Improving Machine Translation with Large Language Models: A Preliminary Study with Cooperative Decoding

Anonymous ACL submission

Abstract

Contemporary translation engines built upon the encoder-decoder framework have reached a high level of development, while the emergence of Large Language Models (LLMs) has disrupted their position by offering the potential for achieving superior translation quality. To understand *in which scenarios LLMs outperform traditional NMT systems and how to leverage their strengths*, we first conduct a comprehensive analysis to assess the strengths and limitations of various commercial NMT systems and MT-oriented LLMs. Our findings indicate that neither NMT nor MT-oriented LLMs can effectively address all the translation issues, but MT-oriented LLMs can serve as a promising complement to the NMT systems. Building upon these insights, we explore hybrid methods and propose **Co**operative **Dec**oding (CoDec), which treats NMT systems as a pretranslation model and MT-oriented LLMs as a supplemental solution to handle complex scenarios beyond the capability of NMT alone. The results on the WMT22 test sets and a newly collected test set WebCrawl demonstrate the effectiveness and efficiency of CoDec, highlighting its potential as a robust solution for combining NMT systems with MT-oriented LLMs in machine translation.

1 Introduction

Over the years, the encoder-decoder framework has established Neural Machine Translation (NMT) models as the prevailing standard, achieving impressive translation quality through extensive training on large-scale, high-quality parallel data (Vaswani et al., 2017; Freitag and Firat, 2020; Fan et al., 2021). Commercial machine translation engines, e.g., Google Translate, are already proficient in addressing the majority of translation requirements. More recently, generative large language models (LLMs) like OpenAI's GPT-4 have demonstrated remarkable performance in language understanding and generation (Brown et al., 2020; OpenAI, 2023), prompting researchers to explore their potential application in various domains including machine translation (Ouyang et al., 2022; Zhu et al., 2023; Hendy et al., 2023).

Research on LLMs in machine translation can be categorized into two distinct approaches. One approach treats LLMs as black-box interfaces and leverages them under in-context learning (ICL) (Hendy et al., 2023; Zhu et al., 2023; Lin et al., 2022; Agrawal et al., 2022). However, this method limits further customization and improvement due to privacy concerns related to user data. A more promising approach involves fine-tuning relatively smaller LLMs (i.e., fewer than 13B parameters) to create LLMs specifically tailored for MT (Zeng et al., 2023; Zhang et al., 2023; Jiao et al., 2023).

This study aims to investigate the following research question: *In which scenarios do MT-oriented LLMs demonstrate superior performance, and how can we leverage the strengths of LLMs to enhance translation quality?*

To begin, we conduct a comprehensive analysis to gain insights into the characteristics of translations generated by commercial NMT systems and MT-oriented LLMs. Our findings reveal that commercial NMT systems excel at producing adequate translations in specific domains or languages. Conversely, MT-oriented LLMs demonstrate proficiency in generating authentic-sounding translations and handling infrequent words that are not effectively processed by NMT systems. In summary, MT-oriented LLMs can serve as valuable fallback systems in cases where the output of commercial NMT systems is unsatisfactory.

To complement NMT with MT-oriented LLMs, recently, Hendy et al. (2023) introduced the Hybrid Threshold approach, which employs the NMT system as the primary translation system. When the translation fails to meet the quality threshold determined by the quality estimation (QE) module (e.g., COMETkiwi), the GPT-like model is utilized

to provide alternative translations. Despite its effectiveness, there are still two issues: decoding latency and quality estimation. Neural language models typically employ autoregressive decoding, which relies on sequential execution rather than parallel processing. Due to the substantial computational requirements at each decoding step, decoding latency becomes a significant concern when integrating LLMs into the translation pipeline. For quality estimation, current reference-free metrics face challenges in accurately aligning with human judgment at the segment level and are insufficient for distinguishing the top-performing machine translation systems (Freitag et al., 2021, 2020; Ma et al., 2019; Rei et al., 2022a). Besides, their inclusion in the cascaded framework can introduce additional complexity, which needs to be used with caution in industrial applications.

To address the issues above, we propose a method called **Co**operative **Dec**oding (CoDec). In CoDec, the NMT system functions as the front-end module, generating an initial translation draft for a given input sentence. Subsequently, the MT-oriented LLM serves as an evaluator and evaluates the draft from a language modeling perspective. Specifically, the LLM refines the partial translation starting from a specific position where the token in the draft is not among the top-K token candidates suggested by the LLM. Since the evaluation process takes advantage of parallel computation and the front-end module can handle most situations effectively, CoDec is more efficient compared to using LLMs for complete decoding.

The contributions of this paper are three-fold:

- We conduct in-depth analyses to identify the strengths and weaknesses of traditional NMT systems and MT-oriented LLMs, finding that MT-oriented LLMs can complement NMT systems.
- We propose CoDec, an effective hybrid framework that combines the strengths of NMT systems and MT-oriented LLMs, and demonstrate its effectiveness on the WMT22 test sets and a newly collected test set, WebCrawl.
- We incorporate CoDec into the hybrid threshold pipeline(Hendy et al., 2023) as a comparative system, named CoDec(QE), and demonstrate its superiority over the original pipeline. Notably, CoDec achieves competitive or better performance than CoDec(QE) without the need for an additional quality estimation module, and it offers a significant acceleration advantage, achieving 2x speedup compared to directly using MT-oriented LLMs for generation.

2 Preliminary Experiments

In this section, we conduct a series of analyses to quantitatively investigate *the characteristics of translations from different systems*.

2.1 Setup

Commercial NMT Systems & MT-oriented LLMs. Our focus is the use of MT-oriented LLMs in industrial settings, and the chosen commercial NMT systems consist of Google Translate (*GoogleMT* for brevity)[1] and Microsoft Translate (*MicroMT* for brevity)[2]. Regarding MT-oriented LLMs, we utilize *BayLing-7B* (Zhang et al., 2023). We directly use the translations released on GitHub[3]. Additionally, we develop an in-house MT-oriented LLM. We fine-tune the tigerbot-13b-base[4] with TIM (Zeng et al., 2023) as the final MT-oriented LLM, denoted as *TIM*. More details can be found in Appendix A.

Automated MT Metrics. The MT Metrics shared task recommends the use of neural network-based metrics for evaluation, as they have demonstrated a high correlation with human evaluation and are robust to domain shifts. We utilize COMET-22 (wmt22-COMET-da) (Rei et al., 2021), and COMETkiwi (wmt22-COMET-kiwi-da) (Rei et al., 2022b) for reference-free quality estimation. We also report ChrF (Popovic, 2015) and SacreBLEU (Papineni et al., 2002) for completeness.

2.2 Analyses on WMT22 test sets

To prevent data leakage (Garcia et al., 2023), we analyze the latest WMT22 test sets. Detailed statistics of it are reported in Appendix B.

