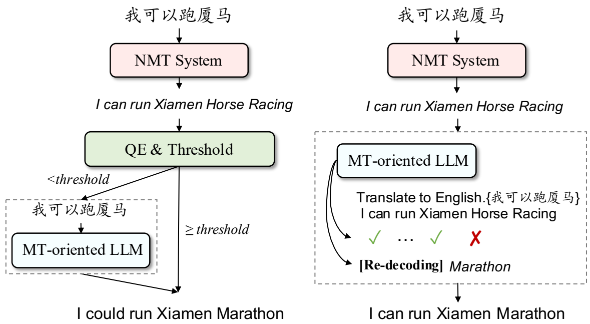
Improving Machine Translation with Large Language Models: A Preliminary Study with Cooperative Decoding

Anonymous ACL submission

Abstract

Contemporary translation engines based on the encoder-decoder framework have made significant strides in development. However, the emergence of Large Language Models (LLMs) has disrupted their position by presenting the potential for achieving superior translation quality. To uncover *the circumstances in which LLMs excel and explore how their strengths can be harnessed to enhance translation quality*, 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 alone can effectively address all the translation issues, but MT-oriented LLMs show promise as a complementary solution to NMT systems. Building upon these insights, we 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. Experimental 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 the field of machine translation.



(a) Hybrid Threshold (b) Our CoDec

Figure 1: **Comparison between the Hybrid Threshold and CoDec frameworks**. CoDec is more efficient than Hybrid Threshold as it eliminates the need for an extra quality evaluation module and autoregressive generation of the whole translation using MT-oriented LLM.

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 and 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 proficient in addressing the majority of translation requirements. More recently, with the emergence of generative large language models (LLMs), the position of traditional NMT models has been challenged (Brown et al., 2020; OpenAI, 2023). While commercial LLMs like OpenAI's GPT-4 currently perform well in translation (Hendy et al., 2023; Zhu et al., 2023; Lin et al., 2022; Agrawal et al., 2022), they are constrained by their interface nature, thereby limiting further customization and improvement due to privacy concerns in industrial applications. 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).

In this context, this study aims to investigate the following research questions: *In which scenarios do MT-oriented LLMs demonstrate superior performance to conventional NMT models, and how can we leverage the strengths of the two paradigms to enhance translation quality?*

To begin, we conduct a comprehensive analysis 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, Hendy et al. (2023) introduced the Hybrid Threshold approach (Figure 1(a)), 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, an alternative translation is generated using a GPT-like model. However, this approach faces two primary challenges. First, existing reference-free metrics struggle to align with human judgment, resulting in inaccuracies being propagated (Freitag et al., 2021, 2020; Ma et al., 2019; Rei et al., 2022a). Second, the integration of neural quality estimation modules and the sequential execution by LLMs leads to increased decoding time (Tay et al., 2023; Xu et al., 2021), which poses concerns for efficient translation in practical applications.

To address the above issues, we propose an efficient implementation approach for system ensembles called **Co**operative **Dec**oding (CoDec). As illustrated in Figure 1(b), 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 both an evaluator and a refiner, which firstly evaluates the draft from a language modeling perspective, and then 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 on the WMT22 test sets and a newly collected test set, WebCrawl, to identify the strengths and weaknesses of traditional NMT systems and MT-oriented LLMs, finding that MT-oriented LLMs can complement NMT systems.
- We present CoDec, a novel hybrid framework that synergizes the strengths of NMT systems and MT-oriented LLMs. By harnessing the complementary capabilities of MT-oriented LLMs, CoDec effectively overcomes the limitations of traditional NMT systems. We promise to release the code, data, and translations of different systems upon acceptance.
- We evaluate the performance of CoDec on various test sets. Our CoDec, without the need for an additional quality estimation module, achieves competitive or even better performance than Hybrid Threshold. Furthermore, CoDec offers a significant acceleration advantage, achieving an acceleration ratio of approximately 2x compared to directly using LLM for generation.

2 Related Work

2.1 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.

2.2 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.

