on scores on FLORES-200. To verify this hypothesis, we conduct experiments using mid/low-resource pivot languages. According to WMT22[2], we select Ukrainian and Croatian as mid- and low-resource languages, respectively. Table 11 shows that using candidates from high-resource languages outperforms those obtained from mid- and low-resource languages. The quality of candidates is presented in Table 12. In conclusion, since high-resource languages can also provide sufficient diversity, we select top-performing paths based on the results on FLORES-200.

Table 11: Comparison with mid- and low-resource languages. *U*, *C*, *E*, and *S* represent candidates from Ukrainian, Croatian, English, and Spanish pivot, respectively.

Method	**Korean→Italian**		
	BLEU	chrF++	COMET
PIVOTE (GPT-4; *U, C*)	15.28	41.78	85.75
PIVOTE (GPT-4; *E, S*)	**16.27**	**42.55**	**86.50**

Table 12: Quality of candidates from each pivot path.

Model	**Korean→Italian**		
	BLEU	chrF++	COMET
Candidate			
NLLB (Ukrainian pivot)	11.95	35.03	82.32
NLLB (Croatian pivot)	12.25	35.93	79.91
NLLB (Spanish pivot)	13.87	38.47	83.71
NLLB (English pivot)	14.77	39.39	81.48

[2] `https://www.statmt.org/wmt22/translation-task.html`

C. Metric for Selecting Translation Paths

To analyze the impact of the path selection metric, we compare results using BLEU and COMET. As Table 13 shows, the performance differences are marginal. This is because both metrics select largely overlapping pivot languages, with only minor differences in ordering, as detailed in Table 14.

Table 13: Impact of the pivot path selection metric.

top-*k*	**Path Selection**	**Korean→Italian**		
		BLEU	chrF++	COMET
top-*1*	COMET	15.98	42.59	86.22
	BLEU	16.20	42.84	85.36
top-*2*	COMET	16.46	42.58	86.66
	BLEU	16.57	43.04	86.37
top-*3*	COMET	16.39	42.41	86.04
	BLEU	16.66	42.85	86.82

D. Prompt Templates

We use the zero-shot prompt template from Hendy et al. (2023) to instruct the LLMs for translation,

```
Translate this sentence from [source language] to [target language], Source: [source sentence]

Target:
```

when ensembling with candidates, we use the prompt template as follows,

```
Ensemble the [source language] sentence with the provided [target language] candidates to create the best possible [target language] translation.

[source language] sentence: [source sentence]

[target language] candidate k: [target candidate]

Please provide only the [target language] translation and no additional text.

[target language] translation:
```

Table 14: Selected top-4 pivot paths from each metric. Scores are from experiments on FLORES-200.

Lang-pair	BLEU	COMET
KO→IT	English (14.82), direct (14.02), Spanish (13.89), Portuguese (13.80)	English (82.89), Spanish (82.70), Indonesian (81.62), Portuguese (81.50)
IT→KO	English (19.34), direct (18.63), Spanish (18.39), Portuguese (18.01)	Spanish (87.32), English (87.07), Portuguese (87.02), French (86.14)
AR→PT	English (28.40), direct (27.15), Spanish (26.60), French (24.96)	direct (85.71), English (85.57), Spanish (85.54), Indonesian (84.94)
PT→AR	English (15.92), direct (15.22), Spanish (14.91), Italian (14.09)	French (82.65), direct (81.36), English (81.04), German (80.44)

E. Illustrations of the Merging Modules

In this section, we provide visual illustrations of the encoder-decoder ensemble architectures, specifically FiD and TRICE. Figure 3 and Figure 4 depict the input formatting and processing pipelines for each architecture, respectively.

F. Open-source LLMs

In experiments with LLM-Blender and EVA, we employ the same models as used in each paper. These open-source LLMs are listed in Table 15.

Table 15: Open-source LLMs along with their respective model sizes.

Model	Model Size
LLM-Blender *(Jiang et al., 2023)*	
Vicuna (Chiang et al., 2023)	13B
Baize (Xu et al., 2023a)	13B
Alpaca (Taori et al., 2023)	13B
Koala (Geng et al., 2023)	13B
Open Assistant (LAION-AI, 2023)	12B
Dolly V2 (Conover et al., 2023)	12B
Flan-T5 (Chung et al., 2022)	11B
MOSS (Sun and Qiu, 2023)	7B
Mosaic MPT (MosaicML, 2023)	7B
StableLM (Stability-AI, 2023)	7B
ChatGLM (Du et al., 2022)	6B
EVA *(Xu et al., 2024)*	
Baichuan2-Chat (Baichuan, 2023)	7B
TigerBot-Chat-V3 (Chen et al., 2023)	7B
Vicuna-V1.5 (Chiang et al., 2023)	7B
Llama-2-Chat (Touvron et al., 2023)	7B

G. Impact of *temperature* in MBR

To investigate the best performance of MBR, we compared it across three different *temperature* configurations: 1.0, 0.5, and 0.0, which were used in prior works by Farinhas et al. (2023), Suzgun et al. (2023), and Peng et al. (2023), respectively.

Tables 16 and 17 show the quality of MBR outputs and hypotheses under different *temperature* settings, respectively. Aligning with the findings of the previous study (Peng et al., 2023), we observed that a lower *temperature* setting achieved better performance. Thus, we set the *temperature* to 0.0 for MBR in our experiments.

Table 16: Impact of *temperature* in MBR decoding.

Method	Korean→Italian		
	BLEU	chrF++	COMET
MBR (*temp*=1.0; Farinhas et al. (2023))	13.53	42.13	86.57
MBR (*temp*=0.5; Suzgun et al. (2023))	13.90	42.19	86.69
MBR (*temp*=0.0; Peng et al. (2023))	**14.10**	**42.24**	**86.70**

Table 17: Average quality of MBR hypotheses.

Method	Korean→Italian		
	BLEU	chrF++	COMET
MBR hypotheses (*temp*=1.0)	13.47 (±0.21)	41.71 (±0.14)	84.94 (±2.62)
MBR hypotheses (*temp*=0.5)	13.86 (±0.13)	42.03 (±0.11)	86.55 (±0.15)
MBR hypotheses (*temp*=0.0)	14.09 (±0.07)	42.21 (±0.06)	86.62 (±0.10)

H. Dataset Statistics

Table 18 shows the dataset statistics for each language pair used in the experiments in Table 4.

Table 18: Dataset statistics.

Lang-pair	Dataset	# Sentences		
		Train	Dev	Test
Distant Language Pairs				
PT ↔ RU	news-commentary v18.1 (Tiedemann, 2012)	66,743	2,000	2,000
NL ↔ RU	news-commentary v18.1 (Tiedemann, 2012)	80,724	2,000	2,000
FR ↔ UK	WikiMatrix v1 (Schwenk et al., 2019)	166,063	2,000	2,000
Similar Language Pairs				
ES ↔ PT	TED 2020 v1 (Reimers and Gurevych, 2020)	315,462	2,000	2,000
UK ↔ RU	TED 2020 v1 (Reimers and Gurevych, 2020)	197,978	2,000	2,000