Self-Retrieval from Distant Contexts for Document-Level Machine Translation

Ziqian Peng[1,2] and **Rachel Bawden**[2] and **François Yvon**[1]
[1]Sorbonne Université & CNRS, ISIR, Paris, France
[2]Inria, Paris, France
{ziqian.peng,francois.yvon}@isir.upmc.fr rachel.bawden@inria.fr

Abstract

Document-level machine translation is a challenging task, as it requires modeling both short-range and long-range dependencies to maintain the coherence and cohesion of the generated translation. However, these dependencies are sparse, and most context-augmented translation systems resort to two equally unsatisfactory options: either to include maximally long contexts, hoping that the useful dependencies are not lost in the noise; or to use limited local contexts, at the risk of missing relevant information. In this work, we study a self-retrieval-augmented machine translation framework (SELF-RAMT), aimed at informing translation decisions with informative local and global contexts dynamically extracted from the source and target texts. We examine the effectiveness of this method using three large language models, considering three criteria for context selection. We carry out experiments on TED talks as well as parallel scientific articles, considering three translation directions. Our results show that integrating distant contexts with SELF-RAMT improves translation quality as measured by reference-based scores and consistency metrics.

1 Introduction

Document-level machine translation (DLMT) is a challenging task, as it requires modeling both short-range and long-range dependencies to maintain the coherence and cohesion of the generated translation. Inter-sentential contexts are indispensable for the handling of phenomena such as co-reference, lexical consistency, textual coherence and cohesiveness, which continue to be challenging for long document translation (Bawden et al., 2018; Maruf et al., 2019; Voita et al., 2019b; Fernandes et al., 2023). Numerous approaches, reviewed in (Maruf et al., 2021; Castilho and Knowles, 2024), have been proposed to integrate these contexts. They include segment concatenation (Tiedemann and Scherrer, 2017; Bawden et al., 2018; Sun et al., 2022), architecture adaptation (Miculicich et al., 2018; Yang et al., 2019; Ma et al., 2020), training strategy optimization (Lupo et al., 2022b; Li et al., 2023; Wu et al., 2024), and multi-pass refinement (Voita et al., 2019a; Yu et al., 2020; Koneru et al., 2024). Past work also shows that various sources of contextual information contribute differently to translation quality; the local source and target context is the main resource for handling anaphoric references and word-sense disambiguation information (Bawden et al., 2018; Gete et al., 2022), whereas the global context, especially on the target side, holds information likely to improve coherence and cohesiveness of the full translated document (Pal et al., 2024).

Recent generative models such as Llama3 (Grattafiori et al., 2024) and GPT4 (OpenAI et al., 2023) can process inputs up to hundreds of thousands of tokens, creating new possibilities for the inclusion of the whole source text, as well as already translated target segments, in the translation context. It however remains an open question whether such architectures, relying on the self-attention-mechanism (Vaswani et al., 2017), are effectively able to identify relevant long-range dependencies and actually improve DLMT (Wang et al., 2024). This is because inter-sentential dependencies can be sparsely distributed within a document, whereas self-attention generates dense patterns spreading out over the entire past text (Tay et al., 2023; Liu et al., 2024). Therefore, most approaches to DLMT still consider a limited context window size, usually up to 1024 tokens or a fixed number of sentences.

In order to capture long-distance dependencies without requiring the attention mechanism to handle overly long contexts, we propose self-retrieval augmentation for machine translation (SELF-RAMT), aiming to take into account both local and global dependencies, regardless of the

Proceedings of the Tenth Workshop on Machine Translation (WMT), pages 220–240
November 8–9, 2025.

document length. In our approach, inter-sentential dependencies are precomputed for the full source document to identify past relevant segments for each translation unit. As soon as they are translated, these segments and their translation become available to inform the subsequent translation choices. In our implementation, which uses large language models (LLMs), such dynamic contexts are taken into account through in-context learning (ICL). Two scenarios are considered: (a) one where in-context examples correspond to correct translations, as in online learning, where the input sentences, once incrementally post-edited by a human translator, become available to revise the model (Álvaro Peris and Casacuberta, 2019),[1] and (b) a fully automatic setup, with imperfect in-context translations, requiring no human intervention. In this context, our main research questions are as follows: (i) how to best identify and retrieve useful context segments, (ii) what improvement to MT quality do these retrieved contexts bring, and (iii) to what extent distant (as opposed to local) contexts actually enhance translation scores. We compare three criteria for context selection (cosine similarity with respect to LaBSE embeddings (COS), Best Match 25 (BM25) and point-wise mutual information (PMI)) and carry out experiments on three LLMs in three language directions (English to German (EN–DE), French (EN–FR), and Chinese (EN–ZH)), analyzing the impact of contexts, especially distant ones. Experimental data includes both TED talks (Cettolo et al., 2012) and a new dedicated parallel test set, MERSENNE, consisting of scientific articles for the EN–FR direction. Scientific articles offer an interesting use case to study term consistency in long document translation. Our investigation reveals that distant contexts retrieved with PMI provide valuable information that increases translation metric scores as well as term consistency. We make our code and data available.[2]

2 Related Work

Document-level MT DLMT research broadly falls into two categories: *Doc2Sent*, which involves translating each sentence individually using intra-document source and/or target context to aid translation, and *Doc2Doc*, which involves translating multiple sentences at once (Popescu-Belis, 2019; Maruf et al., 2021; Castilho and Knowles, 2024). *Doc2Doc* approaches represent a simple strategy for effective context integration, maintaining better consistency and coherence than *Doc2Sent* methods (Li et al., 2020; Sun et al., 2022). However, *Doc2Doc* methods struggle to process very long sequences of sentences, as relevant information is sparse in global contexts (Lupo et al., 2022a; Wang et al., 2023), which can lead to the degradation of translation quality or omitted sentences (Zhuocheng et al., 2023; Li et al., 2023; Peng et al., 2025). To address this problem, several approaches have been explored, including context-aware attention (Maruf et al., 2019; Zheng et al., 2021; Yang et al., 2023), which selects important contexts according to the attention distribution, and dynamic context selection (Kang et al., 2020), which applies a reward model to identify varying numbers of useful context sentences within the local context, constrained by the complexity of reinforced training.

LLM-based DLMT Various LLM-based methods have been proposed for DLMT. Several works explore zero-shot prompting and ICL for *Doc2Doc* MT (Hendy et al., 2023; Karpinska and Iyyer, 2023) and study the best way to train LLMs for DLMT (Xu et al., 2024; Li et al., 2024; Guo et al., 2024; Alves et al., 2024), illustrating the importance of high-quality in-context demonstrations and fine-tuning parallel corpora. LLMs have also been used as post-editors, either through fine-tuning (Koneru et al., 2024; Li et al., 2025; Dong et al., 2025) or prompting for iterative translation refinement (Briakou et al., 2024; Wang et al., 2025a).

When it comes to integrating context, LLMs offer greater flexibility than traditional neural machine translation models. Various multi-aspect prompting techniques have been proposed to enhance the input by incorporating or automatically summarizing relevant information from the context. For instance, DELTA (Wang et al., 2025b) builds a dynamic context for each source sentence, heavily relying on LLM components to extract and assemble relevant information from the available source and target texts (including proper nouns, bilingual summaries of local contexts and relevant past sentences within a predefined context window). However, DELTA is computationally costly (a lot more so than SELF-RAMT), due to the multiple steps required for dynamic context extraction. SENT2SENT++ (Guo et al., 2025) incorporates two types of contexts: a static part, consisting of an au-

[1] We refer to this scenario as *online in-context learning*.

[2] https://anr-matos.github.io/resources.

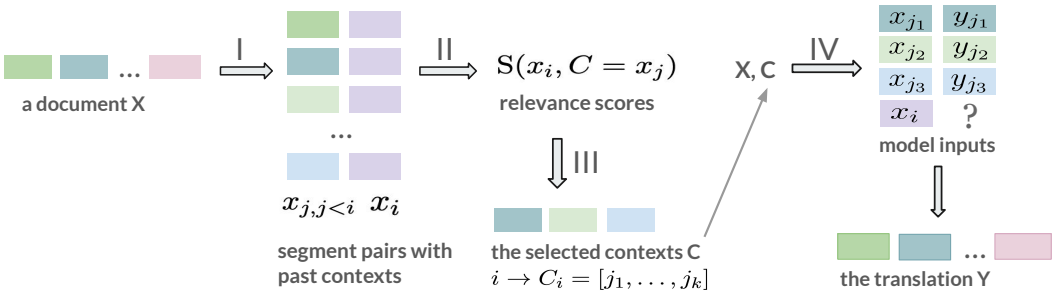

Figure 1: SELF-RAMT framework. It consists of four steps: I. Define the search domain (e.g. the past contexts); II. Compute contextual relevance scores for segments in the search domain; III. Retrieve relevant contexts according to the ranked scores and IV. Integrate the selected contexts to the inputs then generate the translations.

tomatically generated bilingual summary of the full source document, and a dynamic part, composed of the previous source and target sentences. However, the use of static, automatic summaries (a) has the effect of potentially changing the words of the context, which can be detrimental to lexical consistency and (b) does not ensure that the most relevant information is accessible for each source sentence, especially as documents increase in length.

Retrieval-augmented MT Choosing which context to be included in MT can be seen as a type of retrieval-augmented MT. In past works, retrieval-augmented MT systems have mostly been designed to mimic the use of *translation memories* by translators, which has a long history in MT (Kay, 1997). Recent implementations of this idea for neural models encode the target side of relevant example(s) together with the source sentence in an extended translation context (Gu et al., 2018; Bulte and Tezcan, 2019; Xia et al., 2019; Xu et al., 2020; He et al., 2021; Cheng et al., 2022). Variants, relying on both the source and target sides of the retrieved example(s) are proposed by Pham et al. (2020) and Reheman et al. (2023).