Main Results. The experimental results are illustrated in Table 1. We have made the following observations: 1) *GoogleMT* and *MicroMT* showcase excellent performance in many-to-English and English-to-many translation tasks. They consistently outperform the *WMT winner* in most of the

[1] https://translate.google.com/
[2] https://www.bing.com/translator
[3] https://github.com/ictnlp/BayLing/tree/main/exp/translation_benchmark/bayling-7b
[4] https://huggingface.co/TigerResearch/tigerbot-13b-base

System	COMET-22	COMETkiwi	ChrF	SacreBLEU	COMET-22	COMETkiwi	ChrF	SacreBLEU
		DE⇒EN				*EN⇒DE*		
WMT-Best	85.0	81.4	58.5	33.4	87.2	83.6	64.6	**38.4**
GoogleMT	**85.8**	**81.8**	**59.1**	**34.1**	**88.1**	**84.1**	**64.7**	37.5
MicroMT	85.1	81.4	58.8	33.9	87.4	83.7	64.7	37.5
BayLing-7B	83.2	80.1	53.6	28.2	82.1	79.2	53.6	25.7
TIM-13B	83.7	80.2	56.0	31.1	83.2	80.5	55.9	28.5
		ZH⇒EN				*EN⇒ZH*		
WMT-Best	81.0	77.7	**61.1**	**33.5**	86.7	82.0	41.1	44.8
GoogleMT	**82.7**	**79.3**	60.0	29.4	**88.2**	**82.7**	**45.8**	**50.5**
MicroMT	80.3	77.5	57.6	27.8	86.0	81.3	43.1	48.1
BayLing-7B	77.5	75.1	49.9	20.3	84.4	79.6	34.5	38.2
TIM-13B	79.9	76.5	56.2	26.4	86.3	80.9	41.7	46.4
		RU⇒EN				*EN⇒RU*		
WMT-Best	86.0	81.7	68.9	45.1	**89.5**	**84.4**	58.3	32.4
GoogleMT	**86.6**	**82.0**	**69.1**	45.6	**89.5**	84.2	59.5	34.3
MicroMT	85.5	81.1	**69.1**	**45.7**	88.7	83.6	**59.6**	**34.9**
BayLing-7B	82.5	79.3	60.4	34.7	74.7	70.6	35.5	14.8
TIM-13B	83.0	79.8	62.1	37.2	83.1	79.1	48.6	22.8
		JA⇒EN				*EN⇒JA*		
WMT-Best	81.6	80.3	49.8	24.8	89.3	85.8	36.8	27.6
GoogleMT	**84.0**	**81.7**	**51.8**	**26.2**	**90.2**	**86.5**	**37.6**	**28.2**
MicroMT	81.5	80.1	49.5	24.6	88.0	85.3	34.8	25.1
BayLing-7B	72.2	72.5	34.7	11.6	71.2	73.5	9.6	4.5
TIM-13B	80.5	79.3	45.8	21.3	86.2	83.0	27.9	18.2

Table 1: **Experimental results on the WMT22 test sets**.

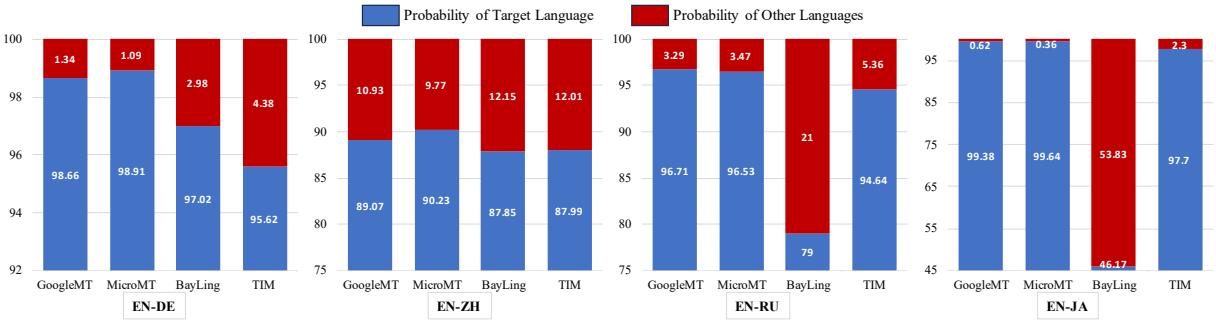

Figure 1: **Off-target rates (%) of translations.** MT-oriented LLMs exhibit a higher prevalence of off-target translations than NMT systems.

language pairs, highlighting the robust capabilities of these well-established translation engines. 2) *TIM* demonstrates comparable performance to the NMT systems in many-to-English, as evident from both reference-based and reference-free COMET scores. This suggests that LLMs, pretrained in English dominant corpus and fine-tuned with a small amount of high-quality parallel data, can exhibit impressive translation capabilities, particularly in many-to-English directions. 3) While a performance gap exists in English-to-many, MT-oriented LLMs still possess the potential for further enhancement. Specifically, *TIM* surpasses *BayLing* significantly across all language pairs and exhibits comparable performance to *MicroMT* on EN⇒ZH. This means that we can enhance the translation performance of MT-oriented LLMs by using a larger number of high-quality parallel sentences and effective fine-tuning methods, enabling them to achieve similar capabilities to commercial NMT systems without the need for model ensemble or rule-based engineering techniques.

Off-target Rates. Off-target indicates translations generated by machines involve segments of wrong languages or code-mixing, presenting a significant challenge in multilingual neural machine translation (Chen et al., 2023b; Zhang et al., 2020). Here, we use langdetect[5] to identify the language of each translation. The off-target rate of a translation is the subtraction of the probability of the

[5]https://github.com/Mimino666/langdetect

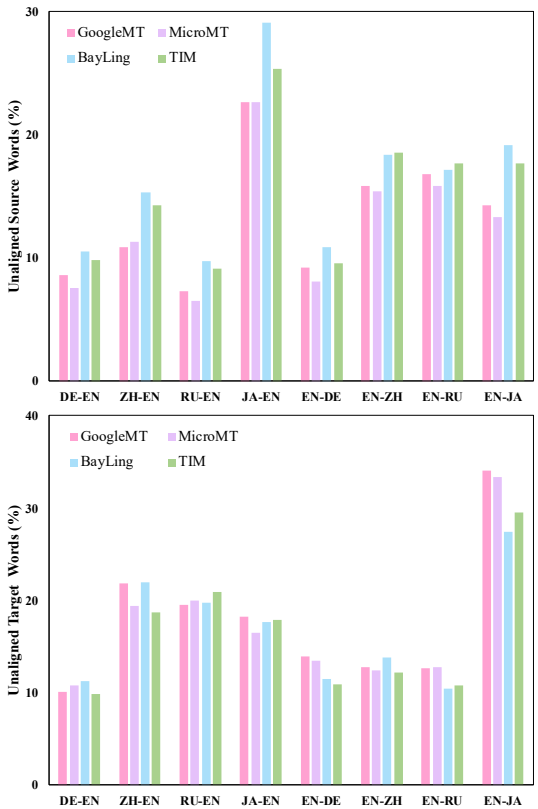

Figure 2: **Comparison of unaligned source words (USW) and unaligned target words (UTW).** LLM's translations demonstrate a significantly greater occurrence of unaligned source words across the test sets than its NMT counterpart.

target language prediction from 1. For a test set, we compute the average off-target rate across all the sentences.

As shown in Figure 1[6], the MT-oriented LLMs generate translations with higher off-target rates than the NMT systems. Specifically, *BayLing* exhibits off-target rates of 21% and 53.83% for EN⇒RU and EN⇒JA translations, respectively, which falls outside the language scope covered by the training data. This highlights a more pronounced off-target issue in LLMs, especially in zero-shot scenarios. In contrast, *TIM* achieves comparable off-target rates to *GoogleMT* and *MicroMT* in both ZH⇒EN and EN⇒ZH translations. We speculate that this can be attributed to TIM's incorporation of millions of corresponding training data, which enhances its ability to handle language switching and produce more accurate translations in these language pairs.

