

DeMPT: Decoding-enhanced Multi-phase Prompt Tuning for Making LLMs Be Better Context-aware Translators

Anonymous ACL submission

Abstract

Generally, the *decoder-only* large language models (LLMs) are adapted to context-aware neural machine translation (NMT) in a concatenating way, where LLMs take the concatenation of the source sentence (i.e., intra-sentence context) and the inter-sentence context as the input, and then to generate the target tokens sequentially. This adaptation strategy, i.e., concatenation mode, considers intra-sentence and inter-sentence contexts with the same priority, despite an apparent difference between the two kinds of contexts. In this paper, we propose an alternative adaptation approach, named **Decoding-enhanced Multi-phase Prompt Tuning (DeMPT)**, to make LLMs discriminately model and utilize the inter- and intra-sentence context and more effectively adapt LLMs to context-aware NMT. First, DeMPT divides the context-aware NMT process into three separate phases. During each phase, different continuous prompts are introduced to make LLMs discriminately model various information. Second, DeMPT employs a heuristic way to further discriminately enhance the utilization of the source-side inter- and intra-sentence information at the final decoding phase. Experiments show that our approach significantly outperforms the concatenation method, and further improves the performance of LLMs in discourse modeling. We will release our code and datasets on GitHub.

1 Introduction

Context-aware neural machine translation (NMT) goes beyond sentence-level NMT by incorporating inter-sentence context at the document level (Zhang et al., 2018; Miculicich et al., 2018; Voita et al., 2018, 2019b,a; Bao et al., 2021; Sun et al., 2022), aiming to address discourse-related challenges such as zero pronoun translation (Wang et al., 2019), lexical translation consistency (Lyu et al., 2021, 2022), and discourse structure (Hu and Wan, 2023). A recent paradigm shift has been witnessed in context-

aware NMT with the emergence of the *decoder-only* large language models (LLMs) (BigScience, 2022; Google, 2022; MetaAI, 2023b,a; OpenAI, 2023). These generative language models, trained on massive data, have gained significant attention in the natural language processing (NLP) community. In adapting LLMs to context-aware NMT, a common strategy involves concatenating multiple source sentences as a prefix and generating translations token-by-token, relying on the prefix and previously predicted target tokens, as shown in Figure 1 (a). However, a critical observation of this strategy reveals a potential drawback – the equal prioritization of the inter- and intra-sentence contexts during token generation. Importantly, the intra-sentence context inherently contains richer parallel semantic information with the target sentence and should be given a higher priority than the inter-sentence context. Consequently, we propose that separately modeling and utilizing the inter- and intra-sentence contexts should explicitly inform LLMs of the document-level context and the current sentence itself, thus being able to prevent the misallocation of attention weights to source-side tokens (Bao et al., 2021; Li et al., 2023). Inspired by the success of prompt tuning (Li and Liang, 2021; Liu et al., 2022; Tan et al., 2022), our alternative approach, named Decoding-Enhanced Multi-phase Prompt Tuning (DeMPT), aims to enhance LLMs’ adaptability to context-aware NMT, as shown in Figure 1 (b).¹

Specifically, we divide the whole procedure of context-aware NMT into three phases: inter-sentence context encoding, intra-sentence context encoding, and decoding. Following Li and Liang (2021); Liu et al. (2022), we sequentially and dif-

¹Following the findings of Bao et al. (2021), which indicate that source-side context is relatively more important for document-level MT compared to target-side context, we focus exclusively on source-side context in this paper. Nonetheless, we provide an additional discussion on integrating target-side context in Appendix M.

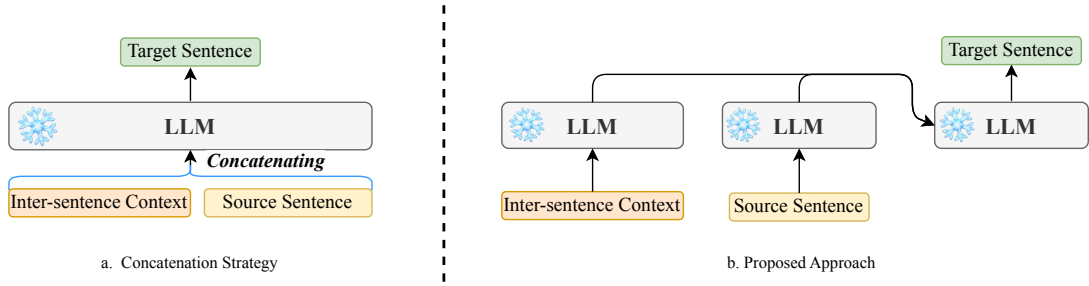


Figure 1: Comparison of different strategies for adapting LLMs to context-aware NMT. The concatenation strategy (*left*) treats inter-sentence and intra-sentence (referred to as the "source sentence" context in the figure) with equal importance. In contrast, our approach (*right*) divides context-aware NMT into three distinct phases, enabling LLMs to selectively model and leverage both inter- and intra-sentence contexts.

ferentially adapt LLMs for each phase, utilizing phase-specific trainable prompts. This phased tuning method enables LLMs to independently capture and model both inter- and intra-sentence contexts, facilitating a better understanding of their differences. Our approach splits the input into three parts without significantly increasing computational load, thus maintaining inference speed comparable to concatenation, as detailed in Appendix D.

Furthermore, during the decoding phase, we propose a heuristic method to emphasize the difference between inter- and intra-sentence contexts, and avoid *long-distance* issue when utilizing inter-sentence context. Specifically, at each decoding step, we use LLMs to predict the next token three times. The decoding states used for each prediction directly concatenate with the representations of two contexts in a discriminative manner. Finally, we combine three probability distributions to search for the next token as the output from the target vocabulary. This method enables LLMs to learn not only to properly capture inter-sentence context in addressing discourse-related issues but also to recognize a difference between inter- and intra-sentence contexts, allowing for effective utilization of both types of contexts.

Our contributions can be summarized as follows:

- We introduce a multi-phase prompt tuning approach that divides context-aware NMT into three phases, enabling LLMs to distinguish between inter- and intra-sentence contexts.
- We introduce an enhanced decoding method that discriminately utilizes both context types. This allows LLMs not only to properly capture inter-sentence context in addressing discourse-related issues, but also to be aware of the importance of the intra-sentence context.

- We validate our approach using llama-2-7b and bloomz-7b1-mt as foundation models, demonstrating its effectiveness across five translation directions. Extensive analyses further highlight the substantial enhancement in LLMs' ability for context-aware MT.

2 Methodology

In this section, we describe our decoding-enhanced multi-phase approach for adapting LLMs to context-aware NMT in details. Specifically, we break down the whole procedure of context-aware NMT into three phases (Section 2.1), i.e., inter-sentence context encoding, intra-sentence encoding, and decoding. Additionally, we discriminatively enhance the utilization of inter- and intra-sentence contexts during the decoding phase (Section 2.2). Finally, we describe our phase-aware prompts and training objective in Section 2.3 and Section 2.4, respectively.

For a given document pair (S, T) with K sentences, we will construct K training instances. Each training instance is denoted as a tuple (C, S, T) . Here $S = x|_k^{|S|}$ represents k -th current source sentence with $|S|$ tokens, i.e., intra-sentence context, and $T = y|_k^{|T|}$ is the k -th target sentence with $|T|$ tokens. C denotes the z previous sentences of S , i.e., the inter-sentence context of S . We denote the hidden size of the LLM as d , and L as the number of transformer layers within it.

2.1 Multi-phase Encoding and Decoding

We implement our approach based on deep prompt tuning (Li and Liang, 2021; Liu et al., 2022). Next, we use training instance (C, S, T) as an example to describe the multi-phase approach. Figure 2 illustrates the procedure of multi-phase prompt tuning.

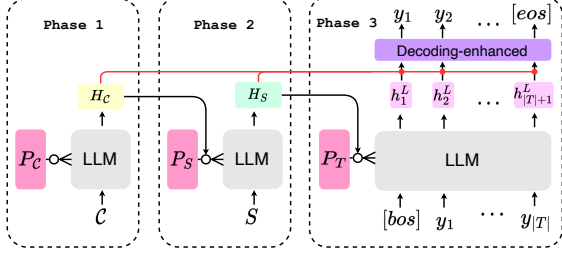


Figure 2: Illustration of pipeline of multi-phase prompt tuning LLM for context-aware NMT. Red lines illustrate the procedure of enhanced decoding phase.