System	COMET	COMETk.	COMET	COMETk.	COMET	COMETk.	COMET	COMETk.
	DE⇒EN		*EN⇒DE*		*ZH⇒EN*		*EN⇒ZH*	
WMT-Best	85.0	81.4	87.2	83.6	81.0	77.7	86.7	82.0
GoogleMT	**85.8**	**81.8**	**88.1**	**84.1**	**82.7**	**79.3**	**88.2**	**82.7**
MicroMT	85.1	81.4	87.4	83.7	80.3	77.5	86.0	81.3
BayLing-7B	83.2	80.1	82.1	79.2	77.5	75.1	84.4	79.6
TIM-13B	84.4	81.0	86.4	83.1	80.8	77.8	87.6	82.3
	RU⇒EN		*EN⇒RU*		*JA⇒EN*		*EN⇒JA*	
WMT-Best	86.0	81.7	**89.5**	**84.4**	81.6	80.3	89.3	85.8
GoogleMT	**86.6**	**82.0**	**89.5**	84.2	**84.0**	**81.7**	**90.2**	**86.5**
MicroMT	85.5	81.1	88.7	83.6	81.5	80.1	88.0	85.3
BayLing-7B	82.5	79.3	74.7	70.6	72.2	72.5	71.2	73.5
TIM-13B	84.2	80.8	86.7	82.5	80.8	79.8	87.5	84.5

Table 1: **Experimental results on the WMT22 test sets**. MT-oriented LLMs have the potential to achieve comparable performance to commercial NMT systems, eliminating the need for rule-based engineering techniques.

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.

3 Preliminary Experiments

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

3.1 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] due to their strong performance and high reproducibility. 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. We follow previous studies (Hendy et al., 2023; Zhu et al., 2023; Zeng et al., 2023; Zhang et al., 2023) to 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) in Table 8 in Appendix.

3.2 Analyses on WMT22 test sets

To prevent data leakage (Garcia et al., 2023), we analyze the WMT22 test sets. Detailed statistics 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. They consistently outperform the *WMT winner* in most of the language pairs, highlighting the robust capabilities of these well-established translation engines. 2) Despite the existing performance gap, MT-oriented LLMs still have untapped potential for further improvement. Notably, *TIM* outperforms *BayLing* by a significant margin across all language pairs. Moreover, *TIM* exhibits slightly inferior performance compared to *MicroMT* on most test sets. This suggests that employing more effective fine-tuning methods with large amounts of high-quality parallel data can enhance the translation capabilities of MT-oriented LLMs, making them close to commercial NMT systems.

[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

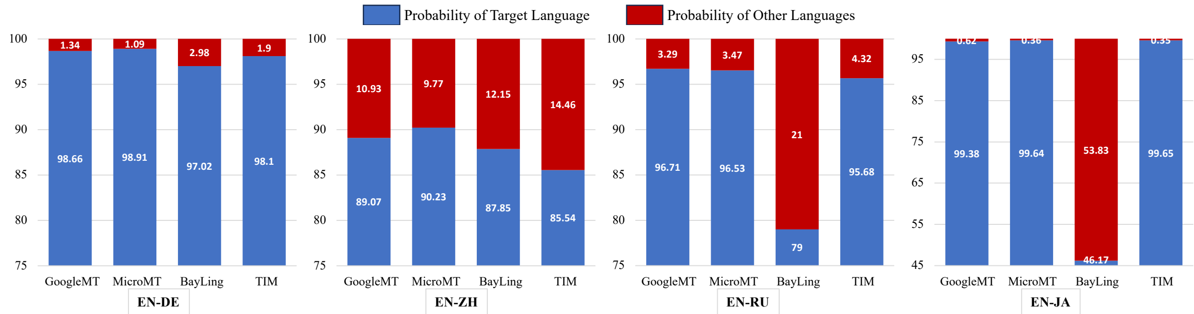


Figure 2: **Off-target rates (%) of translations.** MT-oriented LLMs (i.e., BayLing and TIM) exhibit a higher prevalence of off-target translations than NMT systems.

