Cross-Preference Learning for Sentence-Level and Context-Aware Machine Translation

Anonymous ACL submission

Abstract

Context-aware machine translation (MT) leverages document-level information, yet it does not consistently outperform sentence-level MT, as contextual signals are unevenly beneficial across sentences. Existing training objectives do not explicitly model this variability, limiting a model's ability to adaptively exploit context. In this paper, we propose Cross-Preference Learning (CPL), a preference-based training framework that explicitly captures the complementary benefits of sentence-level and context-aware MT. CPL achieves this by integrating both intra- and cross-condition preferences into the preference optimization objective. The introduction of intra- and cross-condition preferences provides explicit supervision on when and how contextual information improves translation quality. We validate the proposed approach on several public context-aware MT tasks using multiple models, including Qwen3-4B, Qwen3-8B, and Llama-3-8B. Experimental results demonstrate consistent improvements in translation quality and robustness across both input conditions, achieved without any architectural modifications. All data and code will be released.

1 Introduction

Modern machine translation (MT) systems are increasingly expected to operate under multiple input conditions. In some scenarios, only a single source sentence is available, while in others, additional inter-sentence context can be provided to improve translation quality (e.g., document-level or context-aware MT) (Tiedemann and Scherrer, 2017; Maruf et al., 2021). These two settings—sentence-level and context-aware translation—share the same output space but are typically handled by separate models and optimized independently.

A key observation underlying our work is that context-aware translation does not consistently outperform sentence-level translation. As illustrated in Figure 1, context-aware models greatly outperform sentence-level models for a subset of sentences, but also slightly underperform, or even greatly underperform for a non-negligible portion of the data, while performing on par in many cases. This distribution highlights that the two systems often produce diverse yet complementary translations rather than one uniformly dominating the other (Voita et al., 2018; Dong et al., 2025). Consistent with this observation, Table 1 shows that the two systems achieve comparable overall BLEU and COMET scores, while the oracle results indicate substantial headroom when the better output is selected on a per-sentence basis. These findings highlight a key limitation of existing approaches: current MT systems lack an explicit mechanism to decide *when* to rely on context. Instead, models are trained under a fixed input condition, implicitly assuming that more context is always beneficial. This leads to suboptimal behavior when context is noisy, redundant, or even misleading.

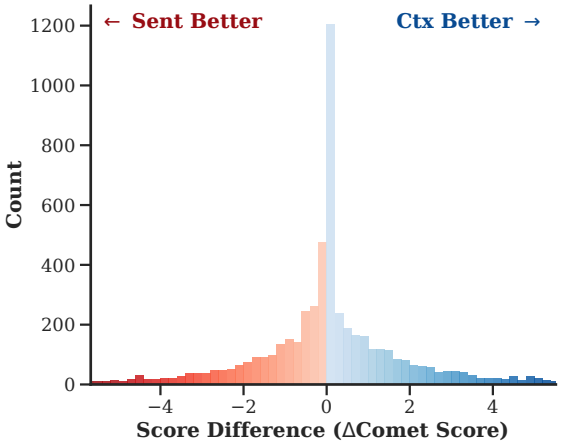

Figure 1: Distribution of COMET differences between sentence-level (Sent) and context-aware (Ctx) English–German translations produced by Qwen3-8B.

Inspired by recent advances in preference optimization, we cast the preference for using inter-

Model	BLEU	COMET
Sentence-level	31.72	86.12
Context-aware	31.60	86.13
Oracle	33.91	87.17

Table 1: Overall English-to-German translation quality of sentence-level and context-aware translation. *Oracle* means we always select the better translation from the two systems for each sentence.

sentence context in translation as a preference optimization problem, addressed via preference learning. However, existing methods such as direct preference optimization (DPO) (Rafailov et al., 2023), contrastive preference optimization (CPO) (Xu et al., 2024), and simplified preference optimization (SimPO) (Meng et al., 2024) focus on single-condition settings, where preference comparisons share the same input configuration. Consequently, these approaches are not designed to account for multiple input conditions sharing the same output space but differ in available information.

To address this limitation, we propose Cross-Preference Learning (CPL) built upon CPO. In addition to the intra-condition preference learning in vanilla CPO, CPL introduces a cross-condition preference optimization (Cross-CPO) to explicitly model the interactions between the two conditions. By ranking sentence-level and context-aware translations, CPL enables the model to learn a shared preference structure across conditions, allowing preference signals from one condition to influence learning in the other. Importantly, CPL operates entirely at the level of the training objective and requires no architectural changes, and can be applied to any shared-parameter translation system.

The main contributions in this paper can be summarized as follows:

- We extend preference-based optimization to jointly support sentence-level and context-aware translation, providing a principled framework for multi-task preference learning.
- We introduce CPL, which incorporates cross-condition preference optimization to explicitly model interactions between the two translation conditions and adaptively leverage context.
- We evaluate CPL on multiple document-level MT tasks and show that a single trained model performs robustly under both sentence-level and context-aware settings, achieving significant improvements in translation quality and cross-condition consistency.

2 Methodology

Figure 2 provides an overview of Cross-Preference Learning (CPL). We first introduce the two translation settings in this work (Figure 2(a)) in Section 2.1, then describe our proposed CPL (Figure 2(b)) in Section 2.2.

2.1 Sentence- and Context-Conditioned Translation

We study a machine translation setting in which a single model operates under two source-side input conditions: sentence-level and context-aware translation. In the former, the model receives only a source sentence x, while in the latter it additionally receives source-side context c, such as neighboring sentences within a document.

The model, parameterized by θ, defines two conditional distributions over target sentences: $p_\theta(y \mid x)$ for sentence-level translation and $p_\theta(y \mid x, c)$ for context-aware translation. For each distribution, we sample two candidates: $\{y_{\mathrm{s}}^1, y_{\mathrm{s}}^2\}$ from $p_\theta(y \mid x)$ and $\{y_{\mathrm{c}}^1, y_{\mathrm{c}}^2\}$ from $p_\theta(y \mid x, c)$.

For each candidate pair, we construct preference labels based on automatic evaluation. Taking sentence-level translation as an example, we score each candidate using the average of sentence-level COMET and document-level COMET (d-COMET) (Vernikos et al., 2022) against the reference. The higher-scoring candidate is treated as the preferred translation y_{s}^+, while the lower-scoring one is treated as the dispreferred translation y_{s}^-. We apply the same procedure to the context-aware candidates to obtain the corresponding preference pair $(y_{\mathrm{c}}^+, y_{\mathrm{c}}^-)$.

Since translations produced under the two conditions are directly comparable, this setup enables preference modeling both within and across conditions. Our goal is to train a single model that performs robustly under both settings.

2.2 Cross-Preference Learning

Next, we present the background for preference-based optimization in Section 2.2.1, followed by intra-condition preference learning (Section 2.2.2) and cross-condition preference learning (Section 2.2.3). Finally, we present the overall CPL objective in Section 2.2.4.

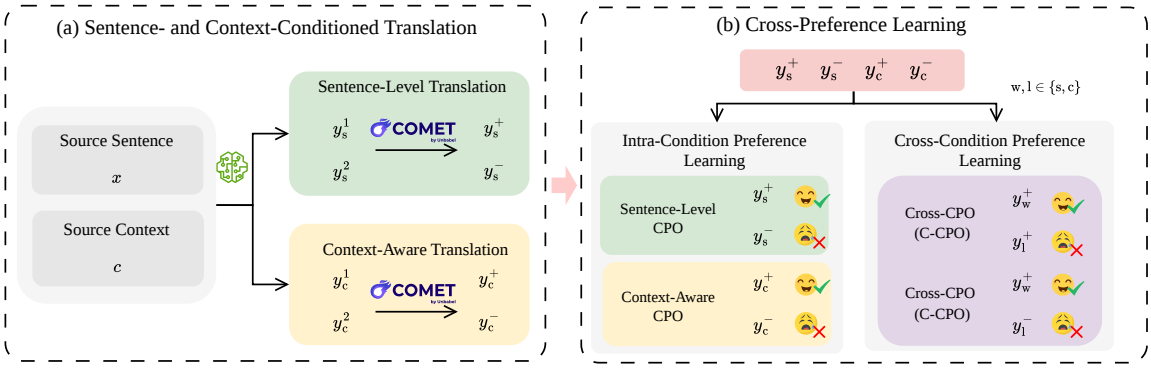

Figure 2: Illustration of our approach.

2.2.1 Preference-Based Optimization Background

Preference-based optimization trains models on pairwise comparisons of candidate translations rather than reference likelihood. For a preference pair (y^+, y^-) and input x, the model is encouraged to assign higher probability to y^+. When context is available, we treat x as (x, c) implicitly.

We adopt contrastive preference optimization (CPO) (Xu et al., 2024). The intra-condition CPO loss for a pair (y^+, y^-) is:

$$\mathcal{L}_{\text{CPO}}(x, y^+, y^-) = -\big[\log \sigma\big(\beta \log p_\theta(y^+ \mid x) - \beta \log p_\theta(y^- \mid x)\big)\big] - \log p_\theta\left(y^+ \mid x\right), \quad (1)$$

where σ is the sigmoid function, β controls preference separation.