Inter-sentence Context Encoding Phase. In the inter-sentence context encoding phase (Phase 1 in Figure 2), we first concatenate all sentences in \mathcal{C} into a sequence, and then utilize the LLM to encode \mathcal{C} by incorporating the trainable prompt:

$$H_C^{1:L} = \text{LLM}(\mathcal{C}, \mathbf{P}_C), \quad (1)$$

where $H_C^{1:L} \in \mathbb{R}^{L \times |\mathcal{C}| \times d}$ is the sequence of activations for \mathcal{C} , $\mathbf{P}_C \in \mathbb{R}^{L \times 2q \times d}$ is the current-phase trainable prompt, and q is a hyper-parameter for the length of the prompt. \mathbf{P}_C aims to adapt the LLM for better modeling the inter-sentence context. Same as basic deep prompting, at the l -th transformer block, we inject corresponding prompt in \mathbf{P}_C into encoding procedure of \mathcal{C} as follows:

$$H_C^l = \text{FFN}(\text{Multi-Attn}(\mathbf{K}_C, \mathbf{V}_C, \mathbf{Q}_C)), \quad (2)$$

$$\mathbf{Q}_C = H_C^{l-1}, \quad (3)$$

$$\mathbf{K}_C = [\mathbf{P}_C[l, :, q, :]; H_C^{l-1}], \quad (4)$$

$$\mathbf{V}_C = [\mathbf{P}_C[l, q, :, :]; H_C^{l-1}], \quad (5)$$

where $H_C^l \in \mathbb{R}^{|\mathcal{C}| \times d}$ is the output of the l -th transformer block. FFN and Multi-Attn are the feed-forward network and multi-head self-attention sub-layers, respectively.² $[\cdot; \cdot]$ and $[\cdot : \cdot]$ are the concatenating and slicing operations, respectively.

Intra-sentence Context Encoding Phase. In the intra-sentence context encoding phase (Phase 2 in Figure 2), the LLM encodes the intra-sentence context S by conditioning on the past activations of the inter-sentence context $H_C^{1:L}$ and trainable prompt:

$$H_S^{1:L} = \text{LLM}(S, H_C^{1:L}, \mathbf{P}_S), \quad (6)$$

where $H_S^{1:L} \in \mathbb{R}^{L \times |S| \times d}$ is the sequence of activations for S , and $\mathbf{P}_S \in \mathbb{R}^{L \times 2q \times d}$ denotes current-phase prompt. Similarly, at the l -th transformer

²For simplicity, we omit the normalization and residual operations in this paper.

block, we incorporate H_C and \mathbf{P}_S into the encoding procedure of S as follows:

$$H_S^l = \text{FFN}(\text{Multi-Attn}(\mathbf{K}_S, \mathbf{V}_S, \mathbf{Q}_S)), \quad (7)$$

$$\mathbf{Q}_S = H_S^{l-1}, \quad (8)$$

$$\mathbf{K}_S = [\mathbf{P}_S[l, :, q, :]; H_C^{l-1}; H_S^{l-1}], \quad (9)$$

$$\mathbf{V}_S = [\mathbf{P}_S[l, q, :, :]; H_C^{l-1}; H_S^{l-1}], \quad (10)$$

where H_S^l is output of the l -th transformer block, which fuses H_C^{l-1} , the $l-1$ layer output of the inter-sentence context encoding.

Decoding Phase. In the decoding phase (Phase 3 in Figure 2), given the past activations H_S and trainable prompt, we call the LLM again to generate the hidden state for predicting the probability of the target sentence:

$$H_T^{1:L} = \text{LLM}(T, H_S^{1:L}, \mathbf{P}_T), \quad (11)$$

where $H_T^{1:L} \in \mathbb{R}^{L \times |T| \times d}$ is the sequence of activations for T , and $\mathbf{P}_T \in \mathbb{R}^{L \times 2q \times d}$ is current-phase prompt. Similarly, we inject S and \mathbf{P}_T into the decoding procedure of T as follows:

$$H_T^l = \text{FFN}(\text{Multi-Attn}(\mathbf{K}_T, \mathbf{V}_T, \mathbf{Q}_T)), \quad (12)$$

$$\mathbf{Q}_T = H_T^{l-1}, \quad (13)$$

$$\mathbf{K}_T = [\mathbf{P}_T[l, :, q, :]; H_S^{l-1}; H_T^{l-1}], \quad (14)$$

$$\mathbf{V}_T = [\mathbf{P}_T[l, q, :, :]; H_S^{l-1}; H_T^{l-1}], \quad (15)$$

where $H_T^l \in \mathbb{R}^{|T| \times d}$ is the decoding state of the l -th transformer block. Finally, we refer the t -th decoding state as h_t^L (i.e., $H_T^L = h_t^L |_{t=1}^{|T|+1}$) which is used to predict the next token y_t :

$$p(y_t | S, \mathcal{C}, y_{<t}) = \text{Softmax}(h_t^L W), \quad (16)$$

where $W \in \mathbb{R}^{d \times |\mathcal{V}|}$ is parameter of LLM-Head layer and $|\mathcal{V}|$ is the vocabulary size.

2.2 Enhanced Decoding Phase

As shown in Figure 2, both the inter-sentence context representation $H_C^{1:L}$ and the intra-sentence context representation $H_S^{1:L}$ are used as keys and values when generating hidden states of next phase. Meanwhile, hidden states of decoding phase, i.e., $h_i^L |_{i=1}^{|T|}$ are used to predict next tokens. On the one hand, while the decoding hidden states incorporate both inter- and intra-sentence contexts, there is no explicit differentiation between the two when predicting next tokens. On the other hand, the inter-sentence context representation $H_C^{1:L}$ and decoding hidden states $H_T^{1:L}$ are mediated by hidden states

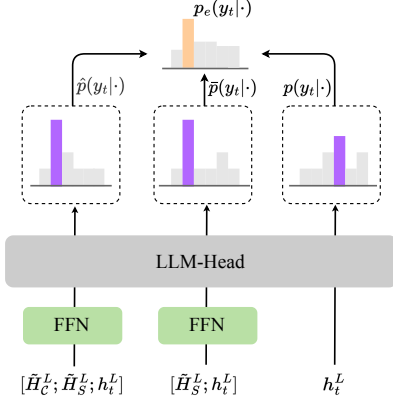


Figure 3: Illustration of the procedure of our proposed decoding-enhanced approach at the t -th decoding step.

of phases 2, i.e., $H_S^{1:L}$. This may result in a *long-distance* issue such that the inter-sentence context are not properly aligned by target-side tokens.

Therefore, to address above two issues, we propose an enhanced decoding phase with an aim to more effectively utilize both the inter- and intra-sentence contexts. Inspired by Kuang et al. (2018), we move both the two types of inter- and intra-sentence contexts closer to target words to achieve a tight interaction between them. Specifically, we concatenate the decoding states with the two types of representations to predict the next target words. As shown in Figure 3, the enhanced next word prediction p_e is a combination of three distributions:

$$p_e(y_t|S, \mathcal{C}, y_{<t}) = \lambda_1 \times \hat{p}(y_t|S, \mathcal{C}, y_{<t}) + \lambda_2 \times \bar{p}(y_t|S, \mathcal{C}, y_{<t}) + (1 - \lambda_1 - \lambda_2) \times p(y_t|S, \mathcal{C}, y_{<t}), \quad (17)$$

where λ_1 and λ_2 control the contribution of $\hat{p}(y_t|\cdot)$ and $\bar{p}(y_t|\cdot)$, respectively, which can be further formulated as:

$$\hat{p}(y_t|S, \mathcal{C}, y_{<t}) = \text{Softmax}(\hat{h}_t^L W), \quad (18)$$

$$\bar{p}(y_t|S, \mathcal{C}, y_{<t}) = \text{Softmax}(\bar{h}_t^L W), \quad (19)$$

$$\hat{h}_t^L = \text{FFN}([\tilde{H}_C^L; \tilde{H}_S^L; h_t^L]), \quad (20)$$

$$\bar{h}_t^L = \text{FFN}([\tilde{H}_S^L; h_t^L]), \quad (21)$$

where W is same as in Eq. 16, $\tilde{H}_S^L \in \mathbb{R}^d$ and $\tilde{H}_C^L \in \mathbb{R}^d$ are the averaged H_S^L and H_C^L at token level, respectively.³ To further identify the effect of inter- and

³Notably, the computation of \hat{p} and \bar{p} does not require a full decoding forward pass. It involves solely an FFN layer (two linear transformation layers and a ReLU activation layer), an LLM-Head layer (a linear transformation layer), and a softmax function layer.

intra-sentence context in this strategy, we provide an ablation study about \hat{p} and \bar{p} in Appendix J.