LLM-based MT systems seamlessly accommodate examples through in-context learning, where examples of the translation task (the source and target sides of parallel samples) (Radford et al., 2019) are inserted into the prompt. The optimal selection of in-context examples has also been the focus of recent research (Moslem et al., 2023; Vilar et al., 2023; Zhang et al., 2023; Bawden and Yvon, 2023; Agrawal et al., 2023; Cui et al., 2024; Zebaze et al., 2025), also analyzed by Zaranis et al. (2024) and Bouthors et al. (2024).

A key difference with SELF-RAMT is that these methods retrieve examples from external resources, instead of the input sequence, with the aim to find similar examples that can be easily edited. Several retrieval-augmented architectures have also been proposed, e.g., by Rubin and Berant (2024), to retrieve relevant contextual information from very long input documents. These approaches have been evaluated in language modeling tasks, but, to the best of our knowledge, have not yet been applied to MT.

3 Augmenting MT with Self-retrieval

3.1 A Self-retrieval Framework for MT

As illustrated in Figure 1, SELF-RAMT involves translating each segment of an input document X with relevant contexts retrieved within X. It consists of four steps:

I. Defining the Search Domain We consider a Doc2Sent scenario, translating sentences $\mathbf{x}_i$ in a document $X = \langle \mathbf{x}_1 \ldots \mathbf{x}_T \rangle$ using previous context sentences $\{\mathbf{x}_j, j < i\}$ in X.

II. Contextual Relevance Scores We compute contextual relevance scores $\mathrm{S}(\mathbf{x}_i, \mathbf{x}_j)$ of candidate segments $\mathbf{x}_j$ for each $\mathbf{x}_i$, with the aim of improving the consistency and coherence of the resulting translations. Details are in Section 3.2.

III. Context Retrieval For each $\mathbf{x}_i$, $\mathbf{x}_j$ is selected as a contextual segment if $\mathrm{S}(\mathbf{x}_i, \mathbf{x}_j)$ is among the top K relevance scores. Additionally, $\mathbf{x}_j$ is disregarded if $\mathrm{S}(\mathbf{x}_i, \mathbf{x}_j) \leq \tau$, where τ represents the minimum value such that $\mathbf{x}_j$ is relevant to $\mathbf{x}_i$ for score S (see Section 3.2). The resulting list of selected sentences, which we refer to as C_i, constitutes a dynamic context containing up to K sentences.

IV. Context-aware Translation Contextual sentences selected in step III are included as few-shot demonstrations in the LLM prompt in the order

in which they appear in the original text. We use specific prompts for each LLM. Details about the prompt selection and the decoding process are given in Appendix B.

3.2 Context Selection Criteria

Multiple criteria can be used to identify the relevant contextual sentences (reviewed in (Bouthors et al., 2024) for retrieval-based MT). In our approach, contextual relevance is assessed based on source side similarity between segments, which enables us to pre-compute the relevant contexts for all $\mathbf{x}_i$ prior to translation. Our hypothesis is that if $\mathbf{x}_j, j < i$ is sufficiently similar to $\mathbf{x}_i$, then $(\mathbf{x}_j, \mathbf{y}_j)$ will contain useful information when generating $\mathbf{y}_i$. This enables us to vary the retrieval score while keeping the translation infrastructure unchanged. It is therefore simpler than the proposal of Wang et al. (2025b), where contexts are dynamically updated during the generation process. In our experiments, we consider three contextual relevance scores:

COS We compute the cosine similarity between the sentence embeddings using LaBSE (Feng et al., 2022). Only positive cosine similarities are taken into account (i.e. $\tau = 0$).

BM25 We adapt BM25L (Lv and Zhai, 2011), a variant with length normalisation[3] of the Best Match 25 (BM25) relevance score (Robertson and Zaragoza, 2009), which is a go-to method for retrieving lexically relevant segments in large data stores. Implementation details are in Appendix B.2.

PMI To better reflect contextual relevance, we also consider an alternative inspired by the Pointwise Cross-Mutual Information (P-CXMI) (Fernandes et al., 2021, 2023) and the likelihood difference (Shi et al., 2024; Pombal et al., 2024). We identify relevant contexts based on the point-wise mutual information (PMI) between $\mathbf{x}_i$ and $\mathbf{x}_j$, defined as:

$$\mathrm{PMI}(\mathbf{x}_i, \mathbf{x}_j) = -\frac{1}{l_i} \sum_{t=1}^{l_i} \log \frac{P_C(x_{i,t}|x_{i,<t})}{P_C(x_{i,t}|x_{i,<t}, \mathbf{x}_j)},$$

where l_i is the length of $\mathbf{x}_i$. In other words, $\mathrm{PMI}(\mathbf{x}_i, \mathbf{x}_j)$ measures how much the knowledge of $\mathbf{x}_j$ reduces the uncertainty about $\mathbf{x}_i$ for some autoregressive language model P_C. We disregard $\mathbf{x}_j$ if $\mathrm{PMI}(\mathbf{x}_i, \mathbf{x}_j) \leq \tau$ with $\tau = 0$.

[3]Our code uses the Python implementation of Lù (2024).

Baseline We compare these relevance scores to four baselines: (i) Zero-shot, vanilla sentence-level MT, reflecting the basic non-contextual MT ability of LLMs, (ii) Past-K, where the local context is composed of K previous sentences, (iii) Random-K, where we randomly select K past sentences, and (iv) Indep-K (K independent examples generated by an LLM, the same for all sentences). More details are given in Appendix B.

4 Experimental Settings

4.1 Datasets

Our experiments rely on two test sets described below.

IWSLT Following Wang et al. (2025b) and Guo et al. (2025), we take the test sets of IWSLT2017[4] (Cettolo et al., 2012) as our test sets, for three translation directions: English to German (DE), French (FR) and Chinese (ZH), with respectively 10, 12 and 12 TED talks.

MERSENNE Due to the scarcity of long document-level data, we curated a set of 23 published scientific articles and their translations for the EN–FR language pair.[5] These articles are segmented into sentences and aligned into parallel articles. We refer to this test set as MERSENNE. More details on data preparation are given in Appendix A.

Statistics of test sets Table 1 reports the number of full documents and the number of sentences in our test sets. It also includes the average, minimum, and maximum length of sentences in LLAMA tokens, for the source and target languages. The average number of sentences is 119 for TED talks from IWSLT, and 192 for articles from MERSENNE.

4.2 Models and Inference Settings

We evaluate our framework with three open-weight medium-size multilingual LLMs using ICL: Llama3.1-8B-Instruct[6] (LLAMA) (Grattafiori et al., 2024), EuroLLM-9B-Instruct[7] (EUROLLM) (Martins et al., 2024, 2025), and Qwen2.5-7B-Instruct[8] (QWEN) (Qwen et al., 2025). These models do

[4]https://wit3.fbk.eu/2017-01-d

[5]https://www.centre-mersenne.org/

[6]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[7]https://huggingface.co/utter-project/EuroLLM-9B-Instruct

[8]https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

	IWSLT			MERSENNE
	en-de	en-fr	en-zh	en-fr
#doc	10	12	12	23
#sent	1138	1455	1459	4417
mean	20/25	21/26	20/24	36/53
min	2/2	2/2	2/2	1/1
max	106/143	93/121	93/117	256/348

Table 1: Statistics of IWSLT and MERSENNE, including the number of documents (#doc) and sentences (#sent). 'mean', 'min', and 'max' correspond respectively to the average, minimum, or maximum length of sentences, measured in LLAMA tokens.

not contain IWSLT nor parallel articles from MERSENNE in their pre-training data. Our experimental pipelines relies on vLLM (Kwon et al., 2023), an efficient framework for text generation. Decoding is performed with a beam width of 5 and a maximum number of new tokens of 256. To determine the impact of K (the maximum number of selected contexts), we vary K from 0 to 6. Regarding context selection, we compute PMI using LLAMA. For the Indep-K baseline, we generate 6 examples in the style of TED talks again using LLAMA. More details regarding the experimental setup are in Appendix B.

4.3 Metrics for DLMT

To evaluate general translation quality, we primarily rely on COMET (Rei et al., 2022) and its document-level variant (d-COMET) (Vernikos et al., 2022), with the reference-based model `wmt22-comet-da`. We also report BLEU[9] (Papineni et al., 2002) and SLIDE (Raunak et al., 2024) with `wmt22-cometkiwi-da` with a window size of 8 sentences and a stride of 6. For lexical consistency, we compute Lexical Translation Consistency Ratio (LTCR) (Lyu et al., 2021; Wang et al., 2025b) for proper nouns annotated using spaCy[10] and aligned using awesome-align (Dou and Neubig, 2021). We also conduct case studies to examine the effectiveness of SELF-RAMT.

5 Examining Context Selection Strategies

In this section, we aim to answer the following questions: (a) how effective are the context selection scores to identify relevant contexts? and (b) how similar are the retrieved segments when retrieval is performed in the source text or in the target text? As discussed above, generating coherent texts ideally requires taking the target context into account. As only the source text is initially available, it is important to verify that source-based retrieval is a reliable substitute for target-based retrieval, simulated using the oracle reference.

[9]We use SacreBLEU (Post, 2018) with signature: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.0. We use the default zh tokenizer for translations into Chinese.

[10]https://spacy.io/usage

		EN			FR		
K	rand.	COS	PMI	BM25	COS	PMI	BM25
1	0.52	0.12	0.01	0.22	0.09	0.10	0.32
2	0.52	0.17	0.09	0.26	0.15	0.10	0.30
3	0.52	0.19	0.11	0.28	0.21	0.11	0.31
4	0.52	0.23	0.10	0.28	0.23	0.11	0.34
5	0.51	0.24	0.10	0.29	0.25	0.12	0.34
6	0.51	0.25	0.10	0.29	0.25	0.13	0.36

Table 2: Extraction error rate on MERGEDTED using COS, PMI, and BM25, for K from 1 to 6, computed on the source (EN, left) or target texts (FR, right).

