Bridging the Consistency Gap: A Noun Bank Augmented Framework for Low-Cost Long Text Translation

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have advanced machine translation, yet long-text translation still faces a practical tension between document-level lexical consistency and inference cost. Recent document-level methods improve coherence by expanding context, maintaining multi-level memories, or injecting multi-granularity knowledge, but such broad context mechanisms often introduce substantial prompt redundancy and limit parallel inference. We propose **NoBAT** (**No**un **B**ank **A**ugmented **T**ranslation), a lightweight framework that distills document-level lexical constraints into a compact global Noun Bank. NoBAT first constructs and disambiguates the Noun Bank serially, and then retrieves only relevant noun constraints for local, parallel translation. Experiments on IWSLT2017 and Novels-XL with multiple LLM backbones show that NoBAT consistently improves lexical translation consistency while maintaining competitive COMET scores. Compared with DelTA, NoBAT further reduces total token consumption by **6.3**×, demonstrating a cost-effective trade-off among consistency, translation quality, and efficiency.

1 Introduction

Large Language Models (LLMs) have brought substantial progress to machine translation (MT), especially in sentence-level and short-context settings (Garcia and Firat, 2022; Hendy et al., 2023; Jiao et al., 2023; Zhang et al., 2023; Cheng et al., 2025; Zheng et al., 2025b,a; Google Translate Research Team et al., 2026; Zheng et al., 2026). However, long-form translation remains challenging because documents require discourse-level consistency across distant contexts, including lexical cohesion, entity identity, topic continuity, and local coherence (Bawden et al., 2018; Wang et al., 2024; Zhao et al., 2025). Recurring names, locations, and domain terms are particularly important lexical anchors: once translated inconsistently, they can noticeably degrade readability even when individual sentences are fluent.

Recent LLM-based document-level MT studies inject broader context through prompting, in-context learning, source-primed multi-turn conversation, multilingual contextualization, and document-level LLM adaptation (Wang et al., 2023; Wu and Hu, 2023; Wu et al., 2024; Choudhary et al., 2025; Cui et al., 2024; Hu et al., 2025; Ramos et al., 2025). Yet simply expanding the context window is not a complete solution: longer prompts increase token consumption and latency, may introduce noisy context, and can still suffer from omissions or hallucinations when translating large chunks (Karpinska and Iyyer, 2023; Dale et al., 2023). Thus, practical long-text MT systems must balance translation quality, consistency, prompt efficiency, and parallelizability.

Memory- and knowledge-augmented approaches are more targeted but still costly. DelTA (Wang et al., 2025) maintains Proper Noun Records, Bilingual Summary, Long-Term Memory, and Short-Term Memory during online sentence-by-sentence translation, improving consistency but making decoding inherently serial and repeatedly serializing heterogeneous memory states into prompts. Sub-DocTrans (Hong et al., 2025) segments documents into coherent sub-documents and injects bilingual summaries, themes, proper nouns, topics, and transition hints, which improves coherence but introduces non-trivial prompt and inference overhead.

These studies reveal a cost–consistency mismatch in current document-level MT: while document-level context and memory are useful, many measurable consistency errors concentrate on sparse lexical anchors such as recurring nouns, entities, and terminology. Maintaining broad, frequently updated document states is often unnecessary when a translation unit only needs a small subset of lexical constraints. This motivates a lighter mechanism that constructs compact global lexical

constraints once and retrieves only relevant entries locally.

To address this gap, we propose **NoBAT** (**No**un **B**ank **A**ugmented **T**ranslation), a lightweight framework for low-cost long-text translation. NoBAT first builds and disambiguates a global Noun Bank from coarse document segments, then translates finer-grained units with limited local context and only the noun constraints retrieved from the finalized bank. Because the Noun Bank remains static during the second stage, translation units can be decoded independently, enabling parallel inference while avoiding repeated multi-level memory updates.

We evaluate NoBAT on IWSLT2017 (Akiba et al., 2004) and Novels-XL (Wang et al., 2024) with several models. Results show that NoBAT consistently improves lexical translation consistency ratios (LTCR-1 and LTCR-1_f) (Wang et al., 2025) while maintaining competitive sentence- (Rei et al., 2020) and document-level (Vernikos et al., 2022) COMET scores. Compared with DelTA, NoBAT also reduces total token consumption by **6.3**×, highlighting its practical advantage for cost-sensitive long-text translation.

The primary contributions of this paper are summarized as follows:

- We propose **NoBAT**, a lightweight noun-bank-augmented framework that reframes long-text translation consistency as targeted noun-level constraint modeling rather than broad context expansion or frequently updated multi-level memory maintenance.
- We design a two-stage serial-to-parallel translation mechanism: NoBAT first constructs and disambiguates a compact global Noun Bank, and then retrieves only locally relevant noun constraints for parallel final-stage translation.
- Experiments on IWSLT2017 and Novels-XL with multiple LLM backbones show that NoBAT improves lexical translation consistency while maintaining competitive COMET scores and reducing token consumption by 6.3× compared with DelTA.

2 Related Work

Document-level Machine Translation. Document-level machine translation (DocMT) extends translation beyond isolated sentences by using cross-sentence context and discourse information to resolve ambiguity, preserve lexical cohesion, and improve coherence. Recent corpus, benchmark, and training studies further show that document-level context is essential for evaluating and improving long-form translation quality (Pal et al., 2024; Ma et al., 2025; Wang et al., 2024; Zhao et al., 2025). For LLM-based DocMT, existing methods commonly rely on Doc2Doc generation or sentence/segment-level decoding with additional context. Context prompting, in-context learning, source-primed multi-turn conversation, and multilingual contextualization have all been explored to improve discourse-aware translation (Choudhary et al., 2025; Cui et al., 2024; Hu et al., 2025; Ramos et al., 2025). As discussed by Hong et al. (2025), Doc2Doc methods can better exploit global context but may introduce omissions or hallucinations, while sentence-level decoding improves coverage but often weakens discourse phenomena such as pronoun translation and inter-sentential coherence.

Memory Mechanisms for Long-context Translation. Memory mechanisms are widely used to help LLMs retain and reuse information beyond short local context. General LLM memory studies store long-term user or task information, train agents to manage memory, or reshape long-context processing through explicit memory operations (Zhong et al., 2023; Yu et al., 2025; Yan et al., 2025). In document-level MT, memory is typically used to cache document signals needed for coherent sentence-by-sentence translation. DelTA (Wang et al., 2025) continuously retrieves and updates Proper Noun Records, Bilingual Summary, Long-Term Memory, and Short-Term Memory, improving consistency but making decoding inherently serial and repeatedly serializing heterogeneous memory states into prompts. Doc-Guided Sent2Sent++ similarly introduces doc-guided memory to strengthen sentence-level translation with document information (Guo et al., 2025b), while SubDocTrans (Hong et al., 2025) injects several types of global and local knowledge for each sub-document. Overall, these studies confirm the importance of discourse context and memory for DocMT, yet their broad memory designs can introduce non-trivial token cost and latency. This motivates a more focused memory mechanism that stores only the consistency-critical information, such as document-level noun translations.