2.3 Phase-aware Prompts

We emphasize the LLM needs to play various roles across three phases, and maintaining similar prompts across different phases may not be reasonable. Thus, we empower LLM to distinguish different phases by introducing a type embedding and a transfer layer⁴ for these prompts:

$$\mathbf{P}_r = (\tanh(\mathbf{O}_r W_1)) W_2 + \text{TypeEmb}(r), \quad (22)$$

where $\mathbf{O}_r \in \mathbb{R}^{L \times 2q \times d}$ is randomly initialized prompt, $W_1, W_2 \in \mathbb{R}^{d \times d}$ are the trainable parameters, and $\text{TypeEmb}(\cdot)$ is type embeddings layer of the prompts. $r \in \{\mathcal{C}, S, T\}$ represents either phase 1, phase 2, or phase 3.

2.4 Training Objective

We employ the cross-entropy loss as the training objective of our model. Given a training instance (\mathcal{C}, S, T) , its training loss is defined as:

$$\mathcal{L}(\mathcal{C}, S, T) = -\frac{1}{|T|} \sum_{t=1}^{|T|} \log p_e(y_t|S, \mathcal{C}, y_{<t}). \quad (23)$$

Notably, the parameters in LLM, including W in Eq. 16, 18, 19, are frozen during training.

3 Experimentation

We build our approach upon two open-source LLMs, i.e., llama-2-7b⁵ and bloomz-7b1-mt⁶. We verify the effectiveness of our proposed approach on five translation tasks, including {Chinese (ZH), French (FR), German (DE), Spanish (ES), Russian (RU)} \rightarrow English (EN).

3.1 Experimental Settings

Datasets and Preprocessing. The corpus of all translation tasks is extracted from News-Commentary-v18. For LLM-based models, We use the tokenizer of foundation models to process the input data and no other preprocessing is performed. See Appendix A for more details on

⁴Unlike the multi-layer perceptrons (MLPs) used for reparameterization, our transfer layer shares parameters across all prompts, reducing the number of trainable parameters. Table 3 compares the trainable parameters of various tuning methods, and Appendix J analyzes the effect of the transfer layer.

⁵<https://huggingface.co/meta-llama/Llama-2-7b-hf>

⁶<https://huggingface.co/bigscience/bloomz-7b1-mt>

Model	ZH→EN		FR→EN		DE→EN		ES→EN		RU→EN		Average	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
⊙Trans.	29.86	0.8406	38.53	0.8545	41.44	0.8682	48.74	0.8783	32.25	0.8169	38.16	0.8517
<i>Traditional context-aware NMT models</i>												
⊙MR-Trans.	30.61	0.8413	38.72	0.8533	42.11	0.8693	49.69	0.8812	33.27	0.8211	38.88	0.8532
+ mBART	32.69	0.8601	42.01	0.8759	44.61	0.8840	51.67	0.8831	36.39	0.8459	41.39	0.8698
⊙G-Trans.	30.99	0.8411	38.96	0.8524	42.46	0.8658	49.68	0.8794	33.59	0.8201	39.14	0.8518
+ mBART	32.99	0.8597	42.02	0.8764	44.81	0.8836	52.07	0.8911	36.83	0.8461	41.74	0.8714
<i>llama-2-7b as foundation model</i>												
⊙MT-LoRA	27.43	0.8511	38.18	0.8647	40.96	0.8712	47.52	0.8733	33.00	0.8311	37.42	0.8583
⊙MT-PT	31.32	0.8565	41.92	0.8675	43.56	0.8752	51.32	0.8819	35.46	0.8333	40.72	0.8629
⊙CMT-PT	31.13	0.8387	42.01	0.8699	43.11	0.8762	51.66	0.8823	35.91	0.8396	40.76	0.8613
⊙MPT	*33.21	0.8645	†43.11	0.8744	*43.88	0.8824	†52.01	0.8913	†36.49	0.8456	41.74	0.8716
⊙DeMPT	*33.89	0.8658	†43.71	0.8816	*44.69	0.8899	†53.10	0.8979	†36.55	0.8438	42.39	0.8758
<i>blomz-7b1-mt as foundation model</i>												
⊙MT-LoRA	25.79	0.8466	35.67	0.8601	35.17	0.8522	46.32	0.8644	28.01	0.8012	34.21	0.8449
⊙MT-PT	30.99	0.8520	40.49	0.8661	37.76	0.8579	50.68	0.8823	30.27	0.8106	38.04	0.8539
⊙CMT-PT	30.82	0.8504	40.31	0.8639	38.01	0.8601	50.26	0.8832	29.80	0.8108	37.84	0.8537
⊙MPT	*31.81	0.8601	*41.11	0.8766	†38.99	0.8669	*51.33	0.8910	*30.99	0.8201	38.85	0.8629
⊙DeMPT	*32.46	0.8649	*41.92	0.8790	†40.06	0.8703	*52.25	0.8990	*31.79	0.8253	39.70	0.8677

Table 1: Results of different systems on sacreBLEU and COMET metrics. **DeMPT/MPT** is our proposed Multi-phase Prompt Tuning approach *with/without* Decoding-enhanced strategy (in Sec. 2.2). Scores with **bold** indicate the best performance. * (or †) indicates the gains are statistically significant over MT-PT (or CMT-PT) with $p < 0.01$ (Koehn, 2004). ⊙ and ⊙ indicate the model is *context-agnostic* and *context-aware*, respectively.

splitting, preprocessing and statistics of datasets. Besides, we provide a discussion for scales of the training set in Appendix I.

Baselines. In addition to *traditional context-agnostic* (Trans.) and *context-aware* (G-Trans (Bao et al., 2021) and MR-Trans (Sun et al., 2022) with or without pre-training setting, i.e., + mBART (Liu et al., 2020)) NMT models with *encoder-decoder* architecture,⁷ our primary comparison focuses on the following three LLM-based alternatives: 1) **MT-LoRA**: It is a tuned LLM adapted to NMT task via the tuning method of Low-Rank Adaptation (Hu et al., 2022), which makes large-scale pre-training models adapt to a new task by injecting a trainable rank decomposition matrices into each layer of the Transformer architecture; 2) **MT-PT**: It is a tuned LLM adapted to NMT task via the deep prompt tuning with MLPs reparameterization,⁸ which only tunes continuous prompts with a frozen language model; 3) **CMT-PT**: It indiscriminately utilizes inter- and intra-sentence context via the concatenation strategy, as depicted in Figure 1 (a). Similar to MT-PT, it is also a tuned LLM

⁷Please refer to Appendix C for more introduction about the G-Trans and MR-Trans.

⁸We attempt to remove reparameterization but experience a significant decline in performance.

via the deep prompt tuning with MLPs reparameterization. Among them, MT-LoRA and MT-PT are *context-agnostic* systems while CMT-PT is a *context-aware* system. For a fair comparison, we ensure that all context-aware models built upon LLM, including CMT-PT, MPT, and DeMPT, incorporate identical inter-sentence context. We provide more discussion in utilization of various inter-sentence contexts in Appendix L and M.

Model Setting and Training. For all *encoder-decoder* Transformer models, including Trans., MR-Trans and G-Trans⁹, we implement them upon Fairseq (Ott et al., 2019). For MT-LoRA models, we set the rank of trainable matrices as 16 which performs best in our preliminary experiment. For all MT-PT models, CMT-PT models, and our models, we set the prompt length q as 64.¹⁰ For the incorporation of inter-sentence context in CMT-PT and our models, we consider a dynamic z , in which the total tokens are no more than 256. In enhanced decoding, we consider the three next word predictions to be equally important

⁹For G-Trans, we use their official implementation upon Fairseq. Code: <https://github.com/baoguangsheng/g-transformer>.

¹⁰we provide more discussion in Appendix E about the prompt length.

Model	ZH→	FR→	DE→	ES→	RU→	Avg.
⊙Trans.	47.63	54.41	58.29	62.52	48.79	54.33
<i>Traditional context-aware NMT model</i>						
⊙MR-Trans.	48.51	55.55	59.02	63.51	49.88	55.29
+ mBART	50.66	58.01	61.99	66.01	54.11	58.15
⊙G-Trans.	48.99	55.31	59.23	63.99	50.09	55.52
+ mBART	50.98	57.88	61.97	66.21	54.33	58.27
<i>llama-2-7b as foundation model</i>						
⊙MT-LoRA	44.83	54.52	57.72	62.18	49.06	53.66
⊙MT-PT	49.49	57.87	60.89	65.02	52.59	57.17
⊙CMT-PT	49.53	58.27	61.23	65.89	53.34	57.65
⊙MPT	51.56	59.56	62.15	67.14	54.18	58.92
⊙DeMPT	52.68	60.33	63.11	67.95	54.94	59.80
<i>bloomz-7b1-mt as foundation model</i>						
⊙MT-LoRA	43.23	51.82	51.12	61.77	43.29	50.25
⊙MT-PT	49.48	56.81	55.40	64.71	46.14	54.51
⊙CMT-PT	49.61	57.05	55.81	65.12	46.09	54.74
⊙MPT	50.22	57.93	56.69	66.25	47.29	55.68
⊙DeMPT	50.62	58.30	57.34	67.12	48.00	56.28

Table 2: Results of different systems on BlonDe metric.

by setting both λ_1 and λ_2 to 1/3. We provide an analysis of λ and more training details in Appendix K and B, respectively.