K	rand.	COS	PMI	BM25
1	0.03	0.39	0.54	0.43
2	0.05	0.47	0.52	0.53
3	0.07	0.60	0.46	0.56
4	0.08	0.61	0.46	0.58
5	0.09	0.63	0.55	0.62
6	0.11	0.64	0.61	0.62

Table 3: Cover rate on MERGEDTED using COS, PMI, and BM25, for K from 1 to 6.

Method To assess the context selection criteria for their sensitivity to coherence, we challenge their ability to distinguish sentences extracted from the same documents from other noise segments. Starting with a set of document pairs (X^1, X^2) both containing n sentences in the same language, we randomly shuffle sentences from X^1 with those of X^2, resulting in a combined document $X^{1,2}$. We then retrieve, from the first $2n-1$ segments of $X^{1,2}$, the K most relevant segments for the last sentence ($\mathbf{x}_{2n}$), using each relevance score, and compute the extraction error rate r, defined as the proportion of selected sentences that do not belong to the same document as $\mathbf{x}_{2n}$. For a set of N documents $\{X_l^{1,2}, l = 1 \dots N\}$, from which we retrieve the K context sentences $\{c_{l,1}, \dots, c_{l,K}\}$ for each $\mathbf{x}_{2n}$, r is computed as follows:

$$r = \frac{\sum_{l=1}^{N} \sum_{j=1}^{K} \mathbb{I}(\mathrm{doc}(c_j) \neq \mathrm{doc}(\mathbf{x}_{2n}))}{N \times K},$$

where $\mathrm{doc}(\mathrm{c})$ returns the document index of its input segment c.

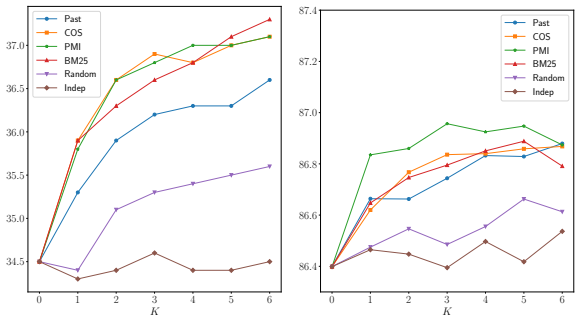

Figure 2: BLEU (left) and s-COMET (right) scores for IWSLT EN–FR translated using EuroLLM, with source and reference as contexts, for K from 0 to 6.

To compare the relevance scores computed using source and target texts, we also compute Kendall's τ (Kendall, 1938, 1945) between the relevance ranking respectively induced on the source and target texts, and average over the N documents. Finally, we also report the cover rate, defined as the ratio of context sentences c_j recognized by retrieval using both $X^{1,2}$ and $Y^{1,2}$, which is the reference translation of $X^{1,2}$, among all selected contexts retrieved from $Y^{1,2}$.[11]

MergedTED We artificially construct shuffled documents $X^{1,2}$ and their translations $Y^{1,2}$ from the 12 EN–FR talks from IWSLT. We consider the first 30 sentences of each talk as a pseudo-document, and include all 66 possible pairs as (X^1, X^2).[12] We refer to the resulting corpus as MergedTED.

Results For question (a), Table 2 reports the extraction error rate for COS, PMI and BM25, derived from MergedTED. We observe that, with the exception of PMI, error rates are quite high: already for $K = 1$, about 12% (resp. 22%) of the sentences retrieved by COS (resp. BM25) do not belong to the same talk as the focus sentence. In comparison, PMI error rates increase more slowly with the value of K. Regarding (b), we note that the error rates computed in the target language (FR) are only slightly higher than in the source (EN). Table 3 reports the cover rates of contextual segments selected in $X^{1,2}$ and in $Y^{1,2}$. Using the source document, the context criteria identify around half of the relevant contexts determined with the reference target document. Regarding the ranking of potential contexts, Kendall's τ between the relevance scores derived from the source and the reference are 0.55, 0.45 and 0.55 for COS, PMI and BM25 respectively. These results provide an empirical support of our main hypotheses, in particular they confirm that we can effectively perform context selection by only looking at the source side of the input documents, yet identify relevant dependencies on the target side.

[11] For the cover rate, we only count segments appearing in the same document as $\mathbf{x}_{2n}$.

[12] We consider 10 different shuffled versions for each pair (X^1, X^2), then report the average r and Kendall's τ.

6 Results and Analyses

6.1 MT quality with SELF-RAMT

The Impact of K To determine the optimal value for K and the best relevance criteria, we examined the BLEU and s-COMET scores for K from 0 (i.e. Zero-shot) to 6, with pairs of source and reference as contexts. Figure 2 displays representative results with the EN–FR translation of IWSLT using EuroLLM, where the context-augmented translations perform better than Random-K and Indep-K. Furthermore, compared to s-COMET, BLEU scores distinguish better translations using selected contexts from Past-K, indicating that these contexts lead to greater lexical similarity between the translations and the references. For a trade-off between quality and complexity, we take $K = 3$ for the following experiments and analysis.

Comparing Relevance Scores We perform context-aware evaluations on IWSLT, reported in Table 4. The results show that all reference-based metrics, including BLEU, s-COMET, and d-COMET, classify PMI as the best or the second-best relevance score for all models and translation directions. In contrast, SLIDE scores are less conclusive, ranking PMI as the top-2 best systems 7 out of 9 times. This suggests that PMI performs better than COS and BM25. LTCR only prefers PMI for EN–FR, while for EN–DE and EN–ZH, baseline methods give higher scores. This highlights a small issue with this metric, when we use the oracle reference as context. Assume that the MT engine translates the first instance of a term x as y_1, different from the reference version (y_2); then, for all the subsequent instances of the same term, we may retrieve the translation of the first instance (y_2), making the system more inclined to generate the same translation (y_2), which is what we want. This will introduce a discrepancy between the first instance (y_1) and the remaining ones, which will be penalized by LTCR. By comparison, the baseline system may appear more consistent.

		EuroLLM					Llama					Qwen				
		BLEU	s-comet	d-comet	slide	LTCR	BLEU	s-comet	d-comet	slide	LTCR	BLEU	s-comet	d-comet	slide	LTCR
DE	0-shot	27.8	85.4	75.8	81.8	95.8	24.7	82.7	72.3	79.5	95.2	22.8	81.6	70.6	77.2	91.2
	indep	28.1	85.4	75.8	81.8	**96.0**	24.4	82.7	71.9	79.5	95.9	22.8	81.2	69.7	76.5	91.8
	rand.	28.9	*85.7	*76.3	81.8	95.9	25.1	*83.3	*73.0	*79.9	**96.7**	22.6	*82.1	71.1	77.7	**92.8**
	past	29.7	*86.0	*76.8	$^{\Delta}$**82.0**	95.8	26.0	*83.7	*73.4	79.9	93.6	24.0	*82.6	*71.7	77.7	90.3
	COS	**30.2**	*86.1	*76.9	81.8	93.3	**26.8**	*83.9	*73.7	*80.0	94.5	24.4	*83.1	*72.5	**78.0**	88.8
	PMI	30.1	***86.2**	***77.1**	81.8	94.9	26.7	***84.2**	***74.1**	***80.3**	92.6	**25.1**	***83.2**	***72.8**	77.8	87.1
	BM25	29.5	*86.1	*76.8	81.8	93.6	26.6	*83.8	*73.6	79.8	94.9	24.9	*83.0	*72.4	77.9	89.5
FR	0-shot	40.1	86.4	76.8	83.3	88.9	36.5	84.4	74.0	81.8	87.2	34.5	83.9	73.2	80.7	89.6
	indep	41.3	86.4	77.0	83.4	89.9	36.5	84.3	74.0	81.8	87.5	34.8	84.0	*73.6	80.9	88.7
	rand.	41.5	86.5	*77.2	83.3	88.0	37.2	*84.8	*74.5	82.0	88.7	35.2	*84.4	*73.9	*81.2	89.6
	past	42.4	*86.7	*77.6	**83.5**	90.5	38.0	*84.9	*74.8	82.0	85.5	36.5	*84.5	*74.3	*81.4	88.4
	COS	42.8	*86.8	*77.7	83.3	89.7	38.6	*85.0	*75.0	$^{\delta}$**82.1**	86.8	36.7	*84.5	*74.3	$^{\Delta}$80.8	90.0
	PMI	**43.2**	***87.0**	***77.9**	**83.5**	**91.0**	38.6	***85.2**	***75.3**	$^{\delta}$**82.1**	**90.7**	37.1	*84.9	*74.8	*81.4	**93.9**
	BM25	43.1	*86.8	*77.7	83.4	90.4	**39.0**	*85.1	*75.2	82.0	87.9	**37.4**	***85.0**	***74.9**	***81.6**	90.8
ZH	0-shot	30.1	84.4	73.3	**81.3**	75.5	28.3	83.2	70.6	79.2	75.9	29.2	83.2	71.6	78.6	79.0
	indep	30.5	*84.7	*73.8	**81.3**	78.9	29.1	*83.6	*71.9	***79.7**	**76.4**	29.8	*83.7	*72.5	*79.5	81.3
	rand.	30.8	*84.7	*74.0	81.2	78.2	29.4	*83.5	*71.6	79.1	76.3	30.3	*84.1	*73.0	*79.4	**81.5**
	past	31.6	*85.0	*74.4	81.2	**79.7**	30.6	*83.9	*72.4	79.3	74.7	31.5	*84.6	*73.7	*79.4	78.9
	COS	32.0	*84.9	*74.3	80.3	73.2	31.4	*83.9	*72.5	79.2	73.3	32.0	*84.5	*73.7	*79.5	75.9
	PMI	**32.4**	***85.2**	***74.7**	81.0	74.0	**31.6**	***84.0**	***72.9**	79.3	75.1	**32.2**	***84.7**	***74.1**	***79.7**	75.1
	BM25	32.3	*85.0	*74.4	80.6	73.8	31.5	*83.9	*72.4	78.9	72.0	32.1	*84.5	*73.6	*79.5	73.7

Table 4: Results for IWSLT (source and reference as context). We mark MT systems that are significantly better than zero-shot in COMET-based scores, for sentences excluding the first 20 ones ($^{\Delta}$), all sentences ($^{\delta}$), or in both cases (*), with p-value < 0.05. The top two scores are marked in bold (best) and underlined (second-best).