2.2.2 Intra-Condition Preference Learning

We extend contrastive preference optimization (CPO), originally formulated for a single input condition, to the sentence- and context-conditioned translation. This extension enables preference-based optimization to be applied consistently across the two related tasks within a shared-parameter model.

CPO is applied independently to the selected preference pairs for each condition: (y_s^+, y_s^-) for sentence-level translation under $p_\theta(y \mid x)$, and (y_c^+, y_c^-) for context-aware translation under $p_\theta(y \mid x, c)$. Let $\mathcal{P}_s$ and $\mathcal{P}_c$ denote the selected sets of sentence-level and context-aware preference pairs, respectively. Intra-condition preference learning minimizes the following objective:

$$\mathcal{L}_{\text{intra}}(\theta) = \mathbb{E}_{(y^+,y^-)\sim\mathcal{P}_s}\left[\mathcal{L}_{\text{CPO}}\left(x, y^+, y^-\right)\right] + \mathbb{E}_{(y^+,y^-)\sim\mathcal{P}_c}\left[\mathcal{L}_{\text{CPO}}\left((x, c), y^+, y^-\right)\right], \quad (2)$$

where $\mathcal{L}_{\text{CPO}}$ is the CPO loss defined in Eq. 1, computed over sentence-level and context-aware preference pairs, respectively.

The intra-condition loss $\mathcal{L}_{\text{intra}}(\theta)$ enables effective preference-based learning within each translation condition while sharing model parameters across both tasks. Appendix A presents the details of preference pair selection for both $\mathcal{P}_s$ and $\mathcal{P}_c$.

2.2.3 Cross-Condition Preference Learning

Intra-condition preference learning optimizes sentence-level and context-aware translation independently, and therefore does not explicitly relate the two conditioned outputs. To exploit interactions across conditions, we introduce cross-condition preference learning, which directly compares sentence-level and context-aware translations for the same source sentence.

For each source sentence x with context c, we identify the best translation y_w^+ from four candidates, i.e., (y_s^+, y_s^-) and (y_c^+, y_c^-). The input condition that produces y_w^+ is defined as the winning condition (w), while the other serves as the losing condition (l), where $\text{w}, \text{l} \in \{\text{s}, \text{c}\}$. Then we construct two cross-condition preference pairs: (y_w^+, y_l^+) and (y_w^+, y_l^-). These pairs encourage the model to consistently prefer the winning-condition output over both strong and weak outputs from the losing condition. We denote the corresponding inputs as x_w^+, x_l^+, and x_l^-.

Let $\mathcal{P}_{\text{cr}}$ denote the set of cross-condition preference pairs. The cross-condition preference loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{cross}}(\theta) \quad (3)\\ = &\mathbb{E}_{(y_w^+,y_l^+)\sim\mathcal{P}_{\text{cr}}}\mathcal{L}_{\text{C-CPO}}\left(\{x_w^+, x_l^+\}, y_w^+, y_l^+\right)\\ + &\mathbb{E}_{(y_w^+,y_l^-)\sim\mathcal{P}_{\text{cr}}}\mathcal{L}_{\text{C-CPO}}\left(\{x_w^+, x_l^-\}, y_w^+, y_l^-\right),\end{aligned}$$

where the cross-CPO (C-CPO) loss is defined as:

$$\mathcal{L}_{\text{C-CPO}}(\{x^+, x^-\}, y^+, y^-) \quad (4)$$
$$= -\big[\log \sigma\big(\beta \log p_\theta(y^+ \mid x^+) - \beta \log p_\theta(y^- \mid x^-)\big)\big] - \log p_\theta(y^+ \mid x^+).$$

As shown, cross-CPO adapts the standard CPO objective to allow comparisons between translations generated under different conditioning signals while preserving the contrastive learning objective. Overall, the cross-condition loss $\mathcal{L}_{\text{cross}}(\theta)$ encourages the model to assign higher probability to preferred translations across input conditions, explicitly modeling interaction between sentence-level and context-aware outputs. Appendix A presents the details of preference pair selection for $\mathcal{P}_{\text{cr}}$.

2.2.4 CPL objective

CPL unifies intra- and cross-condition preference signals into a single objective, jointly optimizing sentence-level and context-aware translation under different input conditions.

The CPL objective is defined as the sum of the corresponding preference-based losses defined in Eq. 2 and Eq. 3:

$$\mathcal{L}_{\text{CPL}}(\theta) = \mathcal{L}_{\text{intra}}(\theta) + \mathcal{L}_{\text{cross}}(\theta). \quad (5)$$

By jointly optimizing intra-condition and cross-condition preference pairs, CPL enables a single shared-parameter model to perform robustly under both sentence-level and context-aware translation settings, without requiring architectural changes.

3 Experimentation

3.1 Experimental Settings

Datasets. Following recent work, we use the News Commentary v18.1 from WMT25,[1] which features parallel text with document boundaries. Our experiments cover six translation directions: from English (En) to German (De), Spanish (Es), French (Fr), Italian (It), Dutch (Nl), and Russian (Ru). Refer to Appendix B for more details of the datasets.

Similar to Xu et al. (2024), we constructed a small-scale dataset for the cold-start fine-tuning stage. For each direction, we extract a contiguous block of 20,000 sentences sequentially, ensuring the availability of context for context-aware training. Specifically, for the context-aware samples, we prepend the preceding source text, truncated to the last 256 tokens (Lyu et al., 2024), as the input context.

Following Xu et al. (2024) and Wang et al. (2025a), we use the high-quality validation sets to source the preference pairs. Appendix B shows the statistics of datasets used for CPL learning and the preference pairs used.

Models and Settings. We select Qwen3-4B[2] (Yang et al., 2025), Qwen3-8B[3] and Meta Llama-3-8B-Instruct[4] (AI@Meta, 2024) as the foundation open-source LLMs. For detailed fine-tuning and hyper-parameter settings, please refer to Appendix C.

Baselines. We compare our approach against the following baselines:

- **Sent-Level.** The base LLM translates each sentence independently. **Sent-Level$_{\text{tuned}}$** is fine-tuned on sentence-level data. Building on this model, we additionally apply CPO using sentence-level preference pairs, yielding **Sent-Level$_{\text{CPO}}$**.
- **Ctx-Aware.** The base LLM translates using the current sentence and its preceding context. **Ctx-Aware$_{\text{tuned}}$** is fine-tuned on context-aware data. Similarly, **Ctx-Aware$_{\text{CPO}}$** applies CPO with context-aware preference pairs.
- **Sent+Ctx$_{\text{tuned}}$.** A naive multi-condition baseline jointly fine-tuned on both sentence-level and context-aware data using standard likelihood training. **Sent+Ctx$_{\text{tuned+}}$** further incorporates the dataset used for extracting CPL preference pairs during fine-tuning.

All prompt details are provided in Appendix I.

Metrics. We primarily report *sentence-level COMET* scores (Rei et al., 2020). Specifically, we use the reference-based metric `wmt22-comet-da` (Rei et al., 2022a). In addition, we report supplementary evaluation metrics, including sentence-level BLEU (s-BLEU) (Post, 2018) and document-level COMET (d-COMET), in Appendix D. Since our model and most baselines support both input conditions, we evaluate and report performance for both sentence-level (*Sent*) and context-aware (*CTX*) translation.