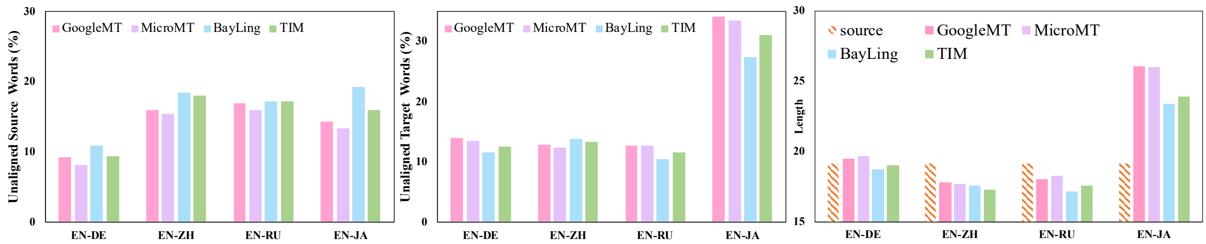


Figure 3: **Comparison of unaligned source words, unaligned target words, and the length of translations.** MT-oriented LLMs consistently generate translations that are noticeably shorter in length and have a higher occurrence of unaligned source words across the test sets when compared to NMT models.

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

As depicted in Figure 2[6], the MT-oriented LLMs tend to produce translations with higher off-target rates compared to 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 notably lower off-target rates in EN⇒RU and EN⇒JA compared to *BayLing*. We speculate that this can be attributed to TIM's incorporation of corresponding training data, which enhances its ability to handle language switching and produce more accurate translations.

Unaligned Source/Target Words. To assess 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. The left portion of Figure 3 illustrates that the MT-oriented LLMs incur a notably larger number of unaligned source words across 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), indicating the possibility of increased paraphrasing or a higher degree of inadequacy, such as omitted or inserted content. However, back to the middle part of Figure 3, 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.

Additionally, we calculate the average word count in the generated translations. As depicted in Figure 3, MT-oriented LLMs tend to produce

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

[6]Due to limited space, we only present the results for English-to-Many translations here. The results for Many-to-Enligsh can be found in Figure 6 in Appendix.

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 4: **Examples of free translation generated by MT-oriented LLM.** MT-oriented LLMs often produce shorter translations with significant paraphrasing, maintaining the original meaning while using different words and sentence structures.

Model	**ZH⇒EN** COMET/COMETk.	**EN⇒ZH** COMET/COMETk.
GoogleMT	64.4/59.1	71.2/60.5
MicroMT	59.2/57.4	68.9/62.9
TIM	**65.1/61.9**	**74.6/64.9**

Table 2: **Experimental results on WebCrawl test sets**. LLMs hold promise as potential fallback systems when NMT systems fail to meet quality expectations.

shorter sentences, utilizing concise and precise language. Humans often use concise language, especially in conversations, which is abundant in the training corpus of LLMs. This influence may result in LLMs generating shorter translations.

Figure 4 presents several examples that highlight translation differences. For instance, the phrase “frustration shouting” should be translated as “大喊大叫 (scream in frustration)”. While *MicroMT* aims for fidelity by using translation augmentation segments like “我非常沮丧 (I feel extremely frustrated)”, *TIM* demonstrates a better understanding of the entire sentence and provides 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 generate shorter translations with substantial paraphrasing, where the original text is rephrased using different words and sentence structures while preserving the same meaning.

3.3 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 directions. To acquire the data, we follow the process outlined in Appendix E.

Main Results. Similarly, we compute COMET and COMETkiwi[7] for NMT systems and TIM on the WebCrawl test sets[8]. As shown in Table 2, 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.

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[9] to calculate the perplexity for the two groups. The perplexity for sentences in which *GoogleMT* excels is 38.61, whereas for sentences in which *TIM* performs better, it is 45.51. The MT-oriented LLM showcases superior proficiency in handling complex source language sentences, as reflected by

[7]We also show the other metrics (e.g., USW) in Table 7 in Appendix, the phenomena observed are consistent with the analysis in Section 3.2, demonstrating the generalizability of our findings.

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

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

System	Translation
	Terminology/abbreviations
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。
	Complex，Repetition-containing
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” Ahhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh being to To to to toto to Eclips e Love！!
TIM	Save the drama drought \ “Total Eclipse of Love” ah ah ah ah ah two kissing monsters!!