3 Method

3.1 Problem Formulation and Overview

Given a source document consisting of a sequence of sentences $D_s = \{s_1, s_2, \ldots, s_N\}$, the goal of document-level machine translation is to generate a target document $D_t = \{t_1, t_2, \ldots, t_N\}$ that maximizes not only sentence-level fidelity but also discourse-level coherence. Existing memory-augmented agents (e.g., DELTA (Wang et al., 2025)) tackle this by maintaining a set of dense, heterogeneous memories $\mathcal{M} = \{M_{\text{short}}, M_{\text{long}}, M_{\text{bilingualsummary}}, M_{\text{noun}}\}$, updating them autoregressively:

$$t_i, \mathcal{M}_i = f_{\text{LLM}}(s_i, \mathcal{M}_{i-1}) \tag{1}$$

While effective, this autoregressive memory dependency enforces a strictly serial decoding paradigm and incurs $\mathcal{O}(N)$ sequential prompt expansions, leading to severe latency and token redundancy.

To circumvent this, NoBAT introduces the two-panel pipeline in Figure 1. The blue upper panel converts the source document into a high-quality global Noun Bank through a serial segment-level pass: each segment is translated as a draft, relevant entries are retrieved from the accumulating bank, and the resulting noun evidence is merged and disambiguated with historical bilingual contexts. The green lower panel then fixes this final bank and uses it as a lightweight lexical constraint store for sentence-level parallel translation. The overall procedure is summarized in Algorithm 1, and the prompt templates for Stage 1, Stage 2, and noun disambiguation are provided in Appendix A.2. We formalize the pipeline as a two-step mapping: (1) serial construction and disambiguation of a global Noun Bank $\mathcal{B}$, and (2) non-autoregressive, retrieval-augmented decoding conditioned only on local source context and relevant noun constraints.

3.2 Stage 1: Serial Noun Bank Construction

As shown in the upper panel of Figure 1, Stage 1 is responsible for producing the final Noun Bank before sentence-level decoding begins. Given the input multi-sentence text $D_s = \{s_1, s_2, \ldots, s_N\}$, we first merge consecutive sentences under a token budget τ into K coarse segments $\mathcal{S} = \{S_1, S_2, \ldots, S_K\}$:

$$\mathcal{S} = \text{Merge}(D_s, \tau), \quad |S_k| \leq \tau. \tag{2}$$

Each segment S_k follows the "LLM–draft–noun evidence" flow in the figure: it is translated once to produce a segment-level draft T_k and a set of noun entries $E_k = \{(x_j, y_j)\}_{j=1}^{|E_k|}$, where x_j is a source proper noun or named entity and y_j is its target translation. The draft is not used as the final output; instead, it provides bilingual evidence for constructing and validating noun translations.

For the first segment, no previous bilingual context or bank entry is available, so the LLM directly reads S_1 and generates both the draft translation and noun entries:

$$T_1, E_1 = \mathcal{M}_\theta(S_1 \mid \mathcal{P}_{\text{stage1}}). \tag{3}$$

The initial bank and history are obtained by merging E_1 with the empty state:

$$\begin{aligned} \mathcal{B}_1, \mathcal{H} &= \text{Merge}_{\text{disambig}}(\mathcal{B}_0, \mathcal{H}, E_1, S_1, T_1), \\ \mathcal{B}_0 &= \emptyset. \end{aligned} \tag{4}$$

For every subsequent segment S_k ($k > 1$), the LLM follows the "Context–Retrieve–LLM–Draft" path in Figure 1. It is conditioned on the current segment, the bilingual previous segment (S_{k-1}, T_{k-1}), and noun entries retrieved from the accumulating bank:

$$R_k = \text{Retrieve}(\mathcal{B}_{k-1}, S_k), \tag{5}$$

$$T_k, E_k = \mathcal{M}_\theta(S_k \mid S_{k-1}, T_{k-1}, R_k, \mathcal{P}_{\text{stage1}}). \tag{6}$$

Here, the bilingual previous segment preserves local discourse continuity, while R_k enforces noun translations that have already been established. After each call, the extracted entries and the segment-level draft are sent to the accumulating bank and disambiguation module:

$$\mathcal{B}_k, \mathcal{H} = \text{Merge}_{\text{disambig}}(\mathcal{B}_{k-1}, \mathcal{H}, E_k, S_k, T_k). \tag{7}$$

The "History" component in the figure corresponds to a context history $\mathcal{H}_x$ maintained for each source proper noun x. Whenever an entry (x, y) is extracted from segment S_k, the bilingual evidence (S_k, T_k) is appended to $\mathcal{H}_x$:

$$\mathcal{H}_x \leftarrow \mathcal{H}_x \cup \{(S_k, T_k)\}. \tag{8}$$

If x is not yet in the bank, NoBAT directly inserts (x, y). If x already exists with the same translation, only its context history is updated. When a conflict occurs, i.e., the existing translation $y_{\text{old}} = \mathcal{B}_{k-1}[x]$ differs from the newly extracted translation y_{new}, NoBAT invokes the merge-and-disambiguate step shown in Figure 1. The disambiguation prompt

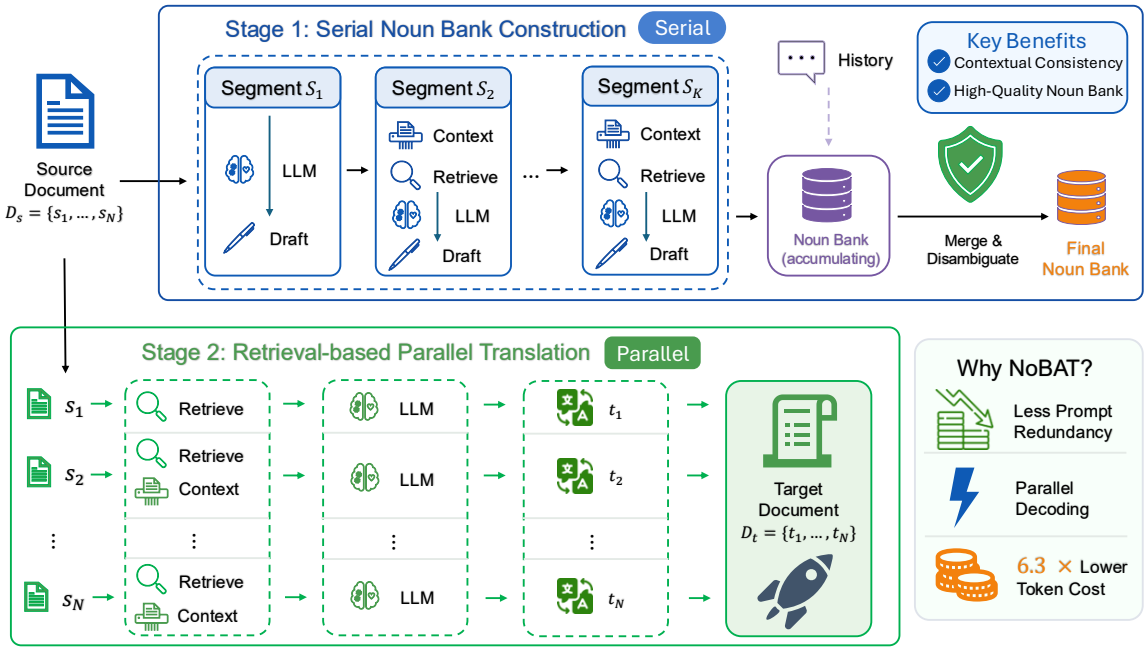

Figure 1: Overview of NoBAT. In Stage 1, merged segments S_k are processed serially: the current segment, bilingual previous context (S_{k-1}, T_{k-1}), and retrieved noun entries are fed to the LLM to produce a draft translation and noun evidence, which are accumulated, merged, and disambiguated into a final Noun Bank. In Stage 2, each sentence s_n independently retrieves relevant entries from the fixed bank and is translated with its previous-j-sentence source context C_n, enabling parallel decoding with reduced prompt redundancy.