Evaluation. We use sacreBLEU (accuracy-related metric)¹¹ (Post, 2018), COMET (semantics-related metric) with the wmt22-comet-da model¹² (Rei et al., 2020), and BlonDe (discourse-related metric) (Jiang et al., 2022) as evaluation metrics.¹³

3.2 Experimental Results

The main experimental results are presented in Tables 1 and 2. Additionally, a comparison of the number of trainable parameters is presented in Table 3 across different tuning methods. When examining llama-2-7b and focusing on context-agnostic models, we find that the Transformer models (Trans.) generally outperform LLMs with LoRA tuning (MT-LoRA) in most translation directions based on BLEU score. However, the MT-LoRA models surpass Trans. in COMET, indicating that translations from LLMs may better align with human preferences. Additionally, the MT-PT models exhibit superior performance compared to the MT-LoRA models across BLEU, COMET, and BlonDe metrics. This improvement could be attributed to the more trainable parameters in the

¹¹Signature: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1

¹²<https://github.com/Unbabel/COMET>

¹³We provide more discourse-related evaluation in Appendix F.

	MT-LoRA	MT-PT/CMT-PT	MPT/DeMPT
Trainable Para.	0.12%	13.87%	3.11%

Table 3: Proportion of trainable parameters against total parameters for different tuning methods.

MT-PT models (13.87% vs. 0.12%).

Importantly, by comparing MT-PT and CMT-PT, we observe that CMT-PT which indiscriminately leverages the inter- and intra-sentence context with the concatenation way, even hurts performance for certain translation tasks. For example, the CMT-PT models, despite excelling in discourse-related BlonDe scores (averaging 57.65 vs. 57.17), underperforms in BLEU and COMET compared to the MT-PT models. In contrast, our context-aware MPT and DeMPT models outperform all LLM baselines across all translation tasks in three metrics. For example, our MPT models achieve an average gain of 0.98/0.0103/1.27 in BLEU/COMET/BlonDe compared to the CMT-PT models. Our decoding-enhance strategy further enhances the capacity of LLMs, with DeMPT outperforming MPT with an average gain of 0.65/0.0042/0.88. Compared to G-Trans. (+mBART) or MR-Trans (+mBART), DeMPT also demonstrates either superior or comparable performance across all language pairs.

Finally, we observe a similar performance trend among MT models built upon bloomz-7b1-mt. It also indicates that models built upon llama-2-7b outperform those utilizing bloomz-7b1-mt, suggesting that llama-2-7b serves as a more robust foundation model for translation tasks.

4 Discussion

In this section, we use bloomz-7b1-mt as the foundation model to discuss our approach.¹⁴ See Appendix D~M for further discussions.

4.1 Effect of Length of Inter-sentence Context

For efficient training, we define the inter-sentence context in Section 2 as previous sentences with a total tokens not exceeding 256. We are curious about the potential impact of inter-sentence length on the performance of our approach. Consequently, we extend the inter-sentence context length from 256 to 1024 and assess the performance of our approach in the ZH→EN task. Figure 4 shows

¹⁴Considering page limitation and the consumption of GPUs resources and training time, we use the ZH→EN task as a representative to report the BLEU and BlonDe scores.

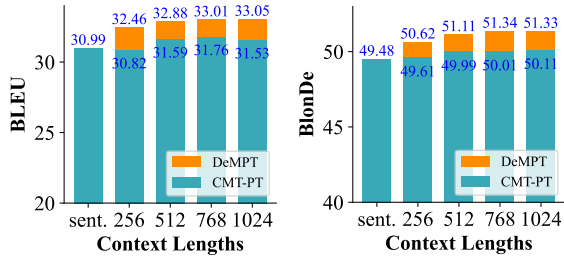


Figure 4: Performance of CMT-PT and our DeMPT on ZH→EN test set when using different inter-sentence context lengths.

the performance trend of the CMT-PT model and our DeMPT model. As the length of the inter-sentence context increases, both models exhibit a slight enhancement in both BLEU and BlonDe scores. Interestingly, our model with a 256-token inter-sentence context outperforms the CMT-PT model with a 1024-token inter-sentence context in both BLEU and BlonDe scores. This further suggests the effectiveness of our approach in harnessing the capabilities of LLMs for context-aware NMT compared to the concatenation strategy.

4.2 Effect of Multi-phase Strategy

Our multi-phase strategy divides the whole translation into three phases: phase 1 for encoding inter-sentence, phase 2 for encoding intra-sentence, and phase 3 for decoding current sentence. To assess the effect of multi-phase strategy, we compare its performance with two contrasting strategies: merging the first two phases into one (i.e., *Merging 1&2*) and merging all phases into a single one (i.e., *Merging 1&2&3*).¹⁵ Note that in both the two contrasting strategies, we replace the enhanced next word prediction $p_e(y_t|\cdot)$ (Eq. 17) in decoding phase with $p(y_t|\cdot)$. Table 4 presents the performance of different phrasing strategies. Comparing *Merging 1&2* and *Merging 1&2&3*, it shows that separating the decoding phrase from the encoding marginally improves the performance in both BLEU and BlonDe. Importantly, the comparison of MPT and *Merging 1&2* tells that separating the encoding of inter- and intra sentence achieves more gains across all metrics.

Meanwhile, we conjecture that another benefit of multi-phasing strategy lies in the robustness to the noise contained in document-level context. To test the conjecture, we replace the original inter-

¹⁵When merging them all into one, it equals CMT-PT, i.e., the concatenate strategy.

Model	BLEU	COMET	BlonDe
MPT	31.81	0.8601	50.22
DeMPT	32.46	0.8649	50.62
<i>Merging 1&2&3</i>	30.82	0.8504	49.61
<i>Merging 1&2</i>	31.01	0.8503	49.91

Table 4: Comparison of performances when using different phrasing strategies on ZH→EN test set.

Model	BLEU	COMET	BlonDe
<i>Merging 1&2&3</i>	30.82	0.8504	49.61
w/ rnd. CTX	28.63	0.8402	48.01
DeMPT	32.46	0.8649	50.62
w/ rnd. CTX	31.56	0.8581	49.71

Table 5: Comparison of performance when using gold or random inter-sentence context on ZH→EN test set.

sentence context with a random inter-sentence context, meaning we randomly select some sentences from other documents to serve as the inter-sentence context. As shown in Table 5, the performance of both the *Merging 1&2&3* and DeMPT models consistently deteriorates when exposed to random context (w/ rnd. CTX). However, the decline is more pronounced for *Merging 1&2&3* than for DeMPT (-2.19/0.0102/1.60 vs -0.90/0.0068/0.91). This suggests that DeMPT, owing to its multi-phase strategy, exhibits more robustness in utilizing inter-sentence context in contrast to *Merging 1&2&3*.

4.3 Human Evaluation

We use the Direct Assessment (DA) method (Graham et al., 2017) to manually assess the quality of translations generated by DeMPT and CMT-PT. In this assessment, human evaluators compare the meaning of the MT output with a human-produced reference translation, working within the same language. Specifically, we randomly select 5 documents with a total of 200 groups of sentences from the ZH→EN test set. To avoid potential bias in evaluation, we recruit 6 professional translators and ensure each translation from DeMPT or CMT-PT is scored twice by two translators. Table 6 shows the DA scores for CMT-PT and DeMPT. Our DeMPT outperforms CMT-PT by 7.14 DA score, providing strong evidence for the effectiveness of our approach. Further details and results regarding the DA can be found in Appendix H.

Model	Score_1	Score_2	Average
CMT-PT	79.00	80.17	79.59
DeMPT	86.17 (+7.17)	87.30 (+7.13)	86.73 (+7.14)

Table 6: Human DA scores for CMT-PT and DeMPT on ZH→EN translation task.

5 Related Work

Due to limited space, we omit the discussion on conventional context-aware MT, focusing instead on LLM-based context-aware MT and prompt tuning for LLMs. Besides, considering our DeMPT’s inspiration from MSP (Tan et al., 2019), we offer further discussion on their differences.