Context		BLEU	s-comet	d-comet	slide	LTCR
	0-shot	55.7	89.5	86.6	74.6	**92.8**
SRC+REF	rand.	58.7	*89.6	*86.8	*74.8	91.7
	past	59.8	*89.8	*87.0	***74.9**	91.8
	COS	60.8	*89.9	*87.2	***74.9**	90.9
	PMI	**61.0**	***90.0**	***87.3**	***74.9**	91.5
	BM25	60.5	*89.9	*87.2	***74.9**	91.2
SRC+MT	rand.	55.8	89.5	86.7	*74.7	91.4
	past	55.0	***89.6**	*86.8	*74.8	92.3
	COS	56.2	***89.6**	***86.9**	*74.8	92.6
	PMI	**56.3**	***89.6**	***86.9**	***74.9**	**92.8**
	BM25	56.2	***89.6**	***86.9**	*74.8	92.7

Table 5: Results for MERSENNE using EuroLLM. * indicates significant gains as in Table 4.

SRC+MT		BLEU	s-comet	d-comet	slide	LTCR
DE	0-shot	27.8	85.4	75.8	81.8	95.8
	rand.	28.5	*85.7	*76.2	82.0	96.0
	past	28.5	*85.7	*76.4	***82.1**	**96.6**
	COS	28.3	*85.7	*76.2	82.0	96.5
	PMI	**28.9**	***85.9**	***76.6**	***82.1**	95.4
	BM25	28.3	*85.7	*76.3	81.9	95.2
FR	0-shot	40.1	86.4	76.8	83.3	88.9
	rand.	41.2	86.5	*77.2	**83.4**	89.2
	past	41.2	86.5	*77.3	**83.4**	87.6
	COS	41.2	86.4	*77.2	**83.4**	92.1
	PMI	**41.8**	***86.6**	***77.5**	**83.4**	90.9
	BM25	41.5	86.4	*77.2	**83.4**	**92.4**
ZH	0-shot	30.1	84.4	73.3	**81.3**	75.5
	rand.	30.6	*84.6	*73.7	81.2	83.1
	past	30.7	84.6	*73.6	81.2	85.1
	COS	30.7	84.5	*73.6	81.2	85.5
	PMI	**30.8**	***84.7**	***73.9**	81.2	**90.0**
	BM25	30.2	84.2	73.2	80.1	88.8

Table 6: Results for IWSLT (source and MT as contexts) using EuroLLM. * indicates significant gains as in Table 4.

Note that this problem does not arise when we retrieve automatic translations, where we see the benefits of SELF-RAMT more clearly.

The evaluation results for MERSENNE and IWSLT translated with reference and automatic translations as context (using EuroLLM) are given in Tables 5 and 6. These results lead to the same conclusion that PMI is a good criterion for retrieving relevant past segments. As the LTCR scores rank the context selection methods differently across test sets, we conduct a follow-up case study presented in Section 6.3, which better highlights the contribution of selected contexts to term consistency. The complete scores for all MT systems are in Tables 11 and 12 in Appendix C.

6.2 Analysis of Context Contribution

Distance Contexts are Retrieved Contextual information plays a crucial role in DLMT. To analyze the contribution of relevant contexts (especially distant ones) to the translation quality, we bin sentences according to the distance (in number of sentences) to their most distant selected context. For example, all selected contexts of sentences in the group 0–20 are retrieved from the past 20 sentences, while there are contexts more distant than 64 sentences for members of the group 64–256.

We then measure the ratio between the effective number of contexts occurring within a given interval (e.g., $20 - 40$) and the maximal possible number of contexts in that interval.[13] This analysis

[13]Which depends on the sentence position in a document: sentences in the initial paragraphs only have access to a re-

Range	rand. nb	rand. ratio	COS nb	COS ratio	PMI nb	PMI ratio	BM25 nb	BM25 ratio
0-20	2066	0.26	3322	0.41	4575	0.57	3432	0.42
20-40	2069	0.26	1708	0.21	1407	0.17	1568	0.19
40-80	2685	0.34	2028	0.25	1417	0.18	2116	0.27
80-120	900	0.21	740	0.17	498	0.12	663	0.16
120-256	356	0.20	278	0.16	179	0.10	297	0.17

Table 7: Distance between the translated sentence and selected contexts, for sentences appearing after the 40^{th} sentence in IWSLT with $K = 3$. 'ratio' denotes the effective value ('nb') normalized by the number of selected contexts for sentences that have access to the corresponding distance interval.

IWSLT Range	DE	FR	ZH
0-20	337	390	373
20-40	241	350	352
40-64	125	223	236
64-256	235	252	258

MERSENNE Range	FR
0-20	966
20-40	770
40-64	650
64-128	995
128-320	570

Table 8: Retrieval statistics with respect to the distance between the translated sentence and selected contexts, for sentences appearing after the 20^{th} sentence. Selection is performed with PMI and $K = 3$, for IWSLT (left) and MERSENNE (right).

is performed for the TED talks test set (IWSLT), for $K = 3$. We exclude from the analysis the first 40 sentences in each document, as the context they can access is limited. The corresponding statistics are in Table 7.

We observe that about half of the retrieved contexts are in the past 20 sentences, while a sizable portion of contexts are chosen in more distant part of the document (in the $20 - 80$ range); more remote sentences, with a distance larger than 80 are also frequently selected. There is a clear variance between relevance scores: COS retrieves closer segments on average, whereas PMI and BM25 are more likely to extract more remote sentences.

Distant Contexts Matter For each group of sentences, we now compare the translations generated with PMI and with the baseline methods. Table 8 reports the corresponding retrieval statistics for these experiments, where we again group sentences by their position index in the source text.

Translation scores are in Table 9, where we compute the d-COMET difference between baselines and PMI-based retrieval. The upper part of Table 9 shows that, for translations of IWSLT with

stricted contexts, while sentences occurring in the last position have a much larger set of contextual segments to chose from.

SRC	The hemihedria is, moreover, non-superposable.
REF	L'hémiédrie est, en outre, non superposable.
0-shot	La hemihedrie est, en outre, non superposable.
Past	De plus, l'hémimorphie n'est pas superposable.
PMI	De plus, l'hémiédrie n'est pas superposable.

Figure 3: Translations of the 72^{nd} sentence of a MERSENNE article, using EUROLLM with MT as target-side context. The correct translations of "hemihedria" are underlined.

SRC: And he found a match.
REF: 结果找到了一个配对！ *(a match)*
0-shot: 他找到了合适的人。*(an appropriate person)*
Past: 他找到了匹配的物品。*(a matched object)*
PMI: 他找到了匹配的物种。*(a matched species)*

The contexts of PMI:
English: It turns out that different species have slightly different structures of collagen, so if you get a collagen profile of an unknown bone, you can compare it to those of known species, and, who knows, maybe you get a match.
Chinese: 事实证明，不同物种的胶原蛋白结构略有不同，因此如果你得到一块未知骨头的胶原蛋白谱，你可以将其与已知物种的胶原蛋白谱进行比较，谁知道，也许能找到匹配。
English: So she shipped him one of the fragments, FedEx.
Chinese: 于是她通过联邦快递把其中一块碎片寄给了他。
English: LN: And he processed it, and compared it to 37 known and modern-day mammal species.
Chinese: LN: 他对它进行了处理，并将其与 37 种已知和现代哺乳动物进行了比较。

Figure 4: EN–ZH translations of the 45^{th} sentence from an IWSLT talk, using EUROLLM with MT as target-side context. PMI retrieves relevant contexts for "a match" that corresponds to a specie (see Figure 7 in Appendix C for more details.)

source and reference as contexts, integrating remote contexts selected by PMI leads to better translation quality than Zero-shot, Random and Past. The bottom part reports the results for translations with source and automatic translations as contexts. In this setting, translation using EUROLLM with PMI is still better than Zero-shot, Random and Past, with lesser performance gains. In contrast, for QWEN, PMI appears to do less well than Past. The performance of LLAMA depends on the language pair, PMI being best only for EN–ZH translations.

We report a similar analysis in Table 10 for MERSENNE corpus. The results show that PMI outperforms Zero-shot and Random for all models using reference or automatic translations as target side context. In all cases except four, PMI is better than Past in d-COMET, especially for QWEN.