Table 3: **Case Study**. We present examples of several translation challenges that pose difficulties for NMT systems but are effectively mitigated by MT-oriented LLMs.

higher perplexity scores.

Case Study. In Table 3, 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.

3.4 Discussion

The analysis in Section 3.2 demonstrates the accuracy of commercial NMT systems, likely due to their extensive training and cross-attention capability. MT-oriented LLMs, known for their paraphrastic nature, can further enhance NMT's ability to handle figurative text translations. Moreover, MT-oriented LLMs excel on WebCrawl test sets (Section 3.3), particularly with specialized terminology and ill-formed sentences. This suggests that an effective hybrid framework combining NMT and LLMs can handle challenging input domains and figurative text. Based on these insights, 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?*

Hybrid Threshold (Hendy et al., 2023) employs the NMT model as the primary translation system, with a quality estimation module (e.g., COMETkiwi) to assess the translation. If the quality falls below a certain threshold, the GPT-like model is used as an alternative translation engine. However, it faces two main challenges: autoregressive decoding latency and reliable quality estimation. It is crucial for the practical implementation of the hybrid approaches to ensure efficient decoding and high translation quality.

4 Cooperative Decoding

We propose an innovative cooperative decoding approach[10], which leverages an NMT model as a pretranslation model and incorporates an MT-oriented LLM as a quality assessment module and fallback model if needed. 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 parallel computing. 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

[10]We illustrate clear processing of cooperative decoding in Figure 7 in Appendix.

Method	ZH⇒EN COMET	COMETk.	Token/s	Speedup	Ratio	EN⇒ZH COMET	COMETk.	Token/s	Speedup	Ratio
GoogleMT	76.8	72.8	-	-	-	81.9	74.5	-	-	-
TIM	75.6	72.3	21.8	1.0×	-	83.0	76.0	20.7	1.0×	-
CoDec-4	76.7	73.0	28.5	1.3×	24.44	83.3	**76.1**	24.1	1.2×	21.64
CoDec-8	**77.1**	73.2	32.0	1.5×	38.83	**83.4**	**76.1**	25.8	1.3×	32.69
CoDec-16	**77.1**	**73.3**	38.7	1.8×	55.11	83.1	76.0	29.7	1.4×	46.06
CoDec-32	**77.1**	73.2	47.9	2.2×	67.36	83.0	75.8	33.6	1.6×	57.06
CoDec-64	77.0	73.1	57.7	2.7×	76.23	82.7	75.6	38.7	1.9×	66.25
CoDec-128	77.0	73.0	73.5	3.4×	84.29	82.6	75.4	45.5	2.2×	74.35

Table 4: **Effect of different values of k (Eq. 1) for CoDec**. We present the results on ZH⇒EN and EN⇒ZH including COMET-22, 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. The choice of k should be considered to strike a balance between performance and efficiency.

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

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

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.

5 Experiments

5.1 Main Results

We merge the WMT22 and WebCrawl test sets to simulate the distribution of translation requests in real-world scenarios. For CoDec, we use GoogleMT as the NMT system, and TIM as the MT-oriented LLM. Detailed setup can be found in Appendix F.

Effect of different values of k. Intuitively, as k increases, cooperative decoding can accept a wider range of tokens in NMT translations during the verification stage. As a result, less content needs to be re-decoded by LLMs, leading to a reduction in processing time. Here, we examine the performance of CoDec under various values of k.

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 larger k, *CoDeC-128* achieves a 3.4x and 2.2x speedup over *TIM* in ZH⇔EN. This signifies that CoDec effectively reduces decoding latency while maintaining translation quality. Besides, our CoDec-(*) models exhibit superior performance compared to both *GoogleMT* and *TIM*. This highlights the potential of cooperative decoding in improving translation accuracy and overall system performance. Moreover, models with lower values of k, such as *CoDec-8*, achieve better translation quality, suggesting that the choice of k should be considered to strike a balance between performance and efficiency.