LLM-based Context-aware Machine Translation. While traditional context-aware neural machine translation (NMT) has seen considerable progress in recent years (Jean et al., 2017; Wang et al., 2017; Voita et al., 2018; Maruf et al., 2019; Kang et al., 2020; Bao et al., 2021; Sun et al., 2022; Bao et al., 2023), the effective integration of large language models (LLMs) to model inter-sentence context and enhance context-aware translation remains an area of limited exploration. Existing studies mainly focus on the assessment of LLMs’ ability in discourse modeling. For example, Wang et al. (2023) approach context-aware NMT as a task involving long sequence generation, employing a concatenation strategy, and conduct comprehensive evaluations of LLMs such as ChatGPT and GPT-4. Their focus includes the impact of context-aware prompts, comparisons with translation models, and an in-depth analysis of discourse modeling ability. Similarly, Karpinska and Iyyer (2023) engage professional translators to evaluate LLMs’ capacity in context-aware NMT. In contrast, Wu et al. (2024) compare the effectiveness of various parameter-efficient fine-tuning methods on moderately-sized LLMs for context-aware NMT. Besides, Wu and Hu (2023) explore the prompt engineering with GPT language models specifically for document-level (context-aware) MT while Li et al. (2024) experiment with combining sentence-level and document-level translation instructions of varying lengths to fine-tune LLMs. Differently, Koneru et al. (2023) propose a post-editing approach to enhance LLMs’ capacity in utilization of inter-sentence context in document-level MT.

Prompt Tuning for Large Language Model. Liu et al. (2021) and Li and Liang (2021) propose to

make LLMs adapt to various tasks by adding trainable prompts (also called continuous prompts) to the original input sequences. In this paradigm, only the continuous prompts are updated during training. Liu et al. (2022) further introduces deep prompt tuning, extending the idea by inserting trainable prompts into all layers of LLMs, rather than just the embedding layer. While these approaches provide a general framework, we focus on enhancing LLM performance specifically for inter-sentence context modeling in context-aware NMT.

Discussion with MSP. Tan et al. (2022) propose a multi-phase tuning approach (MSP) to enhance the sentence-level translation performance of a multilingual GPT. Our DeMPT mainly differs from MSP in the following aspects: 1) DeMPT adopts a phase-aware prompt to enable distinctive modeling for different inputs, namely inter-sentence contexts, intra-sentence contexts, and the target sentence, a feature not present in MSP; 2) DeMPT incorporates a decoding-enhanced strategy to further improve the effectiveness of utilizing different context information, a capability not available in MSP; 3) DeMPT is designed to alleviate discourse problems in context-aware LLM-based machine translation tasks, rather than addressing sentence-level machine translation tasks as in the case of MSP; 4) DeMPT is designed to adapt LLMs rather than smaller pre-trained model used in MSP.

6 Conclusion

In this paper, we have examined the hypothesis that it is crucial to differentially model and leverage inter-sentence context and intra-sentence context when adapting LLMs to context-aware NMT. This stems from our observation that intra-sentence context exhibits a stronger correlation with the target sentence compared to inter-sentence context, owing to its richer parallel semantic information. To this end, we have proposed a novel decoding-enhanced multi-phase prompt tuning (DeMPT) approach to make LLMs aware of the differences between inter- and intra-sentence contexts, and further improve LLMs’ capacity in discourse modeling. We have evaluated our approach using two foundation models and present experimental results across five translation directions. Experimental results and discussions have demonstrated a significant enhancement in the performance of LLMs in context-aware NMT, manifesting as improved translation accuracy and a reduction in discourse-related issues.

569 Limitations

570 Owing to resource limitations, our work is re-
571 stricted to moderate-scale LLMs, specifically those
572 with 7 billion parameters, and a confined window
573 size of inter-sentence context. It is imperative to
574 acknowledge that the results of our research may
575 differ when employing larger models and extended
576 window sizes for inter-sentence contexts. Consid-
577 ering that English text forms the main body of the
578 training data for LLMs, this paper only focuses on
579 the English-centric translation tasks. The results of
580 non-English-centric translation tasks may vary. We
581 acknowledge these limitations and consider them
582 as avenues for future exploration. Besides, follow-
583 ing the finding of Bao et al. (2021), we focus solely
584 on the source-side inter-sentence context in this
585 work. We will explore more about the integration
586 of target-side inter-sentence context in the future.

587 References

- 588 Guangsheng Bao, Zhiyang Teng, and Yue Zhang. 2023.
589 Target-side augmentation for document-level ma-
590 chine translation. In *Proceedings of ACL*, pages
591 10725–10742.
- 592 Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing
593 Chen, and Weihua Luo. 2021. G-transformer for
594 document-level machine translation. In *Proceedings
595 of ACL*, pages 3442–3455.
- 596 BigScience. 2022. Bloom: A 176b-parameter open-
597 access multilingual language model. *Computing Re-
598 search Repository*, arXiv:2211.05100.
- 599 Google. 2022. Palm: Scaling language modeling with
600 pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.
- 601 Yvette Graham, Qingsong Ma, Timothy Baldwin, Qun
602 Liu, Carla Parra, and Carolina Scarton. 2017. Im-
603 proving evaluation of document-level machine trans-
604 lation quality estimation. In *Proceedings of EACL*,
605 pages 356–361.
- 606 Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-
607 Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu
608 Chen. 2022. LoRA: Low-rank adaptation of large
609 language models. In *Proceedings of ICLR*.
- 610 Xinyu Hu and Xiaojun Wan. 2023. Exploring discourse
611 structure in document-level machine translation. In
612 *Proceedings of EMNLP*, pages 13889–13902.
- 613 Sebastien Jean, Stanislas Lauly, Orhan Firat, and
614 Kyunghyun Cho. 2017. Does neural machine trans-
615 lation benefit from larger context? *Computing Re-
616 search Repository*, arXiv:1704.05135.

- Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dong-
dong Zhang, Jian Yang, Haoyang Huang, Rico Sen-
rich, Ryan Cotterell, Mrinmaya Sachan, and Ming
Zhou. 2022. BlonDe: An automatic evaluation
metric for document-level machine translation. In
Proceedings of NAACL, pages 1550–1565, Seattle,
United States. 617
618
619
620
621
622
623
- Xiaomian Kang, Yang Zhao, Jiajun Zhang, and
Chengqing Zong. 2020. Dynamic context selection
for document-level neural machine translation via
reinforcement learning. In *Proceedings of EMNLP*,
pages 2242–2254. 624
625
626
627
628
- Marzena Karpinska and Mohit Iyyer. 2023. Large lan-
guage models effectively leverage document-level
context for literary translation, but critical errors per-
sist. In *Proceedings of WMT*, pages 419–451. 629
630
631
632
- Philipp Koehn. 2004. Statistical significance tests for
machine translation evaluation. In *Proceedings of
EMNLP*, pages 388–395. 633
634
635
- Sai Koneru, Miriam Exel, Matthias Huck, and Jan
Niehues. 2023. Contextual refinement of translations:
Large language models for sentence and document-
level post-editing. arXiv:1701.06549. 636
637
638
639
- Shaohui Kuang, Junhui Li, António Branco, Weihua
Luo, and Deyi Xiong. 2018. Attention focusing for
neural machine translation by bridging source and
target embeddings. In *Proceedings of ACL*, pages
1767–1776. 640
641
642
643
644
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021.
The power of scale for parameter-efficient prompt
tuning. In *Proceedings of EMNLP*, pages 3045–
3059. 645
646
647
648
- Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning:
Optimizing continuous prompts for generation. In
Proceedings of ACL-IJCNLP, pages 4582–4597, On-
line. 649
650
651
652
- Yachao Li, Junhui Li, Jing Jiang, Shimin Tao, Hao Yang,
and Min Zhang. 2023. P-Transformer: Towards Bet-
ter Document-to-Document Neural Machine Transla-
tion. *IEEE/ACM Transactions on Audio, Speech, and
Language Processing*, 31:3859–3870. 653
654
655
656
657
- Yachao Li, Junhui Li, Jing Jiang, and Min Zhang.
2024. Enhancing document-level translation of large
language model via translation mixed-instructions.
Computing Research Repository, arXiv:2401.08088. 658
659
660
661
- Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengx-
iao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning:
Prompt tuning can be comparable to fine-tuning
across scales and tasks. In *Proceedings of ACL*, pages
61–68. 662
663
664
665
666
- Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding,
Yujie Qian, Zhilin Yang, and Jie Tang. 2021. Gpt
understands, too. *Computing Research Repository*,
arXiv:2103.10385. 667
668
669
670