[6]We only show the results of English-to-Many due to the limited space, more detailed results are reported in Figure 6.

System	COMETk.↑	Sent Hal.↓	Token Hal.↓
GoogleMT	**79.3**	62.6	7.6
MicroMT	77.5	58.8	**5.6**
BayLing	75.1	73.6	13.2
TIM	76.5	**57.3**	6.5

Table 2: **Hallucination phenomenon**. BayLing exhibits severe hallucination in translations compared to GoogleMT and MicroMT, while TIM effectively alleviates this issue.

Unaligned Source/Target Words. To measure the literalness of the translation, we follow (Raunak et al., 2023; Hendy et al., 2023) to calculate the number of source and target words that do not align on a word-to-word basis. More details can be found in Appendix C. Figure 2 shows that the MT-oriented LLMs incur a notably larger number of unaligned source words over the test sets than the NMT counterpart. We examine the top six part-of-speech (POS) tags of the unaligned source words (in Appendix D). The difference mainly lies in nouns (NN) and adjectives (JJ), which suggests the possibility of either increased paraphrasing or a higher degree of inadequacy, such as omitted or inserted content.

However, back to Figure 2, the number of unaligned target words of the MT-oriented LLMs does not significantly differ from those of NMT systems, suggesting that the adequacy of translations produced by LLMs is comparable to NMT. We provide several examples highlighting the differences in translations in Appendix G, showing that MT-oriented LLMs tend to generate shorter translations (see Appendix E) with much paraphrasing, where the original text is rephrased using different words and sentence structures while preserving the same meaning.

Hallucination. The hallucination phenomenon in machine translation is that the translation system generates words, phrases, or sentences not present in the source text or deviates significantly from the intended meaning. We use the hallucination detector provided by Zhou et al. (2021) on the ZH⇒EN test set. We report the COMETkiwi, sentence-level, and token-level hallucination scores in Table 2. We can see that *BayLing*'s hallucination phenomenon is significantly more severe than *GoogleMT* and *MicroMT*, and *TIM* effectively mitigates this issue. This demonstrates that the hallucination is pronounced in simple supervised fine-tuned LLMs, while can be effectively alleviated through care-

System	GoogleMT	MicroMT	TIM
	ZH⇒EN		
COMETk.↑	59.05	57.44	**61.93**
#Length	56.81	51.38	62.66
Off-Target↓	1.15%	1.15%	0.84%
USW↓	13.87%	13.70%	15.94%
UTW↓	31.29%	25.08%	28.59%
	EN⇒ZH		
COMETk.↑	60.46	62.92	**64.90**
#Length	48.51	47.99	45.07
Off-Target↓	15.55%	14.08%	13.37%
USW↓	21.76%	20.23%	26.83%
UTW↓	16.55%	13.64%	15.34%

Table 3: **Experimental results on WebCrawl test sets**. MT-oriented LLMs can serve as valuable fallback systems in cases where the quality of commercial NMT systems is unsatisfactory.

fully designed fine-tuning methods such as training with comparison in (Zeng et al., 2023).

2.3 Analyses on Web Crawl test sets

The WMT22 test set is meticulously screened and annotated, with source sentences free of errors and from common domains. While the sentences have strong syntactic structures and grammatical correctness, real-world translation scenarios may not always have these ideal conditions. To reflect practical challenges, we collected a challenging test set from the open domain through web crawling. Here, we focus on Chinese⇔English and English⇔Chinese directions. To acquire the data, we follow the process outlined in Appendix H.

Main Results. Similarly, we compute various evaluation metrics for NMT systems and TIM on the WebCrawl test sets[7]. As shown in Table 3, *GoogleMT* outperforms *MicroMT* in the ZH⇒EN direction, while *MicroMT* outperforms *GoogleMT* in the EN⇒ZH direction. This indicates that different NMT systems exhibit a certain level of complementarity in handling translation issues that they individually struggle with, highlighting the benefit of the model ensemble. Moreover, it is noteworthy that *TIM* demonstrates significant improvements in both ZH⇒EN and EN⇒ZH directions. This surprising finding suggests that MT-oriented LLMs can serve as valuable fallback systems in cases where the quality of commercial NMT systems is unsatisfactory. Moreover, the phenomena observed in the other metrics (e.g., USW and UTW) are consistent with the analysis in Section 2.2, further demonstrating the universality of our findings.

Case Study. In Appendix I, we provide several examples that are hard for *GoogleMT* to handle but are solved well by *TIM*. It shows that the NMT system struggles to understand the meaning of some professional terms and fails to produce suitable translations for ill-informed text. In contrast, MT-oriented LLM demonstrates its superiority in handling such issues, which can be attributed to its enhanced ability to comprehend rare, specialized words, and informal texts.

To further support our hypothesis, we calculate the COMETkiwi scores of the translations generated by *GoogleMT* and *TIM* against the source text, selecting a group of sentences where *GoogleMT* has higher scores than *TIM* by more than 3 points, and another group where *TIM* has higher scores than *GoogleMT*. To mitigate the impact of sentence lengths, we retain only those sentences containing fewer than 60 tokens. Next, we use gpt2-large[8] to calculate the perplexity for the two groups. The perplexity for sentences in which *GoogleMT* excels is 38.60, whereas for sentences in which *TIM* performs better, it is 45.53. The MT-oriented LLM showcases superior proficiency in handling complex source language sentences, as reflected by lower perplexity scores.

3 Method

The analyses in Section 2.2 show that commercial NMT systems generate accurate translations due to their extensive training and cross-attention capability, while LLM's paraphrastic nature can enhance NMT's ability to handle figurative text translations. Moreover, MT-oriented LLMs perform well on WebCrawl test sets (Section 2.3), particularly for specialized terminology or ill-formed sentences. This suggests that an effective hybrid framework combining NMT and LLMs can handle challenging input domains and figurative text. Inspired by this, we propose to investigate an effective hybrid framework to answer the second question: *How can we effectively harness the capabilities of LLMs to enhance translation quality?*

3.1 Hybrid Threshold

Hendy et al. (2023) propose *Hybrid Threshold* (see Figure 11 (b)), that combine NMT and GPT systems. In particular, Hybrid Threshold employs the

[7]We select TIM as the representative for MT-oriented LLMs, due to its better performance on the WMT22 test set.

[8]https://huggingface.co/gpt2-large

NMT model as the primary translation system, with a quality estimation module utilizing COMETkiwi to assess the translation. If the quality falls below a certain threshold, the GPT model is used as an alternative translation source. However, the practical implementation of it faces two main challenges: autoregressive decoding latency and reliable quality estimation. Addressing these challenges is crucial for the practical implementation of the hybrid approaches, ensuring efficient decoding and high translation quality.

3.2 Cooperative Decoding

To alleviate the issues above, we propose a novel cooperative decoding approach. This approach leverages an NMT model as a pretranslation model and incorporates an MT-oriented LLM as a quality assessment module and fallback model if needed. The overview of cooperative decoding is shown in Figure 12, and we will give a detailed description of each step in the following.

Step1: Draft Generation. Given an input source sentence x, the NMT system generates the translation o using an autoregressive decoding strategy like beam search. The difference is that the translation o is considered as a draft and requires further confirmation or modification before being used as the final output.