6.3 A Case study: Lexical Consistency

We illustrate the benefits of retrieving relevant contexts for the adequate translation in context-dependent cases. A first example is in Figure 4: to translate *a match*, which corresponds here to a

	dist	DE EURO	DE LLAMA	DE QWEN	FR EURO	FR LLAMA	FR QWEN	ZH EURO	ZH LLAMA	ZH QWEN
SRC+REF										
PMI − sent	0-20	*1.4	*2.3	*2.7	*1.0	*1.2	*1.7	*1.6	*2.8	*2.7
	20-40	*1.4	*1.4	*1.5	*1.3	*1.6	*1.7	*1.5	*2.2	*2.2
	40-64	*1.4	*2.2	*2.1	*0.9	*1.5	*1.9	*1.5	*2.9	*2.9
	64-256	*1.5	*2.2	*2.4	*1.3	*1.9	*1.6	*1.1	*2.0	*3.1
PMI − rand	0-20	*1.1	*1.4	*2.0	*0.7	*1.3	0.5	*0.9	*2.1	*1.6
	20-40	*0.9	0.7	*1.3	*1.0	*0.7	*1.0	*0.9	*1.5	*0.6
	40-64	0.3	*1.2	*2.9	0.5	0.6	*1.0	*1.1	*1.2	*1.5
	64-256	*1.2	*1.7	*1.8	*1.0	*0.8	0.3	0.5	*1.3	*1.5
PMI − past	0-20	0.3	*0.8	0.8	0.2	*0.9	0.3	0.2	0.4	0.1
	20-40	0.2	*0.7	0.7	*0.5	0.5	*0.7	0.4	*0.8	0.3
	40-64	0.3	0.5	*1.8	0.5	*0.7	*1.5	0.4	0.5	*1.0
	64-256	0.6	*1.3	*1.9	0.4	0.4	-0.4	-0.2	*0.9	*0.8
SRC+MT										
PMI − sent	0-20	*0.7	0.2	0.5	*0.5	-0.2	0.4	*0.7	*1.5	*1.6
	20-40	*0.9	-0.2	-0.7	*0.8	-0.3	0.6	*0.7	*1.2	*0.9
	40-64	*1.6	-0.3	0.8	0.4	0.2	0.7	0.4	*1.3	*1.1
	64-256	*0.8	-0.7	-0.3	*0.8	*0.7	*1.2	*0.7	*1.1	*1.7
PMI − rand	0-20	0.5	0.3	0.4	0.3	0.2	0.3	0.1	*1.2	*0.8
	20-40	*0.7	-0.5	-0.7	*0.5	-0.5	*0.7	*0.6	*1.3	-0.1
	40-64	0.3	-0.6	*2.2	0.3	-0.3	-0.2	0.3	0.5	0.1
	64-256	0.6	-0.5	-0.2	0.3	0.2	0.2	0.4	*1.0	0.6
PMI − past	0-20	0.1	-0.5	-0.8	0.1	-0.1	-0.2	0.2	0.5	-0.1
	20-40	-0.0	-0.4	*-1.0	*0.3	-0.5	0.2	0.4	*0.7	-0.3
	40-64	*1.0	*-0.9	0.6	0.4	0.3	0.7	0.4	0.2	-0.4
	64-256	0.1	*-0.8	-0.2	-0.1	-0.0	-0.1	-0.0	*0.8	0.5

Table 9: Average differences between the d-COMET of translations using PMI and three baselines. From top to bottom these are: zero-shot (sent), random (rand), past (past). We only consider all sentences occurring past the 20^{th} sentence in IWSLT articles, and take into account the maximum distance between the selected contexts and the current sentence. * marks a significant difference with p-value < 0.05.

	dist	SRC+REF EURO	SRC+REF LLAMA	SRC+REF QWEN	SRC+MT EURO	SRC+MT LLAMA	SRC+MT QWEN
sent	0-20	*0.6	*0.9	*1.4	*0.3	*0.5	*0.8
	20-40	*0.7	*0.8	*1.6	*0.4	*0.3	*0.9
	40-64	*0.9	*0.7	*1.9	*0.5	0.3	*1.2
	64-120	*0.7	*0.9	*1.5	*0.2	*0.3	*0.8
	120-320	*0.6	*1.0	*1.6	-0.0	*0.4	*0.6
rand	0-20	*0.3	*0.5	*0.8	0.0	*0.2	*0.3
	20-40	*0.3	*0.5	*0.5	*0.2	*0.3	*0.3
	40-64	*0.6	*0.5	*0.9	*0.3	0.2	*0.5
	64-120	*0.5	*0.6	*0.7	0.1	0.0	*0.3
	120-320	*0.4	*0.9	*0.7	-0.0	0.2	-0.1
past	0-20	-0.0	*0.1	*0.5	-0.0	0.0	*0.3
	20-40	*0.2	*0.2	*0.5	0.1	0.0	*0.2
	40-64	*0.4	0.1	*0.6	0.1	*-0.3	*0.4
	64-120	*0.3	*0.4	*0.5	0.1	0.1	*0.2
	120-320	0.2	*0.4	*1.0	-0.1	0.0	0.4

Table 10: Average differences between the d-COMET of translations using PMI and three baselines. From top to bottom these are: zero-shot (sent), random (rand), past (past). We only consider all sentences occurring past the 20^{th} sentence in MERSENNE articles, and take into account the maximum distance between the selected contexts and the current sentence. * marks significant difference with p-value < 0.05.

matched species, PMI retrieves the relevant contexts and generates the correct translation, while Zero-shot translates it as "an appropriate person" and Past considers it as a matched object. Complementary details and another Chinese example are presented in Appendix C.

Distant contexts are necessary to ensure lexical consistency, especially when translating full articles, and we do observe some evidence for improved translation consistency. For example, in one particular MERSENNE article, the term *hemihedria* appears 12 times,[14] with a consistent French reference translation *hémiédrie*. Table 3 shows the translation of the 72^{nd} sentence using EUROLLM taking source and MT as context.

PMI achieves consistent translation for this term, while Past fails due to the absence of remote contexts, resulting in 11 translation errors with 4 variants. Zero-shot translates them into 6 different forms with only one correct translation, despite its high LTCR score.

Figure 8 and Figure 9 in Appendix C provide the contexts selected using PMI for the 72^{nd} and the 159^{th} sentences respectively, including the sentences in lines 51 and 58 that contain the term *hemihedria*.

7 Conclusion

To achieve better document-level machine translation for extra long documents, we propose the SELF-RAMT framework. By retrieving relevant sentences in a local translation memory composed

[14] Lines 26, 39, 51, 53, 56, 58, 72, 78, 89, 122, 159, 160.

of sentences that have already been translated, and possibly post-edited, in the same document, we expect to generate translations that are globally more consistent. We carry out experiments to translate full TED talks and a novel parallel corpus consisting of scientific articles, using three LLMs with in context learning, to integrate the retrieved past contexts.

We further analyze the influence of distant contexts on translation quality, according to their distance from the current sentence. Our findings show that incorporating distant contexts, selected using context criteria such as PMI, can be useful for better lexical consistency. Distant context also seem to improve the general translation quality, as measured by reference-based scores.

Limitations

In this work, we only considered three open-weight models, excluding close-source models such as GPT4 (OpenAI et al., 2023) the use of which makes results difficult, if possible at all, to reproduce for scientific purposes. The translation of low-resource languages is not included in our experiments, although our framework could be beneficial in that scenario. In our experimental settings, we process each document sentence per sentence, defining the translation unit and the contextually relevant sentences based on this predefined segmentation. Segmenting and processing input documents in larger chunks, made of several consecutive sentences, is left for future work. Our evaluation was based mainly on derivatives of BLEU and COMET metrics, with a variant of LTCR and some case studies. While we reckon that some better metrics should be applied to better reflect document translation quality, measuring for example, the consistency and coherence of the translated text, we also can only regret that such standard evaluation metrics have not yet developed nor adopted by the community.

Ethics Statement

This work has conducted experiments and analysis using open source models, tools, and corpus. We see no ethical problem with this study.

Acknowledgments

This work was supported by the French national agency (ANR) as part of the MaTOS project under reference ANR-22-CE23-0033.[15] Rachel Bawden was also partly funded by her chair position in the PRAIRIE institute funded by ANR as part of the "Investissements d'avenir" programme under reference ANR19-P3IA-0001. The authors wish to thank Célia Vaudaine and Caroline Rossi for giving access to the MERSENNE corpus. The authors are also grateful to the anonymous reviewers for their insightful comments and suggestions, and to Paul Lerner and Lichao Zhu for their review and feedback on a preliminary draft of this work.

[15] http://anr-matos.github.io/

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland. European Association for Machine Translation.

Maxime Bouthors, Josep Crego, and François Yvon. 2024. Retrieving examples from memory for retrieval augmented neural machine translation: A systematic comparison. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3022–3039, Mexico City, Mexico. Association for Computational Linguistics.

Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024. Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1301–1317, Miami, Florida, USA. Association for Computational Linguistics.

Bram Bulte and Arda Tezcan. 2019. Neural fuzzy repair: Integrating fuzzy matches into neural machine

translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1800–1809, Florence, Italy.

Sheila Castilho and Rebecca Knowles. 2024. A survey of context in neural machine translation and its evaluation. *Natural Language Processing*, page 1–31.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Xin Cheng, Shen Gao, Lemao Liu, Dongyan Zhao, and Rui Yan. 2022. Neural machine translation with contrastive translation memories. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3591–3601, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Menglong Cui, Jiangcun Du, Shaolin Zhu, and Deyi Xiong. 2024. Efficiently exploring large language models for document-level machine translation with in-context learning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10885–10897, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Yichen Dong, Xinglin Lyu, Junhui Li, Daimeng Wei, Min Zhang, Shimin Tao, and Hao Yang. 2025. Two intermediate translations are better than one: Fine-tuning LLMs for document-level translation refinement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14917–14933, Vienna, Austria. 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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. When does translation require context? a data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 606–626, Toronto, Canada. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6467–6478, Online. Association for Computational Linguistics.

Harritxu Gete, Thierry Etchegoyhen, David Ponce, Gorka Labaka, Nora Aranberri, Ander Corral, Xabier Saralegi, Igor Ellakuria, and Maite Martin. 2022. TANDO: A corpus for document-level machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3026–3037, Marseille, France. European Language Resources Association.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, and Bobbie Chern et al. 2024. The Llama 3 Herd of Models. Preprint arXiv:2407.21783.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O.K. Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *International Conference on Learning Representations*.

Jiaxin Guo, Yuanchang Luo, Daimeng Wei, Ling Zhang, Zongyao Li, Hengchao Shang, Zhiqiang Rao, Shaojun Li, Jinlong Yang, Zhanglin Wu, and Hao Yang. 2025. Doc-Guided Sent2Sent++: A Sent2Sent++ Agent with Doc-Guided memory for Document-level Machine Translation. Preprint arXiv:2501.08523.

Jiaxin Guo, Hao Yang, Zongyao Li, Daimeng Wei, Hengchao Shang, and Xiaoyu Chen. 2024. A novel paradigm boosting translation capabilities of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 639–649, Mexico City, Mexico. Association for Computational Linguistics.