671	Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey	Xin Tan, Longyin Zhang, Deyi Xiong, and Guodong	725
672	Edunov, Marjan Ghazvininejad, Mike Lewis, and	Zhou. 2019. Hierarchical modeling of global context	726
673	Luke Zettlemoyer. 2020. Multilingual denoising pre-	for document-level neural machine translation. In	727
674	training for neural machine translation. <i>Transac-</i>	<i>Proceedings of EMNLP-IJCNLP</i> , pages 1576–1585.	728
675	<i>tions of the Association for Computational Linguis-</i>		
676	<i>tics</i> , 8:726–742.		
677	Xinglin Lyu, Junhui Li, Zhengxian Gong, and Min	Zhixing Tan, Xiangwen Zhang, Shuo Wang, and Yang	729
678	Zhang. 2021. Encouraging lexical translation consis-	Liu. 2022. MSP: Multi-stage prompting for making	730
679	tency for document-level neural machine translation.	pre-trained language models better translators. In	731
680	In <i>Proceedings of EMNLP</i> , pages 3265–3277.	<i>Proceedings of ACL</i> , pages 6131–6142.	732
681	Xinglin Lyu, Junhui Li, Shimin Tao, Hao Yang, and	Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob	733
682	Min Zhang. 2022. Modeling consistency preference	Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz	734
683	via lexical chains for document-level neural machine	Kaiser, and Illia Polosukhin. 2017. Attention is all	735
684	translation. In <i>Proceedings of EMNLP</i> , pages 6312–	you need. In <i>Proceedings of NIPS</i> , pages 5998–6008.	736
685	6326.		
686	Sameen Maruf, André F. T. Martins, and Gholam-	Elena Voita, Rico Sennrich, and Ivan Titov. 2019a.	737
687	reza Haffari. 2019. Selective attention for context-	Context-aware monolingual repair for neural ma-	738
688	aware neural machine translation. In <i>Proceedings of</i>	chine translation. In <i>Proceedings of EMNLP-</i>	739
689	<i>NAACL</i> , pages 3092–3102.	<i>IJCNLP</i> , pages 877–886.	740
690	MetaAI. 2023a. Llama 2: Open foundation and fine-	Elena Voita, Rico Sennrich, and Ivan Titov. 2019b.	741
691	tuned chat models. <i>Computing Research Repository</i> ,	When a good translation is wrong in context: Context-	742
692	arXiv:2307.09288.	aware machine translation improves on deixis, ellip-	743
693	MetaAI. 2023b. Llama: Open and efficient foundation	sis, and lexical cohesion. In <i>Proceedings of ACL</i> ,	744
694	language models. <i>ArXiv</i> , abs/2302.13971.	pages 1198–1212.	745
695	Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas,	Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan	746
696	and James Henderson. 2018. Document-level neural	Titov. 2018. Context-aware neural machine transla-	747
697	machine translation with hierarchical attention net-	tion learns anaphora resolution. In <i>Proceedings of</i>	748
698	works. In <i>Proceedings of EMNLP</i> , pages 2947–2954.	<i>ACL</i> , pages 1264–1274.	749
699	OpenAI. 2023. Gpt-4 technical report. <i>Computing</i>	Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang,	750
700	<i>Research Repository</i> , arXiv:2303.08774.	Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023.	751
701	Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan,	Document-level machine translation with large lan-	752
702	Sam Gross, Nathan Ng, David Grangier, and Michael	guage models. In <i>Proceedings of EMNLP</i> , pages	753
703	Auli. 2019. fairseq: A fast, extensible toolkit for	16646–16661.	754
704	sequence modeling. In <i>Proceedings of NAACL-HLT:</i>	Longyue Wang, Zhaopeng Tu, Xing Wang, and Shum-	755
705	<i>Demonstrations</i> .	ing Shi. 2019. One model to learn both: Zero pro-	756
706	Matt Post. 2018. A call for clarity in reporting BLEU	noun prediction and translation. In <i>Proceedings of</i>	757
707	scores. In <i>Proceedings of WMT</i> , pages 186–191.	<i>EMNLP-IJCNLP</i> , pages 921–930.	758
708	Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase,	Longyue Wang, Zhaopeng Tu, Andy Way, and Qun	759
709	and Yuxiong He. 2020. Zero: Memory optimiza-	Liu. 2017. Exploiting cross-sentence context for neu-	760
710	tions toward training trillion parameter models. In	ral machine translation. In <i>Proceedings of EMNLP</i> ,	761
711	<i>SC20: Proceedings of High Performance Computing,</i>	pages 2826–2831.	762
712	<i>Networking, Storage and Analysis</i> , pages 1–16.	Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George	763
713	Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon	Foster, and Gholamreza Haffari. 2024. Adapt-	764
714	Lavie. 2020. COMET: A neural framework for MT	ing large language models for document-level ma-	765
715	evaluation. In <i>Proceedings of EMNLP</i> , pages 2685–	chine translation. <i>Computing Research Repository</i> ,	766
716	2702.	arXiv:2401.06468.	767
717	Rico Sennrich, Barry Haddow, and Alexandra Birch.	Yangjian Wu and Gang Hu. 2023. Exploring prompt en-	768
718	2016. Neural machine translation of rare words with	gineering with GPT language models for document-	769
719	subword units. In <i>Proceedings of ACL</i> , pages 1715–	level machine translation: Insights and findings. In	770
720	1725.	<i>Proceedings of the Eighth Conference on Machine</i>	771
721	Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao,	<i>Translation</i> , pages 166–169.	772
722	Shujian Huang, Jiajun Chen, and Lei Li. 2022. Re-	Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei	773
723	thinking document-level neural machine translation.	Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018.	774
724	In <i>Findings of ACL</i> , pages 3537–3548.	Improving the transformer translation model with	775
		document-level context. In <i>Proceedings of EMNLP</i> ,	776
		pages 533–542.	777

A Datasets

Splitting, Preprocessing and Statistics of Datasets. For all translation tasks, we randomly select 80% document pairs from the corpus as the training set. Both the test set and validation set include 150 document pairs each, randomly sampled from the remaining 20% of document pairs in the corpus. Regarding sentence preprocessing across all datasets for LLM-based models, we segment the sentences with the tokenizer from the respective foundation model. No additional preprocessing steps are performed. For *encoder-decoder* Transformer models, we segment the source and target sentences into sub-words by a BPE model with 30K merged operations (Sennrich et al., 2016). We provide the detailed statistic in Table 7. Datasets are downloaded from <https://data.statmt.org/news-commentary/v18>.

B Training Details

For all *encoder-decoder* Transformer NMT models, we use the transformer-base setting as in Vaswani et al. (2017), where the learning rate is set to $1e-4$ with an inverse-square schedule and warmup steps of 4000, and use Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. For the other special training settings in G-Trans and MR-Trans, we keep consistent with that provided in their paper. All Transformer NMT models are trained on $4 \times$ NVIDIA V100 32GB GPUs with a batch size of 4096. For the models with prompt tuning in Section 3, including MT-PT, CMT-PT, MPT and DeMPT models, the length of the trainable prompt is set as 64. During both training and inference, the model generates only the current target sentence, operating in a many-to-one translation mode. For all fine-tuning models in this paper, we set the training epoch to 4, and the warm-up rate to 0.1. We use the log learning rate decay strategy with a maximum learning rate of $5e-5$. We collate a mini-batch by counting the total tokens inside the batch and set the batch size as 4096. All fine-tuning models are trained on $4 \times$ NVIDIA A800 GPUs with Deespeed Zero 2 offload setting (Rajbhandari et al., 2020).¹⁶

C Traditional Context-aware Models

In this paper, we implement G-Transformer (G-Trans) (Bao et al., 2021) and Multi-Resolution Transformer (MR-Trans) (Sun et al., 2022) as rep-

resentatives of traditional context-aware models for comparison. For ease of understanding, we provide a brief introduction to these two models in this section.

G-Transformer. The transformer model equips a group attention on the lower encoder/decoder layer and a combined attention on the top encoder/decoder layer. For a sentence being translated with its inter-sentence context, the group attention helps maintain locality bias by focusing on intra-sentence context. Meanwhile, the combined attention effectively integrates boundary information, enhancing the translation process with inter-sentence context.

Multi-Resolution Transformer. The Transformer model does not include any additional modules specifically for modeling inter-sentence context. Instead, it only uses a mixed training set that comprises both sentence-level and document-level instances with varying numbers of sentences. Training on this mixed set allows the Transformer model to handle both sentence-level and document-level translation tasks. In this paper, we implement its Document-to-Sentence variant, which uses all preceding contexts as the source and the current sentence as the target.

D Comparison of Inference Speed

Table 8 compares the inference speed of different models on ZH→EN translation task. Our MPT and DeMPT models, dividing the context-aware NMT process into three separate phases, demonstrates comparable inference speed to the single-phase MT-PT and CMT-PT models, with only a marginal drop of 0.02 seconds per sentence in decoding. This illustrates the efficiency of our approach without introducing significant computational overhead.