Step2: Verification. We feed o into the MT-oriented LLM in a forward process, which fully utilizes the parallel computing of the architecture. The procedure is the same as training LLMs, and we can obtain a probability distribution v_t at each position, which is modeled as $P(o_t|o_{<t}, x)$. The distribution can be regarded as the confidence of the LLM given the specific prefix of the draft o, and we use it to verify the prediction of the NMT model. One straightforward approach for verification is to check whether the token with the highest probability matches the prediction of NMT. If o is fortunately exactly the same with $\{\text{argmax}(v_1), ..., \text{argmax}(v_n)\}$, the inference will finish with o as the final translation. However, high-quality generation does not follow a distribution of the highest probability of the next tokens, and the tokens in o that can be regarded as accurate may appear outside of the top-1 selection, like in beam search. To address this issue, we relax the matching constraint using the top-k candidates of the LLM and define the verification criterion as

$$o_t \in \text{top}-k(v_t). \tag{1}$$

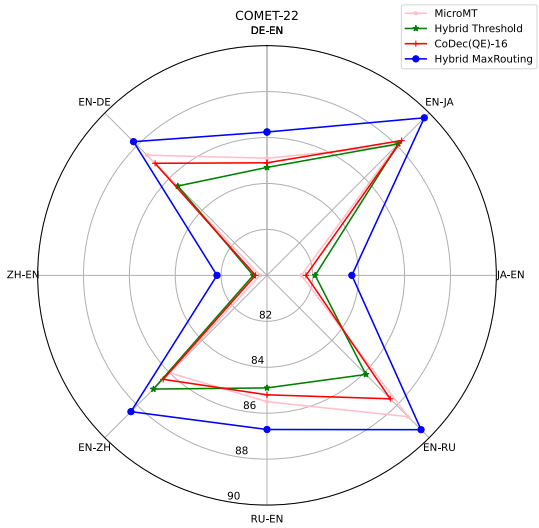

Figure 3: **Comparison of MicroMT and different hybrid frameworks**. We report COMET-22 scores here and report COMETkiwi in Figure 13.

Step3: Re-decoding. The verification is performed from left to right, and we end the verification once there is a situation that does not meet the verification criteria, i.e., $o_{t'} \notin \text{top}-k(v_{t'})$. Then, we feed the verified prefix $o_{t'-1}$ into the MT-oriented LLM and use it to re-decode the subsequent sequence. Compared to totally replacing NMT models with MT-oriented LLMs, our cooperative decoding can speed up the whole inference process due to the expensive cost of autoregressive decoding. The speedup is more significant when the longer draft is accepted. Moreover, the cooperative mechanism alleviates the issue of inaccuracy of LLMs by exploiting the output of NMT models.

4 Experiments

4.1 Main Results

Detailed setup can be found in Appendix J. The performance of the different hybrid frameworks is compared in Figure 3. The Hybrid Max-Routing approach, depicted in Figure 11(a), establishes an upper bound by selecting the best translation from either system based on the evaluation metric COMETkiwi. The significant gap between *MicroMT* and *Hybrid MaxRouting* across all test sets

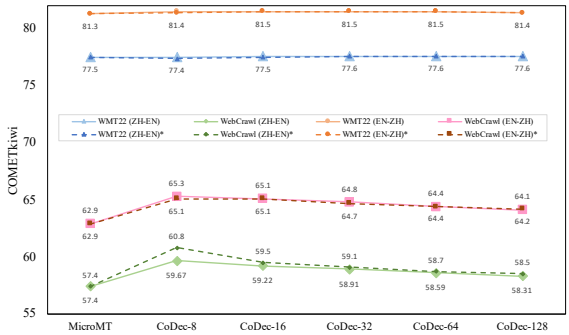

Figure 4: **Performance of MicroMT, CoDec(QE), and CoDec with different values of k (Eq. 1).** We show the CoMETkiwi scores on WMT22 and WebCrawl test sets. The solid line denotes the results obtained from CoDec(QE), whereas the dashed line represents the results obtained from CoDec.

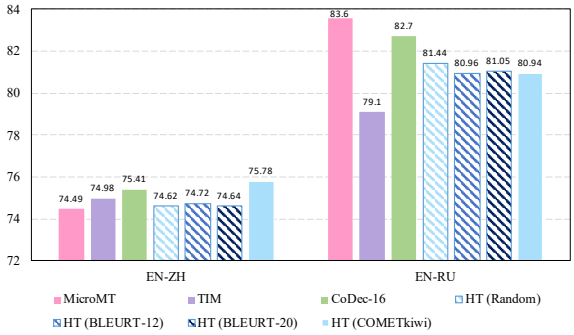

Figure 5: **Performance of MicroMT, TIM, CoDec-16, and Hybrid Threshold with different quality estimation methods.** We show the COMETkiwi scores on the merge of the WMT22 and WebCrawl test sets in EN⇒ZH, and WMT22 EN⇒RU.

supports our assertion that MT-oriented LLMs can serve as valuable fallback systems for NMT. Additionally, it is evident that *Hybrid Threshold* does not consistently yield improvements (e.g., on EN⇒DE and EN⇒RU directions), as its effectiveness relies on the accuracy of the QE module and the quality of LLM translations used as substitutes. This suggests that the simplistic threshold-based replacement in *Hybrid Threshold* may be sub-optimal. Intuitively, introducing CoDec to Hybrid Threshold (i.e., CoDec(QE)) partially addresses these limitations. The observed improvement is thanks to the valid-then-decoding mechanism employed in our CoDec. By validating the NMT translation first and then refining it, we can retain high-quality translation segments while rectifying errors. Overall, the results highlight the benefits of CoDec in terms of both inference efficiency and translation quality.

CoDec(QE) vs. CoDec. We compare the performance of *CoDec(QE)* and *CoDec* with different values of k (Eq. 1). In Figure 4, we observe minimal performance variation on the WMT22 test sets, and we speculate that this is because when the NMT system itself produces translations of sufficiently high quality, there is no need for further refinement by MT-oriented LLMs. However, as k increases, both *CoDec(QE)* and *CoDec* exhibit gradual decreases in performance on the WebCrawl test sets. The underlying reason is that a higher value of k represents a more relaxed evaluation standard during verification. Consequently, it allows more erroneous segments from the NMT translations, leading to a decline in overall translation quality.

CoDec vs. Hybrid Threshold. We compare the performance of CoDec and Hybrid Threshold using different QE methods: *HT(Random)*, where we randomly replace 50% MicroMT's translations with TIM's translations; *HT(BLEURT-12)*, which uses BLEURT-20-D12[9] as the QE method; *HT(BLEURT-20)*, which employs BLEURT-20[10] as the QE method; and *HT(COMETkiwi)*. We evaluate different methods on the combined WMT22 and WebCrawl test sets in EN⇒ZH in which *TIM* outperforms *MicroMT*, and WMT22 EN⇒RU in which *TIM* underperforms *MicroMT*.

As depicted in Figure 5, Hybrid Threshold demonstrates improvement on the EN⇒ZH test set but exhibited a significant decrease on the EN⇒RU test set. Additionally, Hybrid Thresholds with different QE modules yield varying performances. This confirms that the performance of Hybrid Threshold relies on the precision of introduced QE modules and the quality of LLM translations employed as a replacement. In contrast, *CoDec-16* outperforms most of the Hybrid Threshold models, showing that the QE modules may not be necessary. This also indicates that our method achieves a better balance between efficiency and effectiveness.