$\mathcal{P}_{\text{disambig}}$ provides the source proper noun, the two candidate translations, and selected bilingual contexts from $\mathcal{H}_x$:

$$y^* = \mathcal{M}_\theta(x, y_{\text{old}}, y_{\text{new}}, \text{Select}(\mathcal{H}_x) \mid \mathcal{P}_{\text{disambig}}). \tag{9}$$

The selection function keeps the first occurrence and the most recent h occurrences of the noun, so the decision considers both the initial naming convention and the local evidence around the conflict. The LLM is instructed to output only one chosen translation, and the bank is updated as $\mathcal{B}_k[x] = y^*$. This procedure turns noisy draft-level noun candidates into a high-quality final Noun Bank, preventing later segments from introducing inconsistent variants while allowing the canonical translation to be corrected when stronger document-level evidence appears. After the serial pass, the finalized bank is $\mathcal{B} = \mathcal{B}_K$.

3.3 Stage 2: Retrieval-based Parallel Translation

The lower panel of Figure 1 shows how the finalized Noun Bank is used after Stage 1. NoBAT returns to the original sentence sequence $D_s = \{s_1, \dots, s_N\}$ and translates each sentence with a compact prompt consisting of the current sentence, a short source-side context, and relevant noun entries from the fixed bank. Because the bank is no longer updated, no sentence depends on the translation output of another sentence, enabling the parallel decoding path shown in the figure. For each source sentence s_n, we define its local source-side context as the previous j sentences:

$$C_n = \{s_i \mid \max(1, n-j) \le i < n\}. \tag{10}$$

Thus, C_n is empty for the first sentence, contains only available preceding sentences near the beginning of the document, and contains exactly j preceding sentences for $n > j$.

For each sentence, the retrieval box in Figure 1 selects only the noun constraints that occur in the current sentence:

$$R_n = \text{Retrieve}(\mathcal{B}, s_n) = \{(x, y) \in \mathcal{B} \mid x \in s_n\}. \tag{11}$$

The final sentence-level translation is then generated as:

$$t_n = \mathcal{M}_\theta(s_n \mid C_n, R_n, \mathcal{P}_{\text{stage2}}). \tag{12}$$

Because C_n contains only source-side preceding sentences and $\mathcal{B}$ is static, generating t_n does not depend on any other target-side output t_m. Therefore

Algorithm 1: The Overall Framework of NoBAT

Input: Source multi-sentence text $D_s = \{s_1, \ldots, s_N\}$, LLM $\mathcal{M}_\theta$, segment budget τ, context size j, history size h
Output: Target multi-sentence text $D_t = \{t_1, \ldots, t_N\}$

/* Stage 1: Serial Noun Bank Construction */
$\mathcal{S} \leftarrow \mathrm{Merge}(D_s, \tau)$ */* Merge sentences into $S_1, \ldots, S_K$ */*
$\mathcal{B}_0 \leftarrow \emptyset, \mathcal{H} \leftarrow \emptyset$ */* Initialize bank and context history */*
for $k = 1$ **to** K **do**
  if $k = 1$ **then**
    $T_k, E_k \leftarrow \mathcal{M}_\theta(S_k \mid \mathcal{P}_{\text{stage1}})$ */* No previous context */*
  else
    $R_k \leftarrow \mathrm{Retrieve}(\mathcal{B}_{k-1}, S_k)$ */* Retrieve noun constraints */*
    $T_k, E_k \leftarrow \mathcal{M}_\theta(S_k \mid S_{k-1}, T_{k-1}, R_k, \mathcal{P}_{\text{stage1}})$
  end if
  $\mathcal{B}_k, \mathcal{H} \leftarrow \mathrm{Merge}_{\text{disambig}}(\mathcal{B}_{k-1}, \mathcal{H}, E_k, S_k, T_k)$ */* Update histories and resolve conflicts */*
end for
$\mathcal{B} \leftarrow \mathcal{B}_K$ */* Final Noun Bank */*

/* Stage 2: Retrieval-based Parallel Translation */
for $n = 1$ **to** N **in parallel do**
  $C_n \leftarrow \{s_i \mid \max(1, n-j) \leq i < n\}$ */* Previous j source sentences */*
  $R_n \leftarrow \mathrm{Retrieve}(\mathcal{B}, s_n)$ */* Retrieve entries in s_n */*
  $t_n \leftarrow \mathcal{M}_\theta(s_n \mid C_n, R_n, \mathcal{P}_{\text{stage2}})$ */* Sentence-level translation */*
end for
$D_t \leftarrow \bigoplus_{n=1}^{N} t_n$ */* Concatenate in sentence order */*

all sentence translations can be decoded in parallel:

$$D_t = \bigoplus_{n=1}^{N} t_n. \tag{13}$$

where $\oplus$ denotes concatenation in the original sentence order. This serial-to-parallel design explains the key benefits highlighted in Figure 1: Stage 1 concentrates document-level consistency modeling into a one-time noun-bank construction pass, while Stage 2 avoids repeatedly prompting with broad memories and retrieves only sentence-relevant constraints, reducing prompt redundancy and token cost.

4 Experiments

4.1 Settings

Datasets. Following the experimental protocols of DelTA (Wang et al., 2025) and SubDocTrans (Hong et al., 2025), we evaluate NoBAT on two document-level translation benchmarks: IWSLT2017 (Akiba et al., 2004) and Novels-XL (GuoFeng) (Wang et al., 2024). For IWSLT2017, we use the tst2017 TED-talk test documents and consider eight bidirectional directions: En⇔Zh, En⇔De, En⇔Fr, and En⇔Ja. The sentence-aligned document structure of IWSLT2017 allows us to evaluate both sentence-level quality and document-level consistency. For Novels-XL, we use the GuoFeng Chinese–English web-fiction TEST_2 set, which contains long-form narrative documents and is therefore suitable for evaluating lexical consistency in ultra-long text translation.

Baseline Method. We compare NoBAT with representative sentence-, context-, document-, and memory-based translation strategies. **Sentence** translates each input sentence independently without document context. **Context** (Wu et al., 2024) augments the current sentence with the previous three source-target sentence pairs, following common local-context prompting practice. **Doc2Doc** (Wang et al., 2023) translates a block of 10 consecutive sentences in one generation turn, representing a direct document-context prompting baseline. **DelTA** (Wang et al., 2025) translates sentence by sentence with multi-level memories, including proper noun records, bilingual summaries, long-term memory, and short-term memory. We also consider **SubDocTrans** (Hong et al., 2025), which injects multi-granularity knowledge such as summaries, themes, proper nouns, topics, and transition hints. Since SubDocTrans is not publicly released, we report it only where comparable scores are available and omit it for newly evaluated backbone models. Furthermore, we also include results from `NLLB-3.3B` (Costa-jussà et al., 2024) and `GoogleTrans`[1] for comparison.