E Effect of Prompt Length

As our approach is implemented based on deep prompt tuning, next we compare the impact of the trainable prompt length for MT-PT, CMT-PT, and our DeMPT. Figure 7 shows the performance curves when increasing the prompt length from 32 to 128. We observe that increased prompt length tends to enhance performance for both BLEU and BlonDe, yet the gains exhibit diminishing returns. This finding is consistent with that in Li and Liang (2021); Lester et al. (2021); Tan et al. (2022). We also observe that DeMPT with a prompt length of

¹⁶<https://github.com/microsoft/DeepSpeed>

Dataset	ZH→EN		FR→EN		DE→EN		ES→EN		RU→EN	
	#Doc	#Sent	#Doc	#Sent	#Doc	#Sent	#Doc	#Sent	#Doc	#Sent
Training	8,622	342,495	7,915	310,489	8,417	333,201	9,677	378,281	7,255	272,100
Validation	150	6,061	150	5,890	150	5,866	150	5,782	150	5,691
Test	150	5,747	150	5,795	150	5,967	150	5,819	150	5,619

Table 7: Statistics of training, validation, and test sets for five translation tasks. #Doc and #Sent denote the numbers of *Document* and *Sentence*, respectively.

Score	Criterion
0-20	The translation is completely incorrect and unclear , with only a few words or phrases being correct. It is totally unreadable and difficult to understand .
21-40	The translation has very little semantic similarity to the source sentence, with key information missing or incorrect. It has numerous unnatural and unfluent expressions and grammatical errors .
41-60	The translation can express part of the key semantics but has many non-key semantic errors. It lacks fluency and idiomaticity .
61-80	The translation can express the key semantics but has some non-key information errors and significant grammatical errors. It lacks idiomaticity .
81-100	The translation can express the semantics of the source sentence with only a few non-key information errors and minor grammatical errors. It is fluent and idiomatic .

Figure 5: Scoring criterion for Direct Assessment. We group the score into five ranges, i.e., 0-20, 21-40, 41-60, 61-80, 81-100.

Model	Speed	BLEU
MT-PT	0.75 <i>sec/sent.</i>	30.99
CMT-PT	0.77 <i>sec/sent.</i>	30.82
MPT	0.78 <i>sec/sent.</i>	31.81
DeMPT	0.79 <i>sec/sent.</i>	32.46

Table 8: Comparison of inference speed on ZH→EN translation task. Speed is measured on the test set using 4 GPUs. *sec/sent.* means seconds spent for decoding each sentence. Note that the reparameterization is not needed during inference (Li and Liang, 2021).

64 outperforms both MT-PT and CMT-PT with a prompt length of 128 on both metrics, suggesting the superiority of our approach over the concatenation strategy in enhancing LLMs’ capacity for context-aware NMT.

F Performance on Contrastive Test Set

We evaluate the models’ ability to resolve discourse inconsistencies using the contrastive test set proposed by (Voita et al., 2019a), which focuses on four discourse phenomena such as deixis, lexicon

consistency (lex.c), ellipsis inflection (ell.infl), and verb phrase ellipsis (ell.VP) in English→Russian translation. Within the test set, each instance comprises a positive translation and several negative ones that vary by only one specific word. The purpose of the contrastive test set is to assess whether a model is more inclined to generate a correct translation as opposed to incorrect variations. Table 9 lists the accuracy of translation prediction on the contrastive test set for MT-PT, CMT-PT and DeMPT. Compared to the context-agnostic MT-PT model, both context-aware CMT-PT and DeMPT models show substantial improvements across the four discourse phenomena. Additionally, DeMPT demonstrates the best performance, surpassing CMT-PT by an average accuracy margin of 3.8.

G Effect of Various Contexts for Decoding-enhanced Strategy

We conduct an ablation study on the ZH-EN translation direction using the bloomz-7b-mt model as the foundation model to clarify the effect of the three probabilities p in Equation 17, i.e., the effect of various contexts for the heuristic decoding-

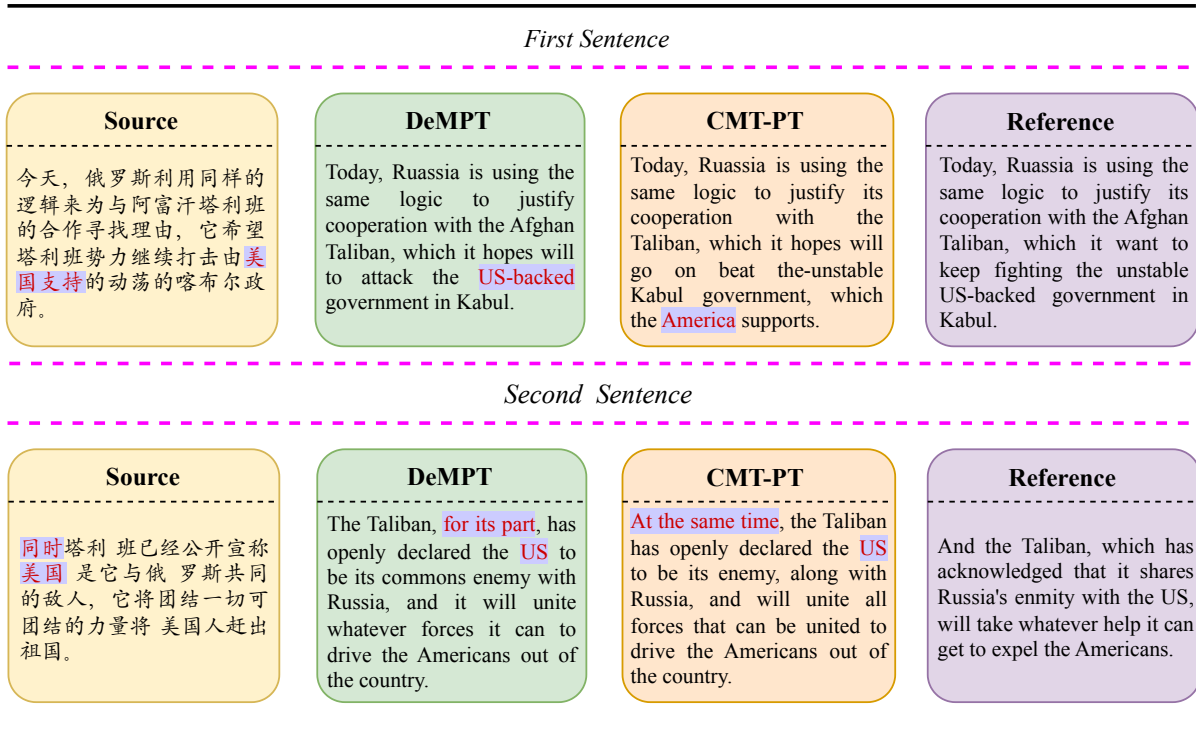


Figure 6: A case study for the CMT-PT model and our DeMPT model on ZH→EN translation task.

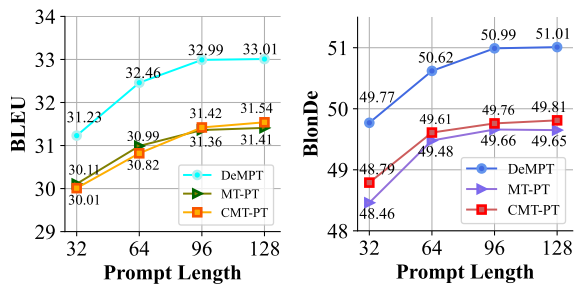


Figure 7: Performance of MT-PT, CMT-PT, and our DeMPT on ZH→EN test set when using different lengths of the trainable prompts.

enhanced strategy. From the Table 10, we observe that removing \hat{p} , i.e., *w/o* \hat{p} , leads to a significant degradation in the discourse-related metric, namely the BlonDe. This is because the integration enhances the utilization of the inter-sentence context during the decoding phase. We are additionally, removing results in the most substantial degeneration in BLEU metric. This observation demonstrates that our heuristic decoding-enhanced strategy can distinctively improve the utilization of various contexts during the decoding phase.

H Details of Human Evaluation

Criterion and Recruitment. Given a source sentence, its translation from MT (i.e., CMT-PT and

Model	deixis	lex.c	ell.infl	ell.VP	Avg.
MT-PT	50.0	45.7	53.0	28.6	44.3
CMT-PT	80.2	46.1	74.3	75.3	68.9
DeMPT	80.1	55.7	75.9	79.3	72.7

Table 9: Accuracy [%] of translation prediction for four discourse phenomena on the English → Russian contrastive test set.

our DeMPT), and its human-produced reference translation, the evaluators are asked to give a score ranging from 0 to 100. Figure 5 presents the detailed criterion of scoring. We recruit evaluators from professional translators with at least five years of experience in translation.