CoDec vs. Hybrid Threshold. In our comparison between CoDec and Hybrid Threshold, we utilize different Quality Estimation (QE) methods, including *HT(Random)*, where 50% of GoogleMT's translations are randomly replaced with TIM's translations, *HT(BLEURT-12)*, which uses BLEURT-20-D12[11] as the QE method; *HT(BLEURT-20)*, which employs BLEURT-20[12] as the QE method; and *HT(COMETk.)*. Additionally, CoDec is integrated into the Hybrid Threshold pipeline as a comparative system, referred to as *HT(COMETk.) w/ CoDec*. Furthermore, we follow Hendy et al. (2023) to use Hybrid Max-Routing to establish an upper bound by selecting

[11] https://huggingface.co/lucadiliello/BLEURT-20-D12
[12] https://huggingface.co/lucadiliello/BLEURT-20

Model	ZH⇒EN COMET/COMETk.	EN⇒ZH COMET/COMETk.
GoogleMT	76.8/72.8	81.9/74.5
TIM	75.6/72.3	83.0/76.0
HT(Random)	76.2/72.5	82.4/75.2
HT(BLEURT-12)	76.3/72.8	82.6/75.1
HT(BLEURT-20)	76.3/72.8	82.7/75.2
HT(COMETk.)	76.5/73.1	83.3/**76.2**
w/ CoDec	**77.1/73.3**	**83.4/76.2**
CoDec-8	**77.1**/73.2	**83.4**/76.1
Max-Routing	77.4/74.3	84.0/76.5

Table 5: **Comparison among CoDec-8 and Hybrid Threshold with different QE methods**. Different QE methods in Hybrid Threshold (HT) show varying performances, whereas CoDec surpasses most HT models. Our CoDec achieves a better balance between efficiency and effectiveness.

the best translation from either system based on the COMETkiwi.

The performance comparison in Table 5 reveals a notable performance disparity between *GoogleMT* and *Max-Routing*. This result supports our assertion that MT-oriented LLMs can play a crucial role as reliable fallback systems for NMT systems. Moreover, the different QE modules employed in Hybrid Threshold yield varying performances, highlighting the dependence of Hybrid Threshold's performance on the precision of the QE modules and the quality of LLM translations used as replacements. In contrast, *CoDec-8* surpasses most of the Hybrid Threshold models and achieves competitive results with *HT(COMETk.) w/ CoDec*, suggesting that the QE modules may not be necessary. The findings validate that our approach achieves a better balance between efficiency and effectiveness, resulting in enhanced translation quality without compromising system efficiency.

5.2 Human Evaluation

In addition, we carry out a human evaluation on the WebCrawl EN⇒ZH dataset. A total of 300 sentences are randomly selected from the test set, and two individuals are asked to evaluate the translations produced by GoogleMT, HT(COMETk.), and our CoDec-8. We use the commonly used pairwise comparison method to count the number of better, similar, and worse translations from System 1 rather than System 2. The result of CoDec vs. GoogleMT is 144:115:41, while the result of CoDec vs. HT(COMETk.) is 106:130:64. It shows that our CoDec significantly outperforms the commercial NMT system and performs better than the Hybrid Threshold without an additional quality evaluation module.

Model	DE⇒EN COMET/ChrF	DE⇒EN Suc.	ZH⇒EN COMET/ChrF	ZH⇒EN Suc.
Lingua Custodia	73.5/61.8	62.2	60.9/32.6	74.7
$\text{UEDIN}_{\text{LLM}}$	81.3/60.0	58.8	**75.7/41.2**	75.3
GoogleMT	80.3/54.3	55.0	75.3/41.0	67.1
TIM w/o term	79.6/54.0	54.1	73.8/38.5	58.6
TIM w/ term	**82.3**/65.2	**82.5**	73.4/39.4	**85.0**
CoDec-8	80.7/56.1	59.0	75.3/41.0	76.4

Table 6: **Performance on WMT23 terminology translation**. "Suc." denotes Terminology Success Rate. Our CoDec combines NMT's superior translation quality with the constrained translation capabilities of MT-oriented LLMs.

5.3 Terminology Translation

Unlike conventional NMT models, MT-oriented LLMs enable them to exploit instructions to handle various translation scenarios. Here, we apply CoDec to assess the effectiveness of incorporating instructions in a dedicated terminology translation test set obtained from WMT23[13]. The result is shown in Table 6, evaluated by COMET, ChrF, and Terminology Success Rate. The data statistics and details of baselines can be found in Appendix G.