Backbone Models and Inference. We conduct experiments with four LLM backbones: `GPT-5.4`[2], `GPT-4o-mini`[3], `Qwen3.6-35B-A3B`[4] (Qwen Team, 2026), and `DeepSeek-R1-Distill-Qwen-32B`[5] (Guo et al., 2025a). For a fair comparison, all comparable systems use the same backbone model and decoding configuration within each experiment. Among them, only `DeepSeek-R1-Distill-Qwen-32B` is used with thinking enabled, while `GPT-5.4`, `GPT-4o-mini`, and `Qwen3.6-35B-A3B` are run in their fast-response mode. Unless otherwise specified, all models use the providers' recommended default parameters.

In NoBAT, the first stage merges input sentences into text segments under a token budget $\tau = 200$ and serially constructs the global Noun Bank with

[1] https://py-googletrans.readthedocs.io/en/latest/
[2] https://developers.openai.com/api/docs/models/gpt-5.4
[3] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/
[4] https://huggingface.co/Qwen/Qwen3.6-35B-A3B
[5] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B

System	En->Xx				Xx->En			
	sCOMET	dCOMET	LTCR-1	LTCR-1_f	sCOMET	dCOMET	LTCR-1	LTCR-1_f
NLLB	82.11	6.36	74.56	81.87	84.10	6.98	79.03	90.76
GOOGLE	80.41	5.83	81.38	84.72	80.17	5.96	81.43	90.81
`GPT-5.4`								
Sentence	85.75	6.84	86.88	90.70	85.51	7.14	85.33	95.62
Context	86.13	6.85	86.61	92.04	85.97	7.15	87.95	96.49
Doc2Doc	–	6.92	86.26	91.50	–	7.20	84.95	94.70
DELTA	86.15	**6.97**	86.52	91.48	**86.14**	**7.26**	88.82	**97.93**
SubDocTrans	–	–	–	–	–	–	–	–
NoBAT	**86.45**	6.96	**88.04**	**92.23**	85.73	7.24	**90.12**	97.91
`Qwen3.6-35B-A3B`								
Sentence	85.62	6.87	81.88	88.28	85.20	7.13	83.87	91.86
Context	85.80	6.84	84.23	91.06	**85.55**	7.17	83.33	92.43
Doc2Doc	–	7.03	80.25	87.89	–	**7.32**	83.68	93.95
DELTA	85.32	**7.04**	86.44	91.29	85.07	7.19	86.71	**96.22**
SubDocTrans	–	–	–	–	–	–	–	–
NoBAT	**85.94**	6.94	**86.61**	**91.56**	85.34	7.23	**88.45**	96.10
`GPT-4o-mini`								
Sentence	81.51	6.35	78.59	85.07	84.01	6.99	81.42	91.34
Context	84.78	6.65	80.01	86.99	84.95	7.15	84.40	94.34
Doc2Doc	–	6.75	80.54	85.39	–	7.01	83.50	93.39
DELTA	**85.85**	**6.80**	81.80	86.33	85.26	**7.24**	85.25	95.89
SubDocTrans	85.79	6.76	83.14	87.43	**85.27**	**7.24**	86.33	96.32
NoBAT	85.49	6.73	**87.12**	**93.66**	85.09	7.19	**88.05**	**97.22**
`DeepSeek-R1-Distill-Qwen-32B`								
Sentence	83.16	6.42	74.17	82.75	83.47	6.99	77.71	86.98
Context	82.89	6.45	79.10	85.77	83.65	7.07	83.61	92.37
Doc2Doc	–	6.70	78.53	85.46	–	7.25	82.88	92.24
DELTA	83.29	6.51	81.85	86.76	83.39	7.07	84.85	93.54
SubDocTrans	**84.50**	**6.71**	80.78	88.44	**84.88**	**7.28**	86.05	**95.86**
NoBAT	83.66	6.42	**89.14**	**93.15**	84.50	7.07	**87.72**	95.39

Table 1: Main results on IWSLT2017. Scores are averaged over En⇒Xx and Xx⇒En directions, where Xx denotes Zh, De, Fr, and Ja. Doc2Doc does not produce sentence-aligned outputs, so sCOMET is not reported. The best performance in each block is highlighted in **bold**, while the second-best is underlined. Part of the data is sourced from Wang et al. (2025) and Hong et al. (2025).

the most recent $h = 3$ occurrences of the noun for disambiguation. The second stage translates each sentence in parallel by conditioning on the current sentence, the previous $j = 3$ source sentences as local context, and noun constraints retrieved from the finalized Noun Bank.

Evaluation Metrics. We evaluate translation quality with sentence-level COMET (sCOMET) with the model `Unbabel/wmt22-comet-da`[6] and document-level COMET (dCOMET) with the model `wmt21-comet-qe-mqm`[7]. The sCOMET measures sentence-level adequacy and fluency for aligned outputs, while dCOMET evaluates document-level translation quality. To measure lexical consistency, we adopt LTCR-1 and LTCR-1_f following DELTA (Wang et al., 2025). LTCR-1 checks whether repeated entity translations remain identical to their first occurrence, whereas LTCR-1_f is a relaxed fuzzy variant that treats containment-based variants as consistent, e.g., a full name and its abbreviated form.

[6] https://github.com/Unbabel/COMET/

[7] https://huggingface.co/NataliaKhaidanova/wmt21-comet-qe-mqm

4.2 Main Results

Test Results on IWSLT2017. Table 1 reports the averaged results on IWSLT2017 over four En⇒Xx and four Xx⇒En directions, following prior work (Wang et al., 2025; Hong et al., 2025). Detailed per-language results are provided in Appendix A.1,

System	sCOMET	dCOMET	LTCR-1	LTCR-1_f	sCOMET	dCOMET	LTCR-1	LTCR-1_f
		GPT-5.4				Qwen3.6-35B-A3B		
Sentence	77.91	3.32	71.35	77.66	77.96	3.41	71.35	83.49
Context	78.52	3.27	74.87	86.20	78.93	3.31	80.67	86.95
Doc2Doc	–	**3.36**	80.19	86.13	–	3.42	72.19	90.24
DELTA	78.68	3.35	86.89	95.63	78.25	**3.43**	**87.92**	93.72
SubDocTrans	–	–	–	–	–	–	–	–
NoBAT	**79.38**	**3.36**	**88.23**	**96.07**	**79.10**	3.39	87.43	**95.65**
		GPT-4o-mini				DeepSeek-R1-Distill-Qwen-32B		
Sentence	77.87	3.10	58.82	70.59	75.28	3.10	57.77	63.11
Context	78.56	3.19	64.32	74.37	75.76	3.17	62.32	74.88
Doc2Doc	–	2.96	82.04	91.62	–	3.26	60.49	81.46
DELTA	78.77	3.34	**88.94**	96.48	76.55	3.22	84.54	93.72
SubDocTrans	**78.89**	**3.35**	81.55	94.66	**78.06**	**3.42**	**89.05**	**96.52**
NoBAT	78.52	3.29	85.57	**96.63**	77.85	3.21	83.73	95.69

Table 2: Main results on Novels-XL (GuoFeng) in the Zh⇒En direction. Doc2Doc does not produce sentence-aligned outputs, so sCOMET is not reported. The best performance in each block is highlighted in **bold**, while the second-best is underlined. Part of the data is sourced from Wang et al. (2025) and Hong et al. (2025).