Qiuxiang He, Guoping Huang, Qu Cui, Li Li, and Lemao Liu. 2021. Fast and accurate neural machine translation with translation memory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3170–3180, Online. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. Preprint arXiv:2302.09210.

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 the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2242–2254, Online. Association for Computational Linguistics.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.

Martin Kay. 1997. The proper place of men and machines in language translation. *Machine Translation*, 12(3–23).

Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Maurice G. Kendall. 1945. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251.

Sai Koneru, Miriam Exel, Matthias Huck, and Jan Niehues. 2024. Contextual refinement of translations: Large language models for sentence and document-level post-editing. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2711–2725, Mexico City, Mexico. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 2020. Does multi-encoder help? a case study on context-aware neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3512–3518, Online. Association for Computational Linguistics.

Chen Li, Meishan Zhang, Xuebo Liu, Zhaocong Li, Derek Wong, and Min Zhang. 2024. Towards demonstration-aware large language models for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13868–13881, Bangkok, Thailand. Association for Computational Linguistics.

Yachao Li, Junhui Li, Jing Jiang, Shimin Tao, Hao Yang, and Min Zhang. 2023. P-transformer: Towards better document-to-document neural machine translation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 31:3859–3870.

Zongyao Li, Zhiqiang Rao, Hengchao Shang, Jiaxin Guo, Shaojun Li, Daimeng Wei, and Hao Yang. 2025. Enhancing large language models for document-level translation post-editing using monolingual data. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8830–8840, Abu Dhabi, UAE. Association for Computational Linguistics.

Lei Liu and Min Zhu. 2022. Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38(2):621–634.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2022a. Divide and rule: Effective pre-training for context-aware multi-encoder translation models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4557–4572, Dublin, Ireland. Association for Computational Linguistics.

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2022b. Focused concatenation for context-aware neural machine translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 830–842, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yuanhua Lv and ChengXiang Zhai. 2011. When documents are very long, bm25 fails! In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, page 1103–1104, New York, NY, USA. Association for Computing Machinery.

Xinglin Lyu, Junhui Li, Zhengxian Gong, and Min Zhang. 2021. Encouraging lexical translation consistency for document-level neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3265–3277, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xing Han Lù. 2024. Bm25s: Orders of magnitude faster lexical search via eager sparse scoring. Preprint arXiv:2407.03618.

Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020. A simple and effective unified encoder for document-level machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3505–3511, Online. Association for Computational Linguistics.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Manuel Faysse, Pierre Colombo, François Yvon,

Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. EuroLLM-9B: Technical Report. Preprint arXiv:2506.04079.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe. Preprint arXiv:2409.16235.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3092–3102, Minneapolis, Minnesota. Association for Computational Linguistics.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2).

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

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

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A lightweight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90, Online. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, and Lenny Bogdonoff et al. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Proyag Pal, Alexandra Birch, and Kenneth Heafield. 2024. Document-level machine translation with large-scale public parallel corpora. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13185–13197, Bangkok, Thailand. Association for Computational Linguistics.

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

Ziqian Peng, Rachel Bawden, and François Yvon. 2025. Investigating length issues in document-level machine translation. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 4–23, Geneva, Switzerland. European Association for Machine Translation.

Minh Quang Pham, Jitao Xu, Josep Crego, François Yvon, and Jean Senellart. 2020. Priming neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 516–527, Online. Association for Computational Linguistics.

José Pombal, Sweta Agrawal, Patrick Fernandes, Emmanouil Zaranis, and André F. T. Martins. 2024. A context-aware framework for translation-mediated conversations. Preprint arXiv:2412.04205.

Andrei Popescu-Belis. 2019. Context in neural machine translation: A review of models and evaluations. Preprint arXiv:1901.09115.

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

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, and Rui Men et al. 2025. Qwen2.5 technical report. Preprint arXiv:2412.15115.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI blog.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. TransQuest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Vikas Raunak, Tom Kocmi, and Matt Post. 2024. SLIDE: Reference-free evaluation for machine translation using a sliding document window. In *Proceedings of the 2024 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 205–211, Mexico City, Mexico. Association for Computational Linguistics.

Abudurexiti Reheman, Tao Zhou, Yingfeng Luo, Di Yang, Tong Xiao, and Jingbo Zhu. 2023. Prompting neural machine translation with translation memories. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

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. 2022. 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.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Ohad Rubin and Jonathan Berant. 2024. Retrieval-pretrained transformer: Long-range language modeling with self-retrieval. *Transactions of the Association for Computational Linguistics*, 12:1197–1213.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico. Association for Computational Linguistics.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3537–3548, Dublin, Ireland. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2023. Efficient Transformers: A Survey. *ACM Computing Surveys*, 55(6):1–28.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. Context-aware monolingual repair for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 877–886, Hong Kong, China. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019b. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Kuang-Da Wang, Teng-Ruei Chen, Yu Heng Hung, Shuoyang Ding, Yueh-Hua Wu, Yu-Chiang Frank Wang, Chao-Han Huck Yang, Wen-Chih Peng, and Ping-Chun Hsieh. 2025a. Plan2align: Predictive planning based test-time preference alignment in paragraph-level machine translation. Preprint arXiv:2502.20795.

Longyue Wang, Zefeng Du, Wenxiang Jiao, Chenyang Lyu, Jianhui Pang, Leyang Cui, Kaiqiang Song, Derek Wong, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking and improving long-text translation with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7175–7187, Bangkok, Thailand. Association for Computational Linguistics.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023. Augmenting language models with long-term memory. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yutong Wang, Jiali Zeng, Xuebo Liu, Derek F. Wong, Fandong Meng, Jie Zhou, and Min Zhang. 2025b. DelTA: An online document-level translation agent based on multi-level memory. In *The Thirteenth International Conference on Learning Representations*.

Minghao Wu, Yufei Wang, George Foster, Lizhen Qu, and Gholamreza Haffari. 2024. Importance-aware

data augmentation for document-level neural machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 740–752, St. Julian's, Malta. Association for Computational Linguistics.

Mengzhou Xia, Guoping Huang, Lemao Liu, and Shuming Shi. 2019. Graph based translation memory for neural machine translation. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Jitao Xu, Josep Crego, and Jean Senellart. 2020. Boosting neural machine translation with similar translations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1580–1590, Online. Association for Computational Linguistics.

Jian Yang, Yuwei Yin, Shuming Ma, Liqun Yang, Hongcheng Guo, Haoyang Huang, Dongdong Zhang, Yutao Zeng, Zhoujun Li, and Furu Wei. 2023. Hanoit: Enhancing context-aware translation via selective context. In *Database Systems for Advanced Applications: 28th International Conference, DASFAA 2023, Tianjin, China, April 17–20, 2023, Proceedings, Part III*, page 471–486, Berlin, Heidelberg. Springer-Verlag.

Zhengxin Yang, Jinchao Zhang, Fandong Meng, Shuhao Gu, Yang Feng, and Jie Zhou. 2019. Enhancing context modeling with a query-guided capsule network for document-level translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1527–1537, Hong Kong, China. Association for Computational Linguistics.

Lei Yu, Laurent Sartran, Wojciech Stokowiec, Wang Ling, Lingpeng Kong, Phil Blunsom, and Chris Dyer. 2020. Better document-level machine translation with Bayes' rule. *Transactions of the Association for Computational Linguistics*, 8:346–360.

Emmanouil Zaranis, Nuno M. Guerreiro, and Andre Martins. 2024. Analyzing context contributions in LLM-based machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14899–14924, Miami, Florida, USA. Association for Computational Linguistics.

Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. 2025. In-context example selection via similarity search improves low-resource machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1222–1252, Albuquerque, New Mexico. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: a case study. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. 2021. Towards making the most of context in neural machine translation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

Zhang Zhuocheng, Shuhao Gu, Min Zhang, and Yang Feng. 2023. Addressing the length bias challenge in document-level neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11545–11556, Singapore. Association for Computational Linguistics.

Álvaro Peris and Francisco Casacuberta. 2019. Online learning for effort reduction in interactive neural machine translation. *Computer Speech & Language*, 58:98–126.

A Mersenne

Due to the scarcity of complete parallel scholarly documents, we constructed MERSENNE, which consists of 23 articles in English and their translation into French prepared by the Mersenne Center.[16] Nineteen of them report recent research in the geosciences domain and the remaining four belong to the chemical sciences. The translations are human post-edits of an initial machine-translated version.

For each article, we first convert the curated html page to plain texts using `pandoc`.[17] Extra empty lines and the symbol xa0 are removed before normalizing the texts to the NFC[18] format through `unicodedata`.[19] We then extract the article from the processed plain text, excluding equations and tables, which are reserved for future exploitation. We segment the texts into sentences and align the sentences to parallel articles. For sentence segmentation, we use Trankit (Nguyen et al., 2021) as it recognizes lists of citations well. Our alignment tool is derived from BertAlign (Liu and Zhu, 2022), which supports many-to-many alignments. All the alignments matched with an empty string were checked and manually adjusted whenever needed.

[16] https://www.centre-mersenne.org/

[17] https://pandoc.org/

[18] Normalization Form Canonical Composition.

[19] https://docs.python.org/3/library/unicodedata.html

We subsequently evaluated the aligned sentence pairs using TransQuest (Ranasinghe et al., 2020). All alignment scores were above 0.75, suggesting that sentence alignment is mostly correct and that all aligned sentences can be kept. Statistics regarding the MERSENNE parallel articles can be found in Table 1.

B Experimental Details

This section presents details about the inference settings, the computation of relevance scores, and the prompt patterns for ICL.

B.1 Inference

In our experiments, we obtain automatic translations through ICL using vLLM (Kwon et al., 2023) in bfloat16. We set the beam width to 5 and the temperature to 0. The minimum and maximum number of new tokens are 1 and 256 respectively.