The results indicate that the use of terminology information in instructions, as demonstrated by *TIM w/ term*, enables MT-oriented LLMs to achieve constrained machine translation, resulting in more accurate domain-specific terminology in the translated output. When compared to *Lingua Custodia* and $\textit{UEDIN}_{\text{LLM}}$ (Semenov et al., 2023), *CoDec-8* combines the advantages of higher translation quality offered by NMT and the constrained translation capabilities of MT-oriented LLMs. This combination leads to higher-quality translations while maintaining a higher Terminology Success Rate.

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.

[13]https://wmt-terminology-task.github.io/

7 Limitations

This paper primarily concentrates on enhancing translation performance for medium and high-resource language pairs. Further investigation is required to analyze the translation characteristics of different systems in low-resource languages, which we defer to future research.

Additionally, the draft translations were validated by directly utilizing the top-k candidates predicted by the target MT-oriented LLM. We acknowledge that the implementation of more meticulously designed token-level validation methods has the potential to further enhance CoDec, and we consider it as an avenue for future exploration.

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.

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.

Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. 2023. Findings of the wmt 2023 shared task on machine translation with terminologies. In *Proceedings of the Eight Conference on Machine Translation (WMT)*. 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.

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[14], such as the newstest2017-2021 of German⇔English, Chinese⇔English, Russian⇔English, and Jappanese⇔English. In addition, we have incorporated high-quality bilingual sentence pairs in Chinese⇔English, German⇔English, and Russian⇔English, resulting in a total of two million sentences in our training data. 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[15] 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 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[16]. We use jieba[17] and MeCab[18] for Chinese and Japanese, respectively. We use *awesome-*

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

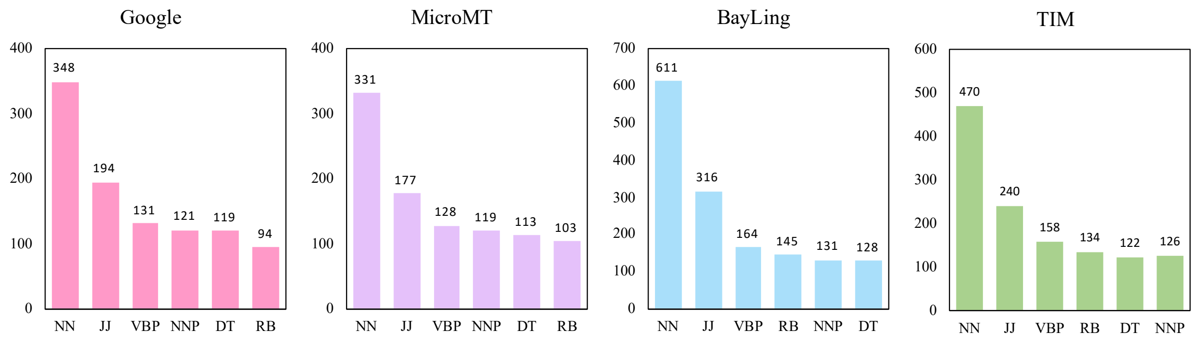


Figure 5: **POS tags of unaligned source words.** We show the top-6 POS tags of USW on the WMT22 EN⇒ZH test set, and the incremental USW of MT-oriented LLMs mainly lies in nouns (NN) and adjectives (JJ).

align[19] (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 5 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 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.

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

System	**GoogleMT**	**MicroMT**	**TIM**
	ZH⇒EN		
#Length	56.81	51.38	52.09
Off-Target↓	1.08%	1.04%	1.82%
USW↓	13.87%	13.70%	16.52%
UTW↓	31.29%	25.08%	27.91%
	EN⇒ZH		
#Length	48.51	47.99	46.57
Off-Target↓	15.55%	14.08%	22.41%
USW↓	21.76%	20.23%	25.70%
UTW↓	16.55%	13.64%	16.39%

Table 7: **Experimental results on WebCrawl test sets**. The phenomena observed are consistent with the analysis in Section 3.2, demonstrating the generalizability of our findings.