Setting	sCOMET	dCOMET	LTCR-1	LTCR-1_f
Full	78.52	**3.29**	**85.57**	**96.63**
w/o Stage 1	**78.59**	3.28	62.80	76.81
w/o Stage 2	78.43	3.20	79.32	93.26
w/o Context (S1)	78.48	3.22	82.69	96.15
w/o Context (S2)	78.15	3.19	82.29	94.25
w/o Disambiguation	78.56	3.25	83.33	95.23

Table 3: Ablation results on Novels-XL with GPT-4o-mini. S1 and S2 denote the Noun Bank construction stage and the retrieval-based translation stage, respectively. The best performance in each block is highlighted in **bold**, while the second-best is underlined.

where Tables 4 and 5 break down the results for Zh, De, Fr, and Ja separately. The overall trend is that NoBAT brings the most consistent improvements to lexical consistency while preserving competitive translation quality.

Across all four backbone models, NoBAT consistently achieves the best LTCR-1 scores in both translation directions, showing that the Noun Bank provides robust document-level lexical constraints. Compared with DELTA, NoBAT improves LTCR-1 by 1.52, 0.17, 5.32, and 7.29 points for En⇒Xx under GPT-5.4, Qwen3.6-35B-A3B, GPT-4o-mini, and DeepSeek-R1-Distill-Qwen-32B, respectively. In the reverse Xx⇒En direction, the corresponding gains over DELTA are 1.30, 1.74, 2.80, and 2.87 points. For En⇒Xx, NoBAT also obtains the best LTCR-1_f under every backbone, with particularly large gains on GPT-4o-mini and DeepSeek-R1-Distill-Qwen-32B. For Xx⇒En, NoBAT ranks first or second on LTCR-1_f in all model blocks, while DELTA or SubDocTrans sometimes slightly leads on the fuzzy metric. In terms of COMET-based translation quality, NoBAT remains competitive with context- and memory-augmented baselines. It achieves the best sCOMET for GPT-5.4 and Qwen3.6-35B-A3B in En⇒Xx, and its dCOMET scores are generally close to the best-performing systems. The only clear quality trade-offs appear for smaller or weaker backbones, where SubDocTrans or DELTA sometimes attains slightly higher COMET scores; nevertheless, NoBAT still substantially improves LTCR metrics, showing a favorable balance between translation adequacy and lexical consistency.

Test Results on Novels-XL (GuoFeng). Table 2 summarizes the Zh⇒En results on Novels-XL (GuoFeng), a long-form web-fiction benchmark with denser entity recurrence. NoBAT performs especially well on GPT-5.4, achieving the best or tied-best scores on all four metrics and improving DELTA from 86.89 to 88.23 on LTCR-1 and from 95.63 to 96.07 on LTCR-1_f.

With `Qwen3.6-35B-A3B`, NoBAT obtains the best sCOMET and LTCR-1_f, and is second only to DELTA on LTCR-1. With `GPT-4o-mini`, NoBAT reaches the best LTCR-1_f score of 96.63 and the second-best LTCR-1, although DELTA and SubDocTrans retain advantages on COMET-based quality metrics. For `DeepSeek-R1-Distill-Qwen-32B`, SubDocTrans gives the strongest overall scores, but NoBAT still ranks second on sCOMET and LTCR-1_f, indicating that noun-level memory remains effective even when broader knowledge augmentation is competitive.

Overall, the revised results show a clear trade-off between broad context injection and focused lexical memory. The Context baseline usually improves over Sentence, but its gains are limited by the short previous-three-sentence window and are not sufficient for long-range entity recurrence. Doc2Doc can improve dCOMET in some blocks, yet it loses sentence alignment and does not reliably improve LTCR scores. Compared with these approaches, NoBAT decouples global lexical memory from local generation: it first normalizes recurring nouns into a compact document-level bank and then retrieves only relevant entries during translation. This design yields more stable consistency improvements while keeping translation quality close to strong document-level baselines.

4.3 Ablation Study

We conduct ablation experiments on Novels-XL with `GPT-4o-mini` to isolate the contribution of each component. As shown in Table 3, removing Stage 1 causes the largest degradation in consistency: LTCR-1 drops from 85.57 to 62.80 and LTCR-1_f drops from 96.63 to 76.81. This confirms that serial Noun Bank construction is the key source of global lexical consistency. Removing Stage 2 also weakens performance, especially on LTCR-1, indicating that the retrieval-based second pass is necessary to apply the global constraints to final sentence-level translation.

The context ablations show that local context complements but cannot replace the Noun Bank. Removing context from Stage 1 slightly reduces consistency, while removing context from Stage 2 leads to larger drops in both LTCR-1 and LTCR-1_f, suggesting that local discourse context is more directly useful when generating final translations. Finally, removing disambiguation decreases LTCR-1 by 2.24 points and LTCR-1_f by 1.40 points, demonstrating that resolving entity conflicts during bank merging helps maintain canonical translations across long documents. A qualitative Novels-XL example in Table 6 of Appendix A.3 further illustrates how this disambiguation step normalizes conflicting entity translations for later retrieval.

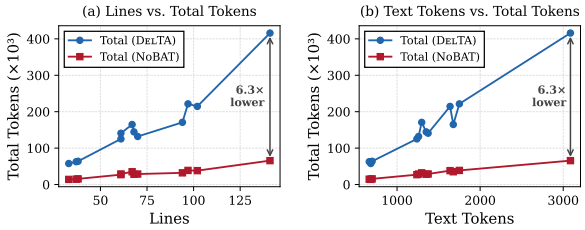

Figure 2: Token consumption comparison between DELTA and NoBAT. NoBAT reduces total token usage by replacing repeated multi-level memory prompting with compact Noun Bank construction and local retrieval.

4.4 Token Consumption Comparison

Beyond translation quality and consistency, we further compare the token consumption of NoBAT and DELTA. As shown in Figure 2, DELTA's total token usage grows rapidly as document length increases, because its online translation process repeatedly retrieves and updates multi-level memories. In contrast, NoBAT keeps the memory compact by storing only noun-level constraints and retrieves only entries relevant to each translation unit. This design substantially reduces redundant prompt tokens and achieves about a 6.3× lower total token cost in the compared long-text settings.

5 Conclusion

This paper introduced NoBAT, a lightweight Noun Bank Augmented framework for low-cost long text translation that decouples document-level lexical constraint construction from final translation. By serially building and disambiguating a compact global Noun Bank and then retrieving only sentence-relevant noun constraints for parallel decoding, NoBAT preserves document-wide lexical consistency while reducing prompt redundancy. Experiments on IWSLT2017 and Novels-XL show that NoBAT consistently improves LTCR-1 and LTCR-1_f across multiple LLM backbones, keeps COMET scores competitive with strong document-level baselines, and reduces total token consumption by about 6.3× compared with DELTA, offering a practical trade-off among consistency, translation quality, parallelism, and inference cost.

Limitations

Despite its effectiveness, NoBAT has several limitations. It focuses on noun-level lexical consistency and does not explicitly model broader discourse phenomena such as pronoun resolution, ellipsis, style, or paragraph-level coherence. It also depends on LLM-based noun extraction and disambiguation, so missed or incorrect entries may affect later retrieval. In addition, Stage 1 remains serial and may become a bottleneck for extremely long documents. Finally, our evaluation covers limited datasets, language directions, and backbones; future work will explore richer discourse memories and broader deployment settings.

Ethics Statement

This work builds upon publicly available datasets that are widely used in the community. These datasets do not contain personally identifying information or offensive content. No new user-generated or sensitive data was collected or processed in this study. Our use of these datasets complies with their original licenses and intended use.