[1] https://www2.statmt.org/wmt25/translation-task.html

[2] https://huggingface.co/Qwen/Qwen3-4B

[3] https://huggingface.co/Qwen/Qwen3-8B

[4] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

#	Model	En ⇒ De		En ⇒ Es		En ⇒ Fr		En ⇒ It		En ⇒ Nl		En ⇒ Ru		*Average*	
		Sent	*CTX*	*Sent*	*CTX*	*Sent*	*CTX*	*Sent*	*CTX*	*Sent*	*CTX*	*Sent*	*CTX*	*Sent*	*CTX*
							Qwen3-4B								
1	Sent-Level	82.51	-	84.91	-	83.13	-	86.94	-	82.71	-	83.12	-	83.89	-
2	Sent-Level$_{\text{tuned}}$	84.41	-	86.88	-	84.62	-	87.97	-	84.88	-	84.58	-	85.56	-
3	Sent-Level$_{\text{CPO}}$	85.14	-	87.43	-	85.18	-	88.66	-	85.88	-	86.19	-	86.41	-
4	Ctx-Aware	-	83.25	-	86.98	-	84.47	-	87.81	-	83.47	-	83.86	-	84.97
5	Ctx-Aware$_{\text{tuned}}$	-	84.31	-	86.85	-	84.64	-	88.01	-	84.85	-	84.86	-	85.59
6	Ctx-Aware$_{\text{CPO}}$	-	85.38	-	87.56	-	85.32	-	88.78	-	85.91	-	86.20	-	86.52
7	Sent+Ctx$_{\text{tuned}}$	84.66	84.79	87.07	87.02	84.60	84.61	88.18	88.27	85.16	85.08	85.04	84.95	85.79	85.79
8	Sent+Ctx$_{\text{tuned+}}$	84.58	84.60	87.11	87.12	84.74	84.78	88.39	88.39	85.31	85.32	84.88	84.83	85.84	85.84
9	CPL (Ours)	85.92	85.91	87.75	87.86	85.84	85.82	89.08	89.06	86.52	86.57	86.50	86.56	86.94	86.96
10	- IntraLoss	85.53	85.53	87.63	87.63	85.56	85.52	88.92	88.95	86.20	86.21	86.09	85.97	86.66	86.64
11	- CrossLoss	85.58	85.66	87.59	87.73	85.54	85.66	88.81	88.91	86.46	86.46	86.41	86.49	86.73	86.82
							Qwen3-8B								
1	Sent-Level	85.13	-	87.59	-	85.25	-	88.70	-	85.93	-	85.76	-	86.39	-
2	Sent-Level$_{\text{tuned}}$	86.12	-	87.73	-	85.82	-	89.12	-	87.07	-	86.34	-	87.03	-
3	Sent-Level$_{\text{CPO}}$	86.26	-	88.20	-	86.14	-	89.44	-	87.67	-	87.33	-	87.51	-
4	Ctx-Aware	-	85.05	-	87.72	-	85.35	-	88.75	-	86.09	-	85.59	-	86.43
5	Ctx-Aware$_{\text{tuned}}$	-	86.13	-	87.78	-	85.80	-	88.99	-	87.11	-	86.39	-	87.03
6	Ctx-Aware$_{\text{CPO}}$	-	86.54	-	88.29	-	86.26	-	89.47	-	87.60	-	87.32	-	87.58
7	Sent+Ctx$_{\text{tuned}}$	86.14	86.24	87.82	87.86	85.71	85.78	89.16	89.22	87.30	87.33	86.51	86.47	87.11	87.15
8	Sent+Ctx$_{\text{tuned+}}$	86.24	86.34	87.81	87.94	85.80	85.89	89.25	89.23	87.31	87.47	86.46	86.29	87.15	87.19
9	CPL (Ours)	87.19	87.23	88.47	88.51	86.56	86.63	89.81	89.99	88.22	88.28	87.71	87.75	87.99	88.06
10	- IntraLoss	86.85	86.85	88.34	88.35	86.38	86.34	89.68	89.67	87.86	87.83	87.41	87.26	87.75	87.72
11	- CrossLoss	86.88	87.00	88.33	88.39	86.39	86.54	89.74	89.74	87.97	87.99	87.35	87.43	87.78	87.85
							Llama-3-8B-Instruct								
1	Sent-Level	82.72	-	84.40	-	83.07	-	86.11	-	85.20	-	82.15	-	83.94	-
2	Sent-Level$_{\text{tuned}}$	86.00	-	87.59	-	85.44	-	88.99	-	88.22	-	85.82	-	87.01	-
3	Sent-Level$_{\text{CPO}}$	86.49	-	88.10	-	85.92	-	89.55	-	88.63	-	86.71	-	87.57	-
4	Ctx-Aware	-	51.33	-	56.62	-	50.41	-	58.05	-	53.51	-	53.05	-	53.83
5	Ctx-Aware$_{\text{tuned}}$	-	86.10	-	87.67	-	85.54	-	89.10	-	88.28	-	85.89	-	87.10
6	Ctx-Aware$_{\text{CPO}}$	-	86.72	-	88.15	-	86.19	-	89.56	-	88.69	-	86.77	-	87.68
7	Sent+Ctx$_{\text{tuned}}$	86.42	86.56	87.84	87.85	85.76	85.85	89.28	89.31	88.61	88.62	86.40	86.17	87.39	87.39
8	Sent+Ctx$_{\text{tuned+}}$	86.43	86.51	87.92	87.93	85.95	85.97	89.42	89.37	88.74	88.75	86.26	86.23	87.45	87.46
9	CPL (Ours)	87.09	87.19	88.43	88.45	86.60	86.65	89.82	89.85	88.93	89.00	87.42	87.43	88.05	88.10
10	- IntraLoss	87.00	87.03	88.29	88.32	86.44	86.47	89.69	89.77	88.79	88.79	87.16	86.94	87.90	87.89
11	- CrossLoss	86.98	87.11	88.17	88.16	86.41	86.42	89.69	89.71	88.94	88.95	87.35	87.39	87.92	87.96

Table 2: Results of different systems on COMET metric. Best scores are shown in dark red and second-best scores are in light red . *Sent* and *CTX* denote sentence-level and context-aware inputs, respectively.

3.2 Main Results

We begin with an important observation from Table 2. Across all models and training settings, sentence-level and context-aware translations (#2 vs. #5) achieve very similar COMET scores. This suggests that simply providing document-level context does not consistently improve average translation quality. Such behavior has also been observed in recent work, where the benefits of context are often sparse and localized (Dong et al., 2025; Lyu et al., 2024). This observation motivates our approach: rather than assuming context is always beneficial, we aim to adaptively learn when sentence-level or context-aware translation should be preferred. Based on the table, we further analyze the results as follows:

- **Effect of fine-tuning and joint training.** Among the baselines, fine-tuned LLMs consistently outperform their untuned counterparts (e.g., #2 vs. #1 and #5 vs. #4). Joint fine-tuning on both sentence- and context-conditioned data further improves performance for both tasks (#7 vs. #2 and #7 vs. #5). For example, with Llama-3-8B-Instruct, joint training yields average COMET gains of 0.38 and 0.29 over individually fine-tuned models, indicating that we establish a strong and competitive baseline.

- **Impact of incorporating additional training data.** Extending joint fine-tuning with the subset of data used for CPL (Sent+Ctx$_{\text{tuned+}}$, #8) leads to only marginal improvements, suggesting that performance gains are not simply driven by additional data, but require more effective learning signals. Moreover, using the additional training data for preference learning, both Sent-Level$_{\text{CPO}}$ (#3) and Ctx-Aware$_{\text{CPO}}$ (#6) achieve more improvement for both sentence-level and context-aware translation, respectively.

- **Effectiveness of CPL.** Compared with all baselines, CPL (#9) achieves substantial improvements. Notably, when compared to Sent+Ctx$_{\text{tuned+}}$ (#8), which leverages the same

training data, CPL yields consistent gains for both sentence-level and context-aware translation. For instance, CPL improves average COMET by 1.10/1.12, 0.84/0.87, and 0.60/0.64 for sentence-level/context-aware translation using Qwen3-4B, Qwen3-8B, and Llama-3-8B-Instruct, respectively. This demonstrates the effectiveness of Cross-Preference Learning beyond conventional fine-tuning. Moreover, CPL (#9) outperforms the two other strong baselines Sent-Level$_{\text{CPO}}$ (#3) and Ctx-Aware$_{\text{CPO}}$ (#6) which are augmented with preference learning on the same training data.

- **Ablation studies.** Ablation results (#9 vs. #10 vs. #11) show that both intra-condition loss and cross-condition loss contribute to CPL's performance. Removing either component degrades performance, with intra-condition loss having a slightly larger impact, indicating its complementary role in stabilizing preference learning.

- **Consistency across model scales.** Finally, the observed performance trends are consistent across all three LLM backbones, suggesting that CPL is robust and model-agnostic, and can be effectively applied to different pre-trained models.

4 Discussion

In this section, we use Qwen3-8B as the foundation model to discuss our approach. Unless otherwise specified, we report the average sentence-level COMET scores across all six translation tasks. We also provide case studies and other related discussions in Appendix E to G.

4.1 Does CPL Learn When to Use Context?

We examine whether CPL can exploit useful context while remaining robust to noisy context. We first consider the En⇒De test set from NewsCommentary and ask GPT-5 (GPT-5-2025-08-07) (OpenAI, 2025) to label whether the source-side context is helpful for translating each sentence. We retain 696 out of 5,932 sentences whose context is labeled as *High*, indicating that the context is highly necessary for accurate translation.[5] To further validate the findings, we also analyze the En⇒De test set from IWSLT 2017,[6] where contextual information tends to play a larger role due to stronger discourse dependencies. Using the same labeling procedure, we select 2,183 sentences out of 7,945 for evaluation.

Dataset	Sent	Ctx-gold	Ctx-random
NewsCommentary	82.50	82.71	82.52
IWSLT	78.06	78.70	78.02

Table 3: Performance (document-level COMET) on sentences with useful context. *Ctx-random* replaces the original context with randomly sampled one.

Table 3 reports the results. When the original context is used (Ctx-gold), context-aware CPL consistently outperforms sentence-level CPL with 0.21 and 0.64 d-COMET improvement on the two test sets, indicating that the model can effectively leverage informative context. When the context is replaced with randomly sampled sentences from other documents to simulate noisy context (Ctx-random), the performance of context-aware CPL becomes nearly identical to that of sentence-level translation. This trend holds across both datasets, with a larger gain on IWSLT, where contextual dependencies are stronger. Overall, these results suggest that CPL can benefit from informative context while remaining robust when contextual signals are unreliable.