In this way, we collected a total of 889 Chinese sentences and 1195 English sentences as our final test set, named *WebCrawl test sets*. We hire 2 annotators who have degrees in English Linguistics to annotate translations. Before formal annotation, annotators were asked to annotate 100 samples randomly extracted from the dataset, and based on average annotation time we set a fair salary (i.e., 30 dollars per hour) for them.

F Setup

For CoDec, we use GoogleMT as the NMT system, and TIM as the MT-oriented LLM. In particular, we set the threshold as the 50th percentile of COMETkiwi scores of GoogleMT (Hendy et al., 2023). 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	ChrF	SacreBLEU	ChrF	SacreBLEU	ChrF	SacreBLEU	ChrF	SacreBLEU
	DE⇒EN		*EN⇒DE*		*ZH⇒EN*		*EN⇒ZH*	
WMT-Best	58.5	33.4	64.6	**38.4**	**61.1**	**33.5**	41.1	44.8
GoogleMT	**59.1**	**34.1**	**64.7**	37.5	60.0	29.4	**45.8**	**50.5**
MicroMT	58.8	33.9	64.7	37.5	60.0	29.4	**45.8**	**50.5**
BayLing-7B	53.6	28.2	53.6	25.7	49.9	20.3	34.5	38.2
TIM-13B	56.9	31.7	60.8	33.2	56.8	26.9	42.4	46.9
	RU⇒EN		*EN⇒RU*		*JA⇒EN*		*EN⇒JA*	
WMT-Best	68.9	45.1	58.3	32.4	49.8	24.8	36.8	27.6
GoogleMT	**69.1**	**45.7**	59.5	34.3	**51.8**	**26.2**	**37.6**	**28.2**
MicroMT	**69.1**	**45.7**	**59.6**	**34.9**	49.5	24.6	34.8	25.1
BayLing-7B	60.4	34.7	35.5	14.8	34.7	11.6	9.6	4.5
TIM-13B	65.7	40.4	54.6	28.5	46.3	21.6	29.6	19.7

Table 8: Experimental results on the WMT22 test sets.

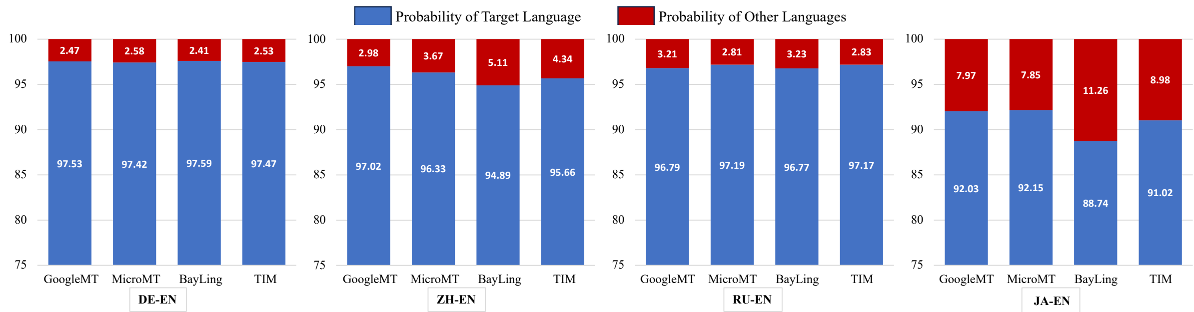


Figure 6: **Off-target rates (%) of translations.** MT-oriented LLMs exhibit a higher prevalence of off-target translations than NMT systems.

G Terminology Translation

Terminology translation is an extensively encountered application scenario, where the NMT (Neural Machine Translation) model is expected to precisely handle the provided domain-specific terminology. In this experiment, we use the prompt "`{srcWord} means {tgtWord}. Translate the following sentences from {src} to {tgt}, and muse use the given word translations.{line}`" for inference of TIM. The numbers of sentences on Zh⇒En and De⇒En are 2640 and 2963, respectively. The average numbers of terms per segment on Zh⇒En and De⇒En are 3.8 and 1.1, respectively. We only highlight a few systems that achieved the best performance on specific metrics in the competition findings (Semenov et al., 2023). Lingua Custodia, which utilizes a specialized Transformer architecture to ensure the inclusion of given terminology in the translation. Additionally, the $\text{UEDIN}_{\text{LLM}}$ employs ChatGPT with prompts specifically designed for terminology translation.