References

Yasuhiro Akiba, Marcello Federico, Noriko Kando, Hiromi Nakaiwa, Michael Paul, and Jun'ichi Tsujii. 2004. Overview of the IWSLT evaluation campaign. In *Proceedings of the First International Workshop on Spoken Language Translation: Evaluation Campaign*, Kyoto, Japan.

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.

Shanbo Cheng, Yu Bao, Qian Cao, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, Wenhao Zhu, Jingwen Chen, Zhichao Huang, Tao Li, Yifu Li, Huiying Lin, Sitong Liu, Ningxin Peng, Shuaijie She, Lu Xu, Nuo Xu, Sen Yang, and 7 others. 2025. Seed-x: Building strong multilingual translation llm with 7b parameters. *Preprint*, arXiv:2507.13618.

Ritvik Choudhary, Rem Hida, Masaki Hamada, Hayato Futami, and Toshiyuki Sekiya. 2025. Exploring context strategies in LLMs for discourse-aware machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24382–24391.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

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.

David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loïc Barrault, and Marta R. Costa-jussà. 2023. Halomi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation. *Preprint*, arXiv:2305.11746.

Xavier Garcia and Orhan Firat. 2022. Using natural language prompts for machine translation. *Preprint*, arXiv:2202.11822.

Google Translate Research Team, Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, Markus Freitag, and David Vilar. 2026. TranslateGemma Technical Report.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025a. DeepSeek-r1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638.

Jiaxin Guo, Yuanchang Luo, Daimeng Wei, Ling Zhang, Zongyao Li, Hengchao Shang, Zhiqiang Rao, Shaojun Li, Jinlong Yang, Zhanglin Wu, and Hao Yang. 2025b. Doc-guided sent2sent++: A sent2sent++ agent with doc-guided memory for document-level machine translation. *Preprint*, arXiv:2501.08523.

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.

Hanghai Hong, Yibo Xie, Jiawei Zheng, and Xiaoli Wang. 2025. Subdoctrans: Enhancing document-level machine translation with plug-and-play multi-granularity knowledge augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14490–14506.

Hanxu Hu, Jannis Vamvas, and Rico Sennrich. 2025. Source-primed multi-turn conversation helps large language models translate documents. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23702–23712.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is ChatGPT a good translator? yes with GPT-4 as the engine. *Preprint*, arXiv:2301.08745.

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.

Paweł Ma, Yusuf Can Semerci, Jan Scholtes, and Gerasimos Spanakis. 2025. You are what you train: Effects of data composition on training context-aware machine translation models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27414–27437.

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.

Qwen Team. 2026. Qwen3.6-35B-A3B: Agentic coding power, now open to all.

Miguel Moura Ramos, Patrick Fernandes, Sweta Agrawal, and André F. T. Martins. 2025. Multilingual contextualization of large language models for document-level machine translation. *Preprint*, arXiv:2504.12140. COLM 2025.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

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.

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.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661.

Yutong Wang, Jiali Zeng, Xuebo Liu, Derek Wong, Fandong Meng, Jie Zhou, and Min Zhang. 2025. Delta: An online document-level translation agent based on multi-level memory. In *International Conference on Learning Representations*, volume 2025, pages 15708–15731.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation. *Preprint*, arXiv:2401.06468.

Yangjian Wu and Gang Hu. 2023. Exploring prompt engineering with GPT language models for document-level machine translation: Insights and findings. In *Proceedings of the Eighth Conference on Machine Translation*, pages 166–169.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Hinrich Schütze, Volker Tresp, and Yunpu Ma. 2025. Memory-r1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. *Preprint*, arXiv:2508.19828.

Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiying Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, and Hao Zhou. 2025. Memagent: Reshaping long-context llm with multi-conv rl-based memory agent. *Preprint*, arXiv:2507.02259.

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*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

Xiying Zhao, Zhoufutu Wen, Zhixuan Chen, Jingzhe Ding, Jianpeng Jiao, Shuai Li, Xi Li, Danni Liang, Shengda Long, Qianqian Liu, Xianbo Wu, Hongwan Gao, Xiang Gao, Liang Hu, Jiashuo Liu, Mengyun Liu, Weiran Shi, Chenghao Yang, Qianyu Yang, and 4 others. 2025. Discox: Benchmarking discourse-level translation task in expert domains. *Preprint*, arXiv:2511.10984.

Mao Zheng, Zheng Li, Tao Chen, Bo Lv, Mingrui Sun, Mingyang Song, Jinlong Song, Hong Huang, Decheng Wu, Hai Wang, Yifan Song, Yanfeng Chen, and Guanwei Zhang. 2026. Hy-mt2: A family of fast, efficient and powerful multilingual translation models in the wild. *Preprint*, arXiv:2605.22064.

Mao Zheng, Zheng Li, Tao Chen, Mingyang Song, and Di Wang. 2025a. Hy-mt1.5 technical report. *Preprint*, arXiv:2512.24092.

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. 2025b. Hunyuan-mt technical report. *Preprint*, arXiv:2509.05209.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2023. Memorybank: Enhancing large language models with long-term memory. *Preprint*, arXiv:2305.10250.

A Appendix

A.1 Detailed IWSLT2017 Results

Tables 4 and 5 provide the per-language IWSLT2017 results that underlie the averaged scores in Table 1. Table 4 reports En⇒Zh, En⇒De, En⇒Fr, and En⇒Ja results, while Table 5 reports the reverse Xx⇒En directions. The best and second-best scores within each backbone block are marked in **bold** and <u>underlined</u>, respectively.

A.2 Prompt Templates

This appendix presents the prompt templates used by NoBAT. Figure 3 shows the Stage 1 prompt, which asks the LLM to translate a coarse text segment while extracting proper-noun translation pairs under historical noun constraints and preceding bilingual context. Figure 4 shows the Stage 2 prompt for sentence-level translation, where the model receives only the current source sentence, the previous j source sentences, and retrieved Noun Bank entries. Figure 5 shows the disambiguation prompt used when a source proper noun receives conflicting target translations; the LLM selects the more appropriate canonical translation based on accumulated bilingual contexts.

A.3 Qualitative Disambiguation Case

Table 6 presents a Novels-XL example that illustrates how NoBAT resolves an inconsistent entity translation during Noun Bank construction. The same highlighted source entity is first translated as "Yan RUSHENG" and later as "Yan Rusheng"; the disambiguation step selects the canonical form and stores it in the Noun Bank, so subsequent translation units can retrieve and reuse the normalized entity translation.

A.4 Code Availability

During the review stage, our code is submitted together with the supplementary ZIP archive. Upon acceptance, we will publicly release the code.

Prompt for Stage 1 Noun Bank Construction

You are an {src_lang}-{tgt_lang} bilingual expert, responsible for two tasks: translating a {src_lang} text segment and annotating proper nouns. First, refer to the historical translations of proper nouns and the preceding bilingual text to ensure consistent proper noun translations and style, then translate the current {src_lang} text segment into {tgt_lang}; second, annotate all proper nouns in the {src_lang} source and their corresponding {tgt_lang} translations.