4.2 Effect of different values of k

Intuitively, as k increases, cooperative decoding can accept a larger range of tokens of NMT translations during the verification stage. Consequently, less content must be re-decoded by LLMs, resulting in reduced processing time. Here, we investigate the performance of CoDec under different values

[9] https://huggingface.co/lucadiliello/BLEURT-20-D12
[10] https://huggingface.co/lucadiliello/BLEURT-20

Method	ZH⇒EN				EN⇒ZH			
	COMETk.	Token/s	Speedup	$\text{Ratio}_{\text{acc}}$	COMETk.	Token/s	Speedup	$\text{Ratio}_{\text{acc}}$
MicroMT	71.03	-	-	-	74.49	-	-	-
TIM	71.85	21.72	1.0×	-	74.98	20.67	1.0×	-
CoDec-8	71.85	34.23	1.6×	65.03	75.37	27.49	1.3×	62.09
CoDec-16	71.74	39.62	1.8×	75.59	75.41	30.86	1.5×	72.44
CoDec-32	71.64	45.91	2.1×	82.46	75.29	34.70	1.7×	80.14
CoDec-64	71.51	51.72	2.4×	87.97	75.14	39.06	1.9×	85.28
CoDec-128	71.43	57.10	2.6×	91.76	75.07	43.55	2.1×	89.14

Table 4: **Effect of different values of k (Eq. 1) for CoDec**. We present the results on ZH⇒EN and EN⇒ZH including COMETkiwi, decoding speed measured by tokens per second, decoding speedup, and the ratio of the number of tokens accepted at the verification stage to the total tokens of the draft.

of k. We merge the WMT22 and WebCrawl test sets to simulate the distribution of translation requests in real-world scenarios. As shown in Table 4, with the increase of k, the ratio of tokens accepted on average and the decoding speed increase consistently. With a large k, *CoDeC-128* achieves 2.6x speedup over the baselines, i.e., *MicroMT* in both ZH⇔EN and *TIM* in EN⇒ZH, and also obtains higher COMETkiwi scores. This signifies that CoDec effectively reduces decoding latency and maintains translation quality. On the other hand, the models with lower k achieve better translation quality such as *CoDec-8*, indicating that k should be chosen carefully to get the balance of performance and efficiency.

5 Related Work

5.1 Large Language Models on Machine Translation

Research on Large Language Models (LLMs) for machine translation can be broadly divided into two categories: utilizing LLMs as an interface and optimizing them for specific translation tasks. For the former, Hendy et al. (2023) evaluate ChatGPT, GPT3.5, and text-adavinci-002 in eighteen translation directions, while Zhu et al. (2023) assess XGLM, BLOOMZ, OPT, and ChatGPT across 202 directions and 102 languages. Other researchers explore strategies for selecting translation exemplars (Lin et al., 2022; Agrawal et al., 2022) and incorporating external knowledge (Lu et al., 2023) to enhance GPT translation. Fine-tuning smaller models (e.g., 7B) specifically for translation tasks has attracted increasing attention (Zeng et al., 2023; Zhang et al., 2023; Jiao et al., 2023). Diverging from existing approaches, our research focuses on examining the capabilities and limitations of commercial NMT systems and MT-oriented LLMs and developing efficient hybrid frameworks that leverage their respective strengths.

5.2 Accelerate Generation for Large Language Models

Efforts to improve the inference efficiency of LLMs have been ongoing for several years(Tay et al., 2023; Xu et al., 2021), leveraging techniques such as knowledge distillation(Hinton et al., 2015; Jiao et al., 2020; Wang et al., 2020), quantization(Shen et al., 2020; Sun et al., 2020), pruning(Fan et al., 2020), and others (Kim and Cho, 2021; Lei, 2021). The most related work is to leverage speculative execution (Burton, 1985; Hennessy and Patterson, 2012) for the speedup of autoregressive models. Stern et al. (2018) propose to decode several tokens in parallel to accelerate greedy decoding. For LLMs, speculative decoding (Chen et al., 2023a; Leviathan et al., 2023) uses an additional draft model and generates sequences with sampling. Yang et al. (2023) copy some tokens from retrieved reference text to the decoder, which are validated with output probabilities. Santilli et al. (2023) reframe MT's standard greedy autoregressive decoding procedure with a parallel formulation. We are pioneers in using speculative execution as a fusion approach for commercial NMT systems and MT-oriented LLMs, without requiring an auxiliary quality estimation module or modifications to the target LLMs' parameters.

6 Conclusion

We explore the strengths of both NMT and LLM and propose CoDec that integrates the two to achieve superior performance compared to existing hybrid frameworks. Notably, our CoDec offers reduced decoding latency compared to relying solely on LLMs for inference, and it does not require any modifications to the target LLMs.

7 Limitations

In this paper, the top-K candidates predicted by the target MT-oriented LLM were directly used for verification of the draft translation. We acknowledge that more carefully designed token-level validation methods can further improve CoDec and CoDec(QE), which we leave as future work.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation. *CoRR*, abs/2212.02437.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

F. Warren Burton. 1985. Speculative computation, parallelism, and functional programming. *IEEE Trans. Computers*, 34(12):1190–1193.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. 2023a. Accelerating large language model decoding with speculative sampling. *CoRR*, abs/2302.01318.

Liang Chen, Shuming Ma, Dongdong Zhang, Furu Wei, and Baobao Chang. 2023b. On the off-target problem of zero-shot multilingual neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9542–9558, Toronto, Canada. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. 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, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

Angela Fan, Edouard Grave, and Armand Joulin. 2020. Reducing transformer depth on demand with structured dropout. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Markus Freitag and Orhan Firat. 2020. Complete multilingual neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 550–560, Online. Association for Computational Linguistics.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Xavier Garcia, Yamini Bansal, Colin Cherry, George F. Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. *CoRR*, abs/2302.01398.

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. *CoRR*, abs/2302.09210.

John L. Hennessy and David A. Patterson. 2012. *Computer Architecture - A Quantitative Approach, 5th Edition*. Morgan Kaufmann.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Parrot: Translating during chat using large language models. *CoRR*, abs/2304.02426.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Gyuwan Kim and Kyunghyun Cho. 2021. Length-adaptive transformer: Train once with length drop, use anytime with search. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6501–6511. Association for Computational Linguistics.

Tao Lei. 2021. When attention meets fast recurrence: Training language models with reduced compute. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7633–7648, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19274–19286. PMLR.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *EMNLP 2022*, pages 9019–9052.

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. 2023. Chain-of-dictionary prompting elicits translation in large language models. *CoRR*, abs/2305.06575.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

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. Association for Computational Linguistics.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation, WMT@EMNLP 2015, 17-18 September 2015, Lisbon, Portugal*, pages 392–395. The Association for Computer Linguistics.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan. 2023. Do GPTs produce less literal translations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1041–1050, Toronto, Canada. 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, José G. C. de Souza, Duarte M. Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luísa Coheur, and André F. T. Martins. 2022b. COMET-22: unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 578–585. Association for Computational Linguistics.

Ricardo Rei, Ana C. Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro G. Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-ist 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation, WMT@EMNLP 2021, Online Event, November 10-11, 2021*, pages 1030–1040. Association for Computational Linguistics.

Andrea Santilli, Silvio Severino, Emilian Postolache, Valentino Maiorca, Michele Mancusi, Riccardo Marin, and Emanuele Rodola. 2023. Accelerating transformer inference for translation via parallel decoding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12336–12355, Toronto, Canada. Association for Computational Linguistics.

Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. 2020. Q-BERT: hessian based ultra low precision quantization of BERT. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY,*

USA, February 7-12, 2020, pages 8815–8821. AAAI Press.