For the decoding process, we propose a multi-turn decoding algorithm to access the incrementally generated target-side contexts. This involves translating the i^{th} sentences of all documents from the same batch in parallel, and updating the pre-selected contexts for the $(i+1)^{th}$ sentences with the generated translations. When contexts comprise source texts and reference translations of past sentences, we integrate them in the LLM prompts and decode them all at once.

Regarding context selection, we compute PMI scores for source texts using LLAMA, performing compuations in bfloat32 for IWSLT and in bfloat16 for MERSENNE and MERGEDTED.

B.2 Implementation of BM25L

In practice, given a document X, we preprocess each sentence,[20] then compute the term frequency and the inverse document frequency (IDF) for each term in the whole document. We also precomputed an IDF for terms from the training split of IWSLT-2016 (Cettolo et al., 2012)[21] for the EN–FR language pair. Therefore, the IDF of a term in X is replaced by the precomputed values if available.

B.3 Prompt Patterns

We integrate the bilingual context for DLMT in a few-shot template for ICL. As all our MT engines are instruction-tuned, we use the chat template. Regarding prompt design, we empirically tested the performance of the Past-K baseline using different prompt patterns on the IWSLT TST2010 and TST2011 test sets, for $K \in \{2, 6\}$. After disregarding some patterns that lead to over-generation, we selected the prompt templates in Figure 6, where we integrate the few-shot contexts into system prompts for EUROLLM and LLAMA, and into user prompts for QWEN.

For each sentence, we apply the prompt pattern without context to perform zero-shot translation. For the Indep-K baseline, we maintain K-shot demonstrations for all sentences. In practice, we generate 6 examples in the style of TED talks using LLAMA,[22] then integrate the first K examples as K-shot demonstrations.

C Complementary Evaluation Results

Translation Quality Based on the complete evaluation scores for IWSLT with source and automatic translations as contexts (see Table 11), and MERSENNE (see Table 12), we can confirm that in general PMI performs better than COS and BM25, as discussed in Section 6.1. On the other hand, we also observed that the performance of LLAMA is worse than EUROLLM. This quality degradation has a strong influence on DLMT using multi-turn decoding.

Case Study In this section, we provide additional examples that illustrate the need to integrate remote contexts for term consistency when translating long documents, and relevance criteria such as PMI can identify such contexts. This analysis is complementary to the one in Section 6.3. For example, a consistent translation of the term *hemihedria* in the 159^{th} sentence of a MERSENNE article requires contexts more distant than the past 30 sentences, and contexts with the K highest PMI scores include these relevant but remote sentences (see Figure 9). Figure 5 displays another example that contextual information from the past 10^{th} sentence is required for the consistent translation of "the High Arctic".

Figure 7 gives complementary information for the example in Figure 4, including the two successive sentences of the source sentence to be trans-

[20] The preprocessing consists of lowercasing, stop-word removal and stemming using PyStemmer: https://pypi.org/project/PyStemmer/.

[21] https://wit3.fbk.eu/2016-01

[22] We use the following chat template integrated in the system prompt "You are a good translation assistant!" and the user prompt "Give six few-shot examples to assist the translation from English to {tgt_lang} in the style of TED talks."

		EuroLLM					Llama					Qwen				
		BLEU	s-comet	d-comet	slide	LTCR	BLEU	s-comet	d-comet	slide	LTCR	BLEU	s-comet	d-comet	slide	LTCR
DE	0-shot	27.8	85.40	75.77	81.83	95.85	24.7	82.73	72.25	79.52	95.23	22.8	81.64	70.61	77.21	91.16
	indep	28.1	85.42	75.81	81.83	96.02	24.4	82.69	71.91	79.46	95.88	22.8	81.23	69.71	76.54	91.83
	rand.	28.5	*85.70	*76.18	81.98	96.01	24.4	82.88	72.41	79.73	95.08	22.7	81.75	70.65	77.43	93.48
	past	28.5	*85.69	*76.43	***82.07**	**96.57**	24.5	*83.09	*72.71	***80.08**	96.64	**23.0**	81.93	71.07	77.74	91.61
	COS	28.3	*85.69	*76.24	82.00	96.50	24.5	82.85	72.27	79.71	**97.95**	22.9	***82.15**	$^\delta$**71.19**	77.81	**95.72**
	PMI	**28.9**	***85.92**	***76.62**	*82.06	95.38	24.2	82.68	72.16	*79.97	96.04	22.3	81.69	70.75	77.72	91.58
	BM25	28.3	*85.72	*76.32	81.94	95.24	**24.9**	***83.16**	***72.76**	79.72	97.82	22.9	81.72	70.74	**77.90**	93.75
FR	0-shot	40.1	86.40	76.84	83.34	88.89	36.5	84.36	73.98	81.78	87.20	34.5	83.87	73.20	80.68	89.64
	indep	41.3	86.39	77.03	83.37	89.89	36.5	84.28	73.97	81.77	87.53	34.8	84.03	*73.58	80.88	88.66
	rand.	41.2	86.48	*77.18	83.41	89.21	36.5	**84.40**	74.03	81.79	88.81	34.7	84.02	73.45	*81.17	89.91
	past	41.2	86.52	*77.32	83.43	87.65	**36.7**	84.32	74.08	**81.93**	88.23	35.0	83.97	*73.71	$^\delta$81.12	89.46
	COS	41.2	86.43	*77.16	83.43	92.14	**36.7**	84.23	73.89	**81.93**	92.53	35.1	84.07	73.52	$^\Delta$81.05	93.35
	PMI	**41.8**	***86.64**	***77.53**	**83.45**	90.87	**36.7**	84.27	73.99	81.84	**93.63**	**35.2**	84.07	*73.85	81.03	**94.71**
	BM25	41.5	86.43	*77.18	83.43	**92.43**	36.6	84.29	**74.12**	81.73	92.22	35.1	**84.17**	***73.91**	***81.35**	94.38
ZH	0-shot	30.1	84.42	73.31	81.28	75.52	28.3	83.21	70.62	79.17	75.88	29.2	83.24	71.56	78.63	79.02
	indep	30.5	*84.65	*73.76	**81.30**	78.91	**29.1**	***83.59**	***71.89**	***79.70**	76.40	29.8	*83.73	*72.53	*79.54	81.34
	rand.	30.6	*84.62	*73.65	81.22	83.08	26.6	83.10	70.87	79.14	78.70	29.9	*83.90	*72.49	*79.39	79.44
	past	30.7	84.60	*73.64	81.18	85.12	26.9	83.20	*71.33	79.30	82.89	30.1	***84.02**	***72.98**	***79.64**	87.26
	COS	30.7	84.49	*73.56	81.20	85.50	27.1	82.98	*71.04	78.96	81.92	30.0	*83.93	*72.69	*79.55	**89.94**
	PMI	**30.8**	***84.69**	***73.86**	81.23	**90.02**	27.3	$^\Delta$83.46	*71.86	79.35	**86.14**	**30.3**	*84.00	*72.90	*79.57	86.11
	BM25	30.2	84.21	73.23	80.15	88.79	26.8	83.12	*71.23	79.14	85.02	**30.3**	*83.87	*72.59	*79.22	88.37

Table 11: Evaluation for the translation of IWSLT, translated using source and automatic translation as context, for $K = 3$. We mark significantly positive difference between context-augmented methods and zero-shot MT for COMET-based scores, for sentences excluding the first 20 ones ($^\Delta$), all sentences ($^\delta$), or in both cases (*), with p-value < 0.05.

		EuroLLM					Llama					Qwen				
		BLEU	s-comet	d-comet	slide	LTCR	BLEU	s-comet	d-comet	slide	LTCR	BLEU	s-comet	d-comet	slide	LTCR
REF	0-shot	55.7	89.48	86.64	74.60	**92.85**	51.0	88.41	85.34	73.20	**92.59**	47.6	87.58	84.18	71.99	90.12
	rand.	58.7	*89.60	*86.78	*74.82	91.73	52.6	*88.65	*85.62	*73.39	92.09	51.8	*88.25	*85.01	*72.93	89.79
	past	59.8	*89.80	*87.02	*74.88	91.79	53.8	*88.89	*85.89	***73.52**	91.28	53.1	*88.37	*85.10	*72.91	**90.74**
	COS	60.8	*89.94	*87.22	*74.88	90.89	55.0	*89.03	*86.04	***73.52**	91.85	55.4	*88.68	*85.54	***73.33**	90.06
	PMI	**61.0**	***89.97**	***87.26**	***74.91**	91.54	**55.2**	***89.14**	***86.16**	*73.51	92.32	**55.8**	*88.80	*85.67	*73.26	90.68
	BM25	60.5	*89.91	*87.17	*74.90	91.21	54.9	*89.11	*86.09	*73.44	91.32	55.4	***88.82**	***85.70**	*73.24	89.98
MT	0-shot	55.7	89.48	86.64	74.60	**92.85**	51.0	88.41	85.34	73.20	92.59	47.6	87.58	84.18	71.99	90.12
	rand.	55.8	89.51	86.72	*74.75	91.45	51.5	*88.55	*85.52	73.31	92.87	48.4	*88.01	*84.74	*72.81	90.63
	past	55.0	*89.56	*86.84	*74.85	92.34	51.5	*88.63	***85.69**	***73.57**	92.89	48.3	*87.91	*84.74	*72.97	92.09
	COS	56.2	*89.59	*86.87	*74.82	92.64	51.5	*88.61	*85.61	*73.44	93.62	48.7	*88.04	*84.90	*73.06	92.62
	PMI	**56.3**	*89.61	***86.90**	***74.88**	92.77	**51.6**	*88.64	*85.67	*73.56	**93.96**	**49.0**	***88.09**	***85.01**	***73.16**	92.79
	BM25	56.2	***89.63**	***86.90**	*74.85	92.69	**51.6**	***88.66**	*85.66	*73.45	93.94	48.5	*88.02	*84.89	*73.06	**92.92**

Table 12: Evaluation for the translation of MERSENNE, translated using bilingual contexts with source and reference (REF, top) or source and automatic translation (MT, bottom), for $K = 3$. We mark significantly positive difference between context-augmented methods and zero-shot MT for COMET-based scores, for sentences excluding the first 20 ones ($^\Delta$), all sentences ($^\delta$), or in both cases (*), with p-value < 0.05.