4.2 Analysis with Additional Quality Metrics

In addition to BLEU and COMET, we evaluate CPL using complementary metrics that target aspects of translation quality not fully captured by standard accuracy-based evaluation, namely discourse coherence, faithfulness, and fluency. These properties are particularly important for context-aware translation.

Following Li et al. (2023) and Dong et al. (2025), we measure discourse coherence (Coh.) by computing the cosine similarity between SimCSE sentence embeddings (Gao et al., 2021) of neighboring translated sentences. Higher scores indicate better consistency across sentence boundaries. To assess faithfulness, we report ALTI+ (Ferrando et al., 2022), which uses a pretrained NLLB model to detect under-translation and hallucination by analyzing attribution patterns during generation.[7] Higher

[5] See Appendix H for details of the labeling process.

[6] https://huggingface.co/datasets/IWSLT/iwslt2017/tree/main/data

[7] https://facebookresearch.github.io/stopes/docs/eval/alti

Model	Coh. ↑	ALTI+ ↑	Fluency ↑
Qwen3-8B	73.28	54.49	3.81
Sent+Ctx$_{\text{tuned}}$	74.78	54.67	3.90
CPL (ours)	75.24	54.90	4.04

Table 4: Additional quality analysis using coherence (Coh.), ALTI+, and GPT-based fluency.

ALTI+ scores correspond to fewer faithfulness issues. Finally, we follow Sun et al. (2025b) and evaluate translation fluency using GPT-5 (GPT-5-2025-08-07) (OpenAI, 2025) as an external judge , which provides a holistic assessment of grammaticality and naturalness.[8] Among them, both Coh. and Fluency are document-level metrics as document-level context is needed in evaluation.

Table 4 reports results on complementary quality metrics. Sent+Ctx$_{\text{tuned}}$ improves coherence, faithfulness, and fluency over the base model, indicating more effective use of document context. CPL further achieves the best performance across all metrics, with higher discourse coherence, fewer faithfulness issues as measured by ALTI+, and improved fluency. These consistent gains suggest that CPL enhances translation quality beyond standard COMET evaluation.

4.3 Evaluation with GPT-as-Judge

To obtain a complementary quality assessment beyond automatic metrics, we conduct an evaluation using GPT-as-Judge. We randomly sample 1,200 sentences from the test set. For each instance, we provide GPT-5 with the source sentence, its source-side context, and the reference translation. We then present four candidate translations in random order: the sentence-level and context-aware outputs from both Sent+Ctx$_{\text{tuned}}$ and CPL (i.e., systems #7 and #9 in Table 2). The order of candidates is randomized to mitigate positional bias.

GPT-5 is asked to score each translation along four dimensions: *contextual coherence*, *fluency*, *faithfulness*, and *terminology expression*, each on a 0-25 scale, with the overall score computed as their sum. The evaluation prompt is provided in Figure 9 in Appendix J. We report the average overall scores across all evaluated sentences.

As shown in Table 5, CPL consistently outperforms Sent+Ctx$_{\text{tuned}}$ under both sentence-level and context-aware settings, with gains of +1.99 and +1.58, respectively. These improvements indicate that CPL enhances translation quality across multiple dimensions. Importantly, the gains are observed under both input conditions, suggesting that CPL provides overall quality improvements rather than benefiting only from traditional automatic metrics. Furthermore, the context-aware variant of CPL slightly outperforms its sentence-level counterpart, achieving the highest overall score. At the instance level, the context-aware CPL produces the best translation for 30.0% of the evaluated sentences, compared to 26.8% for sentence-level CPL.

[8]Due to the high cost of API-based fluency evaluation (about $50 per direction), we report only the average scores for En⇒De and En⇒Ru.

Model	Sent	CTX
Sent+Ctx$_{\text{tuned}}$	85.74	86.31
CPL	87.73	87.89

Table 5: Evaluation results of GPT-as-Judge.

Metric Used	Sent	CTX
BLEU	86.96	87.16
d-COMET	87.97	87.98
s-COMET	87.97	88.03
(d- + s-COMET)/2	87.99	88.06

Table 6: CPL performance (in COMET) using different metrics for selecting intra- and cross-condition preference pairs.

4.4 Effect of Preference Selection Metrics

Cross-Preference Learning relies on ranking candidate translations to construct informative intra-condition and cross-condition preference pairs. By default, we use the average of document-level COMET (d-COMET) and sentence-level COMET (s-COMET) as the selection metric. To assess the sensitivity of CPL to this design choice, we compare several alternatives, including d-COMET alone, s-COMET alone, and BLEU (Post, 2018).

Table 6 shows the performance. Among the COMET-based variants, performance differences are negligible, indicating that CPL is robust to different granularities of COMET scoring. In contrast, using BLEU as the selection metric leads to a noticeable performance drop. This is expected, as BLEU relies on surface-level n-gram overlap and is less aligned with semantic adequacy and discourse-level quality, which are crucial for selecting informative preference pairs.

4.5 Effect of Cross-Condition Decomposition

In cross-condition preference learning, we include two types of preference constraints: $(y_{\text{w}}^{+}, y_{\text{l}}^{+})$ and

Model	Sent	CTX
CPL	87.99	88.06
CPL w/o (y_w^+, y_l^+)	87.98	88.01
CPL w/o (y_w^+, y_l^-)	87.88	87.91

Table 7: Effect of removing individual cross-condition preference constraints.

(y_w^+, y_l^-). To analyze their respective contributions, we conduct an ablation study in which we remove one of the two preference pairs from the cross-condition loss.

As shown in Table 7, removing either constraint causes a small but consistent performance drop for both sentence-level and context-aware translation, indicating that both are beneficial. The degradation is slightly larger when removing (y_w^+, y_l^-), suggesting that contrasting against a weaker losing candidate provides a marginally stronger signal. These results suggest that the two cross-condition constraints play complementary roles and that CPL is not overly sensitive to any single preference relation. This redundancy contributes to the robustness of CPL, allowing CPL to maintain stable performance even when individual cross-condition signals are partially removed.

5 Related Work

LLM-based Document-level Machine Translation. With the emergence of LLMs, DocMT has been revisited by leveraging LLMs' strong contextual modeling and long-context capabilities. Early studies primarily evaluated prompt-based LLM performance on DocMT, analyzing how models such as ChatGPT handle extended source-side context (Wang et al., 2023; Karpinska and Iyyer, 2023). Subsequent work proposes more structured prompting and inference-time strategies, including retrieval-augmented generation (Cui et al., 2024), agent-based frameworks with memory mechanisms (Wang et al., 2025b; Guo et al., 2025), and iterative translation schemes that preserve discourse information across segments (Hu et al., 2025).

Another line of research focuses on supervised fine-tuning (SFT) to adapt LLMs explicitly for DMT. These methods explore instruction mixing (Li et al., 2024), multi-stage fine-tuning with monolingual and parallel documents (Wu et al., 2024a), sentence-level fine-tuning evaluated at the document level (Stap et al., 2024), and decoding- or prompt-enhanced training strategies (Lyu et al., 2024). Dong et al. (2025) refine translations by jointly leveraging intermediate outputs from sentence-level and document-level translation.

Different from above studies, in this work we explicitly address the interaction between sentence-level and context-aware translation outputs and provide mechanisms for selectively leveraging context.

Preference Learning for Machine Translation. Preference learning has long been explored in MT as an alternative to likelihood-based training (Hopkins and May, 2011; Cherry and Foster, 2012), using ranking-based or pairwise objectives to better align outputs with human judgments. Recent advances in LLM alignment, including RLHF (Christiano et al., 2017), Direct Preference Optimization (DPO; Rafailov et al., 2023), and Contrastive Preference Optimization (CPO; Xu et al., 2024), have renewed interest in this paradigm. Several studies adapt these techniques to MT by constructing preference pairs from human annotations or automatic metrics, showing improvements in translation adequacy and fluency (Agrawal et al., 2024; Wu et al., 2024b; Yang et al., 2024; Cui et al., 2025; Sun et al., 2025a). In contrast to prior work, which applies preference learning within a single translation setting, our approach extends preference learning to a dual-condition scenario by explicitly modeling interactions between sentence-level and context-aware translation.

6 Conclusion

In this paper, we study sentence-level and context-aware machine translation from a unified perspective and show that, while their overall performance is often similar, their outputs exhibit substantial and systematic differences. Motivated by this observation, we propose Cross-Preference Learning (CPL), a preference-based optimization framework that jointly models intra-condition and cross-condition translation preferences, enabling a single model to leverage document context adaptively. Experiments across multiple document-level translation tasks demonstrate that CPL consistently improves both sentence-level and context-aware translation quality, enhances cross-condition consistency, and remains robust to context noise and metric choices. Our analysis further shows that CPL provides a simple yet effective way to align translation behavior across input conditions without architectural changes, offering a promising direction for robust and context-adaptive machine translation.