Mitchell Stern, Noam Shazeer, and Jakob Uszkoreit. 2018. Blockwise parallel decoding for deep autoregressive models. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 10107–10116.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. MobileBERT: a compact task-agnostic BERT for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2158–2170, Online. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2023. Efficient transformers: A survey. *ACM Comput. Surv.*, 55(6):109:1–109:28.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS 2017*, pages 5998–6008.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jingjing Xu, Wangchunshu Zhou, Zhiyi Fu, Hao Zhou, and Lei Li. 2021. A survey on green deep learning. *CoRR*, abs/2111.05193.

Nan Yang, Tao Ge, Liang Wang, Binxing Jiao, Daxin Jiang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Inference with reference: Lossless acceleration of large language models. *CoRR*, abs/2304.04487.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2023. TIM: teaching large language models to translate with comparison. *CoRR*, abs/2307.04408.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, and Yang Feng. 2023. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. *CoRR*, abs/2306.10968.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *CoRR*, abs/2304.04675.

A Training details of TIM

we develop an in-house MT-oriented LLM, trained on human-written validation data from previous WMT competitions[11], such as the newstest2017-2021 of German⇔English, Chinese⇔English, Russian⇔English, and Jappanese⇔English, along with millions of high-quality bilingual sentence pairs in Chinese⇔English. According to the data license of WMT22, the data released for the General MT task can be freely used for research purposes. We fine-tune the tigerbot-13b-base[12] with TIM (Zeng et al., 2023) as the final MT-oriented LLM.

B WMT22 test sets

To prevent data leakage (Garcia et al., 2023), we analyze the latest WMT22 test sets, consisting of recent content from diverse domains including news, social media, e-commerce, and conversation. The test sets consist of the following number of samples for each language pair: German-to-English (DE⇒DE) - 1984 samples, English-to-German (EN⇒DE) - 2037 samples, Chinese-to-English (ZH⇒EN) - 1875 samples, English-to-Chinese (EN⇒ZH) - 2037 samples, Russian-to-English (RU⇒EN) - 2016 samples, English-to-Russian (EN⇒RU) - 2037 samples, Japanese-to-English (JA⇒EN) - 2008 samples, English-to-Japanese (EN⇒JA) - 2037 samples.

C Unaligned Source/Target Words.

For English and German, we utilize the Moses tokenizer[13]. We use jieba[14] and MeCab[15] for Chinese and Japanese, respectively. We use *awesome-*

[11] https://www.statmt.org/wmt22/translation-task.html
[12] https://huggingface.co/TigerResearch/tigerbot-13b-base
[13] https://github.com/moses-smt/mosesdecoder/tree/master/scripts/tokenizer
[14] https://github.com/fxsjy/jieba
[15] https://github.com/SamuraiT/mecab-python3

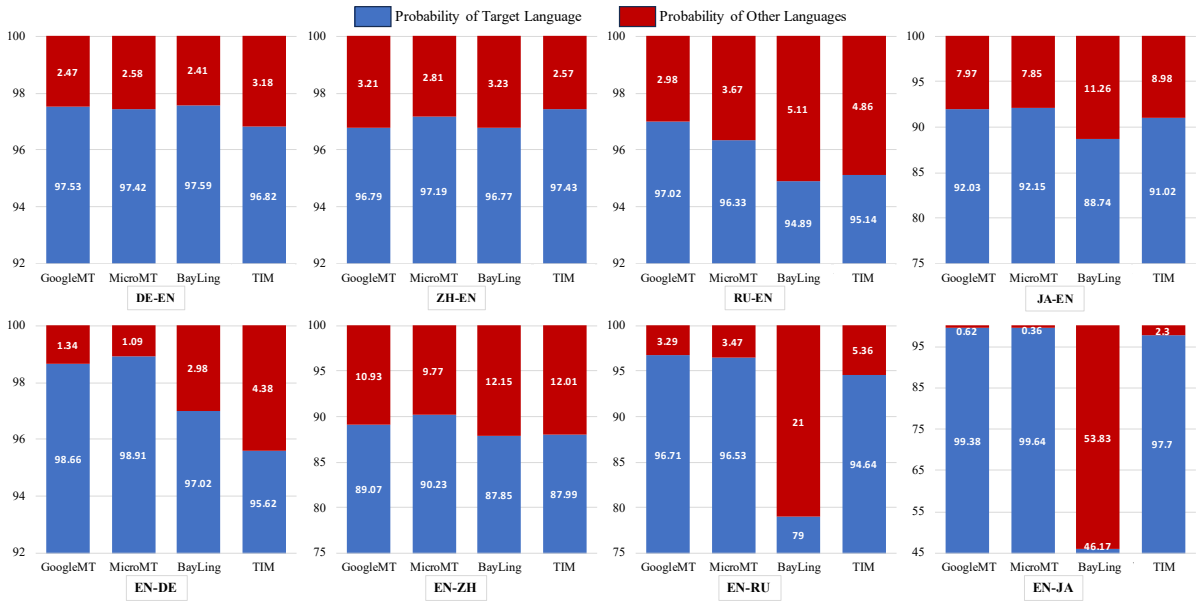

Figure 6: **Off-target rates (%) of translations.** MT-oriented LLMs exhibit a higher prevalence of off-target translations than NMT systems.

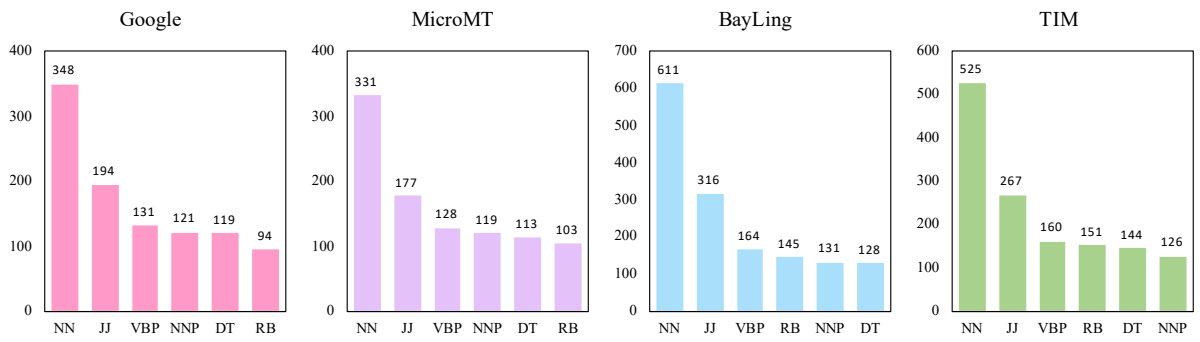

Figure 7: **POS tags of unaligned source words.** We show the top-6 POS tags of USW on the WMT22 EN⇒ZH test set, and the incremental USW of MT-oriented LLMs mainly lies in nouns (NN) and adjectives (JJ).

align[16] (Dou and Neubig, 2021) to obtain the word alignments.

Unaligned source words (USW) indicate the number of words in the source text that have no corresponding translation in the target sentence. Unaligned target words (UTW) assess the degree to which words are potentially added or inserted into the translation without any basis or support from the source sentence.

D Pos tags of Unaligned Source Words

We examine the top six part-of-speech (POS) tags by NLTK toolkit (Bird et al., 2009) of the unaligned source words (Figure 7 in Appendix D), and the difference mainly lies in nouns (NN) and adjectives (JJ). This observation suggests the possibility of either increased paraphrasing or a higher degree of inadequacy, such as omitted or inserted content.