<|im_start|>system
Translate the following English source text to Chinese, considering the provided English context and its Chinese translations:
English: that Natalia had dug out of the High Arctic belonged to ...
Chinese: Natalia 从加拿大北极高地挖出的东西属于……
English: So this camel would have been about nine feet tall, weighed around a ton.
Chinese: 因此，这只骆驼身高约 9 英尺，重约 1 吨。
English: But chances are the postcard image you have in your brain is one of these, the dromedary, quintessential desert creature -- hangs out in sandy, hot places like the Middle East and the Sahara, has a big old hump on its back for storing water for those long desert treks, has big, broad feet to help it tromp over sand dunes.
Chinese: 但是，很可能你脑海中浮现的明信片图像是其中一种，即单峰驼，这种动物是典型的沙漠生物，生活在中东和撒哈拉等炎热干燥的地方，背上有一大块驼峰，用来储存水分以应对漫长的沙漠探险，还有大而宽的脚，帮助它在沙丘上行走。<|im_end|>
<|im_start|>user
English: So how on earth would one of these guys end up in the High Arctic?
Chinese:<|im_end|>
<|im_start|>assistant

Figure 5: Selected contexts using PMI for the 56^{th} sentence from a IWSLT talk for EN–ZH translation. It included the 46^{th} sentence that contains the first mention of "the High Arctic".

lated ("And he found a match"), and also the contexts used in Past and PMI, with $K = 3$. The contexts selected using PMI, stating that the person can "compare it to those of known species" and "get a match", produce an adequate translation of "a match".

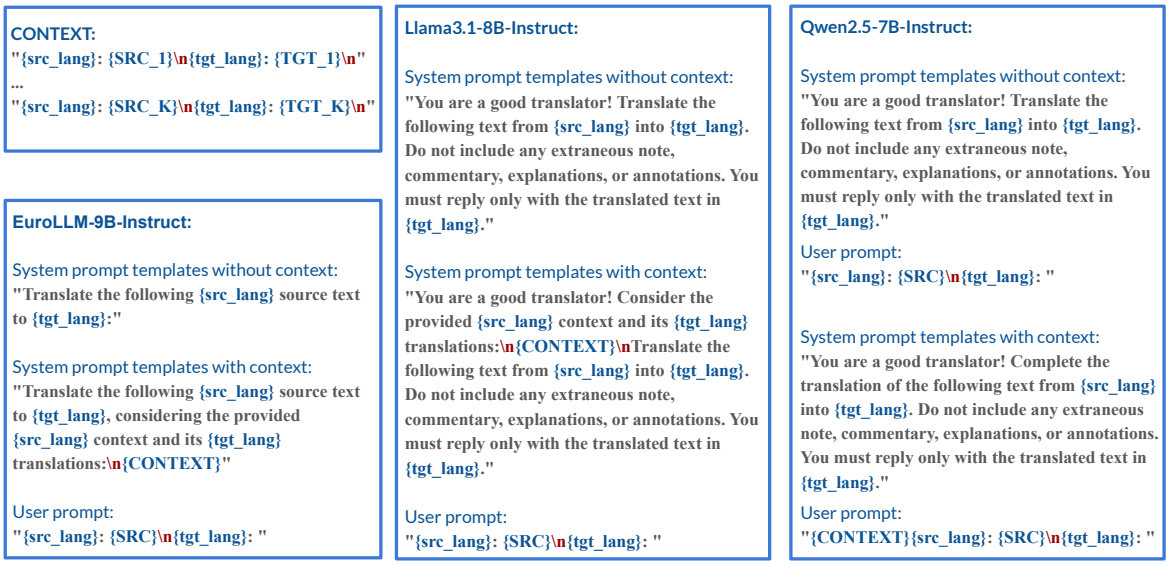

Figure 6: The prompt patterns for EuroLLM, Llama and Qwen applied in our experiments.

SRC: #1 And he found a match. #2 that Natalia had dug out of the High Arctic belonged to ... #3 a camel. **0-shot:** #1 他找到了合适的人。*(an appropriate person)* #2 纳塔莉亚从北极地区挖出来的东西属于…… #3 骆驼 **Past:** #1 他找到了匹配的物品。*(a matched object)* #2 纳塔莉亚从高北极地区挖出来的东西属于…… #3 一只骆驼。 **PMI**: #1 他找到了匹配的物种。*(a matched species)* #2 Natalia 从加拿大北极高地挖出的东西属于…… #3 一只骆驼。	**The contexts of Past for #1:** **English**: So she shipped him one of the fragments, FedEx. **Chinese**: 于是她用联邦快递把碎片寄给了他。 **English**: NR: Yeah, you want to track it. It's kind of important. **Chinese**: NR：是的，你想追踪它。这很重要。 **English**: LN: And he processed it, and compared it to 37 known and modern-day mammal species. **Chinese**: LN：然后他进行了处理，并将其与37种已知和现代哺乳动物进行了比较。	**The contexts of PMI for #1:** **English**: It turns out that different species have slightly different structures of collagen, so if you get a collagen profile of an unknown bone, you can compare it to those of known species, and, who knows, maybe you get a match. **Chinese**: 事实证明，不同物种的胶原蛋白结构略有不同，因此如果你得到一块未知骨头的胶原蛋白谱，你可以将其与已知物种的胶原蛋白谱进行比较，谁知道，也许能找到匹配。 **English**: So she shipped him one of the fragments, FedEx. **Chinese**: 于是她通过联邦快递把其中一块碎片寄给了他。 **English**: LN: And he processed it, and compared it to 37 known and modern-day mammal species. **Chinese**: LN：他对它进行了处理，并将其与 37 种已知和现代哺乳动物进行了比较。

Figure 7: EN–ZH translations of the 45^{th} sentence from an IWSLT talk, using EuroLLM with MT as target-side context. PMI retrieves relevant contexts in the 41^{th} sentence for "a match", which means a matched species.

<|im_start|>system
Translate the following English source text to French, considering the provided English context and its French translations:
English: In the memoir of May 22, 1848[4], it is the "tartrates" and the "paratartrates" which are primarily considered, but the young scientist seeks fruitful generalisations: "It will be said, and rightly so: All organic substances that deviate from the plane of polarisation when they are dissolved will therefore enjoy hemihedria.
French: Dans le mémoire du 22 mai 1848 [4], ce sont les « tartrates » et les « paratartrates » qui sont principalement considérés, mais le jeune scientifique cherche des généralisations fructueuses : « On dira, et à juste titre : toutes les substances organiques qui dévient du plan de polarisation lorsqu'elles sont dissoutes profiteront donc de l'hémiédrie.
English: It was even by studying this latter property that I was assured of the hemihedria, which I then realised via careful observation of the crystalline form.
French: C'est même en étudiant cette dernière propriété que j'ai été convaincu de l'hémiédrie, que j'ai ensuite confirmée par une observation attentive de la forme cristalline.
English: At this stage, Pasteur had moved away from the morphological study of the crystals to the study of the possible rotations of the plane of polarisation which they induced, which had led him to better characterise the hemihedria of tartrates.
French: À ce stade, Pasteur s'était éloigné de l'étude morphologique des cristaux pour étudier les rotations possibles du plan de polarisation qu'ils induisaient, ce qui l'avait conduit à mieux caractériser l'hémiédrie des tartrates.<|im_end|>
<|im_start|>user
English: The hemihedria is, moreover, non-superposable.
French:<|im_end|>
<|im_start|>assistant

Figure 8: Selected contexts for the 72^{nd} sentence of an MERSENNE article using PMI, including the 51^{st}, 56^{th} and 58^{th} sentences containing the term *hemihedria*.

<|im_start|>system
Translate the following English source text to French, considering the provided English context and its French translations:
English: His appointment to the University of Lille, in an industrial environment that led him to study amyl alcohols, helped to reorient his scientific activity, but he remained mainly driven by his hypothesis that “molecular dissymmetry” was the prerogative of the living.
French: Sa nomination à l'université de Lille, dans un environnement industriel qui l'a conduit à étudier les alcools amyles, a contribué à réorienter son activité scientifique, mais il est resté principalement guidé par son hypothèse selon laquelle la « dissymétrie moléculaire » était l'apanage du vivant.
English: In the memoir of May 22, 1848[4], it is the "tartrates" and the "paratartrates" which are primarily considered, but the young scientist seeks fruitful generalisations: "It will be said, and rightly so: All organic substances that deviate from the plane of polarisation when they are dissolved will therefore enjoy hemihedria.
French: Dans le mémoire du 22 mai 1848 [4], ce sont les « tartrates » et les « paratartrates » qui sont principalement considérés, mais le jeune scientifique cherche des généralisations fructueuses : « On dira, et à juste titre : toutes les substances organiques qui dévient du plan de polarisation lorsqu'elles sont dissoutes profiteront donc de l'hémiédrie.
English: At this stage, Pasteur had moved away from the morphological study of the crystals to the study of the possible rotations of the plane of polarisation which they induced, which had led him to better characterise the hemihedria of tartrates.
French: À ce stade, Pasteur s'était éloigné de l'étude morphologique des cristaux pour étudier les rotations possibles du plan de polarisation qu'ils induisaient, ce qui l'avait conduit à mieux caractériser l'hémiédrie des tartrates.<|im_end|>
<|im_start|>user
English: This preceded the second thesis, devoted to amyl alcohols, where Pasteur examined the crystallographic question thus posed, and where he writes to have not succeeded in inducing crystalline hemihedria.
French:<|im_end|>
<|im_start|>assistant

Figure 9: Selected contexts for the 159^{th} sentence of an MERSENNE article using PMI, including the 51^{st} and 58^{th} sentences containing the term *hemihedria*.