Limitations

While Cross-Preference Learning (CPL) shows consistent improvements for both sentence-level and context-aware translation, several limitations remain. First, CPL relies on high-quality preference pairs derived from automatic metrics (e.g., COMET), which may propagate metric biases and affect learning. Second, our approach requires sampling multiple candidates per sentence and computing their scores, which increases computational cost during training. Third, although CPL can leverage context selectively, its ability to handle extremely long documents or complex discourse phenomena has not been fully explored. Finally, our evaluations focus on a limited set of language pairs and LLM backbones; further studies are needed to validate CPL on broader languages and model scales.

References

Sweta Agrawal, José G. C. De Souza, Ricardo Rei, António Farinhas, Gonçalo Faria, Patrick Fernandes, Nuno M Guerreiro, and Andre Martins. 2024. Modeling user preferences with automatic metrics: Creating a high-quality preference dataset for machine translation. In *Proceedings of EMNLP*, pages 14503–14519.

AI@Meta. 2024. Llama 3 model card.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of NAACL-HLT*, pages 427–436.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proceedings of NeurIPS*, pages 4302–4310.

Guofeng Cui, Pichao Wang, Yang Liu, Zemian Ke, Zhu Liu, and Vimal Bhat. 2025. CRPO: Confidence-reward driven preference optimization for machine translation. In *Findings of ACL*, pages 560–574.

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 ACL*, pages 10885–10897.

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 ACL*, pages 14917–14933.

Javier Ferrando, Gerard I. Gállego, Belen Alastruey, Carlos Escolano, and Marta R. Costa-jussà. 2022. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer. In *Proceedings of the EMNLP*, pages 8756–8769.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of EMNLP*, pages 6894–6910.

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. *CoRR*, abs/2501.08523.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Proceedings of EMNLP*, pages 1352–1362.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. In *Proceedings of ICLR*.

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

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of WMT*, pages 419–451.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of ACL*, pages 12286–12312.

Yachao Li, Junhui Li, Jing Jiang, and Min Zhang. 2024. Enhancing document-level translation of large language model via translation mixed-instructions. *CoRR*, abs/2401.08088.

Xinglin Lyu, Junhui Li, Yanqing Zhao, Min Zhang, Daimeng Wei, Shimin Tao, Hao Yang, and Min Zhang. 2024. DeMPT: Decoding-enhanced multi-phase prompt tuning for making LLMs be better context-aware translators. In *Proceedings of EMNLP*, pages 20280–20295.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation. *ACM Computing Surveys*, 54(2).

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. In *Proceedings ofNeurIPS*, pages 124198–124235.

OpenAI. 2025. Gpt-5 system card.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of WMT*, pages 186–191.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Proceedings of NeurIPS*, pages 53728–53741.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of WMT*, pages 578–585.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of EMNLP*, pages 2685–2702, Online.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of WMT*, pages 634–645.

David Stap, Eva Hasler, Bill Byrne, Christof Monz, and Ke Tran. 2024. The fine-tuning paradox: Boosting translation quality without sacrificing LLM abilities. In *Proceedings of ACL*, pages 6189–6206.

Haoxiang Sun, Ruize Gao, Pei Zhang, Baosong Yang, and Rui Wang. 2025a. Enhancing machine translation with self-supervised preference data. In *Proceedings of ACL*, pages 23916–23934.

Yirong Sun, Dawei Zhu, Yanjun Chen, Erjia Xiao, Xinghao Chen, and Xiaoyu Shen. 2025b. Fine-grained and multi-dimensional metrics for document-level machine translation. In *Proceedings of NAACL-HLT: Student Research Workshop*, pages 1–17.

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.

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 WMT*, pages 118–128.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of ACL*, pages 1264–1274.

Hao Wang, Linlong Xu, Heng Liu, Yangyang Liu, Xiaohu Zhao, Bo Zeng, Liangying Shao, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025a. Beyond single-reward: Multi-pair, multi-perspective preference optimization for machine translation. *arXiv preprint arXiv:2510.13434*.

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 EMNLP*, pages 16646–16661.

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 *Proceedings of ICLR*.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024a. Adapting large language models for document-level machine translation. *CoRR*, abs/2401.06468.

Qiyu Wu, Masaaki Nagata, Zhongtao Miao, and Yoshimasa Tsuruoka. 2024b. Word alignment as preference for machine translation. In *Proceedings of EMNLP*, pages 3223–3239.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: pushing the boundaries of llm performance in machine translation. In *Proceedings of ICML*, pages 55204–55224.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, and Others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Guangyu Yang, Jinghong Chen, Weizhe Lin, and Bill Byrne. 2024. Direct preference optimization for neural machine translation with minimum Bayes risk decoding. In *Proceedings of NAACL-HLT: Short Paper*, pages 391–398.

Task	Test		CPL	
	# Doc	# Sent	# Doc	# Sent
En ⇒ De	150	5,967	150	5,628
En ⇒ Es	150	5,819	150	5,782
En ⇒ Fr	150	5,795	150	5,890
En ⇒ It	150	5,749	150	5,948
En ⇒ Nl	151	5,992	151	5,925
En ⇒ Ru	150	5,619	150	5,691

Table 8: Statistics of test sets for the translation tasks.

Task	$\mathcal{P}_{\text{s}}$	$\mathcal{P}_{\text{c}}$	$\mathcal{P}_{\text{cr}}$
En ⇒ De	4,143	4,277	8,086
En ⇒ Es	4,963	5,051	8,736
En ⇒ Fr	4,768	4,877	8,829
En ⇒ It	4,923	5,024	8,877
En ⇒ Nl	3,904	4,072	8,056
En ⇒ Ru	4,016	4,116	7,841

Table 9: Statistics of preference pairs in cross-condition preference learning.

A Preference Pair Selection

We construct the sentence-level preference set $\mathcal{P}_{\text{s}}$ by filtering the preferred–dispreferred translation pairs $(y_{\text{s}}^{+}, y_{\text{s}}^{-})$ obtained in Section 2.1. A pair is retained only if it satisfies the following criteria: 1) both translations contain between 6 and 100 words; 2) both translations achieve quality scores higher than 0.3; and 3) the score margin between the two translations lies in the range $[0.2, 10]$.

These constraints filter out degenerate, low-quality, or weakly distinguishable pairs, ensuring that retained pairs provide reliable and informative preference signals. The same strategy is applied to construct the context-aware preference set $\mathcal{P}_{\text{c}}$.

Similarly, all preference pairs in the cross-preference set $\mathcal{P}_{\text{cr}}$ are required to satisfy the same selection criteria.

B Dataset

Table 8 shows the statistics of the test sets of six translation tasks, and the dataset used for CPL preference selection.

Table 9 shows the statistics of the preference datasets used in the six translation tasks. Specifically, $\mathcal{P}_{\text{cr}}$ denotes the total number of preference pairs, including both types of cross-condition preference pairs, $(y_{\text{w}}^{+}, y_{\text{l}}^{+})$ and $(y_{\text{w}}^{+}, y_{\text{l}}^{-})$.

C SFT and CPL Settings

During SFT, we fine-tune the model using the LoRA approach (Hu et al., 2021). The LoRA rank is set to 16. The batch size is 256, and the learning rate is fixed at 5×10^{-5}. We train the model for a single epoch.

During CPL, we also adopt LoRA-based fine-tuning with the LoRA rank set to 16. The batch size is 128, and the learning rate is set to 5×10^{-5}. The model is trained for 2 epochs. The β parameter is set to 0.1, following Rafailov et al. (2023) and Xu et al. (2024).

D Experimental Results in s-BLEU and d-COMET

Table 10 shows the performance in document-level COMET (Vernikos et al., 2022) using the reference-based `wmt22-comet-da` (Rei et al., 2022a). Table 11 shows the performance in sentence-level SacreBLEU (Post, 2018).

E Distributional Analysis of Context Utility

To better understand how Cross-Preference Learning affects the interaction between sentence-level and context-aware translation, we analyze the distribution of sentences for which one input condition substantially outperforms the other. Specifically, we categorize sentences into five groups based on the COMET score difference between the context-aware and sentence-level outputs, denoted as $\Delta = \text{COMET}_{\text{ctx}} - \text{COMET}_{\text{sent}}$. The categories are defined as follows: *clearly better* ($\Delta \geq 1.0$), *better* ($0.5 \leq \Delta < 1.0$), *on par* ($-0.5 < \Delta < 0.5$), *worse* ($-1.0 < \Delta \leq -0.5$), and *clearly worse* ($\Delta \leq -1.0$).

As shown in Figure 3, the untuned base model, Qwen3-8B, exhibits a large proportion of sentences in the *clearly better, CB* and *clearly worse, CW* categories, indicating that document context can substantially help some sentences while severely degrading others.

Joint fine-tuning (Sent+Ctx$_{\text{tuned}}$) mitigates this effect by increasing the proportion of *on par, P* cases, suggesting more balanced use of context. CPL further strengthens this trend, achieving the highest share of *on par, P* sentences while simultaneously reducing extreme improvements and degradations. Compared to the Qwen3-8B base model and Sent+Ctx$_{\text{tuned}}$, CPL produces more balanced and stable performance between sentence-level and context-aware translation, reducing extreme gains or degradations associated with context usage.