E Average Length.

We calculate the average number of words in the generated translations. As depicted in Figure 8, MT-oriented LLMs tend to produce shorter sentences, employing concise and precise wording to convey the meaning of the source sentence. The underlying reason is that humans tend to use concise language when explaining things, especially in conversations, and such data is abundant in the corpus used to train LLMs. In contrast, NMT models primarily rely on parallel data and strive to encompass translations for each token within the source sentence.

In Appendix F, we also compare the fluency of translations produced by different systems. Intu-

[16] https://github.com/neulab/awesome-align

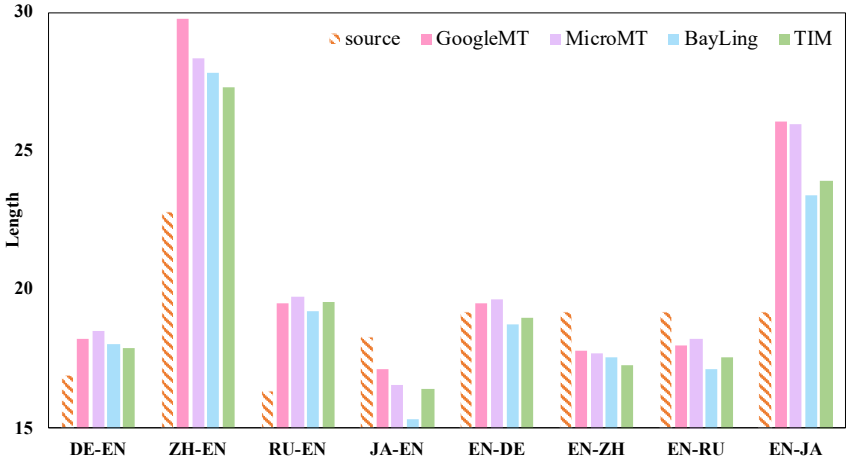

Figure 8: **Average Length of Translations.** MT-oriented LLMs tend to produce shorter sentences, i.e., using compact and precise wording to express the meaning of the source sentence.

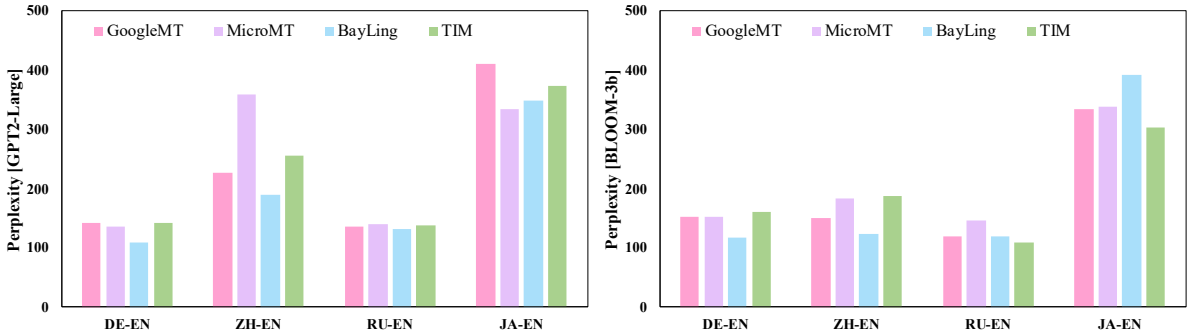

Figure 9: **Translation Fluency.** MT-oriented LLMs do not exhibit a significant advantage compared to NMT systems.

itively, MT-oriented LLMs can demonstrate superior fluency as measured by perplexity, although the results do not reflect this. This is because it is adequate to fully train the NMT systems using the available sentence pairs for the selected languages (German, Chinese, Russian, and Japanese).

F Translation Fluency.

We measure translation fluency using gpt2-large[17] and BLOOM-3b[18]. We assess fluency in many-to-English translation directions since both models are primarily trained on English text. In Figure 9, we observe that the four systems exhibit varying degrees of fluency in different translation directions. Surprisingly, MT-oriented LLMs do not exhibit a significant advantage. This discrepancy can be attributed to the availability of abundant bilingual and monolingual data for the selected four languages. The NMT systems, benefiting from extensive language modeling during training, can generate fluent sentences in these language pairs.

G Examples of free translation generated by MT-oriented LLM

Figure 10 provides several examples highlighting the differences in translations. For instance, the phrase "frustration shouting", is expected to be translated to "大喊大叫 (scream in frustration)". *MicroMT*, in its pursuit of fidelity, employs translation augmentation segments such as "我非常沮丧 (I feel extremely frustrated)". In contrast, *TIM* exhibits a superior understanding of the complete sentence and offers more accurate translations. However, in the third example, *TIM* overlooks the inclusion of the expression "the receptive, super-alert brain of a tiny child" from the source text, resulting in a certain degree of translation oversight. In summary, MT-oriented LLMs tend to gen-

[17] https://huggingface.co/gpt2-large
[18] https://huggingface.co/bigscience/bloom-3b

Eample	System	Translation	USW	Reference
Were you able to try purchasing **on the computer on the website**?	MicroMT	您是否可以尝试**在网站上的计算机**上购买?	-	您能使用电脑从网站购买吗?
	TIM	你能在网站上尝试购买吗?	computer	
After much **frustration shouting** at my watch during phone calls so I could be heard and / or understood.	MicroMT	在打电话时对着手表大喊大叫之后，**我非常沮丧**，这样我就可以被听到和/或理解。	-	在为了让别人能听到和/或听懂我说的话而在打电话时对着手表大喊大叫之后。
	TIM	在电话里对着我的手表**大喊大叫**，这样我才能被听到和/或被理解。	frustration	
What kind of message does that send into **the receptive , super-alert brain of** a tiny child ?	MicroMT	这会向一个小孩子的**接受性、超级警觉的大脑**发送什么样的信息?	-	这会向一个接受能力强、高度警觉的小孩子传递什么样的信息呢?
	TIM	这会给**幼小的孩子**传递什么样的信息?	brain, receptive, super-alert	

Figure 10: **Examples of free translation generated by MT-oriented LLM.**

erate translations with much paraphrasing, where the original text is rephrased using different words and sentence structures while preserving the same meaning.

H WebCrawl test sets

To acquire the data, we follow the process outlined below:

- We extract snippets from web pages and use an in-house sentence segmentation tool to split them into individual sentences.
- We employ sensitive word filters, language identification tools, length ratio checks, and perplexity scores to filter out sentences of lower quality.
- We utilize Google Translator to obtain translations of the sentences, with a primary focus on the Chinese⇔English directions.
- We calculate COMETkiwi scores and retain sentences with scores below 65.

In this way, we collected a total of 889 Chinese sentences and 1195 English sentences as our final test set, named *WebCrawl test sets*.

I Case study

These issues can be roughly categorized into three types: 1) professional terms expressed in abbreviations, 2) ill-informed text, and 3) complex sentences containing repeated segments. For instance, GoogleMT struggles to understand the meaning of "Art.18 GDPR" and "厦马", resulting in incorrect translations such as "艺术。GDPR第18条" and "Xiamen Horse Racing". Besides, GoogleMT fails to produce suitable translations for ill-informed text (e.g., "no. 6 fi lbert"), code, and repetitive text (e.g., "啊啊啊啊啊啊啊"). In contrast, TIM demonstrates its superiority in handling such issues, which can be attributed to its enhanced ability to comprehend rare, specialized words, and informal texts.