Historical translations of proper nouns:
{hist_info}

Preceding bilingual text:
<{src_lang} text> {src_context}
<{tgt_lang} text> {tgt_context}

Here are some examples for you:
Example 1:
<{src_lang} source>
美国国家航空航天局的开普勒任务已经发现了围绕着其他恒星的数千颗潜在的行星，这也表明了地球只是银河系中数十亿行星中的一颗。
<Translation>NASA's Kepler mission has discovered thousands of potential planets around other stars, indicating that Earth is but one of billions of planets in our galaxy.</Translation>
<Proper nouns>"美国国家航空航天局" - "NASA", "开普勒" - "Kepler", "地球" - "Earth"</Proper nouns>

Example 2:
<{src_lang} source>
我开车回到家，在哈尔滨的寒冬，大约午夜时分，我开车从城镇一边到另一边，去看望我的朋友张杰，门廊上的温度计显示零下40度——不需要知道是摄氏度还是华氏度，到了零下40度，两个温度显示都一样——天气非常冷。
<Translation>I had just driven home, it was around midnight in the dead of Harbin winter, I had been visiting my friend, Zhang Jie, across town, and the thermometer on the front porch read minus 40 degrees -- and don't bother asking if that's Celsius or Fahrenheit, minus 40 is where the two scales meet -- it was very cold.</Translation>
<Proper nouns>"哈尔滨" - "Harbin", "张杰" - "Zhang Jie", "摄氏度" - "Celsius", "华氏度" - "Fahrenheit"</Proper nouns>

Task 1: Translation
Translate the following {src_lang} source text to {tgt_lang} and output the result only (no explanation) within <Translation> </Translation>.

Task 2: Proper Noun Annotation
If there are no proper nouns, answer "N/A". Output the result within <Proper nouns> </Proper nouns>.

Now complete the two tasks without explanation:
<{src_lang} source> {src}

Figure 3: Prompt template for Stage 1 Noun Bank construction.

Prompt for Stage 2 Parallel Translation

You are an {src_lang}-{tgt_lang} bilingual expert, translating a very long {src_lang} document. Given the historical translation of some proper nouns and source texts preceding the current sentence, translate the current {src_lang} source sentence into {tgt_lang}. Please ensure that the translations of proper nouns in the source sentence are consistent with their historical translation, and the translation style remains consistent as well.

Historical translations of proper nouns:
{hist_info}

Preceding texts:
<{src_lang} text> {src_context}

Now translate the following {src_lang} source sentence to {tgt_lang}. Only output the translation result, do not explain.
<{src_lang} source> {src}

Figure 4: Prompt template for Stage 2 retrieval-based parallel translation.

System	sCOMET	dCOMET	LTCR-1	LTCR-1_f	sCOMET	dCOMET	LTCR-1	LTCR-1_f
	En ⇒ Zh				En ⇒ De			
NLLB	76.81	6.20	71.68	84.96	84.15	6.64	91.76	99.61
GOOGLE	78.46	5.76	89.45	92.36	80.23	5.78	93.55	99.19
`GPT-5.4`								
Sentence	84.77	6.68	82.13	87.97	85.53	6.95	93.18	99.24
Context	85.09	6.74	**92.27**	**95.97**	85.77	6.89	91.65	97.62
Doc2Doc	–	6.81	87.24	94.29	–	6.97	90.26	98.87
DELTA	85.34	**6.88**	83.80	89.84	85.79	**7.02**	92.19	**100.00**
SubDocTrans	–	–	–	–	–	–	–	–
NoBAT	**85.70**	**6.88**	88.85	94.93	**86.06**	6.99	**93.33**	98.88
`GPT-4o-mini`								
Sentence	82.13	6.43	78.04	91.89	81.41	6.39	90.70	98.84
Context	84.36	6.68	78.95	93.42	84.83	6.70	92.66	99.61
Doc2Doc	–	6.60	82.33	88.35	–	**6.90**	91.05	99.22
DELTA	84.94	**6.81**	85.52	91.72	**85.47**	6.79	92.19	**100.00**
SubDocTrans	**85.22**	6.80	86.09	94.04	85.00	6.71	**93.46**	99.23
NoBAT	84.69	6.77	**91.63**	**97.74**	84.80	6.74	89.25	98.51
`Qwen3.6-35B-A3B`								
Sentence	84.83	6.88	75.65	86.51	85.30	6.91	88.16	98.09
Context	85.19	6.87	77.89	89.45	85.37	6.85	87.59	96.61
Doc2Doc	–	7.02	73.54	84.19	–	**7.02**	85.66	97.35
DELTA	84.98	**7.06**	83.04	87.88	84.49	7.00	87.96	96.99
SubDocTrans	–	–	–	–	–	–	–	–
NoBAT	**85.39**	6.96	**85.12**	**92.08**	**85.44**	6.96	**90.26**	**98.50**
`DeepSeek-R1-Distill-Qwen-32B`								
Sentence	84.47	6.79	77.09	86.91	81.79	6.33	83.98	94.17
Context	83.89	6.78	76.60	85.11	81.24	6.32	85.32	93.58
Doc2Doc	–	6.92	79.58	87.54	–	**6.60**	87.35	96.33
DELTA	83.92	6.87	82.67	88.33	81.78	6.38	88.67	98.05
SubDocTrans	**85.09**	**7.00**	84.35	90.14	**83.22**	6.56	91.37	**98.43**
NoBAT	84.46	6.73	**93.24**	**97.10**	82.16	6.32	**92.80**	98.10
	En ⇒ Fr				En ⇒ Ja			
NLLB	85.35	6.23	87.84	89.86	82.12	6.37	46.94	53.06
GOOGLE	82.21	5.53	88.61	91.46	80.74	6.23	53.92	55.88
`GPT-5.4`								
Sentence	85.61	6.43	**94.28**	**94.92**	87.09	7.31	60.20	62.24
Context	86.45	6.46	90.00	91.23	87.21	7.31	**65.96**	**68.87**
Doc2Doc	–	6.48	92.98	94.51	–	7.40	56.30	59.66
DELTA	86.05	**6.54**	92.99	94.58	87.43	**7.44**	56.17	60.67
SubDocTrans	–	–	–	–	–	–	–	–
NoBAT	**86.51**	**6.54**	93.49	94.42	**87.55**	7.42	50.00	55.55
`GPT-4o-mini`								
Sentence	80.79	5.82	88.89	90.91	81.72	6.74	56.73	58.65
Context	85.10	6.14	90.52	92.81	84.84	7.09	57.89	62.11
Doc2Doc	-	6.24	**91.00**	94.00	-	7.25	57.78	60.00
DELTA	86.38	**6.28**	90.94	93.85	**86.61**	**7.32**	58.54	59.76
SubDocTrans	**86.50**	6.26	89.97	91.22	86.42	7.28	**63.04**	**65.22**
NoBAT	86.20	6.22	90.06	**95.83**	86.25	7.20	58.41	61.38
`Qwen3.6-35B-A3B`								
Sentence	85.60	6.36	92.76	93.71	86.74	7.35	57.57	59.84
Context	86.01	6.33	92.63	94.47	86.62	7.31	**68.18**	71.81
Doc2Doc	–	6.54	90.85	92.42	–	**7.52**	56.89	63.79
DELTA	84.97	**6.61**	**96.03**	**96.64**	86.83	7.51	66.40	**73.43**
SubDocTrans	–	–	–	–	–	–	–	–
NoBAT	**86.02**	6.47	94.24	95.15	**86.89**	7.38	61.01	64.40
`DeepSeek-R1-Distill-Qwen-32B`								
Sentence	83.45	5.96	81.34	85.56	82.91	6.59	54.26	64.34
Context	83.42	6.03	84.12	89.89	82.99	6.65	**70.34**	**74.48**
Doc2Doc	–	**6.24**	88.50	92.65	–	7.03	58.67	65.33
DELTA	83.84	6.07	**94.95**	**97.16**	83.61	6.71	61.11	63.49
SubDocTrans	**84.93**	6.23	84.59	92.46	**84.77**	**7.05**	62.81	72.73
NoBAT	84.48	6.03	90.59	94.04	83.54	6.60	68.21	71.31

Table 4: Translation quality on TED (IWSLT17 tst2017), En⇒xx directions.