#	Model	En ⇒ De		En ⇒ Es		En ⇒ Fr		En ⇒ It		En ⇒ Nl		En ⇒ Ru		*Average*	
		Sent	*CTX*	*Sent*	*CTX*	*Sent*	*CTX*	*Sent*	*CTX*	*Sent*	*CTX*	*Sent*	*CTX*	*Sent*	*CTX*
							Qwen3-4B								
1	Sent-Level	79.00	-	81.29	-	79.50	-	83.26	-	79.35	-	79.23	-	80.27	-
2	Sent-Level$_{\text{tuned}}$	81.55	-	84.25	-	81.74	-	84.81	-	81.94	-	81.02	-	82.86	-
3	Sent-Level$_{\text{CPO}}$	82.40	-	84.90	-	82.37	-	85.57	-	82.98	-	82.73	-	83.49	-
4	Ctx-Aware	-	80.26	-	84.46	-	81.59	-	84.38	-	80.29	-	80.22	-	81.87
5	Ctx-Aware$_{\text{tuned}}$	-	81.52	-	84.26	-	81.84	-	84.85	-	81.95	-	81.38	-	82.88
6	Ctx-Aware$_{\text{CPO}}$	-	82.74	-	85.03	-	82.60	-	85.74	-	83.15	-	82.83	-	83.68
7	Sent+Ctx$_{\text{tuned}}$	81.78	82.00	84.44	84.46	81.70	81.83	85.04	85.20	82.24	82.20	81.53	81.43	83.04	83.14
8	Sent+Ctx$_{\text{tuned+}}$	81.72	81.85	84.49	84.55	81.87	81.96	85.29	85.37	82.43	82.53	81.31	81.26	82.85	82.92
9	CPL (Ours)	83.23	83.25	85.22	85.38	83.02	83.08	86.00	86.04	83.82	83.84	83.01	83.09	84.26	84.32
10	- IntraLoss	82.82	82.83	85.10	85.11	82.76	82.78	85.81	85.91	83.39	83.48	82.64	82.53	83.98	84.02
11	- CrossLoss	82.86	83.02	85.07	85.24	82.76	82.95	85.76	85.93	83.67	83.72	82.92	83.01	84.02	84.17
							Qwen3-8B								
1	Sent-Level	82.15	-	85.01	-	82.38	-	85.51	-	83.08	-	82.04	-	83.36	-
2	Sent-Level$_{\text{tuned}}$	83.37	-	85.17	-	82.99	-	86.06	-	84.37	-	82.88	-	84.14	-
3	Sent-Level$_{\text{CPO}}$	83.65	-	85.69	-	83.41	-	86.44	-	85.07	-	83.91	-	84.69	-
4	Ctx-Aware	-	82.29	-	85.15	-	82.59	-	85.56	-	83.40	-	81.99	-	83.50
5	Ctx-Aware$_{\text{tuned}}$	-	83.43	-	85.25	-	83.04	-	86.00	-	84.48	-	82.97	-	84.20
6	Ctx-Aware$_{\text{CPO}}$	-	83.95	-	85.79	-	83.56	-	86.53	-	85.03	-	83.97	-	84.81
7	Sent+Ctx$_{\text{tuned}}$	83.39	83.58	85.27	85.35	82.89	83.03	86.15	86.23	84.69	84.74	83.03	82.97	84.24	84.32
8	Sent+Ctx$_{\text{tuned+}}$	83.49	83.68	85.24	85.41	82.98	83.13	86.25	86.25	84.72	84.91	82.90	82.71	84.26	84.35
9	CPL (Ours)	84.63	84.69	86.00	86.05	83.82	83.92	86.78	87.05	85.69	85.84	84.32	84.36	85.21	85.32
10	- IntraLoss	84.23	84.32	85.85	85.88	83.64	83.64	86.69	86.69	85.32	85.33	83.98	83.87	84.95	84.95
11	- CrossLoss	84.29	84.44	85.83	85.91	83.68	83.85	86.84	86.81	85.42	85.45	83.98	84.02	85.01	85.08
							Llama-3-8B-Instruct								
1	Sent-Level	78.32	-	80.56	-	79.22	-	81.46	-	81.20	-	77.15	-	79.65	-
2	Sent-Level$_{\text{tuned}}$	83.26	-	85.07	-	82.54	-	85.93	-	85.63	-	82.21	-	84.11	-
3	Sent-Level$_{\text{CPO}}$	83.86	-	85.63	-	83.15	-	86.58	-	86.12	-	83.27	-	84.77	-
4	Ctx-Aware	-	44.93	-	50.38	-	44.08	-	51.81	-	46.62	-	45.98	-	47.30
5	Ctx-Aware$_{\text{tuned}}$	-	83.39	-	85.20	-	82.80	-	86.02	-	85.73	-	82.34	-	84.25
6	Ctx-Aware$_{\text{CPO}}$	-	84.14	-	85.67	-	83.47	-	86.61	-	86.20	-	83.38	-	84.91
7	Sent+Ctx$_{\text{tuned}}$	83.69	83.84	85.34	85.38	82.95	83.05	86.29	86.29	86.14	86.20	82.77	82.64	84.53	84.57
8	Sent+Ctx$_{\text{tuned+}}$	83.67	83.81	85.39	85.43	83.12	83.20	86.42	86.36	86.30	86.34	82.70	82.73	84.60	84.65
9	CPL (Ours)	84.44	84.59	85.98	86.04	83.84	83.95	86.83	86.89	86.49	86.58	83.99	83.99	85.26	85.34
10	- IntraLoss	84.41	84.45	85.79	85.84	83.62	83.76	86.63	86.76	86.34	86.37	83.65	83.54	85.07	85.12
11	- CrossLoss	84.42	84.56	85.65	85.71	83.69	83.75	86.68	86.73	86.48	86.55	83.89	83.98	85.14	85.21

Table 10: Results of different systems on document-level COMET metric. Best scores are shown in dark red and second-best scores are in light red . *Sent* and *CTX* denote sentence-level and context-aware inputs, respectively.

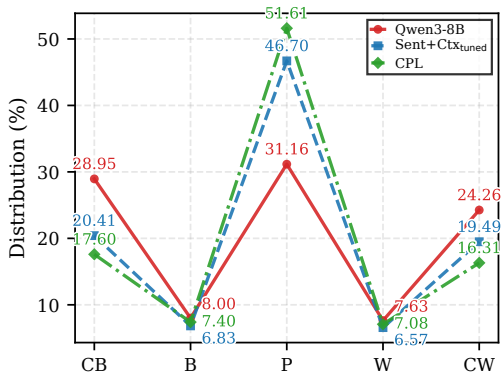

Figure 3: Distribution (%) of sentences by COMET differences between context-aware and sentence-level translations. CB/B/P/W/CW = *Clearly Better*, *Better*, *On Par*, *Worse*, *Clearly Worse*.

F Leveraging Both Outputs via Reranking

An additional advantage of CPL is that it naturally supports both sentence-level and context-aware translation within a unified framework. Given a source sentence x and its context c, CPL produces two candidate translations: a sentence-conditioned output y and a context-conditioned output y'. Rather than committing to a single translation mode, these two candidates can be reranked, and the better one selected as the final output.

We report the results in Table 12. The first two rows show the standalone performance of CPL's sentence-level and context-aware outputs. We then evaluate three reranking strategies. *Prob.-based Reranking* selects the candidate with higher model likelihood, reflecting the model's internal confidence. *COMETKiwi-based Reranking* uses a reference-free quality estimator (Rei et al., 2022b), enabling automatic selection without gold references. *Oracle Reranking* selects the candidate with higher reference-based COMET score and serves as an upper bound. As shown, all reranking strategies consistently improve over individual outputs, demonstrating that sentence-level and context-aware translations provide complementary strengths. Prob.-based reranking yields modest gains, while KiwiCOMET achieves larger improvements, suggesting that semantic quality estimation