J Setup

For CoDec, we use MicroMT as the NMT system, and TIM as the MT-oriented LLM. As an important baseline, we incorporate our cooperative decoding into the Hybrid Threshold approach (Hendy et al., 2023), which we refer to as CoDec(QE). In particular, we set the threshold as the 50th percentile of COMETkiwi scores of MicroMT (Hendy et al., 2023) and set k as 16. We use the MT-oriented LLM to generate the translation only when the COMETkiwi score of the NMT translation is under the threshold. We use beam search with a beam size of 4 for TIM during inference. The decoding and speed measurement processes are performed on a single A100 GPU.

System	Translation
	Terminology/abbreviations
Source	Art. 18 GDPR: Right to restriction of data processing if the requirements Art. 18 para 1 lit. a to d are fulfilled.
GoogleMT	艺术。 GDPR 第 18 条：如果满足第 18 条的要求，则有权限制数据处理。 18 段 1 字。 a到d均满足。
TIM	《通用数据保护条例》第 18 条：如果满足第 18 条第 1 款 a 至 d 项的要求，则有权限制数据处理。
Source	刚好准备去厦门旅游，还能顺便跑个厦马,
GoogleMT	I was just about to go to Xiamen for a trip, and I could also run the Xiamen Horse Racing.
TIM	Just ready to go to Xiamen for a trip, and can also run the Xiamen Marathon,
	Ill-informed text
Source	Use a no. 6 fi lbert to create the illusion of the rungs and the back of the chair on the left.
GoogleMT	使用否。 6 榛子创建左边的梯级和椅背的幻觉。
TIM	用 6 号画笔在左边的椅子上创造梯级和椅背的错觉。
Source	批《道路机动车辆生产企业及产品公告》中，江淮
GoogleMT	In the batch of "Announcement of Road Motor Vehicle Manufacturers and Products", JAC
TIM	In the "Road Motor Vehicle Manufacturers and Products Announcement", Jianghuai
	Complex，Repetition-containing
Source	let mut v = vec![10, 20, 30]; let handle = thread::spawn(‖ v.push(10););
GoogleMT	让 mut v = vec![10, 20, 30];让句柄 = thread::spawn(‖ v.push(10););
TIM	let mut v = vec! [10,20,30]; let handle = thread::spawn (‖ v.push (10););
Source	拯救剧荒\《爱之全蚀》啊啊啊啊啊啊啊两个亲亲怪！！
GoogleMT	Save the drama\“Total Eclipse of Love” Ahhhh

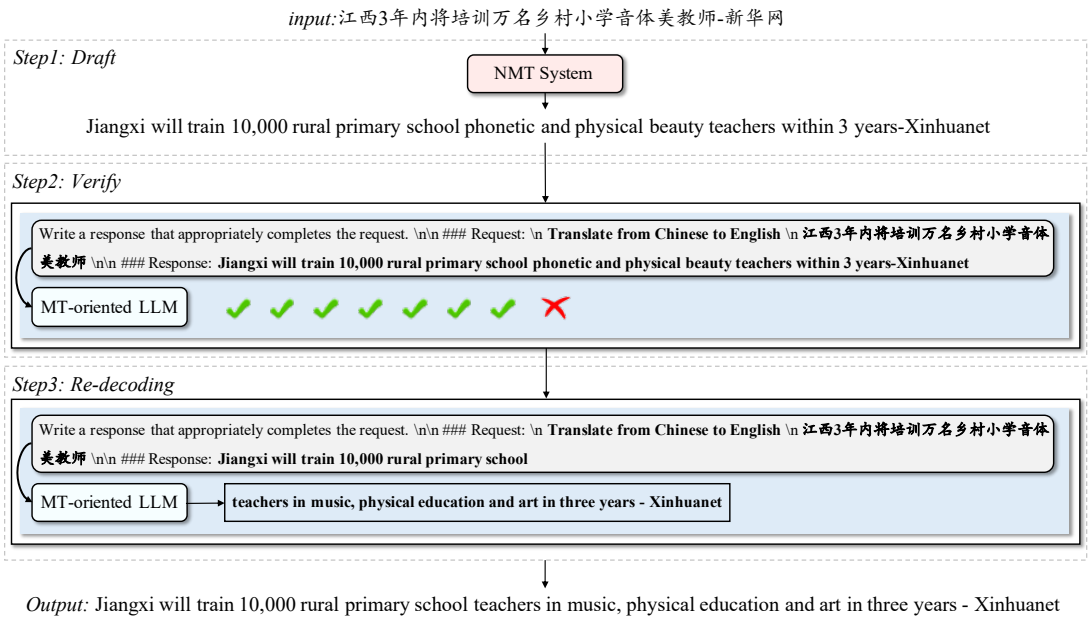

Figure 12: **Cooperative Decoding.** The NMT model generates the initial translation (referred to as *draft*), and the MT-oriented LLM assesses the quality of the *draft* and takes over from the error position, performing verification and re-decoding steps (*Verify* and *Re-decoding*).

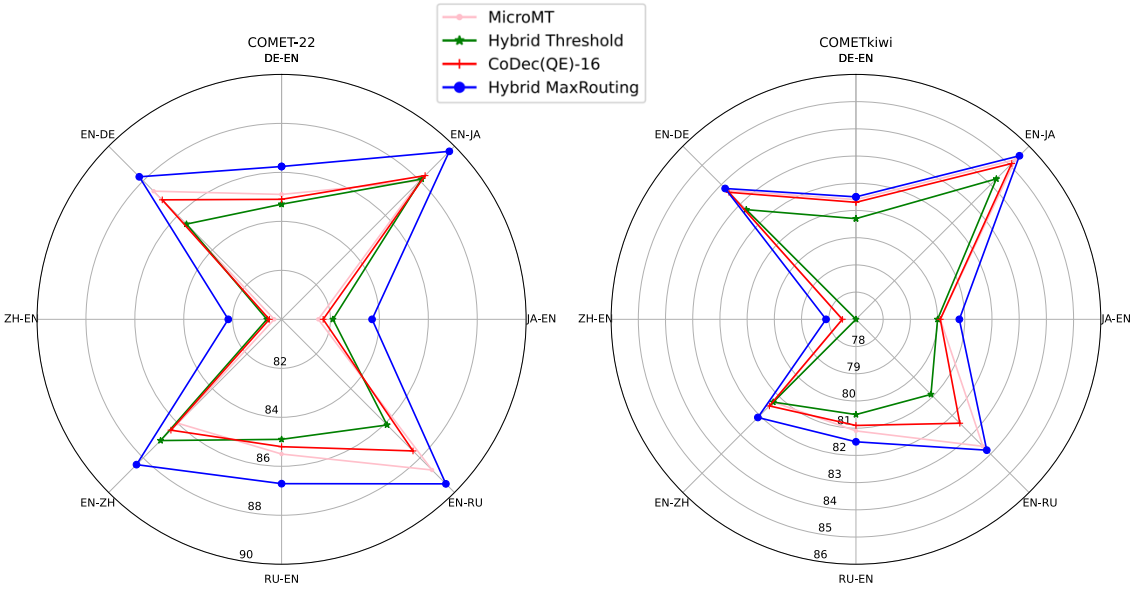

Figure 13: **Comparison of MicroMT and different hybrid frameworks**. We report COMET and COMETkiwi scores.