Statistics of Translation Errors. We manually count the number of bad cases from our DeMPT model. The bad cases fall into two categories: (1) the DA score is 60 or lower; (2) the DA score is lower than that of the translation from CMT-PT. The main types of the bad cases are **Mistranslation** (Mis.), **Unnoticed Omission** (UO), **Inappropriate Expression** (IE), and **Grammatical Error** (GE). We present detailed statistics in Table 11. The statistics indicate the bad cases mainly come from Mistranslation and Unnoticed Omission. Meanwhile, our DeMPT model outperforms the CMT-PT

Model	BLEU	COMET	BlonDe
MT-PT	30.99	0.8520	49.48
CMT-PT	30.82	0.8504	49.61
DeMPT	32.46	0.8649	50.62
<i>w/o \hat{p}</i>	32.33	0.8629	50.29
<i>w/o \bar{p}</i>	32.11	0.8641	50.51

Table 10: Comparison of performances of the DeMPT when removing different probabilities p in decoding-enhanced strategy.

Group	Type of Bad Case				
	Mis.	UO	IE	GE	Total (Perc.)
1	6	3	1	2	12 (6.0%)
2	9	7	6	5	27 (13.5%)

Table 11: Statistics of bad cases from our DeMPT model on ZH→EN translation task. *Perc.* denotes the percentage of bad cases against the total of DA cases.

model in 86.5% DA cases.

Case Study. We present a case in Figure 6 to illustrate how our DeMPT model outperforms the CMT-PT model. In this case, we compare the translations of two consecutive sentences from our model and the CMT-PT model. First, we notice that the CMT-PT model translates the source word 美国 in the two sentences into *US* and *America*, respectively. However, our model **consistently** translates them into *US*. Second, our model uses *for its part*, a phrase with more **coherent preference**, as the translation of 同时, instead of *At the same time* adopted in the translation from the CMT-PT model. Both of them demonstrate the superiority of our proposed approach in discourse modeling.

I Effect of Dataset Scales

We conduct an experiment to analyze the impact of training dataset scales on the concatenating strategy (CMT-PT) and the multi-phased, decoding-enhanced strategy (DeMPT). To do this, we expand the ZH→EN training set with additional document-level data from the LDC.¹⁷ Specifically, we randomly selected 200K and 400K sentence pairs with their inter-sentence context from the LDC and combined them with the existing ZH→EN training set

¹⁷The training data set consists of LDC2002T01, LDC2004T07, LDC2005T06, LDC2005T10, LDC2009T02, LDC2009T15, and LDC2010T03.

Model	BLEU	COMET	BlonDe
CMT-PT	30.82	0.8504	49.61
+ 200K	31.21	0.8521	49.88
+ 400K	31.73	0.8555	50.11
DeMPT	32.46	0.8649	50.62
+ 200K	32.77	0.8663	50.99
+ 400K	33.56	0.8701	51.47

Table 12: Comparison of performances of CMT-PT and DeMPT trained on the different scales of corpus for the ZH→EN translation task.

to train the CMT-PT and DeMPT models.

Table 12 lists the performances of CMT-PT and DeMPT when extending scales of the training set into 500K (300K +200K) and 700K (300k + 400K). We observe increasing the scale of the training set consistently boosts the performance of DeMPT and CMT-PT. However, our DeMPT significantly outperforms CMT-PT across all three metrics.

J Effect of Transfer Layer and Type Embedding

As in Eq. 22 within Section 2.3, we introduce two sublayers: a non-linear transfer sublayer and a type embedding sublayer for the trainable prompt in each phase. This design enhances the awareness of LLMs regarding the distinctions in inputs across the three tuning phases, allowing them to adapt to specific roles at each phase. We investigate the effect of these two sublayers.

As shown in Table 15, our observations reveal that the transfer sublayer holds greater importance than the type embedding sublayer. Removing either the non-linear transfer sublayer (*w/o Transfer.*) or the type embedding sublayer (*w/o Embed.*) results in a performance drop of 0.84/0.0048/0.39 or 0.45/0.0036/0.007 in BLEU/COMET/BlonDe metrics.

K Effect of Hyperparameter λ

Due to the limited computational resources, we do not perform extensive experiments to find the optimal combination of λ_1 and λ_2 for different translation tasks, simply setting them to be equal. For example, verifying each combination of λ_1 and λ_2 requires 10 experiments (5×2 for the number of translation directions and foundation models).

Model	BLEU	COMET	BlonDe
MT-PT	30.99	0.8520	49.48
CMT-PT	30.82	0.8504	49.61
DeMPT	32.46	0.8649	50.62
<i>w/o</i> Transfer.	31.62	0.8601	50.23
<i>w/o</i> Embed.	32.01	0.8613	50.55
<i>w/o</i> CTX.	31.98	0.8593	49.89

Table 13: Comparison of performances of the DeMPT variants on ZH→EN test set. *w/o* Trans. or *w/o* Embed. denotes the variant without the non-linear transfer sublayer or type embedding sublayer in Eq. 22. *w/o* CTX. means the inter-sentence context is not available, i.e., context-agnostic DeMPT system.

Model (DeMPT)	BLEU	COMET	BlonDe
$\lambda_1=1/3, \lambda_2=1/3, \text{ZH}\rightarrow\text{EN}$	32.46	0.8649	50.62
$\lambda_1=1/4, \lambda_2=1/3, \text{ZH}\rightarrow\text{EN}$	32.51	0.8653	50.31
$\lambda_1=1/3, \lambda_2=1/3, \text{FR}\rightarrow\text{EN}$	41.92	0.8790	58.30
$\lambda_1=1/4, \lambda_2=1/3, \text{FR}\rightarrow\text{EN}$	41.82	0.8785	57.92

Table 14: Comparison of performances of the DeMPT with different combinations of λ_1 and λ_2 on ZH→EN and FR→EN test sets.

Therefore, we carry out targeted experiments using a combination of λ_1 and λ_2 on ZH→EN and FR→EN only here.

The results are reported in Table 14. We use a smaller value for λ_1 here and observe that the BlonDe scores are more sensitive to changes λ_1 compared to BLEU and COMET. For example, a smaller λ_1 results in -0.31 and -0.38 for ZH→EN and FR→EN, respectively. This sensitivity may be reasonable because λ_1 is used for adjusting the utilization of inter-sentence context.

L Effect of Inter-sentence Context

We implement the context-agnostic (sentence-level) DeMPT system to analyze the effect of the inter-sentence context and differences with MSP. More specifically, we replace the input of LLMs in the inter-sentence context encoding phase with the intra-sentence context. In other words, we encode the intra-sentence context twice to keep the multi-phase tuning strategy in DeMPT while making the inter-sentence context unavailable.

As shown in the last row of Table 15 (i.e., *w/o* CTX), we find that the inter-sentence context is crucial for the alleviation of discourse-related is-

Model	d-BLEU	d-COMET	d-BlonDe
MT-PT (<i>m2o</i>)	34.19	0.8216	49.48
CMT-PT (<i>m2o</i>)	34.06	0.8211	54.68
DeMPT (<i>m2o</i>)	35.76	0.8316	55.97
CMT-PT (<i>m2m</i>)	34.13	0.8256	55.34

Table 15: Comparison of performances of the models with different translation modes, i.e., with/without target-side inter-sentence context, on ZH→EN test set.

sues. The BlonDe score drops by 0.73 when the inter-sentence context is unavailable. Meanwhile, our DeMPT also significantly improves the performance of LLMs in context-agnostic MT, e.g., + 0.99 BLEU score and + 0.0073 COMET score compared to the MT-PT model.

M Effect of Target-side Inter-sentence Context

To enable a fair comparison, we incorporate only the source-side inter-sentence context for the model with the concatenating strategy, i.e., the CMT-PT model in the many-to-one (*m2o*) translation mode, as shown in Tables 1 and 2. To further investigate the effect of target-side inter-sentence context for the concatenating strategy, we compare the CMT-PT model in the many-to-many (*m2m*) translation mode to the models in the many-to-one translation mode, for the ZH→EN translation task when using the bloomz-7b1-mt as the foundation model.

Different from the results in Tables 1 and 2, we report the document-level BLEU, BlonDe, and COMET scores for all models here due to the unavailability of sentence-level alignment for many-to-many model. From the experimental results, we observe that the CMP-PT (*m2m*) model outperforms the CMP-PT (*m2o*) model (mostly significant in terms of the d-BlonDe metric), which demonstrates the effectiveness of the target context in addressing discourse issues. However, the CMP-PT (*m2m*) model still underperforms the DeMPT model across three metrics.