#	Model	En ⇒ De		En ⇒ Es		En ⇒ Fr		En ⇒ It		En ⇒ Nl		En ⇒ Ru		*Average*	
		Sent	*CTX*	*Sent*	*CTX*	*Sent*	*CTX*	*Sent*	*CTX*	*Sent*	*CTX*	*Sent*	*CTX*	*Sent*	*CTX*
							Qwen3-4B								
1	Sent-Level	12.64	-	28.95	-	22.16	-	22.75	-	14.78	-	14.38	-	19.28	-
2	Sent-Level$_{\text{tuned}}$	24.92	-	40.05	-	31.94	-	35.38	-	24.38	-	22.20	-	31.33	-
3	Sent-Level$_{\text{CPO}}$	28.99	-	41.73	-	35.16	-	36.25	-	28.61	-	26.49	-	32.87	-
4	Ctx-Aware	-	11.09	-	28.98	-	21.21	-	21.35	-	12.72	-	13.05	-	18.07
5	Ctx-Aware$_{\text{tuned}}$	-	26.73	-	40.02	-	32.79	-	35.20	-	24.46	-	23.26	-	31.84
6	Ctx-Aware$_{\text{CPO}}$	-	27.71	-	41.29	-	34.94	-	35.14	-	30.30	-	24.99	-	32.40
7	Sent+Ctx$_{\text{tuned}}$	28.11	28.95	41.33	39.42	33.54	33.38	36.65	36.56	27.51	28.32	24.89	25.52	33.43	33.33
8	Sent+Ctx$_{\text{tuned+}}$	28.42	29.15	41.47	40.74	32.84	33.46	37.26	37.52	29.74	28.17	24.79	25.19	32.42	32.37
9	CPL (Ours)	29.38	29.46	42.03	41.99	35.50	35.63	35.87	35.32	33.16	33.24	27.38	27.39	35.19	35.13
10	- IntraLoss	29.10	30.03	41.64	41.75	35.11	35.27	36.00	36.69	32.71	32.80	26.30	27.18	34.91	35.31
11	- CrossLoss	30.56	29.98	41.87	41.87	34.89	34.97	36.37	36.01	32.91	32.21	26.78	26.80	35.32	35.01
							Qwen3-8B								
1	Sent-Level	16.95	-	35.96	-	25.91	-	27.32	-	18.39	-	18.41	-	23.82	-
2	Sent-Level$_{\text{tuned}}$	31.72	-	42.65	-	35.48	-	38.59	-	34.66	-	28.66	-	35.29	-
3	Sent-Level$_{\text{CPO}}$	33.47	-	43.79	-	37.40	-	38.90	-	35.01	-	30.12	-	36.45	-
4	Ctx-Aware	-	14.44	-	31.60	-	23.43	-	24.58	-	16.16	-	14.78	-	20.83
5	Ctx-Aware$_{\text{tuned}}$	-	31.60	-	42.55	-	35.44	-	38.44	-	34.14	-	28.61	-	35.13
6	Ctx-Aware$_{\text{CPO}}$	-	33.34	-	43.38	-	37.07	-	39.05	-	34.29	-	30.12	-	36.21
7	Sent+Ctx$_{\text{tuned}}$	31.91	31.88	43.57	43.62	36.06	35.96	39.77	39.62	35.58	35.75	29.19	28.83	36.01	35.94
8	Sent+Ctx$_{\text{tuned+}}$	32.38	32.42	43.65	43.86	36.22	36.12	40.34	40.42	34.87	36.39	29.14	28.72	36.10	36.32
9	CPL (Ours)	32.28	33.26	44.22	44.24	37.80	37.78	38.80	39.73	36.51	35.75	30.03	30.06	36.61	36.80
10	- IntraLoss	33.79	33.83	43.93	44.12	37.54	37.73	39.16	39.50	36.47	34.94	30.32	30.35	36.87	36.74
11	- CrossLoss	33.70	33.82	44.01	44.13	37.11	37.33	39.57	39.44	36.47	35.86	29.85	30.24	36.79	36.80
							Llama-3-8B-Instruct								
1	Sent-Level	26.29	-	35.30	-	31.16	-	31.59	-	30.37	-	24.33	-	29.84	-
2	Sent-Level$_{\text{tuned}}$	31.80	-	42.96	-	35.10	-	38.97	-	38.31	-	27.93	-	35.85	-
3	Sent-Level$_{\text{CPO}}$	33.34	-	44.15	-	37.51	-	39.93	-	38.47	-	29.39	-	37.13	-
4	Ctx-Aware	-	13.33	-	18.12	-	15.80	-	16.27	-	13.98	-	13.27	-	15.13
5	Ctx-Aware$_{\text{tuned}}$	-	32.11	-	43.20	-	35.32	-	39.26	-	38.50	-	28.14	-	36.09
6	Ctx-Aware$_{\text{CPO}}$	-	33.49	-	44.13	-	37.67	-	39.42	-	38.46	-	29.55	-	37.12
7	Sent+Ctx$_{\text{tuned}}$	32.96	33.15	44.40	44.34	36.27	36.55	40.33	40.37	40.13	40.29	28.74	28.65	37.14	37.23
8	Sent+Ctx$_{\text{tuned+}}$	32.94	33.32	44.36	44.48	36.59	36.58	40.67	40.82	40.55	40.59	28.91	28.75	37.34	37.42
9	CPL (Ours)	33.87	33.95	44.17	44.37	37.82	38.09	39.86	40.09	39.23	39.14	29.93	29.73	37.48	37.56
10	- IntraLoss	33.70	34.17	44.18	44.47	37.14	37.93	39.40	39.62	39.16	39.09	29.87	30.04	37.24	37.55
11	- CrossLoss	32.90	33.75	44.09	44.05	37.24	37.43	39.44	39.52	39.09	39.01	29.48	29.34	37.04	37.18

Table 11: Results of different systems on sentence-level BLEU metric. Best scores are shown in dark red and second-best scores are in light red. *Sent* and *CTX* denote sentence-level and context-aware inputs, respectively.

Model	BLEU	COMET
Sent-Level	36.61	87.99
Ctx-Aware	36.80	88.06
+ Prob.	37.14	88.19
+ COMETKiwi	37.13	88.42
+ Oracle	38.63	88.73

Table 12: Performance of reranking sentence-level and context-aware translations produced by CPL.

is more effective than likelihood alone.

G Case Study

Figure 4 presents two examples where CPL produces better translations than Sent+Ctx$_{\text{tuned}}$. In Case 1, both the sentence-level and context-aware variants of Sent+Ctx$_{\text{tuned}}$ incorrectly translate *delusions* as *Delirien*. In contrast, while the sentence-level variant of CPL makes a similar mistake, the context-aware variant correctly translates *delusions* as *Wahnvorstellungen*. This shows that the additional context helps CPL select the more appropriate translation instead of the less suitable *Delirium*. In Case 2, CPL again demonstrates more effective use of context than Sent+Ctxtuned. The sentence-level variant of Sent+Ctxtuned omits the discourse connective and is less well-formed, while its context-aware variant only partially improves the output by adding *Angesichts dieser Zahlen* but still uses semantically weaker terminology *Verfassungsschutzmaßnahmen* and an imprecise predicate *aufrechterhalten wollen*. In contrast, CPL shows a clear improvement with context: although its sentence-level variant already captures the correct predicate *unterstützen*, the context-aware variant further preserves the full structure of *support retaining* as *die Aufrechterhaltung ... unterstützen* and better captures the legal meaning associated with *Roe* as *verfassungsgemäße Schutzgarantie*. This highlights CPL's ability to leverage context for both semantic precision and structural fidelity.

H Identifying the Usefulness of Context

To identify test instances for which source-side context is genuinely useful, the prompt shown in Figure 5 specifically asks whether the current sen-

Case 1

[Source]
If a dissident demonstrated against the Soviet system at least twice - say, by distributing illicit writings - the "delusions" signified schizophrenia.

[Source-side Context]
Anyone who believed that the political system should be changed was suffering from "reformist delusions." . . . Soviet authorities relied on an idiosyncratic definition of schizophrenia . . .

[Sent+Ctx$_{\text{tuned}}$ (Sent)]
Wenn ein Dissident gegen das sowjetische System mindestens zweimal demonstrierte - etwa indem er illegale Schriften verbreitete - bedeutete das „Delirien“ Schizophrenie.

[Sent+Ctx$_{\text{tuned}}$ (Ctx)]
Wenn ein Dissident gegen das sowjetische System mindestens zweimal protestiert hatte - etwa durch die Verteilung illegaler Schriften - so wurden die „Delirien“ als Schizophrenie interpretiert.

[CPL (Sent)]
Wenn ein Dissident mindestens zweimal gegen das sowjetische System demonstriert hatte – etwa indem er illegale Schriften verbreitete –, so bedeutete das „Delirium“, dass er an Schizophrenie litt.

[CPL (Ctx)]
Wenn ein Dissident mindestens zweimal gegen das sowjetische System demonstrierte – etwa durch die Verbreitung illegaler Schriften –, dann wurden die „Wahnvorstellungen“ als Zeichen für Schizophrenie interpretiert.

[Reference]
Wenn sich ein Dissident mindestens zweimal gegen das sowjetische System äußerte, indem er beispielsweise verbotene Schriften in Umlauf brachte, dann deuteten die „Wahnvorstellungen" auf Schizophrenie hin.

Case 2

[Source]
Given these figures, it is no wonder that nearly two-thirds of Americans support retaining the constitutional protection that Roe provides.

[Source-side Context]
Between the Roe decision and 2017 · · · more than 58 million abortions were performed in the US. · · · one in four American women will have an abortion · · ·

[Sent+Ctx$_{\text{tuned}}$ (Sent)]
Dass fast zwei Drittel der Amerikaner die Verfassungsschutz durch Roe bewahren wollen, ist kein Wunder.

[Sent+Ctx$_{\text{tuned}}$ (Ctx)]
Angesichts dieser Zahlen ist es kein Wunder, dass fast zwei Drittel der Amerikaner die Verfassungsschutzmaßnahmen, die Roe bietet, aufrechterhalten wollen.

[CPL (Sent)]
Angesichts dieser Zahlen ist es kein Wunder, dass fast zwei Drittel der Amerikaner die Aufrechterhaltung der Verfassungsschutzgarantie unterstützen, die Roe bietet.