H Different from traditional NMT with additional language models

Traditional language models, such as causal language models are usually used as decoder initialization or reranking to improve fluency. We do not consider the prediction probabilities of LLMs during the decoding process of NMT. Instead, we treat LLMs as independent translation systems and introduce speculative execution as a fusion approach for NMT systems and MT-oriented LLMs.

I About speedup

The time consumption of the Hybrid Threshold is the sum of the inference time for both the NMT systems and the MT-oriented LLM, whereas the CoDec requires only the inference time of the NMT systems and a small amount of calculation of the LLM. Considering the relatively negligible time consumption of Google Translate, we did not specifically factor in its inference time in our analysis, as it does not significantly impact the overall performance comparison.

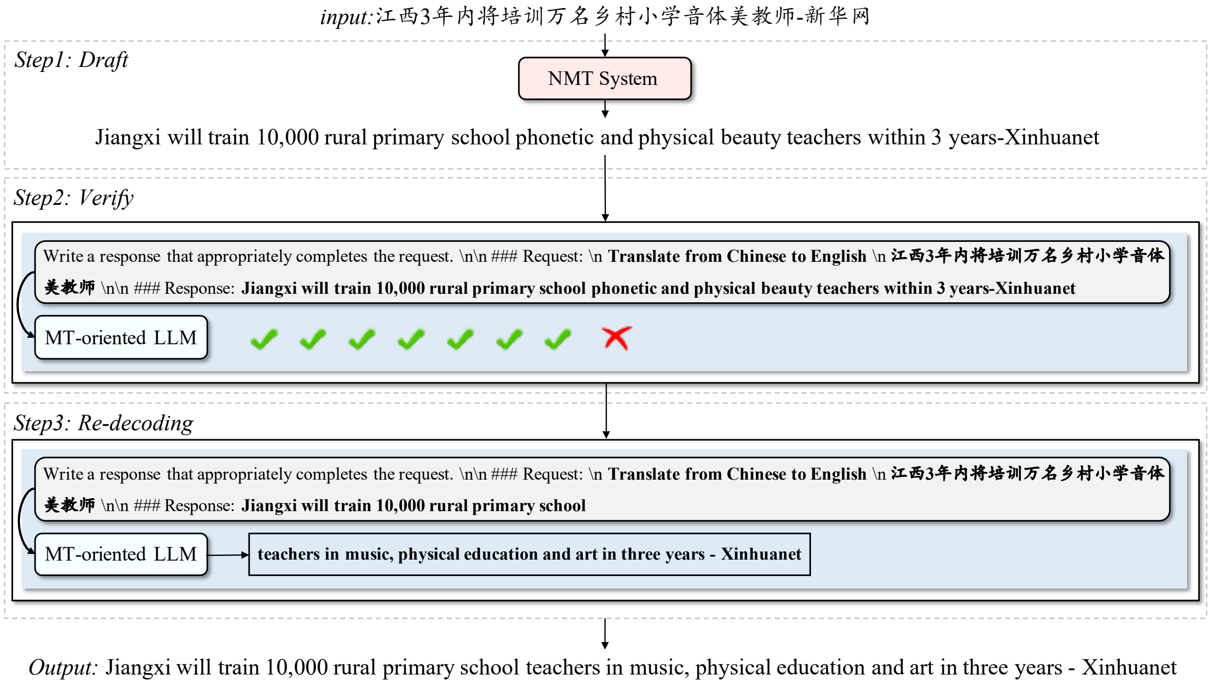


Figure 7: **Cooperative Decoding.** The NMT model generates the initial translation (referred to as *draft*), and the MT-oriented LLM assesses the quality of the *draft* and takes over from the error position, performing verification and re-decoding steps (*Verify* and *Re-decoding*).

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。
	Ill-informed text
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););

Table 9: **Case Study**. We present examples of several translation challenges that pose difficulties for NMT systems but are effectively mitigated by MT-oriented LLMs.