System	sCOMET	dCOMET	LTCR-1	LTCR-1_f	sCOMET	dCOMET	LTCR-1	LTCR-1_f
	Zh ⇒ En				De ⇒ En			
NLLB	82.14	7.01	75.31	88.27	85.63	7.23	95.98	98.85
GOOGLE	78.06	5.68	72.89	86.14	82.31	6.47	96.53	98.27
`GPT-5.4`								
Sentence	84.59	7.35	80.22	92.65	86.53	7.24	95.20	99.40
Context	85.36	7.33	83.42	93.14	86.79	7.26	**99.41**	**100.00**
Doc2Doc	–	7.38	80.11	93.75	–	7.32	95.37	97.68
DELTA	**85.47**	**7.44**	83.79	96.08	**86.97**	**7.36**	98.83	**100.00**
SubDocTrans	–	–	–	–	–	–	–	–
NoBAT	85.04	7.43	**85.71**	**96.57**	86.57	7.32	98.84	**100.00**
`GPT-4o-mini`								
Sentence	83.55	7.24	71.93	84.80	85.11	7.17	98.20	**100.00**
Context	83.96	7.35	78.24	91.18	86.12	7.27	98.88	**100.00**
Doc2Doc	–	7.15	79.62	92.36	–	7.19	95.24	97.62
DELTA	84.10	**7.47**	79.41	94.71	**86.61**	7.31	98.32	**100.00**
SubDocTrans	**84.33**	7.41	**83.15**	**95.51**	86.20	**7.35**	98.84	**100.00**
NoBAT	83.92	7.38	80.68	94.88	86.31	7.29	**99.40**	**100.00**
`Qwen3.6-35B-A3B`								
Sentence	84.35	7.33	78.03	87.28	86.34	7.24	98.83	98.83
Context	**84.74**	7.37	81.60	92.52	**86.67**	7.28	97.12	99.42
Doc2Doc	–	**7.51**	80.00	91.76	–	**7.39**	94.21	98.26
DELTA	84.10	7.24	83.05	94.35	86.48	7.38	97.67	**100.00**
SubDocTrans	–	–	–	–	–	–	–	–
NoBAT	84.69	7.42	**87.42**	**96.00**	86.44	7.35	**99.42**	99.42
`DeepSeek-R1-Distill-Qwen-32B`								
Sentence	82.76	7.22	60.12	82.66	84.42	7.18	94.57	98.45
Context	83.11	7.30	68.79	83.24	84.70	7.24	95.71	96.43
Doc2Doc	–	7.42	81.14	90.86	–	**7.36**	97.01	98.80
DELTA	83.00	7.31	83.62	94.35	85.42	7.30	98.19	99.40
SubDocTrans	**84.22**	**7.50**	**85.14**	**96.00**	85.74	7.34	**98.83**	**99.42**
NoBAT	83.57	7.27	82.02	90.44	**85.92**	7.22	97.63	98.81
	Fr ⇒ En				Ja ⇒ En			
NLLB	87.59	6.79	93.56	97.42	81.02	6.90	51.27	78.48
GOOGLE	84.64	6.19	95.63	96.83	75.67	5.50	60.67	82.00
`GPT-5.4`								
Sentence	88.22	6.85	95.95	97.97	82.70	7.13	68.26	92.21
Context	88.51	6.85	**98.01**	**100.00**	83.22	7.17	68.86	92.21
Doc2Doc	–	6.90	95.93	98.47	–	7.20	66.86	88.37
DELTA	**88.64**	**6.96**	97.53	**100.00**	**83.47**	**7.27**	73.68	**95.32**
SubDocTrans	–	–	–	–	–	–	–	–
NoBAT	88.37	6.94	97.50	99.50	82.94	**7.27**	**77.19**	**95.32**
`GPT-4o-mini`								
Sentence	87.32	6.77	94.42	97.85	80.04	6.76	61.11	82.72
Context	87.72	6.81	94.85	97.42	82.00	7.17	65.62	88.75
Doc2Doc	–	6.83	94.42	98.28	–	6.86	64.71	85.29
DELTA	88.13	**6.90**	96.58	98.72	82.20	7.29	66.67	90.12
SubDocTrans	**88.15**	6.88	97.99	**100.00**	**82.38**	**7.33**	65.34	89.77
NoBAT	88.02	6.87	**98.49**	**100.00**	82.09	7.23	**72.72**	**93.75**
`Qwen3.6-35B-A3B`								
Sentence	87.95	6.83	94.47	95.97	82.16	7.13	62.35	84.70
Context	**88.25**	6.83	**98.99**	**99.49**	**82.53**	7.22	52.09	76.64
Doc2Doc	–	**6.98**	97.48	98.99	–	**7.39**	60.35	85.79
DELTA	88.05	6.90	96.48	98.99	81.64	7.24	67.66	**91.01**
SubDocTrans	–	–	–	–	–	–	–	–
NoBAT	88.01	6.91	97.47	98.98	82.22	7.25	**68.02**	89.53
`DeepSeek-R1-Distill-Qwen-32B`								
Sentence	86.75	6.78	94.33	96.91	80.40	6.91	52.35	71.76
Context	86.85	6.83	96.43	**99.49**	80.74	7.08	66.06	87.27
Doc2Doc	–	6.90	95.92	98.47	–	7.32	58.43	80.72
DELTA	87.07	6.84	**97.56**	99.02	80.72	7.09	65.06	86.14
SubDocTrans	**87.80**	**6.92**	92.83	98.44	**81.77**	**7.35**	67.65	90.59
NoBAT	87.46	6.80	97.47	**99.49**	81.06	6.99	**72.67**	**92.44**

Table 5: Translation quality on TED (IWSLT17 tst2017), xx⇒En directions.

Prompt for Noun Bank Disambiguation

You are a {src_lang}-{tgt_lang} bilingual expert. A proper noun has two conflicting translations. Based on all the bilingual contexts below where this noun appears, choose the more appropriate translation.

Proper noun ({src_lang}): {src_ent}
Translation A: {old_tgt}
Translation B: {new_tgt}

Contexts:
{contexts}

Reply with ONLY the chosen translation (A or B's content), nothing else.

Figure 5: Prompt template for Noun Bank disambiguation.

	Earlier occurrence	Later occurrence
Source	第1898章 不思量，自难忘（四）言如生甩给她一个白眼，“温栩栩，你自恋了。”…	…看着栩栩脸红害羞，言如生即兴奋又懊恼，兴奋是他体内荷尔蒙自动反应，懊恼是法子他内心。
Translation	Chapter 1898: Without Thinking, It's Hard to Forget (Part 4) Yan RUSHENG gave her a white eye, “Wen Xuxu, you're being self-absorbed.” …	…Seeing Xuxu blush and feel shy, Yan Rusheng felt both excited and annoyed—excitement from his hormones reacting automatically, annoyance stemming from his inner self.
Disambiguation	言如生: Yan RUSHENG → Yan Rusheng	

Table 6: A disambiguation case from Novels-XL. The Noun Bank normalizes an inconsistent entity translation and stores the canonical form for later occurrences.