[CPL (Ctx)]
Angesichts dieser Zahlen ist es kein Wunder, dass fast zwei Drittel der Amerikanerinnen und Amerikaner die Aufrechterhaltung der verfassungsgemäßen Schutzgarantie, die Roe bietet, unterstützen.

[Reference]
Angesichts dieser Zahlen verwundert es nicht, dass zwei Drittel aller Amerikanerinnen und Amerikaner sich für den Bestand des verfassungsmäßigen Schutzes aussprechen, der seit Roe besteht.

Figure 4: Case studies illustrating that CPL benefits from informative context.

tence can still be translated accurately and naturally without access to the preceding source context, and whether the context provides necessary disambiguating information. The evaluation covers four dimensions: coreference, lexical ambiguity, discourse/logical relation, and specific referential or

background knowledge. Each dimension is scored from 0 to 3, and the total score is mapped to `Low`, `Medium`, or `High` context necessity.

To obtain consistent structured outputs at scale, we use batched API inference with a strict JSON schema (Figure 6). The schema requires the model to return the total score, the necessity label, a brief explanation of the main translation risk without context, and the most likely error that would arise if the context were ignored.

I Prompts

In all experiments presented in this paper, we employ a unified set of prompt templates to ensure fair comparison.

As illustrated in Figure 7, the sentence-level template consists of a system instruction and the user input, where the source sentence is inserted into the `{src_sent}` placeholder. In contrast, as shown in Figure 8, following Wang et al. (2023), the context-aware template augments the input by introducing a `{context_text}` placeholder before the source sentence, explicitly instructing the model to incorporate inter-sentence context during translation.

J GPT-as-Judge

To obtain a more comprehensive and fine-grained quality assessment, we design a detailed evaluation prompt for the judge model. The prompt explicitly specifies the evaluation inputs, scoring dimensions, scoring criteria, and required output format, enabling the model to assess each candidate translation in a more controlled and consistent manner across multiple aspects of quality. The full prompt is shown in Figure 9.

Prompt Template

You are a computational linguist with expertise in `{tgt_lang}`. Your task is to evaluate how strongly the translation of the current sentence depends on its preceding context in a machine translation training example.

Your job is **not** to translate the sentence. Instead, determine whether the **[current sentence]** can still be translated accurately and naturally into `{tgt_lang}` *without any context*, and then judge whether the provided **[context]** supplies *necessary disambiguating information*.

Please carefully distinguish the following two cases:
(1) The context only makes the translation more natural or complete, but the sentence can still be translated mostly correctly without it → **Low dependency**.
(2) Without the context, the sentence would likely be mistranslated due to unresolved ambiguity, incorrect discourse relation, wrong coreference resolution, or incorrect word sense selection → **Medium/High dependency**.

Evaluate the sentence on the following four dimensions, assigning a score from 0 to 3 for each:

1. Coreference Dependency
0: No pronoun/reference, or the reference is clear without context.
1: Minor referential uncertainty, but translation is unlikely to be seriously affected.
2: Context helps determine the referent; without it, the translation may be awkward or partially wrong.
3: Context is necessary to determine referent identity, gender, number, or grammatical role; omission would likely cause mistranslation.

2. Lexical Ambiguity
0: The key lexical meaning is clear.
1: Slight ambiguity, but the most likely translation remains stable.
2: Context clearly helps disambiguate the intended sense; without it, the translation may be inaccurate.
3: Context is necessary to identify the correct sense; omission would likely cause mistranslation.

3. Discourse / Logical Relation
0: The sentence is logically self-contained.
1: Weak discourse connection, but it does not materially affect translation.
2: Context helps determine tone or discourse relation such as contrast, causality, or continuation.
3: Without context, the discourse relation would likely be misinterpreted, affecting translation choices.

4. Specific Referential / Background Knowledge
0: No context-specific referent or background knowledge is needed.
1: Slight background dependence, but it does not materially affect translation.
2: Context helps identify a specific person, entity, event, or concept.
3: Context is necessary to identify the intended referent; omission would likely cause substantial misunderstanding.

The total score is the sum of the four dimension scores (0–12).
Map the total score to `necessity_level` as follows:
0–3 → `Low`
4–7 → `Medium`
8–12 → `High`

Output requirements:
- In `bottleneck_analysis.explanation`, briefly describe the most likely failure point if the context were unavailable.
- In `zero_shot_failure_case`, provide the most likely incorrect interpretation or mistranslation that would arise if the context were completely ignored.
- If there is essentially no obvious failure risk, output `"None"` for `zero_shot_failure_case`.

Return the result strictly following the provided schema, and do not output any additional text.

[Context]
`{context_text}`

[Current sentence]
`{src_sent}`

Figure 5: Prompt used for context-dependency evaluation of sentence-level translation.

Output JSON Schema

```
{
  "name": "context_dependency_eval_schema",
  "strict": true,
  "schema": {
    "type": "object",
    "additionalProperties": false,
    "properties": {
      "total_dependency_score": {
        "type": "integer",
        "minimum": 0,
        "maximum": 12
      },
      "necessity_level": {
        "type": "string",
        "enum": ["High", "Medium", "Low"]
      },
      "bottleneck_analysis": {
        "type": "object",
        "additionalProperties": false,
        "properties": {
          "has_grammatical_risk": {"type": "boolean"},
          "has_semantic_risk": {"type": "boolean"},
          "explanation": {"type": "string"}
        },
        "required": [
          "has_grammatical_risk",
          "has_semantic_risk",
          "explanation"
        ]
      },
      "zero_shot_failure_case": {"type": "string"}
    },
    "required": [
      "total_dependency_score",
      "necessity_level",
      "bottleneck_analysis",
      "zero_shot_failure_case"
    ]
  }
}
```

Figure 6: Strict JSON schema used to constrain model outputs in batched API inference.

Prompt Template

Translate the text enclosed in triple quotes from `{src_lang}` into `{tgt_lang}` without any explanation.

`{src_lang}`: ''' `{src_sent}` '''

Figure 7: Prompt used for Sentence-level translation (Sent-Level).

Prompt Template

[`{src_lang}`]:`{context_text}`

Given the provided context, translate the following `{src_lang}` sentence to `{tgt_lang}`.

[`{src_lang}`]:`{src_sent}`

Figure 8: Prompt used for Context-Aware translation (Ctx-Aware).

Prompt Template

You are an expert bilingual translation evaluator.

You will receive one JSON object describing a single translation evaluation task. Evaluate 4 candidate translations independently.

Input fields:
- `doc_id`: document id
- `sent_id`: sentence id
- `lang_pair`: translation direction, e.g. `en-de` means translating English into German
- `prompt_ctx`: broader contextual prompt or discourse context
- `prompt_sent`: sentence-level translation prompt
- `input`: source sentence
- `ref`: reference translation of the current sentence
- `candidates`: exactly four translations: `Candidate_A`, `Candidate_B`, `Candidate_C`, `Candidate_D`

Evaluation principles:
- Use `ref` only as a helpful reference for the intended meaning. Multiple valid translations may exist. Do not penalize harmless variation if meaning, contextual fit, and fluency are preserved.

For each candidate, consider:
1. `input`
2. `prompt_sent`
3. `prompt_ctx`
4. `ref` (Use `ref` only as a helpful reference for the intended meaning. Multiple valid translations may exist.)
5. the candidate itself

Score each candidate on 4 dimensions, 25 points each (total 100):

1. Contextual Appropriateness (25)
Judge whether the translation is fully appropriate given `prompt_ctx`, including topic, discourse framing, tone/register, communicative setting, and context-dependent interpretation. Does it correctly resolve ambiguities, pronouns, gender agreements, referential links, or terminology that are dictated by the surrounding context?
- 22–25: fully context-appropriate
- 16–21: mostly appropriate, with minor contextual mismatch
- 0–15: contextually inappropriate, inconsistent, or fails to resolve context-dependent meaning correctly

2. Faithfulness and Accuracy (25)
Judge whether the candidate accurately conveys the meaning of `input`, using `ref` as a strong anchor.
- 22–25: highly accurate
- 16–21: mostly accurate, minor deviations
- 0–15: major semantic errors, omissions, additions, or distortions

3. Fluency and Naturalness (25)
Judge grammaticality, idiomaticity, readability, and naturalness in the target language.
- 22–25: fluent and natural
- 16–21: understandable but somewhat awkward
- 0–15: noticeable language problems

4. Terminology and Expression Quality (25)
Judge lexical choice, phrasing, precision, and stylistic suitability, especially for key expressions.
- 22–25: precise and well-phrased
- 16–21: acceptable but somewhat suboptimal
- 0–15: weak, awkward, or inappropriate wording

Output requirements:
- Score each candidate and provide one concise overall reasoning.
- In contextual reasoning, explicitly mention how `prompt_ctx` affects the judgment.
- Ensure `total` equals the sum of the four dimension scores.
- Return the result strictly following the provided JSON schema without extra text.

Figure 9: Prompt used for GPT-as-Judge.